Continuous Optimization

Chapter 3: Constrained Optimization

1 Definitions

In this chapter we will consider constrained optimization problems with the following shape

$$\begin{aligned} \min \;& f(x) \\ \text{s.t.} \;& x \in C \end{aligned} \tag{1}$$

Definition 1.1 (Convex Set). *A set C is said to be convex if given $x_1, x_2 \in C$ and $\lambda \in [0,1]$, then $\lambda x_1 + (1-\lambda)x_2 \in C$.*

Definition 1.2 (Convex Function). *A function $f : C \to \mathbb{R}$ defined on a convex set C is said to be convex if given $x_1, x_2 \in C$ and $\lambda \in [0,1]$, then*

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

Definition 1.3 (Strictly Convex Function). *A function $f : C \to \mathbb{R}$ defined on a convex set C is said to be strictly convex if given $x_1, x_2 \in C$ and $\lambda \in [0,1]$, then*

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2).$$

A function is called concave if $-f$ is convex and strictly concave if $-f$ is strictly convex.
Now, given Δ_k the unit-simplex, that is the subset of $\mathbb{R}^k$ comprising all nonnegative vectors whose sum is 1, i.e.,

$$\{\lambda \in \mathbb{R}^k : \lambda \ge 0, e^t \lambda = 1\},$$

we can provide the following very useful result by Jensen's.

Theorem 1.1 (Jensen's Inequality). *Let $f : C \to \mathbb{R}$ be a convex function over a convex set C. Then for any $x_1, x_2, \ldots, x_k \in C$ and $\lambda \in \Delta_k$ we have*

$$f\left(\sum_{i=1}^k \lambda_i x_i\right) \le \sum_{i=1}^k \lambda_i f(x_i). \tag{2}$$

Proof. We will prove (2) by induction on k. For $k = 1$ the result is obvious ($f(x_1) \le f(x_1) \;\; \forall x_1 \in C$). We now assume that (2) holds for k and we will prove that it also holds for $k+1$. Suppose we have $x_1, x_2, \ldots, x_{k+1} \in C$ and $\lambda \in \Delta_{k+1}$, we will show that $f(z) \le \sum_{i=1}^{k+1} \lambda_i f(x_i)$ with $z = \sum_{i=1}^{k+1} \lambda_i x_i$. If $\lambda_{k+1} = 1$, then $z = x_{k+1}$ and (2) is obvious. If $\lambda_{k+1} < 1$, then

$$\begin{aligned} f(z) &= f\left(\sum_{i=1}^k \lambda_i x_i + \lambda_{k+1} x_{k+1}\right) \\ &= f\left((1-\lambda_{k+1}) \sum_{i=1}^k \frac{\lambda_i}{1-\lambda_{k+1}} x_i + \lambda_{k+1} x_{k+1}\right) \\ &\le (1-\lambda_{k+1}) f(v) + \lambda_{k+1} f(x_{k+1}), \end{aligned}$$

with $v = \sum_{i=1}^k \frac{\lambda_i}{1-\lambda_{k+1}} x_i$. Since $\sum_{i=1}^k \frac{\lambda_i}{1-\lambda_{k+1}} = \frac{1-\lambda_{k+1}}{1-\lambda_{k+1}} = 1$, it follows that v is a convex combination of k points from C, hence by the induction hypotesis we have that $f(v) \le \sum_{i=1}^k \frac{\lambda_i}{1-\lambda_{k+1}} f(x_i)$, which combined with the equality above yields

$$f(z) \le \sum_{i=1}^{k+1} \lambda_i f(x_i).$$

□

2 Characterizations of Convex Functions

Theorem 2.1 (Gradient characterization of convex functions)**.** *Let $f \in \mathrm{C}^1(C)$, where C is convex. Then f is convex over C if and only if*

$$f(x) + \nabla f(x)^T (y - x) \leq f(y) \quad \forall x, y \in C. \tag{3}$$

Proof. Exercise. □

Proposition 2.1 (Sufficiency of stationarity under convexity)**.** *Let $f \in \mathrm{C}^1(C)$, where $C \subseteq \mathbb{R}^n$ is convex. Suppose that $\nabla f(x^*) = 0$ for some $x^* \in C$. Then x^* is a global minimizer of f over C.*

