Lecture 2

Smooth Optimization: The gradient descent method

In this lecture we will talk about first order optimization methods that are gradient based. The aim is to solve the optimization problem

$$\min_{x \in X} f(x), \tag{2.1}$$

where f is differentiable. To this purpose we rely on *iterative methods*, that is methods that builds a sequence $(x_k)_{k \in \mathbb{N}}$ iteratively, that is by starting with an initial guess x_0 and then defining x_{k+1} by applying some explicit rule on the previous $x_k, \cdots, x_0$.

Among the several methods, the first that comes to mind is the one that uses the current point x_k and a *descent direction* at that point, that is, a unit vector u along with the derivative of f at x_k is negative

$$D_u f(x_k) := \lim_{t \to 0} \frac{f(x_k + tu) - f(x_k)}{t} < 0. \tag{2.2}$$

Indeed in such case, by the same definition of limit, we can find a sufficiently small $t_k > 0$ such that $f(x_k + t_k u) - f(x_k) < 0$; so defining

$$x_{k+1} = x_k + t_k u \tag{2.3}$$

will diminish the objective function f. We note that for any unit vector u we have $D_u f(x) = \langle \nabla f(x_k), u \rangle$ and, by the Cauchy-Schwartz inequality,

$$-\|\nabla f(x_k)\| \leq \langle \nabla f(x_k), u \rangle \leq \|\nabla f(x_k)\|. \tag{2.4}$$

So, there exists the *steepest* descent direction which is

$$-\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, \tag{2.5}$$

since

$$\min_{\|u\|=1} \langle \nabla f(x_k), u \rangle = -\|\nabla f(x_k)\| = \left\langle \nabla f(x_k), -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} \right\rangle.$$

Methods that uses that direction are generally called gradient descent methods. They have the form

$$x_{k+1} = x_k - \gamma \nabla f(x_k).$$

We note that if at some iteration, say k, this method does not make any progress, we have $x_k = x_k - \gamma \nabla f(x_k)$ and hence $\nabla f(x_k) = 0$, that is x_k is a minimizer of f.

2.1 Differentiability and convexity

We recall the definition of differentiable functions. Throughout the chapter X will be an Euclidean space.

Let $f\colon X \to]-\infty, +\infty]$ be a proper extended real-valued function and let $x_0 \in \operatorname{int}(\operatorname{dom} f)$. Then f *is (Gâteaux) differentiable at* x_0 if there exists a vector $\nabla f(x) \in X$ such that

$$(\forall\, v \in X) \qquad \lim_{t\to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = \langle \nabla f(x_0), v\rangle. \tag{2.6}$$

In such case $\nabla f(x)$ is called the *gradient of* f *at* x_0. Thus, f admits *directional derivatives* at x_0 in every direction v and the directional derivatives depend linearly from v. When f is differentiable at every point of a subset $A \subset \operatorname{int}(\operatorname{dom} f)$ we say that f *is differentiable on* A.

Remark 2.1.1. In case $X = \mathbb{R}^d$, if we take $v = e_i$ (the canonical basis of $\mathbb{R}^d$), then we get $\langle \nabla f(x_0), e_i\rangle = \partial_i f(x_0)$ and hence $\nabla f(x_0) = \big(\partial_1 f(x_0), \dots, \partial_d f(x_0)\big)$.

Theorem 2.1.2 (Fermat's rule)**.** *Let* $f\colon X \to]-\infty, +\infty]$ *be a proper convex function. Let* $x \in \operatorname{int}(\operatorname{dom} f)$ *and suppose that* f *is differentiable at* x*. Then the following statements are equivalent:*

(i) x *is a minimizer of* f*;*

(ii) $\nabla f(x) = 0$.

Proposition 2.1.3 (Characterizations of convexity)**.** *Let* $f\colon X \to]-\infty, +\infty]$ *be a proper extended real-valued function such that* $\operatorname{dom} f$ *is open and convex. Suppose that* f *is differentiable on* $\operatorname{dom} f$*. Then the following are equivalent statements.*

(i) f *is convex.*

(ii) *For every* $x, y \in \operatorname{dom} f$*,* $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$*.*

(iii) *For every* $x, y \in \operatorname{dom} f$*,* $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$*.*

In case f *is twice differentiable on* $\operatorname{dom} f$*, the previous statements are equivalent to*

(iv) *for every* $x \in \operatorname{dom} f$ *and for every* $v \in X$*,* $\langle \nabla^2 f(x) v, v\rangle \geq 0$*.*

Proof. (i) $\Rightarrow$ (ii): Let $x, y \in \operatorname{dom} f$. Then, for every $\lambda \in]0, 1]$, we have $f((1-\lambda)x + \lambda y) \leq (1-\lambda) f(x) + \lambda f(y)$ and hence

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x). \tag{2.7}$$

Thus, by (2.6) and (2.7), we get $\langle \nabla f(x), y - x\rangle \leq f(y) - f(x)$.

