Convex Optimization

Vincent Divol

1 Convexification of the $0-1$ loss

In the previous chapter, we studied in detail the properties of the empirical risk minimizer $\hat{f}_{\mathcal{F}}$ in binary classification. This estimator is defined as the minimizer of the functional

$$f \in \mathcal{F} \mapsto \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\mathbf{y_i} \neq f(\mathbf{x_i})\}. \tag{1}$$

This functional is highly discontinuous, and is actually piecewise constant on each output $A(z_1,\dots,z_n) := \{f \in \mathcal{F},\ \forall i = 1,\dots,n,\ f(\mathbf{x_i}) = z_i\}$ for $z_1,\dots,z_n \in \{-1,1\}$. The number of such sets is exactly $\mathcal{N}_{\mathcal{F}}(\mathbf{x_1},\dots,\mathbf{x_n})$. If we believe that Sauer's lemma is tight (and it is in many cases), then this number is exponential in the VC dimension. Even for very simple sets, such as the class of linear classifiers, this number will be very large even for moderate dimension of the input space $\mathcal{X}$. This computational blow up shows the impossibility of minimizing (1) in many situations.

To overcome this problem, we make a simple remark. The set $\{-1,1\}$ is a subset of $\mathbb{R}$. Therefore, we can consider the classification task as an instance of a regression task, choose a predictor $g : \mathcal{X} \to \mathbb{R}$, and obtain a classifier by letting $f = \mathrm{sgn} \circ g$, where sgn is the sign function (equal to $+1$ on $[0,+\infty)$ and to -1 on $(-\infty,0)$). Minimizing the $0-1$ risk of the classifier $f = \mathrm{sgn} \circ g$ amounts to minimizing the function

$$g \mapsto \mathbb{E}_P[\mathbf{1}\{\mathrm{sgn}(g(\mathbf{x})) \neq \mathbf{y}\}] = \mathbb{E}_P[\mathbf{1}\{g(\mathbf{x})\mathbf{y} < 0\}]. \tag{2}$$

The loss function $\ell(y,y') = \mathbf{1}\{yy' < 0\}$ is still not continuous, making the minimization of the empirical risk as difficult as before. However, this loss can be replaced by other convex losses that are similar to this one. This

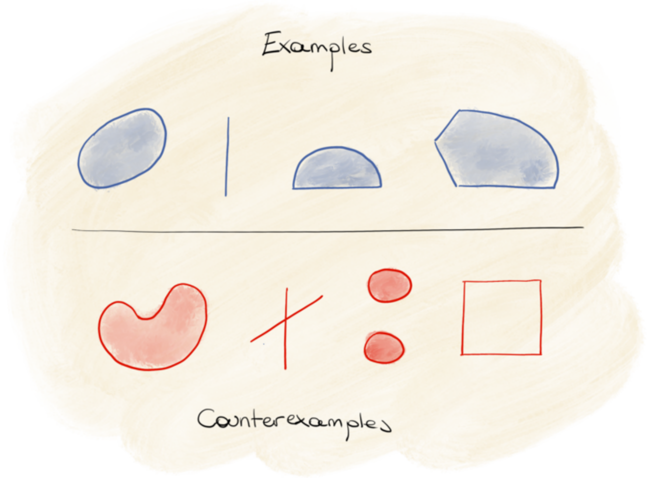


Figure 1: Examples and counterexamples of convex sets.

process is referred to as the convexification of the loss. We will come back to this problem in Section 5 after introducing the necessary background on convex optimization.

2 Convex functions

We start by recalling some elementary definitions.

Definition 2.1 (Convex set)**.** *A set $A \subset \mathbb{R}^d$ is* ***convex*** *if, given two points x and y in A, the segment joining x and y is included in A:*

$$\forall x, y \in A,\ \forall t \in [0,1],\ tx + (1-t)y \in A. \tag{3}$$

Let $A \subset \mathbb{R}^D$ and let f be a function defined on A. We define the **epigraph** of f as

$$\operatorname{epi}(f) := \{(x,t) \in A \times \mathbb{R} :\ f(x) \geq t\}. \tag{4}$$

Definition 2.2 (Convex function)**.** *Let A be a convex set and let $f : A \to \mathbb{R}$. We say that f is* ***convex*** *if* $\operatorname{epi}(f) \subset \mathbb{R}^{d+1}$ *is convex. Equivalently, f is convex if*

$$\forall x, y \in A,\ \forall t \in [0,1],\ f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{5}$$

We call A the domain of f, denoted by $\operatorname{dom}(f)$.

