Mathematical background for Machine Learning

Mathurin Massias

Contents

Notation As in the slides, as much as possible, we use upper case bold $\mathbf{X}$ for matrices, lowercase $\mathbf{x}$ for vectors and lower case nonbold x for scalars. Uppercase non bold is for random variables. Calligraphic uppercase $\mathcal{X}$ is for sets. This is a general convention, with potential exceptions.

Abbreviations L/RHS: left/right hand side. wlog: without loss of generality.

1 Expectation, variance and covariance

1.1 Definitions and facts

- We say that the random variable X taking values in $\mathbb{R}^d$ has *density* $p : \mathbb{R}^d \to \mathbb{R}$ if for any $\mathcal{A} \subset \mathbb{R}^d$, $\mathbb{P}(X \in \mathcal{A}) = \int_{\mathbf{x} \in A} p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$. A probability density therefore integrates to 1: $\int_{\mathbf{x}} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1$.
- The expectation of said random variable X is then $\mathbb{E}[X] = \int_{\mathbf{x}} \mathbf{x} \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \in \mathbb{R}^d$.

- The expectation is linear: if X is a random variable in $\mathbb{R}^d$ and $\mathbf{A}$ is a fixed (nonrandom) matrix in $\mathbb{R}^{n\times d}$, then $\mathbb{E}[\mathbf{A}X] = \mathbf{A}\mathbb{E}[X]$. In particular, if X has expectation $\mathbf{0}_d$, so does the random variable $\mathbf{A}X$.

- Expectation and trace commute: if X is a matrix-valued random variable, $\mathbb{E}[\mathrm{tr}(X)] = \mathrm{tr}[\mathbb{E}(X)]$. This is often used in the following way: if X is a random variable in $\mathbb{R}^d$, $\mathbf{X}^\top\mathbf{X}$ is a scalar, hence equal to its trace, and:

$$\underbrace{\mathbb{E}[\mathbf{X}^\top\mathbf{X}]}_{\in\mathbb{R}} = \mathbb{E}[\mathrm{tr}(\mathbf{X}^\top\mathbf{X})] = \mathbb{E}[\mathrm{tr}(XX^\top)] = \mathrm{tr}(\underbrace{\mathbb{E}[XX^\top]}_{\in\mathbb{R}^{d\times d}}) \tag{1}$$

- If X is a random variable in $\mathbb{R}^d$, its *variance* is a (positive) scalar:

$$\begin{aligned}\mathrm{Var}(X) &:= \mathbb{E}[\|X - \mathbb{E}[X]\|^2] &(2)\\ &= \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2 &(3)\\ &= \mathbb{E}[\mathbf{X}^\top\mathbf{X}] - \mathbb{E}[X]^\top\mathbb{E}[X] \in \mathbb{R}_+ &(4)\end{aligned}$$

  (see how this generalizes the variance of scalar random variables, which is $\mathbb{E}[(X - \mathbb{E}[X])^2]$: we simply add a norm).
  Its **covariance** is a matrix in $\mathbb{R}^{d\times d}$:

$$\begin{aligned}\mathrm{cov}(X) &:= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] &(5)\\ &= \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top \in \mathbb{R}^{d\times d} &(6)\end{aligned}$$

  A covariance matrix is always positive semidefinite (Definition 3.2): for $u \in \mathbb{R}^d$, $u^\top(\mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top)u = \mathbb{E}[\|Xu\|^2] - \|\mathbb{E}[X]u\|^2$. By a computation very similar to (1), the variance is equal to the trace of the covariance.

  From the very definition of the covariance and linearity of expectatione, we see that $\mathrm{cov}(\mathbf{A}X) = \mathbf{A}\,\mathrm{cov}(X)\mathbf{A}^\top$ for any fixed matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$.

- Jensen's inequality: if $\phi : \mathbb{R}^d \to \mathbb{R}$ is a convex function and X a random variable taking values in $\mathbb{R}^d$, with density p,

$$\phi(\mathbb{E}[X]) \le \mathbb{E}[\phi(X)], \quad \text{aka} \quad \phi\left(\int_{\mathbf{x}} \mathbf{x}\,p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \le \int_{\mathbf{x}} \phi(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{7}$$

  Example of application: for the convex $\phi = \|\cdot\|^2$, we get $\|\mathbb{E}[X]\|^2 \le \mathbb{E}[\|X\|^2]$, giving an alternative proof that the variance of X is a positive quantity, since the variance is the RHS minus the LHS.

  Mnemonics: it is frequent to write Jensen's inequality in the wrong direction. To remember the correct direction of the inequality in Jensen, use the above example: if X has 0 mean, then the LHS can be 0. On the other hand, the RHS is the integral of the positive quantity $\|\mathbf{x}\|^2$ and thus must be positive.

1.2 The conditional expectation

This section is not 100% rigorous: what it presents as definitions are rather characterizations in specific settings. A rigorous construction of the conditional expectation, valid in more generic setups, is possible, but it would take us too far for a machine learning course.

To motivate the introduction of the *conditional* expectation, we start of a characterization of the expectation: for a random variable Y taking values in $\mathbb{R}^d$, one has

$$\mathbb{E}[Y] = \operatorname*{argmin}_{\mathbf{c}\in\mathbb{R}^d} \mathbb{E}_Y[\|Y-\mathbf{c}\|^2] \tag{8}$$

The proof is quite simple: writing $\mathbf{e} = \mathbb{E}[Y]$, one has for any $\mathbf{c} \in \mathbb{R}^d$:

$$\begin{aligned}
\mathbb{E}[\|Y-\mathbf{c}\|^2] &= \mathbb{E}[\|Y-\mathbf{e}+\mathbf{e}-\mathbf{c}\|^2] &(9)\\
&= \mathbb{E}[\|Y-\mathbf{e}\|^2] + 2\,\mathbb{E}[\langle Y-\mathbf{e}, \mathbf{e}-\mathbf{c}\rangle] + \mathbb{E}[\|\mathbf{e}-\mathbf{c}\|^2] &(10)\\
&= \mathbb{E}[\|Y-\mathbf{e}\|^2] + 2\,\langle \mathbb{E}[Y]-\mathbf{e}, \mathbf{e}-\mathbf{c}\rangle + \|\mathbf{e}-\mathbf{c}\|^2 &(11)\\
&= \mathbb{E}[\|Y-\mathbf{e}\|^2] + \|\mathbf{e}-\mathbf{c}\|^2 &(12)
\end{aligned}$$

which is greater than $\mathbb{E}[\|Y-\mathbf{e}\|^2]$, with equality only when $\mathbf{c} = \mathbf{e}$.

