Topics in Math of ML, Lecture 8
Optimization with gradient methods

Lénaïc Chizat*

November 3, 2025

1 Introduction

- In empirical risk minimization, we choose a predictor by minimizing the empirical risk over a parameterized set of predictors, potentially with regularization. For a parameterization $\{f_\theta\}_{\theta\in\mathbb{R}^p}$ and a regularizer $\Omega : \mathbb{R}^p \to \mathbb{R}$ (e.g. $\Omega(\theta) = \|\theta\|_2^2$ or $\Omega(\theta) = \|\theta\|_1$), this requires to minimize
$$F(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \Omega(\theta).$$
In optimization, the function $F : \mathbb{R}^p \to \mathbb{R}$ is called the *objective function*.

- In general, the minimizer has no closed form. Even when it has one (e.g. linear predictor and square loss), it could be expensive to compute for large problems (in practice n and p can be of the order of millions or more). We thus resort to iterative algorithms.

- Solving optimization problems to high accuracy is computationally expensive. Which accuracy is satisfying in machine learning? It is sufficient to reach an optimization accuracy of the order of the estimation error (usually of the order $O(1/\sqrt{n})$ or $O(1/n)$, see Lectures 2 and 3).

- In this context, gradient-based optimization methods, called *first order methods* are the most suitable. Today, we present such algorithms and analyze their performance on convex functions[1]. To go further on this topic, one may refer to Nesterov [2018], Bubeck [2015].

2 First order methods

Suppose we want to solve, for a function $F : \mathbb{R}^p \to \mathbb{R}$, the optimization problem
$$\min_{\theta\in\mathbb{R}^p} F(\theta).$$
The basic first-order methods is Gradient Descent (GD) in Alg. 1.

At each iteration, GD requires to compute a "full" gradient $\nabla F(\theta_t)$ which could be costly (typically a time complexity in $O(np)$). An alternative is to instead only compute unbiased stochastic estimations of the gradient g_t, i.e. such that $\mathbf{E}[g_t|\theta_t] = \nabla F(\theta_t)$, which can be much faster to compute. This leads to Stochastic gradient descent (SDG) in Alg. 2.

*EPFL lenaic.chizat@epfl.ch

[1] Note that F above is only guaranteed to be convex if $\theta \mapsto f_\theta$ is linear (provided the usual assumption that $\hat{y} \mapsto \ell(y, \hat{y})$ is convex), which is the case for kernel methods but not for neural networks.

Algorithm 1: Gradient descent (GD)

Choose step-size sequence $(\eta_t)_{t\geq 0}$, pick $\theta_0 \in \mathbb{R}^p$ and for $t = 0, 1, \ldots$, let

$$\theta_{t+1} = \theta_t - \eta_t \nabla F(\theta_t).$$

Algorithm 2: Stochastic gradient descent (SDG)

Choose step-size sequence $(\eta_t)_{t\geq 0}$, pick $\theta_0 \in \mathbb{R}^p$ and for $t \geq 0$, let

$$\theta_{t+1} = \theta_t - \eta_t g_t.$$

SGD in machine learning. There are two ways to use SGD for supervised machine learning:

- (empirical risk minimization) If $F(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_\theta(x_i)) + \Omega(\theta)$ then at iteration t we can choose uniformly at random $i_t \in \{1, \ldots, n\}$ and define $g_t = \nabla_\theta[\ell(y_{i_t}, f_{\theta_t}(x_{i_t}))] + \nabla\Omega(\theta)$. There exists "mini-batch" variants where at each iteration, the gradient is averaged over a random subset of the indices.
- (population risk minimization) If $F(\theta) = \mathbf{E}[\ell(Y, f_\theta(X))]$ then at iteration t we can take a fresh sample (x_t, y_t) and define $g_t = \nabla_\theta[\ell(y_t, f_{\theta_t}(x_t))]$. Here, we *directly minimize the (generalization) risk*. The counterpart is that if we only have n samples, then we can only run n SGD iterations. In this setting, sometimes referred to as *online SGD*, any optimization guarantee (e.g. an error of order $O(1/\sqrt{t})$) directly translates into genereralization guarantee in terms of number of samples (e.g. $O(1/\sqrt{n})$).

