Continuous (convex) optimisation

M2 - PSL / Dauphine / S.U.

Antonin Chambolle, CNRS, CEREMADE

Université Paris Dauphine PSL

Sep.-Dec. 2024

Lecture 1: first order descent methods and convergence rates.

Contents

Some resources:

- Boris Polyak: Introduction to optimization, (1987).
- J.-B. Hiriart-Urruty and C. Lemarechal, Convex Analysis and Minimization Algorithms (1993).
- Yurii Nesterov: Introductory lectures on convex optimization, 2004
- Jorge Nocedal and Stephen J. Wright: Numerical Optimization, 2006.
- Dimitri Bertsekas: Convex Optimization Algorithms. Athena Scientific 2015.
- Amir Beck: First-Order Methods In Optimization, 2019.
- R. Tyrell Rockafellar: Convex analysis, 1970 (1997).
- H. Bauschke and P.L. Combettes: Convex analysis and monotone operator theory in Hilbert spaces (Springer 2011)
- Ivar Ekeland and Roger Temam: Convex analysis and variational problems, 1999.
- Juan Peypouquet: Convex Optimization in Normed Spaces.

Gradient descent

Problem:

$$\min_{x \in \mathcal{X}} f(x)$$

Gradient descent

Problem:

$$\min_{x\in\mathcal{X}} f(x)$$

$\mathcal{X}$ is a vector space, f a real valued function.

(Very elementary) *Algorithm:*

$$x^{k+1} = x^k - \text{“}\textit{a descent direction}\text{”}.$$

Gradient descent

Algorithm:

$x^{k+1} = x^k -$ *"a descent direction"*.

Simplest descent direction: *linearize* f at x^k ($\rightarrow$ "first order" method):

$$f(y) = f(x^k) + df(x^k) \cdot (y - x^k) + o(|y - x^k|)$$

and then find a direction "which points on the good side of $\ker df$" which is tangent to the level set $\{f = f(x^k)\}$.

Gradient descent

Algorithm:

$x^{k+1} = x^k - \text{"}a\ descent\ direction\text{"}$.

Simplest descent direction: *linearize* f at x^k ($\rightarrow$ "first order" method):

$$f(y) = f(x^k) + df(x^k) \cdot (y - x^k) + o(|y - x^k|)$$

and then find a direction "which points on the good side of $\ker df$" which is tangent to the level set $\{f = f(x^k)\}$.

⚠ $df(x^k) \in \mathcal{X}^*$ and "$x^{k+1} = x^k - \tau df(x^k)$" **does not** make any sense.

One would need a mapping $J : \mathcal{X}^* \rightarrow \mathcal{X}$ which can be linear or non-linear and converts df into $(-)$ a descent direction, and then we may write $x^{k+1} = x^k - \tau J(df(x^k))$. We first start with the simplest case.

Gradient descent

In most of the lectures $\mathcal{X}$ is a Hilbert or finite-dimensional Euclidean space. In which case, one can define a *gradient* thanks to the scalar product (and "Riesz' representation theorem").

Gradient descent

In most of the lectures $\mathcal{X}$ is a Hilbert or finite-dimensional Euclidean space. In which case, one can define a *gradient* thanks to the scalar product (and "Riesz' representation theorem").

Definition

The gradient of $f : \mathcal{X} \to \mathbb{R}$ at x is the vector $p = \nabla f(x) \in \mathcal{X}$ such that for all $y \in \mathcal{X}$,

$$df(x) \cdot y = \langle p, y \rangle_{\mathcal{X}}$$

Gradient descent

In most of the lectures $\mathcal{X}$ is a Hilbert or finite-dimensional Euclidean space. In which case, one can define a *gradient* thanks to the scalar product (and "Riesz' representation theorem").

Definition

The gradient of $f : \mathcal{X} \to \mathbb{R}$ at x is the vector $p = \nabla f(x) \in \mathcal{X}$ such that for all $y \in \mathcal{X}$,

$$df(x) \cdot y = \langle p, y \rangle_{\mathcal{X}}$$

Then, $\ker df(x) = \{y : \langle \nabla f(x), y \rangle_{\mathcal{X}} = 0\}$ and

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle_{\mathcal{X}} + o(|y - x|)$$

so that f increases if $y - x$ is small in the direction of $\nabla f(x)$: that is, $-\nabla f(x)$ is a "descent direction".

Gradient descent

Then the iteration:

$$x^{k+1} = x^k - \tau \nabla f(x^k) =: T_\tau(x^k)$$

($\tau > 0$) makes sense and one has

Gradient descent

Then the iteration:

$$x^{k+1} = x^k - \tau \nabla f(x^k) =: T_\tau(x^k)$$

$(\tau > 0)$ makes sense and one has

$$f(x^{k+1}) = f(x^k) - \tau |\nabla f(x^k)|^2_{\mathcal{X}} + o(\tau) < f(x^k).$$

if τ is small enough (but how small?)

Gradient descent: choices for τ

- Optimal: solve $\min_\tau f(x^k - \tau\nabla f(x^k))$. 1D optimization solved with a "line search" is easy but requires many evaluations of f;
- "Armijo" type rule: for instance, find $i \geq 0$ such that $f(x^k - \tau\rho^i\nabla f(x^k)) \leq f(x^k) - c\tau\rho^i|\nabla f(x^k)|^2$, $\rho < 1, c < 1$ fixed: "sufficient descent rule";
- Fixed step $\tau > 0$.

