Lecture Notes - Learning theory: Part I
Empirical risk minimization
and related fields

M. Lerasle and E. Scornet

2019-2020
# Contents

1 Empirical risk minimization .................................................. 5
  1.1 Setting ............................................................................ 5
    1.1.1 Supervised learning .................................................. 6
    1.1.2 Unsupervised learning .............................................. 7
  1.2 Bayes estimator ................................................................. 10
  1.3 Empirical risk minimization ............................................... 12
  1.4 Learning from finite dictionaries ....................................... 13
  1.5 Fast rates ........................................................................ 15
    1.5.1 Fast rates under Bernstein’s condition ....................... 16
    1.5.2 Examples satisfying the Bernstein’s assumption ......... 16

2 Learning with infinite resources ............................................. 19
  2.1 The general framework ...................................................... 19
    2.1.1 Example: Lipschitz loss and linear functionals ............ 19
  2.2 Symmetrization principle .................................................. 20
  2.3 Application to learning theory .......................................... 21
  2.4 The finite case .................................................................. 21
  2.5 Vapnik-Chervonenkis theory ............................................. 22
    2.5.1 VC inequality ............................................................ 23
  2.6 Covering numbers ............................................................. 24
  2.7 The chaining method ......................................................... 26

3 Convex relaxation ................................................................. 29
  3.1 Construction of convex proxy ............................................ 29
  3.2 Link with hard classifiers .................................................. 31
  3.3 Bounding the $\varphi$-excess risk ........................................ 36
  3.4 Boosting ......................................................................... 37
  3.5 Support Vector Machine ................................................... 38
    3.5.1 Reproducing kernel Hilbert space .............................. 38
    3.5.2 Representer theorem ................................................. 40
    3.5.3 Excess risk of $\varphi$-ERM ........................................... 41
CONTENTS

4 Convex optimization 43
  4.1 Convex problems ........................................ 44
  4.2 Gradient descent ........................................ 46
  4.3 Projected gradient descent ............................ 48
  4.4 An exercise .............................................. 50
  4.5 Examples .................................................. 51
    4.5.1 SVM .................................................. 51
    4.5.2 Boosting .............................................. 52
  4.6 Mirror descent .......................................... 53
  4.7 Stochastic optimization ............................... 60
    4.7.1 Stochastic gradient descent ....................... 60
    4.7.2 Stochastic mirror descent ......................... 62

5 Stochastic bandit 65
  5.1 The full information case ............................. 66
  5.2 Upper confidence bound (UCB) policy ................. 68
  5.3 Bounded regret ......................................... 69
  5.4 An exercise .............................................. 70

6 Concentration inequalities 73
  6.1 Chernoff’s bound ........................................ 73
  6.2 Generic approaches .................................... 77
    6.2.1 SubGaussian random variables .................... 77
    6.2.2 SubPoissonian random variables .................. 81
    6.2.3 Bernstein’s inequality ............................. 84
  6.3 Maximal concentration inequalities .................... 87
    6.3.1 Azuma-Hoeffding’s inequality .................... 88
    6.3.2 The bounded difference inequality ............... 89
  6.4 Gaussian concentration inequality .................... 90
Chapter 1

Empirical risk minimization

1.1 Setting

Let $Z$ denote a measurable space where data take values. Let $F$ denote a set of parameters. For any $f \in F$ and $z \in Z$, let $\ell_f(z)$ denote the loss of parameter $f$ at observation point $z$. Let $P$ denote a probability distribution on $Z$. The risk of any parameter $f \in F$ is measured by

$$R(f) = P \ell_f := \mathbb{E}_{Z \sim P} [\ell_f(Z)].$$

This course deals with minimizing the risk function

$$\min_{f \in F} P \ell_f.$$

The distribution $P$ is unknown, so exact minimization of the risk is impossible. Nevertheless, minimizers of the risk will play an important role. When it exists, $f^* \in \arg\min_{f \in F} P \ell_f$ is called an oracle. To approximate $f^*$, a data set $D_N = \{Z_1, \ldots, Z_N\}$ is available. In these notes, $D_N$ always contains i.i.d. random variables with distribution $P$. A natural idea to estimate $f^*$ is to estimate each loss $P \ell_f$ by its empirical estimator $P_N \ell_f = N^{-1} \sum_{i=1}^{N} \ell_f(Z_i)$. Then, $f^*_F$ is estimated by the empirical risk minimizer

$$\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f.$$

If a minimizer does not exist, one can define alternatively, for $\epsilon > 0$, $\epsilon$-minimizers $\hat{f}_{\text{erm}}^{(\epsilon)}$ as any function satisfying

$$P_N \ell_{\hat{f}_{\text{erm}}^{(\epsilon)}} \leq \inf_{f \in F} P_N \ell_f + \epsilon.$$

Most of these notes deal with these estimators. Empirical risk minimization is quite common as can be appreciated from the following examples.
1.1.1 Supervised learning

In supervised learning, data \( z \in Z \) are couples \( z = (x, y) \in \mathcal{X} \times \mathcal{Y} \) and parameters \( f \in F \) are functions \( f : \mathcal{X} \rightarrow \mathcal{Y} \). The most classical problems in supervised learning are classification where \( \mathcal{Y} \) is a discrete set as in binary classification where \( \mathcal{Y} = \{0, 1\} \) and regression where \( \mathcal{Y} \subset \mathbb{R} \) is a continuous subset.

Among classical loss functions, one can mention the 0-1 loss in classification \( \ell_f(x, y) = 1_{\{y \neq f(x)\}} \) and the quadratic loss in regression \( \ell_f(x, y) = (y - f(x))^2 \).

In the couple \((x, y)\), \( x \) is called the input and \( y \) the output. The interpretation of supervised learning is that one wants to predict a typical output \( Y \) associated to the input \( X \) when \((X, Y) \sim P\).

For example, classification algorithms are routinely used to classify spams or help diagnosis.

In supervised learning, the function minimizing the risk \( P \ell_f \) among all functions \( f : \mathcal{X} \rightarrow \mathcal{Y} \) for which this risk is well defined would be an ideal estimator. It is referred to as the Bayes estimator and it is denoted by \( f^* \).

For any \( f \in F \), one has the relations

\[
R(f^*) \leq R(f_F^*) \leq R(f) .
\]

The difference \( \mathcal{E}(f) := R(f) - R(f^*) \) is usually called the excess risk of \( f \). Let us also introduce the notation \( \mathcal{E}_{\ell,F}(f) = R(f) - R(f_F^*) \) which is non negative for any \( f \in F \), so the excess risk can be decomposed

\[
\mathcal{E}(f) = \mathcal{E}(f_F^*) + \mathcal{E}_{\ell,F}(f) .
\]

In particular, for any data driven \( \hat{f} \in F \),

\[
\mathcal{E}(\hat{f}) = \mathcal{E}(f_F^*) + \mathcal{E}_{\ell,F}(\hat{f}) .
\]

The term \( \mathcal{E}(f_F^*) \) is an error unavoidable, sometimes called bias of estimation. Implicitly, working with the set of functions \( F \) means that this error is assumed to be small. Of course, the richest the class \( F \), the larger is the class \( f^* \) of functions for which it is true. The other term \( \mathcal{E}_{\ell,F}(\hat{f}) = R(\hat{f}) - R(f_F^*) \) is random as \( R(\hat{f}) \) is (the expectation defining \( R(f) \) is only taken with respect to \( Z \), independent of \( D_N \), so

\[
R(\hat{f}) = \mathbb{E}_{Z \sim P}[\ell_f(Z)|D_N] .
\]

A large part of these notes aims at providing deterministic upper bounds on \( \mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}) \), \( \Delta_{N,\delta}(F) \) such that, with probability at least \( 1 - \delta \),

\[
\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}) \leq \Delta_{N,\delta}(F) .
\]
1.1. SETTING

One will also sometimes try to bound the expectation $E\left[\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}})\right] \leq \Delta_N(F)$. These inequalities are referred to as oracle inequalities as they compare the risk of the estimator $\hat{f}_{\text{erm}}$ with the one of the oracle $f_F^*$. The term $\Delta_{N,\delta}(F)$ typically goes to 0 when $N \to \infty$, while it grows with the complexity of $F$. It is tempting to minimize the excess risk to choose a small set $F$ (the excess risk is null if $F$ is a singleton), but, even if $\hat{f}_{\text{erm}}$ has a small excess risk, it may have poor prediction properties if the class $F$ is not rich enough. Ideally, the model $F$ should be chosen to realize a tradeoff between the errors $\mathcal{E}(f_F^*)$ and $\mathcal{E}_{\ell,F}(\hat{f})$. These lectures aim at understanding the error $\mathcal{E}_{\ell,F}(\hat{f})$ in this tradeoff.

1.1.2 Unsupervised learning

In unsupervised learning, we don’t observe the output $y$ so data are simply inputs $z = x \in \mathcal{X}$. One wants to learn features of their distribution $P$.

**Multivariate mean estimation** Let $\mathcal{X} \subset \mathbb{R}^d$, and imagine that one wants to estimate the expectation $f_F^* = P[Z]$. Let then $F = \mathbb{R}^d$ and $\ell_f(x) = \|x - f\|^2$, where $\| \cdot \|$ is the Euclidean norm on $\mathbb{R}^d$. For any $f \in F$

$$\|X - f\|^2 = \|X - f_F^*\|^2 + \|f - f_F^*\|^2 - 2(X - f_F^*)^T(f - f_F^*) .$$

Hence

$$P\ell_f = P[\|X - f\|^2] = P[\|X - f_F^*\|^2] + \|f - f_F^*\|^2 \geq P[\|X - f_F^*\|^2] = P\ell_{f_F^*} .$$

Thus

$$\arg\min_{f \in F} P\ell_f = \{f_F^*\}, \quad \mathcal{E}_{\ell,F}(f) = \|f - f_F^*\|^2 .$$

In this example one can check that $\hat{f}_{\text{erm}}$ is simply the empirical mean

$$\hat{f}_{\text{erm}} = \frac{1}{N} \sum_{i=1}^{N} X_i .$$

**Least-squares density estimation** Assume that $P$ has density $f^*$ w.r.t. a known measure $\mu$ on $\mathcal{X}$ and that $f^* \in L^2(\mu)$. Let $(\varphi_i)_{i=1,\ldots,d}$ denote an orthonormal system in $L^2(\mu)$ and let $F$ denote the linear span of the $(\varphi_i)_{i=1,\ldots,d}$. Let $f_F^*$ denote the orthogonal projection of $f^*$ onto $F$ and let

$$\forall f \in L^2(\mu), \quad \ell_f(x) = \|f\|_{L^2(\mu)}^2 - 2f(x) .$$
The key is to remark that, for any $f \in L^2(\mu)$,

$$\langle f^*, f \rangle_{L^2(\mu)} = \int f f^* \, d\mu = P[f] ,$$

and, for any $f \in F$, by definition of $f^*_F$,

$$\langle f^*, f \rangle_{L^2(\mu)} = \langle f^*_F, f \rangle_{L^2(\mu)} .$$

From these relationships,

$$P\ell f = \|f\|_{L^2(\mu)} - 2Pf = \|f\|_{L^2(\mu)} - 2 \langle f^*, f \rangle_{L^2(\mu)}$$
$$= \|f\|_{L^2(\mu)} - 2 \langle f^*_F, f \rangle_{L^2(\mu)} = \|f - f^*_F\|_{L^2(\mu)} - \|f^*_F\|_{L^2(\mu)}$$
$$\geq -\|f^*_F\|_{L^2(\mu)} = P\ell f^*_F .$$

Hence, $f^*_F$ is the oracle in $F$ (unique $\mu$-a.s.). One can also check that

$$f^*_F = \sum_{i=1}^d P[\varphi_i] \varphi_i, \quad \widehat{f}_{\text{erm}} = \sum_{i=1}^d P_N[\varphi_i] \varphi_i .$$

The excess risk is $\mathcal{E}(f) = \|f - f^*\|_{L^2(\mu)}^2$ and

$$\mathcal{E}_{\ell,F}(\widehat{f}_{\text{erm}}) = \|\widehat{f}_{\text{erm}} - f^*_F\|_{L^2(\mu)}^2 = \sum_{i=1}^d ((P_N - P)\varphi_i)^2 . \quad (1.1)$$

This problem has strong connections with multivariate means estimation. Since

$$f^*_F = \sum_{i=1}^d P[\varphi_i] \varphi_i ,$$

estimating $f^*_F$ amounts to estimate the finite dimensional expectation

$$\Gamma_F = P \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_d \end{bmatrix} \in \mathbb{R}^d .$$

The estimator provided in the last section was

$$\widehat{f}_{\text{erm}} = P_N \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_d \end{bmatrix} \in \mathbb{R}^d .$$
1.1. SETTING

The natural embedding of this estimator is thus

$$\hat{f}_{\text{erm}} = \sum_{i=1}^{d} P_N[\varphi_i] \varphi_i.$$ 

Moreover, (1.1) shows that the excess risk of this estimator is the squared Euclidean distance between $f^*_F$ and $\hat{f}_{\text{erm}}$. There are important differences between both problems though. First, in least-squares density estimation, any estimator $\hat{f} \in F$ usually has a bias $\|f^* - \hat{f}\|^2$ while this term in null in multivariate mean estimation. Second, in density estimation, the statistician has the choice of the basis $\varphi_i$ and the dimension $d$. The functions $\varphi_i$ can therefore be chosen bounded, which implies that the random variables $\varphi_i(X_j)$ are bounded. On the other for multivariate mean estimation, the coordinates $X_{i,j}$ can be unbounded.

**Maximum likelihood estimators** To estimate $P$, an alternative to the projection method, widely spread in statistics, is to start with a set $F$ of densities with respect to a given measure $\mu$ (like Lebesgue measure on $\mathbb{R}^d$) and define, for each $f$, the likelihood of the data set $D_N$:

$$L_f(D_N) = \prod_{i=1}^{N} f(X_i).$$

The Maximum Likelihood Estimator is a maximizer of the likelihood

$$\hat{f}_{\text{erm}} \in \arg\max_{f \in F} L_f(D_N).$$

Equivalently, $\hat{f}_{\text{erm}}$ is the minimizer of

$$\frac{1}{N}(-\log(L_f(D_N))) = P_N[-\log(f(X))].$$

Define $\ell_f(x) = -\log(f(x))$ so $\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f$ is an empirical risk minimizer. The risk $P\ell_f$ associated to this loss is therefore

$$P\ell_f = P[-\log(f(X))].$$

To check that this is a natural risk measure, one can compare its value at some $f \in F$ and in the true density $f^*$. It holds

$$P[\ell_f - \ell_{f^*}] = \int f^* \log \left( \frac{f^*}{f} \right) d\mu.$$

It is the Kullback divergence between $f^*$ and $f$, it is always non-negative as, by Jensen’s inequality, since log is concave,

$$\int f^* \log \left( \frac{f}{f^*} \right) d\mu = \mathbb{E} \left[ \log \left( \frac{f(X)}{f^*(X)} \right) \right] \leq \log \left( \mathbb{E} \left[ \frac{f(X)}{f^*(Z)} \right] \right) = \log \left[ \int f d\mu \right] = 0.$$
CHAPTER 1. EMPIRICAL RISK MINIMIZATION

Clustering A connection with classification can be made when the distribution $P$ of $X$ is a mixture. In the simplest case, say that there exist $p \in (0,1)$, $f_0$ and $f_1$ such that the density $f$ of $X$ w.r.t. a given measure $\mu$ on $\mathcal{X}$ is

$$f = (1 - p)f_0 + pf_1 .$$

One way to realize this distribution is to say that $X$ has the distribution of the first marginal of the couple $(X,Y)$, whose distribution is defined as follows: $Y$ has Bernoulli distribution $Y \sim \mathcal{B}(p)$ and, conditioning on $Y = 1$, $X \sim f_1$ while, conditioned on $Y = 0$, $X \sim f_0$. The clustering problem is to find the distribution of $Y|X$: given the observation $x$, one wants to know if this random variable $X$ that generated this observation has been simulated using $f_0$ (in this case, say that $X$ belongs to the class 0) or $f_1$ (in this case, $X$ belongs to the class 1). This problem looks like binary classification, except that the outputs $Y_i$ associated to the inputs $X_i$ are not observed and cannot be used to train the estimators. Clustering is a classification problem without observation (or supervision) of the outputs. This explains the name “unsupervised” learning.

1.2 Bayes estimator

In the remaining of this chapter, we focus on binary classification, which is a supervised learning problem where the output space $\mathcal{Y} = \{0,1\}$. The set of parameters $F$ is a set of classifiers, that is functions $f : \mathcal{X} \to \{0,1\}$. The loss $\ell_f$ is the $0-1$ loss $\ell_f(x,y) = 1_{\{f(x) \neq y\}}$.

Let us first assume that $F$ is the set of all functions $f : \mathcal{X} \to \mathcal{Y}$ and call $\eta$ the regression function, that is a function such that, for any measurable bounded function $\varphi : \mathcal{X} \to \mathbb{R}$, $E[Y \varphi(X)] = E[\eta(X) \varphi(X)]$.

**Definition 1.** The Bayes classifier $f^*$ is defined by

$$\forall x \in \mathcal{X}, \quad f^*(x) = 1_{\{\eta(x) > 1/2\}} .$$

The Bayes classifier minimizes the classification error among all functions $f : \mathcal{X} \to \mathcal{Y}$ as shown by the following theorem.

**Theorem 2.** The risk of the Bayes classifier satisfies

$$R(f^*) = E[\min(\eta(X), (1 - \eta(X)))] \leq 1/2 .$$

Moreover, for any $f : \mathcal{X} \to \{0,1\}$, the excess risk of $f$ defined as $\mathcal{E}(f) := R(f) - R(f^*)$ satisfies

$$\mathcal{E}(f) = P[2\eta - 1|1_{\{f \neq f^*\}}] = E[2\eta(X) - 1|1_{\{f(X) \neq f^*(X)\}}] .$$
1.2. BAYES ESTIMATOR

Proof. By definition, for any function \( f : \mathcal{X} \to \mathcal{Y} \),

\[
R(f) = \mathbb{E}[Y(1-f(X)) + (1-Y)f(X)] \\
= \mathbb{E}[\eta(X)(1-f(X)) + (1-\eta(X))f(X)] .
\]

By definition of the Bayes classifier, it follows that

\[
R(f^*) = \mathbb{E}[\eta(X)1_{\{\eta(X) \leq 1/2\}} + (1-\eta(X))1_{\{1-\eta(X) \leq 1/2\}}] \\
= \mathbb{E}[\min(\eta(X), (1-\eta(X)))] .
\]

As \( \min(\eta(X), (1-\eta(X))) \leq 1/2 \) almost surely, \( R(f^*) \leq 1/2 \) and \( R(f^*) = 1/2 \) iff \( \eta(X) = 1/2 \) almost surely, which happens when \( X \) does not bring information of \( Y \).

Moreover, for any function \( f : \mathcal{X} \to \mathcal{Y} \),

\[
R(f) - R(f^*) = \mathbb{E}[\eta(X)(f^*(X) - f(X)) + (1-\eta(X))(f(X) - f^*(X))] \\
= \mathbb{E}[(2\eta(X) - 1)(f^*(X) - f(X))] .
\]

Now, \((f^*(x) - f(x)) = 0\) if \( f^*(x) = f(x) \), \( f^*(x) - f(x) = 1\) if \( \eta(x) > 1/2 \) and \( f^*(x) \neq f(x) \) if \( \eta(x) \leq 1/2 \) and \( f^*(x) \neq f(x) \).

Overall, \( f^*(x) - f(x) = 1_{\{f^*(x) \neq f(x)\}}\text{sign}(2\eta(x) - 1) \) and

\[
(2\eta(X) - 1)(f^*(X) - f(X)) = |2\eta(X) - 1|1_{\{f^*(X) \neq f(X)\}} .
\]

\( \square \)

Theorem 3 implies that \( f^* \) is the best possible classifier. However, this classifier depends on the unknown distribution \( P \) of the data and is not available in practice. Instead, we want to build a random function \( \hat{f} : \mathcal{X} \to \mathcal{Y} \) such that the excess risk \( \mathcal{E}(\hat{f}) = R(\hat{f}) - R(f^*) \) is small. This task is hard in general as shown by the following negative result.

Theorem 3. If \( \mathcal{X} \) is infinite, then, for any \( N \geq 1 \), for any classifier \( \hat{f}_N \) built with \( \mathcal{D}_N \), for any \( \epsilon > 0 \), there exists a distribution \( P_N \) on \( \mathcal{X} \times \mathcal{Y} \) such that \( R(f^*) = 0 \) and \( \mathcal{E}(\hat{f}_N) \geq 1/2 - \epsilon \).

The proof is left as an exercise. One can for example proceed by completing the following steps. Let \( n, K \) denote integers, \( f : (\mathcal{X} \times \mathcal{Y})^n \to \{0, 1\} \) a classifier. The set \( \mathcal{X} \) being infinite, one can pick \( K \) distinct points \( a_1, \ldots, a_K \) in \( \mathcal{X} \). Denote by \( (\mathbb{P}_q)_{q \in \mathcal{Y}^K} \) denote the set of probability distributions on \( \mathcal{X} \otimes \mathcal{Y} \), such that, for any \( q \in \mathcal{Y}^K \),

\[
\mathbb{P}_q(X = a_j, Y = q_j) = K^{-1} \quad \text{for any} \quad j \in \{1, \ldots, K\} .
\]
1. Show that, for any \( q \in \{0, 1\}^N \), \( \mathbb{P}_q(X = a_j) = 1/K \) for any \( j \in \{1, \ldots, K\} \).

2. Compute \( \mathbb{P}_q(Y \neq f(X)) \) for any classifier \( f \).

3. Show that \( L^*_q = \inf_{f: X \rightarrow Y} \mathbb{P}_q(Y \neq f(X)) = 0 \) and that the infimum is achieved by any \( f^*_q \), such that, for any \( x \in \{a_1, \ldots, a_K\} \),
\[
f^*_q(x) = \sum_{i=1}^K q_i 1_{a_i}(x) .
\]

4. Show that
\[
sup_{\mathbb{P}} \{ \mathbb{E}_{\mathbb{P}^n}[\mathbb{P}(Y \neq f(X; D_n)|D_n)] - L^*_p \} \geq sup_{q \in \mathcal{Y}^K} \mathbb{E}_{\mathbb{P}^n}[\mathbb{P}_q(Y \neq f(X; D_n)|D_n)] \\
\geq \frac{1}{2K} \sum_{q \in \mathcal{Y}^K} \mathbb{E}_{\mathbb{P}^n}[\mathbb{P}_q(Y \neq f(X; D_n)|D_n)] \\
\geq \frac{1}{2K} \sum_{q \in \mathcal{Y}^K} \mathbb{E}_{\mathbb{P}^n}[\mathbb{E}_{\mathbb{P}_q}[1_{\{Y \neq f(X; D_n)\}|X, D_n}] 1_{\{X \neq \{X_1, \ldots, X_n\}\}}] .
\]

5. Show that, for any \( x \notin \{X_1, \ldots, X_n\} \), there exist \( 2^{K-1} \) values of \( q \in \{0, 1\}^K \) such that \( \mathbb{E}_{\mathbb{P}_q}[1_{\{Y \neq f(x; D_n)\}|D_n}] = 0 \) and \( 2^{K-1} \) values of \( q \in \{0, 1\}^K \) such that \( \mathbb{E}_{\mathbb{P}_q}[1_{\{Y \neq f(x; D_n)\}|D_n}] = 1 \).

6. Show that, for any \( q \in \{0, 1\}^K \) and for any \( x \in \{a_1, \ldots, a_K\} \),
\[
\mathbb{E}_{\mathbb{P}_q}[1_{\{X_1 \neq x, X_2 \neq x, \ldots, X_n \neq x\}}] = (1 - 1/K)^n .
\]

7. Conclude.

The important message carried by Theorem 3 is that one cannot hope to bound \( \mathcal{E}(\hat{f}) \) without making assumption on \( P \). In the following, these assumptions will be made via a choice of a subset \( F \). We will implicitly assume that \( \mathcal{E}(f^*_F) \) is not too large and focus on bounding \( \mathcal{E}_{\mathcal{F}}(\hat{f}) \).

### 1.3 Empirical risk minimization

Hereafter, we assume that a subset \( F \) of functions \( f : X \rightarrow Y \) is given, and, assuming that such a function exists, we denote by \( f^*_F \) an “oracle”: \( f^*_F \in F \) and satisfies
\[
f^*_F \in \text{argmin}_{f \in F} R(f) .
\]
The oracle cannot be used as a predictor either, but one can hope to estimate it correctly if $F$ is not too large. Recall that the excess risk of any estimator $\hat{f} \in F$ is decomposed as follows

$$\mathcal{E}(\hat{f}) = \mathcal{E}(f_F^*) + \mathcal{E}_{\ell,F}(\hat{f}). \quad (1.2)$$

The error $\mathcal{E}_{\ell,F}(\hat{f}) = R(\hat{f}) - R(f_F^*)$ is a stochastic term called estimation error that one hopes to bound either in expectation or by a deterministic quantity $\Delta_{N,\delta}(F)$ with probability at least $1 - \delta$. The residual excess risk $\mathcal{E}(f_F^*)$ is a modelisation error unavoidable for any $\hat{f} \in F$ that is comparable to the bias in statistics.

We focus on a particular estimator called empirical risk minimizer, that was originally introduced by Vapnik. The idea is simple, the oracle $f_F^*$ minimizes $P_{\ell_f}$ over $F$. The operator $P$ is unknown in this definition but can be estimated by the empirical mean operator $P_N$ defined for any real valued function $\varphi : X \times Y \to \mathbb{R}$ by $P_N \varphi = (1/N) \sum_{i=1}^N \varphi(X_i, Y_i)$. The unknown risk $P_{\ell_f}$ can be estimated by the empirical risk $P_N \ell_f$ and $f_F^*$ by a minimizer $\hat{f}_{\text{erm}}$ of the empirical risks

$$\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f, \quad P_N \ell_f = \frac{1}{N} \sum_{i=1}^N 1_{(f(X_i) \neq Y_i)}.$$

When exact minimization is hard, one can also define $\epsilon$-ERM $\hat{f}_{\text{erm}}^{(\epsilon)}$ such that

$$P_N \ell_{f_{\text{erm}}^{(\epsilon)}} \leq \min_{f \in F} P_N \ell_f + \epsilon.$$

### 1.4 Learning from finite dictionaries

This section shows a first oracle inequality for the empirical risk minimizer $\hat{f}_{\text{erm}}$ in the elementary case where $F = \{f_1, \ldots, f_M\}$.