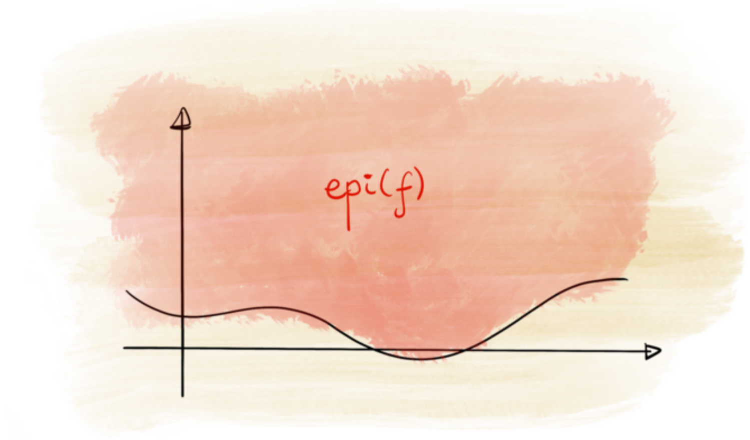


Figure 2: The epigraph of a function f (whose graph is displayed in black).

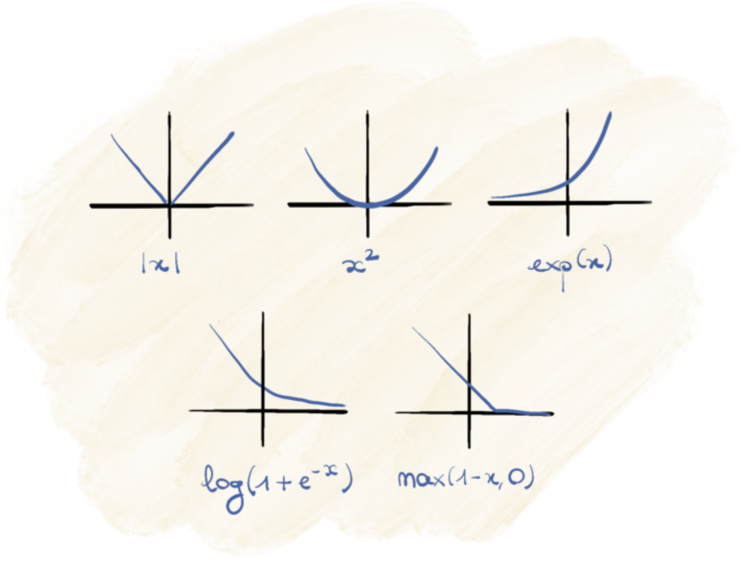


Figure 3: Examples of convex functions.

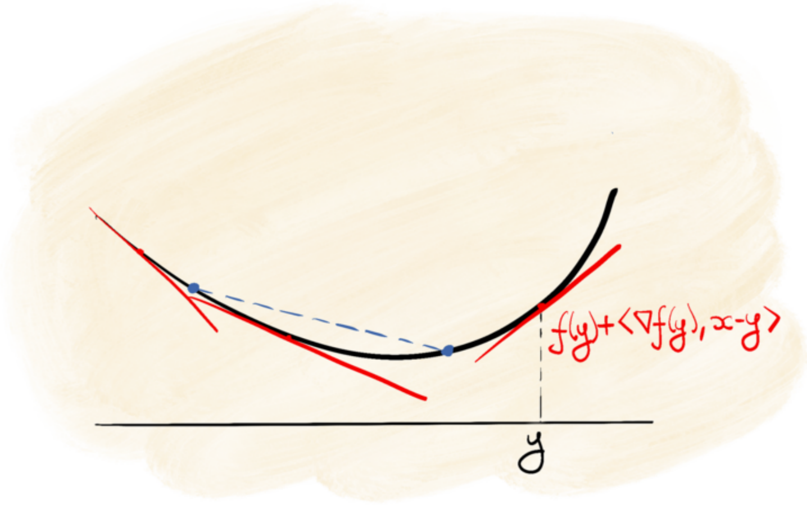


Figure 4: Blue: the chord joining two points of the graph of a convex function is above the graph. Red: the graph of a function f stays above its tangent lines.

Intuitively speaking, a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if, when we let a bead rolls from a point on the graph of f, the bead always arrives to the minimum of f (that is the point of lowest altitude). Alternatively, a function f is convex if a chord joining two points on the graph of f is always ”above” the graph of f.

A convex function is always continuous on the interior of its domain. However, it may not be differentiable (take $f : x \mapsto |x|$). If we assume that it is differentiable, then the differential has to be monotone.

Proposition 2.3. *Let f be a convex differentiable function.*

- *If $d = 1$, then f' is nondecreasing.*
- *If $d \geq 2$, then, for all $x, y \in \mathrm{dom}(f)$, we have*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0 \tag{6}$$

and

$$f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle. \tag{7}$$

Geometrically, this means that a function is convex if its graph is always above its tangent curves (see Figure 2).

