CONTINUOUS OPTIMIZATION FOR MACHINE LEARNING

ALESSANDRO BETTI, IMT SCHOOL FOR ADVACED STUDIES, LUCCA

These notes collects some of the material for the PhD course on *Continuous Optimization for Machine Learning* held from March 17 to March 21, 2024, at the University of Siena. Most of the technical content on the gradient-based optimization techniques is drawn from A. Chambolle's notes on *Continuous Optimization*.[1]

A particularly clear book on optimization is Polyak's *Introduction to Optimization* [1].

1. Statistical ML: Risk, Empirical Risk, Uniform Convergence

The classical problem in ML is that of finding a model $f\colon X \to Y$, where X is an input space, where the perceptual data lives and Y is an output space (in which typically is encoded some symbolic properties of percepts). The goodness of the predictions of f is measured by a loss $\ell\colon Y \times Y \to \mathbb{R}_+$. Let us The point is that on $\Omega := X \times Y$ there is in general a (unknown) probability measure π that assigns probabilities to measurable subsets of Ω. Then we can define the functional risk that is a functional that assigns to a (measurable) function f the real number $L(f)$ defined as

eq:fun-risk

$$f \mapsto L(f) := \mathrm{E}\,\ell \circ (f \times \mathrm{id}) = \int_\Omega \ell(f(x), y)\, d\pi(x, y). \tag{1}$$

where $f \times \mathrm{id}\colon X \times Y \to Y \times Y$ maps $(x, y) \mapsto (f(x), y)$. The catch here is that in any practical situation the measure π is unknown and therefore one cannot directly minimize L to find the best possible model. Usually instead of π what is available is a set of *independently identically distributed samples* from π, i.e. a set of measurements $D = \{(x_1, y_1), \dots (x_N, y_N)\}$. Therefore the best I could do is to find a function $\bar{f}$ based on those samples that in general will be suboptimal in the sense that

eq:minimum-problem-fun-risk

$$\inf\{L(f) : f \in \mathrm{dom}(L)\} \le L(\bar{f}). \tag{2}$$

What are the options to compute $\bar{f}$? And how good can this solution can be in terms of the functional risk? Let us make an explicit example:

exmp:fun-risk

Example 1.1. suppose $X = \mathbb{R}$, $Y = \mathbb{R}$ and $\Omega = \mathbb{R}^2$. The measure π is defined as follows

$$\pi(A) = \frac{1}{\pi} \int_A e^{-(x-y)^2} e^{-x^2}\, dx dy,$$

for every measurable set $A \in \mathbb{R}^2$. Choose $\ell(a, b) = (a - b)^2/2$ and then

$$L(f) = \frac{1}{\pi} \int_{\mathbb{R}^2} (f(x) - y)^2 e^{-(x-y)^2} e^{-x^2}\, dx dy \ge 0.$$

[1] https://www.ceremade.dauphine.fr/~chambolle/data/medias/optimisation/coursopti2024.pdf

Let us now try to understand what is the inf of this risk on the domain:

$$\mathrm{dom}(L) = \{f \in \mathbb{R}^{\mathbb{R}} : L(f) < +\infty\}.$$

For the moment let us try to find a candidate by using the indirect method of the calculus of variations. First of all let us rewrite L so that it is in a canonical form.

$$\begin{aligned} L(f) &= \frac{1}{\pi}\int_{\mathbb{R}^2}(f^2(x) - 2f(x)y + y^2)e^{-(x-y)^2}e^{-x^2}\,dxdy. \\ &= \frac{1}{\pi}\Bigg[\int_{\mathbb{R}}\Bigg[f^2(x)e^{-x^2}\Big(\int_{\mathbb{R}} e^{-(x-y)^2}\,dy\Big) \\ &\qquad - 2f(x)e^{-x^2}\Big(\int_{\mathbb{R}} ye^{-(x-y)^2}\,dy\Big)\Bigg]dx\Bigg] + C \end{aligned}$$

Recalling that $\int_{\mathbb{R}} \exp(-x^2)\,dx = \sqrt{\pi}$ we have

$$\int_{\mathbb{R}} e^{-(x-y)^2}\,dy = \sqrt{\pi}.$$

Similarly

$$\int_{\mathbb{R}} ye^{-(x-y)^2}\,dy = 2x\int_{\mathbb{R}} e^{-z^2}\,dz + 4\int_{\mathbb{R}} ze^{-z^2}\,dz = \sqrt{\pi}x.$$

Therefore, assuming that the minima of L exists these coincides with the minima of

$$\int_{\mathbb{R}}(f^2(x) - 2xf(x))e^{-x^2}\,dx.$$

Stationarity conditions for this functional gives $f^*(x) = x$. Moreover we have immediately (with some easy integration of gaussians) that $L(f^*) = 1/2$.

Now we have to check that our candidate actually is a minimum. To do so notice that

$$\begin{aligned} L(f) &= \frac{1}{\pi}\int_{\mathbb{R}^2}(f(x) - x + x - y)^2e^{-(x-y)^2}e^{-x^2}\,dxdy \\ &= \frac{1}{\pi}\Bigg\{\int_{\mathbb{R}}(f(x) - x)^2e^{-x^2}\Big(\int_{\mathbb{R}} e^{-(x-y)^2}\,dy\Big)\,dx \\ &\qquad + 2\int_{\mathbb{R}}(f(x) - x)e^{-x^2}\Big(\int_{\mathbb{R}}(x-y)e^{-(x-y)^2}\,dy\Big)\,dx \\ &\qquad + \int_{\mathbb{R}} e^{-x^2}\Big(\int_{\mathbb{R}}(x-y)^2e^{-(x-y)^2}\,dy\Big)\,dx\Bigg\} \geq \frac{1}{\pi}(0 + 0 + \pi/2) = 1/2. \end{aligned}$$

This proves that f^* is a minimum of L.

