Algorithmic Foundations of Learning **Lecture 8**

Convex Loss Surrogates. Elements of Convex Theory

Lecturer: Patrick Rebeschini *Version: November 3, 2022*

8.1 Introduction

Let us recall the setting of binary classification:

- Training data $Z_1, \dots, Z_n$ such that $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$;
- Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\}$;
- Loss function $\ell(a, (x, y)) = \phi(a(x), y)$, for $\phi : \{-1, 1\}^2 \to \mathbb{R}_+$.

In the previous lectures we developed results to understand the *statistical* behaviour of the empirical risk minimizer $A^\star$ in the case of the "true" zero-one loss function $\phi(\hat{y}, y) = \mathbf{1}_{\hat{y} \neq y}$. In this case we have, for any $a \in \mathcal{B}$,

$$r(a) = \mathbf{P}(a(X) \neq Y), \qquad R(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{a(X_i) \neq Y_i}.$$

Recall the definitions (we assume minima are attained):

$$a^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} r(a), \qquad a^{\star\star} \in \operatorname*{argmin}_{a \in \mathcal{B}} r(a), \qquad A^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} R(a).$$

Putting together the results in Theorem 6.13 (note that the loss function is bounded in the interval $[0, 1]$, so we can choose $c = 1$), Proposition 4.1, and Theorem 5.6, we proved that for a general class $\mathcal{A}$, possibly with infinitely many classifiers $|\mathcal{A}| = \infty$, we have (constants are not optimized):

$$\mathbf{P}\left(r(A^\star) - r(a^\star) < 27\sqrt{\frac{\mathtt{VC}(\mathcal{A})}{n}} + \sqrt{2\frac{\log(1/\delta)}{n}} \right) \geq 1 - \delta.$$

In the last lecture we also saw how to obtain fast rates in the case when $|\mathcal{A}| < \infty$ and $a^{\star\star} \in \mathcal{A}$.

The analysis that we have developed so far is purely statistical in nature. As such, it does not take into account the *computational* resources (or constraints) at our disposal. In particular, in the analysis of the empirical risk minimization we have implicitly assumed that we have indeed access to a minimzer of the empirical risk $A^\star$. It turns out that computing $A^\star$ with the true loss function is, in general, a NP-hard problem, hence outside of our reach (the computational complexity scales exponentially with the parameters of the problem). It is therefore of interest to approximate the original problem by another problem that is more amenable to computations, and to investigate the price that we pay from a statistical point of view for this approximation. This is what we will achieve today, considering convex relaxations of the original problem.

8.2 Convex Relaxations

To avoid the computational burden associated with the empirical risk minimizer and the true loss function, the main idea that we will explore consists in replacing the original problem with the minimization of a convex upper bound of the true loss over a convex set of hypotheses.

We recall the basic definitions of convexity for functions and sets.

Definition 8.1 (Convex function) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is* convex *if for every $x, \tilde{x} \in \mathbb{R}^d, \lambda \in [0,1]$ we have*

$$\boxed{f(\lambda x + (1-\lambda)\tilde{x}) \leq \lambda f(x) + (1-\lambda) f(\tilde{x})}$$

Definition 8.2 (Convex set) *A set $\mathcal{A}$ is* convex *if for every $a, \tilde{a} \in \mathcal{A}, \lambda \in [0,1]$ we have*

$$\boxed{\lambda a + (1-\lambda)\tilde{a} \in \mathcal{A}}$$

To implement our agenda, we first define the convex upper bounds of the true loss function that we will consider. Note that the true loss function is of the form $\phi(\hat{y}, y) = \varphi^\star(\hat{y}y)$ for a function $\varphi^\star : \mathbb{R} \to \mathbb{R}_+$ defined as $\varphi^\star(u) := \mathbf{1}_{u \leq 0}$.

Definition 8.3 (Convex loss surrogate) *A function $\varphi : \mathbb{R} \to \mathbb{R}_+$ is called a* convex loss surrogate *if it is a convex, non-increasing function such that $\varphi(0) = 1$.*

The reason for this precise definition of convex surrogate is given by Zhang's Lemma, which will state and prove below. Note that from the definition it follows that if φ is a convex loss surrogate, then it is an upper bound on the true loss: $\varphi^\star(u) \leq \varphi(u)$ for every $u \in \mathbb{R}$. The following are examples of popular convex loss surrogates.