**Theorem 4.** For any $\epsilon > 0$, $\epsilon$-ERM over $F = \{f_1, \ldots, f_M\}$ satisfy,

$$\mathbb{E}[\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}})] \leq \sqrt{\frac{2 \log(2M)}{N}} + \epsilon,$$

and, for all $\delta \in (0, 1)$

$$\mathbb{P}\left(\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}) > \sqrt{\frac{2 \log(2M/\delta)}{N}} + \epsilon\right) \leq \delta.$$

**Proof.** The proof relies on the elementary but useful result of Vapnik
Lemma 5 (Vapnik’s Lemma). For any \( \epsilon > 0 \), \( \epsilon \)-ERM satisfy, almost surely,
\[
\mathcal{E}_{\ell,F}(\hat{f}^{(\epsilon)}_{\text{erm}}) \leq 2 \sup_{f \in F} |(P_N - P)\ell_f| + \epsilon.
\]

Proof. By definition of \( \hat{f}^{(\epsilon)}_{\text{erm}} \), \( P_N[\ell^{(\epsilon)}_{\text{erm}} - \ell_{f^*}] \leq \epsilon \), therefore,
\[
\mathcal{E}_{\ell,F}(\hat{f}^{(\epsilon)}_{\text{erm}}) = (P - P_N)\ell^{(\epsilon)}_{\text{erm}} + P_N[\ell^{(\epsilon)}_{\text{erm}} - \ell_{f^*}] + (P_N - P)\ell_{f^*}
\begin{align*}
&\leq (P_N - P)[\ell_{f^*} - \ell^{(\epsilon)}_{\text{erm}}] + \epsilon \\
&\leq 2 \sup_{f \in F} |(P_N - P)\ell_f| + \epsilon .
\end{align*}
\]

By Vapnik’s Lemma, it is sufficient to bound \( \sup_{f \in F} |(P_N - P)\ell_f| \) to prove the theorem. To obtain the result in expectation, we use the following lemma.

Lemma 6 (Pisier-Massart Lemma). Let \( U_1, \ldots, U_M \) denote random variables such that \( \forall s > 0, i \in \{1, \ldots, M\}, \log \mathbb{E}[e^{sU_i}] \leq \frac{s^2\sigma^2}{2} \).

Then
\[
\mathbb{E}[ \max_{i=1,\ldots,M} U_i] \leq \sigma \sqrt{2 \log(M)} .
\]

Proof. By Jensen’s inequality, for any \( s > 0 \),
\[
\mathbb{E}[ \max_{i=1,\ldots,M} U_i] = \frac{1}{s} \mathbb{E}[\log( \max_{i=1,\ldots,M} e^{sU_i})] \leq \frac{1}{s} \log(\mathbb{E}[ \max_{i=1,\ldots,M} e^{sU_i}]) .
\]

Now, as the random variables \( e^{sU_i} \) are all nonnegative,
\[
\mathbb{E}[ \max_{i \in \{1,\ldots,M\}} e^{sU_i}] \leq \mathbb{E} \left[ \sum_{i \in \{1,\ldots,M\}} e^{sU_i} \right] \leq M e^{s^2\sigma^2/2} .
\]

It follows that, for any \( s > 0 \),
\[
\mathbb{E}[ \max_{i=1,\ldots,M} U_i] \leq \frac{2 \log M + s^2\sigma^2}{2s} .
\]

In particular, for \( s = \sqrt{2 \log M/\sigma} \), this yields the result. \( \square \)
By Lemma \([\ref{lemma}])\textsuperscript{2}, the random variables \(\ell_f(Z_i)\) and \(-\ell_f(Z_i)\) belong to SubGau(1/4). Therefore, from Proposition \([\ref{proposition}])\textsuperscript{4}, the \(2M\) random variables
\[
(P_N - P)\ell_f, \ (P - P_N)\ell_f, \quad f \in F,
\]
have log-Laplace transform bounded from above by
\[
\frac{s^2}{2} \sum_{i=1}^{N} \frac{1}{4N^2} = \frac{s^2}{2} \frac{1}{4N}.
\]
It follows therefore from Pisier-Massart’s Lemma that
\[
\mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] \leq \frac{1}{2} \sqrt{\log \frac{2M}{N}}.
\]
The oracle inequality in expectation in the theorem follows therefore from Vapnik’s lemma.

By Theorem \([\ref{theorem}])\textsuperscript{6}, for any \(f \in F\) and any \(t > 0\),
\[
\mathbb{P}((P_N - P)\ell_f > t) \leq e^{-2Nt^2}, \quad \mathbb{P}((P - P_N)\ell_f > t) \leq e^{-2Nt^2}.
\]
Applying a union bounds shows that,
\[
\forall t > 0, \quad \mathbb{P}
\left(
\sup_{f \in F} |(P_N - P)\ell_f| > t
\right)
\leq 2Me^{-2Nt^2}.
\]
Choosing \(t\) solving the equation \(2Me^{-2Nt^2} = \delta\) shows that
\[
\forall \delta \in (0, 1), \quad \mathbb{P}
\left(
\sup_{f \in F} |(P_N - P)\ell_f| > \sqrt{\frac{\log(2M/\delta)}{2N}}
\right)
\leq \delta.
\]
The proof is concluded thanks to Vapnik’s lemma.

\section*{1.5 Fast rates}

To prove faster rates of convergence, besides refined deviation inequalities, the following Bernstein condition will be needed. There exist \(C > 0\) and \(\alpha \in [0, 1]\) such that
\[
\forall f \in F, \quad \text{Var} (\ell_f - \ell_{f_p}) \leq C(P[\ell_f - \ell_{f_p}])^\alpha.
\] (1.4)
The first part of this section shows that the Bernstein condition implies faster rates of convergence for the ERM than \(1/\sqrt{N}\). The second part presents some assumptions under which this condition is satisfied.
CHAPTER 1. EMPIRICAL RISK MINIMIZATION

1.5.1 Fast rates under Bernstein’s condition

**Theorem 7.** Assume that (1.4) holds, then the \( \epsilon \)-ERM \( \hat{f}_{\text{erm}}^{(\epsilon)} \) over the finite dictionary \( F = \{ f_1, \ldots, f_M \} \) satisfies

\[
P \left( \mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}^{(\epsilon)}) \leq \gamma_{\alpha,C} \left( \frac{\log(M/\delta)}{N} \right)^{\frac{1}{2-\alpha}} + 4 \frac{\log(M/\delta)}{N} + \epsilon \right) \geq 1 - \delta .
\]

Here, one can choose \( \gamma_{\alpha,C} = (2 - \alpha) \alpha^{\frac{\alpha}{2-\alpha}} (2C)^{\frac{1}{2-\alpha}} \).

**Proof.** We prove the theorem for the actual \( \hat{f}_{\text{erm}}^{(\epsilon)} \), with \( \epsilon = 0 \). The extension to \( \epsilon > 0 \) is left to the reader. Recall that, from Eq (1.3) in the proof of Vapnik’s Lemma,

\[
\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}^{(\epsilon)}) \leq (P_N - P)(\ell_{f_{\text{erm}}} - \ell_{\hat{f}_{\text{erm}}}^{(\epsilon)}) .
\]  

By Bernstein’s inequality, for any \( f \in F \), with probability at least \( 1 - \delta/M \),

\[
(P_N - P)(\ell_{f_{\text{erm}}} - \ell_{f}) \leq \sqrt{\frac{2\text{Var}(\ell_{f_{\text{erm}}} - \ell_{f}) \log(M/\delta)}{N}} + \frac{2\log(M/\delta)}{N} .
\]

Applying a union bound and Bernstein’s condition (1.4), this shows that, with probability at least \( 1 - \delta \), for any \( f \in F \),

\[
(P_N - P)(\ell_{f_{\text{erm}}} - \ell_{f}) \leq \sqrt{\frac{2C\mathcal{E}_{\ell,F}(f) \log(M/\delta)}{N}} + \frac{2\log(M/\delta)}{N} .
\]

By Minkowsky’s inequality, on the same event,

\[
(P_N - P)(\ell_{f_{\text{erm}}} - \ell_{f}) \leq \frac{1}{2} \mathcal{E}_{\ell,F}(f) + \frac{(2 - \alpha) \alpha^{\frac{\alpha}{2-\alpha}} C^{\frac{1}{2-\alpha}}}{2^{\frac{1}{2-\alpha}}} \left( \frac{\log(M/\delta)}{N} \right)^{1/(2-\alpha)} + \frac{2\log(M/\delta)}{N} .
\]

\[\Box\]

1.5.2 Examples satisfying the Bernstein’s assumption

In this section, we assume that \( f^* \in F \), so \( f_{\text{erm}}^{(\epsilon)} = f^* \) and \( \mathcal{E}_{\ell,F}(f) = \mathcal{E}(f) \). Remark that \( |\ell_f(X,Y) - \ell_{f^*}(X,Y)| \leq 1_{f(X) \neq f^*(X)} \), therefore

\[
\text{Var}(\ell_f - \ell_{f^*}) = \text{Var}(\ell_f - \ell_{f^*}) \leq \mathbb{P}(f(X) \neq f^*(X)) = \mathcal{E}(f) . \quad (1.6)
\]
1.5. FAST RATES

Recall the representation of the excess risk provided in Theorem 2
\[
\mathcal{E}(f) = \mathbb{E}[2\eta(X) - 1|1_{f(X) \neq f^*(X)}] ,
\]
It transpires from this representation that a natural way to bound \( \text{Var}(\ell_f - \ell_{f^*}) \) by \( \mathcal{E}(f) \) is to assume that \( \eta(X) \) does not put too much mass around 1/2. This way, one can bound from below \( |2\eta(X) - 1| \) in the representation of the excess risk. \( \eta(X) = 1/2 \) means that \( X \) does not bring information on \( Y \) or that the classification task. Therefore, assuming that this is not the case corresponds intuitively to assuming that the classification task should be easier. The purpose of this section is to show that, indeed, “margin” assumptions help the ERM to improve its rates of convergence.

**The noiseless case** Assume first that we are in the ideal case where \( \eta(X) = Y \), that is all the relevant information on \( Y \) is contained in \( X \). In this case, \( \eta(X) \in \{0, 1\} \) so, by the representation theorem
\[
\mathcal{E}(f) = \mathbb{E}[2\eta(X) - 1|1_{f(X) \neq f^*(X)}] = \mathbb{P}(f(X) \neq f^*(X)) .
\]
By (1.6), the “ideal” Bernstein’s condition holds with \( C = 1, \alpha = 1 \). From Theorem 8,
\[
\forall \delta \in (0, 1), \quad \mathbb{P}\left( \mathcal{E}(f) \leq \frac{6 \log(M/\delta)}{N} \right) \geq 1 - \delta .
\]

**Massart’s margin condition** A closely related assumption was proposed by Massart. Instead of assuming that all the information is contained in \( X \), which may be restrictive, Massart suggested that it is sufficient to bound away \( \eta \) from 1/2 to check Bernstein’s condition. The following condition is also known as the “hard” margin assumption: there exists \( \gamma \in (0, 1/2] \) such that, almost surely
\[
|\eta(X) - 1/2| \geq \gamma .
\]
Remark that the noiseless case corresponds to \( \gamma = 1/2 \). Under Massart’s condition, it holds that
\[
\mathcal{E}(f) = \mathbb{E}[2\eta(X) - 1|1_{f(X) \neq f^*(X)}] \geq 2\gamma \mathbb{P}(f(X) \neq f^*(X)) .
\]
By (1.6), Bernstein’s condition holds with \( C = 1/(2\gamma), \alpha = 1 \). From Theorem 8,
\[
\forall \delta \in (0, 1), \quad \mathbb{P}\left( \mathcal{E}(f) \leq \left( \frac{1}{\gamma} + 4 \right) \frac{\log(M/\delta)}{N} \right) \geq 1 - \delta .
\]
Mammen-Tsybakov’s margin condition  Mammen and Tsybakov further relaxed the hard margin condition and show that it is actually sufficient to bound the probability that $\eta(X)$ lies in the neighborhood of $1/2$ of size $t$ to check Bernstein’s condition. They introduced the following “soft” margin assumption: there exist $t_0$, $C_0$ and $\nu \in (0, 1)$ such that, for any $t \leq t_0$,

$$\Pr(|\eta(X) - 1/2| \leq t) \leq C_0 t^{1/\nu}.$$ 

Fix some $t \leq t_0$, under Mammen-Tsybakov’s condition

$$\mathcal{E}(f) = \mathbb{E}[|2\eta(X) - 1|1_{f(X) \neq f^*(X)}] \\
\geq \mathbb{E}[|2\eta(X) - 1|1_{|\eta(X) - 1/2| > 0}1_{f(X) \neq f^*(X)}] \\
\geq 2t \Pr(|\eta(X) - 1/2| > t \cap f(X) \neq f^*(X)) \\
\geq 2t(\Pr(f(X) \neq f^*(X)) - \Pr(|\eta(X) - 1/2| \leq t)) \\
\geq 2t(\Pr(f(X) \neq f^*(X)) - 2C_0 t^{1/\nu}).$$

Choose now $t = t_0 \mathcal{E}(f)^{1-\nu}/(1 \vee 4C_0)$ so $t \leq t_0$ and $2C_0 t^{1/\nu} \leq \mathcal{E}(f)/2$. Therefore

$$\frac{2t_0}{1 \vee 4C_0} \mathcal{E}(f)^{1-\nu} \Pr(f(X) \neq f^*(X)) \leq \frac{3\mathcal{E}(f)}{2},$$

that is, by (1.6),

$$\operatorname{Var}(\ell_f - \ell_{f^*}) \leq \frac{3(1 \vee 4C_0)}{4t_0} \mathcal{E}(f)^\nu.$$ 

Bernstein’s condition holds with $C = 3(1 \vee 4C_0)/(4t_0)$, $\alpha = \nu$. From Theorem 7.

For all $\delta \in (0, 1)$,

$$\Pr(\mathcal{E}(f) \leq \gamma_{C_0, t_0} \left(\frac{\log(M/\delta)}{N}\right)^{1/\nu}) \geq 1 - \delta.$$
Chapter 2

Learning with infinite ressources

2.1 The general framework

Assume that \( Z_1, \ldots, Z_N \) are couples \( (X_i, Y_i) \) taking values in \( \mathcal{X} \times [-1, 1] \).
Assume that \( F \) is a set of functions \( f : \mathcal{X} \to [-1, 1] \). Assume that \( \ell_f(z) \in [0, 1] \). The risk of \( f \in F \) is \( R(f) = P\ell_f \).

Oracles \( \hat{f} \) are defined by

\[
\hat{f} \in \arg\min_{f \in F} R(f).
\]

Empirical risk minimizers (ERM) \( \hat{f}_{\text{erm}} \) are defined by

\[
\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f.
\]

From Vapnik’s Lemma, risk bounds for the ERM follow from upper bounds on \( \sup_{f \in F} |(P_N - P)\ell_f| \).

2.1.1 Example: Lipschitz loss and linear functionals

All along this section, besides binary classification, general results will be illustrated on the following example. The importance of this example will become clear in the following chapter.

Assume that the loss \( \ell \) satisfies \( \ell_f(x, y) = c(f(x), y) \), for all \( f : \mathcal{X} \to \mathbb{R}, x \in \mathcal{X} \) and \( y \in \mathcal{Y} \) and that the cost function \( c \) is (uniformly over \( \mathcal{Y} \)) Lipschitz in it’s first argument

\[
\exists L > 0 : \quad \forall u, v \in \mathbb{R}, \forall y \in \mathcal{Y}, \quad c(u, y) - c(v, y) \leq L|u - v| . \tag{2.1}
\]

Let \( B_{\infty} = \{ a \in \mathbb{R}^d : \| a \|_{\infty} = \max_{i \in \{1, \ldots, d\}} |a_i| \leq 1 \} \). Assume furthermore that \( \mathcal{X} = B_1 = \{ x \in \mathbb{R}^d : \| x \|_1 = \sum_{i=1}^d |x_i| \leq 1 \} \) and that

\[
F = \{ f = f_a, a \in B_{\infty} \}, \quad \text{where} \quad \forall a, x \in \mathbb{R}^d, \quad f_a(x) = a^T x .
\]
2.2 Symmetrization principle

From Vapnik’s lemma, the risk of ERM $\hat{f}_{erm}$ will be bounded by $\sup_{f \in F} \left| (P_N - P) \ell_f \right|$. By the bounded difference inequality, see Eq (6.2), it will be enough to bound

$$\mathbb{E}[\sup_{f \in F} (P_N - P) \ell_f] .$$

Bounding this quantity is hard in general. A common step to many methods though is to apply the following symmetrization trick.

**Lemma 8 (Symmetrization Lemma).** For any loss functions $\ell$, any set of parameters $F$,

$$\mathbb{E}[\sup_{f \in F} |(P_N - P) \ell_f|] \leq 2 \mathbb{E}\left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(Z_i) \right| \right] \leq 2 \sup_{\pi \in \mathcal{Z}} \mathbb{E}\left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(z_i) \right| \right] ,$$

where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. Rademacher random variables.

**Proof.** Let $\mathcal{D}_N = (Z'_1, \ldots, Z'_N)$ denote an independent copy of $\mathcal{D}_N$, let $P'_N$ denote the empirical process based on $\mathcal{D}_N'$, that is $P'_N g = N^{-1} \sum_{i=1}^{N} g(Z'_i)$ for any function $g : \mathcal{Z} \to \mathbb{R}$. A key remark underlying the symmetrization principle is that

$$\forall f \in F, \quad P' \ell_f = \mathbb{E}[P'_N \ell_f | \mathcal{D}_N] .$$

By Jensen’s inequality, it follows that

$$\mathbb{E}[\sup_{f \in F} |(P_N - P) \ell_f|] = \mathbb{E}[\sup_{f \in F} |P_N \ell_f - \mathbb{E}[P'_N \ell_f | \mathcal{D}_N]|] \leq \mathbb{E}[\sup_{f \in F} \mathbb{E}[|P_N \ell_f - P'_N \ell_f | | \mathcal{D}_N]|] \leq \mathbb{E}\left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} (\ell_f(Z_i) - \ell_f(Z'_i)) \right| \right] .$$

As the vectors $(\ell_f(Z_i) - \ell_f(Z'_i))_{i \in \{1, \ldots, N\}, f \in F}$ and $(\epsilon_i (\ell_f(Z_i) - \ell_f(Z'_i)))_{i \in \{1, \ldots, N\}, f \in F}$ have the same distribution, it follows that

$$\mathbb{E}[\sup_{f \in F} |(P_N - P) \ell_f|] \leq 2 \mathbb{E}\left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i (\ell_f(Z_i) - \ell_f(Z'_i)) \right| \right] \leq 2 \mathbb{E}\left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(Z_i) \right| \right] .$$

$\square$
2.3 Application to learning theory

Recall that Vapnik’s Lemma implies that the risk of the Empirical Risk Minimizer $\hat{f}_{\text{erm}}$ is bounded from above by

$$E_{\ell,F}(\hat{f}_{\text{erm}}) \leq 2 \sup_{f \in F} |(P_N - P)f| .$$

As $\ell_f \in [0,1]$ the functional $\sup_{f \in F} |(P_N - P)\ell_f|$ belongs to $\mathbb{D}(c)$, with $c_i = 1/N$ for any $i \in \{1, \ldots, N\}$. Therefore, from the bounded difference inequality, see Eq (6.2), for any $\delta \in (0,1)$, with probability larger than $1 - \delta$,

$$\sup_{f \in F} |(P_N - P)\ell_f| \leq E[\sup_{f \in F} |(P_N - P)\ell_f|] + \sqrt{\frac{\log(1/\delta)}{2N}} .$$

By the symmetrization lemma, it follows that

$$\sup_{f \in F} |(P_N - P)\ell_f| \leq 2 \mathcal{R}_N(\ell, F) + \sqrt{\frac{\log(1/\delta)}{2N}} , \quad (2.2)$$

where $\mathcal{R}_N(\ell, F)$ denotes the Rademacher complexity of $F$ with respect to the loss $\ell$, defined by

$$\mathcal{R}_N(\ell, F) = \sup_{z \in 2^N} E\left[\sup_{f \in F} \frac{1}{N} \sum_{i=1}^N \epsilon_i \ell_f(z_i)\right] .$$

2.4 The finite case

Lemma 9. If $F$ is finite, then

$$\mathcal{R}_N(\ell, F) \leq \sqrt{\frac{2\log(2|F|)}{N}} .$$

Proof. Consider the set $B$ of the $|F|$ vectors $(\ell_f(z_i))_{1 \leq i \leq N} \in \mathbb{R}^N$. As their coordinates belong to $[0,1]$, $\Delta(B) = \max_{b \in B} \|b\| \leq \sqrt{N}$. Therefore, the proof terminates with Lemma 10. \qed

Lemma 10. Let $B$ denote a finite subset of vectors in $\mathbb{R}^N$, and let $\Delta(B) = \max_{b \in B} \|b\|$. Then

$$N\mathcal{R}_N(B) = E\left[\max_{b \in B} \left|\sum_{i=1}^N \epsilon_i b_i\right|\right] \leq \Delta(B) \sqrt{2\log(2|B|)} .$$
Proof. For any $b \in B$, let $Z_b = \sum_{i=1}^N \epsilon_i b_i$. As $-|b_i| \leq \epsilon_i b_i \leq |b_i|$, Hoeffding’s Lemma ensures that

$$\forall s > 0, \quad \log(\mathbb{E}[e^{sZ_b}]) = \sum_{i=1}^N \log(\mathbb{E}[e^{s\epsilon_i b_i}]) \leq \frac{s^2 \|b\|^2}{2} \leq \frac{s^2 \Delta(B)^2}{2}.$$ 

Of course the bound holds for the random variable $-Z_b$. By Pisier-Massart’s Lemma, it follows that

$$\mathbb{E}[\max_{b \in B} |Z_b|] \leq \Delta(B) \sqrt{2 \log(2|B|)}.$$ 

$\square$

### 2.5 Vapnik-Chervonenkis theory

The goal of this section is to bound $\mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|]$ in the binary classification setting. For any $f \in F$, let $A_f$ denote the measurable set such that $\{Z \in A_f\} = \{Y \neq f(X)\}$. Let $A_F = \{A_f \in f \in F\}$. Introduce the following Rademacher complexity.

**Definition 11.** The Rademacher complexity of the collection of sets $\mathcal{A}$ in the space $\mathcal{Z}$ is defined as

$$\mathcal{R}_N(\mathcal{A}) = \sup_{z_1, \ldots, z_N \in \mathcal{Z}} \mathbb{E} \left[ \sup_{A \in \mathcal{A}} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i 1_{\{z_i \in A\}} \right| \right].$$

It follows from the symmetrization Lemma that

$$\mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] \leq 2\mathcal{R}_N(\mathcal{A}_F).$$

Introduce, for any $N \geq 1$, the set of vectors

$$T_N(z) = \{(1_{\{z_i \in A_f\}}, \ldots, 1_{\{z_N \in A_f\}})^T, f \in F\}, \quad z = (z_1, \ldots, z_N) \in \mathcal{Z}^N.$$

For any $z \in \mathcal{Z}^N$, $|T_N(z)| \leq 2^N$.

**Definition 12.** $F$ shatters $z \in \mathcal{Z}^N$ if $|T_N(z)| = 2^N$. The shattering coefficients of $F$ is the sequence $(S_n(F))_{n \geq 1}$, where, for any $n \geq 1$,

$$S_n(F) = \sup_{z \in \mathcal{Z}^n} \{|T_n(z)|\}.$$

The VC dimension of $F$ is the largest integer $d$ such that $S_d(F) = 2^d$. We write $VC(F) = d$.

In words, there exists a set of $d$ points $z_1, \ldots, z_d$ in $\mathcal{Z}$ such that $F$ shatters $z = (z_1, \ldots, z_d)$ but there is not a single set of $d + 1$ points that is shattered by $F$. In particular therefore, $F$ does not shatter a single set of $d' > d$ points and the VC dimension is well defined.
2.5. VAPNIK-CHERVONENKIS THEORY

Exercise  Check that the VC dimension of half lines is 2.

2.5.1 VC inequality

Theorem 13. Assume that $F$ has VC dimension $d$, then

$$
\mathbb{E}\left[ \sup_{f \in F} \left| (P_N - P) \ell_f \right| \right] \leq 2 \sqrt{\frac{2d \log(2eN/d)}{N}}.
$$

Proof. Recall that, by the symmetrization trick, it holds

$$
\mathbb{E}\left[ \sup_{f \in F} \left| (P_N - P) \ell_f \right| \right] \leq 2 \mathcal{R}_N(A_F) = \sup_{z \in \mathcal{Z}^N} \frac{1}{N} \mathcal{R}_N(T_N(z)).
$$

For any $z \in \mathcal{Z}^N$, any vector in $T_N(z)$ has coordinates in $\{0, 1\}$, in particular therefore $\Delta(T_N(z)) \leq \sqrt{N}$. Moreover, by definition, for any $z \in \mathcal{Z}^N$, $|T_N(z)| \leq \mathcal{S}_N(F)$, therefore, by Lemma 10,

$$
\mathbb{E}\left[ \sup_{f \in F} \left| (P_N - P) \ell_f \right| \right] \leq 2 \sup_{z \in \mathcal{Z}^N} \mathcal{R}_N(T_N(z)) \leq \sqrt{\frac{2 \log(2\mathcal{S}_N(F))}{N}}.
$$

The following lemma concludes the proof of the theorem.

Lemma 14 (Sauer’s Lemma). If $\text{VC}(F) = d$, then for all $N \geq 1$,

$$
\mathcal{S}_N(F) \leq \sum_{k=0}^{d} \binom{N}{k}.
$$

For any $N \geq d$, in particular,

$$
\mathcal{S}_N(F) \leq \left( \frac{eN}{d} \right)^d.
$$

Together with the BDI, Theorem 13 implies the following corollary.