(ii) $\Rightarrow$ (iii): Let $x, y \in \operatorname{dom} f$. Then we have

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle &\geq 0 \\ f(x) - f(y) - \langle \nabla f(y), x - y\rangle &\geq 0. \end{aligned}$$

Summing, we get $\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq 0$ and hence the statement.

(iii) $\Rightarrow$ (i): Let $x, y \in \operatorname{dom} f$ and define $\phi\colon [0,1] \to \mathbb{R}$, such that, $\phi(\lambda) = f(x+\lambda(y-x))$. Then $\phi(0) = f(x)$, $\phi(1) = f(y)$, and ϕ is differentiable on $[0,1]$ and, for every $\lambda \in [0,1]$, $\phi'(\lambda) = \langle \nabla f(x+\lambda(y-x)), y-x\rangle$. Now, let $\lambda_1, \lambda_2 \in [0,1]$ be such that $\lambda_1 < \lambda_2$. Then

$$\langle \nabla f(x+\lambda_2(y-x)) - \nabla f(x+\lambda_1(y-x)), (\lambda_2-\lambda_1)(y-x)\rangle \geq 0$$

and hence $(\lambda_2-\lambda_1)(\phi'(\lambda_2) - \phi'(\lambda_1)) \geq 0$, which yields $\phi'(\lambda_1) \leq \phi'(\lambda_2)$. Therefore ϕ' is increasing. Now, let $\lambda \in \,]0,1[$. We show that

$$f(x+\lambda(y-x)) \leq f(x) + \lambda\big(f(y) - f(x)\big). \tag{2.8}$$

Indeed, it follows from Lagrange's theorem that there exist $\lambda_1 \in \,]0,\lambda[$ and $\lambda_2 \in \,]\lambda,1[$ such that

$$\frac{\phi(\lambda)-\phi(0)}{\lambda} = \phi'(\lambda_1) \quad \text{and} \quad \frac{\phi(1)-\phi(\lambda)}{1-\lambda} = \phi'(\lambda_2).$$

Thus, since $\phi'(\lambda_1) \leq \phi'(\lambda_2)$, we have $(1-\lambda)(\phi(\lambda)-\phi(0)) \leq \lambda(\phi(1)-\phi(\lambda))$. Rearranging this inequality (2.8) follows.

(iii) $\Rightarrow$ (iv): Let $x \in \operatorname{dom} f$ and $v \in X$. Since $\operatorname{dom} f$ is open, there exists $\delta > 0$ such that, for every $t \in \,]0,\delta]$, $x+tv \in \operatorname{dom} f$ and, because of (iii), $\langle \nabla f(x+tv) - \nabla f(x), tv\rangle \geq 0$, hence, dividing by t^2,

$$(\forall t \in \,]0,\delta]) \quad \left\langle \frac{\nabla f(x+tv) - \nabla f(x)}{t}, v\right\rangle \geq 0. \tag{2.9}$$

Since, by definition, $\nabla^2 f(x)v = \lim_{t\to 0} \big(\nabla f(x+tv) - \nabla f(x)\big)/t$, the statement follows.

(iv) $\Rightarrow$ (iii): Let $x, y \in \operatorname{dom} f$ and define $\phi\colon [0,1] \to \mathbb{R}$ as in the proof of (iii) $\Rightarrow$ (ii). Then, ϕ is twice differentiable and $\phi''(\lambda) = \langle \nabla^2 f(x+\lambda(y-x))(y-x), y-x\rangle \geq 0$. Therefore, ϕ' is increasing in $[0,1]$. Hence $\phi'(0) \leq \phi'(1)$, which means $\langle \nabla f(x), y-x\rangle \leq \langle \nabla f(y), y-x\rangle$. $\square$

Remark 2.1.4. Strict convexity can be characterized by statements (ii) and (iii) of Proposition 2.1.3, where "$\geq$" is replaced by "$>$" and $x \neq y$.

Example 2.1.5. The function $f\colon \mathbb{R} \to \,]-\infty,+\infty]$ defined as $f(x) = -\log x$ if $x > 0$ and $f(x) = +\infty$ if $x \leq 0$ is strictly convex. Indeed if $x > 0$ and $y > 0$, with $x \neq y$, we have $(f'(x) - f'(y))(x-y) = (-1/x + 1/y)(x-y) = (x-y)^2/(xy) > 0$.

Example 2.1.6.

From Proposition 2.1.3 and Proposition 1.3.12 we derive the following result.

Corollary 2.1.7. *Let $f\colon X \to \,]-\infty,+\infty]$ be a proper extended real-valued function such that $\operatorname{dom} f$ is open and convex and let $\mu > 0$. Suppose that f is differentiable on $\operatorname{dom} f$. Then the following statements are equivalent.*

(i) *f is μ-strongly convex.*

(ii) *For every $x, y \in \operatorname{dom} f$, $f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + (\mu/2)\,\|y-x\|^2$.*

(iii) *For every* $x, y \in \operatorname{dom} f$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$.

In case f *is twice differentiable on* $\operatorname{dom} f$*, the previous statements are equivalent to*

(iv) *for every* $x \in \operatorname{dom} f$ *and for every* $v \in X$, $\langle \nabla^2 f(x)v, v \rangle \geq \mu \|v\|^2$.

