8.1 Introduction

Let us recall the setting of binary classification:

- Training data \( Z_1, \ldots, Z_n \) such that \( Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\} \);
- Admissible action set \( A \subseteq B := \{a : \mathbb{R}^d \to \{-1, 1\}\} \);
- 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 \( A^* \) in the case of the “true” zero-one loss function \( \phi(y, \hat{y}) = 1_{\hat{y} \neq y} \). In this case we have, for any \( a \in B \),

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

Recall the definitions:

\[
a^* = \arg\min_{a \in A} r(a), \quad a^{**} = \arg\min_{a \in B} r(a), \quad A^* = \arg\min_{a \in A} R(a).
\]

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 \( A \), possibly with infinitely many classifiers \( |A| = \infty \), we have (constants are not optimized):

\[
P \left( r(A^*) - r(a^*) < 54 \sqrt{\frac{\text{VC}(A)}{n} + \frac{2 \log(1/\delta)}{n}} \right) \geq 1 - \delta.
\]

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

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 \( A^* \). It turns out that computing \( A^* \) with the true loss function is, in general, an 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 our original problem by another problem that is amenable to computations, and to investigate the price that we pay by this approximation from a statistical point of view. 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

\[
  f(\lambda x + (1 - \lambda)\tilde{x}) \leq \lambda f(x) + (1 - \lambda) f(\tilde{x})
\]

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

\[
  \lambda a + (1 - \lambda)\tilde{a} \in A
\]

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 \( \varphi^\star(u) := 1_{u \leq 0} \).

**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 \( \varphi(0) = 1 \) and \( \varphi^\star(u) \leq \varphi(u) \) for every \( u \in \mathbb{R} \).

The reason for this precise definition of convex surrogate is given by Zhang’s Lemma, which will state and prove below. 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 \( B_{soft} := \{a : \mathbb{R}^d \to \mathbb{R}\} \). For a given convex loss surrogate \( \varphi \), if the subset \( A_{soft} \subseteq B_{soft} \) of admissible soft-classifiers is convex, then the empirical \( \varphi \)-risk minimization problem defined as

\[
  R_{\varphi}(a) := \frac{1}{n} \sum_{i=1}^{n} \varphi(a(X_i)Y_i), \quad A_{\varphi}^* := \arg\min_{a \in A_{soft}} R_{\varphi}(a),
\]
is convex and hence 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.** \( A_{\text{soft}} = \{ a(x) = w^T x + b : w \in C_1 \subseteq \mathbb{R}^d, b \in C_2 \subseteq \mathbb{R} \} \), where \( C_1, C_2 \) are convex sets.

- **Majority votes (Boosting).** \( A_{\text{soft}} = \{ a(x) = \sum_{j=1}^m w_j h_j(x) : w = (w_1, \ldots, w_m) \in \Delta_m \} \), where \( \Delta_m \) 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 \( a, \hat{a} \in A_{\text{soft}} \), then \( \lambda a + (1 - \lambda) \hat{a} \in A_{\text{soft}} \) for any \( \lambda \in [0, 1] \), hence satisfying the assumption of a convex set.

Given a soft classifier \( a \in A \), the corresponding hard classifier is defined as \( \text{sign}(a) \in \{-1, 1\} \).

### 8.3 \( \varphi \)-risk Minimisation

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

\[
r_\varphi(a) := E \varphi(a(X)Y), \quad a^*_\varphi := \arg\min_{a \in B_{\text{soft}}} r_\varphi(a),
\]

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

\[
r(a) := E \varphi^*(a(X)Y) = P(a(X) \neq Y), \quad a^* := \arg\min_{a \in B} r(a).
\]

The following result establishes a condition on the convex surrogate loss function that allows the excess \( \varphi \)-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 \( A \) and \( A_{\text{soft}} \) 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 \( \hat{\eta} \in [0, 1] \), \( \hat{a} \in \mathbb{R} \), let

\[
H_{\hat{\eta}}(\hat{a}) := \varphi(\hat{a})\hat{\eta} + \varphi(-\hat{a})(1 - \hat{\eta}), \quad \tau(\hat{\eta}) := \inf_{\hat{a} \in \mathbb{R}} H_{\hat{\eta}}(\hat{a}).
\]