Theorem 15. Let $F$ denote a set of classifiers with VC dimension $d$ and let $\hat{f}_{\text{erm}}$ denote the empirical risk minimizer over $F$. Then, for any $\delta \in (0, 1)$,

$$
\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}) \leq 4 \sqrt{\frac{2d \log(2eN/d)}{N}} + \sqrt{\frac{2 \log(1/\delta)}{N}}.
$$

Remark 16. The set $F$ may have finite VC dimension and be infinite (think about the half lines). Therefore, Theorem 15 extends the slow rates to possibly infinite dictionaries.
2.6 Covering numbers

**Definition 17.** Let $K$ denote a set endowed with a metric $d$ and let $\eta > 0$. An $\eta$-net of $(K, d)$ is a set $V$ such that, for any $x \in K$, there exists $y \in V$ such that $d(x, y) \leq \eta$. The covering numbers of $K, d$ are defined by

$$N(K, d, \eta) = \inf \{ |V| : V \text{ is an } \eta\text{-net of } (K, d) \}.$$ 

Define, for any $f$ and $g$ in $F$, and any $p \geq 1$,

$$d_p^*(f, g) = \left( \frac{1}{N} \sum_{i=1}^N |\ell_f(z_i) - \ell_g(z_i)|^p \right)^{1/p}.$$ 

**Theorem 18.** For any $z \in Z^N$,

$$\mathcal{R}_N^z(F) = \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i \ell_f(z_i) \right| \right] \leq \inf_{\eta > 0} \left\{ \eta + \sqrt{\frac{2 \log(2N(F, d_i^*, \eta))}{N}} \right\}.$$ 

**Proof.** Fix $z \in Z^N$ and $\eta > 0$. Let $V$ denote an $\eta$-net of $(F, d_i^*)$ such that $|V| = N(F, d_i^*, \eta)$. For any $f \in F$, denote by $\pi_V(f) \in V$ an element such that $d(f, \pi_V(f)) \leq \eta$. By the triangular inequality,

$$\mathcal{R}_N^z(F) \leq \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i (\ell_f(z_i) - \ell_{\pi_V(f)}(z_i)) \right| \right] + \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i \ell_{\pi_V(f)}(z_i) \right| \right] \leq \epsilon + \mathbb{E} \left[ \max_{f \in V} \left| \frac{1}{N} \sum_{i=1}^N \epsilon_i \ell_f(z_i) \right| \right].$$ 

The proof terminates thanks to Lemma 9.

In the main example of Section 2.1.1, Theorem 18 yields the following result.

**Theorem 19.** Consider the framework of Section 2.1.1. Then for any $\delta \in (0, 1)$,

$$\mathbb{P} \left( \mathcal{E}_{l,F}(\hat{f}_{\text{err}}) \leq 8 \sqrt{\frac{d \log(2eLN/d)}{N} + \frac{2 \log(1/\delta)}{N}} \right) \geq 1 - \delta.$$ 

**Proof.** Start with a basic lemma.
2.6. COVERING NUMBERS

Lemma 20. For any $f$ in $F$, denote by $a_f \in B_1$ the vector such that, for all $x \in X$, $f(x) = a_f^T x$. For any $f, g \in F$ and any $z \in \mathbb{Z}^N$, and any $p \geq 1$

$$d_p(f, g) = \left( \frac{1}{N} \sum_{i=1}^{N} |\ell_f(z_i) - \ell_g(z_i)|^p \right)^{1/p} \leq L \left( \frac{1}{N} \sum_{i=1}^{N} |(a_f - a_g)^T x_i|^p \right)^{1/p} \leq L \|a_f - a_g\|_\infty .$$

(2.3)

The proof is straightforward. To bound the covering number of $(F, d_1)$, it is therefore sufficient to bound the covering number of $(B_1, d_1)$, where $d_1(a, b) = \|a - b\|_1$, for all $a, b \in \mathbb{R}^d$. This is done in the following lemma.

Lemma 21. For any $\eta > 0$, there exists an $\eta$-net $V$ of $(B_\infty, d_1)$ with cardinality $(2/\eta)^d$.

To build an $\eta$-net of $(F, d_1)$, it is therefore enough to build an $(\eta/L)$-net of $(B_\infty, d_\infty)$. Let $\eta > 0$ and define the integer $k_0$ such that $k_0 - 1$ is the integer part of $(1 - \eta)/\eta$. Let $V_\eta$ denote the set of all vectors in $\mathbb{R}^d$ with coordinates taking values in the grid $G_\eta = \{\eta/2 + kn, k \in \{-k_0, \ldots, k_0 - 1\}\} \subset [-1, 1]$. $V_\eta$ has cardinality $(2k_0)^d \leq (2/\eta)^d$ and, for any $a \in B_\infty$, there exist $v \in V_\eta$ such that

$$\|a - v\|_\infty = \max_{i \in \{1, \ldots, d\}} \min_{g \in G} |a_i - g| \leq \eta .$$

Hence, $V_\eta$ is an $\eta$-net of $(B_\infty, d_\infty)$. From (2.3), it follows that the set $\{f_v, v \in V_{\eta/L}\}$ is an $(\eta/L)$-net of $(F, d_1)$. It follows that $N(F, d_1, \eta) \leq (2L/\eta)^d$. As this holds for any $z \in \mathbb{Z}^N$, it follows from Theorem 18 that the Rademacher complexity of $F$ can be bounded from above by

$$\mathcal{R}_N(\ell, F) \leq \inf_{\eta > 0} \left\{ \eta + \sqrt{\frac{2d \log(4L/\eta)}{N}} \right\} \leq \sqrt{\frac{2d}{N}} \left(1 + \sqrt{\log(2LN/d)}\right) \leq 2 \sqrt{\frac{d \log(2eLN/d)}{N}} .$$

The first inequality follows by taking $\eta = \sqrt{2d/N}$ and the second comes from $1 + \sqrt{x} \leq \sqrt{2(1 + x)}$. Using (2.2) and the symmetrization lemma, it follows that, for all $\delta \in (0, 1)$,

$$\mathbb{P}\left( \sup_{f \in F} |(P_N - P)\ell_f| \leq 4 \sqrt{\frac{d \log(2eLN/d)}{N} + \frac{\log(1/\delta)}{2N}} \right) \geq 1 - \delta .$$
Finally, from Vapnik’s lemma, for all \( \delta \in (0, 1) \),
\[
\mathbb{P} \left( R(\hat{f}_{\text{erm}}) \leq R(\hat{f}) + 8 \sqrt{\frac{d \log(2eLN/d)}{N}} + \sqrt{\frac{2 \log(1/\delta)}{N}} \right) \geq 1 - \delta.
\]

Theorem 19 establishes a risk bound for ERM in a case where the dictionary is infinite. This bound can be refined using a slightly sharper bound of the Rademacher complexity. The chaining method allowing this sharper analysis is at the core of the following section.

2.7 The chaining method

To establish the main result of this section, let
\[
\forall z \in \mathcal{Z}_N, \quad d_z^2(f, g) = \left( \frac{1}{N} \sum_{i=1}^{N} (\ell_f(z_i) - \ell_g(z_i))^2 \right)^{1/2}.
\]

Theorem 22. For any \( z \in \mathcal{Z}_N \),
\[
\mathcal{R}_N^z(F) \leq \inf_{\epsilon > 0} \left\{ 4\epsilon + \frac{12}{\sqrt{N}} \int_0^1 \sqrt{\log(N(F, d_z^2(t)))} \, dt \right\}.
\]

Proof. Fix \( z \in \mathcal{Z}_N \). For any \( j \geq 1 \), denote by \( V_j \) a \( 2^{-j} \)-net of \((F, d_z^2)\) with cardinality \( N(F, d_z^2, 2^{-j}) \). Let \( f = (\ell_f(z_1), \ldots, \ell_f(z_N))^T \) and \( \epsilon = (\epsilon_1, \ldots, \epsilon_N)^T \) so

\[
\mathcal{R}_N^z(F) = \frac{1}{N} \mathbb{E}\left[ \sup_{f \in F} f^T \epsilon \right]
\]

For any \( f \in F \), let \( \pi_j(f) \in V_j \) such that \( d_z^2(f, \pi_j(f)) \leq 2^{-j} \) and \( f_j = (\ell_{\pi_j(f)}(z_1), \ldots, \ell_{\pi_j(f)}(z_N))^T \). Define \( \pi_0(f) = f_0 \in F \), so, for any \( f \in F \) and any \( j_0 \),
\[
\epsilon^T(f - f_0) = \epsilon^T(f - f_{j_0}) + \sum_{j=1}^{j_0} \epsilon^T(f_j - f_{j-1})
\]

In particular therefore,
\[
\mathcal{R}_N^z(F) \leq \frac{1}{N} \mathbb{E}\left[ \sup_{f \in F} \left| \epsilon^T(f - f_{j_0}) \right| \right] + \sum_{j=1}^{j_0} \frac{1}{N} \mathbb{E}\left[ \sup_{f \in F} \left| \epsilon^T(f_j - f_{j-1}) \right| \right].
\]
2.7. THE CHAINING METHOD

First, by Cauchy-Schwarz inequality,

\[
\frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f - f_{j_0})|] \leq \frac{\|\varepsilon\|_2 d_2^j(f, \pi_{j_0}(f))}{\sqrt{N}} \leq 2^{-j_0} .
\]

Second, to bound

\[
\mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|] ,
\]

notice first that there are at most \(|V_j||V_{j-1}| = N(F, d_2^j, 2^{-j+1})N(F, d_2^j, 2^{-j})\) possibles values for the couples \((f_{j-1}, f_j)\). Next, by Lemma 10,

\[
\frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|] \leq \max_{f \in F} \|f_j - f_{j-1}\|_2 \sqrt{2 \log(|V_j||V_{j-1}|)} .
\]

Then

\[
\|f_j - f_{j-1}\|_2 = \sqrt{N} d_2^j(\pi_j(f) - \pi_{j-1}(f)) \leq 3\sqrt{N} 2^{-j} .
\]

Substituting in the previous bound yields

\[
\frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|] \leq 3(2^{-j}) \sqrt{\frac{2 \log(|V_j||V_{j-1}|)}{N}} .
\]

Summing up over \(j\) yields

\[
\sum_{j=1}^{j_0} \frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|] \leq \sum_{j=1}^{j_0} 3(2^{-j}) \sqrt{\frac{2 \log(|V_j|)}{N}} \\
+ \sum_{j=1}^{j_0} 3(2^{-j}) \sqrt{\frac{2 \log(|V_{j-1}|)}{N}} \\
= \sum_{j=1}^{j_0} 6(2^{-j} - 2^{-j-1}) \sqrt{\frac{2 \log(|V_j|)}{N}} \\
+ \sum_{j=1}^{j_0} 3(2^{-j+1} - 2^{-j}) \sqrt{\frac{2 \log(|V_{j-1}|)}{N}} \\
\leq 9 \sum_{j=1}^{j_0} (2^{-j} - 2^{-j-1}) \sqrt{\frac{2 \log(|V_j|)}{N}} \\
= 9 \sum_{j=1}^{j_0} (2^{-j} - 2^{-j-1}) \sqrt{\frac{2 \log(N(F, d_2^j, 2^{-j}))}{N}} .
\]
The function $x \mapsto \sqrt{\log(N(F, d^2_x, x))}$ being non-increasing, this sum can be bounded from above by the integral

$$\sum_{j=1}^{j_0} \frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon_j^f(f_j - f_{j-1})|] \leq \frac{9}{\sqrt{N}} \int_{2^{-j_0-1}}^{1/2} \sqrt{\log(N(F, d^2_x, x))} dx .$$

Choosing $j_0$ such that $2^{-j_0-2} \leq \eta \leq 2^{-j_0-1}$ yields

$$\mathcal{R}^\eta_N(F) \leq 4\eta + \frac{9}{\sqrt{N}} \int_{\eta}^{1} \sqrt{\log(N(F, d^2_x, x))} dx .$$


The advantage of this refined analysis will be clearer in the example of Section 2.1.1.

**Theorem 23.** Consider the framework of Section 2.1.1. There exists an absolute constant $C$ such that, for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}_{erm}) \leq C \left(\sqrt{\frac{d \log(eL)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right) \geq 1 - \delta .$$

**Proof.** Let us first compute $N(F, d^2_x, x)$ for any $x > 0$. Recall that $F = \{f_a, a \in B_\infty\}$. Fix $z \in \mathcal{Z}^N$, then

$$d^2_x(f, g) = \left(\frac{L^2}{N} \sum_{i=1}^{N} ((a_f - a_g)^T x_i)^2\right)^{1/2} \leq L\|a_f - a_g\|_\infty .$$

It follows therefore from Lemma 21 that

$$N(F, d^2_x, x) \leq (2L/x)^d .$$

Therefore, for any $\eta > 0$,

$$\int_{\eta}^{1} \sqrt{\log(N(F, d^2_x, x))} dx \leq \int_{0}^{1} \sqrt{d \log(2L/x)} dx \leq \sqrt{d \log(2L)} + C\sqrt{d} .$$

The Chaining theorem and (2.2) imply therefore that

$$\sup_{f \in F} |(P_N - P)\ell_f| \leq C \left(\sqrt{\frac{d \log(eL)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right) .$$

Vapnik’s Lemma concludes the proof. \(\square\)
Chapter 3
Convex relaxation

Solving the optimization problem defining the ERM

\[ \hat{f}_{\text{erm}} = \arg\min_{f \in F} P_N 1_{\{Y \neq f(X)\}} \]

is computationally demanding. In fact, for most classes \( F \), it simply cannot be done. The basic idea of this chapter is to replace the non-convex optimization problem defining \( \hat{f}_{\text{erm}} \) by a convex one and derive from (approximate) solutions of this problem a classifier with bounded excess risk.

3.1 Construction of convex proxy

Recall the basic definition of convexity.

**Definition 24.** A set \( F \) is convex if, for any \( f \) and \( g \) in \( F \) and any \( t \in (0, 1) \),

\[ tf + (1 - t)g \in F. \]

A function \( h \) is convex on a convex domain \( C \) if, for any \( x \) and \( y \) in \( C \) and any \( t \in (0, 1) \),

\[ h(tx + (1 - t)y) \leq th(x) + (1 - t)h(y). \]

The following chapter will present several algorithms to solve approximately convex optimization problems where one has to minimize a convex function \( \Theta \) over a convex set \( C \). Typically sets of classifiers are not convex and, even if \( F \) is replaced by a convex sets \( F \), the function that should be minimized to define \( \hat{f}_{\text{erm}} \),

\[ \theta_N(f) = P_N 1_{\{Y \neq f(X)\}}, \]

is not convex. Therefore, convex optimization algorithms cannot be used to approximate ERM \( \hat{f}_{\text{erm}} \). The idea of convex relaxation is to replace both \( F \) by a convex set \( \mathcal{F} \) and \( \theta_N \) by a convex function \( \Theta_N \) is such a way that one can derive from a minimizer of \( \Theta_N \) over \( \mathcal{F} \) a classifier with bounded
CHAPTER 3. CONVEX RELAXATION

excess risk. Building these convex problems requires three main steps. The first step, sometimes called spinning, consists in replacing \( Y \in \{0, 1\} \) by \( Y^{(s)} = 2Y - 1 \in \{-1, 1\} \). The spinning variables satisfy

\[
1_{\{Y \neq f(X)\}} = 1_{\{-Y^{(s)}(2f(X) - 1) > 0\}}.
\]

The second step is to replace the class \( F \) of classifiers by a convex class of functions called “soft classifiers”.

**Definition 25.** A soft classifier is a function \( f^{(s)} : \mathcal{X} \to [-1, 1] \). The “hard” classifier \( f^{(h)} \) associated with the soft classifier \( f^{(s)} \) is \( f^{(h)} = (1 + \text{Sign}(f^{(s)}))/2 \).

Let \( \mathcal{F} \) denote a convex set of soft classifiers. Typically, sets \( \mathcal{F} \) of interest will be sets of linear functions \( \mathcal{F} = \{a^T \cdot \mathbf{x} : \mathbf{a} \in \mathcal{C}\} \), where \( \mathcal{C} \) denotes a convex subset of \( \mathbb{R}^d \). It may be that inputs \( X_1, \ldots, X_N \) do not belong to \( \mathbb{R}^d \), in this case, we usually use a set of functions \( f_1, \ldots, f_d \) from \( \mathcal{X} \) to \( \mathbb{R} \) and consider as inputs the vectors

\[
\mathbf{X}_i = \begin{bmatrix} f_1(X_i) \\ \vdots \\ f_d(X_i) \end{bmatrix} \in \mathbb{R}^d.
\]

Classical examples of convex sets \( \mathcal{C} \) here are \( \ell_p \)-balls

\[
\mathbf{B}_p = \{\mathbf{a} \in \mathbb{R}^d : \|\mathbf{a}\|_p \leq 1\}, \quad \text{where} \quad \|\mathbf{a}\|_p = \left( \sum_{i=1}^{d} |a_i|^p \right)^{1/p},
\]

if \( p = +\infty \) the \( \ell_\infty \)-norm is defined as usual as \( \|\mathbf{a}\|_\infty = \max_{1 \leq i \leq d} |a_i| \). One speaks about \( \ell_p \)-aggregation. Another example of interest is when \( \mathcal{C} \) is the simplex

\[
\Delta_d = \{\mathbf{a} \in \mathbb{R}^d : \forall i \in \{1, \ldots, d\}, a_i \geq 0, \sum_{i=1}^{d} a_i = 1\},
\]

one refers to this problem as convex aggregation problem.

The third step is to replace the non-convex loss function \( 1_{\{-Y^{(s)}f^{(s)}(X) > 0\}} \) by a convex surrogate. Let \( \varphi \) denote a convex upper bound of the function \( 1_{\{x > 0\}} \). For example, \( \varphi \) can denote the hinge loss \( \varphi(x) = \max(1 + x, 0) \) or the logistic loss \( \varphi(x) = \log_2(1 + e^x) \).

Given a set \( \mathcal{F} \) of soft classifiers and a convex surrogate \( \varphi \), the \( \varphi \)-ERM is defined by

\[
\hat{f}_\varphi \in \arg\min_{f^{(s)} \in \mathcal{F}} P_N \ell_{\varphi, f^{(s)}}, \quad \text{where} \quad \ell_{\varphi, f^{(s)}}(z) = \varphi(-y^{(s)}f^{(s)}(x)).
\]
3.2 Link with hard classifiers

Define the \( \varphi \)-bayes estimator

\[
    f^*_\varphi \in \arg\min_{f^{(s)}} P\ell_{\varphi,f^{(s)}} ,
\]

where the minimum is taken among all functions \( f^{(s)} : \mathcal{X} \to [-1, 1] \). For any \( x \in \mathcal{X} \), it is clear that

\[
    f^*_\varphi(x) \in \arg\min_{\alpha} \mathbb{E}[\varphi(-Y^{(s)}\alpha)|X = x] .
\]

**Theorem 26.** If \( \varphi \) is differentiable, then, the hard classifier associated to the \( \varphi \)-bayes estimators \( f^{(h)} = (1 + \text{Sign}(f^*_\varphi))/2 \) is the bayes estimator \( f^{(h)} = 1_{\{\eta > 1/2\}} \).

**Proof.** The \( \varphi \)-loss function \( \ell_{\varphi,f^{(s)}}(z) \) is equal to \( \varphi(-f^{(s)}(x)) \) if \( y^{(s)} = 1 \), i.e. when \( y = 1 \) and to \( \varphi(f^{(s)}(x)) \) if \( y = 0 \), thus, for any soft classifier \( f^{(s)} \),

\[
    \mathbb{E}[\ell_{\varphi,f^{(s)}}(X,Y)|X] = \mathbb{E}[Y \varphi(-f^{(s)}(X)) + (1-Y)\varphi(f^{(s)}(X))|X] \\
    = \eta(X)\varphi(-f^{(s)}(X)) + (1-\eta(X))\varphi(f^{(s)}(X)) \\
    = H_\eta(X)(f^{(s)}(X)) ,
\]

(3.1)

where, for any \( \eta \in (0,1) \) and \( \alpha \in \mathbb{R} \), \( H_\eta(\alpha) = \eta\varphi(-\alpha) + (1-\eta)\varphi(\alpha) \), so

\[
    \forall x \in \mathcal{X}, \quad f^*_\varphi(x) \in \arg\min_{\alpha \in \mathbb{R}} H_\eta(x)(\alpha) .
\]

As \( H_\eta \) is differentiable, the minimum is achieved when the derivative is null. As \( H'_\eta(\alpha) = -\eta\varphi'(-\alpha) + (1-\eta)\varphi'(\alpha) \), the condition \( H'_\eta(f^*_\varphi(x)) = 0 \) is equivalent to

\[
    \frac{\eta(x)}{1-\eta(x)} = \frac{\varphi'(f^*_\varphi(x))}{\varphi'(-f^*_\varphi(x))} .
\]

(3.2)

As \( \varphi \) is convex, \( \varphi' \) is non-decreasing. As \( \eta(x) > 1/2 \) is equivalent to \( \varphi'(f^*_\varphi(x)) > \varphi'(-f^*_\varphi(x)) \) by (3.2), it is also equivalent to \( f^*_\varphi(x) > 0 \). \( \Box \)

Perhaps an even more interesting link between the classification problem and its convex relaxation is provided by the following result, which links the excess risk of the hard classifier \( f^{(h)} \) associated to a soft classifier \( f^{(s)} \) with the \( \varphi \)-excess risk of \( f^{(s)} \).

**Lemma 27** (Zhang’s lemma). Let \( \varphi \) denote a non-decreasing, convex, non-negative function such that \( \varphi(0) = 1 \). Denote by \( \tau(\eta) = \inf_{\alpha \in \mathbb{R}} H_\eta(\alpha) \) and assume that there exist constants \( c > 0 \) and \( \gamma \in [0,1] \) such that

\[
    \forall \eta \in [0,1], \quad |\eta - 1/2| \leq c(1 - \tau(\eta))^\gamma .
\]

(3.3)
Then, for any soft classifier \( f(s) \), the associated hard classifier \( f(h) = (1 + \text{Sign}(f(s)))/2 \) satisfies

\[
E(f(h)) \leq 2eE_\phi(f(s))^{\gamma}, \quad E_\phi(f(s)) = \min_f [\ell_\phi; f] - \min_f [\ell_\phi, f].
\]

Before proceeding with the proof of Zhang’s lemma, it is interesting to check the condition (3.3) appearing in this lemma on convex surrogates of interests. Start with the hinge loss. In this case, the following result holds.

**Lemma 28.** Let \( \phi \) denote the hinge loss, that is \( \phi(x) = \max(1+x, 0) \). Then, Condition (3.3) holds with \( c = 1/2 \) and \( \gamma = 1 \).

**Proof.** By definition,

\[
H_\eta(\alpha) = \eta \max(1 - \alpha, 0) + (1 - \eta) \max(1 + \alpha, 0)
\]

\[
= \begin{cases} 
\eta(1 - \alpha) & \text{if } \alpha \leq -1 \\
\eta(1 - \alpha) + (1 - \eta)(1 + \alpha) & \text{if } -1 < \alpha < 1 \\
(1 - \eta)(1 + \alpha) & \text{if } \alpha \geq 1
\end{cases}.
\]

It follows that \( \tau(\eta) = 2 \min(\eta, (1 - \eta)) \). Hence, if \( \eta < 1/2 \),

\[
|\eta - 1/2| = 1/2 - \eta = \frac{1}{2}(1 - 2 \min(\eta, (1 - \eta))) = \frac{1}{2}(1 - \tau(\eta)).
\]

If \( \eta > 1/2 \),

\[
|\eta - 1/2| = \eta - 1/2 = \frac{1}{2}(1 - 2(1 - \eta)) = \frac{1}{2}(1 - 2 \min(\eta, (1 - \eta))) = \frac{1}{2}(1 - \tau(\eta)).
\]

This concludes the proof. \(\square\)

Consider as a second the logistic loss. In this case, the following result holds.

**Lemma 29.** Assume that \( \phi \) is the logistic function, that is \( \phi(x) = \log_2(1 + e^x) \). Then, (3.3) holds with \( \gamma = 1/2 \) and \( c = \sqrt{\log(2)/2} \).

**Proof.** By definition,

\[
H_\eta(\alpha) = \eta \log_2(1 + e^{-\alpha}) + (1 - \eta) \log_2(1 + e^\alpha).
\]

Deriving this expression yields

\[
H'_\eta(\alpha) = \frac{1}{\log 2} \frac{-\eta + (1 - \eta)e^\alpha}{e^\alpha + 1}.
\]
3.2. LINK WITH HARD CLASSIFIERS

It follows that \( H_\eta \) achieves its minimum at \( \alpha = \log(\eta/(1-\eta)) \). This minimum is equal to

\[
\tau(\eta) = H_\eta \left( \log \left( \frac{\eta}{1-\eta} \right) \right) = \eta \log_2 \left( 1 + \frac{1-\eta}{\eta} \right) + (1-\eta) \log_2 \left( 1 + \frac{\eta}{1-\eta} \right)
\]

\[
= -\eta \log_2(\eta) - (1-\eta) \log_2(1-\eta) .
\]

It follows that

\[
1 - \tau(\eta) = \frac{1}{\log 2} (-\log(1/2) + \eta \log(\eta) + (1-\eta) \log(1-\eta))
\]

\[
= \frac{1}{\log 2} \left( \eta \log \left( \frac{\eta}{1/2} \right) + (1-\eta) \log \left( \frac{1-\eta}{1/2} \right) \right) . \tag{3.4}
\]

At this point, the following result is useful.

**Lemma 30** (Pinsker’s inequality). Let \( P, Q \) denote two probability distributions. Denote by \( \mu \) a measure dominating both \( P \) and \( Q \) and let \( p, q \) denote the densities of \( P \) and \( Q \) respectively with respect to \( \mu \). Then

\[
\frac{1}{2} \left( \int |p - q| \, d\mu \right)^2 \leq \int p \log \left( \frac{p}{q} \right) \, d\mu .
\]

**Proof of Pinsker’s inequality.** Let

\[
\psi(x) = x \log(x) - x + 1 .
\]

Then, \( \psi(0) = 1, \psi(1) = 0, \psi'(1) = 0, \psi''(x) = 1/x \geq 0 \), so \( \psi(x) \geq 0 \) for any \( x > 0 \). Let also \( g(x) = (x-1)^2 - (4/3 + 2x/3) \psi(x) \). As \( g(1) = g'(1) = 0 \) and \( g''(x) = -4\psi(x)/3 \leq 0 \), it holds that, for any \( x > 0 \), there exists \( \xi \) such that \( |\xi - 1| < |x - 1| \) and

\[
g(x) = g(1) + g'(1)(x-1) + g''(\xi)(x-1)^2/2 \leq 0 .
\]

Therefore

\[
\forall x > 0, \quad (x-1)^2 \leq \left( \frac{4}{3} + \frac{2x}{3} \right) \psi(x) .
\]
It follows that
\[
\int |p - q| \, d\mu = \int q \left| \frac{p}{q} - 1 \right| \, d\mu \\
\leq \int q \sqrt{\left( \frac{4}{3} + \frac{2p}{3q} \right) \psi \left( \frac{p}{q} \right) \, d\mu} \\
\leq \sqrt{\int \left( \frac{4q}{3} + \frac{2p}{3} \right) \, d\mu} \sqrt{\int q \psi \left( \frac{p}{q} \right) \, d\mu} \\
= \sqrt{2} \sqrt{\int \left( p \log \left( \frac{p}{q} \right) - p + q \right) \, d\mu} \\
= \sqrt{2} \sqrt{\int p \log \left( \frac{p}{q} \right) \, d\mu}.
\]

Going back to (3.4), applying Pinsker’s inequality with the Bernoulli distributions
\[ P = \mathcal{B}(\eta), \quad Q = \mathcal{B}(1/2), \]
gives
\[
2(\eta - 1/2)^2 = \frac{1}{2} \left( \left| \eta - \frac{1}{2} \right| + \left| (1 - \eta) - \frac{1}{2} \right| \right)^2 \\
\leq \eta \log \left( \frac{\eta}{1/2} \right) + (1 - \eta) \log \left( \frac{1 - \eta}{1/2} \right).
\]
Therefore, \( 1 - \tau(\eta) \geq (2/ \log 2)(\eta - 1/2)^2 \). This concludes the proof of the lemma.

