Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

November 18, 2020

Chapter 13

Optimization & Machine Learning: Smooth Optimization

13.1 Motivation in Machine Learning

13.1.1 Unconstraint optimization

In most part of this Chapter, we consider unconstrained convex optimization problems of the form

$$\inf_{x \in \mathbb{R}^p} f(x), \tag{13.1}$$

and try to devise "cheap" algorithms with a low computational cost per iteration to approximate a minimizer when it exists. The class of algorithms considered are first order, i.e. they make use of gradient information. In the following, we denote

$$\operatorname*{argmin}_x f(x) \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^p \,;\, f(x) = \inf f\},$$

to indicate the set of points (it is not necessarily a singleton since the minimizer might be non-unique) that achieve the minimum of the function f. One might have $\operatorname{argmin} f = \emptyset$ (this situation is discussed below), but in case a minimizer exists, we denote the optimization problem as

$$\min_{x \in \mathbb{R}^p} f(x). \tag{13.2}$$

In typical learning scenario, $f(x)$ is the empirical risk for regression or classification, and p is the number of parameter. For instance, in the simplest case of linear models, we denote $(a_i, y_i)_{i=1}^n$ where $a_i \in \mathbb{R}^p$ are the features. In the following, we denote $A \in \mathbb{R}^{n \times p}$ the matrix whose rows are the a_i.

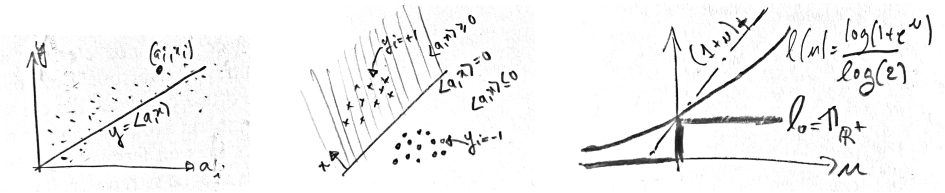


Figure 13.1: Left: linear regression, middle: linear classifier, right: loss function for classification.

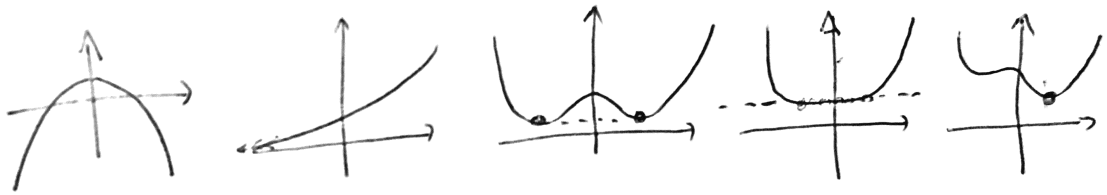


Figure 13.2: Left: non-existence of minimizer, middle: multiple minimizers, right: uniqueness.

13.1.2 Regression

For regression, $y_i \in \mathbb{R}$, in which case

$$f(x) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle x,\, a_i\rangle)^2 = \frac{1}{2}\|Ax - y\|^2, \tag{13.3}$$

is the least square quadratic risk function (see Fig. 13.1). Here $\langle u,\, v\rangle = \sum_{i=1}^{p} u_i v_i$ is the canonical inner product in $\mathbb{R}^p$ and $\|\cdot\|^2 = \langle \cdot,\, \cdot\rangle$.

13.1.3 Classification

For classification, $y_i \in \{-1, 1\}$, in which case

$$f(x) = \sum_{i=1}^{n} \ell(-y_i\langle x,\, a_i\rangle) = L(-\operatorname{diag}(y)Ax) \tag{13.4}$$

where ℓ is a smooth approximation of the 0-1 loss $1_{\mathbb{R}^+}$. For instance $\ell(u) = \log(1 + \exp(u))$, and $\operatorname{diag}(y) \in \mathbb{R}^{n\times n}$ is the diagonal matrix with y_i along the diagonal (see Fig. 13.1, right). Here the separable loss function $L = \mathbb{R}^n \to \mathbb{R}$ is, for $z \in \mathbb{R}^n$, $L(z) = \sum_i \ell(z_i)$.

13.2 Basics of Convex Analysis

13.2.1 Existence of Solutions

In general, there might be no solution to the optimization (13.1). This is of course the case if f is unbounded by below, for instance $f(x) = -x^2$ in which case the value of the minimum is $-\infty$. But this might also happen if f does not grow at infinity, for instance $f(x) = e^{-x}$, for which $\min f = 0$ but there is no minimizer.

In order to show existence of a minimizer, and that the set of minimizer is bounded (otherwise one can have problems with optimization algorithm that could escape to infinity), one needs to show that one can replace the whole space $\mathbb{R}^p$ by a compact sub-set $\Omega \subset \mathbb{R}^p$ (i.e. Ω is bounded and close) and that f is continuous on Ω (one can replace this by a weaker condition, that f is lower-semi-continuous, but we ignore this here). A way to show that one can consider only a bounded set is to show that $f(x) \to +\infty$ when $x \to +\infty$. Such a function is called coercive. In this case, one can choose any $x_0 \in \mathbb{R}^p$ and consider its associated lower-level set

$$\Omega = \{x \in \mathbb{R}^p \;;\; f(x) \leqslant f(x_0)\}$$

which is bounded because of coercivity, and closed because f is continuous. One can actually show that for convex function, having a bounded set of minimizer is equivalent to the function being coercive (this is not the case for non-convex function, for instance $f(x) = \min(1, x^2)$ has a single minimum but is not coercive).

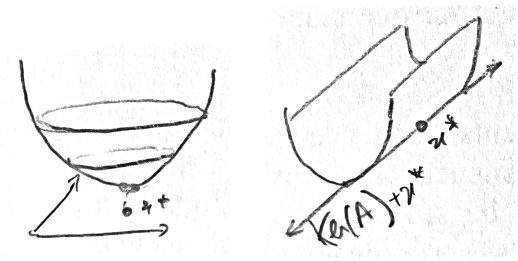


Figure 13.3: Coercivity condition for least squares.

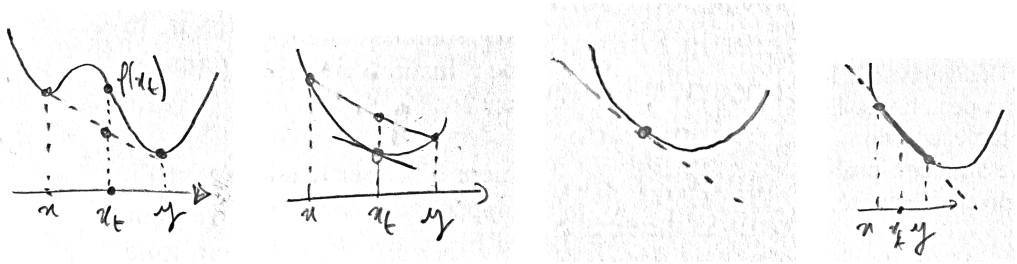


Figure 13.4: Convex vs. non-convex functions ; Strictly convex vs. non strictly convex functions.

Example 1 (Least squares). For instance, for the quadratic loss function $f(x) = \frac{1}{2}\|Ax - y\|^2$, coercivity holds if and only if $\ker(A) = \{0\}$ (this corresponds to the overdetermined setting). Indeed, if $\ker(A) \neq \{0\}$ if $x^\star$ is a solution, then $x^\star + u$ is also solution for any $u \in \ker(A)$, so that the set of minimizer is unbounded. On contrary, if $\ker(A) = \{0\}$, we will show later that the set of minimizer is unique, see Fig. 13.3. If ℓ is strictly convex, the same conclusion holds in the case of classification.

13.2.2 Convexity

Convex functions define the main class of functions which are somehow "simple" to optimize, in the sense that all minimizers are global minimizers, and that there are often efficient methods to find these minimizers (at least for smooth convex functions). A convex function is such that for any pair of point $(x, y) \in (\mathbb{R}^p)^2$,

$$\forall\, t \in [0,1], \quad f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) \tag{13.5}$$

which means that the function is below its secant (and actually also above its tangent when this is well defined), see Fig. 13.4. If $x^\star$ is a local minimizer of a convex f, then $x^\star$ is a global minimizer, i.e. $x^\star \in \operatorname{argmin} f$.