So in a sense, the expectation is the constant that helps the most in predicting Y: it has minimal L2 error. Now suppose we have another random variable X, and we want to know how much information we can get on Y from X. Inspired by Equation (8), we can define the minimization problem

$$\operatorname*{argmin}_{g \text{ measurable}} \mathbb{E}_{X,Y}[\|Y-g(X)\|^2] \tag{13}$$

It turns out that this problem has a unique solution, g^*. It allows defining the conditional expectation, $\mathbb{E}[Y|X] := g^*(X)$; note that it is a random variable.

Two extreme cases help to better grasp the conditional expectation:

1. if Y is X-measurable ($Y = \phi(X)$ with ϕ a measurable function), $\mathbb{E}[Y|X] = Y$
2. if $Y \perp\!\!\!\perp X$, $\mathbb{E}[Y|X] = \mathbb{E}[Y]$

By Equation (13), the conditional expectation is the projection of the random variable Y onto the set of X-measurable random variables, hence one has for all measurable function h, by orthogonality[1]

$$\begin{aligned}
\mathbb{E}_{X,Y}[(Y-g^*(X))h(X)] = 0 &(14)\\
\mathbb{E}_{X,Y}[Yh(X)] = \mathbb{E}[g^*(X)h(X)] &(15)
\end{aligned}$$

which in particular (for which choice of h?) yields:

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y] \tag{16}$$

We conclude with a very important caveat: **the conditional expectation may refer to two things**:

[1] the vanilla property is actually $\mathbb{E}_{X,Y}[(Y-g^*(X))(h(X)-g^*(X))] = 0$, but we use the fact that $h-g^*$ is measurable if and only if h is

1. a random variable, $\mathbb{E}[Y|X] = g^*(X)$, as we've seen so far
2. a function of $\mathbf{x} \in \mathbb{R}^d$, whose value is denoted $\mathbb{E}[Y|X = \mathbf{x}]$, which has value $g^*(\mathbf{x})$. When all random variables have densities, we have the useful formula:

$$\mathbb{E}[Y|X = \mathbf{x}] = \int_{\mathbf{y}} \mathbf{y}\, p_{Y|X}(\mathbf{y}|\mathbf{x})\, \mathrm{d}\mathbf{y} \tag{17}$$

which is to be thought of in comparison of the vanilla expectation $\mathbb{E}[Y] := \int_{\mathbf{y}} \mathbf{y}\, p_Y(\mathbf{y})\, \mathrm{d}\mathbf{y}$: in the conditional expectation, we take the average value of $\mathbf{y}$, but according to $p_{Y|X}(\cdot|\mathbf{x})$, not p_Y.

2 Differential calculus

Again this note is not 100% rigorous: we do not delve into the various notions of differentiability.

Definition 2.1 (Gradient vector)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then for any $\mathbf{x} \in \mathbb{R}^d$, there exists a vector of $\mathbb{R}^d$, denoted $\nabla f(\mathbf{x})$ and called the gradient of f at $\mathbf{x}$, such that for all $\mathbf{h} \in \mathbb{R}^d$,*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + o(\|\mathbf{h}\|) \tag{18}$$

Equation (18) is a linear approximation at f, generalizing the well-known Taylor formula from 1D. The gradient vector is unique: if for some $\mathbf{x} \in \mathbb{R}^d$ there exists $\mathbf{g} \in \mathbb{R}^d$ such that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{h} \rangle + o(\|\mathbf{h}\|) \tag{19}$$

one must have $\mathbf{g} = \nabla f(\mathbf{x})$. The uniqueness provides a practical way to compute the gradient: write $f(\mathbf{x} + \mathbf{h})$, try to isolate $f(\mathbf{x})$ and make a linear term appear, identify the slope of the linear term with $\nabla f(\mathbf{x})$. When the scalar product is the Euclidean one, one has:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_d}(\mathbf{x}) \end{pmatrix} \tag{20}$$

Exercise 2.2. *Compute the gradient of $\mathbf{x} \mapsto \|\mathbf{x}\|^2$ in two ways: with Equation (20), and using Equation (18) with uniqueness of the gradient.*
A bit harder but same spirit: do it for $\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$.

Definition 2.3 (Hessian matrix)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable. Then for any $\mathbf{x} \in \mathbb{R}^d$, there exists a matrix in $\mathbb{R}^{d \times d}$, denoted $\nabla^2 f(\mathbf{x})$ and called the Hessian of f at $\mathbf{x}$, such that for all $\mathbf{h} \in \mathbb{R}^d$,*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2}\langle \mathbf{h}, \nabla^2 f(\mathbf{x})\mathbf{h} \rangle + o(\|\mathbf{h}\|^2) \tag{21}$$

Much like Equation (18), Equation (21) is a local approximation of f, but this time quadratic *(Taylor of order 2).*

If the scalar product is the Euclidean one, one has

$$\nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \right)_{ij} \tag{22}$$

If the second partial derivatives are continuous, the Hessian is symmetric (Schwarz's theorem); this is always the case in ML.
As for the gradient, if the Hessian exists it is unique: if for $\mathbf{x} \in \mathbb{R}^d$ the exists a symmetric $\mathbf{H} \in \mathbb{R}^{d \times d}$ such that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{h} \rangle + \frac{1}{2} \langle \mathbf{h}, \mathbf{H}\mathbf{h} \rangle + o(\|\mathbf{h}\|^2) \tag{23}$$

then $\mathbf{H} = \nabla^2 f(\mathbf{x})$.