As Zhang’s condition is satisfied for standard convex relaxations, it is reasonable. One can now turn turn to the proof of Zhang’s Lemma.

\textbf{Proof of Zhang’s lemma.} Recall first that, from (3.1), for any \( f^{(s)} \),
\[
\mathbb{E}[\ell_{\varphi,f^{(s)}}(X,Y)|X] = H_{\eta(X)}(f^{(s)}(X)).
\]
In particular thus
\[
\mathbb{E}[\ell_{\varphi,f^{*}_x}(X,Y)|X] = \tau(\eta(X)).
\]
Therefore,
\[
P(\ell_{\varphi,f^{(s)}} - \ell_{\varphi,f^{*}_x}) = \mathbb{E}[H_{\eta(X)}(f^{(s)}(X)) - \tau(\eta(X))] . \quad (3.5)
\]
By the representation of the excess risk theorem
\[ \mathcal{E}(f^{(h)}) = \mathbb{E}[|2\eta(X) - 1|1_{\{f^{(h)}(X) \neq f^*(X)\}}] . \]
By definition of the hard classifier \( f^{(h)} \),
\[ 1_{\{f^{(h)}(X) \neq f^*(X)\}} = 1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}} , \]
so
\[ \mathcal{E}(f^{(h)}) = \mathbb{E}[|2\eta(X) - 1|1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}}] . \]
By Zhang’s condition \((\ref{eq:zhang-condition})\),
\[ \mathcal{E}(f^{(h)}) \leq 2c\mathbb{E}[(1 - \tau(\eta(X)))^{\gamma}1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}}] . \]
By Jensen’s inequality,
\[ \mathcal{E}(f^{(h)}) \leq 2c\mathbb{E}[(1 - \tau(\eta(X)))1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}]}^{\gamma} . \quad (3.6) \]
By \((\ref{eq:zhang-condition})\) and \((\ref{eq:zhang-condition})\), Zhang’s lemma is proved if, almost surely,
\[ (1 - \tau(\eta(X)))1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}} \leq H_{\eta}(X)(f^{(s)}(X)) - \tau(\eta(X)) . \quad (3.7) \]
Clearly, for any \( x \) such that \( f^{(s)}(x)(\eta(x) - 1/2) > 0 \), by definition of \( \tau \),
\[ (1 - \tau(\eta(x)))1_{\{f^{(s)}(x)(\eta(x) - 1/2) < 0\}} = 0 \leq H_{\eta}(x)(f^{(s)}(x)) - \tau(\eta(x)) . \]
Thus, it remains to show that, for any \( x \) such that \( f^{(s)}(X)(\eta(X) - 1/2) < 0 \),
\[ (1 - \tau(\eta(x)))1_{\{f^{(s)}(x)(\eta(x) - 1/2) < 0\}} = (1 - \tau(\eta(x))) \leq H_{\eta}(x)(f^{(s)}(x)) - \tau(\eta(x)) , \]
that is that
\[ H_{\eta}(x)(f^{(s)}(x)) \geq 1 . \]
By convexity of \( \varphi \),
\[ H_{\eta}(x)(f^{(s)}(x)) = \eta(x)\varphi(1-f^{(s)}(x))+(1-\eta(x))\varphi(f^{(s)}(x)) \geq \varphi((1-2\eta(x))f^{(s)}(x)) . \]
Therefore, if \( f^{(s)}(x)(\eta(x) - 1/2) < 0 \), as \( \varphi \) is non-decreasing
\[ H_{\eta}(x)(f^{(s)}(x)) \geq \varphi(0) = 1 . \]
This concludes the proof of Zhang’s Lemma. \[ \square \]
### 3.3 Bounding the $\varphi$-excess risk

Thanks to Zhang’s lemma, one can bound the excess risk of hard classifiers by bounding $\varphi$-excess risks of soft classifiers. Recall that from Chapters 1 and 2, by Vapnik’s lemma

$$E_{\ell_{\varphi}}(\hat{f}) \leq \inf_{f \in \mathcal{F}} E_{\ell_{\varphi}}(f) + 2 \sup_{f^{(s)} \in \mathcal{F}} |(P_N - P)\ell_{\varphi,f^{(s)}}| ,$$

Then, by the bounded difference inequality, $\ell_{\varphi,f^{(s)}}$ taking values in $[-1, 1]$, for any $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\sup_{f^{(s)} \in \mathcal{F}} |(P_N - P)\ell_{\varphi,f^{(s)}}| \leq \mathbb{E} \left[ \sup_{f^{(s)} \in \mathcal{F}} |(P_N - P)\ell_{\varphi,f^{(s)}}| \right] + \sqrt{\frac{2 \log(1/\delta)}{N}} .$$

Then, by the symmetrization lemma

$$\mathbb{E} \left[ \sup_{f^{(s)} \in \mathcal{F}} |(P_N - P)\ell_{\varphi,f^{(s)}}| \right] \leq 2R_N(\varphi, \mathcal{F}) ,$$

where

$$R_N(\varphi, \mathcal{F}) = \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \sup_{f^{(s)} \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \varphi(-y_i f^{(s)}(x_i)) \right| \right] .$$

At this point, it is useful to invoke the contraction inequality of Ledoux and Talagrand: If $\phi$ is $L$-Lipschitz and satisfies $\phi(0) = 0$, then for any set $B \subset \mathbb{R}^N$,

$$\mathbb{E} \left[ \sup_{b \in B} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \phi(b_i) \right| \right] \leq 2L \mathbb{E} \left[ \sup_{b \in B} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i b_i \right| \right] .$$

Applying this inequality to $\phi(\cdot) = \varphi(\cdot) - 1$ and assuming that $\varphi$ is $L$-Lipschitz as are both the hinge loss and the logistic loss, it holds $R_N(\varphi, \mathcal{F}) \leq 2LR_N(\mathcal{F})$, where

$$R_N(\mathcal{F}) \leq \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \sup_{f^{(s)} \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i y_i f^{(s)}(x_i) \right| \right] + \mathbb{E} \left[ \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \right| \right] \leq \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \sup_{f^{(s)} \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i f^{(s)}(x_i) \right| \right] + \frac{2\sqrt{2\pi}}{\sqrt{N}} .$$

Ultimately, the hard thresholded estimator $\hat{f} = (1 + \text{Sign}(\hat{f}_{\varphi}))/2$ satisfies

$$\mathcal{E}(\hat{f}) \leq 2c \left( \inf_{f \in \mathcal{F}} \mathcal{E}_{\ell_{\varphi}}(f) \right) + 8LR_N(\mathcal{F}) + \frac{4L\sqrt{2\pi}}{\sqrt{N}} + 2\sqrt{\frac{2 \log(1/\delta)}{N}} \right)^\gamma . \quad (3.8)$$
3.4 Boosting

As in Chapter 2, the tricky part is to bound from above the Rademacher complexity $R_N(\mathcal{F})$ in (3.8). Chapter 2 developed general approaches for this problem. The following sections provide other computations for widely used classes of soft classifiers.

3.4 Boosting

Boosting is an example of linear soft classifiers indexed by the $\ell_1$-ball $B_1$. Let $f_1, \ldots, f_d$ denote a collection of hard classifiers and let $\mathcal{F}$ denotes the convex hull of the classifiers $f_i, -f_i$:

$$\mathcal{F} = \{ \sum_{i=1}^{M} a_i f_i : \sum_{i=1}^{M} a_i \leq 1 \} .$$

In other words, defining the vector valued inputs

$$X_i = \begin{bmatrix} f_1(X_i) \\ \vdots \\ f_d(X_i) \end{bmatrix}, \quad \forall x \in \mathcal{X}, \quad x = \begin{bmatrix} f_1(x) \\ \vdots \\ f_d(x) \end{bmatrix}$$

and the convex set $C = B_1$, $\mathcal{F} = \{ a^T : a \in C \}$.

**Theorem 31.** The Rademacher complexity of the boosting class $\mathcal{F}$ satisfies

$$R_N(\mathcal{F}) \leq \sqrt{\frac{2 \log(4d)}{N}} .$$

In particular, if $\varphi$ is either the hinge loss or the logistic loss and $\hat{f}$ is the hard classifier associated to the $\varphi$-ERM $\hat{f}_\varphi$ over $\mathcal{F}$, it holds

$$\mathcal{E}(\hat{f}) \leq 2c \left( \inf_{f \in \mathcal{F}} \{ \mathcal{E}_\varphi(f) \} + 8L \sqrt{\frac{2 \log(2d)}{N}} + 2 \sqrt{\frac{2 \log(1/\delta)}{N}} \right)^\gamma .$$

**Proof.** By definition,

$$R_N(\mathcal{F}) = \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \sup_{a \in B_1} \left| a^T \left( \frac{1}{N} \sum_{i=1}^{N} \epsilon_i X_i \right) \right| \right] .$$

As, for any vectors $a, b$ in $\mathbb{R}^d$, it holds $a^T b \leq \|a\|_1 \|b\|_\infty$,$$\sup_{a \in B_1} \left| a^T \left( \frac{1}{N} \sum_{i=1}^{N} \epsilon_i X_i \right) \right| \leq \left\| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i X_i \right\|_\infty = \max_{b \in B_1} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i b_i \right|,$$
where \( B_z \) is the set of \( d \) vectors
\[
B_z = \left\{ \begin{bmatrix} e_i^T X_1 \\ \vdots \\ e_i^T X_N \end{bmatrix}, i = 1, \ldots, d \right\}.
\]
The vectors \( b \in B_z \) have coordinates at most 1, so \( \|b\|^2 \leq N \), hence \( \Delta(B_z) \leq \sqrt{N} \), thus, by Lemma \( \square \) it follows that
\[
\mathbb{E} \left[ \max_{b \in B_z} \left\{ \frac{1}{N} \sum_{i=1}^{N} \epsilon_i b_i \right\} \right] \leq \sqrt{\frac{2 \log(2d)}{N}}.
\]
As this holds for any \( z \in \mathcal{Z}^N \), this proves the first part of the theorem, the second part comes from Eq (3.8).

### 3.5 Support Vector Machine

In this section \( \phi \) denotes the hinge loss \( \phi(x) = \max(1 + x, 0) \) and \( F \) is the unit ball of a Hilbert space \( W \).

#### 3.5.1 Reproducing kernel Hilbert space

Start with the definition of positive semi-definite kernels.

**Definition 32** (PSD kernels). A function \( K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is called a positive semi-definite kernel (PSD kernel) if, for any \( x_1, \ldots, x_N \in \mathcal{X} \), the matrix \( K \) such that \( K_{i,j} = K(x_i, x_j) \) is symmetric, positive semi-definite, that is if \( K^T = K \) and, for any \( a \in \mathbb{R}^N \), \( a^T K a \geq 0 \).

Let us provide some examples of PSD kernels on \( \mathbb{R}^d \).

**Example 1.** The function \( K(x, y) = x^T y \) is a PSD kernel on \( \mathbb{R}^d \).

**Example 2.** The function \( K(x, y) = e^{-\|x-y\|^2/(2\sigma^2)} \) is a PSD kernel on \( \mathbb{R}^d \).

(Indication: prove it by recurrence, applying Gauss reduction algorithm for heredity).

**Definition 33.** Let \( W \) denote a Hilbert space of functions \( f : \mathcal{X} \to \mathbb{R} \). A symmetric kernel \( K \) is called reproducing kernel of \( W \) if

(i) for any \( x \in \mathcal{X}, K(x, \cdot) \in W \),

(ii) for any \( x \in \mathcal{X} \) and any \( f \in W \), \( \langle f, K(x, \cdot) \rangle_W = f(x) \).
If \( W \) admits a reproducing kernel \( K \), it is called a reproducing kernel Hilbert space, with kernel \( K \).

The following lemma makes a link between these notions.

**Lemma 34.** A reproducing kernel is positive semi-definite.

**Proof.** Let \( K \) denote a reproducing kernel, \( x \in \mathcal{X}^N \) and \( \mathbb{K} \) denote the associated matrix with entries \( \mathbb{K}_{ij} = K(x_i, x_j) \). Let \( a \in \mathbb{R}^N \). By the reproducing property

\[
 a^T \mathbb{K} a = \left( \sum_{i=1}^{N} a_i K(x_i, \cdot) \right)^2 = \left\| \sum_{i=1}^{N} a_i K(x_i, \cdot) \right\|_W^2 \geq 0 .
\]

**Example 3.** If \( \varphi_1, \ldots, \varphi_M \) denote orthonormal functions in \( L^2(\mu) \), then \( K(x,y) = \sum_{i=1}^{M} \varphi_i(x) \varphi_i(y) \) is a reproducing kernel in the space \( W = \{ f = \sum_{i=1}^{M} a_i \varphi_i, \ a \in \mathbb{R}^M \} \) endowed with the \( L^2(\mu) \) inner product.

It is sufficient to check that \( K \) is a PSD kernel satisfying the reproducing property. First,

\[
 a^T \mathbb{K} a = \sum_{j=1}^{M} \left( \sum_{i=1}^{N} a_i \varphi_j(x_i) \right)^2 \geq 0 .
\]

Hence, \( K \) is PSD. Second, let \( f \in W \), so \( f(x) = \sum_{i=1}^{M} a_i \varphi_i \). Thus

\[
 \langle f, K(x, \cdot) \rangle = \sum_{i=1}^{M} \sum_{j=1}^{M} a_i \varphi_j(x) \varphi_i(\varphi_i)_{L^2(\mu)} = \sum_{i=1}^{M} a_i \varphi_i(x) = f(x) .
\]

**Example 4.** \( K(x,y) = x^T y \) is a reproducing kernel of \( W = \{ w^T, \ w \in \mathbb{R}^d \} \) equipped with the inner product \( \langle w^T, (w')^T \rangle = w^T w' \).

The sup-norm in the RKHS is bounded from above by the norm of the RKHS as shown by the following proposition.

**Proposition 35.** Let \( W \) denote a RKHS with PSD kernel \( K \) such that \( \sup_{x \in \mathcal{X}} K(x, x) = k_\infty < \infty \). Then,

\[
 \forall f \in W, \quad \sup_{x \in \mathcal{X}} |f(x)| \leq \|f\|_W \sqrt{k_\infty} .
\]
Proof. Use the reproducing property to write
\[ |f(x)| = \langle f, K(x, \cdot) \rangle_W . \]

Then by Cauchy-Schwarz inequality,
\[ |f(x)| \leq \|f\|_W \sqrt{\langle K(x, \cdot), K(x, \cdot) \rangle_W} . \]

By the reproducing property,
\[ \langle K(x, \cdot), K(x, \cdot) \rangle_W = K(x, x) \leq k_\infty . \]

\[ \square \]

### 3.5.2 Representer theorem

The key to compute the \( \varphi \)-ERM \( \hat{f}_\varphi \) over the unit ball of the possibly infinite dimensional space \( W \) is that it actually takes value in a finite dimensional space thanks to the following result.

**Theorem 36** (Representer theorem). Let \( W \) denote a RKHS with PSD \( K \) and let \( G : \mathbb{R}^N \to \mathbb{R} \) denote any function. Let \( x_1, \ldots, x_N \in X \). For any \( r > 0 \), let \( B_W(r) = \{ f \in W : \|f\|_W \leq r \} \), let \( W_N = \{ f_a = \sum_{i=1}^N a_i K(x_i, \cdot), \ a \in \mathbb{R}^N \} \) and \( B_{W_N}(r) = \{ f \in W_N : \|f\|_W \leq r \} \). Then

\[
\min_{f \in B_W(r)} G(f(x_1), \ldots, f(x_N)) = \min_{f \in B_{W_N}(r)} G(f(x_1), \ldots, f(x_N))
\]

\[
= \min_{a \in \mathbb{R}^N : a^T K a \leq r^2} G(f_a(x_1), \ldots, f_a(x_N)).
\]

**Proof.** Let \( f \in W \), and write \( f = f_N + f_N^\perp \) with \( f_N \in W_N \) and \( f_N^\perp \in W_N^\perp \). Since \( K(x_i, \cdot) \in W_N \), it holds \( f^T(x_i) = \langle f^T, K(x_i, \cdot) \rangle_W = 0 \) so \( f(x_i) = f_N(x_i) \). Moreover, by Pythagoras theorem \( \|f_N\|_W \leq \|f\|_W \). Hence, for any \( f \in B_W(r) \)

\[ G(f(x_1), \ldots, f(x_N)) = G(f_N(x_1), \ldots, f_N(x_N)) , \]

with \( f_N \in W_N \) and \( \|f_N\|_W \leq r \), so

\[ \min_{f \in B_W(r)} G(f(x_1), \ldots, f(x_N)) \geq \min_{f \in B_{W_N}(r)} G(f(x_1), \ldots, f(x_N)) . \]

Conversely, \( B_{W_N}(r) \subset B_W(r) \) so

\[ \min_{f \in B_W(r)} G(f(x_1), \ldots, f(x_N)) \leq \min_{f \in B_{W_N}(r)} G(f(x_1), \ldots, f(x_N)) . \]
3.5. SUPPORT VECTOR MACHINE

This shows the first part of the theorem. The second is a direct consequence of the reproducing property, since

$$\|f_a\|^2_W = \sum_{1 \leq i,j \leq N} a_i a_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_W = a^T K a .$$

$\square$

3.5.3 Excess risk of $\phi$-ERM

Remark that SVM algorithm can be recast as a $\ell_2$-aggregation problem. Consider as input variables

$$\forall i \in \{1, \ldots, N\}, \quad \mathbf{X}_i = \begin{bmatrix} K(X_1, X_i) \\ \vdots \\ K(X_N, X_i) \end{bmatrix}$$

and, as input space the set of vectors $\mathbf{x} \in \mathbb{R}^d$ such that

$$\forall x \in \mathcal{X}, \quad \mathbf{x} = \begin{bmatrix} K(X_1, x) \\ \vdots \\ K(X_N, x) \end{bmatrix}$$

Define also $\mathcal{B}_W(r) = \{f \in W : \|f\|_W \leq r\}$,

$$\mathcal{B}_{W_N}(r) = \{f \in W_N : \|f\|_W \leq r\} = \{a^T \cdot, a^T K a \leq r^2\} .$$

By the representer theorem, it holds that

$$\argmin_{f \in \mathcal{B}_W(r)} P_N \ell_{\phi,f}(X_i, Y_i) = \argmin_{f \in \mathcal{B}_{W_N}(r)} P_N \ell_{\phi,f}(X_i, Y_i)$$

The following theorem computes the Rademacher complexity of balls in RKHS, from which an excess risk bound for the hard classifier associated with the $\phi$-ERM over this ball is obtained.

**Theorem 37.** Let $W$ denote a RKHS with PSD $K$ such that $k_\infty < \infty$ and let $\mathcal{F} = \mathcal{B}_W(r)$. Then

$$R_N(\mathcal{F}) \leq r \frac{\sqrt{\text{Tr}(K)}}{N} \leq r \sqrt{\frac{k_\infty}{N}} .$$

Denote by $\hat{f} = (1 + \text{Sign}(\hat{f}_\phi))/2$ the hard classifier associated with the $\phi$-ERM

$$\hat{f}_\phi \in \arg\min_{f \in \mathcal{B}_W(r)} P_N \ell_{\phi,f} .$$

For any $\delta \in (0,1)$, with probability $1 - \delta$,

$$\mathcal{E}(\hat{f}) \leq \inf_{f \in \mathcal{F}} \{\mathcal{E}_{\ell_2}(f)\} + 8 \sqrt{\frac{k_\infty}{N}} + 2 \sqrt{\frac{2 \log(1/\delta)}{N}} .$$
Proof. By Cauchy-Schwarz inequality,

\[ R_N(\mathcal{F}) = \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \sup_{f \in \mathcal{B}_W(r)} \left\langle \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot), f \right\rangle_W \right] \]

\[ \leq r \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot) \right\|_W^2 \right] \]

Moreover, by the representing property

\[ \mathbb{E} \left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot) \right\|_W^2 \right] = \frac{1}{N^2} \sum_{1 \leq i, j \leq N} K(x_i, x_j) \mathbb{E}[\epsilon_i \epsilon_j] = \frac{\text{Tr}(\mathbb{K})}{N^2} \leq \frac{k_\infty}{N} . \]

This shows the first part of the theorem. The second part comes from (3.8) combined with the values of \( c \) and \( \gamma \) obtained for the hinge loss right after Zhang’s lemma. \[\square\]
Chapter 4

Convex optimization

This chapter presents various algorithms to solve convex problems. In machine learning, these algorithms are useful to solve the following problems:

\[
\min_{f \in \mathcal{F}} \Theta_N(f), \quad \min_{f \in \mathcal{F}} \Theta(f),
\]

where \( \mathcal{F} \) is a convex set of soft classifiers and, for a convex surrogate \( \varphi \)

\[
\Theta_N(f) = P_N \ell_{\varphi,f} = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y_if(X_i)), \quad \Theta(f) = P\ell_{\varphi,f} = \mathbb{E}[\varphi(-Yf(X))].
\]

There is an important difference between these problems for machine learning. In the first case, convex (deterministic) optimization algorithms are used to obtain an approximation \( \hat{f}_\varphi^{(\epsilon)} \) of the ERM \( \hat{f}_\varphi \) with a deterministic control of the approximation error by \( \epsilon > 0 \). The risk of this \( \epsilon \)-ERM, \( \hat{f}_\varphi^{(\epsilon)} \) estimator is then bounded from above by the sum of \( \epsilon \) (that depends on the number of iterations of the algorithm) and the estimation error of the ERM (that has been bounded from above in the previous chapters). In the second case, stochastic convex approximation algorithms directly produce estimators \( \hat{f} \). The approximation error of these algorithms is the \( \varphi \)-excess risk of \( \hat{f} \). Therefore, stochastic convex optimization does not rely on ERM theory. This error will be bounded in expectation only and not with high probability. There is another important difference between these approaches, while deterministic optimization requires an access to all data simultaneously, these can be made available on line for stochastic approximation, which considerably saves memory resources when handling very large datasets.
4.1 Convex problems

Definition 38 (Convex problems). An optimization problem of the form 
\[ \min_{x \in C} \theta(x) \] is called a convex problem if \( C \) is a convex set and \( \theta \) is a convex 
function.

Assume that the objective function \( \theta \) is defined on a domain \( D \subset \mathbb{R}^d \) and 
that there exists a vector \( c \) such that, for all \( x \in D \), \( \theta(x) = c^T x \). In this 
case, even if \( D \) is not a convex set, the problem \( \min_{x \in D} \theta(x) \) is equivalent to 
a convex problem. To see why, introduce the following definition.

Definition 39 (Convex hull). Let \( D \subset \mathbb{R}^d \). The convex hull of \( D \), \( H(D) \) is 
the set of convex combinations of elements of \( D \):

\[ H(D) = \left\{ \sum_{i=1}^{N} a_i x_i, \ x_i \in D, a \in \Delta_N \right\}. \]

The convex hull of any set \( D \subset \mathbb{R}^d \) is the smallest convex containing \( D \) 
and the following result holds.

Proposition 40. For any \( c \in \mathbb{R}^d \) and any non empty subset \( D \subset \mathbb{R}^d \),

\[ \min_{x \in D} c^T x = \min_{x \in H(D)} c^T x. \]

Proof. Since \( D \subset H(D) \), it is clear that

\[ \min_{x \in D} c^T x \geq \min_{x \in H(D)} c^T x. \]

Hence, it is sufficient to prove the reverse inequality. Let \( x \in H(D) \), by 
definition, there exist \( x_1, \ldots, x_N \) in \( D \) and \( a \in \Delta_N \) such that \( x = \sum_{i=1}^{N} a_i x_i \).
Thus

\[ c^T x = \sum_{i=1}^{N} a_i c^T x_i. \]

Now, all \( c^T x_i \geq \min_{x \in D} c^T x \). As \( a \in \Delta_N \), all \( a_i \geq 0 \) and \( \sum_{i=1}^{N} a_i = 1 \), so

\[ c^T x \geq \sum_{i=1}^{N} a_i \min_{x \in D} c^T x = \min_{x \in D} c^T x. \]

As this holds for all \( x \in H(D) \), it follows that

\[ \min_{x \in H(D)} c^T x \geq \min_{x \in D} c^T x. \]

This concludes the proof of the second inequality, hence, the proof of the 
proposition. \( \square \)
A fundamental tool in this chapter is the notion of sub-gradient.

**Definition 41** (sub-gradient). Let $D \subset \mathbb{R}^d$ and let $\theta : D \to \mathbb{R}$. A vector $g \in \mathbb{R}^d$ is called a sub-gradient of $\theta$ at $x \in D$ if
\[
\forall y \in D, \quad \theta(x) - \theta(y) \leq g^T(x - y) .
\]
The set of sub-gradients of $\theta$ at $x$ is called the sub-differential of $\theta$ at $x$ and is denoted by $\partial \theta(x)$.

Sub-gradients play the roles of gradients but, contrary to gradients, they are well defined at any point $x$ if $\theta$ is convex as shown by the following (admitted) result.

**Theorem 42.** If $\theta : C \to \mathbb{R}$ is convex, then, for any $x \in C$, $\partial \theta(x) \neq \emptyset$. Moreover, if $\theta$ is differentiable at $x \in C$, then $\partial \theta(x) = \{\nabla \theta(x)\}$.

The following result shows that, for convex functions, it is sufficient to show that 0 is a sub-gradient of $\theta$ at $x$ to prove that $x$ is a minimizer of $\theta$.

**Theorem 43.** Let $\theta$ denote a convex function over a convex domain $C$. The following are equivalent.

(i) $x$ is a global minimum of $\theta$.

(ii) $x$ is a local maximum of $\theta$.

(iii) $0 \in \partial \theta(x)$.

**Proof.** By definition of sub-differentials, $0 \in \partial \theta(x)$ if and only if
\[
\forall y \in C, \quad \theta(x) - \theta(y) \leq 0 ,
\]
that is if and only if $x$ is a global minimum of $\theta$.