Convex function are very convenient because they are stable under lots of transformation. In particular, if f, g are convex and a, b are positive, $af + bg$ is convex (the set of convex function is itself an infinite dimensional convex cone!) and so is $\max(f, g)$. If $g : \mathbb{R}^q \to \mathbb{R}$ is convex and $B \in \mathbb{R}^{q\times p}, b \in \mathbb{R}^q$ then $f(x) = g(Bx+b)$ is convex. This shows immediately that the square loss appearing in (13.3) is convex, since $\|\cdot\|^2/2$ is convex (as a sum of squares). Also, similarly, if ℓ and hence L is convex, then the classification loss function (13.4) is itself convex.

Strict convexity. When f is convex, one can strengthen the condition (13.5) and impose that the inequality is strict for $t \in]0,1[$ (see Fig. 13.4, right), i.e.

$$\forall\, t \in]0,1[, \quad f((1-t)x + ty) < (1-t)f(x) + tf(y). \tag{13.6}$$

In this case, if a minimum $x^\star$ exists, then it is unique. Indeed, if $x_1^\star \neq x_2^\star$ were two different minimizer, one would have by strict convexity $f(\frac{x_1^\star + x_2^\star}{2}) < f(x_1^\star)$ which is impossible.

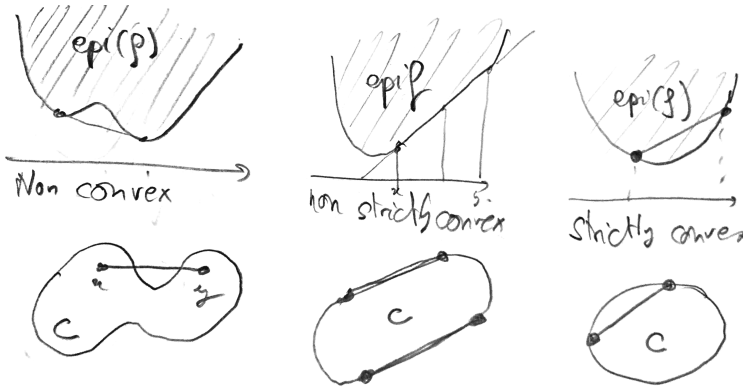


Figure 13.5: Comparison of convex functions $f : \mathbb{R}^p \to \mathbb{R}$ (for $p = 1$) and convex sets $C \subset \mathbb{R}^p$ (for $p = 2$).

Example 2 (Least squares). For the quadratic loss function $f(x) = \frac{1}{2}\|Ax - y\|^2$, strict convexity is equivalent to $\ker(A) = \{0\}$. Indeed, we see later that its second derivative is $\partial^2 f(x) = A^\top A$ and that strict convexity is implied by the eigenvalues of $A^\top A$ being strictly positive. The eigenvalues of $A^\top A$ being positive, it is equivalent to $\ker(A^\top A) = \{0\}$ (no vanishing eigenvalue), and $A^\top Az = 0$ implies $\langle A^\top Az,\, z\rangle = \|Az\|^2 = 0$ i.e. $z \in \ker(A)$.

13.2.3 Convex Sets

A set $\Omega \subset \mathbb{R}^p$ is said to be convex if for any $(x, y) \in \Omega^2$, $(1-t)x + ty \in \Omega$ for $t \in [0, 1]$. The connexion between convex function and convex sets is that a function f is convex if and only if its epigraph $\mathrm{epi}(f) \overset{\text{def.}}{=} \big\{(x, t) \in \mathbb{R}^{p+1} \;;\; t \geqslant f(x)\big\}$ is a convex set.

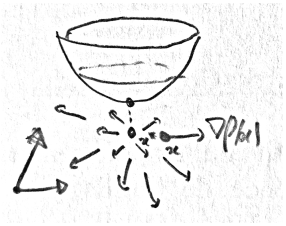
Remark 2 (Convexity of the set of minimizers). In general, minimizers $x^\star$ might be non-unique, as shown on Figure 13.3. When f is convex, the set $\mathrm{argmin}(f)$ of minimizers is itself a convex set. Indeed, if $x_1^\star$ and $x_2^\star$ are minimizers, so that in particular $f(x_1^\star) = f(x_2^\star) = \min(f)$, then $f((1-t)x_1^\star + tx_2^\star) \leqslant (1-t)f(x_1^\star) + tf(x_2^\star) = f(x_1^\star) = \min(f)$, so that $(1-t)x_1^\star + tx_2^\star$ is itself a minimizer. Figure 13.5 shows convex and non-convex sets.

13.3 Derivative and gradient

13.3.1 Gradient

If f is differentiable along each axis, we denote

$$\nabla f(x) \overset{\text{def.}}{=} \left(\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_p}\right)^\top \in \mathbb{R}^p$$



the gradient vector, so that $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$ is a vector field. Here the partial derivative (when they exits) are defined as

$$\frac{\partial f(x)}{\partial x_k} \overset{\text{def.}}{=} \lim_{\eta \to 0} \frac{f(x + \eta\delta_k) - f(x)}{\eta}$$

where $\delta_k = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ is the k^{th} canonical basis vector.

Beware that $\nabla f(x)$ can exist without f being differentiable. Differentiability of f at each reads

$$f(x + \varepsilon) = f(x) + \langle \varepsilon,\, \nabla f(x)\rangle + o(\|\varepsilon\|). \tag{13.7}$$

Here $R(\varepsilon) = o(\|\varepsilon\|)$ denotes a quantity which decays faster than ε toward 0, i.e. $\frac{R(\varepsilon)}{\|\varepsilon\|} \to 0$ as $\varepsilon \to 0$. Existence of partial derivative corresponds to f being differentiable along the axes, while differentiability should hold

for any converging sequence of $\varepsilon \to 0$ (i.e. not along along a fixed direction). A counter example in 2-D is $f(x) = \frac{2x_1x_2(x_1+x_2)}{x_1^2+x_2^2}$ with $f(0) = 0$, which is affine with different slope along each radial lines.

Also, $\nabla f(x)$ is the only vector such that the relation (13.7). This means that a possible strategy to both prove that f is differentiable and to obtain a formula for $\nabla f(x)$ is to show a relation of the form

$$f(x+\varepsilon) = f(x) + \langle \varepsilon,\, g\rangle + o(\|\varepsilon\|),$$

in which case one necessarily has $\nabla f(x) = g$.

The following proposition shows that convexity is equivalent to the graph of the function being above its tangents.

Proposition 38. *If f is differentiable, then*

$$f \text{ convex} \quad \Leftrightarrow \quad \forall (x,x'),\ f(x) \geqslant f(x') + \langle \nabla f(x'),\, x - x'\rangle.$$

Proof. One can write the convexity condition as

$$f((1-t)x + tx') \leqslant (1-t)f(x) + tf(x') \quad \Longrightarrow \quad \frac{f(x+t(x'-x)) - f(x)}{t} \leqslant f(x') - f(x)$$

hence, taking the limit $t \to 0$ one obtains

$$\langle \nabla f(x),\, x' - x\rangle \leqslant f(x') - f(x).$$

For the other implication, we apply the right condition replacing (x, x') by $(x, x_t \overset{\text{def.}}{=} (1-t)x + tx')$ and $(x', (1-t)x + tx')$

$$\begin{aligned} f(x) &\geqslant f(x_t) + \langle \nabla f(x_t),\, x - x_t\rangle = f(x_t) - t\langle \nabla f(x_t),\, x - x'\rangle \\ f(x') &\geqslant f(x_t) + \langle \nabla f(x_t),\, x' - x_t\rangle = f(x_t) + (1-t)\langle \nabla f(x_t),\, x - x'\rangle, \end{aligned}$$

multiplying these inequality by respectively $1-t$ and t, and summing them, gives

$$(1-t)f(x) + tf(x') \geqslant f(x_t).$$

□

13.3.2 First Order Conditions

The main theoretical interest (we will see later that it also have algorithmic interest) of the gradient vector is that it is a necessarily condition for optimality, as stated below.

Proposition 39. *If $x^\star$ is a local minimum of the function f (i.e. that $f(x^\star) \leqslant f(x)$ for all x in some ball around $x^\star$) then*

$$\nabla f(x^\star) = 0.$$

Proof. One has for ε small enough and u fixed

$$f(x^\star) \leqslant f(x^\star + \varepsilon u) = f(x^\star) + \varepsilon\langle \nabla f(x^\star),\, u\rangle + o(\varepsilon) \quad \Longrightarrow \quad \langle \nabla f(x^\star),\, u\rangle \geqslant o(1) \quad \Longrightarrow \quad \langle \nabla f(x^\star),\, u\rangle \geqslant 0.$$

So applying this for u and $-u$ in the previous equation shows that $\langle \nabla f(x^\star),\, u\rangle = 0$ for all u, and hence $\nabla f(x^\star) = 0$. □

Note that the converse is not true in general, since one might have $\nabla f(x) = 0$ but x is not a local mininimum. For instance $x = 0$ for $f(x) = -x^2$ (here x is a maximizer) or $f(x) = x^3$ (here x is neither a maximizer or a minimizer, it is a saddle point), see Fig. 13.6. Note however that in practice, if $\nabla f(x^\star) = 0$ but x is not a local minimum, then $x^\star$ tends to be an unstable equilibrium. Thus most often a gradient-based algorithm will converge to points with $\nabla f(x^\star) = 0$ that are local minimizers. The following proposition shows that a much strong result holds if f is convex.