What can we say about the second derivative of f (should it exist)? Using the previous proposition, for $d = 1$, we should have $f'' \geq 0$. The analogue of the second derivative for $d \geq 2$ is the Hessian matrix $\nabla^2 f$. The good notion of nonnegativity for symmetric matrices is given by the following condition: for every $x \in \mathrm{dom}(f)$ and $u \in \mathbb{R}^d$, we have $u^\top \nabla^2 f(x) u \geq 0$. We then say that $\nabla^2 f(x)$ is positive semi-definite and we write $\nabla^2 f(x) \succcurlyeq 0$. Equivalently, the eigenvalues of $\nabla^2 f(x)$ are nonnegative. If all the eigenvalues are larger than some value α, then we write $\nabla^2 f(x) \succcurlyeq \alpha \mathrm{Id}$. Likewise, if all the eigenvalues are smaller than some number β, then we write $\nabla^2 f(x) \preccurlyeq \beta \mathrm{Id}$. These are respectively equivalent to having $u^\top \nabla^2 f(x) u \geq \alpha \|u\|^2$ and $u^\top \nabla^2 f(x) u \leq \beta \|u\|^2$ for all $u \in \mathbb{R}^d$.

Proposition 2.4. *Assume that f is twice differentiable.*

- *If $d = 1$, then $f'' \geq 0$.*
- *If $d \geq 2$, the Hessian matrix satisfies $\nabla^2 f(x) \succcurlyeq 0$ for every $x \in \mathrm{dom}(f)$.*

Theoretical guarantees for the optimization algorithms presented in the next sections will hold if the convex function f is sufficiently well-behaved. A key ingredient is to assert that f is ”really convex” in a quantitative way.

Definition 2.5. *A real valued convex function f is said to be α-strongly convex if for all $x, y \in \mathrm{dom}(f)$ and $t \in [0,1]$,*

$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) - \frac{\alpha}{2} t(1-t) \|x - y\|^2. \tag{8}$$

For example, a linear function $f : x \mapsto \langle a, x \rangle$ is not strongly convex, as we have the identity $f(tx + (1-t)y) = t f(x) + (1-t) f(y)$ for such a function.

Proposition 2.6. *Assume that f is twice differentiable. The following conditions are equivalent.*

- *The function f is α-strongly convex.*
- *For all $x, y \in \mathrm{dom}(f)$,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|x - y\|^2. \tag{9}$$

- *The function* $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ *is convex.*
- *We have* $\nabla^2 f(x) \succcurlyeq \alpha \mathrm{Id}$ *for every* $x \in \mathrm{dom}(f)$.

Geometrically speaking, a function is α-strongly convex if its graph is sufficiently curved at every point (more precisely, it has curvature at least $1/\alpha$ in all directions). The prototypical example of α-strongly convex function is the function $x \mapsto \alpha\|x\|^2/2$ (this is clear with the third characterization). Note that strong convexity implies that f has a unique minimizer $x^\star$. Indeed, take x far away from 0 and apply (8) to $t = 1/\|x\|$ and $y = 0$. We obtain

$$f(x) \geq \|x\| \left(f(x/\|x\|) + \frac{\alpha}{2}\frac{1}{\|x\|}\left(1 - \frac{1}{\|x\|}\right)\|x\|^2 \right). \tag{10}$$

In particular, $f(x)$ goes to infinity as $\|x\|$ goes to infinity. This implies that the infimum of f is attained on some sufficiently large ball, and by continuity of f the infimum is attained at at least one point. Given two minimizers x_1 and x_2, we have, for any $t \in (0,1)$,

$$\min f \leq f(tx_1 + (1-t)x_2) \leq t \min f + (1-t)\min f - \frac{\alpha}{2}t(1-t)\|x_1 - x_2\|^2. \tag{11}$$

This inequality is possible only if $x_1 = x_2$, implying the uniqueness of the minimizer.

Furthermore, α-strongly convex implies the *Polyak–Lojasiewicz condition* (or PL condition). The PL condition asserts that the function f cannot grow too fast near its minimizer $x^\star$: for all $x \in \mathrm{dom}(f)$,

$$f(x) - f(x^\star) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2. \tag{12}$$

A second condition that ensures good convergence properties is the regularity of the gradient of f.

