Machine learning - Master ICFP 2019-2020

Optimization with gradient methods

Lénaïc Chizat

February 14, 2020

In this lecture, we present optimization algorithms based on gradient descent and analyze their performance on convex functions. References [1, 2].

1 Optimization in machine learning

- In supervised machine learning, we are given n i.i.d. samples $(x_i, y_i)_{i=1}^n$ of a couple of random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$ and the goal is to find a predictor $f : \mathcal{X} \to \mathcal{Y}$ with a small risk

$$\mathcal{R}(f) := \mathbb{E}[\ell(y, f(x))]$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function.

- In the empirical risk minimization approach, we choose the predictor by minimizing the empirical risk over a parameterized set of predictors, potentially with regularization. For a parameterization $\{f_w\}_{w \in \mathbb{R}^p}$ and a regularizer $\Omega : \mathbb{R}^p \to \mathbb{R}$ (e.g. $\Omega(w) = \|w\|_2^2$ or $\Omega(w) = \|w\|_1$), this requires to minimize

$$F(w) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_w(x_i)) + \Omega(w).$$

In optimization, the function $F : \mathbb{R}^p \to \mathbb{R}$ is called the *objective*.

- 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. We thus resort to iterative algorithms.

- Solving optimization problems to high accuracy is computationally expensive. Which accuracy is satisfying in machine learning? If the algorithm returns $\hat{w}$ and $w^* \in \arg\min_w \mathcal{R}(f_w)$, we have the risk decomposition

$$\mathcal{R}(f_{\hat{w}}) - \inf_{w \in \mathbb{R}^p} \mathcal{R}(f_w) = \underbrace{\left\{\mathcal{R}(f_{\hat{w}}) - \hat{\mathcal{R}}(f_{\hat{w}})\right\}}_{\leq \text{ Estimation error}} + \underbrace{\left\{\hat{\mathcal{R}}(f_{\hat{w}}) - \hat{\mathcal{R}}(f_{w^*})\right\}}_{\leq \text{ Optimization error}} + \underbrace{\left\{\mathcal{R}(f_{w^*}) - \hat{\mathcal{R}}(f_{w^*})\right\}}_{\leq \text{ Estimation error}}.$$

It is thus 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).

2 First order optimization algorithms

Suppose we want to solve, for a function $F : \mathbb{R}^p \to \mathbb{R}$, the optimization problem

$$\min_{w \in \mathbb{R}^p} F(w).$$

In today's class, we analyze the following two algorithms, which are often the methods of choice in machine learning.

Algorithm 1 (Gradient descent (GD)) *Choose step-size sequence $(\eta_t)_{t \geq 0}$, pick $w_0 \in \mathbb{R}^p$ and for $t \geq 0$, let*

$$w_{t+1} = w_t - \eta_t \nabla F(w_t).$$

At each iteration, this algorithm requires to compute a "full" gradient $\nabla F(w_t)$ which could be costly. An alternative is to instead only compute unbiased stochastic estimations of the gradient $g_t(w_t)$, i.e. such that $\mathbb{E}[g_t(w_t)|w_t] = \nabla F(w_t)$, which could be much faster to compute. This leads to the following algorithm.

Algorithm 2 (Stochastic gradient descent (SDG)) *Choose step-size sequence $(\eta_t)_{t \geq 0}$, pick $w_0 \in \mathbb{R}^p$ and for $t \geq 0$, let*

$$w_{t+1} = w_t - \eta_t g_t(w_t).$$

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

- (empirical risk minimization) If $F(w) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_w(x_i))$ then at iteration t we can choose uniformly at random $i_t \in \{1, \dots, n\}$ and define $g_t(w) = \nabla_w[\ell(y_{i_t}, f_w(x_{i_t}))]$. There exists "mini-batch" variants where at each iteration, the gradient is averaged over a random subset of the indices.
- (population risk minimization) If $F(w) = \mathbb{E}[\ell(Y, f_w(X))]$ then at iteration t we can take a fresh sample (x_t, y_t) and define $g_t(w) = \nabla_w[\ell(y_t, f_w(x_t))]$. 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.