Figure 13.6: Function with local maxima/minima (left), saddle point (middle) and global minimum (right).

Proposition 40. *If f is convex and $x^\star$ a local minimum, then $x^\star$ is also a global minimum. If f is differentiable and convex,*

$$x^\star \in \operatorname*{argmin}_x f(x) \quad \Longleftrightarrow \quad \nabla f(x^\star) = 0.$$

Proof. For any x, there exist $0 < t < 1$ small enough such that $tx + (1-t)x^\star$ is close enough to $x^\star$, and so since it is a local minimizer

$$f(x^\star) \leqslant f(tx + (1-t)x^\star) \leqslant tf(x) + (1-t)f(x^\star) \quad \Longrightarrow \quad f(x^\star) \leqslant f(x)$$

and thus $x^\star$ is a global minimum.

For the second part, we already saw in (39) the $\Leftarrow$ part. We assume that $\nabla f(x^\star) = 0$. Since the graph of x is above its tangent by convexity (as stated in Proposition 38),

$$f(x) \geqslant f(x^\star) + \langle \nabla f(x^\star),\, x - x^\star \rangle = f(x^\star).$$

□

Thus in this case, optimizing a function is the same a solving an equation $\nabla f(x) = 0$ (actually p equations in p unknown). In most case it is impossible to solve this equation, but it often provides interesting information about solutions $x^\star$.

13.3.3 Least Squares

The most important gradient formula is the one of the square loss (13.3), which can be obtained by expanding the norm

$$\begin{aligned} f(x+\varepsilon) &= \frac{1}{2}\|Ax - y + A\varepsilon\|^2 = \frac{1}{2}\|Ax-y\| + \langle Ax - y,\, A\varepsilon\rangle + \frac{1}{2}\|A\varepsilon\|^2 \\ &= f(x) + \langle \varepsilon,\, A^\top(Ax-y)\rangle + o(\|\varepsilon\|). \end{aligned}$$

Here, we have used the fact that $\|A\varepsilon\|^2 = o(\|\varepsilon\|)$ and use the transpose matrix $A^\top$. This matrix is obtained by exchanging the rows and the columns, i.e. $A^\top = (A_{j,i})_{i=1,\dots,n}^{j=1,\dots,p}$, but the way it should be remember and used is that it obeys the following swapping rule of the inner product,

$$\forall\, (u,v) \in \mathbb{R}^p \times \mathbb{R}^n, \quad \langle Au,\, v\rangle_{\mathbb{R}^n} = \langle u,\, A^\top v\rangle_{\mathbb{R}^p}.$$

Computing gradient for function involving linear operator will necessarily requires such a transposition step. This computation shows that

$$\nabla f(x) = A^\top(Ax - y). \tag{13.8}$$

This implies that solutions $x^\star$ minimizing $f(x)$ satisfies the linear system $(A^\top A)x^\star = A^\top y$. If $A^\star A \in \mathbb{R}^{p\times p}$ is invertible, then f has a single minimizer, namely

$$x^\star = (A^\top A)^{-1}A^\top y. \tag{13.9}$$

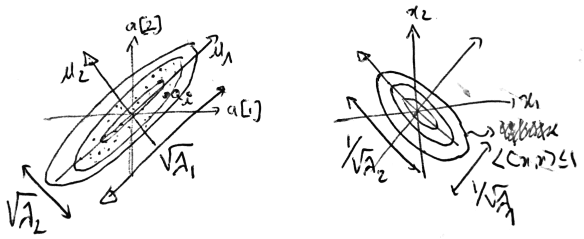


Figure 13.7: Left: point clouds $(a_i)_i$ with associated PCA directions, right: quadratic part of $f(x)$.

This shows that in this case, $x^\star$ depends linearly on the data y, and the corresponding linear operator $(A^\top A)^{-1}A^\star$ is often called the Moore-Penrose pseudo-inverse of A (which is not invertible in general, since typically $p \neq n$). The condition that $A^\top A$ is invertible is equivalent to $\ker(A) = \{0\}$, since

$$A^\top Ax = 0 \quad \Longrightarrow \quad \|Ax\|^2 = \langle A^\top Ax,\, x\rangle = 0 \quad \Longrightarrow \quad Ax = 0.$$

In particular, if $n < p$ (under-determined regime, there is too much parameter or too few data) this can never holds. If $n \geqslant p$ and the features x_i are "random" then $\ker(A) = \{0\}$ with probability one. In this overdetermined situation $n \geqslant p$, $\ker(A) = \{0\}$ only holds if the features $\{a_i\}_{i=1}^n$ spans a linear space $\mathrm{Im}(A^\top)$ of dimension strictly smaller than the ambient dimension p.

13.3.4 Link with PCA

Let us assume the $(a_i)_{i=1}^n$ are centered, i.e. $\sum_i a_i = 0$. If this is not the case, one needs to replace a_i by $a_i - m$ where $m \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n a_i \in \mathbb{R}^p$ is the empirical mean. In this case, $\frac{C}{n} = A^\top A/n \in \mathbb{R}^{p\times p}$ is the empirical covariance of the point cloud $(a_i)_i$, it encodes the covariances between the coordinates of the points. Denoting $a_i = (a_{i,1}, \ldots, a_{i,p})^\top \in \mathbb{R}^p$ (so that $A = (a_{i,j})_{i,j}$) the coordinates, one has

$$\forall\, (k,\ell) \in \{1,\ldots,p\}^2, \quad \frac{C_{k,\ell}}{n} = \frac{1}{n}\sum_{i=1}^n a_{i,k}a_{i,\ell}.$$

In particular, $C_{k,k}/n$ is the variance along the axis k. More generally, for any unit vector $u \in \mathbb{R}^p$, $\langle Cu,\, u\rangle/n \geqslant 0$ is the variance along the axis u.

For instance, in dimension $p = 2$,

$$\frac{C}{n} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n a_{i,1}^2 & \sum_{i=1}^n a_{i,1}a_{i,2} \\ \sum_{i=1}^n a_{i,1}a_{i,2} & \sum_{i=1}^n a_{i,2}^2 \end{pmatrix}.$$

Since C is a symmetric, it diagonalizes in an ortho-basis $U = (u_1, \ldots, u_p) \in \mathbb{R}^{p\times p}$. Here, the vectors $u_k \in \mathbb{R}^p$ are stored in the columns of the matrix U. The diagonalization means that there exist scalars (the eigenvalues) $(\lambda_1, \ldots, \lambda_p)$ so that $(\frac{1}{n}C)u_k = \lambda_k u_k$. Since the matrix is orthogononal, $UU^\top = U^\top U = \mathrm{Id}_p$, and equivalently $U^{-1} = U^\top$. The diagonalization property can be conveniently written as $\frac{1}{n}C = U\,\mathrm{diag}(\lambda_k)U^\top$. One can thus re-write the covariance quadratic form in the basis U as being a separable sum of p squares

$$\frac{1}{n}\langle Cx,\, x\rangle = \langle U\,\mathrm{diag}(\lambda_k)U^\top x,\, x\rangle = \langle \mathrm{diag}(\lambda_k)(U^\top x),\, (U^\top x)\rangle = \sum_{k=1}^p \lambda_k \langle x,\, u_k\rangle^2. \tag{13.10}$$

Here $(U^\top x)_k = \langle x,\, u_k\rangle$ is the coordinate k of x in the basis U. Since $\langle Cx,\, x\rangle = \|Ax\|^2$, this shows that all the eigenvalues $\lambda_k \geqslant 0$ are positive.

If one assumes that the eigenvalues are ordered $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$, then projecting the points a_i on the first m eigenvectors can be shown to be in some sense the best linear dimensionality reduction possible (see

next paragraph), and it is called Principal Component Analysis (PCA). It is useful to perform compression or dimensionality reduction, but in practice, it is mostly used for data visualization in 2-D ($m = 2$) and 3-D ($m = 3$).