- **Exponential loss.** $\varphi(u) = e^{-u}$.
- **Hinge loss.** $\varphi(u) = \max\{1-u, 0\}$.
- **Logistic loss.** $\varphi(u) = \log_2(1 + e^{-u})$.

The second step to get a convex problem is to choose a convex set of classifiers. To achieve this, we consider the family of so-called *soft* classifiers $\mathcal{B}_{\text{soft}} := \{a : \mathbb{R}^d \to \mathbb{R}\}$ that outputs elements in $\mathbb{R}$ instead of elements in $\{-1, 1\}$. For a given convex loss surrogate φ, if the subset $\mathcal{A}_{\text{soft}} \subseteq \mathcal{B}_{\text{soft}}$ of admissible soft-classifiers is convex, then the empirical φ-risk minimization problem defined as

$$R_\varphi(a) = \frac{1}{n}\sum_{i=1}^n \varphi(a(X_i)Y_i), \qquad A_\varphi^\star \in \operatorname*{argmin}_{a \in \mathcal{A}_{\text{soft}}} R_\varphi(a),$$

is convex. *Upon additional structural assumptions*, convex problems are amenable to computations, as we will discuss later on.

Common choices of admissible convex soft classifiers are given by the following:

- **Linear functions with convex parameter space.** $\mathcal{A}_{\text{soft}} = \{a(x) = w^\top x + b : w \in \mathcal{C}_1 \subseteq \mathbb{R}^d, b \in \mathcal{C}_2 \subseteq \mathbb{R}\}$, where $\mathcal{C}_1, \mathcal{C}_2$ are convex sets.

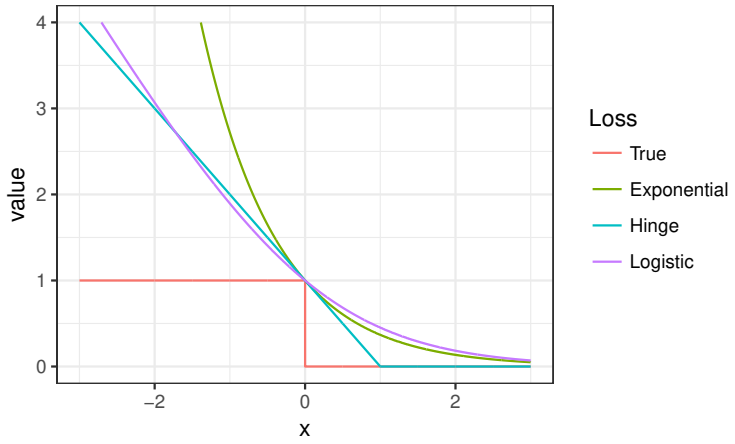

Figure 8.1: Convex loss surrogates.

- **Majority votes (Boosting).** $\mathcal{A}_{\text{soft}} = \{a(x) = \sum_{i=1}^{m} w_j h_j(x) : w = (w_1, \ldots, w_m) \in \Delta_m\}$, where Δ_m is the m-dimensional simplex and $h_1, \ldots, h_m$ are so-called *base classifiers*, functions from $\mathbb{R}^d$ to $\mathbb{R}$.

In both these two cases, it is immediate to verify that if $a, \tilde{a} \in \mathcal{A}_{\text{soft}}$, then $\lambda a + (1-\lambda)\tilde{a} \in \mathcal{A}_{\text{soft}}$ for any $\lambda \in [0,1]$, hence satisfying the assumption of a convex set.