Gradient descent: fixed step

Important remark: Given $\tau > 0$, one can interpret x^{k+1} as the minimizer of a quadratic approximation of f:

$$x^{k+1} = \arg\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle_{\mathcal{X}} + \frac{1}{2\tau}|x - x^k|^2_{\mathcal{X}}.$$

Gradient descent: fixed step

Important remark: Given $\tau > 0$, one can interpret x^{k+1} as the minimizer of a quadratic approximation of f:

$$x^{k+1} = \arg\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle_{\mathcal{X}} + \frac{1}{2\tau}|x - x^k|^2_{\mathcal{X}}.$$

A natural generalization is: obtain x^{k+1} as a minimizer of

$$\min_x f(x^k) + df(x^k) \cdot (x - x^k) + \frac{1}{2\tau} d(x, x^k)^2$$

for d a distance in $\mathcal{X}$. This can serve as a generalization for non-Hilbertian distances ($\rightarrow$ nonlinear gradient descent method), can be used to improve the quadratic approximation of f (2nd order methods, Newton, Quasi-Newton...), or one can even substitute d with more general "divergences".

Conditional Gradient / "Frank-Wolfe" algorithm

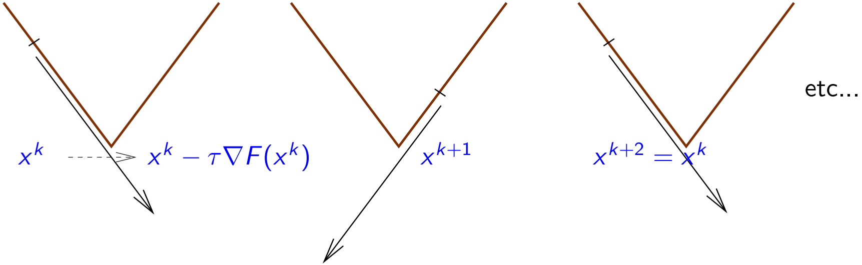
Variant (different from a classical gradient descent method): replace $(1/\tau)|x - x^k|^2_{\mathcal{X}}$ with the *characteristic function* 0 if $x \in X_k$, $+\infty$ else:

$$\min_{x \in X^k} f(x^k) + df(x^k) \cdot (x - x^k)$$

for some set X^k, which could be fixed (a constraint set), or varying, depending on the particular method—a gradient method is recovered if $X^k = B(x^k, \rho_k)$ for some radius $\rho_k > 0$.

(Then in general one chooses x^{k+1} as a convex combination of x^k and the above minimizer.)

Fixed step: Why a Lipschitz gradient is needed



The gradient descent may never converge if the step is too large or the function not smooth enough (here $f(x) = |x|$).

Convergence

Proposition

If f is C^1, bounded from below with ∇f L-Lipschitz

Then: $\nabla f(x^k) \to 0$ as $k \to \infty$ provided $0 < \tau < 2/L$.

Convergence

Proposition

If f is C^1, bounded from below with ∇f L-Lipschitz

Then: $\nabla f(x^k) \to 0$ as $k \to \infty$ provided $0 < \tau < 2/L$.

Proof: One writes:

$$\begin{aligned}
f(x^{k+1}) &= f(x^k) - \int_0^\tau \left\langle \nabla f(x^k - s\nabla f(x^k)), \nabla f(x^k) \right\rangle_{\mathcal{X}} \\
&= f(x^k) - \tau |\nabla f(x^k)|_{\mathcal{X}}^2 + \\
&\qquad \int_0^\tau \left\langle \nabla f(x^k) - \nabla f(x^k - s\nabla f(x^k)), \nabla f(x^k) \right\rangle_{\mathcal{X}} \\
&\qquad\qquad \le f(x^k) - \tau(1 - \tfrac{L\tau}{2}) |\nabla f(x^k)|_{\mathcal{X}}^2 .
\end{aligned}$$

Convergence

Letting $\kappa := \tau(1 - \tau L/2) > 0$, we get:

$$f(x^{k+1}) + \kappa |\nabla f(x^k)|^2_{\mathcal{X}} \leq f(x^k)$$

so that $f(x^k)$ is a decreasing sequence (unless x^k is critical), and summing we get:

$$f(x^n) + \kappa \sum_{k=0}^{n-1} |\nabla f(x^k)|^2_{\mathcal{X}} \leq f(x^0).$$

from which we can deduce the result (letting $n \to \infty$).

Convergence: remarks

Remark: If $\tau = 0$, then the algorithm does not move. If $\tau = 2/L$ it may not converge: example: $f(x) = L|x|^2/2$, $x^0 \neq 0$.

Remark: In the proof we only use that

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle_{\mathcal{X}} \leq L|x - y|^2_{\mathcal{X}}$$

that is, an *upper bound* for the Hessian ($D^2 f \leq LI$), or "$\frac{L}{2}|x|^2_{\mathcal{X}} - f(x)$ *is convex*".

Remark: To get convergence of (a subsequence of) x^k to some critical point one needs in addition to assume that f is "coercive" (for instance, $f(x) \to \infty$ when $|x| \to \infty$).

Remark: If x^* is a minimizer, one also deduces that the gradient is controlled by the objective f:

$$\boxed{\frac{1}{2L}|\nabla f(x^k)|^2_{\mathcal{X}} \leq f(x^k) - f(x^{k+1}) \leq f(x^k) - f(x^*).}$$

(Choosing $\tau = 1/L$.) *(Why am I allowed to do this?)*

Convergence analysis: convex case

In the convex case one can get much more precise results. Convexity will be studied more deeply in a forthcoming lecture, now we just need two properties:

Property

If f is convex, then for any $x, y \in \mathcal{X}$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle_{\mathcal{X}}$$

This is true in fact in any vector space and could be taken as a *definition* of a convex function: a convex function is always above its first order (affine) approximations.
[$\rightarrow$ "Monotonicity" of the gradient]

Convergence analysis: convex case

This is a small piece of a much more general result of (Baillon-Haddad, 1977):

Theorem

If f is convex and ∇f is L-Lipschitz, then for all x, y:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle_{\mathcal{X}} \geq \frac{1}{L} |\nabla f(x) - \nabla f(y)|^2_{\mathcal{X}}.$$

(∇f is said to be "$(1/L)$-co-coercive", or $L^{-1}\nabla f$ "firmly non-expansive".)