The matrix C/n encodes the covariance, so one can approximate the point cloud by an ellipsoid whose main axes are the $(u_k)_k$ and the width along each axis is $\propto \sqrt{\lambda_k}$ (the standard deviations). If the data are approximately drawn from a Gaussian distribution, whose density is proportional to $\exp(\frac{-1}{2}\langle C^{-1}a,\, a\rangle)$, then the fit is good. This should be contrasted with the shape of quadratic part $\frac{1}{2}\langle Cx,\, x\rangle$ of $f(x)$, since the ellipsoid $\left\{x\,;\, \frac{1}{n}\langle Cx,\, x\rangle \leqslant 1\right\}$ has the same main axes, but the widths are the inverse $1/\sqrt{\lambda_k}$. Figure 13.7 shows this in dimension $p = 2$.

13.3.5 Classification

We can do a similar computation for the gradient of the classification loss (13.4). Assuming that L is differentiable, and using the Taylor expansion (13.7) at point $-\operatorname{diag}(y)Ax$, one has

$$\begin{aligned} f(x+\varepsilon) &= L(-\operatorname{diag}(y)Ax - \operatorname{diag}(y)A\varepsilon) \\ &= L(-\operatorname{diag}(y)Ax) + \langle \nabla L(-\operatorname{diag}(y)Ax),\, -\operatorname{diag}(y)A\varepsilon\rangle + o(\|\operatorname{diag}(y)A\varepsilon\|). \end{aligned}$$

Using the fact that $o(\|\operatorname{diag}(y)A\varepsilon\|) = o(\|\varepsilon\|)$, one obtains

$$\begin{aligned} f(x+\varepsilon) &= f(x) + \langle \nabla L(-\operatorname{diag}(y)Ax),\, -\operatorname{diag}(y)A\varepsilon\rangle + o(\|\varepsilon\|) \\ &= f(x) + \langle -A^\top \operatorname{diag}(y)\nabla L(-\operatorname{diag}(y)Ax),\, \varepsilon\rangle + o(\|\varepsilon\|), \end{aligned}$$

where we have used the fact that $(AB)^\top = B^\top A^\top$ and that $\operatorname{diag}(y)^\top = \operatorname{diag}(y)$. This shows that

$$\nabla f(x) = -A^\top \operatorname{diag}(y)\nabla L(-\operatorname{diag}(y)Ax).$$

Since $L(z) = \sum_i \ell(z_i)$, one has $\nabla L(z) = (\ell'(z_i))_{i=1}^n$. For instance, for the logistic classification method, $\ell(u) = \log(1+\exp(u))$ so that $\ell'(u) = \frac{e^u}{1+e^u} \in [0,1]$ (which can be interpreted as a probability of predicting $+1$).

13.3.6 Chain Rule

One can formalize the previous computation, if $f(x) = g(Bx)$ with $B \in \mathbb{R}^{q\times p}$ and $g : \mathbb{R}^q \to \mathbb{R}$, then

$$f(x+\varepsilon) = g(Bx+B\varepsilon) = g(Bx) + \langle \nabla g(Bx),\, B\varepsilon\rangle + o(\|B\varepsilon\|) = f(x) + \langle \varepsilon,\, B^\top \nabla g(Bx)\rangle + o(\|\varepsilon\|),$$

which shows that

$$\nabla(g \circ B) = B^\top \circ \nabla g \circ B \tag{13.11}$$

where "$\circ$" denotes the composition of functions.

To generalize this to composition of possibly non-linear functions, one needs to use the notion of differential. For a function $F : \mathbb{R}^p \to \mathbb{R}^q$, its differentiable at x is a linear operator $\partial F(x) : \mathbb{R}^p \to \mathbb{R}^q$, i.e. it can be represented as a matrix (still denoted $\partial F(x)$) $\partial F(x) \in \mathbb{R}^{q\times p}$. The entries of this matrix are the partial differential, denoting $F(x) = (F_1(x), \ldots, F_q(x))$,

$$\forall\, (i,j) \in \{1,\ldots,q\} \times \{1,\ldots,p\}, \quad [\partial F(x)]_{i,j} \overset{\text{def.}}{=} \frac{\partial F_i(x)}{\partial x_j}.$$

The function F is then said to be differentiable at x if and only if one has the following Taylor expansion

$$F(x+\varepsilon) = F(x) + [\partial F(x)](\varepsilon) + o(\|\varepsilon\|). \tag{13.12}$$

where $[\partial F(x)](\varepsilon)$ is the matrix-vector multiplication. As for the definition of the gradient, this matrix is the only one that satisfies this expansion, so it can be used as a way to compute this differential in practice.

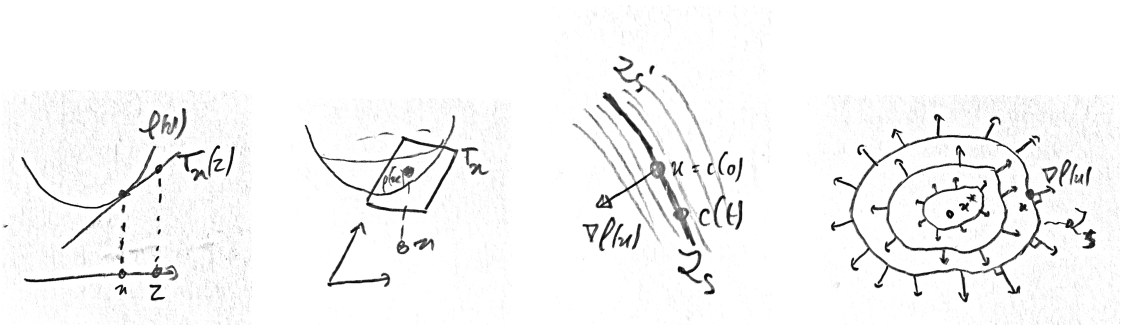


Figure 13.8: Left: First order Taylor expansion in 1-D and 2-D. Right: orthogonality of gradient and level sets and schematic of the proof.

For the special case $q = 1$, i.e. if $f : \mathbb{R}^p \to \mathbb{R}$, then the differential $\partial f(x) \in \mathbb{R}^{1\times p}$ and the gradient $\nabla f(x) \in \mathbb{R}^{p\times 1}$ are linked by equating the Taylor expansions (13.12) and (13.7)

$$\forall\, \varepsilon \in \mathbb{R}^p, \quad [\partial f(x)](\varepsilon) = \langle \nabla f(x), \varepsilon\rangle \quad \Leftrightarrow \quad [\partial f(x)](\varepsilon) = \nabla f(x)^\top.$$

The differential satisfies the following chain rule

$$\partial(G \circ H)(x) = [\partial G(H(x))] \times [\partial H(x)]$$

where "$\times$" is the matrix product. For instance, if $H : \mathbb{R}^p \to \mathbb{R}^q$ and $G = g : \mathbb{R}^q \mapsto \mathbb{R}$, then $f = g \circ H : \mathbb{R}^p \to \mathbb{R}$ and one can compute its gradient as follow

$$\nabla f(x) = (\partial f(x))^\top = ([\partial g(H(x))] \times [\partial H(x)])^\top = [\partial H(x)]^\top \times [\partial g(H(x))]^\top = [\partial H(x)]^\top \times \nabla g(H(x)).$$

When $H(x) = Bx$ is linear, one recovers formula (13.11).

13.4 Gradient Descent Algorithm

13.4.1 Steepest Descent Direction

The Taylor expansion (13.7) computes an affine approximation of the function f near x, since it can be written as

$$f(z) = T_x(z) + o(\|x - z\|) \quad \text{where} \quad T_x(z) \stackrel{\text{def.}}{=} f(x) + \langle \nabla f(x),\, z - x\rangle,$$

see Fig. 13.8. First order methods operate by locally replacing f by T_x.

The gradient $\nabla f(x)$ should be understood as a direction along which the function increases. This means that to improve the value of the function, one should move in the direction $-\nabla f(x)$. Given some fixed x, let us look as the function f along the 1-D half line

$$\tau \in \mathbb{R}^+ = [0, +\infty[\longmapsto f(x - \tau \nabla f(x)) \in \mathbb{R}.$$

If f is differentiable at x, one has

$$f(x - \tau\nabla f(x)) = f(x) - \tau\langle \nabla f(x),\, \nabla f(x)\rangle + o(\tau) = f(x) - \tau\|\nabla f(x)\|^2 + o(\tau).$$

So there are two possibility: either $\nabla f(x) = 0$, in which case we are already at a minimum (possibly a local minimizer if the function is non-convex) or if τ is chosen small enough,

$$f(x - \tau\nabla f(x)) < f(x)$$

which means that moving from x to $x - \tau\nabla f(x)$ has improved the objective function.