Given a soft classifier $a \in \mathcal{A}$, the corresponding *hard* classifier is defined as it sign:

$$\text{sign}(a) \in \{-1, 1\}.$$

8.3 φ-risk Minimisation

Before addressing the question of computing $A_\varphi^\star$ in the case of convex problems (i.e., when φ is a convex surrogate and $\mathcal{A}_{\text{soft}}$ is a convex set), we focus on understanding the *statistical* relationship between the *excess φ-risk* $r_\varphi(a) - r_\varphi(a_\varphi^{\star\star})$ achieved by a soft classifier $a \in \mathcal{B}_{\text{soft}}$ with respect to a convex loss surrogate φ, where

$$r_\varphi(a) = \mathbf{E}\,\varphi(a(X)Y), \qquad a_\varphi^{\star\star} \in \operatorname*{argmin}_{a \in \mathcal{B}_{\text{soft}}} r_\varphi(a),$$

and the excess risk $r(\text{sign}(a)) - r(a^{\star\star})$ achieved by the corresponding hard classifier $\text{sign}(a)$ with respect to the true loss function $\varphi^\star(u) = \mathbf{1}_{u \leq 0}$, where we recall the definitions

$$r(a) = \mathbf{E}\,\varphi^\star(a(X)Y) = \mathbf{P}(a(X) \neq Y), \qquad a^{\star\star} \in \operatorname*{argmin}_{a \in \mathcal{B}} r(a).$$

The following result establishes a condition on the convex surrogate loss function that allows the excess φ-risk to dominate the excess risk for the original problem. This result shows that, even in the case of binary classification, one can safely consider learning problems with convex loss functions and be guaranteed not to lose much by doing so.

Remark 8.4 *Zhang's Lemma is a statement about excess risks, not about estimation errors, cf. Lecture 1. In particular, the admissible sets $\mathcal{A}$ and $\mathcal{A}_{soft}$ do not appear in the lemma, and there is no need to contemplate convexity of the admissible sets. Only convexity of the loss surrogates is needed.*

Lemma 8.5 (Zhang) *Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a convex loss surrogate. For any $\tilde{\eta} \in [0,1]$, $\tilde{a} \in \mathbb{R}$, let*

$$H_{\tilde{\eta}}(\tilde{a}) := \varphi(\tilde{a})\tilde{\eta} + \varphi(-\tilde{a})(1-\tilde{\eta}), \qquad \tau(\tilde{\eta}) := \inf_{\tilde{a}\in\mathbb{R}} H_{\tilde{\eta}}(\tilde{\alpha}).$$

Assume that there exist $c > 0$ and $\nu \in [0,1]$ such that

$$\boxed{\left|\tilde{\eta} - \frac{1}{2}\right| \le c(1-\tau(\tilde{\eta}))^\nu \qquad \text{for any } \tilde{\eta} \in [0,1]}$$

Then, for any $a : \mathbb{R}^d \to \mathbb{R}$ we have

$$\boxed{\underbrace{r(\mathrm{sign}(a)) - r(a^{\star\star})}_{\substack{\text{excess risk}\\ \text{hard classifier}}} \le 2c(\underbrace{r_\varphi(a) - r_\varphi(a_\varphi^{\star\star})}_{\substack{\text{excess } \varphi\text{-risk}\\ \text{soft classifier}}})^\nu}$$

Proof: By Example 1.8, the Bayes decision rule $a^{\star\star}$ reads

$$a^{\star\star}(x) = \operatorname*{argmin}_{\hat{y}\in\{-1,1\}} \mathbf{E}[\varphi^\star(\hat{y}Y)|X=x] = \operatorname*{argmax}_{\hat{y}\in\{-1,1\}} \mathbf{P}(Y=\hat{y}|X=x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ -1 & \text{if } \eta(x) \le 1/2 \end{cases}$$

with $\eta(x) := \mathbf{P}(Y=1|X=x)$. Using the convention

$$\mathrm{sign}(u) := \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{if } u \le 0 \end{cases}$$

we have

$$\{\mathrm{sign}(a(X)) \neq a^{\star\star}(X)\} = \{\mathrm{sign}(a(X))a^{\star\star}(X) \le 0\} \subseteq \{a(X)a^{\star\star}(X) \le 0\} = \{a(X)(\eta(X) - 1/2) \le 0\},$$