3 Analysis of GD for Ordinary Least Squares

We start with a case where the analysis is explicit: ordinary least squares. Let $X \in \mathbb{R}^{n \times d}$ be the design matrix, assumed injective, and $y \in \mathbb{R}^n$ the observations. The least squares estimator w^* minimizes

$$\frac{1}{2n} \|Xw - y\|_2^2.$$

Using results from Lecture 2, the excess risk is, using $\Sigma = \frac{1}{n} X^\top X$:

$$F(w) = \frac{1}{2}(w - w^*)\Sigma(w - w^*).$$

Decrease of objective. Then gradient descent iterates with fixed step-size $\eta_t = \eta$ are:

$$w_{t+1} = w_t - \eta \nabla F(w_t) = w_t - \eta \Sigma (w_k - w^*).$$

We diagonalize $\Sigma = PDP^\top$ with $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, we define $v_k = P^\top (w_k - w^*)$ which evolves as

$$v_{k+1} = (\mathrm{Id} - \eta D) v_k \quad \Rightarrow v_k[j] = (1 - \eta \lambda_j)^t v_0[j].$$

In terms of excess risk, we have

$$F(w_k) = v_k^\top D v_k = \sum_{j=1}^d \lambda_j |1 - \eta \lambda_j|^{2t} v_0[j]^2 \leq \Big(\max_j |1 - \eta \lambda_j| \Big)^{2t} F(w_0).$$

Choice of step-size. If we want the fastest asymptotic rate, we need to choose η that minimizes the contraction ratio. Writing $\alpha = \min\{\lambda_j\}$ and $\beta = \max\{\lambda_j\}$ and the *condition number* $\kappa = \beta/\alpha$, we obtain

$$\min_\eta \max_j |1 - \eta \lambda_j| = \min_\eta \max\{\eta\beta - 1, 1 - \eta\alpha\} = \frac{\beta - \alpha}{\beta + \alpha} = \frac{\kappa - 1}{\kappa + 1}$$

with the minimizer $\eta = 2/(\beta + \alpha)$. In practice, we do not know α, but we can upper bound $\beta = \sup_{\|u\|_2 \leq 1} u^\top \Sigma u$ by $\tilde{\beta} := \frac{1}{n} \sum_{i=1}^n \|x_i\|_2^2$, and we still get an exponential convergence in $O(|1 - \alpha/\tilde{\beta}|^{2t})$.

To go further You can play with the interactive graphs in this article `https://distill.pub/2017/momentum/` (paragraph "First Steps: Gradient Descent") [3]. For an introductory analysis of SGD on quadratic functions, see `https://francisbach.com/the-sum-of-a-geometric-series-is-all-you-need/`.

4 Convex functions

We now wish to analyze GD (and later SGD) in a broader setting. We will always assume convexity, although these algorithms are also used (and can sometimes also be analyzed) when this assumption does not hold. In what follows, except for the examples, f denotes the objective and x or y its variables (they do not stand anymore for a predictor or training variables).

Definition 1 (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 1 (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, $f_w(x)$ is linear in w, and Ω is convex.

It is also worth emphasizing on the following property (immediate from the definition).

Proposition 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*

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

5 Analysis of GD for strongly convex and smooth functions

Definition 2 (Strong convexity) *A differentiable function f is said α-strongly convex, with $\alpha > 0$, 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 [1]). This property implies that f admits a unique minimizer x^*, which is characterized by $\nabla f(x^*) = 0$. Moreover, this guarantees that the gradient is large when a point is far from optimality:

Lemma 1 *If f is differentiable and α-strongly convex with minimizer x^*, then it holds*

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

Proof The right-hand side in Definition 2 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. ■

Definition 3 (Smoothness) *A differentiable function f is said β-smooth iff*

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

This is equivalent to f having a β-Lipschitz gradient, i.e. $\|\nabla f(x) - \nabla f(y)\|_2^2 \leq \|x - y\|_2^2$, $\forall x, y \in \mathbb{R}^p$. For twice differentiable functions, this is equivalent to $-\beta \mathrm{Id} \preceq \nabla^2 f \preceq \beta \mathrm{Id}$ (see [1]).

