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
$$\inf_{t \le T} \left\|\nabla f(x(t))\right\|^2 \le \frac{f(x_0) - f^*}{T}.$$
3. Assume that f satisfies the Polyak-Lojasciewicz inequality for some $\mu > 0$:
$$\forall x, \quad f(x) - f^* \le \frac{1}{2\mu} \left\|\nabla f(x(t))\right\|^2.$$
Prove that $f(x(t))$ converges to f^*, and give the convergence rate.

Exercise 2. Let $d \ge 0$ 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 $\lambda_{\max} = \lambda_1 \ge \cdots \ge \lambda_d = \lambda_{\min} > 0$. 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 $\|x_n - x^*\| \le (1 - \frac{\lambda_{\min}}{\lambda_{\max}})^n \|x^*\|$.

This is what we call *linear* convergence, and $1 - \frac{\lambda_{\min}}{\lambda_{\max}}$ is the rate of convergence. The quantity $\kappa = \frac{\lambda_{\min}}{\lambda_{\max}} \in [0, 1]$ 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 $\|x_n - x^*\| \le \frac{1}{10}$? And to get $\|x_n - x^*\| \le \frac{1}{100}$?

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 $n \ge 0$ we have $(1 - \mu)^{2n} \mu \le \frac{1}{2n+1}$. Deduce that
$$f(x_n) - f(x^*) \le \frac{1}{2(2n+1)\rho} \left\|x^*\right\|^2.$$

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

Solutions

Exercise 1.

1. Computation with the chain rule.

2. We have $\int_0^T \|\nabla f(x(t))\|^2 \,\mathrm{d}t = -\int_0^T \phi'(t)\mathrm{d}t = \phi(0)-\phi(T) \leq \phi(0)-f^*$. The function $T \mapsto \int_0^T \|\nabla f(x(t))\|^2 \,\mathrm{d}t$ is non decreasing and upper bounded, so it converges when $T \to +\infty$, which shows the integrability. Moreover,
$$\inf_{t\leq T} \|\nabla f(x(t))\|^2 \leq \frac{1}{T}\int_0^T \|\nabla f(x(t))\|^2 \,\mathrm{d}t \leq \frac{f(x_0)-f^*}{T}$$
according to the previous computation.

3. The PL inequality applied to $x(t)$ can be rewritten as $\phi(t) - f^* \leq -\frac{1}{2\mu}\phi'(t)$. Then, applying Grönwall's inequality leads to
$$\phi(t) - f^* \leq e^{-\frac{1}{2\mu}t}(\phi(0)-f^*).$$

Exercise 2.

1. We have $\nabla f(x) = Ax - b$ so the Hessian of f is A, which is a positive definite matrix. Therefore, f is strictly convex. This implies that x^* is characterized by the equation $\nabla f(x^*) = 0$ (if this equation has a solution). We obtain $x^* = A^{-1}b$ and $f(x^*) = -\frac{1}{2}b^\top A^{-1}b$.

2. Using the expression of $\nabla f(x_n)$, we can rewrite x_{n+1} as
$$x_{n+1} = (I-\rho A)x_n + \rho b.$$
Substracting x^*, which satisfies $x^* = x^* - \rho\nabla f(x^*) = (I-\rho A)x^* + \rho b$, to both sides of the equation, we obtain
$$x_{n+1} - x^* = (I-\rho A)(x_n - x^*),$$
so that
$$x_n - x^* = (I-\rho A)^n(x_0 - x^*) = -(I-\rho A)^n x^*$$
as $x_0 = 0$. A sufficient condition for x_n to converge to x^* is then that the module of all the eigenvalues of $I-\rho A$ is strictly smaller than 1. The eigenvalues of $I-\rho A$ are of the form $1-\rho\lambda$ with λ an eigenvalue of A, so our condition is satisfied if and only if
$$1-\rho\lambda_{\max} > -1.$$
We obtain the sufficient condition
$$\rho < \frac{2}{\lambda_{\max}}.$$

3. According to question 2, we have $\|x_n - x^*\| \leq \|I-\rho A\|_2^n \|x^*\|$, where $\|\cdot\|_2$ is the operator norm. Recall that the operator norm of a symmetric matrix is the maximum of the modules of its eigenvalues. The largest eigenvalue (in module) of $I-\rho A$ is $1-\rho\lambda_{\min} = 1-\frac{\lambda_{\min}}{\lambda_{\max}}$, which allows us to conclude.

