Exercises: gradient descent

Pierre Ablin

1 Gradient flows

We let $f : \mathbb{R}^p \to \mathbb{R}$ a differentiable function. Starting from $x_0 \in \mathbb{R}^p$, gradient descent with step-size $\eta > 0$ iterates

$$x_{n+1} = x_n - \eta \nabla f(x_n). \tag{1}$$

The behavior of such algorithm is more easily understood by looking at the gradient *flow*, which is the Ordinary Differential Equation (ODE), starting from $x(0) = x_0$:

$$\dot{x}(t) = -\nabla f(x(t)). \tag{2}$$

Indeed, Eq (1) is an Euler discretization of the gradient flow equation with step η, and as such we have $x_n \simeq x(\eta n)$.

1.1

We define $\phi(t) = f(x(t))$. Show that we have

$$\phi'(t) = -\|\nabla f(x(t))\|^2$$

1.2

We assume that f is bounded from below by f^*. Demonstrate that the function $t \to \|\nabla f(x(t))\|^2$ is integrable, and that

$$\inf_{t \leq T} \|\nabla f(x(t))\|^2 \leq \frac{f(x_0) - f^*}{T}.$$

1.3

Assume that f satisfies the Polyak-Lojasciewicz inequality for some $\mu > 0$:

$$\forall w, \;\; f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Demonstrate that $f(x(t))$ converges to f^*, and give the convergence rate.

2 Gradient descent in a simple case

We let $p \geq 0$, and consider a vector $b \in \mathbb{R}^p$ and a matrix $A \in \mathbb{R}^{p \times p}$. We assume that A is a symmetric matrix with positive eigenvalues $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_p = \lambda_{\min} > 0$. We define the following *quadratic* objective function:

$$f(x) = \frac{1}{2} x^\top A x - b^\top x$$

Exercise 1: Show that this function is convex, and that its gradient is given by $\nabla f(x) = Ax - b$. Find the analytical expression of its minimizer x^*, and of $f(x^*)$.

We now consider the sequence of iterates of gradient descent with a step size $\rho > 0$, starting from $x_0 = 0$:

$$\text{For } n \geq 0: \quad x_{n+1} = x_n - \rho \nabla f(x_n)$$

Exercise 2: Obtain a closed form expression for x_n and give a condition on ρ for this sequence to converge to 0.

In the following, we assume that $\rho = \frac{1}{\lambda_{\max}}$.

Exercise 3: Demonstrate that $\|x_n - x^*\| \leq (1 - \frac{\lambda_{\min}}{\lambda_{\max}})^n \|x^*\|$.

This is what we call *linear* convergence, and $1 - \frac{\lambda_{\min}}{\lambda_{\max}}$ is the rate of convergence.

The quantity $\kappa = \frac{\lambda_{\min}}{\lambda_{\max}}$ is called the *conditioning* of the matrix A, and, by extension, of the function f. This number is always between 0 and 1. The closer it is to one, the faster gradient descent converges.

Here, if for instance $\kappa = \frac{1}{2}$, then the convergence is very fast: $\|x_n - x^*\| \leq \frac{1}{2^n}\|x^*\|$, every iteration halves the error. However, in some cases we can have some very poorly conditioned problems.

Exercise 4: Assume that $\kappa = \frac{1}{1000}$, and that $\|x^*\| = 1$. How many iterations of gradient descent are needed to reach an error $\|x_n - x^*\| \leq \frac{1}{10}$? and to get $\|x_n - x^*\| \leq \frac{1}{100}$?

In these badly conditioned case, it would be useful to obtain a bound on the error that does not depend on the conditioning of the problem. To get such a bound, we look at another measure of the error, $f(x_n) - f(x^*)$.

Exercise 5: Show that for all $\mu \in [0, 1]$ and all n we have $(1 - \mu)^{2n}\mu \leq \frac{1}{2n+1}$. Deduce that

$$f(x_n) - f(x^*) \leq \frac{1}{(2n+1)\rho}\|x^*\|^2$$

This is what we call *sub-linear* convergence. Note that this rate of convergence does not get worse when $\lambda_{\min}$ goes to 0: it does not depend on the conditioning of the problem.