1

Introduction

If one observes some random variable ξ (which can be a random vector or a random process) with unknown distribution, the basic problem of statistical inference is to take a decision about some quantity s related to the distribution of ξ, for instance estimate s or provide a confidence set for s with a given level of confidence. Usually, one starts from a genuine estimation procedure for s and tries to get some idea of how far it is from the target. Since generally speaking the exact distribution of the estimation procedure is not available, the role of probability theory is to provide relevant approximation tools to evaluate it. In the situation where $\xi = \xi^{(n)}$ depends on some parameter n (typically when $\xi = (\xi_1, \ldots, \xi_n)$, where the variables $\xi_1, \ldots, \xi_n$ are independent), *asymptotic theory* in statistics uses limit theorems (Central Limit Theorems, Large Deviation Principles, etc.) as approximation tools when n is large. One of the first examples of such a result is the use of the CLT to analyze the behavior of a *maximum likelihood estimator* (*MLE*) on a given regular parametric model (independent of n) as n goes to infinity. More recently, since the seminal works of Dudley in the 1970s, the theory of probability in Banach spaces has deeply influenced the development of asymptotic statistics, the main tools involved in these applications being limit theorems for empirical processes. This led to decisive advances for the theory of asymptotic efficiency in semiparametric models for instance and the interested reader will find numerous results in this direction in the books by Van der Vaart and Wellner [120] or Van der Vaart [119].

1.1 Model Selection

Designing a genuine estimation procedure requires some prior knowledge on the unknown distribution of ξ and choosing a proper model is a major problem for the statistician. The aim of model selection is to construct data-driven criteria to select a model among a given list. We shall see that in many situations motivated by applications such as signal analysis for instance, it is useful to

allow the size of the models to depend on the sample size n. In these situations, classical asymptotic analysis breaks down and one needs to introduce an alternative approach that we call *nonasymptotic*. By nonasymptotic, we do not mean of course that large samples of observations are not welcome but that the size of the models as well as the size of the list of models should be allowed to be large when n is large in order to be able to warrant that the statistical model is not far from the truth. When the target quantity s to be estimated is a function, this allows in particular to consider models which have good approximation properties at different scales and use model selection criteria to choose from the data what is the best approximating model to be considered. In the past 20 years, the phenomenon of the concentration of measure has received much attention mainly due to the remarkable series of works by Talagrand which led to a variety of new powerful inequalities (see in particular [112] and [113]). The main interesting feature of concentration inequalities is that, unlike central limit theorems or large deviations inequalities, they are indeed *nonasymptotic*. The major issue of this series of Lectures is to show that these new tools of probability theory lead to a *nonasymptotic* theory for model selection and illustrate the benefits of this approach for several functional estimation problems. The basic examples of functional estimation frameworks that we have in mind are the following.

- **Density estimation**

  One observes $\xi_1, \ldots, \xi_n$ which are i.i.d. random variables with unknown density s with respect to some given measure μ.

- **Regression**

  One observes $(X_1, Y_1), \ldots, (X_n, Y_n)$ with

$$Y_i = s(X_i) + \varepsilon_i,\ 1 \leq i \leq n.$$

  One assumes the *explanatory* variables $X_1, \ldots, X_n$ to be independent (but nonnecessarily i.i.d.) and the regression errors $\varepsilon_1, \ldots, \varepsilon_n$ to be i.i.d. with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. s is the so-called *regression function*.

- **Binary classification**

  As in the regression setting, one still observes independent pairs

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

  but here we assume those pairs to be copies of a pair (X, Y), where the response variables Y take only two values, say: 0 or 1. The basic problem of *statistical learning* is to estimate the so-called Bayes classifier s defined by

$$s(x) = \mathbb{1}_{\eta(x) \geq 1/2}$$

  where η denotes the regression function, $\eta(x) = \mathbb{E}[Y \mid X = x]$.

- **Gaussian white noise**

  Let $s \in \mathbb{L}_2\left([0,1]^d\right)$. One observes the process $\xi^{(n)}$ on $[0,1]^d$ defined by

$$d\xi^{(n)}(x) = s(x) + \frac{1}{\sqrt{n}} dB(x), \, \xi^{(n)}(0) = 0,$$

where B denotes a Brownian sheet. The level of noise ε is here written as $\varepsilon = 1/\sqrt{n}$ for notational convenience and in order to allow an easy comparison with the other frameworks.