4. We obtain $n \simeq 2300$ for $1/10$ and $n \simeq 4600$ for $1/100$.

5. Denoting $g(\mu) \coloneqq (1-\mu)^{2n}\mu$, we have $g'(\mu) = (1-\mu)^{2n-1}(1-(2n+1)\mu)$. g reaches its maximum on $[0,1]$ either at 0, at 1, or at a point where $g'(\mu)$ is zero. The only candidates are therefore $0, 1$ and $2n+1$. But $g(0) = g(1) = 0$ so the maximum is reached at $2n+1$, as we have $g(2n+1) = \left(\frac{2n}{2n+1}\right)^{2n}\frac{1}{2n+1} > 0$. We conclude by noticing that $g(2n+1) \leq \frac{1}{2n+1}$. Then, to prove the inequality, we write
$$\begin{aligned} f(x_n) - f^* &= \frac{1}{2}(x_n - x^*)^\top A(x_n - x^*) \\ &= \frac{1}{2}(x^*)^\top A(I-\rho A)^{2n}x^* \\ &\leq \frac{1}{2}\|x^*\|^2 \sup_{\lambda\in[\lambda_{\min},\lambda_{\max}]} \lambda(1-\rho\lambda)^{2n} \\ &\leq \frac{1}{2\rho(2n+1)}\|x^*\|^2 \end{aligned}$$
by applying the previous inequality to $\mu \coloneqq \rho\lambda$.

Exercise 3.

1. Proceed by induction. You should obtain $P_{t+1}(X) = 1 - \sum_{s=0}^{t} \gamma_{t,s} X P_s(X)$.

2. If (P_t) is a family such that P_t is of degree t and $P_t(0) = 1$ and $x_t - x^* = P_t(H)(x_0 - x^*)$, the family $(1, XP_0, XP_1, \ldots, XP_t)$ is a basis of $\mathbb{R}_{t+1}[X]$, so there are coefficients $\gamma_{t,s}$ such that $P_{t+1}(X) = 1 - \sum_{s=0}^{t} \gamma_{t,s} X P_s(X)$. Note that the coefficient on the polynomial 1 must be equal to 1 as $P_{t+1}(0) = 1$. We can then write the same computation as in question 1 but reversed, to obtain that $x_{t+1} = x_0 - \sum_{s=0}^{t} \gamma_{t,s} \nabla f(x_s)$.

3. The best method minimises $\|P_t(H)\|_2$ for any H whose eigenvalues are in $[\mu, L]$. We have $\|P_t(H)\|_2 = |P_t(\lambda)|$ for some eigenvalue λ of H, so the polynomial P_t should minimise $\sup_{\lambda \in [\mu, L]} |P_t(\lambda)|$.

4. We go from P_t to Q_t by denoting $\tilde{Q}_t(X) \coloneqq P_t(\frac{L+\mu}{2} - \frac{L-\mu}{2} X)$ and then dividing $\tilde{Q}_t$ by $\sup_{x \in [-1,1]} \left|\tilde{Q}_t(x)\right|$.

5. If Q_t is a polynomial of degree t such that $\sup_{x \in [-1,1]} |Q_t(x)| = 1$ and $Q_t(\frac{L+\mu}{L-\mu}) > T_t(\frac{L+\mu}{L-\mu})$, there exists a small $\varepsilon > 0$ such that $(1-\varepsilon) Q_t(\frac{L+\mu}{L-\mu}) > T_t(\frac{L+\mu}{L-\mu})$. The polynomial $\Delta \coloneqq T_t - (1-\varepsilon) Q_t$ zeroes out t times in $[-1, 1]$, as $|T_t|$ reaches 1 exactly $t+1$ times on $[-1, 1]$ and $(1-\varepsilon)|Q_t|$ never reaches 1 in this interval. Moreover, $T_t(1) = 1 > (1-\varepsilon) Q_t(1)$ and $T_t(\frac{L+\mu}{L-\mu}) < (1-\varepsilon) Q_t(\frac{L+\mu}{L-\mu})$ so Δ has one additional zero in $[1, \frac{L+\mu}{L-\mu}]$. But Δ is of degree t (and is non zero), which leads to a contradiction.