Time complexity The time complexity of an algorithm is

$$\underbrace{(\text{time to compute one iteration})}_{\text{smaller for SGD than GD}} \times \underbrace{(\text{nb of iterations})}_{\text{studied next}}.$$

Link with gradient flow Writing the update as $(\theta_{t+1} - \theta_t)/\eta = -\nabla g_t(\theta_t)$, we have formally, as $\eta \to 0$ and setting the pseudo-time $\tilde{t} = t \cdot \eta$ that the optimization dynamics converges to a solution of the gradient flow ordinary differential equation (ODE):

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{t}}\theta(\tilde{t}) = -\nabla F(\theta(\tilde{t})), \quad \theta(0) = \theta_0.$$

Today, we directly tackle the true GD and SGD iterates but in the next lectures, we will study this continuous dynamics, because it often leads to shorter and more elegant arguments (which can often be converted into statements about GD or SGD, by "discretizing the proof"). Note that for SGD, there are more precise continuous models that involve a *stochastic* differential equation, see https://francisbach.com/rethinking-sgd-noise/.

To go further You can play with the interactive graphs in this article https://distill.pub/2017/momentum/ (paragraph "First Steps: Gradient Descent") Goh [2017]. For an introduction on the analysis of SGD for quadratic functions, see https://francisbach.com/the-sum-of-a-geometric-series-is-all-you-need/.

3 GD on smooth functions

Following the optimization convention, from now on we rename the optimization variable from θ to x and the objective function from F to f. Let us discuss first some assumptions on the objective functions.

Definition 3.1 (Smoothness). *A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is said β-smooth iff*

$$|f(y) - f(x) - \nabla f(x)^\top (y - x)| \leq \frac{\beta}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathbb{R}^p$$

This is equivalent to f having a β-Lipschitz gradient, i.e. $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta\|x - y\|_2$, $\forall x, y \in \mathbb{R}^p$. For twice differentiable functions, this is equivalent to $-\beta \mathrm{Id} \preceq \nabla^2 f \preceq \beta \mathrm{Id}$ (see Nesterov [2018]).

Lemma 3.2 (Descent lemma). *Assume f is β-smooth and let $x_{t+1} = x_t - \eta \nabla f(x_t)$ for some $\eta \leq 1/\beta$. Then*

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2.$$

Proof. From the smoothness assumption,

$$\begin{aligned} f(x_t - \eta \nabla f(x_t)) &\leq f(x_t) + \nabla f(x_t)^\top (-\eta \nabla f(x_t)) + \frac{\beta}{2}\|\eta \nabla f(x_t)\|_2^2 \\ &= f(x_t) - \eta\Big(1 - \frac{\beta\eta}{2}\Big)\|\nabla f(x_t)\|_2^2 \\ &\leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2. \end{aligned}$$

□

- This shows that if GD converges then the limit must be a stationary point. Moreover, if f is smooth and lower-bounded, the inequality $\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 \leq \frac{2}{T\eta}(f(x_0) - \inf f)$ shows that GD must visit some "quasi-stationnary points". These guarantees are quite vague and this is the most we can say in general.
- This lemma suggests that the natural identification between the pseudo-time of the ODE and the iteration count of GD is $\tilde{t} \sim t/\beta$ (because the step-size of GD has to be of order $1/\beta$).

3.1 Reminders on convex functions

[if needed; skipped in class]

Definition 3.3 (Convex function). *A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is said* convex *iff*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \qquad \forall x, y \in \mathbb{R}^p. \tag{1}$$

If f is twice-differentiable, this is equivalent to requiring $\nabla^2 f(x) \succeq 0$, $\forall x \in \mathbb{R}^p$ (here $\succeq$ denotes the semidefinite partial ordering – also called Loewner order – characterized by $A \succeq B \Leftrightarrow A - B$ is positive semidefinite). A more general definition of convexity is that $\forall x, y \in \mathbb{R}^p$ and $\alpha \in [0, 1]$,

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y).$$

Exercise: show that if f is differentiable, this is equivalent to our definition. The following inequality appears frequently in the proofs involving convexity.