[**Definition**: An operator $A : \mathcal{X} \to \mathcal{X}$ is "firmly non expansive" if and only if $|A(x) - A(y)|^2 + |(I - A)(x) - (I - A)(y)|^2 \leq |x - y|^2$ for all x, y.

Show that this is equivalent to being 1-co-coercive: $\langle Ax - Ay, x - y \rangle \geq |Ax - Ay|^2$ for all x, y.]

Convergence analysis: convex case

Proof: assuming f is C^2 (the general case is easily recovered by approximation with smooth functions), we use: $0 \leq D^2 f \leq LI$ (because f is convex, and because ∇f is L-Lipschitz). We let:

$$\nabla f(x) - \nabla f(y) = \int_0^1 D^2 f(y + s(x-y))(x-y)ds =: A(x-y).$$

with $A = \int_0^1 D^2 f(y + s(x-y))ds$ symmetric, $0 \leq A \leq LI$. Then:

$$|\nabla f(x) - \nabla f(y)|^2 = |A(x-y)|^2 = \left\langle AA^{1/2}(x-y), A^{1/2}(x-y)\right\rangle \leq L\left\langle A^{1/2}(x-y), A^{1/2}(x-y)\right\rangle \leq L\langle A(x-y), x-y\rangle = L\langle \nabla f(x) - \nabla f(y), x-y\rangle$$

which is the result. □

Convergence analysis: convex case

Lemma:

If f is convex with L-Lipschitz gradient, then the mapping $T_\tau = I - \tau\nabla f$ is a weak contraction when $0 \le \tau \le 2/L$ (that is, T_τ is 1-Lipschitz, or "non-expansive").

Convergence analysis: convex case

Lemma:

If f is convex with L-Lipschitz gradient, then the mapping $T_\tau = I - \tau\nabla f$ is a weak contraction when $0 \le \tau \le 2/L$ (that is, T_τ is 1-Lipschitz, or "non-expansive").

Proof: we write

$$
\begin{aligned}
|T_\tau x - T_\tau y|^2 &= |x-y|^2 - 2\tau\langle x-y, \nabla f(x) - \nabla f(y)\rangle + \tau^2|\nabla f(x) - \nabla f(y)|^2 \\
&\le |x-y|^2 - \frac{2\tau}{L}\left(1 - \frac{\tau L}{2}\right)|\nabla f(x) - \nabla f(y)|^2 \\
&\le |x-y|^2
\end{aligned}
$$

provided $2\tau/L(1-\tau L/2) \ge 0$.

Convex case: remark

Remark: T_τ is “averaged” for $0 < \tau < 2/L$, that is:

$$T_\tau = \theta T_{2/L} + (1-\theta)I$$

for $\theta = \tau L/2 \in]0,1[$ where, by the previous Lemma, $T_{2/L} = (I - \frac{2}{L}\nabla f)$ is 1-Lipschitz.

Convex case: remark

Remark: T_τ is "averaged" for $0 < \tau < 2/L$, that is:

$$T_\tau = \theta T_{2/L} + (1 - \theta) I$$

for $\theta = \tau L/2 \in]0,1[$ where, by the previous Lemma, $T_{2/L} = (I - \frac{2}{L}\nabla f)$ is 1-Lipschitz.

The convergence of the iterates of this class of operators will be proved later on. It will follow that $x^k \rightharpoonup x^*$ a minimizer of f ($\Leftrightarrow$ a fixed point of T_τ), if it exists.

We now rather establish a convergence rate.

Convergence rate in the convex case

As already mentioned, if f is convex, one has:

$$f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle$$

for any minimizer x^*, so that:

$$\frac{f(x^k) - f(x^*)}{|x^* - x^k|_{\mathcal{X}}} \leq |\nabla f(x^k)|_{\mathcal{X}}.$$

Convergence rate in the convex case

As already mentioned, if f is convex, one has:

$$f(x^*) \geq f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle$$

for any minimizer x^*, so that:

$$\frac{f(x^k) - f(x^*)}{|x^* - x^k|_{\mathcal{X}}} \leq |\nabla f(x^k)|_{\mathcal{X}}.$$

Combined with:

$$f(x^{k+1}) - f(x^*) + \kappa |\nabla f(x^k)|^2_{\mathcal{X}} \leq f(x^k) - f(x^*) =: \Delta_k$$

we obtain (as $|x^{k+1} - x^*| = |T_\tau x^k - T_\tau x^*| \leq |x^k - x^*| \leq |x^0 - x^*|$):

$$\Delta_{k+1} \leq \Delta_k - \frac{\kappa}{|x^0 - x^*|^2_{\mathcal{X}}} \Delta_k^2$$