In all of the above examples, one observes some random variable $\xi^{(n)}$ with unknown distribution which depends on some quantity $s \in \mathcal{S}$ to be estimated. One can typically think of s as a function belonging to some space $\mathcal{S}$ which may be infinite dimensional. For instance

- In the density framework, s is a density and $\mathcal{S}$ can be taken as the set of all probability densities with respect to μ.
- In the i.i.d. regression framework, the variables $\xi_i = (X_i, Y_i)$ are independent copies of a pair of random variables (X, Y), where X takes its values in some measurable space $\mathcal{X}$. Assuming the variable Y to be square integrable, the regression function s defined by $s(x) = \mathbb{E}[Y \mid X = x]$ for every $x \in \mathcal{X}$ belongs to $\mathcal{S} = \mathbb{L}^2(\mu)$, where μ denotes the distribution of X.

One of the most commonly used method to estimate s is minimum contrast estimation.

1.1.1 Minimum Contrast Estimation

Let us consider some empirical criterion γ_n (based on the observation $\xi^{(n)}$) such that on the set $\mathcal{S}$

$$t \to \mathbb{E}[\gamma_n(t)]$$

achieves a minimum at point s. Such a criterion is called an *empirical contrast* for the estimation of s. Given some subset S of $\mathcal{S}$ that we call a *model*, a *minimum contrast estimator* $\widehat{s}$ of s is a minimizer of γ_n over S. The idea is that, if one substitutes the empirical criterion γ_n to its expectation and minimizes γ_n on S, there is some hope to get a sensible estimator of s, at least if s belongs (or is close enough) to model S. This estimation method is widely used and has been extensively studied in the asymptotic parametric setting for which one assumes that S is a given parametric model, s belongs to S and n is large. Probably, the most popular examples are maximum likelihood and least squares estimation. Let us see what this gives in the above functional estimation frameworks. In each example given below, we shall check that a given empirical criterion is indeed an empirical contrast by showing that the associated natural loss function

$$\ell(s,t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)] \tag{1.1}$$

is nonnegative for all $t \in \mathcal{S}$. In the case where $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$, we shall define an empirical criterion γ_n in the following way:

$$\gamma_n(t) = P_n[\gamma(t,.)] = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,\xi_i),$$

so that it remains to specify for each example what is the adequate function γ to be considered.

- **Density estimation**

One observes $\xi_1, \ldots, \xi_n$ which are i.i.d. random variables with unknown density s with respect to some given measure μ. The choice

$$\gamma(t,x) = -\ln(t(x))$$

leads to the *maximum likelihood criterion* and the corresponding loss function ℓ is given by

$$\ell(s,t) = \mathbf{K}(s,t),$$

where $\mathbf{K}(s,t)$ denotes the Kullback–Leibler information number between the probabilities $s\mu$ and $t\mu$, i.e.,

$$\mathbf{K}(s,t) = \int s \ln\left(\frac{s}{t}\right)$$

if $s\mu$ is absolutely continuous with respect to $t\mu$ and $\mathbf{K}(s,t) = +\infty$ otherwise. Assuming that $s \in \mathbb{L}_2(\mu)$, it is also possible to define a *least squares criterion* for density estimation by setting this time

$$\gamma(t,x) = \|t\|^2 - 2t(x)$$

where $\|.\|$ denotes the norm in $\mathbb{L}_2(\mu)$ and the corresponding loss function ℓ is in this case given by

$$\ell(s,t) = \|s-t\|^2,$$

for every $t \in \mathbb{L}_2(\mu)$.

- **Regression**

One observes $(X_1, Y_1), \ldots, (X_n, Y_n)$ with

$$Y_i = s(X_i) + \varepsilon_i,\ 1 \le i \le n,$$

where $X_1, \ldots, X_n$ are independent and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. Let μ be the arithmetic mean of the distributions of the variables $X_1, \ldots, X_n$, then least squares estimation is obtained by setting for every $t \in \mathbb{L}_2(\mu)$

$$\gamma(t,(x,y)) = (y - t(x))^2,$$

and the corresponding loss function ℓ is given by

$$\ell(s,t) = \|s-t\|^2,$$

where $\|.\|$ denotes the norm in $\mathbb{L}_2(\mu)$.