Moreover, if $x$ is a local minimum of $\theta$, there exists $r > 0$ such that, $\forall y \in C$ satisfying $\|y - x\| \leq r$, $\theta(x) \leq \theta(y)$. Let $y' \in C$, there exists $\epsilon \in (0, 1)$ such that $y = x + \epsilon(y' - x) \in C$ and satisfies $\|y - x\| \leq r$. Therefore
\[
\theta(y) \geq \theta(x) .
\]
By convexity of $\theta$,
\[
\theta(y) = \theta(\epsilon y' + (1 - \epsilon)x) \leq \epsilon \theta(y') + (1 - \epsilon)\theta(x) .
\]
Thus
\[
\epsilon \theta(y') + (1 - \epsilon)\theta(x) \geq \theta(x) ,
\]
which is equivalent to
\[
\theta(y') \geq \theta(x) .
\]
As this holds for any $y' \in C$, this shows that $x$ is a global minimizer of $\theta$. 

4.2 Gradient descent

Informally, if \( \|y - x\| \) is small then \( \theta(y) \) is close to its linear approximation

\[
\theta(y) \approx \theta(x) + \nabla \theta(x)^T (y - x).
\]

Therefore, if, for a small \( \eta > 0 \), \( y = x - \eta \nabla \theta(x) \)

\[
\theta(y) \approx \theta(x) - \eta \| \nabla \theta(x) \|^2 \leq \theta(x).
\]

Gradient descent exploits this heuristics, building recursively a sequence of approximate minimizers of \( \theta \) as follows.

**Algorithm 1:** Gradient descent (GD).

```
input : x_1 \in C: initial point, (\eta_t)_t: step sizes, T: stopping time
1 for t = 1 to t = T, do
2 \[ \text{find } g_t \in \partial \theta(x_t), \]
3 \[ x_{t+1} = x_t - \eta_t g_t. \]
4 end
5 Return \( x_T = T^{-1} \sum_{t=1}^{T} x_t \) or \( x_T \in \text{argmin}_{x \in \{x_1, \ldots, x_T\}} \theta(x). \)
```

Theorem 44. Assume that \( \theta \) is convex, \( L \)-Lipschitz on \( \mathbb{R}^d \) and denote by \( x^* \in \text{argmin}_{x \in \mathbb{R}^d} \theta(x) \). If \( \|x_1 - x^*\| \leq R \) and, for any \( t \geq 1 \), \( \eta_t = \eta = R/(L\sqrt{T}) \), then, the outputs of GD satisfy

\[
\theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}, \quad \theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}.
\]

Proof. Let \( t \in \{1, \ldots, T\} \), then, by definition of a sub-gradient,

\[
\theta(x_t) - \theta(x^*) \leq g_t^T (x_t - x^*).
\]

By definition of the gradient descent algorithm, \( g_t = \eta_t^{-1}(x_{t+1} - x_t) \), so

\[
\theta(x_t) - \theta(x^*) \leq \frac{1}{\eta_t} (x_{t+1} - x_t)^T (x_t - x^*) \leq \frac{1}{\eta_t} (x_{t+1} - x_t)^T (x_t - x^*) \leq \frac{1}{2\eta_t} (\|x_{t+1} - x_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2).
\]

Using the equality \( a^T b = (\|a\|^2 + \|b\|^2 - \|a - b\|^2)/2 \), it follows that

\[
\theta(x_t) - \theta(x^*) \leq \frac{1}{2\eta_t} (\|x_{t+1} - x_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2).
\]
4.2. GRADIENT DESCENT

By definition of the gradient descent algorithm, it follows that
\[ \theta(x_t) - \theta(x^*) \leq \frac{1}{2\eta_t}(\eta_t^2 \|g_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) .\]

As \( \theta \) is \( L \)-Lipschitz, the sub-gradient \( g_t \) has \( L^2 \)-norms bounded from above by \( L \), thus
\[ \eta_t(\theta(x_t) - \theta(x^*)) \leq \frac{1}{2}(\eta_t^2 L^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) . \tag{4.1} \]

Summing up these inequalities and dividing by \( \sum_{t=T_0}^T \eta_t \) yields
\[ \sum_{t=T_0}^T \eta_t(\theta(x_t) - \theta(x^*)) \leq \frac{1}{2} \left( \frac{L^2 \sum_{t=T_0}^T \eta_t^2}{\sum_{t=T_0}^T \eta_t} + \frac{\|x_1 - x^*\|^2}{\sum_{t=T_0}^T \eta_t} \right) . \]

By convexity of \( \theta \), this yields
\[ \theta\left( \frac{\sum_{t=T_0}^T \eta_t x_t}{\sum_{t=T_0}^T \eta_t} \right) - \theta(x^*) \leq \frac{1}{2} \left( \frac{L^2 \sum_{t=T_0}^T \eta_t^2}{\sum_{t=T_0}^T \eta_t} + \frac{R^2}{\sum_{t=T_0}^T \eta_t} \right) . \tag{4.2} \]

Taking \( T_0 = 1 \) and \( \eta_t = \eta \) as in the theorem yields
\[ \theta(x_t) - \theta(x^*) \leq \frac{1}{2} \left( L^2 \eta + \frac{R^2}{\eta T} \right) . \]

With the choice of \( \eta \), this gives
\[ \theta(x_t) - \theta(x^*) \leq \frac{LR}{\sqrt{T}} . \]

Moreover, going back to (4.1), we get
\[ \eta(\theta(x_T) - \theta(x^*)) \leq \frac{1}{2} \left( L^2 \eta^2 + \frac{R^2}{T} \right) = \frac{LR}{\sqrt{T}} . \]

\[ \square \]

**Remark 45.** Remark that, from (4.2), the choice \( \eta_t = GR/(L\sqrt{T}) \) and \( T_0 = T/2 \) would give
\[ \theta\left( \frac{\sum_{t=T_0}^T \eta_t x_t}{\sum_{t=T_0}^T \eta_t} \right) - \theta(x^*) \leq \frac{1}{2} \left( \frac{L^2 \sum_{t=T/2}^T \eta_t^2}{\sum_{t=T/2}^T \eta_t} + \frac{R^2}{\sum_{t=T/2}^T \eta_t} \right) \leq C(G) \frac{RL}{\sqrt{T}} . \]

Hence, it is possible to build an estimator with similar rates of convergence in terms of \( R, L \) and \( T \) with step sizes independent of \( T \).
4.3 Projected gradient descent

If one want to minimize a function \( \theta \) over a closed convex subset \( C \) of \( \mathbb{R}^d \), it seems natural to build a sequence of estimates taking values in \( C \). In general, gradient descent algorithms may build sequences taking values outside \( C \). To avoid this, one can at each step choose \( x_{t+1} \) as the projection of \( x_t - \eta g_t \) onto \( C \). For this to make sense, let us first show that such a projection exists.

**Theorem 46 (Projection onto closed convex sets).** Let \( C \) denote a closed convex subset of \( \mathbb{R}^d \). For any \( x \in \mathbb{R}^d \), there exists an element \( \pi_C(x) \in C \) such that

\[
\forall y \in C, \quad \|x - \pi_C(x)\| \leq \|x - y\| .
\]

Moreover, \( \pi_C(x) \) is the unique element of \( C \) such that

\[
\forall y \in C, \quad \langle \pi_C(x) - x, \pi_C(x) - y \rangle \leq 0 .
\]

**Proof.** Let \( x_0 \in C, r = \|x - x_0\| \) and \( B = \{ y \in \mathbb{R}^d : \|y - x\| \leq r \} \). The set \( C \cap B \) is compact and, clearly

\[
\min_{y \in \mathbb{C}} \|x - y\| = \min_{y \in C \cap B} \|x - y\| .
\]

As the function \( \|x - \cdot\| \) is continuous on the compact \( C \cap B \), it achieves its minimum and there exists a minimizer \( \pi_C(x) \) of this function. Moreover, for any \( y \in C \) and \( t \in (0, 1) \), \( \pi_C(x) - t(\pi_C(x) - y) \in C \), so

\[
\|x - \pi_C(x)\|^2 \leq \|x - \pi_C(x) + t(\pi_C(x) - y)\|^2 = \|x - \pi_C(x)\|^2 + t^2\|\pi_C(x) - y\|^2 - 2t\langle \pi_C(x) - x, \pi_C(x) - y \rangle .
\]

Therefore, for any \( t \in (0, 1) \),

\[
\langle \pi_C(x) - x, \pi_C(x) - y \rangle \leq t\|\pi_C(x) - y\|^2 .
\]

Letting \( t \to 0 \) shows that

\[
\langle \pi_C(x) - x, \pi_C(x) - y \rangle \leq 0 .
\]

Imagine now that there exists a point \( \pi \in C \) such that

\[
\forall y \in C, \quad \langle \pi - x, \pi - y \rangle \leq 0 .
\]

Thus, in particular,

\[
\langle \pi_C(x) - x, \pi_C(x) - \pi \rangle \leq 0 ,
\]

\[
\langle x - \pi, \pi_C(x) - \pi \rangle \leq 0 .
\]
Summing these inequalities shows that
\[ \|p_C(x) - p\|^2 \leq 0. \]
This shows unicity of \( p_C(x) \) and concludes the proof of the theorem.

We can now provide the “projected” gradient descent algorithm.

```
input : \( x_1 \in C \): initial point, \((\eta_t)\): step sizes, \( T \): stopping time
1 for \( t = 1 \) to \( t = T \), do
  2 find \( g_t \in \partial \theta(x_t) \),
  3 \( y_{t+1} = x_t - \eta_t g_t \),
  4 \( x_{t+1} = p_C(y_{t+1}) \).
end
6 Return \( x_T = T^{-1} \sum_{i=1}^{T} x_i \) or \( x_T \in \text{argmin}_{x \in \{x_1, \ldots, x_T\}} \theta(x) \).

Algorithm 2: Projected gradient descent (PGD).
```

The following result shows performance of the projected gradient descent algorithm.

**Theorem 47.** Let \( C \) denote a compact non-empty subset of \( \mathbb{R}^d \) with diameter smaller than \( R \). Assume that \( \theta \) is convex, \( L \)-Lipschitz and denote by \( x^* \in \text{argmin}_{x \in C} \theta(x) \). If, for any \( t \geq 1 \), \( \eta_t = \eta = R/(L\sqrt{T}) \), then, the outputs of PGD satisfy
\[ \theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}, \quad \theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}. \]

**Proof.** The proof starts as the one for gradient descent.
\[
\theta(x_t) - \theta(x^*) \leq g_t^T (x_t - x^*)
\]
\[
= \frac{1}{\eta} (x_t - y_{t+1})^T (x_t - x^*)
\]
\[
= \frac{1}{\eta} \left( \|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x^* - y_{t+1}\|^2 \right) \quad (4.3)
\]
Now,
\[ \|x^* - y_{t+1}\|^2 = \|x^* - x_{t+1}\|^2 + \|x_{t+1} - y_{t+1}\|^2 - 2 \langle x_{t+1} - x^*, x_{t+1} - y_{t+1} \rangle. \]
As \( x^* \in C \) and \( x_{t+1} = p_C(y_{t+1}) \), by Theorem 46,
\[ \langle x_{t+1} - x^*, x_{t+1} - y_{t+1} \rangle \leq 0, \]
hence,
\[
\|x^* - y_{t+1}\|^2 \geq \|x^* - x_{t+1}\|^2 + \|x_{t+1} - y_{t+1}\|^2 \geq \|x^* - x_{t+1}\|^2 .
\]

Plugging this bound into (4.3) shows that
\[
\theta(x_t) - \theta(x^*) \leq \frac{1}{\eta} \left( \eta^2 \|g_t\|^2 + \|x_t - x^*\|^2 - \|x^* - x_{t+1}\|^2 \right) .
\]

Summing up yields
\[
\frac{1}{T} \sum_{t=1}^{T} \theta(x_t) - \theta(x^*) \leq \frac{1}{\eta} \left( \eta^2 L^2 + \frac{\|x_1 - x^*\|^2}{T} \right) \\
\leq \frac{1}{\eta} \left( \eta^2 L^2 + \frac{R^2}{T} \right) = \frac{RL}{T} .
\]

Then both \( \theta(x_T) \) (by convexity of \( \theta \)) and \( \theta(x_T) \) (by definition) are smaller than \( T^{-1} \sum_{t=1}^{T} \theta(x_t) \), which concludes the proof.

\[\Box\]

### 4.4 An exercise

In this section, we assume that \( \Theta \) is strongly convex with respect to the Euclidean norm \( \| \cdot \| \), i.e. that there exists \( \nu > 0 \) such that
\[
\forall x, y \in \mathcal{X}, \quad \Theta(y) - \Theta(x) \leq \langle \nabla \Theta(y), y - x \rangle - \frac{\nu}{2} \| y - x \|^2 .
\]

Moreover, we also assume that \( \nabla \Theta \) is \( \beta \)-Lipschitz, i.e. that
\[
\forall x, y \in \mathcal{X}, \quad \| \nabla \Theta(x) - \nabla \Theta(y) \| \leq \beta \| x - y \| .
\]

Fix \( x \in \mathcal{X}, \eta > 0 \) and a closed convex set \( \mathcal{C} \). Let \( N(x) = x - \eta \nabla \Theta(x), S(x) = \pi_{\mathcal{C}}(N(x)), g(x) = \eta^{-1}(x - S(x)) \).

1. Prove that
\[
\|S(x) - x^*\|^2 = \|x - x^*\|^2 - 2\eta \langle g(x), x - x^* \rangle + \eta^2 \|g(x)\|^2 .
\]

2. Prove that
\[
\Theta(y) - \Theta(x) \leq \langle \nabla \Theta(x), y - x \rangle + \frac{\beta}{2} \| y - x \|^2 .
\]
4.5. **EXAMPLES**

3. Prove that
   \[
   \Theta(S(x)) - \Theta(y) \leq \langle \nabla \Theta(x), S(x) - x \rangle + \frac{\beta}{2} \|S(x) - x\|^2 + \langle \nabla \Theta(x), x - y \rangle - \frac{\nu}{2} \|y - x\|^2 .
   \]

4. Prove that
   \[
   \langle \nabla \Theta(x), S(x) - y \rangle \leq \langle g(x), S(x) - y \rangle .
   \]

5. Prove that
   \[
   \Theta(S(x)) - \Theta(y) \leq \langle g(x), x - y \rangle - \frac{\eta}{2} \|g(x)\|^2 - \frac{\nu}{2} \|y - x\|^2 .
   \]

6. Prove that there exists \( \rho > 0 \) such that
   \[
   \|S(x) - x^*\|^2 \leq (1 - \rho)\|x - x^*\|^2 .
   \]

7. Discuss the convergence rate of PGD.

### 4.5 Examples

#### 4.5.1 SVM

Projected gradient descent is very well adapted to minimize functions over convex sets described by Euclidean geometries such as ellipsoids. This is the case, typically, of SVM where one has to minimize

\[
\min_{\alpha \in \mathbb{R}^N, \alpha^T \mathbf{K} \alpha \leq r^2} \Theta_N(\alpha), \quad \text{where} \quad \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - Y_i^{(s)} \alpha^T \mathbf{K} \mathbf{e}_i) .
\]

We have

\[
\nabla \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{1 - Y_i^{(s)} \alpha^T \mathbf{K} \mathbf{e}_i \geq 0\}} Y_i^{(s)} \mathbf{K} \mathbf{e}_i ,
\]

therefore,

\[
\|\nabla \Theta_N(\alpha)\| \leq \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{K} \mathbf{e}_i\| .
\]

Now, as, for any \( g \in W, \sup_{x \in \mathcal{X}} |g(x)| \leq \|g\|_{W} \sqrt{K_\infty}, \) and \( \|K(x, \cdot)\|^2_W = K(x, x) \leq k_\infty, \)

\[
\|\mathbf{K} \mathbf{e}_i\| = \sqrt{\sum_{j=1}^{N} K(X_i, X_j)^2} \leq \sqrt{\sum_{j=1}^{N} \|K(X_i, \cdot)\|^2_W k_\infty} \leq k_\infty \sqrt{N} .
\]
Therefore, $\Theta_N$ is $L$-Lipschitz, with $L = k_{\infty}\sqrt{N}$. To apply our general bound, it is necessary to compute the diameter $R$ of the set $C = \{\alpha \in \mathbb{R}^N : \alpha^T K \alpha \leq r^2\}$. Denote by $\lambda_{\min}(K)$ the smallest eigenvalue of the positive matrix $K$. For any $\alpha \in C$,

$$r^2 \geq \alpha^T K \alpha \geq \lambda_{\min}(K) \alpha^T \alpha.$$ 

Hence,

$$R = 2 \sup_{\alpha \in C} \sqrt{\alpha^T \alpha} \leq \frac{2r}{\sqrt{\lambda_{\min}(K)}}.$$ 

By Theorem 47, it follows that

$$\Theta_N(\bar{f}_T) - \Theta_N(\bar{f}_\varphi) \leq \frac{LR}{\sqrt{T}} = \frac{2r k_{\infty} \sqrt{N}}{\sqrt{T} \sqrt{\lambda_{\min}(K)}}.$$ 

This error is upper bounded by $1/N$ if the number of steps in the PGD algorithm satisfies

$$T \geq \frac{4r^2 k_{\infty}^2 N^3}{\lambda_{\min}(K)}.$$ 

### 4.5.2 Boosting

On the other hand, PGD is much less efficient to minimize functions on convex sets described by non-Euclidean geometries. To illustrate this, consider the boosting example where one has to minimize

$$\min_{\alpha \in B_1} \Theta_N(\alpha), \quad \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y_i(\alpha^T X_i)).$$

The sub-gradients of $\Theta_N$ are easy to compute

$$g_{\Theta_N}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} g_{\varphi}(-Y_i(\alpha^T X_i))(-Y_i(\alpha^T X_i)).$$

Recall that $X_i$ is a vector in $\mathbb{R}^d$ with entries $f_j(X_i)$ for some soft classifiers $f_j$. Thus $\|X_i\|_{\infty} \leq 1$ and therefore $\|X_i\| \leq \sqrt{d}$. Moreover, as $\varphi$ is $L$-Lipschitz, its sub-gradient $|g_{\varphi}(x)| \leq L$ at any $x \in \mathbb{R}$. It follows that

$$\|g_{\Theta_N}(\alpha)\| \leq L \sqrt{d}.$$ 

Therefore, $\Theta_N$ is $L'$-Lipschitz, with $L' = L \sqrt{d}$. Moreover, the diameter of $B_1$ is $R = 2$, so the error made by PGD after $T$ steps is

$$\frac{RL'}{\sqrt{T}} = \frac{2L \sqrt{d}}{\sqrt{T}}.$$ 

It is smaller than $1/N$ if $T \geq 4L^2dN^2$. In large dimensional problems where $d \gg N$, this algorithm can be improved substantially.
4.6 Mirror descent

Mirror descent algorithms are extensions of projected gradient descent algorithms that can be adapted to non-Euclidean geometries. The first idea of this extension is to bound inner products \( \mathbf{a}^T \mathbf{b} \) using other duality formulae than Cauchy-Schwarz inequality.

**Definition 48.** Let \( | \cdot | \) denote a norm on \( \mathbb{R}^d \). The dual norm \( | \cdot |_* \) is defined by

\[
\forall \mathbf{c} \in \mathbb{R}^d, \quad |\mathbf{c}|_* = \sup_{\mathbf{x} \in \mathbb{R}^d : |\mathbf{x}| \leq 1} \mathbf{x}^T \mathbf{c}.
\]

In particular, for any norm \( | \cdot | \) and any \( \mathbf{a}, \mathbf{b} \) in \( \mathbb{R}^d \), one has \( \mathbf{a}^T \mathbf{b} \leq |\mathbf{a}| |\mathbf{b}|_* \).

The second idea to extend PGD is to introduce potential functions.

**Definition 49 (Potential functions).** A function \( \Phi \) defined on a convex domain \( \mathcal{C} \) is called a potential if there exist a constant \( \nu > 0 \) and a norm \( | \cdot | \) such that \( \Phi \) is \( \nu \)-strongly convex with respect to the norm \( | \cdot | \), that is, if it is convex and

\[
\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}, \forall g \in \partial \Phi(\mathbf{x}), \quad \Phi(\mathbf{y}) \geq \Phi(\mathbf{x}) + g^T(\mathbf{y} - \mathbf{x}) + \frac{\nu}{2} |\mathbf{y} - \mathbf{x}|^2.
\]

This definition strengthens the notion of subgradient which only implies that

\[
\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}, \forall g \in \partial \Phi(\mathbf{x}), \quad \Phi(\mathbf{y}) \geq \Phi(\mathbf{x}) + g^T(\mathbf{y} - \mathbf{x}) .
\]

**Example 1** Let \( \mathcal{C} = \mathbb{R}^d \), \( \Phi(\mathbf{x}) = |\mathbf{x}|^2/2 \). Then \( \Phi \) is differentiable with gradient \( \nabla \Phi(\mathbf{x}) = \mathbf{x} \). For any \( \mathbf{y} \in \mathbb{R}^d \),

\[
\Phi(\mathbf{y}) = \frac{|\mathbf{y}|^2}{2} = \frac{|\mathbf{x}|^2}{2} + \mathbf{x}^T(\mathbf{y} - \mathbf{x}) + \frac{|\mathbf{y} - \mathbf{x}|^2}{2} = \Phi(\mathbf{x}) + \nabla \Phi(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2} |\mathbf{y} - \mathbf{x}|^2 .
\]

Therefore, \( \Phi \) is 1-strongly convex with respect to the Euclidean norm over \( \mathbb{R}^d \). Therefore, it is a potential and, as will be clear, mirror descent is an extension of PGD.
CHAPTER 4. CONVEX OPTIMIZATION

Example 2 Let $C = \Delta_d$ and let $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$. Then $\nabla \Phi(x) = (\log(x_i) + 1)_{1 \leq i \leq d}$ and, for any $y \in \mathbb{R}^d$,

$$
\Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x) = \sum_{i=1}^d y_i \log(y_i) - x_i \log(x_i) - (\log(x_i) - 1)(y_i - x_i)
$$

$$
= \sum_{i=1}^d y_i \log \left( \frac{y_i}{x_i} \right) + y_i - x_i \quad (4.5)
$$

By Pinsker’s inequality, see Lemma 30,

$$
\Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x) \geq \frac{1}{2} \left( \sum_{i=1}^d |x_i - y_i| \right)^2 = \frac{\|x - y\|_1^2}{2}.
$$

Therefore, $\Phi$ is 1-strongly convex with respect to $\| \cdot \|_1$ over $\Delta_d$.

A key step to analyse gradient descent and PGD is the formula

$$
a^T b = \frac{1}{2} (\|a\|^2 + \|b\|^2 - \|a - b\|^2) \quad (4.6)
$$

This formula can be extended to more general potential functions using Bregman divergences.

Definition 50 (Bregman divergence). Let $\Phi$ denote a convex function over a convex domain $C$. Then, for any $x, y$ in $C$ the Bregman divergence of $\Phi$ from $y$ to $x$ is defined by

$$
D_\Phi(y, x) = \Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x) \quad .
$$

$D_\Phi(y, x)$ is therefore the error made when approximating $\Phi(y)$ by the linear approximation $\Phi(x) + \nabla \Phi(x)^T (y - x)$. When $\Phi$ is a potential, its Bregman divergence acts as the square of the norm $\| \cdot \|$. The key to extend formula (4.5) is the following proposition.

Proposition 51. If $\Phi$ is convex on a convex domain $C$ and $x, y, z$ lie in $C$, then

$$
D_\Phi(y, x) + D_\Phi(x, y) = (\nabla \Phi(y) - \nabla \Phi(x))^T (y - x) \quad ,
$$

$$
(\nabla \Phi(x) - \nabla \Phi(y))^T (x - z) = D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y) \quad .
$$
4.6. MIRROR DESCENT

Proof. The proof is completely elementary. The first equality is immediate and, for the second one,

\[ D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y) = \Phi(x) - \Phi(y) - \nabla \Phi(y)^T (x - y) \]
\[ + \Phi(z) - \Phi(x) - \nabla \Phi(x)^T (z - x) \]
\[ - \Phi(z) + \Phi(y) + \nabla \Phi(y)^T (z - y) \]
\[ = \nabla \Phi(y)^T (z - y - x + y) - \nabla \Phi(x)^T (z - x) \]
\[ = (\nabla \Phi(y) - \nabla \Phi(x))^T (z - x) . \]

\[ \Box \]

As for projected gradient descent, we also need to project according to the potential \( \Phi \).

**Definition 52** (Bregman projection). Let \( \Phi \) denote a convex function defined on a closed convex domain \( C \) and let \( C \) denote a closed convex subset of \( C \). The Bregman projection of any \( x \in C \) on \( C \) with respect to the function \( \Phi \) is

\[ \pi^\Phi_C(x) \in \arg\min_{y \in C} D_\Phi(x, y) . \]

**Example 1** The Bregman divergence with respect to \( \Phi = \| \cdot \|^2/2 \) is given by (see Eq (4.4))

\[ D_\Phi(y, x) = \frac{\| x - y \|^2}{2} . \]

Therefore, Bregman projection with respect to \( \Phi \) is the usual projection onto convex sets shown in Theorem 4.6.

**Example 2** Let us consider \( C = (\mathbb{R}_+)^d \), \( C = \Delta_d \subset C \) and \( \Phi(x) = \sum_{i=1}^d x_i \log(x_i) \). In this example, the Bregman divergence is given, see (4.5), by

\[ D_\Phi(y, x) = \sum_{i=1}^d y_i \log \left( \frac{y_i}{x_i} \right) + y_i - x_i . \]

The Bregman projection \( x \) of \( y \in C \) is a minimizer of \( D_\Phi(y, \cdot) \) over \( \Delta_d \). Define the Lagrangian

\[ L_\lambda(x) = D_\Phi(y, x) + \lambda \left( \sum_{i=1}^d x_i - 1 \right) . \]

It holds

\[ \forall i \in \{1, \ldots, d\}, \quad \frac{\partial L_\lambda}{\partial x_i}(x) = -\frac{y_i}{x_i} - 1 + \lambda , \quad \frac{\partial L_\lambda}{\partial \lambda}(x) = \sum_{i=1}^d x_i - 1 . \]
Solving the equations \( \frac{\partial L}{\partial x_i}(x) = 0 \) and \( \frac{\partial L}{\partial x}(x) = 0 \) yields, for all \( i \), \( x_i = y_i / (\lambda - 1) \) and \( (\lambda - 1) = \sum_{i=1}^{d} y_i = \|y\|_1 \), so finally
\[
\pi^\Phi_{\Delta_d}(y) = x = \frac{y}{\|y\|_1}.
\]

As for the usual projections onto convex sets, Bregman projections have a nice characterization in terms of inner products.