Example 2.1.8. Let $A \colon X \to Y$ be a linear operator and let $b \in Y$. Set

$$f \colon X \to Y \quad f(x) = \frac{1}{2} \|Ax - b\|^2 . \tag{2.10}$$

Suppose that A^*A is positive definite (i.e., the minimum eigenvalue of A^*A is strictly positive). Then, for every $x \in X$, $\nabla f(x) = A^*(Ax - b)$ and hence

$$(\forall\, x, y \in X) \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle = \langle A^*A(x - y), x - y \rangle \geq \mu \|x - y\|^2 ,$$

where μ is the minimum eigenvalue of A^*A. Thus, by Corollary 2.1.7, f is μ-strongly convex.

Example 2.1.9. The function defined in Example 2.1.5 is not strongly convex. Indeed if it was so, then there would exists $\mu > 0$ such that, for every $x, y > 0$, $x \neq y$, we would have $(x - y)^2/(xy) = (f'(x) - f'(y))(x - y) \geq \mu(x - y)^2$, that is $1/(xy) > \mu$. But this last statement is false since $1/(xy) \to 0$ as $x \to +\infty$ and $y \to +\infty$.

Remark 2.1.10.

(i) Corollary 2.1.7 establishes that strongly convex functions can be bounded from below at each point by tangent quadratic functions.

(ii) Strongly convex and closed and proper functions have a unique minimizer. Indeed if f is such function, then $f = g + (\mu/2) \|\cdot\|^2$. Since g is closed too, by Theorem 4.4.2, it has an affine minimizer. Thus f is minorized by a quadratic function and hence it is coercive. So, since f is closed, Thereom 1.5.4 ensures that f has minimizers. Unicity comes from Theorem 1.5.5. Note that for the existence of minimizers closedness is necessary even for strongly convex functions. Indeed the function

$$f \colon \mathbb{R} \to \left]-\infty, +\infty\right], \qquad f(x) = \begin{cases} x^2 & \text{if } x > 0 \\ +\infty & \text{if } x \leq 0 \end{cases}$$

does not have any minimizer. The problem is that, even though f is coercive, it is not closed.

(iii) Under the assumption of Corollary 2.1.7, suppose that x_* is a minimizer of f. Then, it follows from Corollary 2.1.7(iii) that, for every $x \in \operatorname{dom} f$, $\langle \nabla f(x), x - x_* \rangle \geq \mu \|x - x_*\|^2$ and hence

$$\boxed{(\forall\, x \in \operatorname{dom} f) \quad \mu \|x - x_*\| \leq \|\nabla f(x)\| .} \tag{2.11}$$

Proposition 2.1.11. *Let $f\colon X \to \left]-\infty,+\infty\right]$ be a proper extended real-valued function such that* $\mathrm{dom} f$ *is open and let $\mu > 0$. Suppose that f is differentiable on* $\mathrm{dom} f$ *and μ-strongly convex. Suppose that $x_* \in \mathrm{dom} f$ is the minimizer of f. Then,*

$$\boxed{(\forall\, x \in \mathrm{dom} f) \quad f(x) - f(x_*) \le \frac{1}{2\mu} \left\|\nabla f(x)\right\|^2.} \tag{2.12}$$

Proof. Let $x \in \mathrm{dom} f$. Then Corollary 2.1.7(ii) yields

$$\begin{aligned} f(x_*) = \min_{y\in\mathrm{dom} f} f(y) &\ge \min_{y\in\mathrm{dom} f} \Big(f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\left\|y - x\right\|^2 \Big) \\ &= f(x) + \min_{y\in\mathrm{dom} f} \frac{1}{2\mu}\big(\left\|\mu(y-x) + \nabla f(x)\right\|^2 - \left\|\nabla f(x)\right\|^2 \big) \\ &\ge f(x) - \frac{1}{2\mu}\left\|\nabla f(x)\right\|^2. \end{aligned}$$

and the statement follows. □

Example 2.1.12. Condition (2.12) can hold even for non-strongly convex functions. Here we provide a significant example for that. Let f be as in Example 2.10 where now we do not assume A to be positive definite. Let b_* be the projection of b onto the range $R(A)$ of A. Let $x \in X$ and $x_* \in \operatorname{argmin} f = \big\{x \in X \,|\, Ax = b_*\big\}$. Then by Pythagoras' theorem, we have

$$f(x) = \frac{1}{2}\big(\left\|Ax - b_*\right\|^2 + \left\|b_* - b\right\|^2 \big).$$

Hence $f_* = \inf_X f = (1/2)\left\|b_* - b\right\|^2$ and

$$f(x) - f_* = \frac{1}{2}\left\|Ax - b_*\right\|^2 = \frac{1}{2}\left\|A(x - x_*)\right\|^2.$$