Proof. Let $z \in C$. Plugging $x = x^*$ and $y = z$ in Theorem 2.1 we obtain that

$$f(z) \geq f(x^*) + \nabla f(x^*)^T (z - x^*),$$

which implies that $f(z) \geq f(x^*)$ because $\nabla f(x^*) = 0$. □

We note that Proposition 2.1 establishes only the sufficiency of the stationarity condition $\nabla f(x^*) = 0$ for guaranteeing that x^* is a global optimal solution. When C is not the entire space, this condition is not necessary, in fact it might be that the points for which $\nabla f(\cdot) = 0$ are not in C. On the other hand, when $C = \mathbb{R}^n$ and f is convex, $\nabla f(x^*) = 0$ is both sufficient and necessary condition for x^* to be a global minimum. We can now establish the conditions under which a twice continuously differentiable function f is convex.

Theorem 2.2 (Second order characterization of convexity)**.** *Let $f \in \mathrm{C}^2(C)$, where $C \subseteq \mathbb{R}^n$ is convex and open. Thus, we have that f is convex iff $\nabla^2 f(x) \succcurlyeq 0 \quad \forall x \in C$.*

Proof. Suppose that $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in C$. We will prove (3) which is enough to establish convexity. Let $x, y \in C$, then by the Mean Value Theorem[2] (Theorem 2.6 from Chapter 1) we get that there exists $z \in [x, y]$ (and hence $z \in C$) for which

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x). \tag{4}$$

Since $\nabla^2 f(z) \succcurlyeq 0$, it follows that $(y - x)^T \nabla^2 f(z)(y - x) \geq 0$, which implies (3). To prove the opposite direction, assume that f is convex over C. Let $x \in C$ and $y \in \mathbb{R}^n$. Since C is open, it follows that $x + \lambda y \in C$, for $0 < \lambda < \epsilon$, where ϵ is a small enough positive constant. Using now the gradient characterization of convex functions (3) we get

$$f(x + \lambda y) \geq f(x) + \lambda \nabla f(x)^T y.$$

In addition, by the quadratic approximation theorem (Theorem 2.4 from Chapter 1), we have that

$$f(x + \lambda y) = f(x) + \lambda \nabla f(x)^T y + \frac{\lambda^2}{2} y^T \nabla^2 f(x) y + o(\lambda^2 ||y||^2),$$

which combined with the above inequality gives

$$\frac{\lambda^2}{2} y^T \nabla^2 f(x) y + o(\lambda^2 ||y||^2) \geq 0 \quad \forall \lambda \in (0, \epsilon).$$

Dividing the latter inequality by λ^2 and taking the limit for $\lambda \to 0^+$, we have

$$y^T \nabla^2 f(x) y \geq 0 \quad \forall y \in \mathbb{R}^n,$$

which concludes the proof. □

The same theorem works with positive definiteness and strict convexity, meaning also that the minimum in this case is unique.

3 Optimization over convex constraints

From now on, we consider (1) where C is convex. On the other hand, we will not always assume also f to be convex. From the convexity of f we have the following two theorems. Notice that the following result is not a direct consequence of Proposition 2.1 as the local (and global) minimum, might be on the boundary of the set and not be stationary (in the sense of unconstrained optimization).

Theorem 3.1 (global=local in convex optimization)**.** *Let $f : C \to \mathbb{R}$ be a convex function over a convex set $C \subseteq \mathbb{R}^n$. Let $x^* \in C$ be a local minimum of f over C. Then x^* is a global minimum of f over C.*

Proof. Since x^* is a local minimum of f over C there exists r such that $f(x) \geq f(x^*)$ for any $x \in C \cap B[x^*, r]$. Now let $y \in C$ with $y \neq x^*$. We want to show that $f(y) \geq f(x^*)$. Let $\lambda \in (0, 1]$ be such that $x^* + \lambda(y - x^*) \in B[x^*, r]$, for instance $\lambda = \frac{r}{||y-x^*||}$. Now, since $x^* + \lambda(y - x^*) \in B[x^*, r] \cap C$, it follows that $f(x^*) \leq f(x^* + \lambda(y - x^*))$, and hence, by convexity of f, also

$$f(x^*) \leq f(x^* + \lambda(y - x^*)) \leq (1 - \lambda)f(x^*) + \lambda f(y)$$