so that by Theorem 7.6 and the assumption of the lemma we find

$$\begin{aligned} r(\mathrm{sign}(a)) - r(a^{\star\star}) &= \mathbf{E}[|2\eta(X)-1|\mathbf{1}_{\mathrm{sign}(a(X))\neq a^{\star\star}(X)}] \\ &\le \mathbf{E}[|2\eta(X)-1|\mathbf{1}_{a(X)(\eta(X)-1/2)\le 0}] \\ &\le 2c\mathbf{E}[(1-\tau(\eta(X)))^\nu \mathbf{1}_{a(X)(\eta(X)-1/2)\le 0}] \\ &= 2c\mathbf{E}[((1-\tau(\eta(X)))\mathbf{1}_{a(X)(\eta(X)-1/2)\le 0})^\nu] \\ &\le 2c\mathbf{E}[(1-\tau(\eta(X)))\mathbf{1}_{a(X)(\eta(X)-1/2)\le 0}]^\nu, \end{aligned}$$

where the last inequality follows from Jensen's inequality, as the function $x \to x^\nu$ is concave for $\nu \in [0,1]$. We will show that for any $x \in \mathbb{R}^d$ we have

$$(1-\tau(\eta(x)))\mathbf{1}_{a(x)(\eta(x)-1/2)\le 0} \le \mathbf{E}[\varphi(a(X)Y)|X=x] - \mathbf{E}[\varphi(a_\varphi^{\star\star}(X)Y)|X=x],$$

from which the proof of the lemma follows by taking expectations and using the tower property. Note that

$$\mathbf{E}[\varphi(a(X)Y)|X=x] = \varphi(a(x))\eta(x) + \varphi(-a(x))(1-\eta(x)) = H_{\eta(x)}(a(x)).$$

By Lemma 1.5 we have

$$a_\varphi^{\star\star}(x) = \operatorname*{argmin}_{\tilde{a}\in\mathbb{R}} \mathbf{E}[\varphi(\tilde{a}Y)|X=x] = \operatorname*{argmin}_{\tilde{a}\in\mathbb{R}} H_{\eta(x)}(\tilde{a})$$

so that

$$\mathbf{E}[\varphi(a_\varphi^{\star\star}(X)Y)|X=x] = \min_{\tilde{a}\in\mathbb{R}} H_{\eta(x)}(\tilde{a}) = \tau(\eta(x)).$$

Thus, the inequality we want to prove reads

$$(1-\tau(\eta(x)))\mathbf{1}_{a(x)(\eta(x)-1/2)\leq 0} \leq H_{\eta(x)}(a(x)) - \tau(\eta(x)).$$

The right-hand side is non-negative, so the inequality holds if $a(x)(\eta(x)-1/2) > 0$. On the other hand, if $a(x)(\eta(x)-1/2) \leq 0$ note that by convexity of φ we have

$$\begin{aligned} H_{\eta(x)}(a(x)) &= \varphi(a(x))\eta(x) + \varphi(-a(x))(1-\eta(x)) \\ &\geq \varphi(a(x)\eta(x) - a(x)(1-\eta(x))) \\ &= \varphi(a(x)(2\eta(x)-1)) \geq \varphi(0) = 1, \end{aligned}$$

where for the last inequality we used that φ is non-increasing and that $\varphi(0) = 1$. ■

As far as the conditions of Zhang's Lemma are concerned, it is easy to verify that the following holds:

- **Exponential loss.** $\tau(\tilde{\eta}) = 2\sqrt{\tilde{\eta}(1-\tilde{\eta})}$, $c = 1/\sqrt{2}$, $\nu = 1/2$.
- **Hinge loss.** $\tau(\tilde{\eta}) = 1 - |1-2\tilde{\eta}|$, $c = 1/2$, $\nu = 1$.
- **Logistic loss.** $\tau(\tilde{\eta}) = -\tilde{\eta}\log_2\tilde{\eta} - (1-\tilde{\eta})\log_2(1-\tilde{\eta})$, $c = 1/\sqrt{2}$, $\nu = 1/2$.

8.4 Elements of Convex Theory

The above discussion was motivated by the desire to obtain a convex problem whose solution is related to the original problem of interest. But why is convexity important?

As the next proposition attests, it turns out that convexity by itself does not yield any advantage from a computational viewpoint, as *any optimization problem can be casted into a convex form.*

Let us recall a few definitions first.