The previous example should have clarified what we mean by minimization of functional risk. In the other hand it is clear that when π is unknown the whole exercise ceases to have meaning. In Machine Learning we usually approximate the measure π making use of the available samples in D. More precisely we replace π with

$$\pi_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{(x_i,y_i)},$$

which is called the empirical distribution. When we compute the risk in Eq. (1) we obtain the *Empirical Risk*:

eq:empirical-risk

$$L_N(f) = \int_\Omega \ell(f(x), y)\, d\pi_N(x, y) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), y_i). \tag{3}$$

Another important difference with respect to what we discussed in Example 1.1 and from the definition of functional risk is the domain in which the optimum of the risk is sought. Almost in all modern ML problems are defined on spaces of parametric functions so that the variational problem in Eq. (2) is replaced with a minimization problem in $\mathbb{R}^n$. So let us denote with $\mathscr{P} \subset \mathrm{dom}(L)$ the space of such parametric functions (it could be for instance the space of neural networks with a specific architecture). Then we could choose

eq:empirical-risk-minimization

$$f^{(N)} \in \arg\min\{L_N(f) : f \in \mathscr{P}\}. \tag{4}$$

Then we ca split the risk as follows:

$$\begin{aligned} L(f^{(N)}) - \inf\{L(f) : f \in \mathrm{dom}(L)\} = {} & (L(f^{(N)}) - \inf\{L(f) : f \in \mathscr{P}\}) \\ & + (\inf\{L(f) : f \in \mathscr{P}\} - \inf\{L(f) : f \in \mathrm{dom}(L)\}), \end{aligned}$$

The first term represent, once we restrict into a specific set of functions $\mathscr{P}$ how difficult is to estimate the optimal function from a finite sample of the distribution, the second term instead measure how well the functions in $\mathscr{P}$ approximate the optimum of the functional risk. The first term can be usually controlled using the uniform laws of large numbers.

In these lectures we will discuss how we can solve the problem in Eq. (4).

2. Gradients

sec:gradients

In this course we are interested in minimization problem of the form

$$\min\{f(x) : x \in X\}, \tag{5}$$

where X is an Hilbert space and $f \in C^1(X; \mathbb{R})$.

Actually most of our interest lies in the case where X is a finite dimensional vector space, so basically $X = \mathbb{R}^n$ (or subsets of $\mathbb{R}^n$). Many of these results however can be extended to the case of Hilbert spaces.

We recall that the *differential* of f at a point x, that we will denote $df(x) \in X^*$ the linear part of the affine approximation of f around the point x, i.e. the linear function that satisfies [2]

$$f(y) = f(x) + \langle df(x), (y - x)\rangle + o(|x - y|).$$

If such linear function exists at a point x we say that f is Fréchet differentiable at x and the differential is sometimes also called Fréchet derivative.

Exercise 2.1. Show that if it exists $df(x)$ is unique.

If in addition the map $x \in X \mapsto df(x) \in X^*$ is defined everywhere and it is continuous we say that $f \in C^1(X, \mathbb{R})$.

[2] X^* is the (topological) dual of X, i.e. the space of all linear functional on X.

Remark (Gateau derivative). The notion of Fréchet derivative extends the usual notion of partial derivative. There is however another notion of differentiability that can be naturally introduced in normed vector spaces. Consider a point $h \in X$ with $h \neq 0$ and define the funtion the function $\phi(t) := f(x+th)$ that maps $t \in I(h) \to f(x+th) \in \mathbb{R}$, where $I(h)$ is a suitable open interval containing the origin. Then if ϕ is differentiable at $t = 0$ we say that f is *Gâteau* differentiable and we define the *Gâteau* derivative along the direction h as

$$df(x;h) := \phi'(0).$$

If f is differentiable at some point x, then

$$f(x+th) = f(x) + t\langle df(x), h\rangle + o(|th|)$$

or

$$\frac{f(x+th) - f(x)}{t} = \langle df(x), h\rangle + o(1)$$

which as $t \to 0$ becomes $df(x;h) = \langle df(x), h\rangle$.

Proposition 2.1. *Let X be Hilbert and $x \in X$ be a point with $df(x) \neq 0$. Let $\nabla f(x)$ be the vector $v \in X$ for which the Riesz representation $\langle df(x), h\rangle = v \cdot h$, $\forall h \in X$ holds. Then $\nabla f(x)$ is the direction in which f increases more rapidly in the sense that*

$$\frac{\nabla f(x)}{|\nabla f(x)|} \in \arg\max\{df(x;h) : h \in X \quad \textit{and} \quad |h| = 1\}.$$

Proof. Let us choose the direction $h = \nabla f(x)/|\nabla f(x)|$, then $df(x;h) = \langle df(x), h\rangle = |\nabla f(x)|$. Moreover we have that by Cauchy Schwartz that

$$|df(x;h)| \leq |\nabla f(x)||h|,$$

and that the equality holds for $h = \alpha \nabla f(x)$ with $\alpha \in \mathbb{R}$. This means that if we restrict to directions with $|h| = 1$ we have

$$\arg\max\{df(x;h) : h \in X \quad \text{and} \quad |h| = 1\} = \left\{ \frac{\nabla f(x)}{|\nabla f(x)|} \right\}.$$

□

This property of the gradient will be key when we start to explore minimization methods and algorithms and it is essentially the starting point of all continuous optimization and all Machine Learning techniques.