Moreover, $\nabla f(x) = A^*(Ax - b_*) = A^*A(x - x_*)$, and hence

$$\left\|\nabla f(x)\right\|^2 = \left\|A^*A(x - x_*)\right\|^2.$$

Thus, inequality (2.12) in this case reduces to

$$(\forall\, x \in X) \quad \mu\left\|A(x - x_*)\right\|^2 \le \left\|A^*A(x - x_*)\right\|^2,$$

which is equivalent to

$$(\forall\, y \in R(A)) \quad \mu\left\|y\right\|^2 \le \left\|A^* y\right\|^2 = \langle AA^* y, y\rangle.$$

Now, we consider a singular value decomposition of A^*

$$A^* y = \sum_{i\in I} \sigma_i \langle y, b_i\rangle c_i,$$

where $(\sigma_i)_{i\in I}$ are the singular values of A^* ($(\sigma_i^2)_{i\in I}$ are the nonzero eigenvalues of AA^*), $(b_i)_{i\in I}$ is an orthonormal basis of $N(A^*)^\perp = R(A)$ and $(c_i)_{i\in I}$ is an orthonormal basis of $R(A^*)$. Then, for every $y \in N(A^*)^\perp$,

$$\left\|A^* y\right\|^2 = \sum_{i\in I} \sigma_i^2 |\langle y, b_i\rangle|^2 \ge \sigma_{\min}^2 \sum_{i\in I} |\langle y, b_i\rangle|^2 = \sigma_{\min}^2 \left\|y\right\|^2.$$

Therefore (2.12) holds, with $\mu = \sigma_{\min}^2$, the minimum nonzero eigenvalue of A^*A.

We now study the property of L-*smoothness*, which means that the gradient of the function is L-Lipschitz continuous. The following theorem provides several characterizations of L-smoothness that will be useful in analyzing the gradient descent method. The implication (i) $\Rightarrow$ (ii) is called the *descent lemma*, whereas the implication (i) $\Rightarrow$ (iv) is called the *Baillon-Haddad theorem*.

Theorem 2.1.13. *Let $f\colon X \to \mathbb{R}$ be a convex differentiable function and let $L \in \mathbb{R}_+$. The following statements are equivalent.*

(i) *For every x and y in X, $\|\nabla f(x) - \nabla f(y)\| \leq L\,\|x-y\|$.*

(ii) *For every x and y in X,*

$$f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \frac{L}{2}\,\|x-y\|^2 . \tag{2.13}$$

(iii) *For every x and y in X,*

$$\frac{1}{2L}\,\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x\rangle \tag{2.14}$$

(iv) *For every x and y in X,*

$$\frac{1}{L}\,\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle \tag{2.15}$$

(v) *For every x and y in X, $\langle \nabla f(x) - \nabla f(y), x-y\rangle \leq L\,\|x-y\|^2$.*

(vi) $\dfrac{L}{2}\,\|\cdot\|^2 - f$ *is convex.*

In case f is twice differentiable on X, the previous statements are equivalent to

(vii) *for every $x \in X$ and for every $v \in X$, $\langle \nabla^2 f(x)v, v\rangle \leq L\,\|v\|^2$.*

(viii) *for every $x \in X$, $\|\nabla^2 f(x)\| \leq L$.*

Proof. (i) $\Rightarrow$ (ii): Let $x, y \in X$ and set $\phi\colon [0,1] \to \mathbb{R}$, $\phi(\lambda) = f(x + \lambda(y-x))$. Then ϕ is continuously differentiable and, for every $\lambda \in [0,1]$, $\phi'(\lambda) = \langle \nabla f(x + \lambda(y-x)), y-x\rangle$.

Thus,

$$
\begin{aligned}
f(y)-f(x)-\langle\nabla f(x),y-x\rangle &= \phi(1)-\phi(0)-\langle\nabla f(x),y-x\rangle \\
&= \int_0^1 \phi'(\lambda)\,\mathrm{d}\lambda-\langle\nabla f(x),y-x\rangle \\
&= \int_0^1 \langle\nabla f(x+\lambda(y-x))-\nabla f(x),y-x\rangle\,\mathrm{d}\lambda \qquad (2.16)\\
&\le \int_0^1 \|\nabla f(x+\lambda(y-x))-\nabla f(x)\|\,\|y-x\|\,\mathrm{d}\lambda \\
&\le L\,\|y-x\|^2\int_0^1 \lambda\,\mathrm{d}\lambda \\
&= \frac{L}{2}\,\|y-x\|^2 .
\end{aligned}
$$

(ii) $\Rightarrow$ (iii): Let $x\in X$ and let $g\colon X\to\mathbb{R}\colon y\mapsto f(y)-\langle\nabla f(x),y\rangle$. Then g is convex and differentiable and, for every $y\in X$, $\nabla g(y)=\nabla f(y)-\nabla f(x)$. Since $\nabla g(x)=0$, x is a minimizer of g. Now let $y\in X$. Using implication (i) $\Rightarrow$ (ii) applied to g,

$$
\begin{aligned}
g(x)=\min_{z\in X} g(z) &\le \min_{z\in X}\Big(g(y)+\langle\nabla g(y),z-y\rangle+\frac{L}{2}\,\|z-y\|^2\Big)\\
&= g(y)-\frac{1}{2L}\,\|\nabla g(y)\|^2 .
\end{aligned}
$$

Substituting the expression of $g(x),g(y)$, and $\nabla g(y)$ into the above inequality, (2.14) follows.