Convergence rate in the convex case

Lemma

Let $(a_k)_k$ be a sequence of nonnegative numbers satisfying for $k \geq 0$:

$$a_{k+1} \leq a_k - c^{-1} a_k^2$$

Then, for all $k \geq 0$, $a_k \leq \frac{c}{k+1}$

Convergence rate in the convex case

Lemma

Let $(a_k)_k$ be a sequence of nonnegative numbers satisfying for $k \geq 0$:

$$a_{k+1} \leq a_k - c^{-1} a_k^2$$

Then, for all $k \geq 0$, $a_k \leq \frac{c}{k+1}$

Proof: if we replace a_k with a_k/c, it becomes $a_{k+1} \leq a_k - a_k^2$: hence we may assume $c = 1$. Then, since $a_k(1 - a_k) \geq a_{k+1} \geq 0$, one has $0 \leq a_k \leq 1$ for all $k \geq 0$.
We show the inequality by induction: for $k = 0$, $a_0 \leq 1$. If $k \geq 1$ and if $ka_{k-1} \leq 1$, then we write that

$$\begin{aligned}
(k+1)a_k &\leq (k+1)(a_{k-1} - a_{k-1}^2) \\
&= (k+1)a_{k-1} - (k+1)a_{k-1}^2 = ka_{k-1} + a_{k-1}(1 - (k+1)a_{k-1}) \\
&\leq 1 + a_{k-1}(1 - (k+1)a_k) \quad (\text{since } 0 \leq a_k \leq a_{k-1}).
\end{aligned}$$

Hence $(1 - (k+1)a_k)(1 + a_{k-1}) \geq 0$, and $(k+1)a_k \leq 1$.

Convergence rate in the convex case: conclusion

Theorem

The gradient descent with fixed step satisfies

$$\Delta_k \le \frac{|x^0 - x^*|_{\mathcal{X}}^2}{\kappa(k+1)}$$

(for $\kappa = \tau(1 - \tau L/2) > 0$).

κ is maximal for $\tau = 1/L$, and the corresponding rate is:

$$\Delta_k \le 2L \frac{|x^0 - x^*|_{\mathcal{X}}^2}{k+1}.$$

The strongly convex case

We anticipate and say that f is *strongly convex* if $D^2 f \geq \gamma I$, $\gamma > 0$. This is also called γ-convex. An equivalent definition, which does not require f to be twice differentiable, is that $f - \frac{\gamma}{2}|x|^2_{\mathcal{X}}$ is convex[1].
In that case (assuming, still, f C^2), there is a simpler convergence proof for the gradient descent, as follows. We let x^* be the minimizer (in this case, which exists and is unique) of f. We write:

[1] ⚠ These definitions are for Euclidean or real Hilbert spaces only! General strong convexity is defined a bit differently, by means of an inequality.

The strongly convex case

$$x^{k+1} - x^* = x^k - x^* - \tau(\nabla f(x^k) - \nabla f(x^*)) = \int_0^1 (I - \tau D^2 f(x^* + s(x^k - x^*))(x^k - x^*)ds$$

hence (using that $(1 - \tau L)I \leq I - \tau D^2 f \leq (1 - \tau\gamma)I$)

$$|x^{k+1} - x^*|_{\mathcal{X}} \leq \max\{1 - \tau\gamma, \tau L - 1\}|x^k - x^*|_{\mathcal{X}}.$$

If f is not C^2 one can still show this by smoothing. The best constant is for $\tau = 2/(L + \gamma)$ and gives, for $q = (L - \gamma)/(L + \gamma) \in [0, 1]$

$$|x^k - x^*|_{\mathcal{X}} \leq q^k |x^0 - x^*|_{\mathcal{X}}.$$

The strongly convex case

$$x^{k+1} - x^* = x^k - x^* - \tau(\nabla f(x^k) - \nabla f(x^*)) = \int_0^1 (I - \tau D^2 f(x^* + s(x^k - x^*))(x^k - x^*) ds$$

hence (using that $(1 - \tau L)I \le I - \tau D^2 f \le (1 - \tau\gamma)I$)

$$|x^{k+1} - x^*|_{\mathcal{X}} \le \max\{1 - \tau\gamma, \tau L - 1\}|x^k - x^*|_{\mathcal{X}}.$$

If f is not C^2 one can still show this by smoothing. The best constant is for $\tau = 2/(L + \gamma)$ and gives, for $q = (L - \gamma)/(L + \gamma) \in [0, 1]$

$$|x^k - x^*|_{\mathcal{X}} \le q^k |x^0 - x^*|_{\mathcal{X}}.$$

This is called a *linear* convergence rate. The contraction factor is

$$q = \frac{1 - \gamma/L}{1 + \gamma/L}$$

where $\gamma/L < 1$ can be thought as the inverse condition number of the problem.

What can we achieve?

We will soon see that these convergence bounds are **not tight**. We give here a very basic approach to the complexity theory for first-order method developed by Nemirovsky and Yudin (see also Nesterov's lecture notes).

The idea is to introduce a "hard problem" and show that no first-order method can solve it faster than a certain rate.

Hard problem / F.O.M.

We consider $L > 0$, $\gamma \geq 0$, $1 \leq p \leq n$, and for $x \in \mathbb{R}^n$, a function of the form:

$$f(x) = \frac{L-\gamma}{8}\left((x_1 - 1)^2 + \sum_{i=2}^{p}(x_i - x_{i-1})^2\right) + \frac{\gamma}{2}|x|^2.$$

A "First Order Method" is such that the iterates x^k belong to the subspace spanned by the gradients of already computed iterates: for $k \geq 0$,

$$x^k \in x^0 + \left\{\nabla f(x^0), \nabla f(x^1), \ldots, \nabla f(x^{k-1})\right\},$$

where x^0 is an arbitrary starting point.

Hard problem

Starting from $x^0 = 0$ in the above problem (whose solution is given, in case $\gamma = 0$, by $x_l^* = 1$, $k = 1, \dots, p$, and 0 for $l > p$), then at the first iteration, only the first component x_1^1 will be updated (since $\partial_i f(x^0) = 0$ for $i \geq 2$), and by induction one can check that at iteration k, $x_l^k = 0$ for $l \geq k+1$: information is transmitted very slowly

Hard problem

The solution satisfies $\nabla f = 0$, therefore is characterized by

$$x_i = \frac{L-\gamma}{L+\gamma}\frac{x_{i+1}+x_{i-1}}{2}, \quad i \le p-1,$$

with $x_0 = 1$ and $x_p = (L-\gamma)/(L+3\gamma)x_{p-1}$. The best possible point at iteration k satisfies this equation for $i \le k$, and $x_{k+1} = 0$.