For a set $\mathcal{T} \subseteq \mathbb{R}^n$, its *convex hull* is defined as

$$\operatorname{conv}(\mathcal{T}) := \left\{ \sum_{j=1}^{m} w_j t_j : w \in \Delta_m, t_1, \ldots, t_m \in \mathcal{T}, m \in \mathbb{N} \right\}.$$

For a function $f : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}$, its *epigraph* is defined as

$$\operatorname{epi}(f) := \{(x,t) \in \mathcal{D} \times \mathbb{R} : f(x) \leq t\}.$$

The following result states that optimizing a linear function $t \in \mathcal{T} \to c^\top t$ over a given set $\mathcal{T} \subseteq \mathbb{R}^n$ coincides with optimizing the linear function over the convex hull of the set, i.e. $\operatorname{conv}(\mathcal{T})$.

Proposition 8.6 *Let $\mathcal{T} \subseteq \mathbb{R}^n$ and $c \in \mathbb{R}^n$. Then,*

$$\begin{aligned} \min_{t\in\mathcal{T}} c^\top t &= \min_{t\in\operatorname{conv}(\mathcal{T})} c^\top t, \\ \max_{t\in\mathcal{T}} c^\top t &= \max_{t\in\operatorname{conv}(\mathcal{T})} c^\top t. \end{aligned}$$

Proof: Let's consider the minimization case—the maximization case follows analogously. One direction is trivial: as $\mathcal{T} \subseteq \operatorname{conv}(\mathcal{T})$, we have

$$\min_{t\in\mathcal{T}} c^\top t \geq \min_{t\in\operatorname{conv}(\mathcal{T})} c^\top t.$$

To prove the other direction we proceed as follows:

$$\begin{aligned}
\min_{t\in\mathrm{conv}(\mathcal{T})} c^\top t &= \min_{m\in\mathbb{N},t_1,\dots,t_m\in\mathcal{T},(w_1,\dots,w_m)\in\Delta_m} c^\top\left(\sum_{j=1}^m w_j t_j\right)\\
&= \min_{m\in\mathbb{N},t_1,\dots,t_m\in\mathcal{T},(w_1,\dots,w_m)\in\Delta_m} \sum_{j=1}^m w_j c^\top t_j\\
&\geq \min_{m\in\mathbb{N},t_1,\dots,t_m\in\mathcal{T},(w_1,\dots,w_m)\in\Delta_m} \sum_{j=1}^m w_j \min_{t\in\mathcal{T}} c^\top t\\
&= \min_{t\in\mathcal{T}} c^\top t,
\end{aligned}$$

where in the last line we used that $\sum_{j=1}^m w_j = 1$. ■

We are now ready to show that any optimization problem can be written in a convex form (here we only state the result regarding minimization).

Proposition 8.7 *Let $f : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}$ be any function. Then,*

$$\min_{x\in\mathcal{D}} f(x) = \min_{(x,t)\in\mathcal{C}} t$$

with $\mathcal{C} = \mathrm{conv}(\mathrm{epi}(f))$.

Proof: We immediately have

$$\min_{x\in\mathcal{D}} f(x) = \min_{(x,t)\in\mathrm{epi}(f)} t.$$

The proof is concluded by applying Proposition 8.6. ■

Note that, in Proposition 8.7, the function $(x,t) \in \mathbb{R}^d \times \mathbb{R} \to t \in \mathbb{R}$ is convex (it is linear!), and that the set $\mathcal{C} = \mathrm{conv}(\mathrm{epi}(f))$ is convex (any convex hull is convex by definition). Therefore, it is not true that convex problems are necessarily easy to solve.

The primary reason why, when combined with *additional assumptions*, convex problems are amenable to computations is related to the fact that convexity allows to infer *global* information from *local* information.

Global information are stored in subgradient, which define uniform lower bounds (in the form of hyperplanes).

Definition 8.8 (Subgradient) *Let $f : \mathcal{C} \subset \mathbb{R}^d \to \mathbb{R}$. A vector $g \in \mathbb{R}^d$ is a* subgradient *of f at $x \in \mathcal{C}$ if*

$$\boxed{f(x) - f(y) \leq g^\top(x-y) \quad \text{for any } y \in \mathcal{C}}$$

The set of subgradients of f at x is denoted $\partial f(x)$.