(iii) $\Rightarrow$ (iv): The statement follows by swapping x and y in (2.14) and summing the resulting inequality with (2.14).

(iv) $\Rightarrow$ (i): Let x and y in X. By the Cauchy-Schwartz inequality we get

$$
\frac{1}{L}\,\|\nabla f(x)-\nabla f(y)\|^2\le\langle\nabla f(x)-\nabla f(y),x-y\rangle\le\|\nabla f(x)-\nabla f(y)\|\,\|x-y\| .
$$

Thus $\|\nabla f(x)-\nabla f(y)\|\le L\,\|x-y\|$.

(ii) $\Rightarrow$ (v): It follows by swapping x and y in (2.13) and summing with (2.13).

(v) $\Rightarrow$ (ii): Let $x,y\in X$. If we define ϕ as in the proof of (i) $\Rightarrow$ (ii), we see that (v) implies that ϕ' is continuous. Therefore, it follows from (2.16) that

$$
\begin{aligned}
f(y)-f(x)-\langle\nabla f(x),y-x\rangle &= \int_0^1 \frac{1}{\lambda}\langle\nabla f(x+\lambda(y-x))-\nabla f(x),\lambda(y-x)\rangle\,\mathrm{d}\lambda\\
&\le \int_0^1 L\,\|y-x\|^2\,\lambda\,\mathrm{d}\lambda\\
&= \frac{L}{2}\,\|y-x\|^2 .
\end{aligned}
$$

(v) $\Leftrightarrow$ (vi): Condition (v) can be equivalently written as

$$
(\forall\, x,y\in X)\quad \langle Lx-\nabla f(x),x-y\rangle\ge 0.
$$

Since $\nabla\big((L/2)\,\|\cdot\|^2-f\big)(x)=Lx-\nabla f(x)$, the statement follows from Proposition 2.1.3. $\square$

Proposition 2.1.14. *Let $f: X \to \mathbb{R}$ be a differentiable function. Then the following are equivalent*

(i) *f is μ-strongly convex and ∇f is Lipschitz continuous with constant L.*

(ii) $\forall x, y \in X,\ \dfrac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2 + \dfrac{\mu L}{L+\mu}\|x-y\|^2 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle.$

Proof. Set $g = f - (\mu/2)\|\cdot\|^2$. Then $\frac{L-\mu}{2}\|\cdot\|^2 - g = (L/2)\|\cdot\|^2 - f$. Therefore It follows from Theorem 2.1.13(vi) that (i) is equivalent to g convex and with gradient $(L-\mu)$-Lipschitz continuous. Then, by Theorem 2.1.13 item (iv), this latter fact can be expressed as

$$(\forall x, y \in X) \quad \frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|^2 \leq \langle \nabla g(x) - \nabla g(y), x-y\rangle. \tag{2.17}$$

and hence, substituting the expressions of $\nabla g(x)$ and $\nabla g(y)$, as

$$(\forall x, y \in X)\ \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y) - \mu(x-y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x-y\rangle - \mu\|x-y\|^2.$$

The statement follows from the latter inequality, by multiplying by $L-\mu$, expanding the square norm on the left hand side and rearranging the terms. □

2.2 Nonexpansive and contractive operators

Definition 2.2.1. Let X be an Euclidean space and let $T: X \to X$. Then

(i) T is *nonexpansive* if

$$(\forall x, y \in X) \quad \|Tx - Ty\| \leq \|x-y\|.$$

(ii) T is a *contraction* if there exists $q \in\,]0,1[$ such that

$$(\forall x, y \in X) \quad \|Tx - Ty\| \leq q\|x-y\|,$$

The first important result concerns contractive mapping.

Theorem 2.2.2 (Banach-Cacioppoli)**.** *Let $T: X \to X$ be a q-contractive mapping for some $0 < q < 1$. Then there exists a unique fixed point of T, that is, a point $x_* \in X$ such that $T(x_*) = x_*$. Moreover, let $x_0 \in X$ and define, iteratively*

$$x_{k+1} = T(x_k). \tag{2.18}$$

Then,

$$(\forall k \in \mathbb{N}) \quad \|x_k - x_*\| \leq q^k\|x_0 - x_*\| \quad \textit{and} \quad \|x_k - x_*\| \leq \frac{q^k}{1-q}\|x_0 - x_1\|. \tag{2.19}$$

Proof. We first note that

$$(\forall\, x, y \in X) \qquad \|x - y\| \leq \frac{1}{1-q}\big(\, \|x - Tx\| + \|y - Ty\|\,\big). \tag{2.20}$$