Definition 2.7. *Let f be a real-valued function. We say that f is β-smooth if it is differentiable and its gradient ∇f is β-Lipschitz:*

$$\forall x, y \in \mathrm{dom}(f),\ \|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|. \tag{13}$$

Proposition 2.8. *Assume that f is twice differentiable. The following conditions are equivalent.*

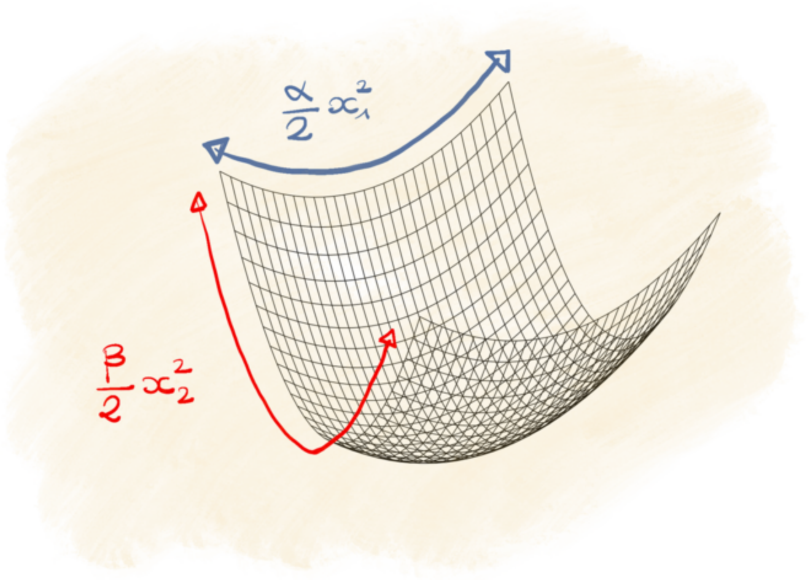


Figure 5: The function $f_{\alpha,\beta}$ is α-strongly convex and β-smooth.

- *The function f is β-smooth.*
- *For all $x, y \in \mathrm{dom}(f)$,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|x - y\|^2. \tag{14}$$

- *We have $\nabla^2 f(x) \preccurlyeq \beta \mathrm{Id}$ for every $x \in \mathrm{dom}(f)$.*

Geometrically, the β-smoothness condition implies that the graph of the function is not too curved. The prototypical example of a function that is both α-strongly convex and β-smooth is the quadratic function

$$f_{\alpha,\beta} : x \in \mathbb{R}^2 \mapsto \frac{\alpha}{2}x_1^2 + \frac{\beta}{2}x_2^2. \tag{15}$$

The graph of this function has an elongated bowl shape, with large width $1/\alpha$ in direction x_1, and small width $1/\beta$ in direction x_2. The Hessian of the function is given by

$$\nabla^2 f_{\alpha,\beta}(x) = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}, \tag{16}$$

showing that it is indeed α-strongly convex and β-smooth, see Figure 2.

3 Gradient descent

We now investigate the problem of finding the minimum of a convex function f. The most important algorithm to find such a minimum is the gradient descent algorithm. The general idea is simple: to find the minimum of a convex function, start at point x, take one step in the direction with largest negative slope, and repeat the procedure.

Definition 3.1. *Let f be a real-valued function and let $x_0 \in \mathbb{R}^d$. Let $T \in \mathbb{N}$ be a number of steps and let $(s_t)_{t=0,\dots,T}$ be a sequence of step sizes. The iterates of the gradient descent on f are defined by*

$$x_{t+1} := x_t - s_t \nabla f(x_t) \tag{17}$$

for $t \in \{0, \dots, T-1\}$.

We are now in position to state the main result of this chapter: the iterates of a gradient descent on a smooth and strongly-convex function converge towards the minimizer of f at a fast rate.

Proposition 3.2 (Convergence of gradient descent for smooth and strongly convex functions)**.** *Let f be a α-strongly convex and β-smooth function defined on $\mathbb{R}^d$. Consider the iterates $(x_t)_{t=0,\dots,T}$ of the gradient descent with constant step-size $s \leq 1/\beta$ and initialization x^0. Let $x^\star$ be the minimizer of f. Then, we have*

$$\begin{aligned} f(x^T) - f(x^\star) &\leq (1-\alpha s)^T (f(x^0) - f(x^\star)) \\ &\leq \exp(-\alpha s T)(f(x^0) - f(x^\star)). \end{aligned} \tag{18}$$

We refer to such a rate of convergence as a **linear** rate of convergence as $\log(f(x^T) - f(x^\star))$ converges at a linear rate to $-\infty$.

Proof. The characterization of β-smoothness (14) implies that

$$\begin{aligned} f(x^{t+1}) = f(x^t - s\nabla f(x^t)) &\leq f(x^t) + \langle \nabla f(x^t), -s\nabla f(x^t)\rangle + s^2 \frac{\beta}{2} \|\nabla f(x^t)\|^2 \\ &= f(x^t) + s\left(\frac{s\beta}{2} - 1\right) \|\nabla f(x^t)\|^2. \end{aligned} \tag{19}$$

Furthermore, by the PL condition (12), we have

$$\|\nabla f(x^t)\|^2 \geq 2\alpha(f(x^t) - f(x^\star)). \tag{20}$$

Plugging in this inequality in (19) yields

$$f(x^{t+1}) - f(x^\star) \leq (f(x^t) - f(x^\star)) \left(1 + 2\alpha s \left(\frac{s\beta}{2} - 1\right)\right). \tag{21}$$