Thus, $\lambda f(x^*) \leq \lambda f(y)$, which concludes the proof. □

Theorem 3.2 (Convexity of the optimal set in convex optimization)**.** *Let $f : C \to \mathbb{R}$ be a convex function with $C \subseteq \mathbb{R}^n$ convex. Then, the set of optimal solutions of the problem* (1)*, which we denote by X^* is convex. Moreover, if f is strictly convex over C, then there exists at most one optimal solution.*

Proof. If $X^* = \emptyset$, the result follows trivially. Suppose that $X \neq \emptyset$ and denote the optimal value of f by f^*. Let $x, y \in C$ with $\lambda \in [0, 1]$. Then, by convexity $f(\lambda x + (1 - \lambda)y) \leq \lambda f^* + (1 - \lambda)f^* = f^*$, hence $\lambda x + (1 - \lambda)y$ is also optimal, i.e., it belongs to X^*, establishing the convexity of X^*. Suppose now that f is strictly convex and X^* is nonempty, and suppose by contradiction that there are 2 points x, y in X^*. Then $\lambda x + (1 - \lambda)y \in C$, and by the strict convexity of f we have

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) = f^*,$$

which is a contradiction to the fact that f^* is the optimal value. □

3.1 Stationarity

Note that the following definition and the following theorem are given also for the more general case in which f is not convex.

Definition 3.1 (Stationary points of convex constrained problems)**.** *Let $f \in \mathrm{C}^1(C)$, where C is closed and convex. Then x^* is a stationary point of* (1) *if*

$$\nabla f(x^*)(x - x^*) \geq 0 \ \forall x \in C. \tag{5}$$

In words, this means that there are no feasible descent directions of f at x^*. This suggests that stationarity is in fact a necessary condition for a local minimum of (1).

Theorem 3.3 (Stationarity as necessary optimality condition of a convex constrained problem)**.** *Let $f \in \mathrm{C}^1(C)$, where C is closed and convex and let x^* be a local minimum of* (1)*. Then x^* is a stationary point of* (1)*.*

Proof. Let x^* be a local minimum of f and assume by contradiction that is not a stationary point of (1). Then there exists $x \in C$ such that $\nabla f(x^*)(x - x^*) < 0$. Therefore, $f'(x, d) < 0$, where $d = x - x^*$. Hence, by Lemma 1.1 of Chapter 2, there exists $\epsilon \in (0, 1)$ such that $f(x^* + td) < f(x^*) \ \forall t \in (0, \epsilon)$. Since C is convex, we have that $x^* + td = (1 - t)x^* + tx \in C$, leading to the conclusion that x^* is not a local optimum of (1), which is a contradiction. □

Theorem 3.4 (Stationarity as necessary and sufficient optimality condition for a convex problem)**.** *Let $f \in \mathrm{C}^1(C)$, where C is closed and convex and f is also convex. Let x^* be a local minimum of* (1)*. Then x^* is a stationary point of* (1) *iff x^* is a optimal solution of* (1)*.*

Proof. The necessity of the stationarity condition follows from Theorem 3.3. To prove the sufficiency, assume that x^* is a stationary point of (1) and let $x \in C$. Then, the gradient characterization of convex functions (3) and stationarity of x^*, we get

$$f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*) \geq f(x^*),$$

which concludes the proof. □

Unfortunately, (5) is not an easy condition to check, we need something else.

