Exercises: differential calculus

Pierre Ablin

1 Convexity: general results

1.1

Show that a sum of smooth functions is smooth. What is the corresponding smoothness constant?

Show that the sum of strongly convex functions is strongly convex. What is the corresponding strong convexity constant ?

1.2

Show that $x \to \|x\|$ is convex, where $\|\cdot\|$ is any norm on $\mathbb{R}^d$.

1.3

Let $f : \mathbb{R}^d \to \mathbb{R}$ convex. Show that $g(x) = f(Ax + b)$ is convex, where $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. If f is μ-strongly convex, is g strongly convex? If so, what is a strong convexity constant of g? If f is L-smooth, is g smooth? If so, what is a smoothness constant of g?

Hint: You can demonstrate, and then use the fact that $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$ and $\sigma_{\max}(AB) \leq \sigma_{\max}(A)\sigma_{\max}(B)$ for two square matrices A, B.

1.4

Let $h_1, \dots, h_n : \mathbb{R} \to \mathbb{R}$ some convex function, $X \in \mathbb{R}^{n\times p}$ and define

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} h_i(\langle x_i, w\rangle),$$

where $x_i \in \mathbb{R}^p$ is the n-th row of X. Assume that the h_i are such that $\sup_{t\in\mathbb{R}} h_i''(t) = M < +\infty$. Show that f is smooth, and determine a smoothness constant.

2 Polyak-Lojasciewicz inequality

Let $f : \mathbb{R} \to \mathbb{R}$ be a μ-strongly convex function. Let x^* its arg-minimum. Show that f verifies the Polyak-Lojasciewicz inequality:

$$\forall x \in \mathbb{R}^d, \;\; f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$$

3 Convexity / non-convexity of matrix functions

3.1

Let $m \in \mathbb{R}$ and define $f(x) = \frac{1}{2}(x-m)^2$, $g(a,b) = \frac{1}{2}(ab-m)^2$. What are the gradient/ Hessian of these functions? Are these functions convex ?

3.2

Let $M \in \mathbb{R}^{p \times p}$ and define $f(X) = \frac{1}{2}\|X - M\|^2$, $g(A,B) = \frac{1}{2}\|AB - M\|^2$ where $A, B \in \mathbb{R}^{p \times p}$. What are the gradient/ Hessian of these functions? Are these functions convex ?

Hint: here, it is convenient to write the Hessians as linear operators. For instance for f, we can write $\nabla^2 f(X)(U) = \ldots$ where $\ldots$ is a linear function of $U \in \mathbb{R}^{p \times p}$.

4 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 $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_p = \lambda_{\min}$. We define the following *quadratic* objective function:

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

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. Hint : what recursion does the sequence $y_n = x_n - x^*$ satisfy?

We now use the spectral decomposition of A, and write

$$A = U^\top DU$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ contains the eigenvalues of A and $U \in \mathbb{R}^{p\times p}$ contains the eigenvectors of A. We recall that $UU^\top = U^\top U = I_p$.

Exercise 3: Define $z_n = U(x_n - x^*)$. Show that z_n is given by

$$z_n = (I_p - \rho D)^n z_0$$

Give a condition on ρ for this sequence to converge to 0.

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

Exercise 4: 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 5: 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 6: Show that for all x, $f(x) - f(x^*) = \frac{1}{2}(x - x^*)^\top A(x - x^*)$. Deduce a closed form formula for $f(x_n) - f(x^*)$.

We are now ready to give a bound that does not depend on the conditioning of the problem:

Exercise 7: Show that for all $\mu \in [0,1]$ and all n we have $(1-\mu)^{2n}\mu \leq \frac{1}{2n+1}$. Deduce that

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

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.

5 Solutions

1.1

Let f and g be L_f (resp. L_g)-smooth, and let $h = f + g$. Then, we have

$$\begin{aligned} \nabla^2 h(x) &= \nabla^2 f(x) + \nabla^2 g(x) && (1)\\ &\preceq L_f I_p + L_g I_p && (2)\\ &\preceq (L_f + L_g) I_p && (3) \end{aligned}$$

So h is $(L_f + L_g)$ smooth. Similarly, if f, g are μ_f and μ_g strongly convex, then $f+g$ is $(\mu_f + \mu_g)-$strongly convex.

1.2

Let $\lambda \in [0,1]$ and $x, y \in \mathbb{R}^d$. We have

$$\begin{aligned} \|\lambda x + (1-\lambda)y\| &\leq \|\lambda x\| + \|(1-\lambda)y\| && (4)\\ &\leq \lambda\|x\| + (1-\lambda)\|y\| && (5) \end{aligned}$$

which demonstrates convexity.

1.3

Convexity We have

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= f(A(\lambda x + (1-\lambda)y)) && (6)\\ &= f(\lambda(Ax+b) + (1-\lambda)(Ay+b)) && (7)\\ &\leq \lambda f(Ax+b) + (1-\lambda)f(Ay+b) && (8)\\ &\leq \lambda g(x) + (1-\lambda)g(y) && (9) \end{aligned}$$

so g is convex.

