Chapter 5

First order methods

The preceeding chapter was the occasion to describe one of the a fundamental difference between statistical estimation problems which can or cannot be solved in polynomial time. Efficient numerical solvers exist for $\mathcal{NP}$-hard problems and their use in high dimensional statistics is explored [8]. Nontheless, we will focus on algorithms which are less computationally demanding, and better scale in very large dimensions.

We have seen that large families of convex optimization problems can be solved via generic purpose solvers which have efficient implementations. These solvers have the following properties, in dimension d:

- The cost of a single iteration is of the order of d^3
- They lead to fast converging sequences allowing to obtain very accurate solutions.

In very large dimensions d^3 may be too big to be considered as reasonable and we need cheaper algorithms. This observation motivated the rise of first order methods as efficient alternatives in high dimensional statistics and signal processing. These methods have a long history in applied mathematics and the recent trends in data analysis bolstered new developments

Sources for this chapter include the classic book of Rockafellar [29], the book of Nesterov [23] as well as elements presented in Sébastien Bubeck's book [10]. Good references on this topic include the surveys [14, 2] which is very close to the statistical matters presented in these notes.

5.1 Gradient descent

In this section f denotes a continuously differentiable function. The gradient descent algorithm can be described as follows, choose $x_0 \in \mathbb{R}^d$ and iterate for $k \in \mathbb{N}$:

$$x_{k+1} = x_k - s_k \nabla f(x_k) \tag{5.1}$$

Each iteration costs a call to the gradient with a vector addtion which. A vector addition costs of the order of d operations. Hence it is much cheaper than the d^3 operations required to run interior point methods. For example, if one is given a computational budget of the order of d^2, then one can implement d steps of gradient descent while Newton step simply cannot be considered. We review basic theoretical results known for the gradient method for convex optimization.

5.1.1 Dynamical systems intuition

The minimizing properties of gradient descent are easily seen in continuous time.

Proposition 5.1.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be twice differentiable such with compact sublevel sets. Consider the differential equation, for $x_0 \in \mathbb{R}^p$,*

$$\dot{x}(t) = -\nabla f(x(t)) \tag{5.2}$$
$$x(0) = x_0. \tag{5.3}$$

Then, there exists a solution to the initial value problem defined for all $t > 0$.

- *$\int_0^{+\infty} \|\nabla f(x(t))\|_2^2 dt < +\infty$ and $\liminf_{t\to\infty} \|\nabla f(x(t))\| = 0$.*
- *Any accumulation point $\bar{x}$ of the trajectory satisfies $\nabla f(\bar{x}) = 0$.*
- *If in addition f is convex, set $f^* = \inf_{x\in\mathbb{R}^p} f(x)$ and assume that it is attained at x^*, we have for any $t \in \mathbb{R}$, $t > 0$,*

$$f(x(t)) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2t}.$$

And $x(t) \underset{t\to\infty}{\to} \bar{x}$ where $\bar{x}$ is a global minimizer of f.

Proof. First note that ∇f is continuous and locally Lipschitz so that there is a unique maximal solution to the initial value problem (Cauchy-Lipschitz). By differentiation, we obtain, for any t in the interval of definition of the solution,

$$\frac{d}{dt}\left(f(x(t))\right) = \dot{x}(t)^T \nabla f(x(t)) = -\|\nabla f(x(t))\|_2^2.$$

We deduce that $f(x(t)) \leq f(x_0)$ and by compacity the trajectory remains bounded and is defined for all $t > 0$ (sortie de tout compact). Integrating between 0 and $T > 0$, we obtain $f(x(T)) = f(x(0)) - \int_0^T \|\nabla f(x(t))\|_2^2 dt$. The function f is decreasing allong the trajectory and bounded below by compacity. This proves the first point

The second point is left as an exercise.

For the third point, using convexity of f, we obain

$$\frac{d}{dt}\|x(t) - x^*\|_2^2 = -2\left\langle \nabla f(x(t)), x(t) - x^* \right\rangle \leq 2\left(f(x^*) - f(x(t))\right) < 0.$$

Integrading between 0 and $t > 0$, we obtain

$$t(f(x(t)) - f^*) \leq \int_0^t f(x(s)) - f^* ds \leq \frac{1}{2}\|x(0) - x^*\|_2^2$$

where the first inequality follows because f is decreasing allong the trajectory. For the convergence of $x(t)$, we have

$$\frac{d}{dt}\|x(t) - x^*\|_2^2 \leq 2\left(f(x^*) - f(x(t))\right) < 0,$$

so that $\|x(t) - x^*\|_2^2$ is decreasing and the trajectory remains bounded. Since $x(t)$ has at least one accumulation point which attains the minimum of f, $x(t)$ must converge to this point (Opial's Lemma). □

5.1.2 Convergence of gradient descent

We start with the following Lemma which proof is left as an exercise.