Having $s \leq 1/\beta$ implies that $2\alpha s \left(\frac{s\beta}{2} - 1\right) \leq -\alpha s$. Iterating this inequality, we obtain the conclusion. □

If we choose the step size s as the largest value allowed in the proposition, that is $s = 1/\beta$, we see that the linear rate of convergence is exactly equal to $\kappa^{-1} = \alpha/\beta$. The quantity $\kappa = \beta/\alpha$ is called the **condition number** of f. This corresponds to an upper bound on the ratio between the largest eigenvalue of the Hessian of f and its smallest. The dependence in κ in Proposition 3.2 indicates that functions f with a large condition number κ are harder to minimize.

More often, we want to control the number of iterations of the gradient descent needed to find a point x with $f(x) - f(x^\star) \leq \varepsilon$ for some fixed $\varepsilon > 0$. Proposition 3.2 implies that $T = \log(\varepsilon^{-1})/\kappa$ operations are needed. When minimizing the empirical risk, we want to compute a predictor that performs almost as well as the empirical risk minimizer $\hat{f}_{\mathcal{F}}$. Considerations in the previous chapter shows that one can expect the excess risk of $\hat{f}_{\mathcal{F}}$ to be at least of order $1/\sqrt{n}$. Therefore, taking ε of order $1/\sqrt{n}$ will most of the time be enough in our setting. This means that roughly $\log(n)/\kappa$ iterations are needed.

It is possible to relax some assumptions in the previous convergence result. For instance, if f is not strongly convex, one still has convergence of the iterates of the gradient descent, although at a much slower rate.

Proposition 3.3 (Convergence of gradient descent for smooth functions)**.** *Let f be a convex β-smooth function defined on $\mathbb{R}^d$. Consider the iterates $(x_t)_{t=0,\dots,T}$ of the gradient descent with constant step-size $s = 1/\beta$ and initialization x^0. Assume that f has a global minimizer $x^\star$. Then, we have*

$$f(x^T) - f(x^\star) \leq \frac{\beta\|x^0 - x^\star\|^2}{2T} \tag{22}$$

The dependence is not exponential in T anymore, but only polynomial. In particular, without the strong convexity assumption, $O(\beta/\varepsilon)$ iterations are needed, a number that is polynomial (and not logarithmic) in ε^{-1}. For $\varepsilon = 1/\sqrt{n}$, this leads to $O(\beta\sqrt{n})$ iterations. The proof of Proposition 3.3 is more delicate than the previous one and we do not include it. It can be found in [Bansal and Gupta, 2019, Theorem 3.3].

Consider a smooth convex function f that is strongly convex only on a neighborhood of its minimizer $x^\star$. Proposition 3.3 implies that gradient descent converges in a small number of steps to a point in this neighborhood. Then, gradient descent will linearly converge to the minimizer. Therefore, gradient descent will naturally adapt to the degree of convexity of the function, without the need to design any complicated procedure to choose the size step: we refer to this phenomenon as gradient descent being adaptive.

4 Newton's method

The gradient descent relies on a first order approximation of f: $f(x_0 + h)$ locally looks like $f(x_0) + \langle \nabla f(x_0), h\rangle$, so to decrease f, we should make a step in the direction minimizing the quantity $\langle \nabla f(x_0), h\rangle$ (and this direction is given by $-\nabla f(x_0)$). What if we try to write a second-order approximation of f around x_0? This yields to a second-order method called **Newton's method**. We have

$$f(x_0 + h) \approx P_{f,x_0}(h) = f(x_0) + \langle \nabla f(x_0), h\rangle + \frac{1}{2} h^T \nabla^2 f(x_0) h. \tag{23}$$

Let us find the minimum of the quadratic function P_{f,x_0}. Its gradient is equal to $\nabla f(x_0) + \nabla^2 f(x_0)h$. Therefore, should the Hessian be invertible at x_0 (for instance if f is strongly convex), then the minimum of P_{f,x_0} is attained at $h = (\nabla^2 f(x_0))^{-1} \nabla f(x_0)$.

Definition 4.1. *Let f be a real-valued function and let $x_0 \in \mathbb{R}^d$. Assume that the Hessian of f is always invertible. Let $T \in \mathbb{N}$ be a number of steps. A step of Newton's method is given by*

$$x^{t+1} := x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t), \tag{24}$$

for $t \in \{0, \dots, T-1\}$.

To obtain convergence guarantees on Newton's method, we will need Lipschitz continuity of the Hessian matrix $\nabla^2 f$.