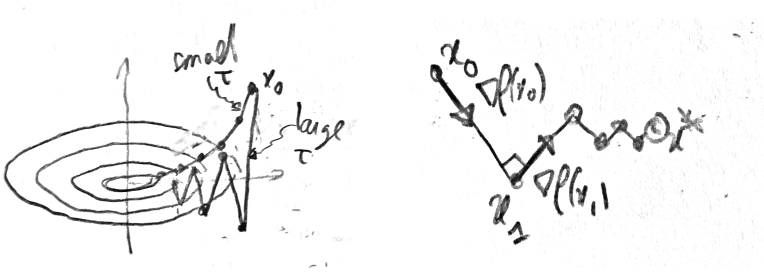


Figure 13.9: Influence of τ on the gradient descent (left) and optimal step size choice (right).

Remark 3 (Orthogonality to level sets). The level sets of f are the sets of point sharing the same value of f, i.e. for any $s \in \mathbb{R}$

$$\mathcal{L}_s \overset{\text{def.}}{=} \{x \; ; \; f(x) = s\}.$$

At some $x \in \mathbb{R}^p$, denoting $s = f(x)$, then $x \in \mathcal{L}_s$ (x belong to its level set). The gradient vector $\nabla f(x)$ is orthogonal to the level set (as shown on Fig. 13.8 right), and points toward level set of higher value (which is consistent with the previous computation showing that it is a valid ascent direction). Indeed, lets consider around x inside $\mathcal{L}_s$ a smooth curve of the form $t \in \mathbb{R} \mapsto c(t)$ where $c(0) = x$. Then the function $h(t) \overset{\text{def.}}{=} f(c(t))$ is constant $h(t) = s$ since $c(t)$ belong to the level set. So $h'(t) = 0$. But at the same time, we can compute its derivate at $t = 0$ as follow

$$h(t) = f(c(0) + tc'(0) + o(t)) = h(0) + \delta\langle c'(0), \nabla f(c(0))\rangle + o(t)$$

i.e. $h'(0) = \langle c'(0), \nabla f(x)\rangle = 0$, so that $\nabla f(x)$ is orthogonal to the tangent $c'(0)$ of the curve c, which lies in the tangent plane of $\mathcal{L}_s$ (as shown on Fig. 13.8, right). Since the curve c is arbitrary, the whole tangent plane is thus orthogonal to $\nabla f(x)$.

Remark 4 (Local optimal descent direction). One can prove something even stronger, that among all possible direction u with $\|u\| = r$, $r\frac{\nabla f(x)}{\|\nabla f(x)\|}$ becomes the optimal one as $r \to 0$ (so for very small step this is locally the best choice), more precisely,

$$\frac{1}{r}\operatorname*{argmin}_{\|u\|=r} f(x+u) \overset{r\to 0}{\longrightarrow} -\frac{\nabla f(x)}{\|\nabla f(x)\|}.$$

Indeed, introducing a Lagrange multiplier $\lambda \in \mathbb{R}$ for this constraint optimization problem, one obtains that the optimal u satisfies $\nabla f(x+u) = \lambda u$ and $\|u\| = r$. Thus $\frac{u}{r} = \pm\frac{\nabla f(x+u)}{\|\nabla f(x+u)\|}$, and assuming that ∇f is continuous, when $\|u\| = r \to 0$, this converges to $\frac{u}{\|u\|} = \pm\frac{\nabla f(x)}{\|\nabla f(x)\|}$. The sign $\pm$ should be $+1$ to obtain a maximizer and -1 for the minimizer.

13.4.2 Gradient Descent

The gradient descent algorithm reads, starting with some $x_0 \in \mathbb{R}^p$

$$x_{k+1} \overset{\text{def.}}{=} x_k - \tau_k \nabla f(x_k) \tag{13.13}$$

where $\tau_k > 0$ is the step size (also called learning rate). For a small enough τ_k, the previous discussion shows that the function f is decaying through the iteration. So intuitively, to ensure convergence, τ_k should be chosen small enough, but not too small so that the algorithm is as fast as possible. In general, one use a fix step size $\tau_k = \tau$, or try to adapt τ_k at each iteration (see Fig. 13.9).

Remark 5 (Greedy choice). Although this is in general too costly to perform exactly, one can use a "greedy" choice, where the step size is optimal at each iteration, i.e.

$$\tau_k \overset{\text{def.}}{=} \operatorname*{argmin}_{\tau} h(\tau) \overset{\text{def.}}{=} f(x_k - \tau\nabla f(x_k)).$$

Here $h(\tau)$ is a function of a single variable. One can compute the derivative of h as

$$h(\tau+\delta) = f(x_k - \tau\nabla f(x_k) - \delta\nabla f(x_k)) = f(x_k - \tau\nabla f(x_k)) - \langle \nabla f(x_k - \tau\nabla f(x_k)),\, \nabla f(x_k)\rangle + o(\delta).$$

One note that at $\tau = \tau_k$, $\nabla f(x_k - \tau\nabla f(x_k)) = \nabla f(x_{k+1})$ by definition of x_{k+1} in (13.13). Such an optimal $\tau = \tau_k$ is thus characterized by

$$h'(\tau_k) = -\langle \nabla f(x_k),\, \nabla f(x_{k+1})\rangle = 0.$$

This means that for this greedy algorithm, two successive descent direction $\nabla f(x_k)$ and $\nabla f(x_{k+1})$ are orthogonal (see Fig. 13.9).

Remark 6 (Armijo rule). Instead of looking for the optimal τ, one can looks for an admissible τ which guarantees a large enough decay of the functional, in order to ensure convergence of the descent. Given some parameter $0 < \alpha < 1$ (which should be actually smaller than $1/2$ in order to ensure a sufficient decay), one consider a τ to be valid for a descent direction d_k (for instance $d_k = -\nabla f(x_k)$) if it satisfies

$$f(x_k + \tau d_k) \leqslant f(x_k) + \alpha\tau\langle d_k,\, \nabla f(x_k)\rangle \tag{13.14}$$

For small τ, one has $f(x_k + \tau d_k) = f(x_k) + \tau\langle d_k,\, \nabla f(x_k)\rangle$, so that, assuming d_k is a valid descent direction (i.e $\langle d_k,\, \nabla f(x_k)\rangle < 0$), condition (13.14) will always be satisfied for τ small enough (if f is convex, the set of allowable τ is of the form $[0, \tau_{\max}]$). In practice, one perform gradient descent by initializing τ very large, and decaying it $\tau \leftarrow \beta\tau$ (for $\beta < 1$) until (13.14) is satisfied. This approach is often called "backtracking" line search.

13.5 Convergence Analysis

13.5.1 Quadratic Case

Convergence analysis for the quadratic case. We first analyze this algorithm in the case of the quadratic loss, which can be written as

$$f(x) = \frac{1}{2}\|Ax - y\|^2 = \frac{1}{2}\langle Cx,\, x\rangle - \langle x,\, b\rangle + \text{cst} \quad \text{where} \quad \begin{cases} C \overset{\text{def.}}{=} A^\top A \in \mathbb{R}^{p\times p}, \\ b \overset{\text{def.}}{=} A^\top y \in \mathbb{R}^p. \end{cases}$$

We already saw that in (13.9) if $\ker(A) = \{0\}$, which is equivalent to C being invertible, then there exists a single global minimizer $x^\star = (A^\top A)^{-1}A^\top y = C^{-1}u$.

Note that a function of the form $\frac{1}{2}\langle Cx,\, x\rangle - \langle x,\, b\rangle$ is convex if and only if the symmetric matrix C is positive semi-definite, i.e. that all its eigenvalues are non-negative (as already seen in (13.10)).

Proposition 41. *For $f(x) = \langle Cx,\, x\rangle - \langle b,\, x\rangle$ (C being symmetric semi-definite positive) with the eigenvalues of C upper-bounded by L and lower-bounded by $\mu > 0$, assuming there exists $(\tau_{\min}, \tau_{\max})$ such that*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tilde{\tau}_{\max} < \frac{2}{L}$$

then there exists $0 \leqslant \tilde{\rho} < 1$ such that

$$\|x_k - x^\star\| \leqslant \tilde{\rho}^\ell \|x_0 - x^\star\|. \tag{13.15}$$

The best rate $\tilde{\rho}$ is obtained for

$$\tau_\ell = \frac{2}{L+\mu} \quad \Longrightarrow \quad \tilde{\rho} \overset{\text{def.}}{=} \frac{L-\mu}{L+\mu} = 1 - \frac{2\varepsilon}{1+\varepsilon} \quad \textit{where} \quad \varepsilon \overset{\text{def.}}{=} \mu/L. \tag{13.16}$$

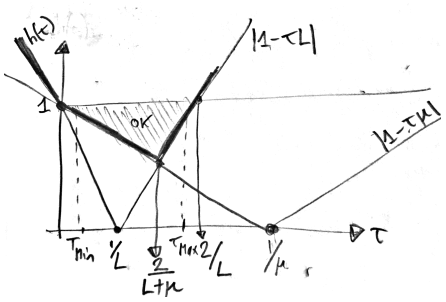


Figure 13.10: Contraction constant $h(\tau)$ for a quadratic function (right).