Exercise 2.4. *Show that for any skew-symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{x}^\top \mathbf{A} \mathbf{x} = 0$. Is there uniqueness of $\mathbf{H}$ in Equation (23) if we do not require that $\mathbf{H}$ is symmetric?*

Exercise 2.5. *Compute the Hessian of $\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{x}\|^2$ in two ways: with Equation (23) and with Equation (22).*
A bit harder, but not too much and very useful: do the same for $\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$.

Finally, when f is vector-valued, there is a generalization of the gradient, called the Jacobian.

Definition 2.6 (Jacobian)**.** *The Jacobian of $f : \mathbf{x} \in \mathbb{R}^n \mapsto \begin{pmatrix} f_1(\mathbf{x}) \\ \dots \\ f_m(\mathbf{x}) \end{pmatrix}$ is the matrix of partial derivatives:*

$$J_f(\mathbf{x}) = \left(\frac{\partial f_i}{\partial x_j}(\mathbf{x}) \right)_{ij} \in \mathbb{R}^{m \times n} \ . \tag{24}$$

and one has (with uniqueness):

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + J_f(\mathbf{x})\, \mathbf{h} + o(\|\mathbf{h}\|) \tag{25}$$

Exercise 2.7. *Compute the Jacobian of $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ for $\mathbf{A} \in \mathbb{R}^{n \times d}$.*

⚠ For real-valued functions ($m = 1$), the Jacobian is the *transposed* gradient. Else, the Jacobian consists of stacked transposed gradients:

$$J_f(\mathbf{x}) = \begin{pmatrix} —\nabla f_1(\mathbf{x})^\top— \\ \dots \\ —\nabla f_m(\mathbf{x})^\top— \end{pmatrix} \in \mathbb{R}^{m \times n} \tag{26}$$

The Jacobian of the gradient is the Hessian (under the conditions of Schwarz's theorem).

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

$$J_{\nabla f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{x}) \end{pmatrix} \in \mathbb{R}^{n \times n}$$

Proposition 2.8 (Chain rule)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^p$ be differentiable functions. Then their composition $g \circ f$ is differentiable and*

$$\underbrace{J_{g \circ f}}_{\in \mathbb{R}^{p \times d}}(\mathbf{x}) = \underbrace{J_g(f(\mathbf{x}))}_{\in \mathbb{R}^{p \times n}} \underbrace{J_f(\mathbf{x})}_{\in \mathbb{R}^{n \times d}} \ . \tag{27}$$

If $d = n = p = 1$ you recover the well-known $(g \circ f)'(x) = g'(f(x))f'(x)$.

Exercise 2.9. *Use the chain rule to compute the Jacobian of $\mathbf{x} \mapsto \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$. Deduce its gradient.*

3 Linear algebra

Theorem 3.1 (Spectral theorem)**.** *Let $\mathbf{M} \in \mathbb{R}^{d \times d}$ be a symmetric real matrix. Then it can be diagonalized in an orthonormal basis: there exists $\mathbf{U} \in \mathbb{R}^{d \times d}$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ such that $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathrm{Id}_d$ and*

$$\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \tag{28}$$

If we write $\mathbf{u}_i$ for the columns of $\mathbf{U}$, then the equation also writes $\mathbf{M} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$.

This is the most important theorem in this section. It is very convenient to diagonalize a matrix in order to raise it to the power k, as $\mathbf{M}^k = \mathbf{U}^\top \mathbf{\Lambda}^k \mathbf{U}$, and to simplify computations: up to a change of basis, a symmetric real matrix can be considered diagonal.

Definition 3.2. *A matrix $\mathbf{M}$ is said to be positive semidefinite (PSD, written $\mathbf{M} \succeq 0$) if it is symmetric, and for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^\top \mathbf{M} \mathbf{x} \geq 0$. If the inequality holds strictly for all nonzero $\mathbf{x}$, $\mathbf{M}$ is said to be positive definite (PD, written $\mathbf{M} \succ 0$).*

Note that by definition, a PSD matrix must be symmetric; also notice that for a generic $\mathbf{A} \in \mathbb{R}^{d \times d}$, the antisymmetric part of $\mathbf{A}$ has no influence of $\mathbf{x}^\top \mathbf{A} \mathbf{x}$, so it's not really a restriction.

Example 3.3 (How to create PSD matrices?)**.** *The typical way to create a PSD matrix is to pick $\mathbf{M} = \mathbf{A}^\top \mathbf{A}$ for $\mathbf{A}$ of any shape, because then $\mathbf{x}^\top \mathbf{M} \mathbf{x} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A} \mathbf{x} = \|\mathbf{A}\mathbf{x}\|^2 \geq 0$. By this formulation, we deduce that $\mathbf{A}^\top \mathbf{A}$ is PD if and only if $\mathbf{A}$ is injective.*

🔧 This type of computations pops up a lot in covariance matrices.

The following is a useful characterization of PSD matrices, rather than using the definition.

Proposition 3.4 (Characterization of PSD matrices). *A symmetric matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ (therefore diagonalizable by Theorem 3.1) is PSD if and only if all its eigenvalues are positive.*

Definition 3.5 (Loewner order). *There is a partial order on symmetric matrices: for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ symmetric, we write $\mathbf{A} \succeq \mathbf{B}$ if and only if $\mathbf{A} - \mathbf{B} \succeq 0$. This order is not total (find an example of two matrices which cannot be compared).*

Apart from PSD matrices, another class of very useful matrices are rank one matrices.

Proposition 3.6. *The matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ is rank one if and only if it writes $\mathbf{M} = \mathbf{x}\mathbf{y}^\top$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, both nonzero. If $\mathbf{M}$ is both symmetric and rank one, then it writes $\mathbf{x}\mathbf{x}^\top$ or $-\mathbf{x}\mathbf{x}^\top$ for $\mathbf{x} \in \mathbb{R}^d$.*

Remark 3.7 (Writing a matrix as sum of rank one matrices). *Let $\mathbf{M} \in \mathbb{R}^{n\times p}$, $\mathbf{N} \in \mathbb{R}^{p\times q}$. Then $\mathbf{M}\mathbf{N} = \sum_{i=1}^{p} \mathbf{M}_{:i}\mathbf{N}_{i:}$.*

The eigenvalue decomposition is a very powerful tool, but not all matrices can be diagonalized. The singular value decomposition (SVD) is a more versatile tool, generalizing the eigenvalue decomposition, that applies to any matrix (even non-square ones).