Proposition 4.2. *Let f be a twice differentiable real-valued convex function defined on $\mathbb{R}^d$ with $\|\nabla^2 f(x)u - \nabla^2 f(y)u\| \leq \gamma\|x-y\|\|u\|$ for every $x, y, u \in \mathbb{R}^d$. Assume further that f is α-strongly convex and β-smooth, with unique minimizer $x^\star$. Consider the iterates $(x_t)_{t=0,\dots,T}$ of the Newton's method with initialization x^0. Then, we have*

$$\|x^{t+1} - x^\star\| \leq \frac{\gamma}{2\alpha}\|x^t - x^\star\|^2. \tag{25}$$

In particular, if $\|x^0 - x^\star\| \leq \alpha/\gamma$, then we have

$$f(x^T) - f(x^\star) \leq \frac{\beta\alpha}{\gamma} 2^{-2^{T+1}}. \tag{26}$$

Proof. The first step consists in rewriting the update:

$$\begin{aligned} x^{t+1} - x^\star &= x^t - x^\star - (\nabla^2 f(x^t))^{-1}\nabla f(x^t) \\ &= x^t - x^\star - (\nabla^2 f(x^t))^{-1}\int_0^1 \nabla^2 f(x^\star + \theta(x^t - x^\star))(x^t - x^\star)\mathrm{d}\theta \\ &= (\nabla^2 f(x^t))^{-1}(\nabla^2 f(x^t))(x^t - x^\star) \\ &\qquad - (\nabla^2 f(x^t))^{-1}\int_0^1 \nabla^2 f(x^\star + \theta(x^t - x^\star))(x^t - x^\star)\mathrm{d}\theta \\ &= (\nabla^2 f(x^t))^{-1} G^t (x^t - x^\star), \end{aligned}$$

where $G^t = \int_0^1 (\nabla^2 f(x^t) - \nabla^2 f(x^\star + \theta(x^t - x^\star))\mathrm{d}\theta$. The operator norm of G^t is controlled:

$$\begin{aligned} \left\|G^t\right\|_{\mathrm{op}} &\leq \int_0^1 \left\|\nabla^2 f(x^t) - \nabla^2 f(x^\star + \theta(x^t - x^\star)\right\|_{\mathrm{op}} \mathrm{d}\theta \\ &\leq \gamma \int_0^1 (1-\theta)\|x^t - x^\star\|\mathrm{d}\theta \leq \frac{\gamma}{2}\|x^t - x^\star\|. \end{aligned}$$

Therefore, we obtain

$$\|x^{t+1} - x^\star\| \leq \left\|(\nabla^2 f(x^t))^{-1}\right\|_{\mathrm{op}} \frac{\gamma}{2}\|x^t - x^\star\|^2 \leq \frac{\gamma}{2\alpha}\|x^t - x^\star\|^2. \tag{27}$$

To iterate this relation, remark that it can be written as

$$c\|x^{t+1} - x^\star\| \leq (c\|x^{t+1} - x^\star\|)^2, \tag{28}$$

where $c = \frac{\gamma}{2\alpha}$. This yields

$$c\|x^T - x^\star\| \leq (c\|x^0 - x^\star\|)^{2^T} \leq 2^{-2^T}, \tag{29}$$

where we use the condition $\|x^0 - x^\star\| \leq \alpha/\gamma$. Eventually, β-smoothness in $x^\star$ implies that

$$\begin{aligned} f(x^T) - f(x^\star) &\leq \langle \nabla f(x^\star), x^T - x^\star \rangle + \frac{\beta}{2}\|x^T - x^\star\|^2 \\ &= \frac{\beta}{2}\|x^T - x^\star\|^2 \leq \frac{\beta\alpha}{\gamma} 2^{-2^{T+1}}. \end{aligned} \tag{30}$$

□

Newton's method converges much faster than gradient descent. We refer to this behavior as a quadratic convergence (because (25) states that each iterate is quadratically closer to the minimum than the previous one). Only $O(\frac{\log\log\varepsilon^{-1}}{\beta\alpha/\gamma})$ iterations are needed to obtain an error ε. In the empirical risk minimization context with n observations, this translates to roughly $\log\log n$ iterations. However, each iteration requires to compute the inverse of the Hessian matrix $\nabla^2 f(x)$. In dimension d, this takes $O(d^3)$ operations using Gauss-Jordan elimination. If d is large (and it is in many applications!), then this cost is prohibitive. Note however that methods that are *much* smarter than Gauss-Jordan elimination are used in practice to compute one step of Newton's method. Another drawback of Newton's method is that it will totally break down should f not be strongly convex. On the opposite, even without strong convexity, we still have convergence guarantees (although slower) for gradient descent (Proposition 3.3).