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

\[
|\hat{\eta} - \frac{1}{2}| \leq c(1 - \tau(\hat{\eta}))^{\nu} \quad \text{for any } \hat{\eta} \in [0, 1]
\]

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

\[
\underbrace{r(\text{sign}(a)) - r(a^*)}_{\text{excess risk hard classifier}} \leq \underbrace{2c(r_\varphi(a) - r_\varphi(a^*_\varphi))}_{\text{excess } \varphi \text{-risk soft classifier}}^{\nu}
\]
Proof: By Example 1.5, the Bayes decision rule $a^{**}$ reads

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

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

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

we have

$$\{\text{sign}(a(X)) \neq a^{**}(X)\} = \{\text{sign}(a(X))a^{**}(X) \leq 0\} \subseteq \{a(X)a^{**}(X) \leq 0\} = \{a(X)(\eta(X) - 1/2) \leq 0\},$$

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

$$r(\text{sign}(a)) - r(a^{**}) = \mathbb{E}[|2\eta(X) - 1|\mathbb{1}_{\text{sign}(a(X)) \neq a^{**}(X)}]$$

$$\leq \mathbb{E}[|2\eta(X) - 1|\mathbb{1}_{a(X)(\eta(X) - 1/2) \leq 0}]$$

$$\leq 2e\mathbb{E}[(1 - \tau(\eta(X)))^\nu \mathbb{1}_{a(X)(\eta(X) - 1/2) \leq 0}]$$

$$= 2e\mathbb{E}[(1 - \tau(\eta(X)))^\nu \mathbb{1}_{a(X)(\eta(X) - 1/2) \leq 0}]$$

$$\leq 2e\mathbb{E}[(1 - \tau(\eta(X)))^\nu \mathbb{1}_{a(X)(\eta(X) - 1/2) \leq 0}]^\nu,$$

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

$$(1 - \tau(\eta(x)))\mathbb{1}_{a(X)(\eta(X) - 1/2) \leq 0} \leq \mathbb{E}[\varphi(a(X)Y)|X = x] - r_\varphi(a_{\varphi}^{**}),$$

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

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

By Lemma 1.3 we have

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

so that

$$r_\varphi(a_{\varphi}^{**}) = \min_{\tilde{a} \in \mathbb{R}} H_{\eta(x)}(\tilde{a}) = \tau(\eta(x)).$$

Thus, the inequality we want to prove reads

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

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 $\varphi$ we have

$$H_{\eta(x)}(a(x)) = \varphi(a(x))\eta(x) + \varphi(-a(x))(1 - \eta(x))$$

$$\geq \varphi(a(x)\eta(x) - a(x)(1 - \eta(x)))$$

$$= \varphi(a(x)(2\eta(x) - 1)) \geq \varphi(0) = 1,$$

where for the last inequality we used that $\varphi$ is non-increasing.

As far as the condition in Zhang’s Lemma is concerned, it is easy to verify that the following holds:

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

Convexity allows to infer global information from local information. This is the primary reason why convex problems are typically (not always!) amenable to computations.

Global information are stored in subgradient, which we now define.

**Definition 8.6 (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
\[
 f(x) - f(y) \leq g^\top (x - y) \quad \text{for any} \ y \in \mathcal{C}.
\]
The set of subgradients of \( f \) at \( x \) is denoted \( \partial f(x) \).

The above inequality can be written as
\[
 f(y) \geq f(x) + g^\top (y - x) \quad \text{for any} \ y \in \mathcal{C}.
\]
Thus, each subgradient defines a plane 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, the gradient evaluated at that point is also a subgradient. In this case, local information (i.e., gradients; recall that derivatives are defined locally) provides global information (i.e., subgradients).

**Theorem 8.7 (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 \( x \in \text{int}(\mathcal{C}), \partial f(x) \neq \emptyset \). Furthermore, if \( f \) is convex and differentiable at \( x \), then \( \nabla f(x) \in \partial f(x) \).