Proposition 3.8 (Singular value decomposition). *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $r = \mathrm{rank}(\mathbf{A})$. Then there exist an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$, an orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d\times d}$ and a diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{n\times d}$, with r nonzero decreasing diagonal entries $\sigma_1 \geq \ldots \geq \sigma_r > 0$, such that:*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top . \tag{29}$$

The σ_i's are called the singular values *of $\mathbf{A}$. The columns of $\mathbf{U}$ (resp. $\mathbf{V}$), denoted $\mathbf{u}_i$ (resp. $\mathbf{v}_i$) and are called the left (resp. right)* singular vectors *of $\mathbf{A}$.*

Since $\mathbf{\Sigma}$ only has r nonzero entries, one often writes the SVD under its compact form, *$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ with $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}, \mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{V} \in \mathbb{R}^{d\times r}$.* ⚠ *Note that in that case $\mathbf{U}$ and $\mathbf{V}$ are no longer orthogonal: $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathrm{Id}_r$ but $\mathbf{U}\mathbf{U}^\top \neq \mathrm{Id}_n$ and $\mathbf{V}\mathbf{V}^\top \neq \mathrm{Id}_n$.*
See Figure 1 for an illustration of the various ways to write the SVD.

Finally, the SVD can also be written as a sum of r matrices of rank 1:

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top . \tag{30}$$

The SVD allows generalizing the notion of inverse to noninvertible and non square matrices, through the Moore-Penrose pseudoinverse.

Definition 3.9. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$, admitting as* compact *singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Its Moore-Penrose pseudoinverse is, by definition,*

$$\mathbf{A}^\dagger := \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^\top = \sum_{i=1}^{r} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^\top \in \mathbb{R}^{d\times n}. \tag{31}$$

⚠ *compared to the SVD of $\mathbf{A}$, σ_i becomes $1/\sigma_i$, but also the orders of $\mathbf{u}_i$ and $\mathbf{v}_i$ are reversed.*

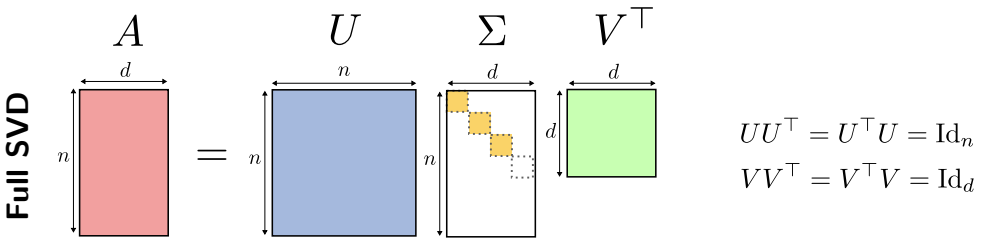

Figure 1: Full and compact SVD of $\mathbf{A} \in \mathbb{R}^{n\times d}$, in the $n \geq d$ setting. Top: with full matrices U and V. Bottom: with square $\mathbf{\Sigma}$ of size $r \times r$ (only nonzero entries). Picture courtesy of Titouan Vayer.

It generalizes the inverse in the following sense: if $\mathbf{A}$ is invertible, then its inverse and its Moore-Penrose pseudoinverse coincide.

Definition 3.10 (Matrix norms)**.** *We mainly consider two norms on matrices:*

1. *the Frobenius norm, which is the Euclidean norm (= it is associated to a scalar product), denoted $\|\cdot\|_F$ or simply $\|\cdot\|$ (when there is not index to the norm, in the vast majority of cases it means the Euclidean norm)*

$$\|\mathbf{M}\|_F = \sqrt{\sum_{i,j} M_{ij}^2} = \sqrt{\langle \mathbf{M}, \mathbf{M}\rangle} \tag{32}$$

   with $\langle \mathbf{M}, \mathbf{N}\rangle := \operatorname{tr}(\mathbf{M}^\top \mathbf{N}) = \sum_{ij} M_{ij}N_{ij}$

2. *the operator norm,*

$$\|\mathbf{M}\|_2 = \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{M}\mathbf{x}\|}{\|\mathbf{x}\|} \tag{33}$$

   where the norms $\|\mathbf{M}\mathbf{x}\|$ and $\|\mathbf{x}\|$ are Euclidean norms on vectors. A useful (straightforward) property of the operator norm is that $\|\mathbf{M}\mathbf{x}\|_2 \leq \|\mathbf{M}\|_2\|\mathbf{x}\|_2$ for all $\mathbf{x}$.

These norms are related to the singular values: using a compact SVD for $\mathbf{M}$, namely $\mathbf{M} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top$ with $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$,

$$\operatorname{tr} \mathbf{M}^\top\mathbf{M} = \operatorname{tr} \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \operatorname{tr} \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top = \operatorname{tr} \mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^2 = \operatorname{tr} \mathbf{\Sigma}^2 \tag{34}$$

so the Frobenius norm is the square root of the sum of squared singular values, $\sqrt{\sum_{i=1}^r \sigma_i^2}$. On the other hand, for the operator norm

$$\frac{\|\mathbf{M}\mathbf{x}\|^2}{\|\mathbf{x}\|^2} = \frac{\langle \mathbf{M}\mathbf{x}, \mathbf{M}\mathbf{x}\rangle}{\|\mathbf{x}\|^2} = \frac{\langle \mathbf{x}, \mathbf{M}^\top\mathbf{M}\mathbf{x}\rangle}{\|\mathbf{x}\|^2} \leq \frac{\|\mathbf{x}\|\|\mathbf{M}^\top\mathbf{M}x\|}{\|\mathbf{x}\|^2} \tag{35}$$

$$\leq \frac{\|\mathbf{x}\|\lambda_{\max}(\mathbf{M}^\top\mathbf{M})\|\mathbf{x}\|}{\|\mathbf{x}\|^2} \tag{36}$$

$$= \lambda_{\max}(\mathbf{M}^\top\mathbf{M}) \tag{37}$$

Now, if we take $\mathbf{x}$ an eigenvector associated to $\lambda_{\max}(\mathbf{M}^\top\mathbf{M})$, the inequality is tight. So $\|\mathbf{M}\|_2 = \sqrt{\lambda_{\max}(\mathbf{M}^\top\mathbf{M})}$, namely the largest singular values of $\mathbf{M}$. We conclude by noticing that this proves that the operator norm is always smaller than the Frobenius norm.