3. Some Remarks on Curves of Maximal Slope

In this section we will try to give a more intrinsic and broad interpretation of gradient methods showing how it is possible to generalize the ideas behind gradient descent in settings where it is not even possible to properly define what a gradient. Interestingly the concept of *curve of maximal slope* can still be given in many circumstances.

The main motivation for this discussion is the fact that in Hilbert spaces, as we discussed in Section 2 we can talk of gradient flows, i.e., roughly speaking solutions of:

eq:GF (6) $$u'(t) = -\nabla\phi(u(t)).$$

The reason why this dynamics are of interested is because of the the following formal computation

$$\frac{d}{dt}\phi(u(t)) = \nabla\phi(u(t)) \cdot u'(t) = -|\nabla\phi(u(t))|^2 \leq 0$$

which means that ϕ decreases along the trajectory $t \mapsto u(t)$.

The interesting fact is that these ideas can be extended also in much more less structured spaces like *metric spaces* (and beyond). In this setting the generalization of gradient flows are called *curve of maximal slope* which was originally introduced in [3].

In metric spaces it is possible to define quantities that are related those employed in Eq. (6); more precisely to the modulus of that equation:

eq:GF-mod

$$|u'(t)| = |\nabla\phi(u(t))|. \tag{7}$$

The metric analogue of $|u'(t)|$ will be the *metric derivative* while the analogue of $|\nabla\phi(u(t))|$ would be an upper gradient of ϕ. Remarkably even thought it seems that in going from Eq. (6) to (7) there is loss of information since we are going from a relation between vector to a relation between scalar the information can be recovered by looking at the derivative of the energy.

To understand this let us consider the case of Hilbert spaces and consider again

$$\frac{d}{dt}\phi(u(t)) = u'(t) \cdot \nabla\phi(u(t)),$$

now, from Eq. (6) we know that $u'(t)$ and $\nabla\phi(u(t))$ lies on the same line, and therefore by Cauchy-Schwartz $|u'(t) \cdot \nabla\phi(u(t))| = |u'(t)||\nabla\phi(u(t))|$ but Eq. (6) tells us more, since it says that they are opposite so that we can conclude that

$$\frac{d}{dt}\phi(u(t)) = u'(t) \cdot \nabla\phi(u(t)) = -|u'(t)||\nabla\phi(u(t))|$$

Notice that up to now we have only used the information about the relative direction of the vectors $u'(t)$ and $\nabla\phi(u(t))$ but not the relations on their modulus expressed by Eq. (7). If we use this we finally get

$$\frac{1}{2}|u'(t)|^2 + \frac{1}{2}|\nabla\phi(u(t))|^2 = -\frac{d}{dt}\phi(u(t)).$$

Now the point is that we can make sense of this for instance in metric spaces replacing $|u'|^2$ the square of the metric distance and $\nabla\phi(u)$ with an upper gradient $g(u)$. Once we do this the idea is to say that u is a curve of maximal slope if

$$\frac{1}{2}\int_s^t \left(|u'|^2(r) + |g(u(r))|^2\right) dr \leq \phi(u(s)) - \phi(u(t)) \tag{8}$$

for $\mathscr{L}^1$-a.e t, s, with $t < s$.

Define the metric derivative and the local upper gradient.

To get some intuition let us for a moment consider the case where the metric space is $\mathscr{S} := \mathbb{R}^N$. The gradient $\nabla\phi$ of a smooth real functional $\phi\colon \mathscr{S} \to \mathbb{R}$ can be defined taking the derivative of ϕ along regular curves, i.e. we say that

$g = \nabla\phi$ if and only if for every regular curve $v\colon (0,+\infty) \to \mathscr{S}$ we have[3]

$$(\phi \circ v)' = g(v) \cdot v'.$$

3.1. **Gradient Flows.** Mettere l'esempietto del corso a UCE.

3.2. **Minimizing Movements.** Il primo esempio di Gobbino secondo me è emblematico

4. Gradient Descent

The simplest iterative method to minimize a function f, whenever the gradient is defined is to use the *gradient descent algorithm* with step (or learning rate) τ. The algorithm is the following:

(1) Start from $x^0 = x_0 \in X$;
(2) For $k > 0$ compute

eq:GD

$$x^{k+1} = x^k - \tau \nabla f(x^k). \tag{9}$$

Notice that in terms of gradient flows this can be considered the implicit Euler scheme associated to the gradient flow $x' = -\nabla f(x)$. Then we have the following result

thm:GD-plain

Theorem 4.1. *Let $f \in C^1(X;\mathbb{R})$, $\inf f > -\infty$ and ∇f L-Lipschitz. Choose $0 < \tau < 2/L$, then the method in Eq. (9) converges in the sense that $\nabla f(x^k) \to 0$.*

Before proving this result we will recall a standard trick that we will often employ

lemma:simple-fact

Lemma 4.2. *Let $f \in C^1(X;\mathbb{R})$, if ∇f is L-Lipschitz, then*

$$f(x - \tau y) = f(x) - \int_0^\tau \nabla f(x - sy) \cdot y \, ds \quad \forall x, y \in X.$$

Proof. For fixed x, y in X let $\phi(s) := f(x - sy)$. Since f is differentiable on X we have that $f(x - (s+t)y) = f(x - sy) + \nabla f(x - sy) \cdot (-ty) + o(|ty|)$. Dividing by t and letting $t \to 0$ we get $\phi'(s) = -\nabla f(x - sy) \cdot y$. Now using the fundamental theorem of calculus we have $\phi(\tau) - \phi(0) = \int_0^\tau \phi'(s)\, ds$ which is the claim of the lemma. □