**Proposition 53.** For any \( z \in C \) and any \( y \in C \),
\[
(\nabla \Phi(\pi^\Phi_C(y)) - \nabla \Phi(y))^T(\pi^\Phi_C(y) - z) \leq 0.
\]
In addition,
\[
D_\Phi(z, \pi^\Phi_C(y)) \leq D_\Phi(z, y).
\]

**Proof.** For any \( t \in [0, 1] \), let \( h(t) = D_\Phi(\pi^\Phi_C(y) + t(z - \pi^\Phi_C(y)), y) \). As \( \pi^\Phi_C(y) + t(z - \pi^\Phi_C(y)) \in C \) by convexity, \( h \) is minimal at \( t = 0 \), so
\[
0 \leq h'(0) = \nabla_x D_\Phi(x, y)|_{x=\pi^\Phi_C(y)}^T(z - \pi^\Phi_C(y)) .
\]
Now, by construction of \( D_\Phi \),
\[
\nabla_x D_\Phi(x, y)|_{x=\pi^\Phi_C(y)} = \nabla \Phi(\pi^\Phi_C(y)) - \nabla \Phi(y)
\]
Hence,
\[
(\nabla \Phi(\pi^\Phi_C(y)) - \nabla \Phi(y))^T(z - \pi^\Phi_C(y)) \geq 0.
\]
This proves the first item. For the second, it follows from Proposition 51 and the first item that
\[
D_\Phi(\pi^\Phi_C(y), y) + D_\Phi(z, \pi^\Phi_C(y)) - D_\Phi(z, y) \leq 0.
\]
As \( D_\Phi(\pi^\Phi_C(y), y) \geq 0 \), this implies that
\[
D_\Phi(z, \pi^\Phi_C(y)) \leq D_\Phi(z, y).
\]
This concludes the proof of the proposition. \(\Box\)

We are now in position to give the mirror descent algorithm.
input: $x_1 \in \text{argmin}_{x \in \mathcal{C}} \Phi(x)$: initial point, $(\eta_t)_t$: step sizes, $T$: stopping time

1. for $t = 1$ to $t = T$, do
2. find $g_t \in \partial \theta(x_t)$,
3. $\nabla \Phi(y_{t+1}) = \nabla \Phi(x_t) - \eta_t g_t$,
4. $x_{t+1} = \pi^\Phi_C(y_{t+1})$.
5. end
6. Return $\bar{x}_T = T^{-1} \sum_{i=1}^{T} x_t$ or $\bar{x}_T \in \text{argmin}_{x \in \{x_1, \ldots, x_T\}} \theta(x)$.

Algorithm 3: Mirror descent (MD).

Compared with gradient descent, mirror descent uses the gradients of the potential $\Phi$ to map elements of $\mathbb{R}^d$ endowed with the norm $\| \cdot \|$ (called primal) into its dual which is homeomorphic to $\mathbb{R}^d$ endowed with the dual norm $\| \cdot \|$.

Then, mirror descent uses a gradient descent step in the dual space and maps back the images into the primal space. The term mirror originates from this mapping into the dual space. One can analyse MD algorithm almost exactly as PGD.

**Theorem 54.** Assume that $\theta$ is convex on a (closed convex) domain $\mathcal{C}$ and that it is $L$-Lipschitz with respect to the norm $\| \cdot \|$, that is $|g_t| \leq L$ for any $g_t \in \partial \theta(x)$ and any $x \in \mathcal{C}$. Let $C$ denote a closed convex subset of $\mathcal{C}$. Assume that $\Phi$ is $\nu$-strongly convex on $\mathcal{C}$ with respect to $\| \cdot \|$. Let $x^* \in \text{argmin}_{x \in \mathcal{C}} \theta(x)$ and

$$R^2 = \sup_{x \in \mathcal{C}} \Phi(x) - \inf_{x \in \mathcal{C}} \Phi(x) .$$

Then, the outputs of MD algorithm with $\eta_t = (R/L) \sqrt{2/\nu T}$ satisfy

$$\theta(\bar{x}_T) - \theta(x^*) \leq RL \sqrt{\frac{2}{\nu T}}, \quad \theta(\bar{x}_T) - \theta(x^*) \leq RL \sqrt{\frac{2}{\nu T}} .$$

**Proof.** Let $x_o \in C \cap \mathcal{C}$, by definition of the MD update,

$$\theta(x_t) - \theta(x_o) \leq g_t^T (x_t - x_o) = \frac{1}{\eta} (\nabla \Phi(x_t) - \nabla \Phi(y_{t+1}))^T (x_t - x_o) .$$

By Proposition 51, it follows that

$$\theta(x_t) - \theta(x_o) \leq \frac{1}{\eta} \left( D_{\Phi}(x_t, y_{t+1}) + D_{\Phi}(x_o, x_t) - D_{\Phi}(x_o, y_{t+1}) \right) .$$

Now we invoke the following result.
**Proposition 55.** Using updates of MD, if $\Phi$ is $\nu$-strongly convex with respect to $| \cdot |$, then

$$D_\Phi(x_t, y_{t+1}) \leq \frac{\eta^2 |g_t|^2}{2\nu}. $$

**Proof of Proposition 55.** By Proposition 51,

$$D_\Phi(x_t, y_{t+1}) = -D_\Phi(y_{t+1}, x_t) + \langle \nabla \Phi(x_t) - \nabla \Phi(y_{t+1}), (x_t - y_{t+1}) \rangle. $$

By strong convexity, it holds therefore

$$D_\Phi(x_t, y_{t+1}) = -\frac{\nu}{2} |y_{t+1} - x_t|^2 + \langle \nabla \Phi(x_t) - \nabla \Phi(y_{t+1}), (x_t - y_{t+1}) \rangle. $$

Using the MD updates and the definition of the dual norm, one gets

$$D_\Phi(x_t, y_{t+1}) \leq \eta g_t^T (x_t - y_{t+1}) - \frac{\nu}{2} |y_{t+1} - x_t|^2 \leq \eta |g_t| |x_t - y_{t+1}| - \frac{\nu}{2} |y_{t+1} - x_t|^2. $$

The proof terminates by the inequality $\sup_{x \in \mathbb{R}} \{ax - bx^2/2\} = a^2/(2b)$, which holds for any $a \in \mathbb{R}$ and $b > 0$. \qed

By Proposition 53,

$$D_\Phi(x_t, y_{t+1}) \leq \frac{\eta^2 |g_t|^2}{2\nu}. $$

As $\theta$ is $L$-Lipschitz, this implies that

$$D_\Phi(x_t, y_{t+1}) \leq \frac{\eta^2 L^2}{2\nu}. $$

Moreover, by Proposition 53

$$D_\Phi(x_o, x_{t+1}) \leq D_\Phi(x_o, y_{t+1}). $$

Hence,

$$\theta(x_t) - \theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{D_\Phi(x_o, x_t) - D_\Phi(x_o, x_{t+1})}{\eta}. $$

Summing up

$$\frac{1}{T} \sum_{t=1}^T \theta(x_t) - \theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{D_\Phi(x_o, x_1)}{\eta T}. $$
As $x_1$ is a minimizer of $\Phi$ over $C$,
\[ D_\Phi(x_0, x_1) = \Phi(x_0) - \Phi(x_1) - \nabla \Phi(x_1)^T (x_0 - x_1) \leq \Phi(x_0) - \Phi(x_1) \leq R^2. \]
Thus,
\[ \frac{1}{T} \sum_{t=1}^{T} \theta(x_t) - \theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{R^2}{\eta T} \leq LR \sqrt{\frac{2}{\nu T}}. \]
The proof concludes as usual that
\[ \theta(x_T) - \theta(x_o) \leq LR \sqrt{\frac{2}{\nu T}}, \quad \theta(x_T) - \theta(x_o) \leq LR \sqrt{\frac{2}{\nu T}}. \]
Finally, one can let $x_o \to x^*$ to obtain the theorem.

To show the interest of MD, let us go back to the boosting example. Take $\Phi(x) = \sum_{i=1}^{d} x_i \log(x_i)$, so the MD update is equivalent to
\[ \forall i \in \{1, \ldots, d\}, \quad \log(y_{t+1,i}) = \log(x_{t,i}) - \eta g_{t,i}, \quad \text{i.e.} \quad y_{t+1,i} = x_{t,i} e^{-\eta g_{t,i}}, \]
and $x_{t+1} = y_{t+1}/\|y_{t+1}\|_1$. Recall that the function to minimize is
\[ \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y_i^{(s)} \alpha^T X_i), \]
where the vectors $X_i$ are defined thanks to a collection of preliminary soft classifiers $f_1, \ldots, f_d$ by
\[ X_i = \begin{bmatrix} f_1(X_i) \\ \vdots \\ f_d(X_i) \end{bmatrix}. \]
The sub-gradients of this objective function can be computed:
\[ g_{\Theta_N}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} g_{\varphi}(-Y_i^{(s)} \alpha^T X_i)(-Y_i^{(s)} X_i). \]
Now $|\cdot| = \|\cdot\|_1$ has dual norm $|\cdot|_* = \|\cdot\|_\infty$. As $\varphi$ is $L$-Lipschitz, its subgradients $|g_{\varphi}(x)| \leq L$ at any point $x$. As $Y_i^{(s)} \in [-1,1]$ and $\|X_i\|_\infty \leq 1$, it holds that
\[ \|g_{\Theta_N}(\alpha)\|_\infty \leq L. \]
As this holds for any $\alpha \in \Delta_d$, this shows that $\Theta_N$ is $L$-Lipschitz with respect to $\|\cdot\|_1$. Moreover, for any $x \in \Delta_d$, $\Phi(x) \leq 0$ and
\[ \Phi(x) + \log d = \sum_{i=1}^{d} x_i \log(x_i/(1/d)) \geq 0, \]
so

\[ \Phi(x) \geq -\log d. \]

Hence, \( R^2 = \log d \). Theorem 57 shows then that MDA has, after \( T \) steps, an error upper bounded by

\[ L \sqrt{\frac{2 \log d}{T}}. \]

This error is upper bounded by \( 1/N \) as soon as \( T \geq 2L^2N^2 \log d \). Recall that PGD on this example achieved this error after \( 4L^2dN^2 \) steps. When the dimension \( d \) is large, the improvement brought by MD over PGD is substantial.

### 4.7 Stochastic optimization

In this section, consider functions \( \theta : \mathcal{C} \times \mathcal{Z} \to \mathbb{R} \) such that \( \mathcal{C} \) is a convex subset of \( \mathbb{R}^d \) and, given a distribution \( P \) on \( \mathcal{Z} \), for \( P \)-almost \( z \in \mathcal{Z} \), the function \( \theta(\cdot, z) \) is convex. Assume also that, for any \( x \in \mathcal{C} \), \( E[|\theta(x, Z)|] < \infty \), so the convexity assumption implies that the function \( \Theta = E[\theta(\cdot, Z)] \) is convex. The goal of stochastic optimization is to approximate

\[ \min_{x \in \mathcal{C}} \Theta(x) \]

using an i.i.d. sample \( Z_1, \ldots, Z_T \) with distribution \( P \). Remark that the difference with convex optimization is that the objective function \( \Theta \) is unknown and we only have access to noisy versions of it (and its gradients), therefore, one cannot define the estimators

\[ x_T \in \arg\min_{x \in \{x_1, \ldots, x_T\}} \Theta(x) \]

that we used in the convex optimization algorithms. On the other hand, by focusing on the actual expected value of the loss, the performance of the final estimator won’t rely on those of the ERM.

#### 4.7.1 Stochastic gradient descent

Denote by \( \partial \theta(x, Z) \) the sub-differential of \( \theta(\cdot, Z) \) at point \( x \in \mathbb{R} \). The stochastic gradient descent algorithm is described below.
4.7. STOCHASTIC OPTIMIZATION

<table>
<thead>
<tr>
<th>input</th>
<th>( x_1 \in C ): initial point, ((\eta_t)): step sizes, ( T ): stopping time</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Z_1, \ldots, Z_T ) i.i.d. random variables</td>
<td></td>
</tr>
<tr>
<td>1 for ( t = 1 ) to ( t = T ), do</td>
<td></td>
</tr>
<tr>
<td>2 find ( \tilde{g}_t \in \partial \theta(x_t, Z_t) ),</td>
<td></td>
</tr>
<tr>
<td>3 ( y_{t+1} = x_t - \eta_t \tilde{g}_t ),</td>
<td></td>
</tr>
<tr>
<td>4 ( x_{t+1} = \pi_C(y_{t+1}) ).</td>
<td></td>
</tr>
<tr>
<td>end</td>
<td></td>
</tr>
<tr>
<td>6 Return ( x_T = T^{-1} \sum_{i=1}^{T} x_i ).</td>
<td></td>
</tr>
</tbody>
</table>

Algorithm 4: Stochastic gradient descent (SGD).

An important difference between gradient descent and stochastic gradient descent is that all gradients \( \partial \theta(x_t, Z_t) \) for \( t = 1, \ldots, T \) must be available for gradient descent while only one of these gradients is used for stochastic gradient descent. This is particularly interesting for huge datasets as these gradients have to be computed in different servers and may be stored in different places. In this case, requesting all gradients can be both demanding on memory resources and produce non reliable estimators if a server is very slow or crash down. In these situations, stochastic gradient is much more efficient.

**Theorem 56.** Let \( C \) denote a closed, convex subset of \( \mathbb{R}^d \) with diameter smaller than \( R \). Let \( x^* \in \arg\min_{x \in C} \Theta(x) \). Assume that, for any \( x \in C \), any random subgradient \( \tilde{g}_t \in \partial \theta(x, Z) \) satisfies \( \mathbb{E}[\|\tilde{g}_t\|^2] \leq L^2 \). If, for any \( t = 1, \ldots, T \), \( \eta_t = \eta = R/(L\sqrt{T}) \), it holds

\[
\mathbb{E}[\Theta(x_T)] - \Theta(x^*) \leq \frac{RL}{\sqrt{T}}.
\]

**Proof.** Let \( F_t \) denote the \( \sigma \)-algebra generated by \( Z_1, \ldots, Z_{t-1} \) so \( x_t \) is \( F_t \)-measurable. Let \( \tilde{g}_t \in \partial \theta(x_t, Z_t) \) so, by independence of \( Z_t \) and \( x_t \),

\[
g_t = \mathbb{E}[	ilde{g}_t | F_t] \in \partial \Theta(x_t).
\]

Then,

\[
\Theta(x_t) - \Theta(x^*) \leq g_t^T (x_t - x^*)
\]

\[
= \mathbb{E}[g_t^T (x_t - x^*) | F_t]
\]

\[
= \frac{1}{\eta} \mathbb{E}[(x_t - y_{t+1})^T (x_t - x^*) | F_t]
\]

\[
= \frac{1}{2\eta} \mathbb{E}[(\|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2 | F_t]
\]

\[
\leq \frac{1}{2\eta} \mathbb{E}[(\eta^2 \|\tilde{g}_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 | F_t] .
\]
Taking expectations,
\[
\mathbb{E}[\Theta(x_t)] - \Theta(x^*) \leq \frac{\eta L^2}{2} + \frac{\mathbb{E}[\|x_t - x^*\|^2] - \mathbb{E}[\|x_{t+1} - x^*\|^2]}{2}.
\]

We conclude as before by summing up, bounding \(\|x_1 - x^*\|^2 \leq R^2\), and using the convexity of \(\Theta\).

4.7.2 Stochastic mirror descent

Stochastic gradient descent suffers the same issues as gradient descent when \(C\) is a convex adapted to \(\ell_1\)-geometry rather than Euclidean geometry. Fortunately, a stochastic version of general mirror descent algorithms can easily be designed to bypass this issue.

**Algorithm 5:** Stochastic mirror descent (SMD).

Mixing the proofs of SGD and MD yields the following result.

**Theorem 57.** Assume that \(\theta(\cdot, Z)\) is \(P\)-almost surely convex on a domain \(C\) and that, for any \(x \in C\), any random subgradient \(\tilde{g}_t \in \partial \theta(x, Z)\) satisfies \(\mathbb{E}[\|\tilde{g}_t\|^2] \leq L^2\). Let \(C\) be a closed convex subset \(C\). Assume that \(\Phi\) is \(\nu\)-strongly convex on \(C\) with respect to \(\|\cdot\|\). Let \(x^* \in \arg\min_{x \in C} \theta(x)\).

\[
R^2 = \sup_{x \in C} \Phi(x) - \inf_{x \in C} \Phi(x).
\]

Then, the outputs of SMD with \(\eta_t = (R/L)\sqrt{2\nu/T}\) satisfy

\[
\Theta(x_T) - \Theta(x^*) \leq RL\sqrt{\frac{2}{\nu T}}.
\]

**Proof.** The proof starts as for SGD and terminates as for MD. Let \(F_t\) denote the \(\sigma\)-algebra generated by \(Z_1, \ldots, Z_{t-1}\) so \(x_t\) is \(F_t\)-measurable. Let \(\tilde{g}_t \in \partial \theta(x_t, Z_t)\) so, by independence of \(Z_t\) and \(x_t\),

\[
g_t = \mathbb{E}[\tilde{g}_t|F_t] \in \partial \Theta(x_t).
\]
4.7. **STOCHASTIC OPTIMIZATION**

Then, for any $x_o \in C$,

$$
\Theta(x_t) - \Theta(x_o) \leq g_t^T(x_t - x_o) = \mathbb{E}[g_t^T(x_t - x_o) | \mathcal{F}_t]
$$

$$
= \frac{1}{\eta} \mathbb{E}[(\nabla \Phi(x_t) - \nabla \Phi(y_{t+1}))^T(x_t - x_o) | \mathcal{F}_t]
$$

$$
\leq \frac{1}{\eta} \mathbb{E}[D_\Phi(x_t, y_{t+1}) + D_\Phi(x_o, x_t) - D_\Phi(x_o, y_{t+1}) | \mathcal{F}_t]
$$

$$
\leq \frac{1}{\eta} \mathbb{E} \left[ \frac{\eta^2 |g_t|^2}{2\nu} + D_\Phi(x_o, x_t) - D_\Phi(x_o, x_{t+1}) | \mathcal{F}_t \right].
$$

Taking expectations and summing yields

$$
\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^{T} \Theta(x_t) \right] - \Theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{D_\Phi(x_o, x_1)}{\eta T}.
$$

As $x_1$ is a minimizer of $\Phi$, $D_\Phi(x_o, x_1) \leq \Phi(x_o) - \Phi(x_1) \leq R^2$, so, by convexity of $\Theta$,

$$
\mathbb{E} \left[ \Theta(x_t) \right] - \Theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{R^2}{\eta T} = \frac{2LR}{\sqrt{2\nu T}}.
$$

Conclude by letting $x_o \to x^*$. \qed
Chapter 5

Stochastic bandit

The stochastic multi-armed bandit is a model for decision making. There are $K$ arms representing the different actions available at each step. Iteratively, the algorithm chooses an arm $\pi_t \in \{1, \ldots, K\}$ and receives a reward $X_{t, \pi_t}$. For any fixed $k$, $X_{t,k}$ is a random variable centered in $\mu_k$, let also $\mu_* = \max_{k \in \{1, \ldots, K\}} \mu_k$. A policy $\pi$ is a sequence $(\pi_t)_{t \geq 1}$ such that, for any $t \geq 1$, $\pi_t \in \{1, \ldots, K\}$ indicates which arm is played at time $t$. Formally, let $(X_{t,k})_{t \geq 1, k \in \{1, \ldots, K\}}$ and $(U_t)_{t \geq 1}$ denote jointly independent random variables, such that $\mathbb{E}[X_{t,k}] = \mu_k$ for any $t \geq 1$. For some random variable $\pi_t \in \{1, \ldots, K\}$ which is measurable with respect to the sigma-algebra generated by $(U_1, Y_1, \ldots, U_{t-1}, Y_{t-1}, U_t)$, we define $N_k(t) = \sum_{s=1}^t 1(\pi_s = k)$ and observe $Y_t = X_{N_{\pi_t(t), \pi_t}}$. A strategy is a set of functions $\psi = (\psi_t)_{t \geq 1}$ such that the arm $\pi_t$ is pulled at time $t$ according to $\pi_t = \psi_t(U_1, Y_1, \ldots, U_{t-1}, Y_{t-1}, U_t)$.

**Definition 58** (Regret). After $n$ iterations of the algorithm, the regret is defined by

$$R_n = n \mu_* - \mathbb{E} \left[ \sum_{t=1}^n Y_t \right] ,$$

The regret can be equivalently evaluated thanks to the following result.

**Proposition 59.** For any $k \in \{1, \ldots, K\}$, let $\Delta_k = \mu_* - \mu_k$ and for any $n \geq 1$, let also $N_k(n) = \sum_{t=1}^n 1(\pi_t = k)$. Then the regret satisfies

$$R_n = \sum_{k=1}^K \Delta_k \mathbb{E}[N_k(n)] .$$

**Proof.** Decomposing according to the sequence of pulled arms, we get

$$\sum_{t=1}^n Y_t = \sum_{k=1}^K \sum_{s=1}^{N_k(n)} X_{s,k} .$$
Thus, the proof is complete if
\[ E \left[ \sum_{s=1}^{N_k(n)} X_{s,k} \right] = E[N_k(n)] \mu_k . \tag{5.1} \]

We will prove this result using Wald’s formula.

**Lemma 60** (Wald’s formula). Let \( \mathcal{F}_t \) denote a filtration, \( N, X_1, \ldots, X_n \) denote random variables such that

(i) \( N \) takes values in \( \{1, \ldots, n\} \) and \( \{N \geq t\} \in \mathcal{F}_{t-1} \).

(ii) \( X_t \) is independent of \( \mathcal{F}_{t-1} \) and \( E[X_t] = \mu \).

Then
\[ E \left[ \sum_{t=1}^{N} X_t \right] = E[N] \mu . \]

**Proof of Wald’s formula.** It is assumed that \( X_t \) and \( 1_{\{N \geq t\}} \) are independent, thus,
\begin{align*}
E \left[ \sum_{t=1}^{N} X_t \right] &= E \left[ \sum_{t=1}^{n} X_t 1_{\{t \leq N\}} \right] = \sum_{t=1}^{n} E[X_t 1_{\{t \leq N\}}] \\
&= \mu \sum_{t=1}^{T} P(N \geq t) = \mu E[N] .
\end{align*}

\( \square \)

To apply Wald’s formulae, let \( \mathcal{F}_t \) denote the sigma-algebra generated by \( X_{1,k}, \ldots, X_{t,k} \), all \( X_{s,k'} \), avec \( k' \neq k \) and all \( U_s \) so \( X_{t,k} \) is independent of \( \mathcal{F}_{t-1} \) and \( \{N_k(t) \geq t\} = \Omega \setminus \bigcup_{s=1}^{t-1} \{N_k(t) = s\} \) belongs to \( \mathcal{F}_{t-1} \). Therefore, (5.1) holds thanks to Wald’s formula and the proposition is proved. \( \square \)

### 5.1 The full information case

In this section, we consider the following ideal situation where, at each time \( t \), we observe, after choosing an arm \( ^\wedge k \), the vector of all possible rewards \[
\begin{bmatrix}
X_{t,1} \\
\vdots \\
X_{t,K}
\end{bmatrix}.
\]
In this situation, it seems natural to pool an arm at random at
5.1. THE FULL INFORMATION CASE

time $t = 1$ and, at each time $t > 1$, the arm with the best average reward up to this point, that is

$$\pi_t \in \arg\max_{k \in \{1, \ldots, K\}} \sum_{s=1}^{t-1} X_{s,k}.$$  

**Lemma 61.** Assume now that all random variables $X_{s,k}$ are subGaussian, with variance proxy $\sigma^2$. The policy $\pi_t$ defined above has regret

$$R_n \leq \sum_{k: \Delta_k > 0} \Delta_k + 4t^2 \Delta_k.$$

**Proof.** Denote by $k^*$ an arm such that $\mu_{k^*} \geq \mu_k$, for any $k$ and let $k$ be such that $\Delta_k = \mu_{k^*} - \mu_k > 0$, we have

$$N_k(n) = 1_{\pi_1 = k} + \sum_{t=2}^{n} 1_{\{1/t \sum_{s=1}^{t-1} X_{s,k} \geq X_{s,k^*} \geq 0\}}$$

$$= 1_{\pi_1 = k} + \sum_{t=2}^{n} 1_{\{1/t \sum_{s=1}^{t-1} X_{s,k} - X_{s,k^*} > (\mu_k - \mu_{k^*}) > \Delta_k\}}$$

As $X_{s,k} - X_{s,k^*} - (\mu_k - \mu_{k^*})$ is centered and sub-Gaussian with variance proxy $2\sigma^2$,

$$E[1_{\{1/t \sum_{s=1}^{t-1} X_{s,k} \geq X_{s,k^*} \geq 0\} \leq e^{-(t-1)\Delta_k^2/(4\sigma^2)},}$$

Hence,

$$R_n = \sum_{k=1}^{K} \Delta_k E[N_k(n)] \leq \sum_{k: \Delta_k > 0} \Delta_k \left(\frac{1}{K} + \sum_{t=1}^{n-1} e^{-t\Delta_k^2/(4\sigma^2)}\right)$$

$$\leq \sum_{k: \Delta_k > 0} \Delta_k \left(\frac{1}{K} + \sum_{t=1}^{\infty} e^{-t\Delta_k^2/(4\sigma^2)}\right).$$

By a comparison between series and integral, it follows that

$$R_n \leq \sum_{k: \Delta_k > 0} \frac{\Delta_k}{K} + \frac{4t^2}{\Delta_k}.$$  
\[\Box\]
5.2 Upper confidence bound (UCB) policy

Let \( \delta(t) \) denote a non increasing sequence of real numbers. For any \( k \), let

\[
\Omega_k = \left\{ \sup_{t>0} \max_{k: \Delta_k>0} \mu_k \leq \frac{1}{t} \sum_{s=1}^{t} X_{s,k} + \delta(t) \right\} \cap \left\{ \inf_{t>0} \mu_{k^*} \geq \frac{1}{t} \sum_{s=1}^{t} X_{s,k^*} - \delta(t) \right\}.
\]

```
1 for t = 1 to K, do
  1 \( \pi_t = t \),
end
4 for t = K + 1 to n, do
  5 \( N_k(t) = \sum_{s=1}^{t-1} 1_{\{\pi_s = k\}} \)
  5 \( \hat{\mu}_{t,k} = \frac{1}{N_k(t)} \sum_{s=1}^{N_k(t)} X_{s,k} \)
  5 \( \pi_t \in \arg\max_{k \in \{1, \ldots, K\}} \left\{ \hat{\mu}_{t,k} + \delta(N_k(t)) \right\} \).
end
```

Algorithm 6: Upper Confidence Bound (UCB).