4 Convexity and minimizers

This is an extremely condensed version of the notes M2 class "Large scale optimization for Machine Learning", available at https://mathurinm.github.io/assets/2022_ens/class.pdf

4.1 Convexity

Definition 4.1 (Global and local minimizers)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$. A* global minimizer *of f is a point $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$. The set of minimizers of f is denoted* $\operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ *or simply* $\operatorname{argmin} f$*. A* local minimizer *of f is a point $\mathbf{x}^*\mathbb{R}^d$ such that there exists a neighborhood $\mathcal{V}$ of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{V}$; all global minimizers are local minimizers.*

Definition 4.2 (Convexity)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is* convex *if and only if it lies below its cords:*

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \forall \lambda \in]0, 1[\,, \quad f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \ . \tag{38}$$

It is strictly convex *if the inequality (38) holds strictly; strict convexity therefore implies convexity.*

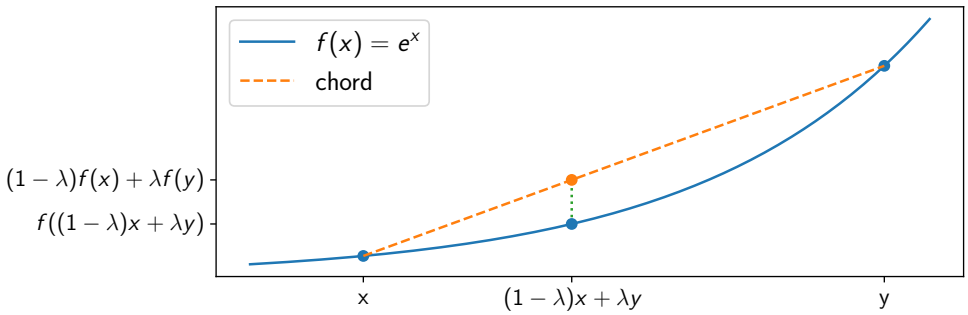

Figure 2: A convex function is below its chords.

🔧 Rather than using the definition, the easiest way to prove that a function is convex, in general, is to use the following characterization.

Proposition 4.3 (2nd order characterization of convex functions)**.** *A twice differentiable function f is convex if and only if its Hessian is PSD everywhere, namely $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \mathbb{R}^d$, or equivalently the eigenvalues of the Hessian at every $\mathbf{x}$ are positive.*

Example 4.4. 🔧 *The Ordinary Least square objective function is convex, since its Hessian is constant equal to $\mathbf{A}^\top \mathbf{A}$, which is PSD (Example 3.3).*

Convex functions are amenable to optimization because of the nice "local to global" properties they enjoy.

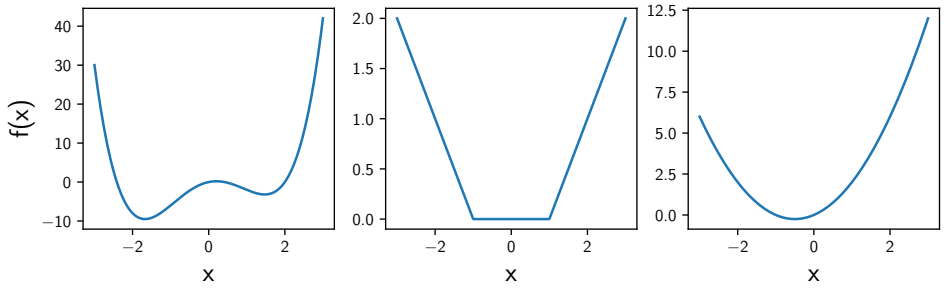

Figure 3: Left to right: nonconvex function, convex but not strictly convex function, and strictly convex function.

Proposition 4.5 (Local minimizers are global minimizers)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then all local minimizers of f are also global minimizers.*

Proposition 4.6 (Above its tangents)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Then*

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ . \tag{39}$$

Geometrically, it tells that f lies above all of its tangents (thinking of it as $\mathbf{x}$ moving and $\mathbf{y}$ being fixed). This property is sometimes called "local to global": from local information $\nabla f(\mathbf{y})$, we get information about the global behavior of f (namely, a lower bound). See Figure 4.

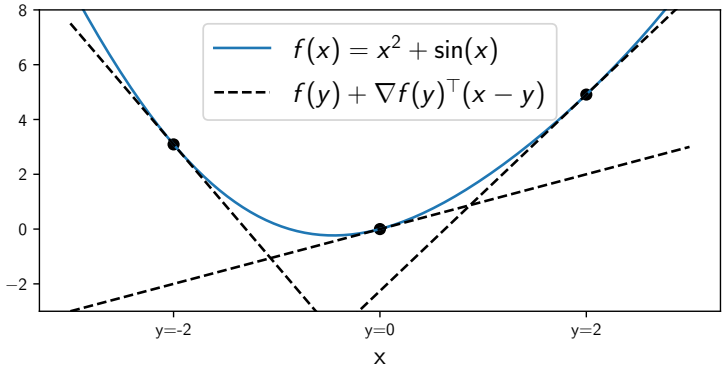

Figure 4: The function $f : x \mapsto x^2 + \sin(x)$ (in blue) is convex since $\nabla^2 f(x) = f''(x) = 2 + \cos(x) \geq 0$ (Proposition 4.3). It is therefore lower bounded by its tangent at y for any y (dashed black lines).