In the next theorem, we show that gradient descent converges exponentially[1] for such problems.

Theorem 1 *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*

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

Proof By smoothness, we have the following descent property, with $\eta_t = 1/\beta$,

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

Rearranging, we get

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

Using Lemma 1, 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*.
- If we only assume that the function is smooth and convex (not strongly convex), then GD with constant step-size $\eta = 1/\beta$ also converges when a minimizer exists, but at a slower rate in $O(1/t)$.
- Choosing the step-size only requires an upper bound β on the smoothness constant (in case it is over-estimated, the convergence rate only degrades slightly).

Example: regularized logistic regression Consider a classification task with $y \in \{-1, +1\}$, the logistic loss $\ell(y, z) = \log(1 + e^{-yz})$, a linear model $f_w(x) = x^\top w$ and regularization $\lambda\|w\|_2^2$. The objective of empirical risk minimization is

$$F(w) = \frac{1}{n}\sum_{i=1}^n \log(1 + e^{-y_i x_i^\top w}) + \lambda\|w\|_2^2.$$

This function is convex and differentiable. It is at least 2λ-strongly convex thanks to the regularization term. Its gradient is

$$\nabla F(w) = \frac{1}{n}\sum_{i=1}^n \frac{-y_i x_i}{1 + e^{y_i x_i^\top w}} + 2\lambda w$$

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

and its Hessian $\nabla^2 F(w) = (\partial_{ij} F(w))_{i,j=1}^d$ is

$$\nabla^2 F(w) = \frac{1}{n} x_i x_i^\top \frac{e^{y_i x_i^\top w}}{(1 + e^{y_i x_i^\top w})^2} + 2\lambda.$$

Thus F is β-smooth with $\beta = (1/n) \sum_{i=1}^n \|x_i\|_2^2 + 2\lambda$. The condition number, which determines the convergence speed, is thus $\kappa = \beta/\alpha = 1 + (1/(2\lambda n)) \sum_{i=1}^n \|x_i\|_2^2$. The regularization, originally introduced to reduce the estimation error, turns out to also help optimization.

6 Analysis of gradient methods on non-smooth problems

We now relax our assumptions and only require Lipschitz continuity, in addition to convexity.

Definition 4 (Lipschitz function) *A function $f : \mathbb{R}^p \to \mathbb{R}$ is said L-Lipschitz continuous iff*

$$|f(y) - f(x)| \leq L\|y - x\|_2^2, \qquad \forall x, y \in \mathbb{R}^p.$$

Exercise: show that if f is differentiable, this is equivalent to the assumption $\|\nabla f(x)\|_2 \leq L$, $\forall x \in \mathbb{R}^p$. Without additional assumptions, this setting is usually referred to as *non-smooth* optimization.

6.1 Convergence rate of GD

Theorem 2 *Assume that f is convex, L-Lipschitz and admits a minimizer x^* that satisfies $\|x^* - x_0\|_2 \leq R$. By chosing $\eta_t = \frac{R}{L\sqrt{t+1}}$ then the iterates $(x_t)_{t \geq 0}$ of GD on f satisfy*

$$\min_{0 \leq s \leq t-1} f(x_s) - f(x^*) \leq RL \frac{2 + \log(t)}{4(\sqrt{t+1} - 1)}.$$

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

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

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

$$\frac{1}{\sum_{s=0}^{t-1} \eta_s} \sum_{s=0}^{s-1} \eta_s \left(f(x_s) - f(x^*)\right) \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{s=0}^{t-1} \eta_s} + L^2 \frac{\sum_{s=0}^{t-1} \eta_s^2}{2\sum_{s=0}^{t-1} \eta_s}.$$

The left-hand side is larger than $\min_{0 \leq s \leq t-1}(f(x_s) - f(x^*))$ (trivially) and than $f(\bar{x}_t) - f(x^*)$ where $\bar{x}_t = (\sum_{s=0}^{t-1} \eta_s x_s)/(\sum_{s=0}^{t-1} \eta_s)$ by Jensen's inequality.