Hessian Let $x \in \mathbb{R}^d$ and consider a small $\varepsilon \in \mathbb{R}^d$. We have

$$
\begin{aligned}
g(x+\varepsilon) &= f(A(x+\varepsilon)+b) && (10)\\
&= f(Ax+b+A\varepsilon) && (11)\\
&= f(Ax+b) + \langle \nabla f(Ax+b), A\varepsilon\rangle + \frac{1}{2}\langle A\varepsilon, \nabla^2 f(Ax+b)A\varepsilon\rangle + \dots && (12)\\
&= f(Ax+b) + \langle A^\top \nabla f(Ax+b), \varepsilon\rangle + \frac{1}{2}\langle \varepsilon, A^\top \nabla^2 f(Ax+b)A\varepsilon\rangle + \dots && (13)
\end{aligned}
$$

from which we deduce

$$
\nabla g(x) = A^\top \nabla f(Ax+b)
$$
$$
\nabla^2 g(x) = A^\top \nabla^2 f(Ax+b)A
$$

Lemma Let $x \in \mathbb{R}^d$. We have $\|ABx\| \le \sigma_{\max}(A)\|Bx\| \le \sigma_{\max}(A)\sigma_{\max}(B)\|x\|$, which shows that $\sigma_{\max}(AB) \le \sigma_{\max}(A)\sigma_{\max}(B)$. Since the singular values of A are the inverse of the singular values of A^{-1}, we have $\sigma_{\min}(A) = \sigma_{\max}(A^{-1})^{-1}$.

Therefore, from the previous lemma, we also have $\sigma_{\min}(AB) = \sigma_{\max}((AB)^{-1})^{-1} \ge \left[\sigma_{\max}(A^{-1})\sigma_{\max}(B^{-1})\right]^{-1} = \sigma_{\min}(A)\sigma_{\min}(B)$

Strong convexity If f is $\mu-$strongly convex, we have $\sigma_{min}(\nabla^2 f(x)) \ge \mu$, so

$$
\sigma_{\min}(\nabla^2 g(x)) \ge \sigma_{\min}(A^\top)\sigma_{\min}(\nabla^2 f(Ax+b))\sigma_{\min}(A) = \sigma_{\min}(A)^2\mu
$$

So g is $(\sigma_{\min}(A)^2\mu)$- strongly convex.

Similarly, if f is $L-$smooth, then g is $(\sigma_{\max}(A)^2 L)$-smooth.

1.4

We have

$$
\begin{aligned}
\nabla^2 f(w) &= \frac{1}{n}\sum_{i=1}^n \nabla^2\left(h_i(\langle x_i, w\rangle)\right) && (14)\\
&= \frac{1}{n}\sum_{i=1}^n h_i''(\langle x_i, w\rangle)x_i x_i^\top \qquad = \frac{1}{n}X^\top D X && (15)
\end{aligned}
$$

where $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix with diagonal coefficients $h_i''(\langle x_i, w\rangle)$. Using the lemma in the previous exercise, we therefore find

$$
\sigma_{\max}(\nabla^2 f) \le \frac{1}{n}\sigma_{\max}(X)^2 M
$$

so f is $(\frac{\sigma_{\max}(X)^2 M}{n})-$smooth.

2

Strong convexity gives

$$
\forall x, y,\ \ f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|x-y\|^2
$$

Taking $y = x - \frac{1}{\mu}\nabla f(x)$, we find $f(x) - f(y) \leq \frac{1}{\mu}\langle \nabla f(x), \nabla f(x)\rangle - \frac{1}{2\mu}\|\nabla f(x)\|^2 = \frac{1}{2\mu}\|\nabla f(x)\|^2$. Since $f(y) \geq f(x^*)$, we conclude

$$f(x) - f(x^*) \leq f(x) - f(y) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

3.1

We have $\nabla f(x) = x - m$ and $\nabla^2 f(x) = 1$. f is therefore convex.

We have $\nabla g(a, b) = [b(ab - m), a(ab - m)]$ and $\nabla^2 g = \begin{bmatrix} b^2 & 2ab - m \\ 2ab - m & a^2 \end{bmatrix}$. g is non-convex because the set of its minimizers, $\{a, b \in \mathbb{R} | ab = m\}$ is non-convex (it is a hyperbole).

3.2

We have $\nabla f(X) = X - M$ and $\nabla^2 f(X) = Id$ (that is to say, for all $U \in \mathbb{R}^{p\times p}$, $\nabla^2 f(X)(U) = U$).