Lemma 5.1.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be continuously differentiable with L-Lipschitz gradient ($L > 0$), then for any $x, y \in \mathbb{R}^p$,*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|_2^2.$$

Proof. Using the fundamental theorem of calculus, we have, for any x, y

$$\begin{aligned} f(y) - f(x) &= \int_{t\in[0,1]} \langle \nabla f((1-t)x + ty), y - x\rangle \, dt \\ &= \int_{t\in[0,1]} \langle \nabla f((1-t)x + ty) - \nabla f(x) + \nabla f(x), y - x\rangle \, dt \\ &= \langle \nabla f(x), y - x\rangle + \int_{t\in[0,1]} \langle \nabla f((1-t)x + t(y)) - \nabla f(x), y - x\rangle \, dt. \end{aligned}$$

We deduce that

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &= \left| \int_{t\in[0,1]} \langle \nabla f((1-t)x + ty) - \nabla f(x), y - x\rangle \, dt \right| \\ &\leq \int_{t\in[0,1]} |\langle \nabla f((1-t)x + ty) - \nabla f(x), y - x\rangle| \, dt \\ &\leq \int_{t\in[0,1]} \|f((1-t)x + ty) - \nabla f(x)\| \times \|y - x\| dt \\ &\leq \int_{t\in[0,1]} tL \times \|y - x\|^2 dt \\ &= \frac{L}{2}\|y - x\|^2, \end{aligned}$$

which proves the result. □

The gradient descent algorithm can be seen as an explicit discretisation of the differential equation (5.3). It preserves the same qualitative properties as seen in the following proposition.

Proposition 5.1.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be continuously differentiable with L-Lipschitz gradient and such that $\inf_{x\in\mathbb{R}^p} f(x) > -\infty$. Consider the algorithm, for $x_0 \in \mathbb{R}^p$ and*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k). \tag{5.4}$$

Then

- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$, *(any accumulation point $\bar{x}$ of the trajectory satisfies $\nabla f(\bar{x}) = 0$).*
- *If in addition f is convex, set $f^* = \inf_{x\in\mathbb{R}^p} f(x)$ and assume that it is attained at x^*, we have for any $k \in \mathbb{N}$, $k > 0$,*

  $$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

  Furthermore x_k converges to $\bar{x}$ a global minimum of f
- *If in addition f is μ-strongly convex, then we have for any $k \in \mathbb{N}$*

  $$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^*).$$

Proof. The ideas are the same, first, the descent Lemma ensures that for any $k \in \mathbb{N}$

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2. \end{aligned} \tag{5.5}$$

Note that that f is decreasing allong the iterates of the algorithm. We have

$$x_{k+1} = \arg\min_{y\in\mathbb{R}^p} f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2, \tag{5.6}$$

so that for all $y \in \mathbb{R}^d$,

$$f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 \tag{5.7}$$

$$= \quad f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + \frac{L}{2}\|y - x_{k+1}\|_2^2. \tag{5.8}$$

We obtain

$$\begin{aligned}
&f(x^*) + \frac{L}{2}\|x^* - x_k\|_2^2 \\
&\geq f(x_k) + \langle \nabla f(x_k), x^* - x_k\rangle + \frac{L}{2}\|x^* - x_k\|_2^2 && \text{convexity} \\
&= f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + \frac{L}{2}\|x_{k+1} - x^*\|_2^2 && (5.8) \\
&\geq f(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x^*\|_2^2, && (5.5)
\end{aligned}$$

By summing up, we obtain for any $K \in \mathbb{N}$, $K \geq 1$,

$$\frac{L}{2}\|x^* - x_0\|_2^2 \geq \sum_{k=1}^{K} f(x_k) - f^* \geq K(f(x_K) - f^*).$$

For the last point we have by strong convexity for any $x \in \mathbb{R}^d$,

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2}\|x^* - x\|_2^2 \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f^*)$$

We have for all $k \in \mathbb{N}$,

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \leq (f(x_k) - f^*)\left(1 - \frac{\mu}{L}\right).$$

□

5.2 Recap on nonsmooth analysis

The following content is treated in greater generality in [29]. In what follows f denotes a lower semi-continuous convex function on $\mathbb{R}^p$ which is finite at least at one point. Lower semi-continuity refers to the fact that the epigraph is closed:

$$\mathrm{epi}_f = \left\{(x, z) \in \mathbb{R}^{p+1},\, z \geq f(x)\right\}.$$

which is expressed equivalently as for any $x \in \mathbb{R}^p$

$$\liminf_{y\to x} f(y) \geq f(x).$$

The function f is allowed to take value $+\infty$, we denote its domain by

$$\mathrm{dom}_f = \{x \in \mathbb{R}^p,\, f(x) < +\infty\},$$

which is a convex set.

Exercise 5.2.1. *Show that a convex function is continuous on the interior of its domain.*

5.2.1 Notion of subgradient

Definition 5.2.1. *For any $x \in \mathrm{dom}_f$, the subgradient of f denotes the set*

$$\partial f(x) = \{v \in \mathbb{R}^p,\, f(y) \geq f(x) + \langle v, y - x\rangle\,,\, \forall y \in \mathbb{R}^p\}\,.$$

For $x \notin \mathrm{dom}_f$, $\partial f(x)$ is set to be empty.