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

$$\min_{0 \leq s \leq t-1} (f(x_s) - f(x^*)) \leq \frac{1}{4(\sqrt{t+1}-1)} \Big(R^2/\tau + \tau L^2 (1 + \log(t)) \Big).$$

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:

$$\sum_{s=0}^{t-1} \frac{1}{\sqrt{s+1}} \geq \int_0^t \frac{ds}{\sqrt{s+1}} = \Big[2\sqrt{s+1} \Big]_0^t = 2\sqrt{t+1} - 2$$

and

$$\sum_{s=0}^{t-1} \frac{1}{s+1} \leq 1 + \sum_{s=2}^{t} \frac{1}{s} \leq 1 + \int_0^t \frac{ds}{s} = 1 + \log(t).$$

- 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.

Example: logistic regression with ℓ_1-regularization. Consider the previous example but with ℓ_1 regularization, giving the objective

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i x_i^\top w}) + \lambda \|w\|_1.$$

This function convex[2]. We have $F(0) = \log(2)/n$ so any minimizer w^* (which exists by coercivity) must satisfy $\|w^*\|_1 \leq \log(2)/(n\lambda)$. Since $\|\cdot\|_2 \leq \|\cdot\|_1$, it follows that $\|w^*\|_2 \leq \log(2)/(n\lambda) =: R$. Using our previous computations, we also have

$$\|\nabla F(w)\|_2 = \Big\| \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i x_i}{1 + e^{y_i x_i^\top w}} + \lambda \operatorname{sign}(w) \Big\|_2 \leq \frac{1}{n} \sum_{i=1}^{n} \|x_i\|_2 + \lambda\sqrt{d} =: L.$$

From these bounds we get explicit step-sizes and convergence guarantees.

[2]It is not differentiable, but the theory above could be adapted to deal with this cases

6.2 Convergence rate of SGD

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

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

Theorem 3 *Assume that f is convex, L-Lipschitz and admits a minimizer x^* that satisfies $\|x^* - x_0\|_2 \leq R$. Assume that the stochastic gradient g satisfies (H1-2). Then, choosing $\eta_t = (R/\sqrt{L^2 + \sigma^2})/\sqrt{t+1}$, the iterates $(x_t)_{t\geq 0}$ of SGD on f satisfy*

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

where $\bar{x}_s = (\sum_{s=0}^{t-1} \eta_s x_s)/(\sum_{s=0}^{t-1} x_s)$.

Proof We follow essentially the same proof as in the deterministic case.

$$\begin{aligned}
\mathbb{E}\Big[\|x_{t+1} - x^*\|_2^2\Big] &= \mathbb{E}\Big[\|x_t - \eta_t g_t(x_t) - x^*\|_2^2\Big] \\
&= \mathbb{E}\Big[\|x_t - x^*\|_2^2\Big] - 2\eta_t \mathbb{E}\Big[g_t(x_t)^\top (x_t - x^*)\Big] + \eta_t^2 \mathbb{E}\Big[\|g_t(x_t)\|_2^2\Big] \\
&= \mathbb{E}\Big[\|x_t - x^*\|_2^2\Big] - 2\eta_t \mathbb{E}\Big[\nabla f(x_t)^\top (x_t - x^*)\Big] + \eta_t^2 \Big(\mathbb{E}\Big[\|\nabla f(x_t)\|_2^2\Big] + \mathbb{E}\Big[\|g_t(x_t) - \nabla f(x_t)\|_2^2\Big]\Big) \\
&\leq \mathbb{E}\Big[\|x_t - x^*\|_2^2\Big] - 2\eta_t \mathbb{E}\Big[\nabla f(x_t)^\top (x_t - x^*)\Big] + \eta_t^2 (G^2 + \sigma^2).
\end{aligned}$$

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

$$\eta_t \mathbb{E}[f(x_t) - f(x^*)] \leq \frac{1}{2}\Big(\mathbb{E}\|x_t - x^*\|^2 - \mathbb{E}\|x_{t+1} - x^*\|^2\Big) + \frac{1}{2}\eta_t^2 (G^2 + \sigma^2). \tag{3}$$

Except for the expectations, this is the same bound that Eq. (2) so we can conclude as in the proof of Theorem 2, *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. ■

References

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

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

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