Convex differentiable functions are nice because it is easy to characterize their minimizers.

Proposition 4.7 (Optimality condition for convex differentiable functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Then*

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \iff \nabla f(\mathbf{x}^*) = 0 \ . \tag{40}$$

🔧 we use this property to compute the minimizer of Ordinary Least Squares, or the optimal centroid positions given cluster assignments in K-means.

4.2 Additional properties: strong convexity and smoothness

To analyze the behavior of optimization algorithm, we will consider functions with additional properties: strong convexity, and smoothness.

Definition 4.8 (Strong convexity)**.** *Let $\mu > 0$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda \in]0, 1[$,*

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2 \ . \tag{41}$$

Since $\mu > 0$, $-\frac{\mu}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2 < 0$ for $\mathbf{x} \neq \mathbf{y}$ and so:

$$\text{strong convexity} \implies \text{strict convexity} \implies \text{convexity.}$$

Proposition 4.9 (First order characterization of strong convexity)**.** *A differentiable function f is μ-strongly convex if and only if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2 \tag{42}$$

Think of this inequality for a $\mathbf{y}$ fixed, and involving two functions of $\mathbf{x}$, f and $\phi_y(\mathbf{x}) := f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$. The function ϕ_y is a convex, isotropic parabola; it is tangent to f at y since $\phi_y(y) = f(\mathbf{y})$ and $\nabla\phi_y(\mathbf{y}) = \nabla f(\mathbf{y})$ (see TD for a proof).

Equation (42) says that for every choice of y, the tangent parabola ϕ_y globally lower bounds f. In that sense, strongly convex functions can be lower bounded by tangent parabolas *(while convex functions are lower bounded by* tangent lines*, see (39)), as illustrated in Figure 5.*

As for convexity, the easiest way to show that a function is strongly convex is through its Hessian.

Proposition 4.10 (Second order characterization of strong convexity)**.** *A twice differentiable function f is μ-strongly convex if and only if $\nabla^2 f(\mathbf{x}) \succeq \mu\mathrm{Id}$ for all $\mathbf{x} \in \mathbb{R}^d$ (meaning that $\nabla^2 f(\mathbf{x}) - \mu\mathrm{Id}$ is PSD everywhere, or equivalently the eigenvalues of the Hessian at every $\mathbf{x}$ are all greater than μ.)*

🔧 we use this to show that the OLS objective function is strongly convex if and only if the design matrix has full row rank ($\mathbf{X}^\top\mathbf{x}$ is invertible) invertible. What about the Ridge objective function?

Proposition 4.11. *A function f is μ-strongly convex if and only if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex:*

$$\nabla^2(f - \frac{\mu}{2}\|\cdot\|^2)(\mathbf{x}) \succeq 0 \iff \nabla^2 f(\mathbf{x}) - \mu\mathrm{Id} \succeq 0 \iff \nabla^2 f(\mathbf{x}) \succeq \mu\mathrm{Id} \tag{43}$$

In this sense, a strongly convex function is so convex *that when you subtract a parabola with positive curvature, it remains convex.*

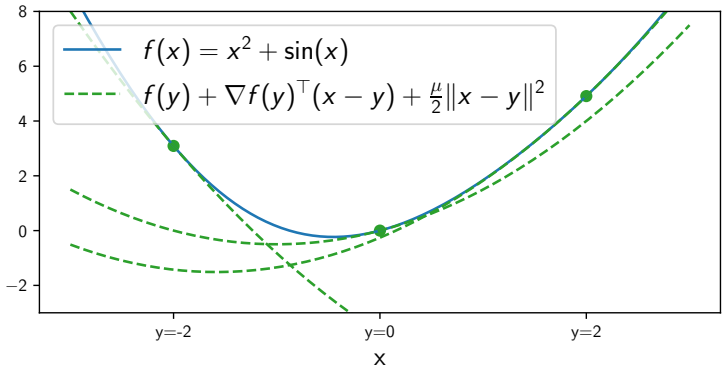

Figure 5: The function $f : x \mapsto x^2 + \sin(x)$ (in blue) is 1-strongly convex since $\nabla^2 f(x) = f''(x) = 2 + \cos(x) \geq 1$ (Proposition 4.10). It can therefore be lower bounded by tangent parabolas of same curvature μ, here displayed in dashed green for three different values of y.

Definition 4.12. *For $L \geq 0$, a differentiable function is said to be L-smooth if and only if its gradient is L-Lipschitz:*

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \tag{44}$$

Proposition 4.13 (Descent lemma)**.** *Let f be a L-smooth function. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \ . \tag{45}$$

As for strong convexity, think of this inequality for a $\mathbf{y}$ fixed, and involving two functions of $\mathbf{x}$, f and $\phi_y(\mathbf{x}) := f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$. The function ϕ_y is a convex, isotropic parabola; it is tangent to f at $\mathbf{y}$ since $\phi_y(\mathbf{y}) = f(\mathbf{y})$ and $\nabla \phi_y(\mathbf{y}) = \nabla f(\mathbf{y})$ (see TD for a proof). Equation (45) says that for every choice of $\mathbf{y}$, the tangent parabola ϕ_y globally upper bounds f. See Figure 6 for an illustration.