We deduce from the definition the generalization of Fermat rule

Theorem 5.2.1. *$x^* \in \arg\min_x f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proposition 5.2.1. *For any $x \in \mathbb{R}^p$, $\partial f(x)$ is a closed convex set. Furthermore, at any $x \in \mathrm{int}(\mathrm{dom}_f)$, $\partial f(x)$ is non empty and bounded*

Proof. Closedness and convexity follow from the definition. Take $x \in \mathbb{R}^p$ and assume that x is in the interior of the domain of f this means that f is finite around x. The set epi_f is convex in $\mathbb{R}^{p+1}$, and $(x, f(x))$ belongs to the boundary of epi_f. Consider a suporting hyperplane of epi_f at $(x, f(x))$ as given by Theorem 4.3.4, this provides a vector $v \in \mathbb{R}^p$ and a number $a \in \mathbb{R}$ such that for all $y \in \mathrm{dom}_f$

$$az + v^T y \geq af(x) + v^T x, \qquad \forall z \geq f(y).$$

If $a = 0$ then v is different from 0 and this provides a supporting hyperplane to dom_f at x which contradicts the fact that f is finite around x. Hence $a \neq 0$. It must holda that $a > 0$ and $\frac{-v}{a}$ provides a subgradient for f. Boundedness follows because for any $v \in \partial f(x)$,

$$f\left(x + v\frac{1}{\|v\|_2^{3/2}}\right) \geq f(x) + \|v\|_2^{1/2},$$

if the set of such v was unbounded, the left hand side should remain finite while the right hand side should diverge to $+\infty$. □

Exercise 5.2.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function, show that ∂f is sequencialy closed in the sence that, for any $\bar{x}$*

$$\left\{v \in \mathbb{R}^p,\, \exists\,(x_k, v_k)_{k\in\mathbb{N}}\,,\, x_k \to \bar{x},\, v_k \to v,\, f(x_k) \to f(\bar{x})\right\} \subset \partial f(\bar{x})$$

Exercise 5.2.3. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, show that f is L-Lipschitz if and only if $\sup_{x\in\mathbb{R}^p,\, v\in\partial f(x)} \|v\|_2 \leq L$.*

Theorem 5.2.2. *Let f be convex and lower semicontinuous and finite at least at one point, then f is the supremum of all its affine minorants: for any $x \in \mathbb{R}^p$*

$$f(x) = \sup_{r\in\mathbb{R}, v\in\mathbb{R}^p} r + v^T x \qquad \text{s.t.} \qquad f(y) \geq r + v^T y,\, \forall y \in \mathbb{R}^p.$$

Proof. epi_f is a closed set in $\mathbb{R}^{p+1}$. Reducing the dimension if necessary and restricting to affine subspaces, we may consider that $\mathrm{int}(\mathrm{dom}_f) \neq \emptyset$. Fix $(x, \mu) \notin \mathrm{epi}_f$, this means that $\mu < \min\{f(x), +\infty\}$. From the separating hyperplane theorem, there exists, $v \in \mathbb{R}^p$, $\beta \in \mathbb{R}$ and $a \in \mathbb{R}$ such that

$$\begin{aligned} &v^T y + \beta z - a \leq 0 \qquad \forall y \in \mathrm{dom}_f,\, z \geq f(y) \\ &v^T x + \beta\mu - a > 0. \end{aligned}$$

If $\beta = 0$, this means that $x \notin \mathrm{dom}_f$. Consider $z \in \mathrm{int}(\mathrm{dom}_f)$ and $\tilde{v} \in \partial f(z)$ (non empty by Proposition 5.2.1), for any $\lambda \geq 0$ and any $y \in \mathrm{dom}_f$

$$\lambda(v^T y - a) + \tilde{v}^T(y - z) + f(z) \leq f(y),$$

So that we have a family of affine minorants of f parametrized by $\lambda \geq 0$. Furthermore, $\lambda(v^T x - a) + \tilde{v}^T(x - z) + f(z)$ can be chosen arbitrarily big as $\lambda \to \infty$ and the supremum is $+\infty$.

Assume that $\beta \neq 0$, then $\beta < 0$ and we have for any $y \in \mathrm{dom}_f$

$$\frac{1}{-\beta}(v^T y - a) \leq f(y)$$

and furthermore $\frac{1}{-\beta}(v^T x - a) > \mu$. We obtain

$$\frac{1}{-\beta}(v^T y + z - a) \leq f(y), \qquad \forall y \in \mathrm{dom}_f$$
$$\mu < \frac{1}{-\beta}(v^T x + z - a) \leq \min\{f(x), +\infty\}$$

since μ is arbitrary, if $f(x)$ is finite, the supremum over all affine lower bounds is $f(x)$, if it is not finite, the supremum is $+\infty$. □

The following is due to Moreau and Rockafellar

Theorem 5.2.3. *For any $x \in \mathrm{int}(\mathrm{dom}_f)$ and any $h \in \mathbb{R}^p$,*

$$D_h f(x) = \sup_{v \in \partial f(x)} \langle v, h \rangle ,$$

where D_h denotes the directional derivative of f,

$$D_h f(x) = \lim_{t>0, t\to 0} \frac{f(x + th) - f(x)}{t}.$$

Proof. By convexity of f, $t \mapsto (f(x + th) - f(x))/t$ is an increasing function of $t > 0$. Indeed, for any $s > t$,

$$f(x + th) = f\left(\left(1 - \frac{t}{s}\right) x + \frac{t}{s}(x + sh)\right) \leq \left(1 - \frac{t}{s}\right) f(x) + \frac{t}{s} f(x + sh),$$

so that

$$\frac{f(x + th) - f(x)}{t} \leq \frac{f(x + sh) - f(x)}{s},$$

Using the Definition 5.2.1, we have for any $v \in \partial f(x)$, any $h \in \mathbb{R}^p$ and $t > 0$,

$$\frac{f(x + th) - f(x)}{t} \geq \langle v, h \rangle$$

which shows by letting $t \to 0$ and taking the supremum on v that

$$D_h f(x) \geq \sup_{v \in \partial f(x)} \langle v, h \rangle .$$

Hence $D_h f(x)$ is well defined for all h and we have one inequality. The function $g \colon h \mapsto D_h f(x)$ is convex, has full domain and is positively homogeneous. By Theorem 5.2.2, we have for any h

$$D_h f(x) = \sup_{r,v} \quad r + v^T h \qquad \text{s.t.} \qquad D_{h'}(x) \geq r + v^T h', \, \forall h' \in \mathbb{R}^p.$$