The above inequality can be written as $f(y) \geq f(x) + g^\top(y-x)$ for any $y \in \mathcal{C}$. Thus, each subgradient defines an hyperplane that *uniformly* bounds the function f from below. This is a form of global information (i.e., the bound holds for any $y \in \mathcal{C}$, not just locally in a neighborhood of x).

Convex functions are important as they always admit subgradients, even if they are not differentiable (in fact, if the domain is a convex set, existence of subgradients at every point is a property that *characterizes* convexity). When a convex function is differentiable at a point, then the gradient evaluated at that point is also a subgradient: local information (i.e., gradients; recall that derivatives are defined locally via limits) provides global information (i.e., subgradients).

Theorem 8.9 (Convexity and subgradients) *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be convex and $f : \mathcal{C} \to \mathbb{R}$. If $\forall x \in \mathcal{C}, \partial f(x) \neq \emptyset$, then f is convex. Conversely, if f is convex, then for any $x \in int(\mathcal{C}), \partial f(x) \neq \emptyset$. Furthermore, if f is convex and differentiable at x, then $\nabla f(x) \in \partial f(x)$.*

Proof: Omitted. ■

The local-to-global property ensured everywhere (in the interior of the domain) by convexity allows to construct at any (local) point x a uniform (global) linear lower bound (hyperplane) on the function that we want to minimize. This immediately suggests an iterative algorithmic procedure to minimize a convex function f. Starting from a given point x, the procedure computes the next point in the following fashion:

1. Compute $\partial f(x)$, a subgradient of f at x, which exists by convexity;
2. Move in the direction of $-\partial f(x)$, where the hyperplane defined by the subgradient decreases.

There are two reasons why the above procedure can not always be tractably implemented (if it were, by Proposition 8.7 we would be able to solve *any problem* in a computationally efficient way!).

In point 1., computing $\partial f(x)$ can be computationally expensive (convexity only guarantees the existence of subgradients, but it does not say anything about how easy they are to be computed!). This is not the case for the applications that we are going to consider in this course, as we will typically deal with convex differentiable functions whose gradients are cheap to compute. The fact that the gradients are cheap to compute is the primary reason why we will only consider first-order methods instead of (with typically better iteration complexity) second-order methods based on the Hessian, which is typically much more expensive to evaluate (so, even if second-order methods require less iterations, each iteration costs more).

In point 2., while the subgradient provides a direction where to move, in general we do not know for how far to move in that direction. To guide the design of algorithms and to establish rates of convergence, additional local-to-global properties are needed that can be used to tune the step size at each iteration. In the next section we describe the most widely-used properties that allow to use gradients of f (local objects) to construct *uniform* (global property) upper and lower bounds on the function f, not necessarily hyperplanes!

We conclude this section with another important property of convex functions. When a convex function is differentiable, the gradient can be used to characterize the minima of the function (note that there could be more than one minimum).

Proposition 8.10 (First order optimality condition) *Let f be convex, and $\mathcal{C}$ be a closed set on which f is differentiable. Then,*

$$\boxed{x^\star \in \operatorname*{argmin}_{x \in \mathcal{C}} f(x) \quad \iff \quad \nabla f(x^\star)^\top (x^\star - y) \leq 0 \quad \text{for any } y \in \mathcal{C}}$$

Proof: First, assume that $\nabla f(x^\star)^\top (x^\star - y) \leq 0$ for all $y \in \mathcal{C}$. Since the gradient is a subgradient, we have

$$f(x^\star) - f(y) \leq \nabla f(x^\star)^\top (x^\star - y) \leq 0,$$

i.e. $f(x^\star) \leq f(y)$ for all $y \in \mathcal{C}$. Second, assume $x^\star \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$. For any $y \in \mathcal{C}$, define the rate of change of f along the line from $x^\star$ to y as $h(t) = f(x^\star + t(y - x^\star))$. Since $x^\star$ is a minimizer, the function f is locally non-decreasing around $x^\star$, i.e. $h'(0) \geq 0$. Thus,

$$h'(0) = \nabla f(x^\star)^\top (y - x^\star) \geq 0,$$

i.e. $\nabla f(x^\star)^\top (x^\star - y) \leq 0$ for all $y \in \mathcal{C}$. ■

When $x^\star$ is an interior point of the set $\mathcal{C}$, $\nabla f(x^\star) = 0$ and the bound in the proposition above holds with equality. In this case the optimality condition is also a useful tool to compute the minimum of convex functions (when solving $\nabla f(x^\star) = 0$ is easy, either analytically or computationally). However, it may also be the case that $x^\star$ lies on the boundary of $\mathcal{C}$, which is what is captured by the general characterization $\nabla f(x^\star)^\top (x^\star - y) \leq 0$.