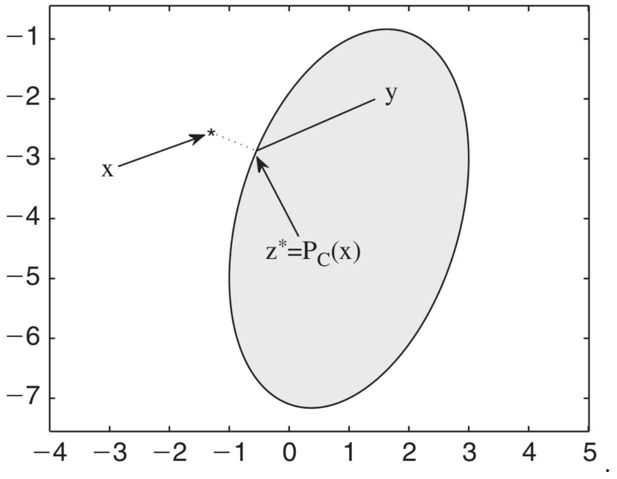

Figure 1: The orthogonal projection operator.

3.2 Orthogonal Projection

We can instead characterize stationary points by using the projection operator. Given a nonempty closed convex set C, the orthogonal projection operator $P_C : \mathbb{R}^n \to C$ is defined by

$$P_C(x) = \text{argmin}\left\{||x - y||^2 : y \in C\right\} \tag{6}$$

The orthogonal projection operator with input x returns the vector in C that is the closest (in ℓ_2-norm) to x. Note that the orthogonal projection operator is defined as a solution of a convex optimization problem, specifically, a minimization of a convex quadratic function subject to a convex feasibility set. The first orthogonal projection theorem states that the orthogonal projection operator is in fact well-defined, meaning that the optimization problem in (6) has a unique optimal solution.

Theorem 3.5 (First Projection Theorem)**.** *Let C be a nonempty closed convex set. Then problem* (6) *has a unique optimal solution.*

Proof. As C is closed and $||x - y||^2$ is coercive, we have that the problem admits at least one solution (by Theorem 3.8 of Chapter 1). Moreover, $||x - y||^2$ is strictly convex as the objective function is quadratic with positive definite Hessian (the identity). Thus, from Theorem 3.2 we get that (6) has a unique solution. □

The second projection theorem, provides an useful characterization of the projection operator. Geometrically it states that for a given closed and convex set C, $x \in \mathbb{R}^n$, and for any $y \in C$, the angle between $x - P_C(x)$ and $y - P_C(x)$ is obtuse. This phenomenon is illustrated in Figure 1.

Theorem 3.6 (Second Projection Theorem)**.** *Let C be a nonempty closed convex set. Then $z = P_C(x)$ iff*

$$(x - z)^T(y - z) \leq 0 \quad \forall\, y \in C. \tag{7}$$

Proof. $z = P_C(x)$ iff it is the optimal solution of (6) iff (by Theorem 3.4)

$$\nabla f(z)^T(y - z) \geq 0 \quad \forall\, y \in C,$$

which concludes the proof as $\nabla f(z) = x - z$. □

Another important property of the orthogonal projection operator is given in the following theorem, which also establishes the so-called nonexpansiveness property of P_C.

Theorem 3.7 (Nonexpansiveness of the projection operator)**.** *Let C be a closed and convex set. Then, for any $v, w \in \mathbb{R}^n$*

a)

$$(P_C(v) - P_C(w))^T(v - w) \geq ||P_C(v) - P_C(w)||^2 \tag{8}$$

b)

$$||P_C(v) - P_C(w)|| \leq ||v - w||. \tag{9}$$

Proof. From Theorem 3.6 we have that for any $x \in \mathbb{R}^n$ and $y \in C$

$$(x - P_C(x))^T (y - P_C(x)) \leq 0.$$

Replacing $x = v$ and $y = P_C(w)$ we have

$$(v - P_C(v))^T (P_C(w) - P_C(v)) \leq 0.$$

Replacing, instead, $x = w$ and $y = P_C(v)$

$$(w - P_C(w))^T (P_C(v) - P_C(w)) \leq 0.$$

Now, summing the two inequalities we get

$$(P_C(w) - P_C(v))^T (v - w + P_C(w) - P_C(v)) \leq 0,$$

and hence,

$$(P_C(v) - P_C(w))^T (v - w) \geq ||P_C(w) - P_C(v)||^2.$$

To prove (9), we note that if $P_C(v) = P_C(w)$, the inequality is trivial. Thus, we assume $P_C(v) \neq P_C(w)$. Then by Cauchy-Schwartz, we have