From positive homogeneity of g, the constraint enforce that for any $t > 0$, $D_{th'}f(x) = tD_{h'} \geq r + tv^T h'$, $\forall h' \in \mathbb{R}^p$ and $D_{h'}f(x) \geq v^T h'$, letting $t \to \infty$, so that r may be chosen to be 0. We deduce that

$$D_h f(x) = \sup_v \quad v^T h \qquad \text{s.t.} \qquad D_{h'}f(x) \geq v^T h', \forall h' \in \mathbb{R}^p.$$

We notice that if $D_{h'}f(x) \geq v^T h'$, $\forall h' \in \mathbb{R}^p$, then $f(x+h') - f(x) \geq v^T h'$ for all $h' \in \mathbb{R}^p$ so that v is a subgradient of f at x. We obtain

$$\begin{aligned} D_h f(x) &= \sup_v \quad v^T h \qquad \text{s.t.} \qquad D_{h'}f(x) \geq v^T h', \forall h' \in \mathbb{R}^p \\ &\leq \sup_{v \in \partial f(x)} \quad v^T h. \end{aligned}$$

□

We deduce from this result that f is differentiable at $x \in \text{int}(\text{dom}_f)$ if and only if $\partial f(x) = \{\nabla f(x)\}$.

5.2.2 Legendre transform

Definition 5.2.2. *Given f convex, the Fenchel-Legendre transform of f is given as follows*

$$f^* \colon z \mapsto \sup_{y \in \mathbb{R}^p} z^T y - f(y)$$

Theorem 5.2.4. *For any f convex, f^* is convex and for any $x, z \in \mathbb{R}^p$*

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

and the preceeding inequality holds if and only if $z \in \partial f(x)$. This is called Fenchel-Young's inequality. Furthermore, if f is lower semicontinuous if and only if $(f^*)^* = f$.

Proof. Convexity follows because f^* is the pointwise supremum of affine functions which are convex and convexity is preserved by pointwise suprema. If we have equality, this means that x attains the minimum of the convex function $y \mapsto f(y) - y^T z$ and we must have zero in the subdifferential of this function at x.

From Fenchel-Young's inequality, we have that $f(x) \geq z^T x - f^*(z)$ for all z so that taking the supremum over z, we obtain $f(x) \geq (f^*)^*(x)$ to get equality, we use Theorem 5.2.2. For any $x \in \mathbb{R}^p$

$$\begin{aligned} (f^*)^*(x) &= \sup_{v \in \mathbb{R}^p} v^T x - f^*(v) \\ &= \sup_{v \in \mathbb{R}^p} v^T x - \sup_{y \in \mathbb{R}^p} v^T y - f(y) \\ &= \sup_{v \in \mathbb{R}^p} v^T x + \inf_{y \in \mathbb{R}^p} f(y) - v^T y \\ &= \sup_{v \in \mathbb{R}^p} v^T x + \sup_{r \in \mathbb{R}} r, \qquad \text{s.t.} \qquad f(y) - v^T y \geq r, \quad \forall y \in \mathbb{R}^p \\ &= \sup_{v, r \in \mathbb{R}^p} v^T x + r, \qquad \text{s.t.} \qquad f(y) \geq r + v^T y, \quad \forall y \in \mathbb{R}^p \end{aligned}$$

Hence f^{**} is the supremum of all affine lower bounds of f. As such it is always lower-semicontinuous since its graph is an intersection of closed sets which is closed. Furthermore, when f is lower-semicontinuous, we obtain $f^{**} = f$. □

Example 5.2.1. *Let $f \colon x \mapsto \max_i x_i$, compute the subgradient of this function.*

5.3 Subgradient descent

Subgradient descent generalizes gradient descent to nonsmooth functions.

Proposition 5.3.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function which attains its infimum and has full domain. Consider the algorithm, for $x_0 \in \mathbb{R}^p$, a sequence of positive numbers $\alpha_k > 0$, $k \in \mathbb{N}$, iterate*

$$x_{k+1} = x_k - \alpha_k v_k \tag{5.9}$$
$$v_k \in \partial f(x_k). \tag{5.10}$$

Then for any global minimizer x^*, setting, $y_k = \sum_{i=0}^k \alpha_i x_i / \left(\sum_{i=0}^k \alpha_i\right)$

$$\min_{i=1,\dots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}$$
$$f(y_k) - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}.$$

Proof. We have for any $k \in \mathbb{N}$

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x^*\|_2^2 &= \frac{1}{2}\|x_k - \alpha_k v_k - x^*\|_2^2 \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k v_k^T (x^* - x_k) + \frac{\alpha_k^2}{2}\|v_k\|_2^2 \\ &\leq \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k (f(x^*) - f(x_k)) + \frac{\alpha_k^2}{2}\|v_k\|_2^2.\end{aligned}$$