8.5 Strong Convexity, Smoothness, Lipschitz Continuity

Convexity alone does not make a problem easy from a computational point of view. As we will see in the next few lectures, a convex problem becomes computationally tractable when convexity is combined (possibly superseded) by other local-to-global geometric properties.

The following definitions hold if the function f is differentiable.

- **Convex:** $f(x) - f(y) \leq \nabla f(x)^\top (x - y)$ for any $x, y \in \mathbb{R}^d$.
- **α-Strongly convex:** There exits $\alpha > 0$ such that $f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{\alpha}{2}\|x - y\|_2^2 \; \forall x, y \in \mathbb{R}^d$.
- **β-Smooth:** There exits $\beta > 0$ such that $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2$ for any $x, y \in \mathbb{R}^d$.
- **γ-Lipschitz:** There exits $\gamma > 0$ such that $\|\nabla f(x)\|_2 \leq \gamma$ for any $x \in \mathbb{R}^d$.

The following characterizations hold if the function f is twice differentiable (a symmetric matrix M satisfies $M \preccurlyeq \beta I$ if and only if the largest eigenvalue of M is less than β; analogously $M \succcurlyeq \alpha I$ if and only if the smallest eigenvalue of M is greater than α).

- **Convex:** $\nabla^2 f(x) \succcurlyeq 0$ for any $x \in \mathbb{R}^d$.
- **α-Strongly convex:** There exits $\alpha > 0$ such that $\nabla^2 f(x) \succcurlyeq \alpha I$ for any $x \in \mathbb{R}^d$.
- **β-Smooth:** There exits $\beta > 0$ such that $\nabla^2 f(x) \preccurlyeq \beta I$ for any $x \in \mathbb{R}^d$.
- **γ-Lipschitz:** There exits $\gamma > 0$ such that $\|\nabla f(x)\|_2 \leq \gamma$ for any $x \in \mathbb{R}^d$.

Going back to loss functions, the following holds if we consider as domain the entire $\mathbb{R}$. Note that strong convexity, smoothness, and Lipschitz continuity are independent properties and they do not imply each others (obviously, strong convexity implies convexity).

	Strongly convex?	**Smooth?**	**Lipschitz continuous?**
Exponential loss	NO	NO	NO
Hinge loss	NO	NO	YES
Logistic loss	NO	YES	YES
Least Square loss	YES	YES	NO

However, for our applications it is typically the case that we only need to consider the loss functions in a compact interval of $\mathbb{R}$, so that we will only need the above properties restricted on a certain set $\mathcal{C} \subseteq \mathbb{R}^d$. See **Problem 3.5** in the Problem Sheets.

Remark 8.11 (Local-to-global properties) *Recall the Taylor series expansion of f around a given $x \in \mathbb{R}^d$:*

$$f(y) \approx f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x) + \cdots$$

A function is convex if it is uniformly bounded from below *by a "personalized" hyperplane (with the slope that depends on x):*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x).$$

A function is α-strongly convex if it is uniformly bounded from below *by a "personalized" hyperplane added to a "not-personalized" quadratic function (the curvature of the quadratic function does not depend on x — note that the position of the quadratic function does depend on x):*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|_2^2.$$

A function is β-smooth if it is uniformly bounded from above *by a "personalized" hyperplane added to a "not-personalized" quadratic function:*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2.$$

A function is γ-Lipschitz if it is uniformly bounded from above and below *by a "not-personalized" double cone (the slope of the cone does not depend on x — note that the position of the cone does depend on x):*

$$f(x) - \gamma\|y - x\|_2 \leq f(y) \leq f(x) + \gamma\|y - x\|_2.$$