Proposition 3.4 (Jensen's inequality)**.** *If $f : \mathbb{R}^p \to \mathbb{R}$ is convex and μ is a probability measure on $\mathbb{R}^p$, then*

$$f\Big(\int x d\mu(x)\Big) \leq \int f(x) d\mu(x).$$

In words: "the image of the average is smaller than the average of the images".

Proof. Let $x^* = \int x d\mu(x)$. By the definition of convexity we have $f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*)$ $\forall x \in \mathbb{R}^p$. Jensen's inequality follows by integrating, and remarking that $\int \nabla f(x^*)^\top (x - x^*) d\mu(x) = 0$. □

The class of convex functions satisfies the following stability properties (exercise):

- If $(f_j)_{j\in[m]}$ are convex and $(\alpha_j)_{j\in[m]}$ are nonnegative, then $\sum_{j=1}^m \alpha_j f_j$ is convex.
- If $f : \mathbb{R}^p \to \mathbb{R}$ is convex and $A : \mathbb{R}^{p'} \to \mathbb{R}^p$ is linear then $f \circ A : \mathbb{R}^{p'} \to \mathbb{R}$ is convex.

Example. Problems of the form Eq. (1) are convex if the loss ℓ is convex in the second variable, $f_\theta(x)$ is linear in θ, and Ω is convex.

4 GD on smooth and strongly convex functions

Example of application: logistic regression with squared ℓ_2-norm regularization.

Definition 4.1 (Strong convexity)**.** *A differentiable function f is said α-strongly convex, with $\alpha > 0$ (or just convex if $\alpha = 0$), iff*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathbb{R}^p$$

For twice differentiable functions, this is equivalent to $\nabla^2 f \succeq \alpha \mathrm{Id}$ (see Nesterov [2018]).

Proposition 4.2. *Assume that $f : \mathbb{R}^p \to \mathbb{R}$ is convex and differentiable. Then $x^* \in \mathbb{R}^p$ is a global minimizer of f iff*

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

If moreover $\alpha > 0$ then f admits a unique minimizer.

Strong convexity guarantees that the gradient is large when a point is far from optimality:

Lemma 4.3 (PL inequality)**.** *If f is differentiable and α-strongly convex with minimizer x^*, then it holds*

$$\|\nabla f(x)\|_2^2 \geq 2\alpha(f(x) - f(x^*)), \quad \forall x \in \mathbb{R}^p. \tag{2}$$

Proof. The right-hand side in Definition 4.1 is strongly convex in y and minimized with $\tilde{y} = x - \frac{1}{\alpha}\nabla f(x)$. Plugging this value into the bound and taking $y = x^*$ in the left-hand side we get

$$f(x^*) \geq f(x) - \frac{1}{\alpha}\|\nabla f(x)\|_2^2 + \frac{1}{2\alpha}\|\nabla f(x)\|_2^2 = f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2.$$

The conclusion follows by rearranging. □

- This inequality, known as Polyak-Łojasiewicz (PL) or as sharpness, is often the only consequence of strong convexity that is needed in the proofs. It is thus often directly assumed (it sometimes holds in certain non-convex settings, at least locally).

- The interest of working with convexity (rather than PL) is that the class of convex functions is more stable (than those satisfying PL), and offer more tools (such as duality tools).

In the next theorem, we show that GD converges exponentially[2] for smooth and strongly convex (or just smooth and PL) functions.

Theorem 4.4. *Assume that f is β-smooth and α-strongly convex. Choosing $\eta_t = 1/\beta$, the iterates $(x_t)_{t\geq 0}$ of GD on f satisfy*

$$f(x_t) - f(x^*) \leq (1-\beta/\alpha)^t (f(x_0) - f(x^*)) \leq \exp(-t\beta/\alpha)(f(x_0) - f(x^*)).$$

Proof. By the descent lemma, we have

$$f(x_{t+1}) - f(x^*) \leq (f(x_t) - f(x^*)) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2.$$

Using the PL inequality, it follows

$$f(x_{t+1}) - f(x^*) \leq (1-\alpha/\beta)(f(x_t) - f(x^*)) \leq \exp(-\alpha/\beta)(f(x_t) - f(x^*)).$$

We conclude by a recursion. □

- We necessarily have $\alpha \leq \beta$. The ratio $\kappa := \beta/\alpha$ is called the *condition number*. The smooth and strongly convex setting is a direct extension of the quadratic case.
- If we only assume that the function is smooth and convex (not strongly convex), then GD with constant step-size $\eta \leq 1/\beta$ also converges when a minimizer exists, but at a slower rate in $O(1/t)$.
- For this class of functions (convex and smooth), there exists 1st-order methods which achieve a $O(1/t^2)$ rate, showing that gradient descent is not optimal. However, these improved algorithms have also drawbacks (lack of adaptivity, instability to noise...)