Indeed $\|x-y\| \leq \|x-Tx\|+\|Tx-Ty\|+\|Ty-y\| \leq \|x-Tx\|+q\,\|x-y\|+\|y-Ty\|$, hence $(1-q)\,\|x-y\| \leq \|x-Tx\|+\|Ty-y\|$ and (2.20) follows. Inequality (2.20) shows that there may exists at most one fixed point of T. Moreover, for every $k, h \in \mathbb{N}$,

$$\begin{aligned}
\|x_k - x_h\| &\leq \frac{1}{1-q}\big(\, \|x_k - x_{k+1}\| + \|x_h - x_{h+1}\|\,\big) \\
&\leq \frac{1}{1-q}\big(\, \big\|T^k(x_0) - T^k(x_1)\big\| + \big\|T^h(x_0) - T^h(x_1)\big\|\,\big) \\
&\leq \frac{1}{1-q}\big(q^k\,\|x_0 - x_1\| + q^h\,\|x_0 - x_1\|\,\big) \\
&\leq \frac{q^k + q^h}{1-q}\,\|x_0 - x_1\|\,.
\end{aligned} \tag{2.21}$$

where we used that T^k is q^k-contractive. Since $0 < q < 1$, q^k and q^h converge to zero as k and h go to $+\infty$. Therefore, $(x_k)_{k\in\mathbb{N}}$ is a Cauchy sequence and hence it converges, say to x_*. Then $Tx_k \to Tx_*$ and $Tx_k = x_{k+1} \to x_*$, so $Tx_* = x_*$, that is, x_* is a fixed point of T. The second inequality in (2.19) follows from (2.21) by letting $h \to +\infty$. The first equality in (2.19) follows from the following chain of inequalities

$$\|x_k - x_*\| = \|Tx_{k-1} - Tx_*\| \leq q\,\|x_{k-1} - x_*\| \leq \cdots \leq q^k\,\|x_0 - x_*\|\,.$$

□

Remark 2.2.3.

(i) Iterative methods of type (2.18) are called *fixed point iterations* or *Picard iterations*.

(ii) Nonexpansive operators, may have no fixed points. For instance, a translation $T = \mathrm{Id} + a$, with $a \neq 0$, does not have any fixed point.

(iii) For nonexpansive operators, even admitting fixed points, the Picard iteration may fail to converge. Indeed, this occurs if we take $T = -\mathrm{Id}$ and start with $x_0 \neq 0$. More generally rotations are nonexpansive operators admitting a fixed points.

Example 2.2.4.

(i) $\alpha\mathrm{Id}$, with $|\alpha| < 1$, is a contractive operator and its only fixed point is zero.

2.3 Convergence analysis

Now we define the gradient descent algorithm for minimizing smooth convex functions. In this section we assume that $f\colon X \to \mathbb{R}$ is convex differentiable with Lipschitz continuous gradient with constant L.

Algorithm 2.3.1. The *gradient descent algorithm* is defined as follows.

$$\begin{array}{l} \text{Let } \gamma > 0 \text{ and } x_0 \in X. \\ \text{For } k = 0, 1, \dots \\ \lfloor \; x_{k+1} = x_k - \gamma \nabla f(x_k). \end{array} \tag{2.22}$$

It is important to note that some restriction on the step-size γ should be required. Indeed if we do gradient descent to the function $f(x) = (L/2)\,\|x\|^2$, we have

$$x_{k+1} = (1 - \gamma L) x_k.$$

Thus if we take $\gamma = 2/L$, we have $x_{k+1} = -x_k$ and the sequence does not converge, unless $x_0 = 0$.

We have the following result.

Proposition 2.3.2. *Let $k \in \mathbb{N}$. Then*

$$\gamma\Big(1 - \frac{\gamma L}{2}\Big)\,\|\nabla f(x_k)\|^2 \le f(x_k) - f(x_{k+1}). \tag{2.23}$$

Thus, if $\gamma \le 2/L$, then $f(x_{k+1}) \le f(x_k)$, that is, the algorithm is descending.

Proof. Since $x_{k+1} - x_k = -\gamma \nabla f(x_k)$, by Theorem 2.1.13(ii), we have

$$f(x_{k+1}) \le f(x_k) - \gamma\langle \nabla f(x_k), \nabla f(x_k)\rangle + \frac{L}{2}\gamma^2\,\|\nabla f(x_k)\|^2$$

and the statement follows. $\square$

Set $T = \mathrm{Id} - \gamma\nabla f \colon X \to X$, where Id is the identity operator on X. Then the gradient descent algorithm (2.22) can be written as a *Picard iteration*

$$x_{k+1} = T x_k \tag{2.24}$$

and the minimizers of f are nothing but the fixed points of T.

We first address the question of when the gradient descent operator T is a contraction. This will provide necessary conditions for applying the Banach fixed point theorem.