By summing up, we obtain

$$\frac{\sum_{i=0}^k \alpha_i (f(x_i) - f^*)}{\sum_{i=0}^k \alpha_i} \leq \frac{\|x_0 - x_*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}$$

and the result follows from convexity of f. □

Corollary 5.3.1. *If f is L-Lipschitz, we have the following convergence result for subgradient method.*

- *If $\alpha_k = \alpha$ is constant, we have*

$$\min_{i=1,\dots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{L^2\alpha}{2}.$$

- *In particular, choosing $\alpha_i = \frac{\|x_0 - x^*\|/L}{\sqrt{k+1}}$, we have*

$$\min_{i=1,\dots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\| L}{\sqrt{k+1}}.$$

- *Choosing $\alpha_k = \|x_0 - x^*\|/(L\sqrt{k})$ for all k, we obtain for all k*

$$\min_{i=1,\dots,k} f(x_k) - f^* = O\left(\frac{\|x_0 - x^*\|_2 L(1 + \log(k))}{\sqrt{k}}\right).$$

5.4 Composite optimization

The subgradient method is slow in practice. Furthermore, convergence depends a lot on step size tuning. In favorable situations there exists better suited algorithms. Good introduction to the topic of proximal algorithms with connection to statistics and signal processing are found in [14, 2].

5.4.1 Motivation

The Lasso estimator is given by

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1.$$

This is the solution of a nonsmooth convex optimization problem. The subgradient method can be used to solve this problem as it can be used to solve any continuous convex optimization problem for which subgradients are available. However, this method is slow and hard to tune in practice. It turns out that the objective function has additional structure which can be leveraged to devise more powerful and easier to implements algorithms. Indeed the objective function is of the form $f + g$ where f is a smooth (quadratic) convex function and g is the ℓ_1 norm, a nonsmooth convex function. Objective functions falling in this classe are sometimes called "composite objectives". Under additional restriction on g (easily computable proximity operator), there exists numerical algorithm which efficiency is comparable to the that of gradient descent for smooth optimization.

5.4.2 Proximity operator

The construction of the following object is due to Jean-Jacques Moreau [20].

Definition 5.4.1. *Given a closed convex function, $f\colon \mathbb{R}^d \mapsto \mathbb{R}$, the proximity operator of f is defined as follows*

$$\mathrm{prox}_f\colon z \mapsto \arg\min_{y\in\mathbb{R}^d} f(y) + \frac{1}{2}\|y - z\|_2^2.$$

By strong convexity, the minimum is attained and is strict.

Note that we have $x = \mathrm{prox}_f(z)$ if and only if $z = \partial f(x) + x$ and the proximity operator is someties denoted $(\partial f + I)^{-1}$.

Exercise 5.4.1. *Describe the* prox *applications for the following functions:*

- *A constant*
- *A linear function*
- *The indicator of a closed convex set C:*

$$\delta\colon x \mapsto \begin{cases} 0 \text{ if } x \in C \\ +\infty \text{ otherwise} \end{cases}$$

- *The function $x \mapsto \frac{1}{2}\|x\|_2^2$*
- *The function $x \mapsto \|x\|_2$*
- *The function $x \mapsto \|x\|_1$*

Exercise 5.4.2. *Let f and g be convex. Show that $\partial(f + g)(x) \supset \partial f(x) + \partial g(x)$ for every x such that $\partial f(x)$ and $\partial g(x)$ are non empty.*

Lemma 5.4.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient and g be convex lower semicontinuous. Fix any $x \in \mathbb{R}^p$ and set*

$$y = \mathrm{prox}_{g/L}\left(x - \frac{1}{L}\nabla f(x)\right).$$

Then, for any $z \in \mathbb{R}^d$,

$$f(z) + g(z) + \frac{L}{2}\|x - z\|_2^2 \geq f(y) + g(y) + \frac{L}{2}\|y - z\|_2^2.$$

Proof. First, the descent Lemma ensures that for any $k \in \mathbb{N}$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \tag{5.11}$$

We have

$$y = \arg\min_{y \in \mathbb{R}^p} f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + g(y), \tag{5.12}$$

so that by strong convexity, for all $z \in \mathbb{R}^d$,

$$\begin{aligned} & f(x) + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|_2^2 + g(z) \\ \geq \quad & f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + g(y) + \frac{L}{2}\|z - y\|_2^2. \end{aligned} \tag{5.13}$$

Combining (5.11) and (5.13), we obtain

$$\begin{aligned} & f(z) + g(z) + \frac{L}{2}\|z - x\|_2^2 \\ & \geq f(x) + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|_2^2 + g(z) && \text{convexity} \\ & \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + g(y) + \frac{L}{2}\|y - z\|_2^2 && (5.13) \\ & \geq f(y) + g(y) + \frac{L}{2}\|y - z\|_2^2, && (5.11) \end{aligned}$$

□

Proposition 5.4.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient and g be convex lower semicontinuous such that $\rho = \inf_{x \in \mathbb{R}^p} f(x) + g(x) > -\infty$ is attained at x^*. Consider the algorithm, for $x_0 \in \mathbb{R}^p$ and*

$$x_{k+1} = \mathrm{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right). \tag{5.14}$$