In case $\gamma = 0$ we find that this point x is affine: $x_i = (1 - i/(k+1))^+$, and $x_i - x_{i-1} = -1/(k+1)$ for $i \le k+1$. Hence

$$f(x) = \frac{L}{8}\sum_{i=1}^{k+1}\frac{1}{(k+1)^2} = \frac{L}{8}\frac{1}{k+1}$$

is the best possible value which can be reached at step k.

Lower bound for hard problem

Using here that $x^0 = 0$ while $x_i^* = 1$ for $i \leq p$ and 0 for $i > p$, one has $|x^0 - x^*|^2 = p$, $f(x^*) = 0$, hence we find:

$$f(x^k) - f(x^*) \geq \frac{L}{8p(k+1)}|x^0 - x^*|^2$$

($k < p$) (while if $k = p$, $x^k = x^*$). For $k = p - 1$ one finds

$$f(x^k) - f(x^*) \geq \frac{L}{8}\frac{|x^0 - x^*|^2}{(k+1)^2}$$

hence no first order method can satisfy a better reverse inequality!

(Dimension-independent) Lower bound

Theorem

For any $n \geq 2$, any $x^0 \in \mathbb{R}^n$, $L > 0$, and $k < n$, there exists a convex, one times continuously differentiable function f with L-Lipschitz continuous gradient, such that for any first-order method, it holds that

$$f(x^k) - f(x^*) \geq \frac{L|x^0 - x^*|^2}{8(k+1)^2},$$

where x^* denotes a minimizer of f.

(*cf* Thms. 2.1.7 and 2.1.13 in Nesterov's book.)

(Dimension-independent) Lower bound

Theorem

For any $n \geq 2$, any $x^0 \in \mathbb{R}^n$, $L > 0$, and $k < n$, there exists a convex, one times continuously differentiable function f with L-Lipschitz continuous gradient, such that for any first-order method, it holds that

$$f(x^k) - f(x^*) \geq \frac{L|x^0 - x^*|^2}{8(k+1)^2},$$

where x^* denotes a minimizer of f.

(*cf* Thms. 2.1.7 and 2.1.13 in Nesterov's book.)

BUT: the gradient descent had $\leq 2L|x^0 - x^*|^2/(k+1)$! which is very far from this lower bound.

Strongly convex case

A similar study (slightly more complicated, in $\mathbb{R}^\infty$) in the strongly convex case shows:

Theorem

For any $x^0 \in \mathbb{R}^\infty \simeq \ell_2(\mathbb{N})$ and $\gamma, L > 0$ there exists a γ-strongly convex, one times continuously differentiable function f with L-Lipschitz continuous gradient, such that for any first order method, it holds that for all k,

$$f(x^k) - f(x^*) \geq \frac{\gamma}{2} q^{2k} |x^0 - x^*|^2$$
$$|x^k - x^*| \geq q^k |x^0 - x^*|$$

where $q = \frac{\sqrt{Q}-1}{\sqrt{Q}+1}$ and where $Q = L/\gamma \geq 1$ is the condition number, and x^* a minimizer of f.

Strongly convex case

Again, here, the gradient descent had: $q = (Q-1)/(Q+1)$. Passing from Q to $\sqrt{Q}$ would be a *huge* improvement.

To have

$$\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^k \le \varepsilon$$

one needs, assuming $Q >> 1$ so that $(\sqrt{Q}-1)/(\sqrt{Q}+1) \approx 1 - 2/\sqrt{Q}$, and $\varepsilon << \gamma$:

$$k \gtrsim \frac{\sqrt{Q}|\log \varepsilon|}{2}$$

iterations. For the gradient descent, we needed

$$k \gtrsim \frac{Q|\log \varepsilon|}{2}$$

iterations. This is the same order of improvement as for the case $\gamma = 0$.

Higher order or accelerated methods

Now, we will consider variants of the gradient descent method and show that the worse case rates of the previous slides can actually be (almost) reached.

Newton method

We first discuss a second order method, which yields *much faster* convergence. However: it is not obvious to find a good starting point, it is computationally too intensive for large problems.

Newton method

The idea is to develop a second order approximation of f at x^k:

$$f(x) = f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{1}{2} \left\langle D^2 f(x^k)(x - x^k), x - x^k \right\rangle + o(|x - x^k|^2).$$

Near a minimizer, one can hope that $D^2 f(x^k) > 0$: we find find x^{k+1} by solving

$$\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2} \left\langle D^2 f(x^k)(x - x^k), x - x^k \right\rangle$$

(Compare with the Gradient descent with step τ in a metric defined by a symmetric positive definite matrix $A > 0$, which would be:

$$\min_x f(x^k) + \left\langle \nabla f(x^k), x - x^k \right\rangle + \tfrac{1}{2\tau} \left\langle A(x - x^k), x - x^k \right\rangle$$

hence we can see Newton's method as a gradient descent in the metric which best approximates the function.)

Newton method

x^{k+1} is given by

$$\nabla f(x^k) + D^2 f(x^k)(x^{k+1} - x^k) = 0 \;\Leftrightarrow\; x^{k+1} = x^k - D^2 f(x^k)^{-1} \nabla f(x^k).$$

Theorem

Assume f is C^2, $D^2 f$ is M-Lipschitz, and $D^2 f \geq \gamma$ (strong convexity). Let $q = \frac{M}{2\gamma^2}|\nabla f(x^0)|$ and assume x^0 is close enough to the minimizer x^*, so that $q < 1$. Then $|x^k - x^*| \leq (2\gamma/M) q^{2^k}$.