$$(P_C(v) - P_C(w))^T (v - w) \leq ||P_C(v) - P_C(w)|| \cdot ||v - w||,$$

which combined with (8) gives

$$||P_C(v) - P_C(w)||^2 \leq ||P_C(v) - P_C(w)|| \cdot ||v - w||,$$

which concludes the proof as $P_C(v) \neq P_C(w)$. □

Coming back to stationarity, let us provide the alternative characterization of a stationary point through the projection operator. Notice that this theorem holds also when f is non-convex.

Theorem 3.8. *Let $f \in C^1(C)$ with C closed and convex and let $s > 0$. x^* is a stationary point of the problem* (1) *iff*

$$x^* = P_C(x^* - s\,\nabla f(x^*)). \tag{10}$$

Proof. By the second projection theorem (Theorem 3.6), we get that $x^* = P_C(x^* - s\,\nabla f(x^*))$ iff

$$(x^* - s\,\nabla f(x^*) - x^*)^T (x - x^*) \leq 0,$$

which concludes the proof, as x^* is a stationary point when $\nabla f(x^*)^T (x - x^*) \geq 0$ □

3.3 Projected Gradient Method

The characterization of stationary points through equation (10) directly suggest a new algorithm for solving convex constrained optimization methods. As we will see later, this algorithm finds stationary points despite f being convex or not.

Algorithm 1: Projected Gradient (PG) Method

```
Input: $x_0 \in \mathbb{R}^n$, $\epsilon > 0$
$k = 0$
while $||x_{k-1} - x_k|| > \epsilon$ do
    Pick a constant step size $t \in (0, \frac{L}{2})$
    $x_{k+1} = P_C(x_k - t_k \nabla f(x_k)$
    $k = k + 1$
```

The proof of convergence of PG is similar to that of GD. In particular, we first prove the Decrease Lemma for constrained optimization problem.

Lemma 3.1 (Decrease Lemma for Convex Constrained Problems)**.** *Let $f \in \mathrm{C}_{\mathrm{L}}^{1,1}(C)$, where C is convex and closed. Then for any $x \in C$ and $t \in (0, \frac{2}{L})$ the following inequality holds*

$$f(x) - f(P_C(x - t\,\nabla f(x))) \geq t\left(1 - \frac{Lt}{2}\right) \left\|\frac{1}{t}(x - P_C(x - t\,\nabla f(x)))\right\|^2.$$

Proof. Exercise. □

It is now convenient to define the gradient mapping as

$$G_M(x) := M\left(x - P_C\left(x - \frac{1}{M}\nabla f(x)\right)\right) \quad \text{with } M > 0. \tag{11}$$

Note that in the unconstrained case $G_M(x) = \nabla f(x)$ so the gradient mapping is an extension of the usual gradient operator. In addition, by Theorem 3.8, $G_M(x) = 0$ iff x is a stationary point of (1). This means that we can look at $\|G_M(x)\|$ as an optimality measure. Moreover, the sufficient decrease stated above can be rewritten as

$$f(x) - f(P_C(x - t\,\nabla f(x))) \geq t\left(1 - \frac{Lt}{2}\right)\left\|G_{\frac{1}{t}}(x)\right\|^2.$$

This generalized sufficient decrease property allows us to prove similar results to those proven in the unconstrained case.

Theorem 3.9 (Convergence of PG method). *Let $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(C)$, with C closed and convex. Let $\{x_k\}_k$ be a sequence generated by Algorithm 1 for solving* (1). *Assume that f is bounded below over C. Then we have the following*

(a) The sequence $\{f(x_k)\}_k$ is nonincreasing. In addition, for any $k \geq 0$, $f(x_{k+1}) < f(x_k)$ unless x_k is a stationary point.

(b) $G_{\frac{1}{t}}(x_k) \to 0$ as $k \to \infty$.