Proof. One iterate of gradient descent reads

$$x_{k+1} = x_k - \tau_\ell(Cx_k - b).$$

Since the solution $x^\star$ (which by the way is unique by strict convexity) satisfy the first order condition $Cx^\star = b$, it gives

$$x_{k+1} - x^\star = x_k - x^\star - \tau_\ell C(x_k - x^\star) = (\mathrm{Id}_p - \tau_\ell C)(x_k - x^\star).$$

If $S \in \mathbb{R}^{p\times p}$ is a symmetric matrix, one has

$$\|Sz\| \leqslant \|S\|_{\mathrm{op}}\|z\| \quad \text{where} \quad \|S\|_{\mathrm{op}} \overset{\text{def.}}{=} \max_k |\lambda_k(S)|,$$

where $\lambda_k(S)$ are the eigenvalues of S and $\sigma_k(S) \overset{\text{def.}}{=} |\lambda_k(S)|$ are its singular values. Indeed, S can be diagonalized in an orthogonal basis U, so that $S = U\,\mathrm{diag}(\lambda_k(S))U^\top$, and $S^\top S = S^2 = U\,\mathrm{diag}(\lambda_k(S)^2)U^\top$ so that

$$\begin{aligned}\|Sz\|^2 &= \langle S^\top Sz,\, z\rangle = \langle U\,\mathrm{diag}(\lambda_k)U^\top z,\, z\rangle = \langle \mathrm{diag}(\lambda_k^2)U^\top z,\, U^\top z\rangle \\ &= \sum_i \lambda_k^2 (U^\top z)_k^2 \leqslant \max_k(\lambda_k^2)\|U^\top z\|^2 = \max_k(\lambda_k^2)\|z\|^2.\end{aligned}$$

Applying this to $S = \mathrm{Id}_p - \tau_\ell C$, one has

$$h(\tau) \overset{\text{def.}}{=} \|\mathrm{Id}_p - \tau_\ell C\|_{\mathrm{op}} = \max_k |\lambda_k(\mathrm{Id}_p - \tau_\ell C)| = \max_k |1 - \tau_\ell\lambda_k(C)| = \max(|1 - \tau_\ell\sigma_{\max}(C)|, |1 - \tau\sigma_{\min}(C)|)$$

For a quadratic function, one has $\sigma_{\min}(C) = \mu, \sigma_{\max}(C) = L$. Figure 13.10, right, shows a display of $h(\tau)$. One has that for $0 < \tau < 2/L$, $h(\tau) < 1$. The optimal value is reached at $\tau^\star = \frac{2}{L+\mu}$ and then

$$h(\tau^\star) = \left|1 - \frac{2L}{L+\mu}\right| = \frac{L-\mu}{L+\mu}.$$

□

Note that when the condition number $\xi \overset{\text{def.}}{=} \mu/L \ll 1$ is small (which is the typical setup for ill-posed problems), then the contraction constant appearing in (13.16) scales like

$$\tilde\rho \sim 1 - 2\xi. \tag{13.17}$$

The quantity ε in some sense reflects the inverse-conditioning of the problem. For quadratic function, it indeed corresponds exactly to the inverse of the condition number (which is the ratio of the largest to smallest singular value). The condition number is minimum and equal to 1 for orthogonal matrices.

The error decay rate (13.15), although it is geometrical $O(\rho^k)$ is called a "linear rate" in the optimization literature. It is a "global" rate because it hold for all k (and not only for large enough k).

If $\ker(A) \neq \{0\}$, then C is not definite positive (some of its eigenvalues vanish), and the set of solution is infinite. One can however still show a linear rate, by showing that actually the iterations x_k are orthogonal to $\ker(A)$ and redo the above proof replacing μ by the smaller non-zero eigenvalue of C. This analysis however leads to a very poor rate ρ (very close to 1) because μ can be arbitrary close to 0. Furthermore, such a proof does not extends to non-quadratic functions. It is thus necessary to do a different theoretical analysis, which only shows a sublinear rate on the objective function f itself rather than on the iterates x_k.

Proposition 42. *For $f(x) = \langle Cx, x\rangle - \langle b, x\rangle$, assuming the eigenvalue of C are bounded by L, then if $0 < \tau_k = \tau < 1/L$ is constant, then*

$$f(x_k) - f(x^\star) \leqslant \frac{\mathrm{dist}(x_0, \mathrm{argmin}\, f)^2}{\tau 8k}.$$

where

$$\mathrm{dist}(x_0, \mathrm{argmin}\, f) \overset{\text{def.}}{=} \min_{x^\star \in \mathrm{argmin}\, f} \|x_0 - x^\star\|.$$

Proof. We have $Cx^\star = b$ for any minimizer $x^\star$ and $x_{k+1} = x_k - \tau(Cx_k - b)$ so that as before

$$x_k - x^\star = (\mathrm{Id}_p - \tau C)^k (x_0 - x^\star).$$

Now one has

$$\frac{1}{2}\langle C(x_k - x^\star), x_k - x^\star\rangle = \frac{1}{2}\langle Cx_k, x_k\rangle - \langle Cx_k, x^\star\rangle + \frac{1}{2}\langle Cx^\star, x^\star\rangle$$

and we have $\langle Cx_k, x^\star\rangle = \langle x_k, Cx^\star\rangle = \langle x_k, b\rangle$ and also $\langle Cx^\star, x^\star\rangle = \langle x^\star, b\rangle$ so that

$$\frac{1}{2}\langle C(x_k - x^\star, x_k - x^\star\rangle = \frac{1}{2}\langle Cx_k, x_k\rangle - \langle x_k, b\rangle + \frac{1}{2}\langle x^\star, b\rangle = f(x_k) + \frac{1}{2}\langle x^\star, b\rangle.$$

Note also that

$$f(x^\star) = \frac{1}{2}\frac{Cx^\star}{x^\star} - \langle x^\star, b\rangle = \frac{1}{2}\langle x^\star, b\rangle - \langle x^\star, b\rangle = -\frac{1}{2}\langle x^\star, b\rangle.$$

This shows that

$$\frac{1}{2}\langle C(x_k - x^\star, x_k - x^\star\rangle = f(x_k) - f(x^\star).$$

This thus implies

$$f(x_k) - f(x^\star) = \frac{1}{2}\langle (\mathrm{Id}_p - \tau C)^k C(\mathrm{Id}_p - \tau C)^k (x_0 - x^\star), x_0 - x^\star\rangle \leqslant \frac{\sigma_{\max}(M_k)}{2}\|x_0 - x^\star\|^2$$

where we have denoted

$$M_k \overset{\text{def.}}{=} (\mathrm{Id}_p - \tau C)^k C(\mathrm{Id}_p - \tau C)^k.$$

Since $x^\star$ can be chosen arbitrary, one can replace $\|x_0 - x^\star\|$ by $\mathrm{dist}(x_0, \mathrm{argmin}\, f)$. One has, for any ℓ, the following bound

$$\sigma_\ell(M_k) = \sigma_\ell(C)(1 - \tau\sigma_\ell(C))^{2k} \leqslant \frac{1}{\tau 4k}$$

since one can show that (setting $t = \tau\sigma_\ell(C) \leqslant 1$ because of the hypotheses)

$$\forall\, t \in [0,1], \quad (1-t)^{2k} t \leqslant \frac{1}{4k}.$$

Indeed, one has

$$(1-t)^{2k} t \leqslant (e^{-t})^{2k} t = \frac{1}{2k}(2kt)e^{-2kt} \leqslant \frac{1}{2k}\sup_{u \geqslant 0} u e^{-u} = \frac{1}{2ek} \leqslant \frac{1}{4k}.$$

□

13.5.2 General Case

We detail the theoretical analysis of convergence for general smooth convex functions. The general idea is to replace the linear operator C involved in the quadratic case by the second order derivative (the hessian matrix).

Hessian. If the function is twice differentiable along the axes, the hessian matrix is

$$(\partial^2 f)(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right)_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}.$$

Where recall that $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ is the differential along direction x_j of the function $x \mapsto \frac{\partial f(x)}{\partial x_i}$. We also recall that $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$ so that $\partial^2 f(x)$ is a symmetric matrix.