This is *very* fast: $q^{2^k} \leq \varepsilon$ if $k \geq \log(|\log \varepsilon| / |\log q|) / \log 2$ (a “quadratic” convergence rate).

Newton method: proof

We start with:

$$\nabla f(x+h) = \nabla f(x) + \int_0^1 D^2 f(x+sh)h\,ds = \nabla f(x) + D^2 f(x)h + \int_0^1 (D^2 f(x+sh) - D^2 f(x))h\,ds$$

so that

$$|\nabla f(x+h) - \nabla f(x) - D^2 f(x)h| \leq \frac{M}{2}|h|^2.$$

Hence

$$|\nabla f(x^{k+1}) - \overbrace{\nabla f(x^k) - D^2 f(x^k)(x^{k+1} - x^k)}^{0}| \leq \frac{M}{2}|x^{k+1} - x^k|^2$$
$$\Rightarrow |\nabla f(x^{k+1})| \leq \frac{M}{2}|D^2 f(x^k)^{-1}|^2 |\nabla f(x^k)|^2 \leq \frac{M}{2\gamma^2}|\nabla f(x^k)|^2$$

Newton method: proof

Hence letting $g_k = |\nabla f(x^k)|$, for all k one has

$$\log g_{k+1} \le 2 \log g_k + \log \frac{M}{2\gamma^2} \;\Rightarrow\; \log g_k \le 2^k \log g_0 + (2^k - 1) \log \frac{M}{2\gamma^2} = 2^k \log q - \log \frac{M}{2\gamma^2}$$

so that

$$|\nabla f(x^k)| \le \frac{2\gamma^2}{M} q^{2^k}.$$

As f is strongly convex, $\left\langle \nabla f(x^k), x^k - x^* \right\rangle \ge \gamma |x^k - x^*|^2$, and we can conclude.

Newton method: proof

Hence letting $g_k = |\nabla f(x^k)|$, for all k one has

$$\log g_{k+1} \le 2\log g_k + \log\frac{M}{2\gamma^2} \;\Rightarrow\; \log g_k \le 2^k \log g_0 + (2^k-1)\log\frac{M}{2\gamma^2} = 2^k\log q - \log\frac{M}{2\gamma^2}$$

so that

$$|\nabla f(x^k)| \le \frac{2\gamma^2}{M} q^{2^k}.$$

As f is strongly convex, $\left\langle \nabla f(x^k), x^k - x^* \right\rangle \ge \gamma |x^k - x^*|^2$, and we can conclude.

This rate is excellent but: it needs a good initialization, it is hard to compute ($\rightarrow$ "quasi"-Newton methods such as "BFGS" try to approximate the inverse Hessian). Let us return to first order methods and try to improve them...

Heavy ball method (Polyak)

The general idea of a multi-step first order method is to have x^{k+1} depending not only on x^k but also on previous iterates. The *heavy ball* method has the general form:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}),$$

$\alpha, \beta \geq 0$.

Inspired by $m\ddot{x} = -\nabla f(x) - c\dot{x}$ equation of a heavy ball in a potential $f(x)$ with a kinetic friction, discretized as:

$$m\frac{x^{k+1} - 2x^k + x^{k-1}}{(\delta t)^2} + c\frac{x^{k+1} - x^k}{\delta t} = -\nabla f(x^k)$$

(which is the same for $\alpha = \frac{(\delta t)^2}{m+c\delta t}$ and $\beta = \frac{m}{m+c\delta t}$).

Heavy ball method

Theorem (Polyak 87)

Let x^* be a (local) minimizer of f such that $\gamma I \le D^2 f(x^*) \le LI$, and choose α, β with $0 \le \beta < 1$, $0 < \alpha < 2(1+\beta)/L$. There exists $q < 1$ such that if $q < q' < 1$ and if x^0, x^1 are close enough to x^*, one has

$$|x^k - x^*| \le c(q') q'^k .$$

Moreover, this is almost optimal: if

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{\gamma})^2}, \beta = \left(\frac{\sqrt{L} - \sqrt{\gamma}}{\sqrt{L} + \sqrt{\gamma}}\right)^2 \quad \textit{then } q = \frac{\sqrt{L} - \sqrt{\gamma}}{\sqrt{L} + \sqrt{\gamma}} = \frac{\sqrt{Q} - 1}{\sqrt{Q} + 1}.$$

Heavy ball method

Remark: This method requires that f is C^2, γ-convex, with L-Lipschitz gradient (at least near a solution x^*):

$$\gamma I \leq D^2 f \leq LI.$$

Proof: we study the iteration of a linearized system near the optimum: close enough to x^*,

$$x^{k+1} = x^k - \alpha D^2 f(x^*)(x^k - x^*) + o(|x^k - x^*|) + \beta(x^k - x^{k-1}),$$

hence $z^k = (x^k - x^*, x^{k-1} - x^*)^T$ satisfies, for $B = D^2 f(x^*)$:

$$z^{k+1} = \begin{pmatrix} (1+\beta)I - \alpha B & -\beta I \\ I & 0 \end{pmatrix} z^k + o(z^k).$$

$\rightarrow$ We study the eigenvalues of the matrix A which appears in this iteration.