Proof. Notice that:

$$f(\mathbf{x}) - f(\mathbf{y}) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))\, \mathrm{d}t = \int_0^1 \langle \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle \, \mathrm{d}t \ . \tag{46}$$

Subtract $\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \int_0^1 \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \, \mathrm{d}t$ on both sides:

$$f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle = \int_0^1 \langle \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \, \mathrm{d}t \ . \tag{47}$$

Then, using $|\int \phi| \leq \int |\phi|$, Cauchy-Schwarz inequality and the fact that ∇f is L-Lipschitz,

$$|f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| = |\int_0^1 \langle \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \, \mathrm{d}t| \tag{48}$$

$$\leq \int_0^1 |\langle \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle| \, \mathrm{d}t \tag{49}$$

$$\leq \int_0^1 \|\nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - \nabla f(\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\| \, \mathrm{d}t \tag{50}$$

$$\leq \int_0^1 L\|\mathbf{y} + t(\mathbf{x} - \mathbf{y}) - f(\mathbf{y})\| \|\mathbf{x} - \mathbf{y}\| \, \mathrm{d}t \tag{51}$$

$$= L\|\mathbf{x} - \mathbf{y}\|^2 \int_0^1 \mathrm{d}t \tag{52}$$

$$= \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{53}$$

□

Proposition 4.14 (Second order characterization of smooth functions)**.** *A twice differentiable function f is L-smooth if and only if $\nabla^2 f(\mathbf{x}) \preceq L\mathrm{Id}$ for all $\mathbf{x} \in \mathbb{R}^d$ (meaning that $L\mathrm{Id} - \nabla^2 f(\mathbf{x})$ is PSD everywhere, or equivalently the eigenvalues of the Hessian at every $\mathbf{x}$ are all smaller than L).*

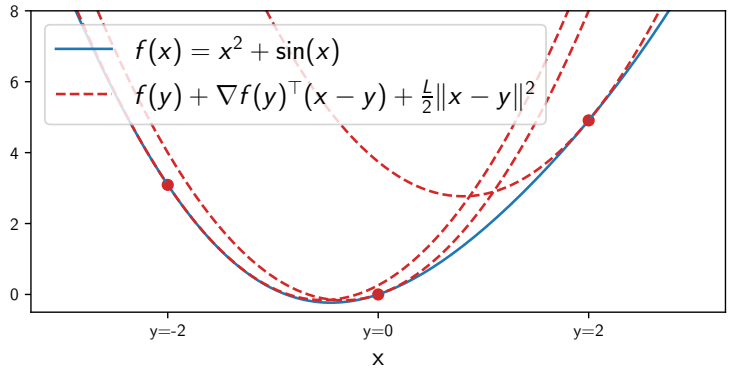

Figure 6: The function $f : x \mapsto x^2 + \sin(x)$ (in blue) is 1-strongly convex since $\nabla^2 f(x) = f''(x) = 2 + \cos(x) \leq 3$ (Proposition 4.14). It can therefore be upper bounded by tangent parabolas of same curvature L, here displayed in dashed red for three different values of y.

Example 4.15. *The OLS objective is L-smooth with $L = \|\mathbf{X}\|_2^2 = \|\mathbf{X}^\top \mathbf{X}\|_2$, where we recall that on matrices, $\|\cdot\|_2$ denotes the operator norm (Definition 3.10). There are (at least) two proofs for this:*

- *its Hessian is constant equal to $\mathbf{X}^\top \mathbf{X}$, whose largest eigenvalue is always smaller than $\lambda_{\max}(\mathbf{X}^\top \mathbf{X})$*

- *denoting f the OLS datafit, $\nabla f(\boldsymbol{\theta}) = \mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$ and so*

$$\|\nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2)\| = \|\mathbf{X}^\top\mathbf{X}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\| \leq \|\mathbf{X}^\top\mathbf{X}\|_2\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \tag{54}$$

using the basic property $\|\mathbf{M}\mathbf{x}\| \leq \|\mathbf{M}\|_2\|\mathbf{x}\|$. So the gradient of OLS is $\|\mathbf{X}^\top\mathbf{X}\|_2$-Lipschitz, which by definition means OLS is $\|\mathbf{X}^\top\mathbf{X}\|_2$-smooth.

4.3 Optimisation algorithms

Gradient descent For a stepsize $\eta > 0$, gradient descent on the function f is defined as:

$$\begin{cases} \mathbf{x}_0 \in \mathbb{R}^d \\ \mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla f(\mathbf{x}_k) \end{cases} \tag{55}$$

Since $\nabla f(\mathbf{x}_k)$ is the direction in which f decreases the fastest, this method makes sense: it moves in the direction that, locally, makes f decrease as fast as possible.

Let us illustrate the behavior of GD on a 2D example: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \begin{pmatrix} L & 0 \\ 0 & \mu \end{pmatrix} \mathbf{x} = \frac{L}{2}u^2 + \frac{\mu}{2}v^2$ – we write $\mathbf{x} = \begin{pmatrix} u \\ v \end{pmatrix}$ in this part because we will reserve $\mathbf{x}_k$ for GD iterates. The minimizer of the function is $(0, 0)$; the function is μ-strongly convex and L-smooth. The gradient of f is $\begin{pmatrix} Lu \\ \mu v \end{pmatrix}$, hence the iterations of GD are:

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix} - \eta \begin{pmatrix} Lu_k \\ \mu v_k \end{pmatrix} = \begin{pmatrix} (1 - L\eta)u_k \\ (1 - \eta\mu)v_k \end{pmatrix} = \begin{pmatrix} (1 - L\eta)^{k+1}u_0 \\ (1 - \eta\mu)^{k+1}v_0 \end{pmatrix} \tag{56}$$

We see that if we want the iterates to converge towards 0, we need to have both $|1-\eta\mu| < 1$ and $|1 - \eta L| < 1$, which is equivalent to $0 < \eta < 2/L$: what limits the stepsize is the highest value of the Hessian (in the general case, this will become the smoothness constant L).

Since for $\eta > 0$, $1 - \eta L \leq 1 - \eta\mu$, we have $\|\mathbf{x}_k\| \leq |1 - \eta\mu|^k\|\mathbf{x}_0\|$: we get very fast convergence to of the iterates to $(0, 0)$, the minimizer of the function. This is called a *convergence rate*[2]. If we take $\eta = 1/L$, the geometric factor reads $(1 - \frac{\mu}{L})$: the closer the two eigenvalues μ and L, the further away from 1 this factor, and thus the faster the convergence.

In the general case (for a L-smooth and μ-strongly convex function), you will prove in TD that GD with stepsize $1/L$ achieves a convergence rate with geometric factor $(1 - \frac{\mu}{L})$, which is a generalization of the result we have derived.