Notice that the theorem above only ensures convergence to a stationary point, which in the non-convex case might not be a global minimum. Also, the rate of convergence of PG is the same as that of GD, that is $\mathcal{O}(\frac{1}{\sqrt{T}})$. If we assume f to be convex, we can instead ensure a faster rate of convergence, moreover, thanks to Theorem 3.4 all stationary points of (1) are global minima.

Theorem 3.10 (Convergence of PG method for convex problems). *Let $f \in \mathrm{C}^{1,1}_{\mathrm{L}}(C)$ be convex, with C closed and convex. Let $\{x_k\}_k$ be a sequence generated by Algorithm 1 for solving* (1). *Assume that the set of optimal solutions X^* is nonempty and that f^* is the optimal value. Then we have the following*

(a) for any $k \geq 0$ and $x^* \in X^*$

$$2t(f(x_{k+1}) - f^*) \leq ||x_k - x^*||^2 - ||x_{k+1} - x^*||,$$

(b) for any $n \geq 0$:

$$f(x_n) - f^* \leq \frac{||x_0 - x^*||}{2tn}.$$

Proof. By the Descent Lemma (for unconstrained optimization, Lemma 2.2 from Chapter 2), we have

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}||x_{k+1} - x_k||^2.$$

Let x^* be a global minimum of (1), then the gradient characterization of convexity (3) implies that $f(x_k) \leq f(x^*) + \nabla f(x_k)^T(x_k - x^*)$, which together with the previous inequality implies that

$$f(x_{k+1}) \leq f(x^*) + \nabla f(x_k)^T(x_k - x^*) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}||x_{k+1} - x_k||^2. \tag{12}$$

By the second projection theorem (7) applied on the projected point x_{k+1}, we have that

$$\left(x_k - t\,\nabla f(x_k) - x_{k+1}\right)^T\left(x^* - x_{k+1}\right) \leq 0$$

so that

$$\nabla f(x_k)^T\left(x_{k+1} - x^*\right) \leq \frac{1}{t}\left(x_k - x_{k+1}\right)^T\left(x_{k+1} - x^*\right).$$

Therefore, from the above inequality, (12) and $t \leq \frac{1}{L}$, we get

$$\begin{aligned}
f(x_{k+1}) &\leq f(x^*) + \nabla f(x_k)^T(x_k - x^*) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}||x_{k+1} - x_k||^2 \\
&= f(x^*) + \nabla f(x_k)^T(x_{k+1} - x^*) + \frac{L}{2}||x_{k+1} - x_k||^2 \\
&\leq f(x^*) + \frac{1}{t}(x_k - x_{k+1})^T\,(x_{k+1} - x^*) + \frac{L}{2}||x_{k+1} - x_k||^2 \\
&\leq f(x^*) + \frac{1}{t}(x_k - x_{k+1})^T\,(x_{k+1} - x^*) + \frac{1}{2t}||x_{k+1} - x_k||^2 \\
&= f(x^*) + \frac{1}{2t}(x_k - x_{k+1})^T\,(x_{k+1} - x^* + x_k - x^*) \\
&= f(x^*) + \frac{1}{2t}(x_k - x^*)^T\,(x_{k+1} - x^* + x_k - x^*) + \frac{1}{2t}(x^* - x_{k+1})^T\,(x_{k+1} - x^* + x_k - x^*) \\
&= f(x^*) + \frac{1}{2t}\left(||x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)
\end{aligned}$$

establishing part (a). To achieve (b), we sum the inequalities (a) for $k = 0, 1, \ldots, n-1$ and obtain

$$||x_n - x^*\|^2 - \|x_0 - x^*\|^2 \leq 2t\sum_{k=0}^{n-1}\left(f(x^*) - f(x_{k+1})\right) \leq 2tn(f(x^*) - f(x_n)),$$

where in the last inequality we used the fact that $f(x_{k+1}) \leq f(x_k)$, which, in turn, is a consequence of the Descent Lemma and the fact that $t \in (0, \frac{1}{L})$. Thus,

$$f(x_n) - f(x^*) \leq \frac{||x_0 - x^*\|^2 - \|x_n - x^*\|^2}{2tn} \leq \frac{||x_0 - x^*\|^2}{2tn}.$$

□

References