5 GD for convex and Lipschitz functions

Example of application: logistic regression with ℓ_1-norm regularization.

Theorem 5.1. *Assume that f is convex, L-Lipschitz and admits a minimizer x^* that satisfies $\|x^* - x_1\|_2 \leq R$. By chosing $\eta_t = \frac{R}{L\sqrt{t}}$ then the iterates $(x_t)_{t\geq 1}$ of GD on f satisfy*

$$\min_{1\leq s\leq t} f(x_s) - f(x^*) \leq RL\frac{2+\log(t)}{4(\sqrt{t+1}-1)} = O\Big(\frac{\log(t)}{\sqrt{t}}\Big).$$

Proof. We look at how x_t approaches x^*. It holds

$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta_t \nabla f(x_t) - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta_t \nabla f(x_t)^\top (x_t - x^*) + \eta_t^2 \|\nabla f(x_t)\|^2.$$

Combining this with the convexity inequality $f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$, it follows

$$\eta_t (f(x_t) - f(x^*)) \leq \frac{1}{2}\Big(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\Big) + \frac{1}{2}\eta_t^2 \|\nabla f(x_t)\|^2. \tag{3}$$

[2] It is also sometimes called geometric convergence, or linear convergence (because it is linear in a "semilogy" plot).

It is sufficient to sum these inequalities to get, for any $x^* \in \mathbb{R}^p$,

$$\frac{1}{\sum_{s=1}^t \eta_s} \sum_{s=1}^t \eta_s \left(f(x_s) - f(x^*)\right) \leq \frac{\|x_1 - x^*\|_2^2}{2\sum_{s=1}^t \eta_s} + L^2 \frac{\sum_{s=1}^t \eta_s^2}{2\sum_{s=1}^t \eta_s}.$$

The left-hand side is larger than $\min_{1\leq s\leq t}(f(x_s) - f(x^*))$ (trivially) and than $f(\bar{x}_t) - f(x^*)$ where $\bar{x}_t = (\sum_{s=1}^t \eta_s x_s)/(\sum_{s=1}^t \eta_s)$ by Jensen's inequality.

The upper bound goes to 0 if $\sum_{s=1}^t \eta_s$ goes to ∞ (to forget the initial condition, the "bias") and $\eta_t \to 0$ (to decrease the "variance" term). Let us choose $\eta_s = \tau/\sqrt{s}$ for some $\tau > 0$. By using the series-integral comparisons below, we get the bound

$$\min_{1\leq s\leq t}(f(x_s) - f(x^*)) \leq \frac{1}{4(\sqrt{t+1}-1)}\Big(R^2/\tau + \tau L^2(1+\log(t))\Big).$$

We choose $\tau = R/L$ (which is suggested by optimizing the previous bound when $\log(t) = 0$) which leads to the result.

In the proof, we used the following series-integral comparisons for decreasing functions:

$$\sum_{s=1}^t \frac{1}{\sqrt{s}} \geq \int_{s=1}^{t+1} \frac{\mathrm{d}t}{\sqrt{t}} = [2\sqrt{s}]_1^{t+1} = 2(\sqrt{t+1}-1)$$

and

$$\sum_{s=1}^t \frac{1}{s} \leq 1 + \sum_{s=2}^t \frac{1}{s} \leq 1 + \int_1^t \frac{ds}{s} = 1 + \log(t).$$

□

- The previous proof scheme is very flexible. It can be extended to:
  - constrained minimization over a convex set (we then insert a projection step at each iteration);
  - non-differentiable convex and Lipschitz objective functions (using sub-gradients, i.e. any vector satisfying Eq. (1) in place of $\nabla f(x_t)$);
  - non-euclidean geometry (for instance multiplicative instead of additive updates);
  - stochastic gradients, as seen below.