Proof of Theorem 4.1. Consider $f(x^{k+1})$, using Eq. (9) we have that $f(x^{k+1}) = f(x^k - \tau \nabla f(x^k))$. Now using Lemma 4.2 with $y = \nabla f(x^k)$ we get:

$$f(x^{k+1}) = f(x^k) - \int_0^\tau \nabla f(x^k - s\nabla f(x^k)) \cdot \nabla f(x^k)\, ds$$

Adding and subtracting $\nabla f(x^k)$ inside the integral in the first term of the scalar product we get

$$f(x^{k+1}) = f(x^k) - \tau |\nabla f(x^k)|^2 + \int_0^\tau (\nabla f(x^k) - \nabla f(x^k - s\nabla f(x^k))) \cdot \nabla f(x^k)$$

[3] We recall that a regular curve v is a map in $C^1((0,+\infty), \mathscr{S})$ such that $v'(s) \neq 0$ for all $s > 0$.

Now using the fact that for all $x \in \mathbb{R}$ $x \le |x|$ and the Lipshitz condition on the gradient of f we get [4]

$$f(x^{k+1}) \le f(x^k) - \tau\left(1 - \frac{L\tau}{2}\right)|\nabla f(x^k)|^2.$$

If we iteratively exploit this property we get

$$f(x^n) \le f(x_0) - \tau\left(1 - \frac{L\tau}{2}\right)\sum_{k=0}^{n-1}|\nabla f(x^k)|^2.$$

If $\tau(1 - L\tau/2) > 0$, i.e. if $\tau < 2/L$ then by hypothesis on the boundedness from below of f we get:

$$-\infty < \inf f \le f(x_n) \le f(x_0) - \tau\left(1 - \frac{L\tau}{2}\right)\sum_{k=0}^{n-1}|\nabla f(x^k)|^2$$

which means that

$$\tau\left(1 - \frac{L\tau}{2}\right)\sum_{k=0}^{n-1}|\nabla f(x^k)|^2 \le f(x_0) - \inf f.$$

This implies that the series $\sum_{k=0}^{+\infty}|\nabla f(x^k)|^2$ since the sequence of its partial sums is an increasing sequence bounded from above. Hence there exists a subsequence (x^{i_n}) such that $|\nabla f(x^{i_n})|^2 \to 0$. □

Exercise 4.1. Clearly the fact that the gradient of f by itself does not ensure convergence of (x^n). This can be proved however if we further assume cohercivity of f. Prove this fact.

4.1. **Convex Case.** Let us start recalling that a function $f\colon X \to \mathbb{R}$ is said to be convex if given two arbitrary points $x, y \in X$ we have

$$f((1-\lambda)x + \lambda y) \le (1-\lambda)f(x) + \lambda f(y) \qquad \forall \lambda \in [0,1]$$

that is to say the values of the functions on the segment joining x to y are always below the line that join $f(x)$ to $f(y)$.

Lemma 4.3. *Let $f \in C^2((a,b),\mathbb{R})$, then f is convex if and only if $f'' \ge 0$ on (a,b).*

Proof. See [4, Theorem 4.4 p.26] □

lem:1d-conv **Lemma 4.4.** *Let $f \in C^2(\mathbb{R}^N;\mathbb{R})$, then f is convex if and only if $D^2 f \ge 0$.*

Proof. Fix x and v in $\mathbb{R}^n$ and let $\psi(t) = f(x + tv)$ for $t \in \mathbb{R}$. Now since f is convex we say that also ψ is convex since

$$\begin{aligned}\psi((1-\lambda)t + \lambda s) = f(x + ((1-\lambda)t + \lambda s)v) &= f((1-\lambda)(x+tv) + \lambda(x+sv)) \\ &\le (1-\lambda)\psi(t) + \lambda\psi(s).\end{aligned}$$

On the other hand by definition the convexity of f is obtained from the convexity of ψ for all x and v in $\mathbb{R}^n$. Now we have $\psi'(t) = \partial_i f(x+tv)v_i$

[4] If $f \in C^2$ the condition

eq:decrese-f
$$(\nabla f(x) - \nabla f(y)) \cdot (x - y) \le L|x-y|^2. \tag{10}$$

with $x = x^k$ and $y = x^k - s\nabla f(x^k)$ is equivalent to say that $D^2 f$ is bounded from above by $L\,\mathrm{Id}$.

and $\psi''(t) = \partial_j\partial_i f(x+tv)v_iv_j = v\cdot D^2f(z)v$ with $z = x+tv$. Form Lemma 4.4 we therefore get that convexity of ψ for all x and v is equivalent to $v\cdot D^2f(z)v \geq 0$ for all v and z in $\mathbb{R}^n$. □

Under this assumption we can be much more precise in the convergence analysis However for Machine Learning purposes with Neural Networks the assumption of convexity usually never met so we won't try to stress and carry on derivation in this case in full details.

One important property that we have in the convex case is the following theorem:

Theorem 4.5 (Baillon-Haddad)**.** *If f is convex and ∇f is L-Lipschitz, then for all $x, y \in X$,*

$$(\nabla f(x) - \nabla f(y))\cdot(x-y) \geq \frac{1}{L}|\nabla f(x) - \nabla f(y)|^2.$$

and ∇f is said to be $1/L$-co-coercive.