**Theorem 62.** Let \( \delta^{-1}(x) = \min\{t \geq 1 : \delta(t) < x\} \). On the event \( \Omega_k \), the UCB policy satisfies

\[
N_k(n) \leq \delta^{-1}(\Delta_k/2).
\]

Therefore, the regret of UCB policy satisfies

\[
R_n \leq \sum_{k=1}^{K} \Delta_k(\delta^{-1}(\Delta_k/2) + n(1 - \mathbb{P}(\Omega_k))).
\]

**Proof.** Assume that there exists \( t \) such that \( \delta(N_k(t)) < \Delta_k/2 \). Then, on \( \Omega_k \), for any \( t' \geq t \),

\[
\hat{\mu}_{t',k} + \delta(N_k(t')) \leq \mu_k + 2\delta(N_k(t')) < \mu_{k^*} \leq \hat{\mu}_{t',k^*} + \delta(N_{k^*}(t')).
\]

Hence, \( \pi_t \neq k \) and the first part of the theorem is proved. The second part is then a consequence of Proposition 59. \( \square \)
By Corollary 84, if $X_{s;k}$ are sub-Gaussian with variance proxy $v^2$ and $\delta(t) = 2v\sqrt{\log(n) + 2\log[\log(4t)]}/t$, $P(\Omega^*) \geq 1 - 1/n$. for this value of $\delta$, $\delta(t) < x$ if

$$t > \frac{12v^2 \log(n)}{x^2},$$

hence,

$$\delta^{-1}(\Delta_k/2) \leq \frac{48v^2 \log(n)}{\Delta_k^2}.$$ \(\text{P}(\Omega^*) \geq 1 - 1/n.\)

Plugging this bound into Theorem 62 shows that UCB satisfies the following result.

**Theorem 63.** Under the notations of Theorem 62, if the rewards are sub-Gaussian with variance proxy $v^2$, the regret of UCB is bounded from above by

$$R_n \leq \sum_{k=1}^{K} \Delta_k + \frac{48v^2 \log(n)}{\Delta_k}.$$ \(\text{P}(\Omega^*) \geq 1 - 1/n.\)

This shows that UCB has regret that is, up the logarithmic term $\log(n)$, comparable with the loss of the full information case given in Lemma 61.

### 5.3 Bounded regret

Consider the case where $\mu_{k^*} = \Delta/2$ and $\mu_k \leq -\Delta/2$ for all $k \in \{1, \ldots, K\} \setminus \{k^*\}$ and the algorithm knows that $\mu_{k^*} > 0$ while all other $\mu_k < 0$. In this case, one can apply the following strategy:

```plaintext
1 for t = 1 to K, do
  2 \hspace{1em} \pi_t = t,
3 end
4 for t = K + 1 to n, do
  5 \hspace{1em} if max_k \hat{\mu}_{t,k} > 0 then
  6 \hspace{2em} \pi_t \in \text{argmax}_k \hat{\mu}_{t,k}
  7 \hspace{1em} end
  8 else
  9 \hspace{2em} \pi_t = 1, \pi_{t+1} = 2
10 end
11 end
```

**Algorithm 7:** Bounded regret policy (BRP).
Theorem 64. If all rewards $X_{t,k} \in \text{SubGau}(v^2)$, the BRP algorithm has regret

$$R_n \leq \sum_{k: \Delta_k > 0} \Delta_k \left( 1 + \frac{16v^2}{\Delta_k^2} \right) \leq \sum_{k: \Delta_k > 0} \Delta_k + \frac{16v^2}{\Delta_k}$$

where $r_\infty = \max\{\Delta_k/\Delta, k = 1, \ldots, K\}$.

Proof. Let $k \neq k^*$.

$$P(\pi_t = k) = P(\pi_t = k, \hat{\mu}_{t,k} > 0) + P(\pi_t = k, \hat{\mu}_{t,k} \leq 0).$$

On one hand,

$$\sum_{t=K+1}^{n} P(\pi_t = k, \hat{\mu}_{t,k} > 0) \leq \sum_{t=K+1}^{+\infty} P(\pi_t = k, \hat{\mu}_{t,k} - \mu_k > \Delta/2) \leq \sum_{s=1}^{+\infty} e^{-s\Delta^2/8v^2} \leq \frac{8v^2}{\Delta^2}.$$

On the other hand,

$$\sum_{t=3}^{n} P(\pi_t = k, \hat{\mu}_{t,k} \leq 0) \leq \sum_{t=3}^{+\infty} P(\pi_t = k, \hat{\mu}_{t,k} - \mu_{k^*} < -\Delta/2) \leq \sum_{s=1}^{+\infty} e^{-s\Delta^2/8v^2} \leq \frac{8v^2}{\Delta^2}.$$

By Proposition 59, it follows that

$$R_n \leq \sum_{k: \Delta_k > 0} \Delta_k \left( 1 + \frac{16v^2}{\Delta_k^2} \right).$$

\[\square\]

5.4 An exercise

We consider in this section the case where the set of arms is $[0, 1]$, so it is in particular infinite. As before, let $(X_{t,x})_{t \geq 1, x \in [0,1]}$ and $(U_t)_{t \geq 1}$ denote jointly independent random variables, such that $E[X_{t,x}] = \mu_x$ for any $t \geq 1$ and
\[ \mathbb{E}[e^{s(X_1-x-\mu_x)}] \leq e^{s^2/2}. \]

Moreover, we assume that \( x \mapsto \mu_x \) is regular in the sense that
\[ \forall x, y \in [0, 1], \quad |\mu_x - \mu_y| \leq C|x - y|^\beta. \]

Let also \( \mu^* = \max_{x \in [0, 1]} \mu_x \). As \([0, 1]\) is infinite, one cannot play all arms, but, as \( \mu_x \) is regular, we decompose \([0, 1]\) into \( K \) intervals \( J_1, \ldots, J_K \) of length \( 1/K \). Then, we build a \( K \)-armed bandit problem by playing the arm \( k \in \{1, \ldots, K\} \) by playing arm \( x \) uniformly at random in \( J_k \). The average reward of this arm is then
\[ \tilde{\mu}_k = K \int_{J_k} \mu_x \, dx. \]

Let then \( \tilde{\mu}_{k^*} = \max\{\tilde{\mu}_k, k = 1, \ldots, K\} \). The regret of the UCB algorithm on this \( K \)-armed bandit is bounded from above thanks to Theorem 63 by
\[ \tilde{R}_n = \sum_{k: \tilde{\Delta}_k > 0} \tilde{\Delta}_k + 48 \log n \frac{\Delta_k}{\Delta_k}, \]
where \( \tilde{\Delta}_k = \tilde{\mu}_{k^*} - \tilde{\mu}_k \).

1. Prove that \( \mu^* - \tilde{\mu}_{k^*} \leq LK^{-\beta} \).

2. Prove that, for any \( \eta > 0 \),
\[ \tilde{R}_n \leq \eta n + \sum_{k: \Delta_k \geq \eta} \Delta_k \mathbb{E}[N_k(n)]. \]

3. Prove that
\[ \sum_{k: \Delta_k \geq \eta} \Delta_k \mathbb{E}[N_k(n)] \leq KL + \frac{48K \log n}{\eta}. \]

4. Prove that there exists a choice of \( K \) such that the UCB algorithm has regret
\[ R_n \leq O\left(n^{(\beta+1)/(2\beta+1)}(\log n)^{\beta/(2\beta+1)}\right). \]
Chapter 6

Concentration inequalities

A concentration inequality for a random variable $X$ is an upper bound on the probabilities $\Pr(X - \mathbb{E}[X] > t)$, for any $t > 0$.

6.1 Chernoff’s bound

A basic tool to prove concentration inequalities, called Chernoff’s bound, is to apply exponential Markov’s inequality and bound from above the Laplace transform.

Theorem 65 (Chernoff’s bound). Let $X$ denote a random variable with finite Laplace transform $\mathbb{E}[e^{sX}] < \infty$, for any $s \in (0, b)$. Then,

$$\forall t > 0, \quad \Pr(X - \mathbb{E}[X] > t) \leq e^{-\psi^*(t)},$$

where $\psi^*(t) = \sup_{s \in (0, b)} \{st - \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}]\}$.

Remark 66. This result implies the following corollary that will be used regularly. If $f(s) \geq \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}]$ and $f^*(t) = \sup_{s \in (0, b)} \{st - f(s)\}$, then

$$\forall t > 0, \quad \Pr(X - \mathbb{E}[X] > t) \leq e^{-f^*(t)}.$$

Proof. Let $t > 0$ and $s \in (0, b)$. As the function $x \mapsto e^{sx}$ is increasing, we have

$$\Pr(X - \mathbb{E}[X] > t) = \Pr(e^{s(X-\mathbb{E}[X])} > e^{st}).$$

By Markov’s inequality, we have

$$\Pr(X - \mathbb{E}[X] > t) \leq \frac{\mathbb{E}[e^{s(X-\mathbb{E}[X])}]}{e^{st}} = e^{-st - \log \mathbb{E}[e^{s(X-\mathbb{E}[X])}]}.$$

As the result holds for any $s \in (0, b)$, optimizing in $s$ yields the result. \qed
Let’s compute the Laplace transform \( \psi(s) = \log \mathbb{E}[e^{sX}] \) and the Fenchel-Legendre transform \( \psi^*(t) \) for several well known distributions.

**Example 5** (Gaussian distributions). Let \( X \sim N(0, \sigma^2) \). If \( \sigma^2 = 1 \),
\[
\mathbb{E}[e^{sX}] = \int_{\mathbb{R}} e^{sx - x^2/2} \frac{dx}{\sqrt{2\pi}} = e^{s^2} \int_{\mathbb{R}} e^{-(x-s)^2/2} \frac{dx}{\sqrt{2\pi}} = e^{s^2}.
\]
If \( \sigma^2 > 0 \), then \( X/\sigma \sim N(0, 1) \), so
\[
\forall s > 0, \quad \mathbb{E}[e^{sX}] = \mathbb{E}[e^{(s/\sigma)(X/\sigma)}] = e^{(s/\sigma)^2}.
\]
It follows that
\[
\psi^*(t) = \sup_{s \in \mathbb{R}^*_+} \left\{ st - s^2 \sigma^2 \right\} = \sup_{s \in \mathbb{R}^*_+} \left\{ -\frac{t^2}{2\sigma^2} - \frac{(s - t/\sigma)^2}{2} \right\} = -\frac{t^2}{2\sigma^2}.
\]
In particular, if \( X_1, \ldots, X_n \) are i.i.d \( N(\mu, \sigma^2) \), \( n^{-1} \sum_{i=1}^n X_i - \mu \sim N(0, \sigma^2/n) \), so
\[
\forall t > 0, \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu > t\right) \leq e^{-\frac{nt^2}{2\sigma^2}}.
\]
Therefore,
\[
\forall t > 0, \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X] > \sqrt{\frac{2\sigma^2 t}{n}}\right) \leq e^{-t}.
\]

**Example 6** (Poisson’s distributions). Let \( X \sim \mathcal{P}(\theta) \). For any \( s > 0 \),
\[
\mathbb{E}[e^{sX}] = e^{-\theta} \sum_{k=0}^{\infty} \frac{e^{sk} \theta^k}{k!} = e^{\theta(e^s - 1)}.
\]
Recall that \( \mathbb{E}[X] = \theta = \text{Var}(X) \). Direct computations show that
\[
\mathbb{E}[e^{s(X - \mathbb{E}[X])}] = e^{\theta(e^s - 1)}.
\]
Let \( f(s) = st - \theta(e^s - 1 - s) \). As \( f'(s) = t - \theta(e^s - 1) \), \( f \) reaches a maximum in \( s = \log(1 + t/\theta) \) and this maximum is equal to
\[
(\theta + t) \log(1 + t/\theta) - t = \theta h(t/\theta),
\]
where \( h(u) = (1 + u) \log(1 + u) - u \). If \( X_1, \ldots, X_n \) are i.i.d. \( \mathcal{P}(\theta) \), we have \( \sum_{i=1}^n X_i \sim \mathcal{P}(n\theta) \), so
\[
\forall t > 0, \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \theta > t\right) \leq e^{-n\theta h(t/\theta)}.
\]
6.1. CHERNOFF’S BOUND

One can check as an exercise that (see also Section 6.2.3)
\[ h(t) \geq \frac{t^2}{2(1 + t/3)} . \]

It follows that
\[ \forall t > 0, \quad P \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \theta > t \right) \leq e^{-\frac{nt^2}{2(1 + t/3)}} . \]

This also implies that
\[ \forall t > 0, \quad P \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] > \sqrt{\frac{2 \text{Var}(X) t}{n} + \frac{2t}{3n}} \right) \leq e^{-t} . \]

**Example 7** (Gamma distributions). Let \( X \sim \Gamma(a, b) \), the density of this distribution with respect to Lebesgue’s measure is
\[ f(x, (a, b)) = \frac{x^{a-1}e^{-x/b}}{b^a \Gamma(a)} . \]

Recall that \( \mathbb{E}[X] = ab \), \( \text{Var}(X) = ab^2 \). We have
\[ \forall s \in (0, 1/b), \quad \mathbb{E}[e^{sX}] = \int \frac{x^{a-1}e^{-x(1/b-s)}}{b^a \Gamma(a)} \, dx = \frac{1}{(1 - bs)^a} . \]

Hence,
\[ \forall s \in (0, 1/b), \quad \mathbb{E}[e^{s(X-\mathbb{E}[X])}] = \frac{e^{-sab}}{(1 - bs)^a} . \]

Let \( f(s) = st + sab + a \log(1 - bs) \), so \( f'(s) = t + ab - \frac{ab}{1-bs} \), thus \( f \) reaches its maximum in \( s = t/(b(ab + t)) \) and this maximum is \( t/b - a \log(1 + t/(ab)) \).

If \( X_1, \ldots, X_n \) are i.i.d. with common distributions \( \Gamma(a, b) \), we have \( \sum_{i=1}^{n} X_i \sim \Gamma(na, b) \), hence
\[ \forall t > 0, \quad P \left( \frac{1}{n} \sum_{i=1}^{n} X_i - ab > abt \right) \leq e^{-na(t-\log(1+t))} . \]

Moreover, for any \( t \in (0, 1/2) \),
\[ t - \log(1 + t) = \sum_{i=2}^{+\infty} \frac{(-1)^k t^k}{k} \leq \frac{t^2}{2(1 - t)} \leq t^2 \, , \]
and, for any \( t > 1/2 \), as \( \log(1 + t) \leq \log(1) + \frac{2t}{3} \), \( t - \log(1 + t) \geq t/3 \), thus,

\[
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} X_i - ab > abt \right) \leq \begin{cases} e^{-\frac{nt^2}{2}} & \text{if } t < 1/2 \\ e^{-\frac{nt^{3/2}}{3}} & \text{if } t \geq 1/2 \end{cases}.
\]

Hence,

\[
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} X_i - ab > abt \right) \leq e^{-\frac{nt^{3/2}}{2+\frac{1}{3}t}}.
\]

This implies in particular that

\[
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] > \sqrt{\frac{2 \text{Var}(X)t}{n} + \frac{1.3bt}{n}} \right) \leq e^{-t}.
\]

**Example 8** (Binomial distributions). Let \( X \sim B(r, \theta) \), we have

\[
\forall s > 0, \quad \mathbb{E}[e^{sX}] = \sum_{k=0}^{r} \binom{r}{k} (e^{s\theta})^{k}(1-\theta)^{r-k} = (1-\theta+\theta e^s)^r.
\]

Hence,

\[
\forall s > 0, \quad \mathbb{E}[e^{s(X-\mathbb{E}[X])}] = (e^{-s\theta}(1-\theta+\theta e^s))^r.
\]

Let \( f(s) = s(t+r\theta) - r \log(1-\theta+\theta e^s) \). We have

\[
f'(s) = t + r\theta - \frac{\theta re^s}{1-\theta+\theta e^s}.
\]

Hence, \( f \) reaches a maximum in \( s = \log[(r\theta + t)(1-\theta)/(\theta(r - r\theta - t)) \) and this maximum is

\[
(r\theta + t) \log \left( \frac{r\theta + t}{r\theta} \right) + (r(1-\theta) - t) \log \left[ \frac{r(1-\theta) - t}{r(1-\theta)} \right].
\]

Introducing, for any \( p, q \) in \((0,1)\), \( \text{kl}(p,q) = p \log(p/q) + (1-p) \log((1-p)/(1-q)) \), we deduce that, for any \( t \in (0, r(1-\theta)) \), \( \psi^*(t) = r \text{kl}(\theta + t/r, \theta) \).

If \( X_1, \ldots, X_n \) are i.i.d. with common distribution \( B(r, \theta) \), on a \( \sum_{i=1}^{n} X_i \sim B(nr, \theta) \),

\[
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} X_i - r\theta > t \right) \leq e^{-nr \text{kl}(\theta+t/r, \theta)}.
\]

One can check that

\[
\text{kl}(\theta + t/r, \theta) \geq \frac{(t/r)^2}{2}, \quad \text{kl}(\theta + t/r, \theta) \geq \frac{(t/r)^2}{2\theta(1-\theta) + t/r}.
\]
so, for any \( t > 0 \),

\[
\Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i - r\theta > rt\right) \leq e^{-\frac{rt^2}{2}}, \quad \Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i - r\theta > rt\right) \leq e^{-\frac{nrt^2}{2(1+b)rt}}.
\]

This implies, for any \( t > 0 \),

\[
\Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] > \sqrt{\frac{2rt}{n}}\right) \leq e^{-t}, \quad \Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i - r\theta > \sqrt{\frac{2\text{Var}(X)t}{n} + \frac{t}{n}}\right) \leq e^{-t}.
\]

In all examples, we proved a deviation bound of the form

\[
\forall t > 0, \quad \Pr\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] > \sqrt{\frac{2\text{Var}(X)t}{n} + C\frac{t}{n}}\right) \leq e^{-t}.
\]

This result holds with \( C = 0 \) for Gaussian distributions. This result can be rewritten, denoting by \( \sigma^2 = \text{Var}(X) \),

\[
\forall t > 0, \quad \Pr\left(\sqrt{n} \frac{\sum_{i=1}^{n} X_i - \mathbb{E}[X]}{\sigma} > \sqrt{2t} + \frac{Ct}{\sigma\sqrt{n}}\right) \leq e^{-t}.
\]

This kind of deviation inequalities precise in a sense the central limit theorem as it shows that the CLT statistics deviates from 0 as a standard Gaussian random variable, up to a correction term of the order \( Ct/\sigma\sqrt{n} \).

### 6.2 Generic approaches

The first examples show that the Chernoff bound could be use to bound from above the quantiles of several usual distributions. As these quantiles are easily available on R or Python, this is not the main reason why deviation inequalities are used (it is however useful to compare numerically these upper bounds with the actual quantiles and with the approximations provided by the CLT). The power of Chernoff bound is that it allows to prove non-asymptotic upper bounds on random variables under much less restrictive assumptions on the random variables than the knowledge of their distributions. In the remaining of the chapter, we provide “generic” conditions on random variables that allow to prove concentration bounds.
6.2.1 SubGaussian random variables

**Definition 67.** A random variable $X$ is called subGaussian if there exists a constant $v^2 > 0$ such that:

$$
\forall s > 0, \quad \mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq e^{\frac{s^2v^2}{2}} .
$$

In this case, we write $X \in \text{SubGau}(v^2)$ and $v^2$ is called the variance proxy of $X$.

Example shows that Gaussian random variables are subGaussian with the variance as a variance proxy.

**Exercise:** Check that, if $X \sim \text{SubGau}(v^2)$, then $\sigma^2 \leq v^2$.

The Chernoff bound yields the following result

**Proposition 68.** If $X \in \text{SubGau}(v^2)$, then

$$
\forall t > 0, \quad \mathbb{P}(X - \mathbb{E}[X] > t) \leq e^{-t^2/(2v^2)}, \quad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2vt}) \leq e^{-t} .
$$

If $X_1, \ldots, X_n$ are independent and subGaussian, $s$ is their empirical mean as shown by the following result.

**Proposition 69 (Tensorization for subGaussian random variables).** Let $X_1, \ldots, X_n$ denote independent and subGaussian random variables such that $X_i \in \text{SubGau}(v_i^2)$ for any $i \in \{1, \ldots, n\}$. Then $\sum_{i=1}^n X_i \in \text{SubGau}(\sum_{i=1}^n v_i^2)$.

In particular, it holds

$$
\forall t > 0, \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \sqrt{\frac{2v^2t}{n}}\right) \leq e^{-t} .
$$

The proof is elementary and left as an exercise.

**Examples of subGaussian random variables**

Besides Gaussian random variables, we have seen in Example that binomial distributions are subGaussian with variance proxy $v^2 = r$. Let us provide another example.

**Proposition 70.** Let $\varepsilon$ denote a Rademacher random variable, that is, with uniform distribution on $\{-1, 1\}$. Then $\varepsilon \in \text{SubGau}(1)$.

**Proof.** Clearly, for any $s \in \mathbb{R}$, $\mathbb{E}[e^{s\varepsilon}] = (e^s + e^{-s})/2$. Developing in power series gives

$$
\mathbb{E}[e^{s\varepsilon}] = \sum_{k=0}^{+\infty} \frac{s^{2k}}{(2k)!}.
$$
Thus, for any $k$, $(2k)! > 2^k k!$, hence,
\[
\mathbb{E}[e^{se}] \leq \sum_{k=0}^{+\infty} \frac{(s^2/2)^k}{k!} = e^{s^2/2}.
\]

One can use Proposition 71 to prove that bounded random variables are subGaussian. This result is known as Hoeffding’s Lemma. To proceed, we use the following symmetrization trick.

**Lemma 71 (Symmetrization).** Let $X$ denote a random variable and let $s > 0$ be such that $\mathbb{E}[e^{sX}] < \infty$. Then, if $X'$ is a random variable independent of $X$, distributed as $X$, and if $\varepsilon$ is a Rademacher random variable, independent from $X$ and $X'$, we have
\[
\mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \mathbb{E}[e^{s(X - X')}] .
\]

The name comes from the fact that the Laplace transform of the possibly non symmetric random variable $(X - \mathbb{E}[X])$ is bounded from above by the Laplace transform of the symmetric random variable $\varepsilon(X - X')$.

**Proof.** We use the following facts:

1. $\mathbb{E}[X] = \mathbb{E}[X'|X]$ which holds by independence of $X$ and $X'$.

2. $X - X'$ is distributed as $\varepsilon(X - X')$, which can be proved by computing the Fourier transform of both variables.

Using the convexity of $x \mapsto e^{sx}$ and Jensen’s inequality yields
\[
\mathbb{E}[e^{s(X - \mathbb{E}[X])}] = \mathbb{E}[e^{s\mathbb{E}[X - X'|X]}] \leq \mathbb{E}[e^{s(X - X')} = \mathbb{E}[e^{s\varepsilon(X - X')}].
\]

We are now in position to state and prove Hoeffding’s lemma

**Lemma 72 (Hoeffding).** Let $X$ denote a random variable taking values in $[a,b]$. Then, $X \in \text{SubGau}((b - a)^2)$.

**Remark 73.** This version of Hoeffding’s result is slightly sub-optimal, with more involved arguments $X \in \text{SubGau}((b - a)^2/4)$, see Lemma 70. The optimal result is interesting as it is precise for Bernoulli random variable $\mathcal{B}(1/2)$. The elementary version that we present now is sufficient in most theoretical applications though.
Proof. By Proposition 70,
\[ \mathbb{E}[e^{s(X-X')}|X,X'] \leq e^{s^2(X-X')^2/2}. \]
As \( X \) and \( X' \) belong to \([a,b]\), it holds \((X-X')^2 \leq (b-a)^2\), so
\[ \mathbb{E}[e^{s(X-X')}|X,X'] \leq e^{s^2(b-a)^2/2}. \]
The result now follows from the symmetrization lemma. \( \square \)

Combining the previous results, one can prove the following result (do it as an exercise. Don’t hesitate to use the version of Hoeffding’s lemma given in Lemma 73).

**Theorem 74** (Hoeffding’s inequality). Assume that \( X_1, \ldots, X_n \) are independent random variables such that each \( X_i \) takes value in \([a_i,b_i]\). Then,
\[ \forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) > \sqrt{\frac{v^2 t}{2n}} \right) \leq e^{-t}, \]
with \( v^2 = n^{-1} \sum_{i=1}^{n} (b_i - a_i)^2 \).

Bounded random variables are subGaussians. More generally, the following result gives a simple condition on the moments of a random variable that is sufficient to prove that it is subGaussian.

**Proposition 75.** Assume that \( X \) is centered with subGeometric moments, meaning that there exists \( b > 0 \) such that
\[ \forall k \geq 2, \quad |\mathbb{E}[X^k]| \leq b^k. \]
Then, \( X \) is subGaussian with variance proxy \( 2b^2 \), that is
\[ \mathbb{E}[e^{sX}] \leq e^{s^2b^2}. \]
In particular, if \( X_1, \ldots, X_n \) are i.i.d. distributed as \( X \), we have
\[ \forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} X_i > 2b\sqrt{\frac{t}{n}} \right) \leq e^{-t}. \]

**Proof.** The existence of \( \mathbb{E}[e^{sX}] \) for any \( s > 0 \) comes from Fubini-Tonelli’s theorem. Let \( s > 0 \), by the symmetrization’s lemma,
\[ \mathbb{E}[e^{sX}] \leq \mathbb{E}[e^{s(X-X')}]. \]
6.2. GENERIC APPROACHES

The random variable \( \varepsilon(X - X') \) being symmetric, for any integer \( k \),
\[ E[(\varepsilon(X - X'))^{2k+1}] = 0. \]

Therefore, using a power series development,
\[ E[e^{sX}] \leq \sum_{k=0}^{+\infty} \frac{s^{2k}E[(X - X')^{2k}]}{(2k)!}. \]

We have
\[ E[(X - X')^{2k}] = \left( \frac{2k}{k} \right) E[X^k]^2 + 2 \sum_{\ell=0}^{k-1} \left( \frac{2k}{\ell} \right) E[X^{2\ell}]E[X^{2(k-\ell)}] \leq b^{2k}2^k. \]

Using the inequality \( (2k)! \geq 2^{k+1}k! \),
\[ E[e^{sX}] \leq \sum_{k=0}^{+\infty} \frac{s^{2k}b^{2k}}{k!} = e^{s^2b^2}. \]

Let us conclude this section with another proof of Hoeffding’s lemma that
provides the optimal constants

**Lemma 76** (Hoeffding). Let \( X \) denote a random variable taking values in \([a, b]\). Then \( X \in \text{SubGau}((b - a)^2/4)\).