6 SGD on convex and Lipschitz functions

Under the same assumptions on the objective, we now study SGD. We assume the following:

- (H1) unbiased gradient: $\mathbf{E}[g_t(x)|x] = \nabla f(x),\ \forall t, x$
- (H2) bounded gradient: $\mathbf{E}[\|g_t(x)\|_2^2|x] \leq B^2,\ \forall t$

Theorem 6.1. *Assume that f is convex, L-Lipschitz and admits a minimizer x^* that satisfies $\|x^* - x_1\|_2 \leq R$. Assume that the stochastic gradient $g_t(x)$ satisfies (H1-2). Then, choosing $\eta_t = R/(B\sqrt{t})$, the iterates $(x_t)_{t\geq 1}$ of SGD on f satisfy*

$$\mathbf{E}\Big[f(\bar{x}_s) - f(x^*)\Big] \leq RB\frac{2+\log(t)}{4(\sqrt{t+1}-1)}.$$

where $\bar{x}_s = (\sum_{s=1}^t \eta_s x_s)/(\sum_{s=1}^t x_s)$.

Proof. We follow essentially the same proof as in the deterministic case. Adding expectations where needed:

$$\begin{aligned}\mathbf{E}\Big[\|x_{t+1}-x^*\|_2^2\Big] &= \mathbf{E}\Big[\|x_t-\eta_t g_t(x_t)-x^*\|_2^2\Big] \\ &= \mathbf{E}\Big[\|x_t-x^*\|_2^2\Big] - 2\eta_t\mathbf{E}\Big[g_t(x_t)^\top(x_t-x^*)\Big] + \eta_t^2\mathbf{E}\Big[\|g_t(x_t)\|_2^2\Big]\end{aligned}$$

We can compute the expectation of the middle term as

$$\begin{aligned}\mathbf{E}\Big[g_t(x_t)^\top(x_t-x^*)\Big] &= \mathbf{E}\Big[\mathbf{E}\big[g_t(x_t)^\top(x_t-x^*)\big|x_t\big]\Big] \\ &= \mathbf{E}\Big[\mathbf{E}\big[g_t(x_t)\big|x_t\big]^\top(x_t-x^*)\Big] \\ &= \mathbf{E}\Big[\nabla f(x_t)^\top(x_t-x^*)\Big].\end{aligned}$$

This leads to

$$\mathbf{E}\Big[\|x_{t+1}-x^*\|_2^2\Big] \leq \mathbf{E}\Big[\|x_t-x^*\|_2^2\Big] - 2\eta_t\mathbf{E}\Big[\nabla f(x_t)^\top(x_t-x^*)\Big] + \eta_t^2 B^2.$$

and thus, combining with the convexity inequality $f(x_t)-f(x^*)\leq\nabla f(x_t)^\top(x_t-x^*)$ it follows

$$\eta_t\mathbf{E}[f(x_t)-f(x^*)] \leq \frac{1}{2}\Big(\mathbf{E}\|x_t-x^*\|^2 - \mathbf{E}\|x_{t+1}-x^*\|^2\Big) + \frac{1}{2}\eta_t^2 B^2. \tag{4}$$

Except for the expectations, this is the same bound that Eq. (3) so we can conclude as in the proof of Theorem 5.1, *mutatis mutandis*. We state our bound in terms of the average iterates because the cost of finding the best iterate could be high in comparison to that of evaluating a stochastic gradient. □

- Averaging of iterates is often done only after a certain number of iterations; in order to forget the initial condition faster (but this should not be done on non-convex problems!).
- When applied to single pass SGD on the population risk, we get that after $t=n$ iterations the excess risk is of the order $O(1/\sqrt{n})$. This leads to a time complexity of $O(tp)=O(np)$ (if ieach iteration has a complexity p), while using GD to get the same accuracy costs $O(tnp)=O(n^2p)$.
- As opposed to the deterministic case, the use of smoothness does not lead to significantly better results.

References

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Gabriel Goh. Why momentum really works. *Distill*, 2(4):e6, 2017.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.