Proof. We will give the proof in the finite dimensional case (so X is basically $\mathbb{R}^N$) and with a stronger regularity assumption: $f \in C^2(X;\mathbb{R})$. A general proof without this assumptions can be given exploiting results of convex analysis. Since f is convex we have that the hessian is positive semidefinite $D^2f \geq 0$ while since its gradient is L-Lipshitz we have that the hessian is bounded by $L\,\mathrm{Id}$; hence $0 \leq D^2f \leq L\,\mathrm{Id}$. Then

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

with $A = \int_0^1 D^2f(y+s(x-y))ds$ symmetric and with $0 \leq A \leq L\,\mathrm{Id}$. Hence since A is positive semidefinite it admits $A^{1/2}$, which commute with A, then:

$$\begin{aligned}|\nabla f(x) - \nabla f(y)|^2 = |A(x-y)|^2 &= AA^{1/2}(x-y)\cdot A^{1/2}(x-y)\\ &\leq L\langle A^{1/2}(x-y), A^{1/2}(x-y)\rangle\\ &\leq L\langle A(x-y), x-y\rangle\\ &= L\langle \nabla f(x) - \nabla f(y), x-y\rangle\end{aligned}$$

which is the result. □

In order to give a result on the convergence rate of gradient descent in the convex case we also need the following additional result

lem:contraction **Lemma 4.6.** *Let f be convex with L-Lipshitz gradient, then the mapping $T_\tau = \mathrm{id} - \tau\nabla f$ is 1-Lipshitz.*

Proof.

$$\begin{aligned}|T_\tau(x) - T_\tau(y)| &= |x-y|^2 - 2\tau(x-y)\cdot(\nabla f(x) - \nabla f(y))\\ &\quad + \tau^2|\nabla f(x) - \nabla f(y)|^2\\ &\leq |x-y|^2 - \frac{2\tau}{L}\left(1 - \frac{\tau L}{2}\right)|\nabla f(x) - \nabla f(y)|^2.\end{aligned}$$

□

prop:conv-char **Proposition 4.7.** *$f \in C^1(\mathbb{R}^N, \mathbb{R})$ is convex if and only if*

$$f(y) \geq f(x) + \nabla f(x)\cdot(y-x). \tag{11}$$

Proof. Assume first that f is convex and define for any $x, y \in \mathbb{R}^N$ the function $\psi(t) := f(x + t(y - x))$ for $t \in [0, 1]$. Since f is convex also ψ is convex with respect to t. As a direct consequence of the definition of a convex function we get that $\psi(1) - \psi(0) \geq \psi'(0)$. Now $\psi'(0) = \nabla f(x) \cdot (y - x)$ from which we get $\nabla f(x) \cdot (y - x) \leq f(y) - f(x)$. On the other hand f is convex if it is convex its restriction along every possible line. □

Now assume that f in addition admits a minimizer x^* then from Proposition 4.7 applied $x = x^k$ and $y = x^*$ we get: $f(x^k) \geq f(x) + \nabla f(x) \cdot (x^* - x^k)$ from where

$$f(x^k) - f(x^*) \leq -\nabla f(x^k) \cdot (x^* - x^k) \leq |\nabla f(x^k)||x^* - x^k|$$

Since x^* is a minimum $T_\tau(x^*) = x^*$, then applying Lemma 4.6 we get $|x^{k+1} - x^0| = |T_\tau(x^k) - T_\tau(x^*)| \leq |x^k - x^*|$ which means that $|x^k - x^*| \leq |x_0 - x^*|$. Therefore $f((x^k) - f(x^*))/|x_0 - x^*| \leq |\nabla f(x^k)|$. Form (10) with $C = \tau(1 - L\tau/2)$ assuming $\tau L < 2$ we have that $C|\nabla f(x^k)|^2 \leq f(x^k) - f(x^{k+1})$ and defining $\delta_k = f(x^k) - f(x^*)$ we finally get

eq:recurrence

$$\Delta_{k+1} \leq \Delta_k - C\frac{\Delta_k^2}{|x_0 - x^*|^2}. \tag{12}$$

lem:recurrence

Lemma 4.8. *Let (a_k) be a sequence of nonnegative numbers satisfying for $k \geq 0$:*

$$a_{k+1} \leq a_k - c^{-1}a_k^2, \quad c > 0$$

Then, for all $k \geq 0$,

$$a_k \leq \frac{c}{k+1}.$$

Proof. First observe that if we replace a_k with ca_k, the property becomes $a_{k+1} \leq a_k - a_k^2$; hence it is enough to prove it for $c = 1$. Then, as $a_k(1 - a_k) \geq a_{k+1} \geq 0$, 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

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

since $0 \leq a_k \leq a_{k-1}$. Hence $(1 - (k+1)a_k)(1 + a_{k-1}) \geq 0$. It follows that $(k+1)a_k \leq 1$. □

Lemma 4.8 applied to Eq. (12) prove the following bound for Δ_k:

$$\Delta_k \leq \frac{|x_0 - x^*|^2}{C(k+1)}.$$

5. Backpropagation

We commented on the fact tha ML models are trained via minimization of the empirical risk, defined in Eq. (3). The training is done with a gradient-based method, hence for a fixed example (x_i, y_i) we need to compute the gradient of $\ell(f(x_i; w), y_i)$, where w are the parameters of the NN. Now

$$\frac{\partial}{\partial w}\ell(f(x_i; w), y_i) = \ell_x(f(x_i; w), y_i) \cdot \frac{\partial}{\partial w} f(x_i; w).$$

Since x_i for the moment is fixed we can define $\varphi(w) := f(x_i; w)$ and our problem therefore become that of computing the gradient of φ. Without knowing anything about the structure of φ typically we would require to compute it numerically, therefore for every component