**Proof.** Let \( \psi(s) = \log E[e^{sX}] \), we have
\[ \psi'(s) = \frac{E[Xe^{sX}]}{E[e^{sX}]}, \quad \psi''(s) = E\left[ X^2 \frac{e^{sX}}{E[e^{sX}]} \right] - \left( E\left[ X \frac{e^{sX}}{E[e^{sX}]} \right] \right)^2. \]

We deduce that \( \psi(0) = \psi'(0) = 0 \) and that \( \psi'' \) is the variance of the distribution \( \mathbb{P} \) with density \( d\mathbb{P}(x) = \frac{e^{sx}}{E[e^{sX}]} \cdot d\mathbb{P}_X(x) \). The distribution \( \mathbb{P} \) is absolutely continuous with respect to the distribution of \( X \), so it has support in \([a, b]\), so
\[ \psi''(s) = \text{Var}_\mathbb{P}(X) \leq \frac{E\left[ \left( X - \frac{b-a}{2} \right)^2 \right]}{4} \leq \frac{(b-a)^2}{4}. \]

Hence,
\[ \psi(s) = \psi(s) - \psi(0) = \int_0^s \psi'(t)dt = \int_0^s (\psi'(t) - \psi'(0))dt = \int_0^s \int_0^t \psi''(u)du dt \]
\[ \leq \int_0^s \int_0^t \frac{(b-a)^2}{4}du dt = \frac{(b-a)^2}{4} \int_0^s tdt = \frac{(b-a)^2}{8}. \]

\[ \square \]
6.2.2 SubPoissonian random variables

Deviation inequalities for subGaussian random variables makes the CLT more precise when the variance proxy and the actual variance are comparable. In the simple example of Bernoulli random variables (binomial with $r = 1$), the variance $\theta(1 - \theta)$ may be much smaller than the variance proxy $1$ when $\theta$ is close to $0$ or $1$. In this section, we present results that allow to precise the CLT in this kind of example.

**Definition 77.** A random variable $X$ is called sub-Poissonian if there exist $v^2 > 0$ and $b \geq 0$ such that

$$\forall s > 0, \quad \log \mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \frac{v^2}{b^2} (e^{bs} - 1 - bs).$$

In this case, we note $X \in \text{SubPoi}(v^2, b)$. If $b = 0$, the previous condition becomes

$$\forall s > 0, \quad \log \mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \frac{v^2 s^2}{2}.$$

This means that $X \in \text{SubGau}(v^2)$.

We have seen that Poisson variables are sub-Poissonian. As for sub-Gaussian random variables, one can provide conditions on the moments of a random variable sufficient to show that it is sub-Poissonian.

**Proposition 78.** Let $X$ denote a random variable such that

$$\forall k \geq 2, \quad \mathbb{E}[(X - \mathbb{E}[X])^2(X - \mathbb{E}[X])^{k-2}] \leq v^2 b^{k-2},$$

then $X \in \text{SubPoi}(v^2, b)$. In particular thus, if $X - \mathbb{E}[X] \leq b$, then $X \in \text{SubPoi}(v^2, b)$.

**Proof.** We use that $(e^x - 1 - x)/x^2$ is non-decreasing, so, for any $x \in \mathbb{R}$,

$$\frac{e^x - 1 - x}{x^2} \leq \frac{e^{x+} - 1 - x_+}{x_+^2}.$$

Let $s > 0$, by a development in power series

$$\mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \mathbb{E}[1 + s(X - \mathbb{E}[X])] + \sum_{k \geq 2} \frac{s^k \mathbb{E}[(X - \mathbb{E}[X])^2(X - \mathbb{E}[X])^{k-2}]}{k!} \leq 1 + \frac{v^2}{b^2} \sum_{k \geq 2} \frac{s^k b^k}{k!} = 1 + \frac{v^2}{b^2} (e^{sb} - 1 - sb).$$

We conclude by the convexity inequality $\log(1 + x) \leq x$. \qed
Proposition 78 grants in particular that any random variable such that $X - \mathbb{E}[X] \leq b$ almost surely is sub-Poissonian. It implies in particular the following result.

**Proposition 79.** Let $X$ denote a non negative random variable, then

$$-X \in \text{SubPoi}(\mathbb{E}[X]^2, \mathbb{E}[X]) .$$

**Proof.** As $X \geq 0$ almost surely, we have $0 \leq (\mathbb{E}[X] - X)_+ \leq \mathbb{E}[X]$ almost surely and such that, for any $k \geq 2$, $\mathbb{E}[(\mathbb{E}[X] - X)_+^k] \leq \mathbb{E}[X]^k$. \hfill \square

Sub-Poissonian random variables satisfy the following result.

**Proposition 80** (Concentration of sub-Poissonian random variables). If $X \in \text{SubPoi}(v^2, b)$, then

$$\forall t > 0, \quad \log \mathbb{P}(X - \mathbb{E}[X] > t) \leq -\frac{v^2}{b^2} h \left( \frac{bt}{v^2} \right) \leq -\frac{t^2}{2(v^2 + bt/3)} . \quad (6.1)$$

This result implies that

$$\forall t > 0, \quad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2v^2t} + \frac{2bt}{3}) \leq e^{-t} .$$

**Proof.** Assume first that $b = 1$. From Example 6, the Laplace transform of $X$ is bounded from above by the one of a Poisson random variable with parameter $v^2$. The computations done in this example ensure that $\psi^*(t) \geq v^2 h(t/v^2)$. Hence, we have

$$\forall t > 0, \quad \mathbb{P}(X - \mathbb{E}[X] > t) \leq e^{-v^2 h \left( \frac{t}{v^2} \right)} .$$

Suppose now that $b > 0$. Then $X/b \in \text{SubPoi}(v^2/b^2, 1)$, hence, for all $t > 0$,

$$\mathbb{P}(X - \mathbb{E}[X] > t) = \mathbb{P} \left( \frac{X}{b} - \mathbb{E} \left[ \frac{X}{b} \right] > \frac{t}{b} \right) \leq e^{-\frac{v^2}{b^2} h \left( \frac{t}{v^2/b^2} \right)} = e^{-\frac{\sigma^2}{b^2} h \left( \frac{bt}{\sigma^2} \right)} .$$

Moreover, as

$$\forall t > 0, \quad h(t) \geq \frac{t^2}{2(1 + t/3)} .$$

We deduce that

$$-\frac{\sigma^2}{b^2} h \left( \frac{bt}{\sigma^2} \right) \leq -\frac{v^2}{b^2} \frac{b^2 t^2}{2\sigma^2 (1 + bt/3\sigma^2)} = -\frac{t^2}{2(\sigma^2 + bt/3)} .$$

\hfill \square
Proposition 81 (Tensorization for sub-Poissonian random variables). Let $X_1, \ldots, X_n$ denote independent random variables such that $X_i \in \text{SubPoi}(b_i)$. Then $\sum_{i=1}^n X_i \in \text{SubPoi}(b_{\max})$, where $b_{\max} = \max\{b_i, i = 1, \ldots, n\}$. In particular,

$$\forall t > 0, \quad \mathbb{P}\left( \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_i] > \sqrt{\frac{2v^2 t}{n} + \frac{2b_{\max} t}{3n}} \right) \leq e^{-t},$$

where $v^2 = n^{-1} \sum_{i=1}^n \text{Var}(X_i)$.

Proof. The key is that the function

$$x \mapsto \frac{e^x - 1 - x}{x^2}$$

is increasing on $\mathbb{R}_+$, as shown by a development in power series. We deduce that, for any $i$ and any $s > 0$,

$$\frac{e^{sb_i} - 1 - sb_i}{b_i^2} \leq \frac{e^{sb_{\max}} - 1 - sb_{\max}}{b_{\max}^2}.$$ 

Denoting by $\sigma_i^2$ the variance of $X_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ the variance of $\sum_{i=1}^n X_i$,

$$\log \mathbb{E}[e^{s \sum_{i=1}^n (X_i - \mathbb{E}[X_i])}] = \sum_{i=1}^n \log \mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}]$$

$$\leq \sum_{i=1}^n \frac{\sigma_i^2 e^{sb_i} - 1 - sb_i}{b_i^2}$$

$$\leq \sigma^2 \frac{e^{sb_{\max}} - 1 - sb_{\max}}{b_{\max}^2}.$$ 

This shows the first result. The second is a consequence of the concentration of sub-Poissonian random variables.

6.2.3 Bernstein’s inequality

Definition 82. A random variable $X$ satisfies Bernstein’s condition if $v^2 > 0$, $b \geq 0$ such that

$$\forall s \in (0, 1/b), \quad \log \mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \frac{v^2 s^2}{2(1 - bs)}.$$ 

In this situation, we write $X \in \text{SubBer}(v^2, b)$. 
Example 9. Suppose that $X \in \text{SubPoi}(v^2, b)$. As, for any $k \geq 2$, $k! \geq 23^{k-2}$, we have

$$\log \mathbb{E}[e^{s(X - \mathbb{E}[X])}] \leq \frac{v^2 + \infty}{2} \sum_{k=2}^{\infty} \frac{b^k s^k}{k!} = \frac{v^2 s^2}{2(1 - bs/3)} .$$

In other words, $X \in \text{SubBer}(v^2, b/3)$.

Proposition 83. Let $h_1(x) = 1 + x - \sqrt{1 + 2x}$. If $X \in \text{SubBer}(v^2, b)$, we have

$$\forall t > 0, \quad \mathbb{P}(X - \mathbb{E}[X] > t) \leq e^{-v^2 s_2(b^2 t)} , \quad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2v^2 t + bt}) \leq e^{-t} .$$

Exercise 1. Show that Proposition 83 implies the last inequality in Eq (6.1).

Proof. Let $t > 0$. To apply Chernoff’s bounds, we define the function

$$f(s) = st - \frac{v^2 s^2}{2(1 - bs)} = s\left(t + \frac{v^2}{2b}\right) + \frac{v^2}{2b^2} - \frac{v^2}{2b^2(1 - bs)} .$$

We have

$$f'(s) = t + \frac{v^2}{2b} - \frac{v^2}{2b(1 - bs)} .$$

Therefore, $f$ reaches a maximum in

$$s = \frac{1}{b} \left(1 - \frac{1}{\sqrt{1 + 2bt/v^2}}\right) .$$

This maximum is equal to

$$f^*(t) = \frac{v^2}{b^2} \left(1 + \frac{bt}{v^2} - \sqrt{1 + \frac{2bt}{v^2}}\right) = \frac{v^2}{b^2} h_1(\frac{bt}{v^2}) .$$

The first result is thus a consequence of Chernoff’s bound.

For the second, write first $h_1(x) = \frac{1 + 2x}{2} - \sqrt{1 + 2x} + \frac{1}{2} = (\sqrt{1 + 2x} - 1)^2$, so that, for any $u > 0$, we have $h_1(x) = u$ if $x = [(1 + \sqrt{2u})^2 - 1]/2 = 2u + u$. Hence, since $h_1^{-1}(u) = \sqrt{2u} + 2u$ and

$$u = \frac{v^2}{b} h_1^{-1}\left(\frac{b^2 u}{v^2}\right) = \sqrt{2v^2 u + bu} .$$

Hence, from the first result,

$$\forall u > 0, \quad \mathbb{P}(X - \mathbb{E}[X] > \sqrt{2v^2 u + bu}) \leq e^{-u} .$$

\qed
Proposition 84 (Tensorization). Let $X_1, \ldots, X_n$ denote independent random variables such that, for all $i \in \{1, \ldots, n\}$, $X_i \in \text{SubBer}(v^2_i, b_i)$. Hence, if $v^2 = n^{-1} \sum_{i=1}^n v^2_i$, $b = \max\{b_i, i = 1, \ldots, n\}$, we have $\sum_{i=1}^n X_i \in \text{SubBer}(nv^2, b)$. In particular, we have therefore

$$\forall t > 0, \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \sqrt{\frac{2v^2 t}{n} + \frac{bt}{n}}\right) \leq e^{-t}.$$  

Proof. Remark that all random variables $X_i \in \text{SubBer}(v^2_i, b)$, so, by independence, for any $s \in (0, 1/b)$,

$$\log \mathbb{E}[e^{s\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}] = \sum_{i=1}^n \log \mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}] \leq \sum_{i=1}^n \frac{s^2 v^2_i}{2(1 - bs)} = \frac{ns^2 v^2}{2(1 - bs)}.$$

This shows the first point. The second point comes from Bernstein’s inequality given in Proposition 83.

The following proposition gives a sufficient condition on the moments of $X$ allowing to check that $X \in \text{SubBer}(v^2, b)$.

Proposition 85. Let $X$ denote a random variable such that

$$\forall k \geq 2, \quad \mathbb{E}[|X - \mathbb{E}[X]|^k] \leq \frac{v^2 t^{k-2} k!}{2}.$$  

Thus $X \in \text{SubBer}(v^2, b)$.

Proof. Let $s \in (0, 1/b)$. Developing in power series, we have

$$\mathbb{E}[e^{s(X - \mathbb{E}[X])}] = 1 + \frac{s^2 v^2}{2} \sum_{k=2}^{+\infty} (sb)^{k-2} = 1 + \frac{s^2 v^2}{2(1 - bs)}.$$  

We conclude by convexity of $1 + x \leq e^x$.

An example of application of this property (which will be very useful) is given by the following result.

Proposition 86. Let $X \in \text{SubGau}(v^2)$, then $(X - \mathbb{E}[X])^2 \in \text{SubBer}(16v^4, 2v^2)$. In particular, if $X_1, \ldots, X_n$ denote random variables such that, for all $i \in \{1, \ldots, n\}$, $\sigma_i^2 = \text{Var}(X_i)$, $X_i \in \text{SubGau}(v^2)$, we have

$$\forall t > 0, \quad \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \{(X_i - \mathbb{E}[X_i])^2 - \sigma_i^2\} > 2v^2 \left(\sqrt{\frac{8t}{n} + \frac{t}{n}}\right)\right) \leq e^{-t}.$$
6.3. MAXIMAL CONCENTRATION INEQUALITIES

Proof. Let us evaluate the moments of \((X - \mathbb{E}[X])^2\). We use the following elementary result whose proof is left as an exercise.

Lemma 87. Let \(X\) denote a positive random variable, with finite expectation, then

\[
\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)dt .
\]

Let \(k \geq 2\), by the lemma,

\[
\mathbb{E}[(X - \mathbb{E}[X])^{2k}] = \int_0^{+\infty} \mathbb{P}(|X - \mathbb{E}[X]| > t^{1/(2k)})dt .
\]

As \(X \in \text{SubGau}(v^2)\), for any \(t > 0\), we have \(\mathbb{P}(|X - \mathbb{E}[X]| > t^{1/(2k)}) \leq 2e^{-t^{1/k}/(2v^2)}\), so

\[
\mathbb{E}[(X - \mathbb{E}[X])^{2k}] = 2 \int_0^{+\infty} e^{-t^{1/k}/(2v^2)}dt .
\]

The change of variables \(u = t^{1/k}/(2v^2)\) gives

\[
\mathbb{E}[(X - \mathbb{E}[X])^{2k}] = 2(2^{2k}v^{2k})k \int_0^{+\infty} u^{k-1}e^{-u}du = \frac{16 v^4 (2v^2)^{k-2}k!}{2} .
\]

The result comes from Proposition 85 since

\[
((X - \mathbb{E}[X])^2 - \sigma^2)_+ \leq (X - \mathbb{E}[X])^2 .
\]

\(\square\)

6.3 Maximal concentration inequalities

Let \(X_1, \ldots\) denote independent centered random variables sub-Gaussian random variables with variance proxy \(v^2\). In the section, we are interested in providing concentration bounds for \(\sup_{t \in [1,T]} \sum_{k=1}^T X_k\). Start with an elementary lemma.

Lemma 88. For any integer \(r \geq 1\),

\[
\forall u > 0, \quad \mathbb{P}\left( \sup_{t \in [r,2r]} \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > u \right) \leq e^{-\frac{u^2}{2v^2}} .
\]
Proof. Fix \( s > 0 \). By assumption, \( e^s \sum_{k=1}^t X_k - tu^2 s^2/2 \) is a super-martingale, so, for any stopping time \( \tau \),
\[
\mathbb{E}[e^s \sum_{k=1}^t X_k - \tau u^2 s^2/2] \leq 1 .
\]
Define \( \tau = \inf\{ t \geq r : \sum_{k=1}^t X_k > \sqrt{tu}\} \), \( \tau \) is a stopping time and \( \sum_{k=1}^\tau X_k > \sqrt{\tau u} \geq \sqrt{ru} \). In addition,
\[
\left\{ \sup_{t \in [r, 2r]} \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > u \right\} = \{ \tau \leq 2r \} .
\]
Thus, by Markov’s inequality
\[
\mathbb{P}\left( \sup_{t \in [r, 2r]} \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > u \right) = \mathbb{P}(\tau \leq 2r)
\leq \mathbb{P}(e^{s\sqrt{\tau u} - ru^2 s^2/2} \geq e^{s\sqrt{ru} - ru^2 s^2})
\leq e^{-\{s\sqrt{ru} - ru^2 s^2\}} .
\]
As this holds for any \( s > 0 \), one can optimize this bound in \( s \) which yields the result.

Corollary 89. Using the notations of Lemma 88, for any integer \( r \) larger than 1, denoting by \( a_t = 8v^2 \log[(\log_2(4t/r))] \),
\[
\forall u > 0, \quad \mathbb{P}\left( \exists t \geq r : \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > \sqrt{u + a_t} \right) \leq e^{-\frac{u}{4v^2}} .
\]

Proof. By a union bound, for any non decreasing \( a_t \),
\[
\mathbb{P}\left( \exists t \geq r : \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > \sqrt{u + a_t} \right) \leq \sum_{\ell=0}^{+\infty} \mathbb{P}\left( \sup_{t \in [2^\ell r, 2^{\ell+1} r]} \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > \sqrt{u + a_{2^\ell r}} \right) .
\]
By Lemma 88,
\[
\mathbb{P}\left( \sup_{t \in [r, 2^\ell r]} \frac{\sum_{k=1}^t X_k}{\sqrt{t}} > \sqrt{u + a_t} \right) \leq \sum_{\ell=0}^{+\infty} e^{-\frac{u+a_{2^\ell r}}{4v^2}} .
\]
Since \( a_t = 8v^2 \log[(\log_2(4t/r))] = 4u^2 \log[(\log_2(t/r) + 1)(\log_2(t/r) + 2)] \), we have
\[
a_{2^\ell r} \geq 4u^2 \log[(\ell + 1)(\ell + 2)] ,
\]
so
\[
\sum_{\ell=0}^{+\infty} e^{-\frac{u+a_{2^\ell r}}{4v^2}} \leq 1 .
\]
6.3. MAXIMAL CONCENTRATION INEQUALITIES

6.3.1 Azuma-Hoeffding’s inequality

The following result is a straightforward extension of Hoeffding’s inequality.

Theorem 90. Assume that, for any \( t \), there exist random variables \( A_t, B_t \), both \( \mathcal{F}_{t-1} \) measurable such that, almost surely, \( A_t \leq \Delta_t \leq B_t \). Then,

\[
\forall u > 0, \quad \mathbb{P}\left( \frac{1}{N} \sum_{t=1}^{N} \Delta_t > u \right) \leq e^{-2N^2 u^2 / \sum_{t=1}^{N} \|B_t - A_t\|_2^2}.
\]

Proof. By the Chernoff bound, it is sufficient to bound from below the Fenchel Legendre transform \( \psi_n^{\Delta}(s) = \log \mathbb{E}[e^{(s/N) \sum_{t=1}^{N} \Delta_t}] = \log \mathbb{E}[e^{(s/N) \sum_{t=1}^{N-1} \Delta_t \mathbb{E}[e^{(s/N)\Delta_N | \mathcal{F}_{N-1}]]}] \).

Conditionally on \( \mathcal{F}_{N-1} \), \( \Delta_N \) is a random variable taking values in \([A_N, B_N]\), therefore, by Hoeffding’s lemma,

\[
\mathbb{E}[e^{(s/N)\Delta_N | \mathcal{F}_{N-1}}] \leq e^{s^2(B_N - A_N)^2/(8N^2)} \leq e^{s^2\|B_N - A_N\|_\infty^2/(8N^2)}.
\]

It follows that

\[
\psi_n^{\Delta}(s) \leq \log \mathbb{E}[e^{(s/N) \sum_{t=1}^{N-1} \Delta_t}] + \frac{\|B_N - A_N\|_\infty^2 s^2}{8N^2}.
\]

By recurrence, we get

\[
\psi_n^{\Delta}(s) \leq \frac{s^2 \sum_{t=1}^{N} \|B_t - A_t\|_\infty^2 s^2}{4N^2}.
\]

The proof is concluded by Proposition 68.

6.3.2 The bounded difference inequality

Azuma-Hoeffding’s inequality can be applied when \( X_n = \mathbb{E}[f(X_1, \ldots, X_n)] \) and \( f \) satisfies the bounded difference property. The result is known as the bounded difference inequality. Let \( N \) be an integer, \( c \in \mathbb{R}_+^N \), \( f : \mathcal{Z}^N \to \mathbb{R} \) is a measurable function and \( Z = (Z_1, \ldots, Z_N) \) is a vector of independent random variables taking values in \( \mathcal{Z} \).

Definition 91. The function \( f \) satisfies the bounded difference property with constants \( c, f \in \mathbb{BD}(c) \), if

\[
\forall x, y \in \mathbb{R}^n, \quad |f(x) - f(y)| \leq \sum_{i=1}^{n} c_i 1_{\{x_i \neq y_i\}}.
\]
**Theorem 92** (Bounded Difference Inequality). Assume that $f \in \mathcal{BD}(c)$, then
\[ \forall t > 0, \quad \mathbb{P}(f(Z) > \mathbb{E}[f(Z)] + t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}. \]

**Proof.** The proof relies on Azuma-Hoeffding’s inequality. Let $\mathcal{F}_t$ denote the sigma-algebra induced by $X_1, \ldots, X_t$, $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and let
\[ \Delta_t = \mathbb{E}[g(Z)|\mathcal{F}_t] - \mathbb{E}[g(Z)|\mathcal{F}_{t-1}] . \]

$\Delta_t$ are by construction martingale increments with respect to the filtration $\mathcal{F}_t$. Let
\[
\begin{align*}
B_t &= \mathbb{E}\left[\sup_{z_t \in Z} g(Z_1, \ldots, Z_{t-1}, z_t, Z_{t+1}, \ldots, Z_N) - g(Z)|\mathcal{F}_{t-1}\right] \\
A_t &= \mathbb{E}\left[\inf_{z_t \in Z} g(Z_1, \ldots, Z_{t-1}, z_t, Z_{t+1}, \ldots, Z_N) - g(Z)|\mathcal{F}_{t-1}\right].
\end{align*}
\]

By construction, $\Delta_t \in [A_t, B_t]$ almost surely and, since $g \in \mathcal{BD}(c)$,
\[ \|B_t - A_t\|_\infty \leq c_i . \]

The theorem follows therefore from a direct application of Azuma-Hoeffding’s inequality.

Assume that $F$ is a class of functions $g : Z \to [0, 1]$ and let
\[ f(Z_1, \ldots, Z_N) = \sup_{g \in F} \sum_{i=1}^N g(Z_i) - P g . \]

Clearly, $f \in \mathcal{BD}(c)$, with $c = (1, \ldots, 1)$. Therefore, the bounded difference inequality implies that
\[
\mathbb{P}\left(\sup_{g \in F} \sum_{i=1}^N g(Z_i) - P g > \mathbb{E}\left[\sup_{g \in F} \sum_{i=1}^N g(Z_i) - P g\right] + t\right) \leq e^{-2t^2/N} . \tag{6.2}
\]

### 6.4 Gaussian concentration inequality

The function $f : \mathbb{R}^d \to \mathbb{R}$ is $L$-Lipschitz with respect to the Euclidean norm if
\[ \forall x, y \in \mathbb{R}^d, \quad |f(x) - f(y)| \leq L\|x - y\| . \]

**Theorem 93.** Let $f : \mathbb{R}^d \to \mathbb{R}$ a $L$-Lipschitz function and let $Z$ denote a standard Gaussian vector in $\mathbb{R}^d$. Then,
\[ \forall t > 0, \quad \mathbb{P}(f(Z) - \mathbb{E}[f(Z)] > t) \leq e^{-t^2/(\sigma^2L^2)} . \]
6.4. GAUSSIAN CONCENTRATION INEQUALITY

**Remark 94.** This result can be sharpened using, for example, the entropy method of Ledoux. The best result one can prove is that, actually,

$$\forall t > 0, \quad P(f(Z) - E[f(Z)] > t) \leq e^{-t^2/(2L^2)}.$$ 

**Proof.** Without loss of generality, one can assume that $L = 1$. Assume moreover that $F$ is differentiable, so its gradient $\|\nabla F(z)\| \leq 1$, for all $z \in \mathbb{R}^d$. Let $Z'$ denote another standard Gaussian vector in $\mathbb{R}^d$, independent from $Z$. We have, by Jensen’s inequality

$$E[e^{s(F(Z) - E[F(Z)])}] \leq e^{E[e^{s(F(Z) - F(Z'))}]}.$$ 

Now, let $g : [0, 1] \to \mathbb{R}$ denote the function $g(\theta) = F(\sin(\theta)Z + \cos(\theta)Z')$. Clearly, $g$ is differentiable, $g(\pi/2) = F(Z)$ and $g(0) = F(Z')$, so,

$$F(Z) - F(Z') = g(\pi/2) - g(0) = \int_0^{\pi/2} g'(\theta) d\theta.$$ 

Now,

$$g'(\theta) = \nabla F(\sin(\theta)Z + \cos(\theta)Z')^T (\cos(\theta)Z - \sin(\theta)Z').$$

The vector

$$\begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} \sin(\theta)Z + \cos(\theta)Z' \\ \cos(\theta)Z - \sin(\theta)Z' \end{bmatrix}$$

is centered, Gaussian, with covariance matrix, as $E[G_1 G_1^T] = (\cos^2(\theta) + \sin^2(\theta))I_d = I_d$ and $E[G_1 G_2^T] = \sin(\theta) \cos(\theta) (E[ZZ^T - Z'(Z')^T]) = 0$,

$$\Sigma = \begin{bmatrix} I_d & 0 \\ 0 & I_d \end{bmatrix}.$$ 

Hence, $G_1$ and $G_2$ are independent. Conditionally on $G_1$, $g'(\theta)$ is a Gaussian random variable, centered, with variance $\|\nabla F(G_1)\|^2 \leq 1$, hence, $E[e^{\pi s g'(\theta)/2} | G_1] \leq e^{\pi^2 s^2/4}$. Therefore, by Jensen’s inequality,

$$E[e^{s(F(Z) - E[F(Z)])}] \leq \frac{1}{\pi} \int_0^{\pi/2} E[e^{\pi s g'(\theta)/2} | G_1] d\theta \leq e^{\pi s^2/4}.$$ 

It means that $F(Z)$ is sub-Gaussian with variance proxy $\sigma^2 = \pi^2/2$. The proof is concluded by Proposition \(\square\).