Then x_k converges to a global minimum and we have for any $k \in \mathbb{N}$, $k > 0$,

$$f(x_k) + g(x_k) - \rho \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

If in addition $f + g$ is μ-strongly convex, we have in addition

$$\|x_{k+1} - x^*\|_2^2 \leq \frac{L}{L + \mu}\|x_k - x^*\|_2^2.$$

Proof. From Lemma 5.4.1 with $x = x_k = z$ and $y = x_{k+1}$ we read that $f + g$ is decreasing along the sequence. From Lemma 5.4.1 with $x = x_k$ z^* and $y = x_{k+1}$ we read that for any $k \in \mathbb{N}$

$$f(x^*) + g(x^*) + \frac{L}{2}\|x_k - x^*\|_2^2 \geq f(x_{k+1}) + g(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x^*\|_2^2.$$

By summing up, we obtain for any $K \in \mathbb{N}$, $K \geq 1$,

$$\frac{L}{2}\|x^* - x_0\|_2^2 \geq \sum_{k=1}^{K} f(x_k) + g(x_k) - \rho \geq K(f(x_K) + g(x_K) - \rho).$$

We also have that $\|x_k - x^*\|_2^2$ is decreasing and the convergence follows (this is Opial's Lemma). For the last statement, Lemma 5.4.1 with $y = x_{k+1}$, $z = x^*$ and $x = x_k$ combined with μ-strong convexity gives

$$\begin{aligned} f(x^*) + g(x^*) + \frac{L}{2}\|x_k - x^*\|_2^2 &\geq f(x_{k+1}) + g(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x^*\|_2^2 \\ &\geq f(x^*) + g(x^*) + \frac{L+\mu}{2}\|x_{k+1} - x^*\|_2^2, \end{aligned}$$

which is the desired result. □

5.5 Acceleration

We have obtained $1/k$ convergence rates for the gradient algorithm and the proximal gradient algorithm. Could we do better?

5.5.1 A lower bound

This is taken from Bubeck's book and originally due to Nesterov. Such lower bounds first appeared in the book of Nemirowski.

Definition 5.5.1. *A first order method to minimize a smooth convex function f when initiated at $x_0 = 0$, produces a sequence of points $(x_i)_{i\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$,*

$$x_{k+1} \in \operatorname{span}\left(\nabla f(x_0), \ldots, \nabla f(x_k)\right).$$

Theorem 5.5.1. *Let $k \leq (d-1)/2$, $L > 0$. There exists a convex function f with L-Lipschitz gradient over $\mathbb{R}^d$, such that for any first order method satisfying definition* (5.5.1)*,*

$$\min_{1\leq s\leq k} f(x_s) - f(x^*) \geq \frac{3L}{32}\frac{\|x_1 - x^*\|^2}{(k+1)^2}.$$

Proof. In this proof for $h : \mathbb{R}^d \to \mathbb{R}$ we denote $h^* = \inf_{x\in\mathbb{R}^d} h(x)$. For $k \leq d$ let $A_k \in \mathbb{R}^{d\times d}$ be the symmetric and tridiagonal matrix defined by

$$(A_k)_{i,j} = \begin{cases} 2, & i = j, i \leq k \\ -1, & j \in \{i-1, i+1\}, i \leq k, j \neq k+1 \\ 0, & \text{otherwise.} \end{cases}$$

We verify that $0 \preceq A_k \preceq 4I$ since

$$x^\top A_k x = 2\sum_{i=1}^{k} x(i)^2 - 2\sum_{i=1}^{k-1} x(i)x(i+1) = x(1)^2 + x(k)^2 + \sum_{i=1}^{k-1}(x(i) - x(i+1))^2.$$

We consider now the following convex function:

$$f(x) = \frac{L}{8}x^\top A_{2k+1}x - \frac{L}{4}x^\top e_1.$$

For any $s = 1, \dots, k$, x_s must lie in the linear span of $e_1, \dots, e_{s-1}$ (because of our assumption on the black-box procedure). In particular for $s \leq k$ we necessarily have $x_s(i) = 0$ for $i = s, \dots, n$, which implies $x_s^\top A_{2k+1}x_s = x_s^\top A_k x_s$. In other words, if we denote

$$f_k(x) = \frac{L}{8}x^\top A_k x - \frac{L}{4}x^\top e_1,$$

We proved that, for all $s \leq k$

$$f(x_s) - f^* = f_k(x_s) - f_{2k+1}^* \geq f_k^* - f_{2k+1}^*.$$

Thus it simply remains to compute the minimizer x_k^* of f_k, its norm, and the corresponding function value f_k^*.