5 Logistic regression

Recall the setting of Section 1. We consider the classification problem, and consider classifiers of the form $x \mapsto \mathrm{sgn}(g(x))$ where $g : \mathcal{X} \to \mathbb{R}$ is some function. The loss that appears in this context is given by $\ell_{01}(g(x), y) =$

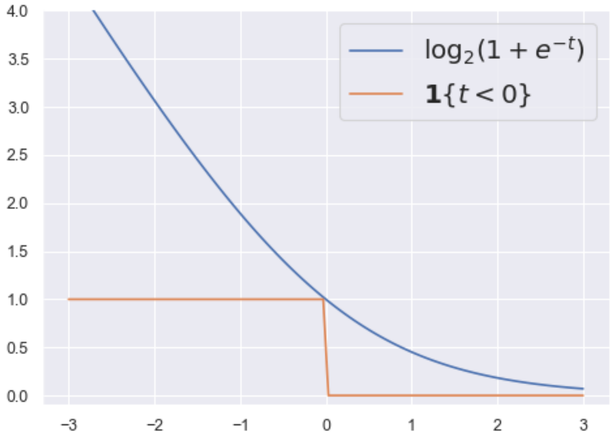


Figure 6: The logistic loss is a convexification of the $0-1$ loss.

$\mathbf{1}\{g(x)y<0\}$. Given n observations $(\mathbf{x_1},\mathbf{y_1}),\ldots,(\mathbf{x_n},\mathbf{y_n})$, the empirical risk is equal to

$$g \mapsto \mathcal{R}_n(g) = \frac{1}{n}\sum_{i=1}^{n} \ell_{01}(g(\mathbf{x_i}),\mathbf{y_i}). \tag{31}$$

A powerful method to find a good classifier g consists in replacing the ℓ_{01} loss by a convex loss function that, hopefully, will lead to similar behaviors. There are a lot of different choices that can act as a surrogate for the ℓ_{01} loss. A popular one is the logistic loss

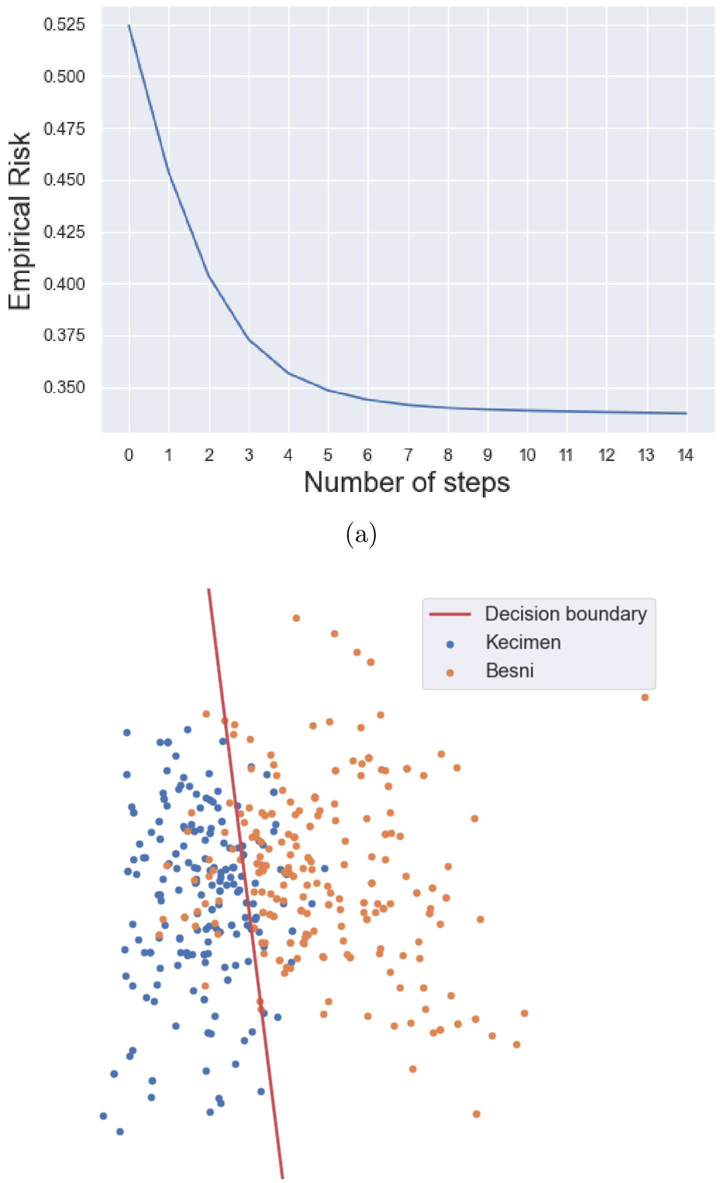
$$\ell_{\log}(y,y') = \log(1+\exp(-yy')) = -\log(\sigma(yy')), \tag{32}$$

where σ is the **sigmoid** function

$$\sigma : t \mapsto \frac{1}{1+\exp(-t)} \in [0,1], \tag{33}$$

see Figure 4. Note that σ satisfies $\sigma(-t) = 1-\sigma(t)$. The corresponding empirical risk is

$$\begin{aligned} g \mapsto \tilde{\mathcal{R}}_n(g) &= \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-\mathbf{y_i}g(\mathbf{x_i}))) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{z_i}\log(1+\exp(-g(\mathbf{x_i}))) \\ &\qquad + (1-\mathbf{z_i})\log(1+\exp(g(\mathbf{x_i}))), \end{aligned} \tag{34}$$