Heavy ball

$$A\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}(1+\beta)I-\alpha B & -\beta I\\ I & 0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} = \rho\begin{pmatrix}x\\y\end{pmatrix}$$

if and only if

$$(1+\beta)x-\alpha Bx-\beta y=\rho x, \quad x=\rho y$$

(and $x, y \neq 0$) hence if $(1+\beta)x-\alpha Bx-(\beta/\rho)x=\rho x$.
We find that

$$Bx = \frac{1}{\alpha}\left(1+\beta-\rho-\frac{\beta}{\rho}\right)x$$

hence $\frac{1}{\alpha}\left(1+\beta-\rho-\frac{\beta}{\rho}\right)=\mu\in[\gamma,L]$ is an eigenvalue of B. We derive the equation

$$\rho^2-(1+\beta-\alpha\mu)\rho+\beta=0$$

which gives two eigenvalues with product β and sum $1+\beta-\alpha\mu$. If $\beta\in[0,1)$ and $-(1+\beta)<1+\beta-\alpha\mu<(1+\beta)$ (extreme cases where $\pm(1,\beta)$ are solutions) then $|\rho|<1$, that is, if $0<\alpha<(2+\beta)/\mu$. Since $\mu<L$ one deduces that if $0\le\beta<1$, $0<\alpha<(2+\beta)/L$, the eigenvalues of A are all in $(-1,1)$ (incidentally, it has $2n$ eigenvalues).

Matrices with spectral radius less than 1

Lemma

Let A be a $N \times N$ matrix and assume that all its eigenvalues (complex or real) have modulus $\leq \rho$. Then for any $\rho' > \rho$, there exists a norm $|\cdot|_*$ in $\mathbb{C}^N$ such that $\|A\|_* := \sup_{|\xi|_* \leq 1} |A\xi|_* < \rho'$.

Matrices with spectral radius less than 1

Proof: up to a change of a basis, A is triangular: there exists P such that

$$P^{-1}AP = T$$

with $T = (t_{i,j})_{i,j}$, $t_{i,i} = \lambda_i$, an eigenvalue, and $t_{i,j} = 0$ if $i > j$. Then, if $D_s = \mathrm{diag}(s, s^2, s^3, \dots, s^N) = (s^i \delta_{i,j})_{i,j}$, $D_s P^{-1} A P D_s^{-1} = (x^s_{i,j})$ with

$$x^s_{i,j} = \sum_{k,l} s^i \delta_{i,k} t_{k,l} s^{-l} \delta_{l,j} = s^{i-j} t_{i,j}$$

and (since $t_{i,j} = 0$ for $i > j$), $x^s_{i,j} \to \lambda_i \delta_{i,j}$ as $s \to +\infty$. Hence, if s is large enough, denoting $|\xi|_\infty = \max_i |\xi_i|$ the ∞-norm,

$$\max_{|\xi|_\infty \le 1} |D_s P^{-1} A P D_s^{-1} \xi|_\infty \le \max_i (|\lambda_i| + (\rho' - \rho)) \le \rho'$$

if s is large. Hence, if $|\xi|_* := |D_s P^{-1} \xi|_\infty$, one has

$$\|A\|_* = \sup_{|\xi|_* \le 1} |A\xi|_* \le \rho'.$$

Heavy ball

It follows, in particular, that if $\rho' < 1$, $\|A^k\|_* \le \|A\|_*^k \le \rho'^k \to 0$ as $k \to \infty$. Applying this to our problem, we see that (choosing $\rho' < 1$)

$$|z^{k+1}|_* = |Az^k + o(z^k)|_* \le (\rho' + \varepsilon)|z^k|_*$$

if $|z^k|_*$ is small enough. Starting from z^0 such that this holds for ε with $\rho' + \varepsilon < 1$, we find that it holds for all $k \ge 0$ and that $|z^{k+1}|_* \le (\rho' + \varepsilon)^k |z^0|_*$, showing the linear convergence. □

Conjugate Gradient

The conjugate gradient is defined as "the best" two-steps method, in the sense that one can define it as follows: given x^k, x^{k-1}, we let $x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$ where α_k, β_k are minimizing

$$\min_{\alpha,\beta} f(x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})).$$

Conjugate Gradient

In particular, we deduce that

$$\left\langle \nabla f(x^{k+1}), \nabla f(x^k) \right\rangle = 0 \quad \text{and} \quad \left\langle \nabla f(x^{k+1}), x^k - x^{k-1} \right\rangle = 0$$

and it also follows

$$\left\langle \nabla f(x^{k+1}), x^{k+1} - x^k \right\rangle = 0.$$

Notice moreover that

$$\nabla f(x^{k+1}) = \nabla f(x^k) - \alpha_k D^2 f(x^k + s(x^{k+1} - x^k))\nabla f(x^k) \\ + \beta_k D^2 f(x^k + s(x^{k+1} - x^k))(x^k - x^{k-1})$$

for some $s \in [0,1]$.

Conjugate Gradient

However, for a general f, it is "conceptual": there is no simple way to compute α_k, β_k. One can show that in *the quadratic case*, that is, if $f(x) = (1/2)\langle Ax, x\rangle - \langle b, x\rangle + c$ (A symmetric), then there are closed forms to compute the parameters. (Using the last formula in the previous slide.)

In that case, one has:

Conjugate Gradient

Lemma

The gradients $p^k = \nabla f(x^k)$ *are all orthogonal.*

Corollary

For a quadratic function, the conjugate gradient is the "best" first order method.

Corollary

A solution is found in $k = \mathrm{rk} A$ *iterations.*

In addition one can show a rate similar to the heavy ball method.