The point x_k^* is the unique solution in the span of $e_1, \dots, e_k$ of $A_k x = e_1$. One can verify (Exercise) that it is defined by $x_k^*(i) = 1 - \frac{i}{k+1}$ for $i = 1, \dots, k$. Thus we have:

$$f_k^* = \frac{L}{8}(x_k^*)^\top A_k x_k^* - \frac{L}{4}(x_k^*)^\top e_1 = -\frac{L}{8}(x_k^*)^\top e_1 = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).$$

Furthermore note that

$$\|x_k^*\|^2 = \sum_{i=1}^{k}\left(1 - \frac{i}{k+1}\right)^2 = \sum_{i=1}^{k}\left(\frac{i}{k+1}\right)^2 \leq \frac{k+1}{3}.$$

Thus one obtains:

$$f_k^* - f_{2k+1}^* = \frac{L}{4}\left(\frac{1}{k+1} - \frac{1}{2k+2}\right) \geq \frac{3L}{32}\frac{\|x_{2k+1}^*\|^2}{(k+1)^2},$$

□

5.5.2 Accelerated algorithm

The previous lower bound shows that there is a gap between the convergence speed of gradient descent for smooth convex functions and the and the lower bound. It remained an open question if the gap was due to gradient descent or if it was due to the fact that the lower bound is loose until Nesterov published in 1983 an algorithm which achieves $1/k^2$ rate [25]. We extend bellow the original proof of Nesterov. An extension to the proximal setting has been developped by Beck and Teboulle in [5].

Theorem 5.5.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient $\inf_{x\in\mathbb{R}^p} f(x) > -\infty$. Consider the algorithm, for $x_{-1} \in \mathbb{R}^p$, set $y_0 = x_{-1}$, $t_1 = 1$ and for $k \in \mathbb{N}$,*

$$\begin{aligned} x_k &= y_k - \frac{1}{L}\nabla f(y_k) \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ y_{k+1} &= x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1}). \end{aligned} \tag{5.15}$$

Then for any $k \in \mathbb{N}$

$$f(x_k) - f^* \leq \frac{4L\|x_0 - x^*\|_2^2}{(k+2)^2}.$$

Proof. We introduce the following notation which is taken from the original proof, for any $k \in \mathbb{N}$,

$$p_k := (t_k - 1)(x_{k-1} - x_k) \qquad \text{so that} \qquad y_{k+1} = x_k - \frac{p_k}{t_{k+1}}$$

First, we have for any $k \geq 1$

$$t_k \geq \frac{1 + \sqrt{4t_{k-1}^2 + 1}}{2} \geq t_{k-1} + \frac{1}{2} \geq t_0 + \frac{k}{2} = 1 + \frac{k}{2}. \tag{5.16}$$

$$(t_{k+1}^2 - t_{k+1}) = t_k^2. \tag{5.17}$$

The main argument of the proof is the following. The sequence $\{z_k\}_{k\in\mathbb{N}}$ defined as

$$z_k := \frac{2t_k^2}{L}(f(x_k) - f^*) + \|p_k - x_k + x^*\|^2, \tag{5.18}$$

is non-increasing and $z_0 \leq 2||x_0 - x^*||^2$. The result can be deduced by combining (5.16) and (5.18).

We have a series of three inequalities.

$$\begin{aligned}
p_{k+1} - x_{k+1} &= p_k - x_k + \frac{t_{k+1}}{L}\nabla f(y_{k+1}) \\
p_{k+1} - x_{k+1} &= (t_{k+1} - 1)(x_k - x_{k+1}) - x_{k+1} \\
&= (t_{k+1} - 1)x_k - t_{k+1}x_{k+1} \\
&= (t_{k+1} - 1)x_k - t_{k+1}\left(y_{k+1} - \frac{1}{L}\nabla f(y_{k+1})\right) \\
&= (t_{k+1} - 1)x_k - t_{k+1}x_k - (t_k - 1)(x_k - x_{k-1}) - \frac{t_{k+1}}{L}\nabla f(y_{k+1}) \\
&= p_k - x_k + \frac{t_{k+1}}{L}\nabla f(y_{k+1})
\end{aligned}$$

This implies

$$\begin{aligned}
\|p_{k+1} - x_{k+1} + x^*\|_2^2 &= \|p_k - x_k + \frac{t_{k+1}}{L}\nabla f(y_{k+1}) + x^*\|_2^2 \\
&= \|p_k - x_k + x^*\|_2^2 + 2\left\langle p_k - x_k + x^*, \frac{t_{k+1}}{L}\nabla f(y_{k+1})\right\rangle \\
&\quad + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
y_{k+1} &= x_k - \frac{p_k}{t_{k+1}} \\
\left\langle p_k - x_k + x^*, \frac{t_{k+1}}{L}\nabla f(y_{k+1})\right\rangle &= \left\langle p_k - y_{k+1} - \frac{p_k}{t_{k+1}} + x^*, \frac{t_{k+1}}{L}\nabla f(y_{k+1})\right\rangle \\
&= \frac{(t_{k+1} - 1)}{L}\left\langle p_k, \nabla f(y_{k+1})\right\rangle + \frac{t_{k+1}}{L}\left\langle x^* - y_{k+1}, \nabla f(y_{k+1})\right\rangle \\
\|p_{k+1} - x_{k+1} + x^*\|_2^2 &= \|p_k - x_k + x^*\|_2^2 + 2\frac{(t_{k+1} - 1)}{L}\left\langle p_k, \nabla f(y_{k+1})\right\rangle \\
&\quad + 2\frac{t_{k+1}}{L}\left\langle x^* - y_{k+1}, \nabla f(y_{k+1})\right\rangle + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2
\end{aligned}$$

From the Lipschitz gradient assumption, we obtain

$$f(x_{k+1}) - f^* \leq f(y_{k+1}) - f^* - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2 \leq \langle \nabla f(y_{k+1}), y_{k+1} - x^*\rangle - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2$$

$$\frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2 \leq f(y_{k+1}) - f(x_{k+1}) \leq f(x_k) - f(x_{k+1}) - \frac{1}{t_{k+1}}\left\langle p_k, \nabla f(y_{k+1})\right\rangle$$