(a)

(b)

Figure 7: (a) Empirical risk at different steps of the gradient descent in the logistic regression model with linear classifiers. (b) Decision boundary of the linear classifier, in the plane given by the two principal components of the dataset.

where $\mathbf{z_i} = \mathbf{1}\{\mathbf{y_i} = 1\}$. Note that this function is convex in g[1].

Link with maximum likelihood estimation

Let $\mathcal{G}$ be a class of real-valued functions defined on $\mathcal{X}$ (for instance $\mathcal{G}$ is the set of linear functions). We consider the following modelization. Assume that there exists $g \in \mathcal{G}$ such that $(\mathbf{x}, \mathbf{y})$ is obtained by sampling $\mathbf{x}$ according to $P_{\mathbf{x}}$, and then letting $\mathbf{y} = 1$ with probability $\sigma(g(\mathbf{x}))$, and $\mathbf{y} = -1$ otherwise.

The likelihood of a set of observations $(\mathbf{x_1}, \mathbf{y_1}), \dots, (\mathbf{x_n}, \mathbf{y_n})$ at $g \in \mathcal{G}$ is given by

$$\prod_{i=1}^{n} \sigma(g(\mathbf{x_i}))^{\mathbf{z_i}} (1 - \sigma(g(\mathbf{x_i})))^{1-\mathbf{z_i}}, \tag{35}$$

where once again $\mathbf{z_i} = \mathbf{1}\{\mathbf{y_i} = 1\}$. The log-likelihood is equal to

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbf{z_i} \log(\sigma(g(\mathbf{x_i}))) + (1 - \mathbf{z_i}) \log(1 - \sigma(g(\mathbf{x_i}))) \\
&= \sum_{i=1}^{n} \mathbf{z_i} \log(\sigma(g(\mathbf{x_i}))) + (1 - \mathbf{z_i}) \log(\sigma(-g(\mathbf{x_i}))) \\
&= -\sum_{i=1}^{n} \mathbf{z_i} \log(1 + \exp(-g(\mathbf{x_i}))) + (1 - \mathbf{z_i}) \log(1 + \exp(g(\mathbf{x_i}))) \\
&= -\tilde{\mathcal{R}}_n(g).
\end{aligned} \tag{36}$$

Therefore, maximizing the log-likelihood is equivalent to minimizing the empirical risk $\tilde{\mathcal{R}}_n$. In particular, the empirical risk minimizer $\hat{g}_{\mathcal{G}}$ is a maximum likelihood estimator! Maximum likelihood esitmators are known to satisfy strong theoretical properties (consistency, asymptotic normality, etc.), justifying the use of the logistic loss as a surrogate for the $0 - 1$ loss.

Example 5.1. In [Çinar et al., 2020], the authors use image processing techniques to extract eight relevant geometric features from two different variety of raisins (Kecimen and Besni). Each grain is then described by a vector in $\mathbb{R}^d$ for $d = 8$ corresponding to those different features (area, perimeter, eccentricity, etc.). We consider the set of affine classifiers $\mathcal{G}_{\mathrm{aff}}$ containing functions

[1] We have defined what it means to be convex for functions defined on $\mathbb{R}^d$. However, the definition (5) can be used as a definition for a convex functionial defined on a space of functions.

g of the form $x \mapsto \theta^\top x + b$ for some $\theta \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Given $n = 450$ grains, we implement a logistic regression to classify the grains into the two varieties. Ten steps of gradient descent are enough to reach the minimizer of the empirical risk $\tilde{R}_n$. We display in Figure 7 the classifier that was obtained (in the plane given by a PCA on the dataset). The accuracy of the classifier is tested on a test dataset, and is equal to 85%.

References

[Bansal and Gupta, 2019] Bansal, N. and Gupta, A. (2019). Potential-function proofs for gradient methods. *Theory of Computing*, 15(1):1–32.

[Çinar et al., 2020] Çinar, İ., Koklu, M., and Taşdemir, Ş. (2020). Classification of raisin grains using machine vision and artificial intelligence methods. *Gazi Mühendislik Bilimleri Dergisi (GMBD)*, 6(3):200–209.