To find the Hessian of g, we let U, V some small $p \times p$ matrices and compute

$$\begin{aligned}
g(A + U, B + V) &= \frac{1}{2}\|(A + U)(B + V) - M\|^2 && (16)\\
&= \frac{1}{2}\|AB + UB + AV + UV - M\|^2 && (17)\\
&= g(A, B) + \langle UB + AV, AB - M\rangle + \langle UV, AB - M\rangle + \frac{1}{2}\langle UB, UB\rangle + \frac{1}{2}\langle AV, AV\rangle + \dots && (18)
\end{aligned}$$

Identifying the terms, we find $\nabla_U g(A, B) = (AB - M)B^\top$ and $\nabla_V g(A, B) = A^\top(AB - M)$. The second order part of the expansion is $\langle UV, AB - M\rangle + \frac{1}{2}\langle UB, UB\rangle + \frac{1}{2}\langle AV, AV\rangle$, and it needs to be rewritten as $\frac{1}{2}\langle U, \nabla g^2_{UU}(A, B)(U)\rangle + \frac{1}{2}\langle V, \nabla g^2_{VV}(A, B)(V)\rangle + \langle U, \nabla g^2_{VU}(A, B)(V)\rangle$.

We therefore find

$$\nabla g^2_{UU}(A, B)(U) = UBB^\top$$

$$\nabla g^2_{VV}(A, B)(V) = A^\top AV$$

$$\nabla g^2_{VU}(A, B)(V) = (AB - M)V^\top$$

4.1

We have

$$f(x + \varepsilon) = f(x) + \langle Ax - b, \varepsilon\rangle + \frac{1}{2}\langle \varepsilon, A\varepsilon \qquad (19)$$

Hence the gradient of f is $Ax - b$, and its Hessian is A. Since A is positive, f is convex. Its minimum x^* is reached when $\nabla f(x^*) = 0$, i.e. $Ax^* = b$, which gives

$$x^* = A^{-1}b$$

We find $f(x*) = -\frac{1}{2}\langle b, A^{-1}b\rangle$

4.2

We have

$$x_{n+1} = x_n - \rho(Ax_n - b) = (I_p - \rho A)x_n + \rho b$$

The sequence y_n therefore satisfies $y_{n+1} = (I_p - \rho A)y_n$, and as a consequence:

$$y_n = (I - \rho A)^n y_0$$

and we find

$$x_n = x^* - (I_p - \rho A)^n x^*$$

4.3

We have $I_p = U^\top U$, hence $I_p - \rho A = U^\top (I_p - \rho D)U$ and $(I_p - \rho A)^n = U^\top (I_p - \rho D)^n U$.

As a consequence:

$$z_n = (I_p - \rho D)^n U x_0 = (I_p - \rho D)^n z_0$$

This sequence converges to 0 when $(I_p - \rho D)^n$ goes to 0. Since this matrix is diagonal, it is equivalent to $|1 - \rho\lambda_i| < 1$ for all i. This gives

$$\begin{cases} \rho > 0 \text{ and} \\ \rho < \frac{2}{\lambda_i} \text{ for all i} \end{cases}$$

In other words, $\rho \in (0, \frac{2}{\lambda_{\max}})$.

4.4

Since U is orthogonal, we have $\|x_n - x^*\| = \|z_n\| \leq \sigma_{\max}((I_p - \rho D)^n)\|z_0\|$. Furthermore, since $(I_p - \rho D)^n$ is diagonal with positive values and largest coefficient $(1 - \frac{\lambda_{\min}}{\lambda_{\max}})^n$, we have $\sigma_{\max}((I_p - \rho D)^n) = (1 - \frac{\lambda_{\min}}{\lambda_{\max}})^n$, and the advertized result follows.

4.5

We have $(1 - \frac{1}{1000})^n < \frac{1}{10}$ when $n > \frac{\log(1/10)}{\log(1 - \frac{1}{1000})} = 2301$. In order to decrease it by a factor 10, we need twice as many iterations.

4.6

Simple computations show that $f(x) - f(x^*) = \frac{1}{2}(x - x^*)^\top A(x - x^*)$. We deduce that

$$f(x_n) - f(x^*) = \frac{1}{2}(x_n - x^*)^\top A(x_n - x^*) \tag{20}$$
$$= \frac{1}{2}(x^*)^\top (I_p - \rho A)^n A(I_p - \rho A)^n x^* \tag{21}$$

And since $(I_p - \rho A)^n$ is a polynomial in A, it commutes with A, and we can write

$$f(x_n) - f(x^*) == \frac{1}{2}(x^*)^\top (I_p - \rho A)^{2n} A x^*.$$

4.7

The maximum of $\mu \mapsto (1 - \mu)^{2n}\mu$ is attained when (by cancelling the derivative) $\mu^* = \frac{1}{2n+1}$. Hence, we have $(1 - \mu)^{2n}\mu \leq (1 - \mu^*)^{2n}\mu^* \leq \frac{1}{2n+1}$.

Once again, using the spectral theorem on A gives

$$f(x_n) - f(x^*) = \frac{1}{2}(Ux^*)^\top (I_p - \rho D)^{2n} D(Ux^*).$$

The prelimenary result shows that the values in $\rho(I_p - \rho D)^{2n} D$ are upper bounded by $\frac{1}{2n+1}$, hence the values in $(I_p - \rho D)^{2n} D$ are upper bounded by $\frac{1}{(2n+1)\rho}$ and we find

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