Using the last three identities, we obtain

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|_2^2 - \|p_k - x_k + x^*\|_2^2 \\
&= 2\frac{(t_{k+1}-1)}{L}\langle p_k, \nabla f(y_{k+1})\rangle + 2\frac{t_{k+1}}{L}\langle x^* - y_{k+1}, \nabla f(y_{k+1})\rangle + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
&\leq 2t_{k+1}\frac{(t_{k+1}-1)}{L}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2\right) \\
&\quad + 2\frac{t_{k+1}}{L}\left(f^* - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2\right) + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
&= 2t_{k+1}\frac{(t_{k+1}-1)}{L}\left(f(x_k) - f^* + f^* - f(x_{k+1})\right) + 2\frac{t_{k+1}}{L}\left(f^* - f(x_{k+1})\right) \\
&= 2\frac{t_k^2}{L}\left(f(x_k) - f^*\right) - 2\frac{t_{k+1}^2}{L}\left(f(x_{k+1} - f^*)\right)
\end{aligned}$$

where we used (5.16) for the last step. This proves that the sequence $(z_k)_{k\in\mathbb{N}}$ is non increasing. It remains to compute z_0,

$$z_0 = \frac{2}{L}(f(x_0) - f^*) + \|x^* - x_0\|^2 \leq 2\|x_0 - x^*\|_2^2.$$

Putting things together

$$f(x_k) - f^* \leq \frac{Lz_0}{2t_k^2} \leq \frac{4L\|x_0 - x^*\|_2^2}{(k+2)^2}.$$

□

5.6 Non convex problems

Most algorithm described in this chapter have extensions to nonconvex problems. In this setting, the only hope is to find first order critical points instead of global minima. The notion of subgradient in this case has to be treated with a lot of care. A reference on the topic is [28].

Exercises

Exercise 5.6.1. *Show that if $f\colon \mathbb{R}^d \mapsto \mathbb{R}$ is $\mathcal{C}^2$ then any accumulation point of the system $\dot{x} = -\nabla f(x)$ is a critical point of f.*

Exercise 5.6.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function,*

- *Show that ∂f is sequencialy closed in the sence that, for any $\bar{x}$*

$$\left\{v \in \mathbb{R}^p,\, \exists\, (x_k, v_k)_{k\in\mathbb{N}}\,,\, x_k \to \bar{x},\, v_k \to v,\, f(x_k) \to f(\bar{x})\right\} \subset \partial f(\bar{x})$$

- *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, show that f is L-Lipschitz if and only if $\sup_{x\in\mathbb{R}^p,\, v\in\partial f(x)} \|v\|_2 \leq L$.*

Exercise 5.6.3.

- *Let $f\colon \mathbb{R} \mapsto \mathbb{R}$ be convex (with full domain), show that for any $s < t < u$,*

$$\frac{f(t) - f(s)}{t - s} \leq \frac{f(u) - f(s)}{u - s} \leq \frac{f(u) - f(t)}{u - t}.$$

- *Deduce that f is continuous on $\mathbb{R}$.*

Exercise 5.6.4.

- *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex (with full domain), show that for any $x \in \mathbb{R}^d$ and $h \in \mathbb{R}^d*$, with $\|h\|_1 < 1$, we have*

$$f(x + h) \leq (1 - \|h\|_1) f(x) + \|h\|_1 \max_{i=1,\dots,d} f(x \pm e_i)$$

  where e_i are elements of the canonical basis.

- *Deduce that f is continuous at x.*

- *What can you say about an extended valued convex function which domain has nonempty interior?*

Exercise 5.6.5. *Let $\|\cdot\|$ be a norm, it is then convex. Its dual norm is defined by*

$$\|z\|_* = \sup z^T x \qquad \textit{such that} \qquad \|x\| \leq 1.$$

Consider the function $f\colon x \mapsto \|x\|$, compute the Legendre transform of f.

Exercise 5.6.6. *Let $f_i\colon \mathbb{R}^d \mapsto \mathbb{R}$ be convex and differentiable on $\mathbb{R}^d$ for $i = 1 \dots n$. Set $F\colon x \mapsto \max_i f_i(x)$. Show that*

$$\partial F(x) = \operatorname{conv}\left(\{\nabla f_i(x),\, f_i(x) = F(x)\}\right).$$

How does this result extend to non differentiable convex functions?

Exercise 5.6.7. *Let $f\colon \mathbb{R}^d \mapsto \mathbb{R}$ be convex with full domain. Show that f is upper bounded if and only if f is constant.*

Exercise 5.6.8. *Describe the* prox *applications for the following functions: a constant, a linear function, the indicator of a closed convex set C:*

$$\delta\colon x \mapsto \begin{cases} 0 \text{ if } x \in C \\ +\infty \text{ otherwise} \end{cases}$$

The function $x \mapsto \frac{1}{2}\|x\|_2^2$, the function $x \mapsto \|x\|_2$, the function $x \mapsto \|x\|_1$

Exercise 5.6.9. *Let f and g be convex. Show that $\partial(f + g)(x) \supset \partial f(x) + \partial g(x)$ for every x such that $\partial f(x)$ and $\partial g(x)$ are non empty. What do you think about the reverse inclusion?*