A differentiable function f is said to be twice differentiable at x if

$$f(x+\varepsilon) = f(x) + \langle \nabla f(x), \varepsilon \rangle + \frac{1}{2} \langle \partial^2 f(x) \varepsilon, \varepsilon \rangle + o(\|\varepsilon\|^2). \tag{13.18}$$

This means that one can approximate f near x by a quadratic function. The hessian matrix is uniquely determined by this relation, so that if one is able to write down an expansion with some matrix H

$$f(x+\varepsilon) = f(x) + \langle \nabla f(x), \varepsilon \rangle + \frac{1}{2} \langle H \varepsilon, \varepsilon \rangle + o(\|\varepsilon\|^2).$$

then equating this with the expansion (13.18) ensure that $\partial^2 f(x) = H$. This is thus a way to actually determine the hessian without computing all the p^2 partial derivative. This Hessian can equivalently be obtained by performing an expansion (i.e. computing the differential) of the gradient since

$$\nabla f(x+\varepsilon) = \nabla f(x) + [\partial^2 f(x)](\varepsilon) + o(\|\varepsilon\|)$$

where $[\partial^2 f(x)](\varepsilon) \in \mathbb{R}^p$ denotes the multiplication of the matrix $\partial^2 f(x)$ with the vector ε.

One can show that a twice differentiable function f on $\mathbb{R}^p$ is convex if and only if for all x the symmetric matrix $\partial^2 f(x)$ is positive semi-definite, i.e. all its eigenvalues are non-negative. Furthermore, if these eigenvalues are strictly positive then f is strictly convex (but the converse is not true, for instance x^4 is strictly convex on $\mathbb{R}$ but its second derivative vanishes at $x = 0$).

For instance, for a quadratic function $f(x) = \langle Cx, x \rangle - \langle x, u \rangle$, one has $\nabla f(x) = Cx - u$ and thus $\partial^2 f(x) = C$ (which is thus constant). For the classification function, one has

$$\nabla f(x) = -A^\top \operatorname{diag}(y) \nabla L(-\operatorname{diag}(y) A x).$$

and thus

$$\begin{aligned} \nabla f(x+\varepsilon) &= -A^\top \operatorname{diag}(y) \nabla L(-\operatorname{diag}(y) A x - -\operatorname{diag}(y) A \varepsilon) \\ &= \nabla f(x) - A^\top \operatorname{diag}(y) [\partial^2 L(-\operatorname{diag}(y) A x)](-\operatorname{diag}(y) A \varepsilon) \end{aligned}$$

Since $\nabla L(u) = (\ell'(u_i))$ one has $\partial^2 L(u) = \operatorname{diag}(\ell''(u_i))$. This means that

$$\partial^2 f(x) = A^\top \operatorname{diag}(y) \times \operatorname{diag}(\ell''(-\operatorname{diag}(y) A x)) \times \operatorname{diag}(y) A.$$

One verifies that this matrix is symmetric and positive if ℓ is convex and thus ℓ'' is positive.

Remark 7 (Second order optimality condition). The first use of Hessian is to decide wether a point $x^\star$ with $\nabla f(x^\star)$ is a local minimum or not. Indeed, if $\partial^2 f(x^\star)$ is a positive matrix (i.e. its eigenvalues are strictly positive), then $x^\star$ is a strict local minimum. Note that if $\partial^2 f(x^\star)$ is only non-negative (i.e. some its eigenvalues might vanish) then one cannot deduce anything (such as for instance x^3 on $\mathbb{R}$). Conversely, if $x^\star$ is a local minimum then $\partial^2 f(x)$

Remark 8 (Second order algorithms). A second use, is to be used in practice to define second order method (such as Newton's algorithm), which converge faster than gradient descent, but are more costly. The generalized gradient descent reads

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

where $H_k \in \mathbb{R}^{p\times p}$ is a positive symmetric matrix. One recovers the gradient descent when using $H_k = \tau_k \mathrm{Id}_p$, and Newton's algorithm corresponds to using the inverse of the Hessian $H_k = [\partial^2 f(x_k)]^{-1}$. Note that

$$f(x_k) = f(x_k) - \langle H_k \nabla f(x_k),\, \nabla f(x_k)\rangle + o(\|H_k \nabla f(x_k)\|).$$

Since H_k is positive, if x_k is not a minimizer, i.e. $\nabla f(x_k) \neq 0$, then $\langle H_k \nabla f(x_k),\, \nabla f(x_k)\rangle > 0$. So if H_k is small enough one has a valid descent method in the sense that $f(x_{k+1}) < f(x_k)$. It is not the purpose of this chapter to explain in more detail these type of algorithm.

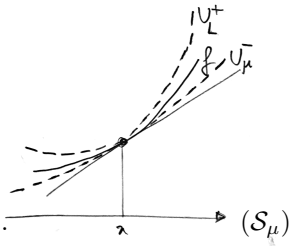
The last use of Hessian, that we explore next, is to study theoretically the convergence of the gradient descent. One simply needs to replace the boundedness of the eigenvalue of C of a quadratic function by a boundedness of the eigenvalues of $\partial^2 f(x)$ for all x. Roughly speaking, the theoretical analysis of the gradient descent for a generic function is obtained by applying this approximation and using the proofs of the previous section.

Smoothness and strong convexity. One also needs to quantify the smoothness of f. This is enforced by requiring that the gradient is L-Lipschitz, i.e.

$$\forall\, (x,x') \in (\mathbb{R}^p)^2, \quad \|\nabla f(x) - \nabla f(x')\| \leqslant L\|x - x'\|. \tag{$\mathcal{R}_L$}$$

In order to obtain fast convergence of the iterates themselve, it is needed that the function has enough "curvature" (i.e. is not too flat), which corresponds to imposing that f is μ-strongly convex

$$\forall\, (x,x'), \in (\mathbb{R}^p)^2, \quad \langle \nabla f(x) - \nabla f(x'),\, x - x'\rangle \geqslant \mu\|x - x'\|^2. \tag{$\mathcal{S}_\mu$}$$



The following proposition express these conditions as constraints on the hessian for $\mathcal{C}^2$ functions.

Proposition 43. *Conditions* ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$) *imply*

$$\forall\, (x,x'), \quad f(x') + \langle \nabla f(x),\, x' - x\rangle + \frac{\mu}{2}\|x - x'\|^2 \leqslant f(x) \leqslant f(x') + \langle \nabla f(x'),\, x' - x\rangle + \frac{L}{2}\|x - x'\|^2. \tag{13.19}$$

If f is of class $\mathcal{C}^2$, conditions ($\mathcal{R}_L$) *and* ($\mathcal{S}_\mu$) *are equivalent to*

$$\forall\, x, \quad \mu \mathrm{Id}_p \preceq \partial^2 f(x) \preceq L \mathrm{Id}_p \tag{13.20}$$

where $\partial^2 f(x) \in \mathbb{R}^{p\times p}$ is the Hessian of f, and where $\preceq$ is the natural order on symmetric matrices, i.e.

$$A \preceq B \quad \Longleftrightarrow \quad \forall\, x \in \mathbb{R}^p, \quad \langle Au,\, u\rangle \leqslant \langle Bu,\, u\rangle.$$

Proof. We prove (13.19), using Taylor expansion with integral remain

$$f(x') - f(x) = \int_0^1 \langle \nabla f(x_t),\, x' - x\rangle \mathrm{d}t = \langle \nabla f(x),\, x' - x\rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, x' - x\rangle \mathrm{d}t$$

where $x_t \overset{\text{def.}}{=} x + t(x' - x)$. Using Cauchy-Schwartz, and then the smoothness hypothesis ($\mathcal{R}_L$)

$$f(x') - f(x) \leqslant \langle \nabla f(x),\, x' - x\rangle + \int_0^1 L\|x_t - x\|\|x' - x\|\mathrm{d}t \leqslant \langle \nabla f(x),\, x' - x\rangle + L\|x' - x\|^2 \int_0^1 t\mathrm{d}t$$

which is the desired upper-bound. Using directly ($\mathcal{S}_\mu$) gives

$$f(x') - f(x) = \langle \nabla f(x),\, x' - x\rangle + \int_0^1 \langle \nabla f(x_t) - \nabla f(x),\, \frac{x_t - x}{t}\rangle \mathrm{d}t \geqslant \langle \nabla f(x),\, x' - x\rangle + \mu \int_0^1 \frac{1}{t}\|x_t - x\|^2 \mathrm{d}t$$

which gives the desired result since $\|x_t - x\|^2/t = t\|x' - x\|^2$. □

The relation (13.19) shows that a smooth (resp. strongly convex) functional is bounded by below (resp. above) by a quadratic tangential majorant (resp. minorant).