6. Heavy Ball method

6.1. **Calculus of Variations.**

7. Stochastic Gradient Descent

Let $f_i := \ell(\cdot, z_i)$ be with L_i-Lipschitz gradient, let us further denote with f the empirical risk $f(\omega) = (1/m)\sum_{i=1}^m f_i(\omega)$. With this notation then the *SGD* algorithm becomes

(1) Choose $\omega^0 \in \mathbb{R}^N$ and set $x^0 = \omega^0$;
(2) For every $k \geq 1$ select $i_k \in \{1, \dots m\}$ sampled with probability $1/m$ and do:
$$x^{k+1} = x^k - \tau \nabla f_{i_k}(x^k).$$

This process involves two set of random variables $X^1, X^2, \dots$ and $I_1, I_2, \dots$. While the I_k are iid with uniform distribution over $\{1, \dots, m\}$, the X^k are clearly not independent. However we notice that the randomness of X only depends on that of I since once we fix ω^0 and we sample $I_1 = i_1, I_2 = i_2, \dots$ the value of X^k is completely determined by Eq. (2). Now let us see what we can say in terms of a convergence analysis of the algorithm[5]. In general for any C^1 function we can write $f(x - \tau y) = f(x) - \int_0^\tau \nabla f(x - sy) \cdot y \, ds$ for any two point x and y, therefore we immediately get for every $j \in \{1, \dots, m\}$

$$\begin{aligned} f_j(X^k - \tau \nabla f_{I_k}(X^k)) &= f_j(X^k) - \int_0^\tau \nabla f_j(X^k - s\nabla f_{I_k}(X^k)) \cdot \nabla f_{I_k}(X^k)\, ds \\ &= f_j(X^k) - \tau \nabla f_j(X^k) \cdot \nabla f_{I_k}(X^k) \\ &\quad + \int_0^\tau \Big(\nabla f_j(X^k) - \nabla f_j(X^k - s\nabla f_{I_k}(X^k)) \Big) \cdot \nabla f_{I_k}(X^k)\, ds. \end{aligned}$$

at this point we can use the fact that if ∇f is Lipshitz we have for any two points x and y $(\nabla f(x) - \nabla f(y))(x - y) \leq |(\nabla f(x) - \nabla f(y))(x - y)| \leq |(\nabla f(x) - \nabla f(y))||(x - y)| \leq L|x - y|^2$; in our case with the choice $x = X^k$ and $y = X^k - s\nabla f_{I_k}(X^k)$ we get

$$\begin{aligned} f_j(X^{k+1}) &\leq f_j(X^k) - \tau \nabla f_j(X^k) \cdot \nabla f_{I_k}(X^k) + L_j |\nabla f_{I_k}|^2 \int_0^\tau s \, ds \\ &= f_j(X^k) - \tau \nabla f_j(X^k) \cdot \nabla f_{I_k}(X^k) + \frac{L_j \tau^2}{2} |\nabla f_{I_k}|^2. \end{aligned}$$

Now summing over all j and dividing by m, once we pose $\overline{L} = \sum_{j=1}^m L_j/m$ we get

$$f(X^{k+1}) \leq f(X^k) - \tau \nabla f(X^k) \nabla f_{I_k}(X^k) + \frac{\overline{L}\tau^2}{2} |\nabla f_{I_k}(X^k)|^2.$$

[5]The following convergence analysis is partially taken from the lecture notes on *Continuous optimization* by Antonin Chambolle available here: `http://www.cmap.polytechnique.fr/~\/antonin/Opti/CoursOpti2020.pdf`

Then the conditioned expectation of $f(X^{k+1})$ given X^k, since I_k are iid with uniform distribution over $1,\dots,m$ is given by

$$\begin{aligned}\mathrm{E}(f(X^{k+1})|X^k) &\le f(X^k) - \tau|\nabla f(X^k)|^2 + \frac{\overline{L}\tau^2}{2}\Big(\frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)|^2\Big)\\ &= f(X^k) - \tau\Big(1-\frac{\overline{L}\tau}{2}\Big)|\nabla f(X^k)|^2 + \frac{\overline{L}\tau^2}{2}\Big(\frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)-\nabla f(X^k)|^2\Big).\end{aligned}$$

Where the last equality follows from the fact that

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)-\nabla f(X^k)|^2 &= \frac{1}{m}\sum_{i=1}^m (\nabla f_i(X^k)-\nabla f(X^k))\cdot(\nabla f_i(X^k)-\nabla f(X^k))\\ &= \frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)|^2 + |\nabla f(X^k)|^2 - \frac{2}{m}\sum_{i=1}^m \nabla f_i(X^k)\cdot\nabla f(X^k)\\ &= \frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)|^2 - |\nabla f(X^k)|^2,\end{aligned}$$

that is an instance of the very well known theorem of the variance of a random variable $\mathrm{E}(X-\mathrm{E}X)^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2$. Which means that

$$\mathrm{E}f(X^{k+1}) \le \mathrm{E}f(X^k) - \tau\Big(1-\frac{\overline{L}\tau}{2}\Big)\mathrm{E}|\nabla f(X^k)|^2 + \frac{\overline{L}\tau^2}{2}\mathrm{E}\Big(\frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)-\nabla f(X^k)|^2\Big).$$