Nesterov's "Accelerated Gradient Descent" (AGD)

(Yu. Nesterov, 1983 / book of 2004)

$x^0 = x^{-1}$ given, x^{k+1} defined by:

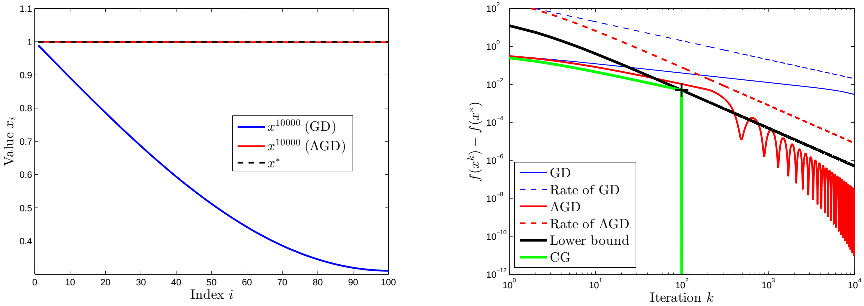
$$\begin{cases} y^k = x^k + \frac{t_k - 1}{t_{k+1}}(x^k - x^{k-1}) \\ x^{k+1} = y^k - \tau \nabla f(y^k) \end{cases}$$

where $\tau = 1/L$ and for instance $t_k = 1 + k/2$. Then,

$$f(x^k) - f(x^*) \leq \frac{2L}{(k+1)^2}|x^0 - x^*|^2$$

$\rightarrow$ optimal. For strongly convex problems, a variant exists with again optimal rate of convergence.

Comparison



Comparison between accelerated vs non-accelerated gradient schemes.

Top: Comparisons of the solutions x of GD and AGD after 10000(!) iterations. Bottom: Rate of convergence for GD, AGD together with their theoretical worst case rates, and the lower bound for smooth optimization. For comparison we also provide the rate of convergence for CG. Note that CG exactly touches the lower bound at $k = 99$ (this is the "hard problem" with $\gamma = 0$, $p = n = 100$)

What about “nonsmooth” problems?

That is: problems where ∇f is not Lipschitz (or even not well defined, later on).

- Non-smooth (“subgradient”) descent;
- Implicit descent.

Nonsmooth problems

Subgradient descent

$$x^{k+1} = x^k - h_k \frac{\nabla f(x^k)}{|\nabla f(x^k)|}.$$

(In practice, the gradient can be replaced with any selection of the "subgradient" if f is not differentiable, definition comes later.)

Nonsmooth problems

Subgradient descent

$$x^{k+1} = x^k - h_k \frac{\nabla f(x^k)}{|\nabla f(x^k)|}.$$

(In practice, the gradient can be replaced with any selection of the "subgradient" if f is not differentiable, definition comes later.)

$$\begin{aligned} |x^{k+1} - x^*|^2 &= |x^k - x^*|^2 - 2\frac{h_k}{|\nabla f(x^k)|}\left\langle \nabla f(x^k), x^k - x^* \right\rangle + h_k^2 \\ &\leq |x^k - x^*|^2 - 2\frac{h_k}{|\nabla f(x^k)|}(f(x^k) - f(x^*)) + h_k^2. \end{aligned}$$

(using $f(x^*) \geq f(x^k) + \left\langle \nabla f(x^k), x^* - x^k \right\rangle$).

Subgradient descent

If we assume in addition f is M-Lipschitz, near x^* at least,

$$\min_{0\le i\le k} f(x^i) - f(x^*) \le M\frac{|x^0 - x^*|^2 + \sum_{i=0}^k h_i^2}{2\sum_{i=0}^k h_i}$$

and choosing $h_i = C/\sqrt{k+1}$ for k iterations, we obtain

$$\min_{0\le i\le k} f(x^i) - f(x^*) \le M\frac{C^2 + |x^0 - x^*|^2}{2C\sqrt{k+1}}$$

(the best choice is $C \sim |x^0 - x^*|$ but this is of course unknown).
In general, one chooses steps such that $\sum_i h_i^2 < +\infty$, $\sum_i h_i = +\infty$, such as $h_i = 1/i$ (actually the best varying choice is rather $h_i \sim 1/\sqrt{i}$, why?). It results in a very slowly converging algorithm which should be used only when there is no other obvious choice.

Nonsmooth problems

Implicit descent

Consider a gradient descent where instead of using the gradient at x^k, one is able to evaluate the gradient in x^{k+1}:

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

- This seems "conceptual", or useless because it seems easy to compute only in situations where $\min f$ also is easy to compute.
- It says that x^{k+1} is a critical point of (and one can ask that it minimises)

  $$f(x) + \frac{1}{2\tau}|x - x^k|^2 \to \min_x = f_\tau(x^k)$$

  and also, that, $x^{k+1} = x^k - \tau \nabla f_\tau(x^k)$.
- Can be showed to always converge to a minimum / critical point with very little assumptions on f. *(Precisions later.)*

Implicit descent

However, we will see that:

- It can be useful for solving composite problems $\min_x f(x) + g(x)$ where f or g is simple and can be treated implicitly;
- Many (simple) algorithms will be, in fact, particular cases of this method (later);
- Sometimes, simply changing the metric makes the method computable, *cf* Lasso problem:

Implicit descent

Example: the LASSO problem

$$\min_x |x|_1 + \frac{1}{2}|Ax - b|^2$$

If $|x|_M^2 = \langle Mx, x\rangle$ and $M = I/\tau - A^*A$, $\tau < 1/|A|^2$, then

$$\min_x \frac{1}{2}|x - x^k|_M^2 + |x|_1 + \frac{1}{2}|Ax - b|^2$$

is solved by

$$x^{k+1} = S_\tau(x^k - \tau A^*(Ax^k - b))$$

where $S_\tau\xi$ is the unique minimizer of

$$\min_x |x|_1 + \frac{1}{2\tau}|x - \xi|^2,$$

called the "shrinkage" operator. This converges with rate $O(1/k)$ to a solution.