Proposition 2.3.3. *Let $f \colon X \to \mathbb{R}$ be a differentiable convex function. Suppose that for some $\gamma > 0$, the operator $T = \mathrm{Id} - \gamma\nabla f$ is a contraction. Then f is strongly convex and its gradient is Lipschitz continuous.*

Proof. Let $x, y \in X$. Then

$$\begin{aligned}
&\|Tx - Ty\|^2 \le q^2\,\|x - y\|^2 \\
&\Leftrightarrow\ \|x - y - \gamma(\nabla f(x) - \nabla f(y))\|^2 \le q^2\,\|x - y\|^2 \\
&\Leftrightarrow\ \|x - y\|^2 + \gamma^2\,\|\nabla f(x) - \nabla f(y)\|^2 - 2\gamma\langle \nabla f(x) - \nabla f(y), x - y\rangle \le q^2\,\|x - y\|^2 \\
&\Leftrightarrow\ (1 - q^2)\,\|x - y\|^2 + \gamma^2\,\|\nabla f(x) - \nabla f(y)\|^2 \le 2\gamma\langle \nabla f(x) - \nabla f(y), x - y\rangle \\
&\Rightarrow\ \begin{cases} \dfrac{1 - q^2}{2\gamma}\,\|x - y\|^2 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle \\[2mm] \dfrac{\gamma}{2}\,\|\nabla f(x) - \nabla f(y)\|^2 \le \langle \nabla f(x) - \nabla f(y), x - y\rangle \end{cases}
\end{aligned}$$

So in virtue of Theorem 2.1.13(iv) and Corollary 2.1.7(iii), f is strongly convex and ∇f is Lipschitz continuous. $\square$

Now we assume that f is strongly convex and with Lipschitz continuous gradient. Then we will prove that there exists an interval of values of γ for which T is a contraction. We will prove this first under the additional hypothesis of twice differentiability.

Theorem 2.3.4 (Convergence 1)**.** *Let $f\colon X \to \mathbb{R}$ be twice differentiable and suppose that f is μ-strongly convex and that ∇f is L-Lipschitz continuous. Then, for every $\gamma > 0$, $T = \mathrm{Id} - \gamma \nabla f$ is Lipschitz continuous with constant*

$$q(\gamma) = \max\{|1-\gamma\mu|, |1-\gamma L|\} = \begin{cases} 1-\gamma\mu & \text{if } \gamma \leq \dfrac{2}{L+\mu} \\ 1-\gamma L & \text{if } \gamma \geq \dfrac{2}{L+\mu}. \end{cases} \tag{2.25}$$

So, T is a contraction if $\gamma \in \,]0, 2/L[$. Therefore the gradient descent algorithm features the following rate of convergence

$$\|x_k - x_*\| \leq q(\gamma)^k \|x_0 - x_*\| \qquad \text{and} \qquad f(x_k) - f(x_*) \leq \frac{L}{2} q(\gamma)^{2k} \|x_0 - x_*\|^2 .$$

For the optimal stepsize $\gamma = 2/(L+\mu)$, we have

$$\|x_k - x_*\| \leq \left(\frac{L-\mu}{L+\mu}\right)^k \|x_0 - x_*\| \qquad \text{and} \qquad f(x_k) - f(x_*) \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x_0 - x_*\|^2 .$$

Proof. The mapping T is differentiable and $T'(x) = \mathrm{Id} - \gamma\nabla^2 f(x)$. By the mean value theorem

$$\forall\, x, y \in X \quad \|Tx - Ty\| \leq q\,\|x-y\| \;\Leftrightarrow\; \forall\, x \in X \quad \|T'(x)\| \leq q.$$

Moreover, $\|T'(x)\| = \sup_{\lambda \in \sigma(\nabla^2 f(x))} |1-\gamma\lambda|$. Since f is μ strongly convex and ∇f is L-Lipschitz continuous,

$$(\forall\, x \in X)(\forall\, u \in X) \quad \mu\,\|u\|^2 \leq \langle \nabla f(x) u, u\rangle \leq L\,\|u\|^2 .$$

So $\sigma(\nabla^2 f(x)) \subset [\mu, L]$ and hence $\|T'(x)\| \leq \max_{\lambda\in[\mu,L]} |1-\gamma\lambda| = q(\gamma)$. It follows from (2.25) that $q(\gamma) < 1 \;\Leftrightarrow\; \gamma \in \,]0, 2/L[$. The inequalities on the values follow from Theorem 2.1.13(ii) with $y = x$ and $x = x_*$. ☐

The hypothesis of twice differentiability can indeed be removed.

Remark 2.3.5. The following two properties are equivalent to Proposition 2.1.14(ii).

(*a*) $\|(\mathrm{Id}-T)x - (\mathrm{Id}-T)y\|^2 + \gamma^2\mu L\,\|x-y\|^2 \leq \gamma(\mu+L)\langle(\mathrm{Id}-T)x - (\mathrm{Id}-T)y, x-y\rangle$.

(*b*) $\|Tx - Ty\|^2 \leq \left(1 - \dfrac{2\gamma\mu L}{\mu+L}\right)\|x-y\|^2 - \left(\dfrac{2}{\gamma(\mu+L)} - 1\right)\|(\mathrm{Id}-T)x - (\mathrm{Id}-T)y\|^2$.