Notice that as it happens fro *GD* this inequality shows that without the last term and when we choose $\tau < 2/\overline{L}$ the $\mathrm{E}f(X^k)$ will indeed decrease, and in that case we would indeed have that $\nabla f(X^n) = 0$ a.s. as n is large. Indeed assuming (as it is natural) that $\inf f > -\infty$ we would have $\mathrm{E}f(X^n) + \tau(1-\overline{L}\tau^2/2)\sum_{k=0}^{n-1}\mathrm{E}|\nabla f(X^k)|^2 \le f(x^0)$. However when the term $\mathrm{E}|\nabla f(X^k)|^2$ becomes comparable to $\mathrm{E}(\sum_{i=1}^m |\nabla f_i(X^k)-\nabla f(X^k)|^2/m)$ the above analysis does not hold anymore and in general we cannot expect convergence (notice that $\mathrm{E}|\nabla f(X^k)|^2$ is of the same order of $|X^{k+1}-X^k|^2$ so we can expect this term to become arbitrarily small). The classical approach to show convergence of this algorithm, due to Robbins and Monro, is to exploit the fact that the unwanted term proportional to the variance of the gradient is of second order in τ, so that the main idea is to choose a vanishing rate $\tau_k \to 0$. With a variable τ_k we get

$$\mathrm{E}f(X^n) \le f(x^0) - \sum_{k=0}^{n-1}\tau_k\Big(1-\frac{\overline{L}\tau_k}{2}\Big)\mathrm{E}|\nabla f(X^k)|^2 + \sum_{k=0}^{n-1}\frac{\overline{L}\tau_k^2}{2}\mathrm{E}\Big(\frac{1}{m}\sum_{i=1}^m |\nabla f_i(X^k)-\nabla f(X^k)|^2\Big).$$

If we additionally require $\overline{L}\tau_k \le 1$ so that $-\tau_k + \overline{L}\tau_k(\tau_k/2) \le -\tau_k/2$, and that the variance of the gradient of f is globally bounded i.e. $\forall x \in \mathbb{R}^N$ we have $\sum_{i=1}^m |\nabla f_i(x)-\nabla f(x)|^2 \le m\sigma^2$, then

$$-\infty < \inf f \le \mathrm{E}f(X^n) \le f(x^0) - \frac{1}{2}\sum_{k=0}^{n-1}\tau_k\mathrm{E}|\nabla f(X^k)|^2 + \frac{\overline{L}\sigma^2}{2}\sum_{k=0}^{n-1}\tau_k^2,$$

which means

$$\frac{1}{2}\sum_{k=0}^{n-1}\tau_k \,\mathrm{E}\,|\nabla f(X^k)|^2 \le f(x^0) - \inf f + \frac{\overline{L}\sigma^2}{2}\sum_{k=0}^{n-1}\tau_k^2.$$

Under the assumption that the series $\sum_{k=0}^{+\infty}\tau_k^2 < +\infty$, the above inequality states that the increasing sequence (since the terms in the sum are positive) of partial sums $\sum_{k=0}^{n-1}\tau_k \,\mathrm{E}\,|\nabla f(X^k)|^2$ is bounded from above and therefore admits a finite limit, which in turn implies the convergence of $\sum_{k=0}^{+\infty}\tau_k \,\mathrm{E}\,|\nabla f(X^k)|^2$. If we now choose $\sum_{k=0}^{+\infty}\tau_k = +\infty$ we immediately have that there must be a subsequence (i_n) such that $\mathrm{E}\,|\nabla f(X^{i_n})|^2 \to 0$. Because strong convergence in L^p implies μ-a.e. convergence for a subsequence in a measure space (see [1] pag. 15–16), then (up to a relabelling of subsequences) it means that

$$\Pr(\lim_{n\to+\infty}\nabla f(X^n) = 0) = 1,$$

meaning that $\nabla f(X^n) \to 0$ a.s. for large n. With the additional hypothesis of f being coercive we have the a.s. convergence to a stationary point.check all this can be terribly wrong

We now show that Eq. (3) also give an estimate on the rate of the convergence of the algorithm. Indeed from Eq. (3) follows that

$$\min_{k\in\{0,\dots,n-1\}}\mathrm{E}\,|\nabla f(X^k)|^2 \le \frac{2(f(x^0)-\inf f) + \overline{L}\sigma^2\sum_{k=0}^{n-1}\tau_k^2}{\sum_{k=0}^{n-1}\tau_k}.$$

Hence the convergence rate is dictated by the ratio $\rho = \sum_{k=0}^{n-1}\tau_k^2/\sum_{k=0}^{n-1}\tau_k$; for example if we choose $\tau_k = 1/(1+k)$ we get $\rho \sim 1/\log n$, while if we choose $\tau_k = 1/\sqrt{1+k}$ we would get $\rho \sim \log n/\sqrt{n}$. Notice that if we fix the number of steps n one can then also use a fixed rate τ; in this case the optimal choice of τ is the one that minimize the rhs of Eq. (4). This correspond to the choice $\tau^2 = (2(f(x^0)-\inf f)/(\overline{L}\sigma^2 n)$ which gives

$$\min_{k\in\{0,\dots,n-1\}}\mathrm{E}\,|\nabla f(X^k)|^2 \le 2\frac{\sqrt{2(f(x^0)-\inf f)\overline{L}}}{\sqrt{n}}\sigma.$$

* * *

Assuming that the above derivation is correct we would have the following results:

Definition 7.1 (Stochastic Gradient Descent)**.** The discrete stochastic process $(X^k, I_k)_{k\ge0}$ is called a SGD if chosen $X^0 = x^0 \in \mathbb{R}^N$, X^{k+1} is chosen as follow

$$X^{k+1} = X^k - \tau_k\nabla f_{I_k}(X^k), \quad \tau_k > 0,$$

with $I_k \sim U(\{1,\cdots,m\})$.