Condition (13.20) thus reads that the singular values of $\partial^2 f(x)$ should be contained in the interval $[\mu, L]$. The upper bound is also equivalent to $\|\partial^2 f(x)\|_{\text{op}} \leqslant L$ where $\|\cdot\|_{\text{op}}$ is the operator norm, i.e. the largest singular value. In the special case of a quadratic function of the form $\langle Cx, x\rangle - \langle b, x\rangle$ (recall that necessarily C is semi-definite symmetric positive for this function to be convex), $\partial^2 f(x) = C$ is constant, so that $[\mu, L]$ can be chosen to be the range of the eigenvalues of C.

Convergence analysis. We now give convergence theorem for a general convex function. On contrast to quadratic function, if one does not assumes strong convexity, one can only show a sub-linear rate on the function values (and no rate at all on the iterates themselves). It is only when one assume strong convexity that linear rate is obtained. Note that in this case, the solution of the minimization problem is not necessarily unique.

Theorem 23. *If f satisfy conditions ($\mathcal{R}_L$), assuming there exists $(\tau_{\min}, \tau_{\max})$ such that*

$$0 < \tau_{\min} \leqslant \tau_\ell \leqslant \tau_{\max} < \frac{2}{L},$$

then x_k converges to a solution $x^\star$ of (13.1) and there exists $C > 0$ such that

$$f(x_k) - f(x^\star) \leqslant \frac{C}{\ell + 1}. \tag{13.21}$$

If furthermore f is μ-strongly convex, then there exists $0 \leqslant \rho < 1$ such that $\|x_k - x^\star\| \leqslant \rho^\ell \|x_0 - x^\star\|$.

Proof. In the case where f is not strongly convex, we only prove (13.21) since the proof that x_k converges is more technical. Note indeed that if the minimizer $x^\star$ is non-unique, then it might be the case that the iterate x_k "cycle" while approaching the set of minimizer, but actually convexity of f prevents this kind of pathological behavior. For simplicity, we do the proof in the case $\tau_\ell = 1/L$, but it extends to the general case. The L-smoothness property imply (13.19), which reads

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

Using the fact that $x_{k+1} - x_k = -\frac{1}{L}\nabla f(x_k)$, one obtains

$$f(x_{k+1}) \leqslant f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 \leqslant f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \tag{13.22}$$

This shows that $(f(x_k))_\ell$ is a decaying sequence. By convexity

$$f(x_k) + \langle \nabla f(x_k), x^\star - x_k\rangle \leqslant f(x^\star)$$

and plugging this in (13.22) shows

$$\begin{aligned} f(x_{k+1}) &\leqslant f(x^\star) - \langle \nabla f(x_k), x^\star - x_k\rangle - \frac{1}{2L}\|\nabla f(x_k)\|^2 && (13.23)\\ &= f(x^\star) + \frac{L}{2}\left(\|x_k - x^\star\|^2 - \|x_k - x^\star - \frac{1}{L}\nabla f(x_k)\|^2\right) && (13.24)\\ &= f(x^\star) + \frac{L}{2}\left(\|x_k - x^\star\|^2 - \|x^\star - x_{k+1}\|^2\right). && (13.25) \end{aligned}$$

Summing these inequalities for $\ell = 0, \ldots, k$, one obtains

$$\sum_{\ell=0}^{k} f(x_{k+1}) - (k+1) f(x^\star) \leqslant \frac{L}{2}\left(\|x_0 - x^\star\|^2 - \|x^{(k+1)} - x^\star\|^2\right)$$

and since $f(x_{k+1})$ is decaying $\sum_{\ell=0}^{k} f(x_{k+1}) \geqslant (k+1) f(x^{(k+1)})$, thus

$$f(x^{(k+1)}) - f(x^\star) \leqslant \frac{L\|x_0 - x^\star\|^2}{2(k+1)}$$

which gives (13.21) for $C \stackrel{\text{def.}}{=} L\|x_0 - x^\star\|^2/2$.

If we now assume f is μ-strongly convex, then, using $\nabla f(x^\star) = 0$, one has $\frac{\mu}{2}\|x^\star - x\|^2 \leqslant f(x) - f(x^\star)$ for all x. Re-manipulating (13.25) gives

$$\frac{\mu}{2}\|x_{k+1} - x^\star\|^2 \leqslant f(x_{k+1}) - f(x^\star) \leqslant \frac{L}{2}\left(\|x_k - x^\star\|^2 - \|x^\star - x_{k+1}\|^2\right),$$

and hence

$$\|x_{k+1} - x^\star\| \leqslant \sqrt{\frac{L}{L+\mu}}\|x_{k+1} - x^\star\|, \tag{13.26}$$

which is the desired result. □

Note that in the low conditioning setting $\varepsilon \ll 1$, one retrieve a dependency of the rate (13.26) similar to the one of quadratic functions (13.17), indeed

$$\sqrt{\frac{L}{L+\mu}} = (1+\varepsilon)^{-\frac{1}{2}} \sim 1 - \frac{1}{2}\varepsilon.$$

13.5.3 Acceleration

The previous analysis shows that for L-smooth functions (i.e. with a hessian uniformly bounded by L, $\|\partial^2 f(x)\|_{\text{op}} \leqslant L$), the gradient descent with fixed step size converges with a speed on the function value $f(x_k) - \min f = O(1/k)$. Even using various line search strategies, it is not possible to improve over this rate. A way to improve this rate is by introducing some form of "momentum" extrapolation and rather consider a pair of variables (x_k, y_k) with the following update rule, for some step size s (which should be smaller than $1/L$)

$$\begin{cases} x_{k+1} = y_k - s\nabla f(y_k) \\ y_{k+1} = x_{k+1} + \beta_k(x_{k+1} - x_k) \end{cases}$$

where the extrapolation parameter satisfies $0 < \beta_k < 1$. The case of a fixed $\beta_k = \beta$ corresponds to the so-called "heavy-ball" method. In order for the method to bring an improvement for the $1/k$ "worse case" rate (which does not means it improves for all possible case), one needs to rather use increasing momentum $\beta_k \to 1$, one popular choice being

$$\beta_k \frac{k-1}{k+2} \sim 1 - \frac{1}{k}.$$

This corresponds to the so-called "Nesterov" acceleration (although Nesterov used a slightly different choice, with the similar $1 - 1/k$ asymptotic behavior).

When using $s \leqslant 1/L$, one can show that $f(x_k) - \min f = O(\frac{\|x_0 - x^\star\|}{sk^2})$, so that in the worse case scenario, the convergence rate is improved. Note however that in some situation, acceleration actually deteriorates the rates. For instance, if the function is strongly convex (and even on the simple case $f(x) = \|x\|^2$), Nesterov acceleration does not enjoy linear convergence rate.

A way to interpret this scheme is by looking at a time-continuous ODE limit when $s \to 0$. On the contrary to the classical gradient descent, the step size here should be taker as $\tau = \sqrt{s}$ so that the time evolves as $t = \sqrt{s}k$. The update reads

$$\frac{x_{k+1} - x_k}{\tau} = (1 - 3/k)\frac{x_k - x_{k-1}}{\tau} - \nabla f(y_k)$$

which can be re-written as

$$\frac{x_{k+1}+x_{k-1}-2x_k}{\tau^2}-\frac{3}{k\tau}\frac{x_k-x_{k-1}}{\tau}+\tau\nabla f(y_k)=0.$$

Assuming $(x_k,y_k)\to(x(t),y(t))$, one obtains in the limit the following second order ODE

$$x''(t)+\frac{3}{t}x'(t)+\nabla f(x(t))=0 \quad \text{with} \quad \begin{cases} x(0)=x_0, \\ x'(0)=0. \end{cases}$$

This corresponds to the movement of a ball in the potential field f, where the term $\frac{3}{t}x'(t)$ plays the role of a friction which vanishes in the limit. So for small t, the method is similar to a gradient descent $x'=-\nabla f(x)$, while for large t, it ressemble a Newtonian evolution $x''=-\nabla f(x)$ (which keeps oscillating without converging). The momentum decay rate $3/t$ is very important, it is the only rule which enable the speed improvement from $1/k$ to $1/k^2$.

Bibliography

[1] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[2] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[3] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[4] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[5] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

[6] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[7] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[8] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[9] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[10] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.

[11] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[12] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier.* Ellipses, 2004.

[13] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.

[14] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[15] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging.* Springer, 2009.

[16] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[17] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis.* Cambridge university press, 2015.