The quantity that governs this convergence speed is the condition number $\kappa = \frac{L}{\mu} \geq 1$: the higher it is, the slower convergence is. This makes sense: with smoothness and strong convexity, we can sandwich our function f between two isotropic parabolas of respective curvatures μ and L: the more these two values are far apart, the more those bounds are

[2]more precisely, a *linear* convergence rate, which is a terrible name (geometric would be better): it is linear in log scale

tight. On the contrary, if $L = \mu$, then our function f is an isotropic parabola, and GD with stepsize $1/L$ converges in one step!

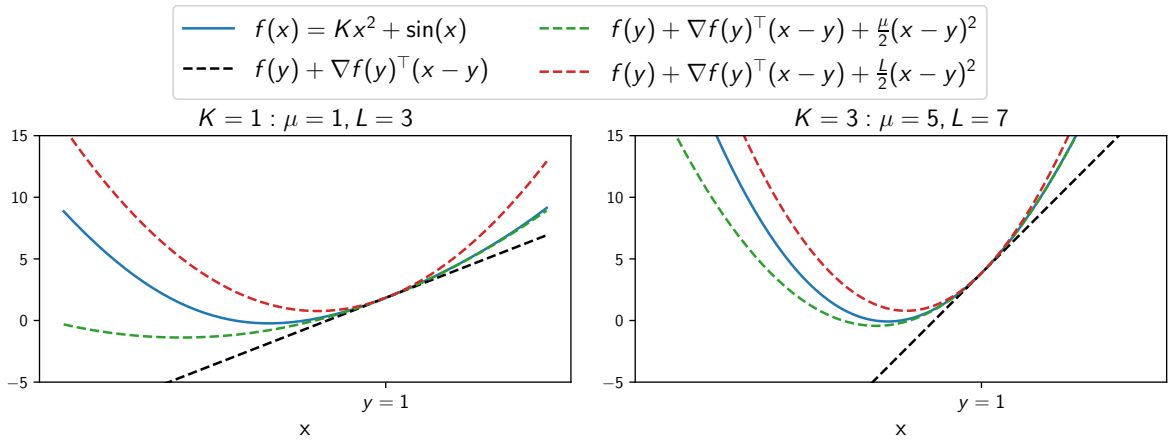

Figure 7: For $K > 0.5$, the function $f : x \mapsto Kx^2 + \sin(x)$ (in blue) is $(2K - 1)$-strongly convex and $(2K + 1)$-smooth since $\nabla^2 f(\mathbf{x}) = f''(x) = 2K + \cos(x)$ is lower bounded by $2K - 1$ and upper bounded by $2K + 1$. Left: $K = 1$, right: $K = 3$. As K increase, the condition number L/μ gets closer to 1, and the sandwiching between upper and lower bounding parabolas becomes more precise, making GD faster to converge.

As a final remark: if f is convex (not strongly) and L-smooth, if a minimizer of f exists, gradient descent with stepsize $0 < \eta < 1/L$ achieves a rate of convergence in $\mathcal{O}(1/k)$ (proof in Appendix A). This is much poorer!

Gradient descent as a majorization-minimization approach

Newton method

A Convergence rate of GD for convex smooth functions

We prove that if f is convex (not strongly) and L-smooth, if a minimizer of f exists, gradient descent with stepsize $0 < \eta < 1/L$ achieves a rate of convergence in $\mathcal{O}(1/k)$

Let $\mathbf{x}^*$ be a minimizer of f. For $k \in \mathbb{N}$, using first convexity of f, second definition of x_{k+1} and then the parallelogram identity,

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \langle \nabla f(\mathbf{x}_k), x_k - \mathbf{x}^* \rangle \tag{57}$$

$$= \frac{1}{\eta} \langle \mathbf{x}_k - x_{k+1}, x_k - \mathbf{x}^* \rangle \tag{58}$$

$$= \frac{1}{2\eta} \left(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{x}_k - x_{k+1}\|^2 \right) \tag{59}$$

$$= \frac{1}{2\eta} \left(\|\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \right) + \frac{\eta}{2} \|\nabla f(\mathbf{x}_k)\|^2 . \tag{60}$$

Summing from $k = 0$ to $t - 1$, we obtain by telescoping cancellation:

$$\sum_{k=0}^{t-1} [f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \frac{1}{2\eta} \left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2 \right) + \frac{\eta}{2} \sum_{k=0}^{t-1} \|\nabla f(\mathbf{x}_k)\|^2 \tag{61}$$

$$\leq \frac{1}{2\eta} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\eta}{2} \sum_{k=0}^{t-1} \|\nabla f(\mathbf{x}_k)\|^2 . \tag{62}$$

To conclude, we use that, since f is L-smooth, $f(\mathbf{x}^*) - f(\mathbf{x}) \leq -\frac{1}{2L} \|\nabla f(\mathbf{x})\|^2$ (inequality obtained by minimizing both sides of the descent lemma, as we do to prove the Polyak-Łojasiewicz inequality from strong convexity in TD), meaning that $\|\nabla f(\mathbf{x})\|^2 \leq \frac{L}{2}(f(\mathbf{x}) - f(\mathbf{x}^*))$. From Equation (62) we get:

$$\sum_{k=0}^{t-1} [f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \frac{1}{2\eta} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \eta L \sum_{k=0}^{t-1} (f(\mathbf{x}_k) - f(\mathbf{x}^*)) \tag{63}$$

$$(1 - \eta L) \sum_{k=0}^{t-1} [f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \frac{1}{2\eta} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \tag{64}$$

Using the descent lemma (Proposition 4.13) with $\mathbf{x} = \mathbf{x}_{k+1}$ and $\mathbf{y} = \mathbf{x}_k$ yields $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \eta(\frac{\eta}{2L} - 1)\|\nabla f(\mathbf{x}_k)\|^2$, meaning that $f(\mathbf{x}_k)$ decreases provided $0 < \eta < 2/L$, and so the LHS in (64) is larger than $t(1 - \eta L)(f(\mathbf{x}_t) - f(\mathbf{x}^*))$.