Proposition 7.1. *Let $f_i\colon \mathbb{R}^N \to \mathbb{R}$ be such that for all $i = 1,\dots,m$ the following assumptions holds*

(1) *$f_i \in C^1(\mathbb{R}^N;\mathbb{R})$; This assumption I'm sure that it can be relaxed*
(2) *∇f_i is L_i-Lipshitz;*
(3) *$\inf f > -\infty$;*
(4) *$\forall x \in \mathbb{R}^N$: $\sum_{i=1}^m |\nabla f_i(x) - \nabla f(x)|^2 \le m\sigma^2$, probably can be relaxed but anyway reasonable*

where we let $f := (1/m)\sum_{i=1}^m f_i$. *Moreover let* $(X^k, I_k)_{k\geq 0}$ *be a SGD. If* $\tau_k \leq 1/\overline{L}$ *for all* $k \geq 0$, $\sum_{k=0}^{\infty} \tau_k = +\infty$ *and* $\sum_{k=1}^{+\infty} \tau_k^2 < +\infty$, *then there exists a subsequence* X^{i_n} *such that* $\nabla f(X^{i_n}) \to 0$ *a. s. for large* n. *If in addition* f *is coercive* X^{i_n} *converges a. s. to a stationary point.*

Proposition 7.2. *Let* f_i *be as in Proposition 1 and* $(X^k, I_k)_{k\geq 0}$ *a SGD, then whenever* $\tau_k \leq 1/\overline{L}$ *for all* $k \geq 0$ *we have*

$$\min_{k\in\{0,\dots,n-1\}} \mathrm{E}\,|\nabla f(X^k)|^2 \leq \frac{2(f(x^0) - \inf f) + \overline{L}\sigma^2 \sum_{k=0}^{n-1} \tau_k^2}{\sum_{k=0}^{n-1} \tau_k}.$$

An easy corollary to Proposition 2 is the remark that leads to Eq. (5).

* * *

Example 7.1. Consider $\{(z_1, y_1), \dots, (z_m, y_m)\}$ where for every $i = 1, \dots m$, $y_i \in \{1, -1\}$, $z_i \in \mathbb{R}^N$ and $f_i(x) := (\alpha(x \cdot z_i) - y_i)^2/2$, with α a sigmoidal activation function: $\alpha(x) := 1/(1 + e^{-x})$. Then $\nabla f_i(x) = (\alpha(x \cdot z_i) - y_i)\alpha'(x \cdot z_i)z_i$ and the hessian matrix is given by $(\mathrm{H} f_i(x))_{jk} = ((\alpha'(x \cdot z_i))^2 + (\alpha(xz_i) - y_i)\alpha''(x \cdot z_i))z_i^j z_i^k$. Given a matrix $A = u \otimes v$ we have $\|A\| = \sup_{x\neq 0} |Ax|/|x|$, now $|Ax|^2/|x|^2 = v^i u^j x^j v^i u^k x^k / x^k x^k = |v|^2 (u \cdot (x/|x|))^2$, but the scalar product has its maximum value when $x = \pm cu$, with $c \in \mathbb{R}$, hence $\|A\| = |v||u|$. This means that $\|\mathrm{H} f_i(x)\| = |((\alpha'(x \cdot z_i))^2 + (\alpha(xz_i) - y_i)\alpha''(x \cdot z_i))||z_i|^2 < (1/16 + 1/(3\sqrt{3}))|z_i|^2$. Then the gradient of f_i are L_i-Lipshitz with $L_i := (1/16 + 1/(3\sqrt{3}))|z_i|^2$. Moreover $\inf f = 0$ and

$$\frac{1}{m}\sum_{i=1}^m |\nabla f_i(x) - \nabla f(x)|^2 < \frac{1}{4m}\sum_{i=0}^m |z_i|^2 + \frac{3}{4}\left(\frac{1}{m}\sum_{i=1}^m |z_i|\right)^2, \quad \forall x \in \mathbb{R}^N.$$

This example show that the hypothesis used to prove the convergence to $\nabla f \to 0$ of SGD are not too restrictive to be applied to real-interest problems. In this example however the function f is not coercive and then we could not expect in general convergence to a stationary point; this inconvenient however can be readily fixed using a quadratic term (weight-decay regularization) to the function as it is shown in the next example.

Example 7.2. We just modify the previous example by letting $f_i(x) = (\alpha(x \cdot z_i) - y_i)^2/2 + |x|^2/2$. Clearly this additional term does not change Lipshitz condition since it simply add to the hessian the identity matrix so that in this case the gradient is Lipshitz with constant $L_i := 1 + (1/16 + 1/(3\sqrt{3}))|z_i|^2$. This addition on the other hand has no effect on the variance of the gradients since it does not change the value of $\nabla f_i(x) - \nabla f(x)$ at all. This time however the function f is also coercive and therefore we expect convergence to a stationary point.

8. Hamiltonian Learning

References

polyak-optimization

[1] Boris T. Polyak, *Introduction to optimization.* Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

ambrosio-gradient

[2] Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures.* Springer Science & Business Media, 2008.

de-giorigi-maximal-slope [3] De Giorgi, Ennio, Antonio Marino, and Mario Tosques, *Problems of evolution in metric spaces and maximal decreasing curve.* Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.(8) 68.3 (1980): 180-187.

rockafellar-convex [4] Rockafellar, R. Tyrrell, *Convex analysis.* Vol. 28. Princeton university press, 1997.