Indeed, multiplying (ii) by $\gamma^2(L+\mu)$ and replacing $\gamma\nabla f$ by $\mathrm{Id}-T$, (*a*) follows. Moreover the equivalence between (*a*) and (*b*) follows again from identity (6.6). Note that if $\gamma(L+\mu)/2 \leq 1$, then (*b*) yields

$$\|Tx - Ty\| \leq \left(1 - \frac{2\gamma\mu L}{\mu+L}\right)^{1/2} \|x-y\|\,, \tag{2.26}$$

where

$$0 < \frac{2\gamma\mu L}{L+\mu} \leq \frac{4\mu L}{(L+\mu)^2} - 1 + 1 = 1 - \left(\frac{L-\mu}{L+\mu}\right)^2 < 1.$$

Therefore, for every $\gamma \in \left]0, 2/(L+\mu)\right]$, T is a *contraction* with the constant given in (2.26). Note that this constant is always worse than that given in (2.25) in the interval $\left]0, 2/(L+\mu)\right[$.

Theorem 2.3.6 (Convergence 2). *Suppose that f is strongly convex with modulus $\mu > 0$ and let x_* be the minimizer of f. Suppose that $0 < \gamma \leq 2/(L+\mu)$. Then, for every $k \in \mathbb{N}$,*

$$\|x_{k+1} - x_*\| \leq \left(1 - \frac{2\gamma\mu L}{L+\mu}\right)^{k/2} \|x_0 - x_*\|, \tag{2.27}$$

where $\big(1 - 2\gamma\mu L/(L+\mu)\big) < 1$. Moreover, the optimal step size in (2.27) is $\gamma = 2/(L+\mu)$ and in such case

$$\|x_{k+1} - x_*\| \leq \left(\frac{L-\mu}{L+\mu}\right)^{k} \|x_0 - x_*\|, \tag{2.28}$$

$$f(x_k) - f(x_*) \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x_0 - x_*\|^2, \tag{2.29}$$

Proof. In Remark 2.3.5 we saw that T is a contraction with constant $\big(1 - 2\gamma\mu L/(L+\mu)\big)^{1/2}$. Since x_* is a fixed point of T, (2.27) follows from Theorem 2.2.2. The minimum value of $\big(1 - 2\gamma\mu L/(L+\mu)\big)$ is reached for γ maximum, that is, $\gamma = 2/(L+\mu)$. In such case, the constant becomes $\big(1 - 4\gamma\mu L/(L+\mu)^2\big)^{1/2} = (L-\mu)/(L+\mu)$. Finally, it follows from Theorem 2.1.13(ii), with $x = x_*$ and $y = x_k$ that

$$f(x_k) - f(x_*) \leq \frac{L}{2}\|x_k - x_*\|^2$$

and (2.29) follows. □

Remark 2.3.7. Under the hypotheses of Theorem 2.3.6, a linear rate (but with a worse constant) can be derived also if $\gamma < 2/L$. Indeed, let $k \in \mathbb{N}$. Then, it follows from Proposition 2.3.2 and Corollary 2.1.11 that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma\left(1 - \frac{L\gamma}{2}\right)\|\nabla f(x_k)\|^2 \\ &\leq f(x_k) - \gamma\mu(2 - L\gamma)(f(x_k) - f(x_*)). \end{aligned}$$

Therefore

$$f(x_{k+1}) - f(x_*) \leq q^2\big(f(x_k) - f(x_*)\big), \tag{2.30}$$

where

$$0 \leq q := \big(1 - \gamma\mu(2 - L\gamma)\big)^{1/2} < 1, \tag{2.31}$$

since $0 < \gamma\mu(2 - L\gamma) \leq \gamma L(2 - \gamma L) = -(\gamma L - 1)^2 + 1 \leq 1$ (we used $\mu \leq L$ in the second inequality). In the end, recalling also (1.17), we have that, for every $k \in \mathbb{N}$,

$$f(x_k) - f(x_*) \leq q^{2k}(f(x_0) - f(x_*)) \qquad \|x_k - x_*\| \leq \sqrt{\frac{2}{\mu}}q^k\sqrt{f(x_0) - f(x_*)}. \tag{2.32}$$

Note that the best (smallest) value of q in (2.31) is given for $\gamma = 1/L$ and it is $q = (1-\mu/L)^{1/2}$. Comparing this result with that given in Theorem 2.3.6, we have that

$$1 - \frac{2\gamma\mu L}{L+\mu} \leq 1 - \gamma\mu(2-\gamma L) \;\Leftrightarrow\; \gamma \in \left]\frac{2}{L}\frac{\mu}{L+\mu}, \frac{2}{L}\right].$$

and $1/L > 2\mu/(L(L+\mu))$. So this analysis provides worse constant than that in Theorem 2.3.6 on the interval $[1/L, 2/L]$.

Example 2.3.8. The first inequality in (2.32) holds also for non-strongly convex functions, provided that inequality (2.12) holds. This is the case of the function considered in Example 2.1.12. Moreover the second inequality in (2.32) is replace by

$$\operatorname{dist}(x_k, \operatorname{argmin} f) \leq \sqrt{\frac{2}{\mu}} q^k \sqrt{f(x_0) - f(x_*)}.$$