TD 4 – Introduction to gradient descent

Mathematics of data

09/10/24

Exercise 1. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Starting from $x_0 \in \mathbb{R}^d$, gradient descent with step-size $\eta > 0$ iterates

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

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

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

Indeed, Equation (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. Define $\phi(t) \coloneqq f(x(t))$. Show that $\phi'(t) = -\left\|\nabla f(x(t))\right\|^2$.
2. We assume that f is bounded from below by $f^* \in \mathbb{R}$. Prove that the function $t \mapsto \left\|\nabla f(x(t))\right\|^2$ is integrable, and that
$$\inf_{t \leq T} \left\|\nabla f(x(t))\right\|^2 \leq \frac{f(x_0) - f^*}{T}.$$
3. Assume that f satisfies the Polyak-Lojasciewicz inequality for some $\mu > 0$:
$$\forall x, \quad f(x) - f^* \leq \frac{1}{2\mu} \left\|\nabla f(x(t))\right\|^2.$$
Prove that $f(x(t))$ converges to f^*, and give the convergence rate.

Exercise 2. Let $d \geq 0$ and consider a vector $b \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d \times d}$. We assume that A is a symmetric matirx with positive eigenvalues $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_d = \lambda_{\min} > 0$. We define the following *quadratic* objective function:

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

1. Show that this function is convex, and compute its gradient. 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 $x_0 = 0$:

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

2. Derive a closed-form expression for x_n and give a condition on ρ for this sequence to converge to x^*.

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

3. Show that $\left\|x_n - x^*\right\| \leq (1 - \frac{\lambda_{\min}}{\lambda_{\max}})^n \left\|x^*\right\|$.

This is what we call *linear* convergence, and $1 - \frac{\lambda_{\min}}{\lambda_{\max}}$ is the rate of convergence. The quantity $\kappa = \frac{\lambda_{\min}}{\lambda_{\max}} \in [0, 1]$ is called the *conditioning* of the matrix A (and of the function f). The closer κ is to 1, the faster gradient descent converges.

4. Assume that $\kappa = \frac{1}{1000}$ and $\|x^*\| = 1$. 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 ill-conditioned cases, it is useful to compute 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, $f(x_n) - f(x^*)$.

5. Show that for all $\mu \in [0, 1]$ and all $n \geq 0$ we have $(1 - \mu)^{2n} \mu \leq \frac{1}{2n+1}$. Deduce that
$$f(x_n) - f(x^*) \leq \frac{1}{2(2n+1)\rho} \left\|x^*\right\|^2.$$

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

Exercise 3. Let $(x^*, f^*) \in \mathbb{R}^d \times \mathbb{R}$ and $H \in \mathbb{R}^{d\times d}$ be a positive definite symmetric matrix. Denote f the associated quadratic form

$$f(x) \coloneqq (x - x^*)^\top H(x - x^*) + f^*.$$

Assume H's eigenvalues λ satisfy $0 < \mu \leq \lambda \leq L$. A general first-order minimization method writes

$$x_{t+1} = x_0 - \sum_{s=0}^{t} \gamma_{t,s} \nabla f(x_s)$$

with $\gamma_{t,t} \neq 0$.

1. Prove that for each of these methods, there exists a sequence of polynomials $(P_t)_{t\in\mathbb{N}}$ with P_t of degree t such that $P_t(0) = 1$ and for all t it holds $x_t - x^* = P_t(H)(x_0 - x^*)$.
2. Prove the reciproqua of question 1.
3. Assuming $\mu \neq 0$, prove that the best method corresponds to a family (P_t) that minimises $\sup_{\lambda\in[\mu,L]} |P_t(\lambda)|$ for all t.
4. Show that finding (P_t) is equivalent to finding a family $(Q_t)_{t\in\mathbb{N}}$ such that $\sup_{x\in[-1,1]} |Q_t(x)| = 1$ and $Q_t(\frac{L+\mu}{L-\mu})$ is maximal.
5. Prove that the t-th Tchebychev polynomial, defined as verifying $\forall\theta$, $T_t(\cos(\theta)) = \cos(t\theta)$, is solution of the problem of question 4.

If you have finished all the exercises, you can move on to the TP4 on github.com/vcastin/teaching