**Proof:** Omitted.

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 8.8 (First order optimality condition)** Let \( f \) be convex, and \( \mathcal{C} \) be a closed set on which \( f \) is differentiable. Then,
\[
x^* \in \arg\min_{x \in \mathcal{C}} f(x) \iff \nabla f(x^*)^\top (x^* - y) \leq 0 \quad \text{for any} \ y \in \mathcal{C}
\]

**Proof:** If we first assume that \( \nabla f(x^*)^\top (x^* - y) \leq 0 \) for all \( y \in \mathcal{C} \), then since the gradient is a subgradient,\[
f(x^*) - f(y) \leq \nabla f(x^*)^\top (x^* - y) \leq 0,
\]
i.e. \( f(x^*) \leq f(y) \) for all \( y \in \mathcal{C} \). Now, if we assume \( x^* \in \arg\min_{x \in \mathcal{C}} f(x) \), we know that \( f \) is locally non-decreasing around \( x^* \). Define \( h(t) = f(x^* + t(y - x^*)) \) for all \( y \in \mathcal{C} \) - the rate of change of \( f \) along the line from \( x \) to \( y \). We require \( h'(0) \geq 0 \) since \( x^* \) is a minimizer. Thus,
\[
h'(0) = \nabla f(x^*)^\top (y - x^*) \geq 0,
\]
i.e. \( \nabla f(x^*)^\top (x^* - y) \leq 0 \) for all \( y \in \mathcal{C} \).

\( \blacksquare \)
When \( x^\ast \) is an interior point of the set \( \mathcal{C} \), \( \nabla f(x^\ast) = 0 \) 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 \( \nabla f(x^\ast) = 0 \) is easy, either analytically or computationally). However, it may also be the case that \( x^\ast \) lies on the boundary of \( \mathcal{C} \), which is what is captured by the more general condition

\[
\nabla f(x^\ast)^\top (x^\ast - y) \leq 0.
\]

In general, the reason why convex functions can be amenable to computations (as far as computing any global minima) is given by the local-to-global phenomenon described above, as the uniform lower bounds on the function that can be constructed at any given point \( x \) (by computing the subgradient at that point) immediately suggests the gradient descent algorithm approach that we will discuss next time.

However, not all convex functions are easy to minimize! To guide the design of algorithms to minimize convex functions and to establish rates of convergence, additional local-to-global properties are needed. We now describe the most widely-used such properties, that allow to use gradients of \( f \) (local property) to construct uniform upper and lower bounds on the function \( f \) (global property).

### 8.5 Strong Convexity, Smoothness, Lipschitz Continuity

Along with convexity, there are other geometrical properties that are typically satisfied by the functions we want to minimize. The following definitions hold if the function \( f \) is differentiable.

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

The following definitions hold if the function \( f \) is twice differentiable (a symmetric matrix \( M \) satisfies \( M \preceq \beta I \) if and only if the largest eigenvalue of \( M \) is less than \( \beta \); analogously \( M \succeq \alpha I \) if and only if the smallest eigenvalue of \( M \) is greater than \( \alpha \)).

- **Convex**: \( \nabla^2 f(x) \succeq 0 \) for any \( x \in \mathbb{R}^d \).
- **\( \alpha \)-Strongly convex**: There exists \( \alpha > 0 \) such that \( \nabla^2 f(x) \succeq \alpha I \) for any \( x \in \mathbb{R}^d \).
- **\( \beta \)-Smooth**: There exists \( \beta > 0 \) such that \( \nabla^2 f(x) \preceq \beta I \) for any \( x \in \mathbb{R}^d \).
- **\( \gamma \)-Lipschitz**: There exists \( \gamma > 0 \) 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).

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>Strongly convex?</th>
<th>Smooth?</th>
<th>Lipschitz continuous?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential loss</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Hinge loss</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Logistic loss</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Least Square loss</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>
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 $C \subseteq \mathbb{R}^d$. See Problem ?? in the Problem Sheets.

**Remark 8.9 (Local-to-global properties)** Recall the Taylor series expansion of $f$ around a given $x \in \mathbb{R}^d$:

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

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

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

A function is $\alpha$-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 $x$):

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

A function is $\beta$-smooth if it is uniformly bounded from above by a “personalized” hyperplane added to a “not-personalized” quadratic function:

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

A function is $\gamma$-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 $x$):

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