- **Binary classification**

One observes independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a pair (X, Y), where Y takes its values in $\{0, 1\}$. We take the same value for γ as in the least squares regression case but this time we restrict the minimization to the set $\mathcal{S}$ of *classifiers* i.e., $\{0, 1\}$-valued measurable functions (instead of $\mathbb{L}_2(\mu)$). For a function t which takes only the two values 0 and 1, we can write

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - t(X_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i \neq t(X_i)}$$

so that minimizing the least squares criterion means minimizing the number of misclassifications on the training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The corresponding minimization procedure can also be called *empirical risk minimization* (according to Vapnik's terminology, see [121]). Setting

$$s(x) = \mathbb{1}_{\eta(x) \geq 1/2}$$

where η denotes the regression function, $\eta(x) = \mathbb{E}[Y \mid X = x]$, the corresponding loss function ℓ is given by

$$\ell(s, t) = \mathbb{P}[Y \neq t(X)] - \mathbb{P}[Y \neq s(X)] = \mathbb{E}\left[|2\eta(X) - 1|\,|s(X) - t(X)|\right].$$

Finally, we can consider the least squares procedure in the Gaussian white noise framework too.

- **Gaussian white noise**

Recall that one observes the process $\xi^{(n)}$ on $[0, 1]^d$ defined by

$$d\xi^{(n)}(x) = s(x) + \frac{1}{\sqrt{n}} dB(x),\ \xi^{(n)}(0) = 0,$$

where W denotes a Brownian sheet. We define for every $t \in \mathbb{L}_2\left([0, 1]^d\right)$

$$\gamma_n(t) = \|t\|^2 - 2\int_0^1 t(x)\, d\xi^{(n)}(x),$$

then the corresponding loss function ℓ is simply given by

$$\ell(s, t) = \|s - t\|^2.$$

1.1.2 The Model Choice Paradigm

The main problem which arises from minimum contrast estimation in a parametric setting is the choice of a proper model S on which the minimum contrast estimator is to be defined. In other words, it may be difficult to guess

what is the right parametric model to consider in order to reflect the nature of data from the real life and one can get into problems whenever the model S is false in the sense that the true s is too far from S. One could then be tempted to choose S as big as possible. Taking S as $\mathcal{S}$ itself or as a "huge" subset of $\mathcal{S}$ is known to lead to inconsistent (see [7]) or suboptimal estimators (see [19]). We see that choosing some model S in advance leads to some difficulties:

- If S is a "small" model (think of some parametric model, defined by 1 or 2 parameters for instance) the behavior of a minimum contrast estimator on S is satisfactory as long as s is close enough to S but the model can easily turn to be false.
- On the contrary, if S is a "huge" model (think of the set of all continuous functions on $[0,1]$ in the regression framework for instance), the minimization of the empirical criterion leads to a very poor estimator of s even if s truly belongs to S.

Illustration (White Noise)

Least squares estimators (LSE) on a linear model S (i.e., minimum contrast estimators related to the least squares criterion) can be computed explicitly. For instance, in the white noise framework, if $(\phi_j)_{1\leq j\leq D}$ denotes some orthonormal basis of the D-dimensional linear space S, the LSE can be expressed as

$$\widehat{s} = \sum_{j=1}^{D} \left(\int_0^1 \phi_j(x)\, d\xi^{(n)}(x) \right) \phi_j.$$

Since for every $1 \leq j \leq D$

$$\int_0^1 \phi_j(x)\, d\xi^{(n)}(x) = \int_0^1 \phi_j(x)\, s(x)\, dx + \frac{1}{\sqrt{n}}\eta_j,$$

where the variables $\eta_1, \ldots, \eta_D$ are i.i.d. standard normal variables, the quadratic risk of $\widehat{s}$ can be easily computed. One indeed has

$$\mathbb{E}\left[\|s - \widehat{s}\|^2\right] = d^2(s, S) + \frac{D}{n}.$$

This formula for the quadratic risk perfectly reflects the model choice paradigm since if one wants to choose a model in such a way that the risk of the resulting least square estimator is small, we have to warrant that the *bias term* $d^2(s,S)$ and the *variance term* D/n are small simultaneously. It is therefore interesting to consider a family of models instead of a single one and try to select some appropriate model among the family. More precisely, if $(S_m)_{m\in\mathcal{M}}$ is a list of finite dimensional subspaces of $\mathbb{L}_2\left([0,1]^d\right)$ and $(\widehat{s}_m)_{m\in\mathcal{M}}$ be the corresponding list of least square estimators, an *ideal* model should minimize $\mathbb{E}\left[\|s - \widehat{s}_m\|^2\right]$ with respect to $m \in \mathcal{M}$. Of course, since we do not know the

bias term, the quadratic risk cannot be used as a model choice criterion but just as a benchmark.

More generally if we consider some empirical contrast γ_n and some (at most countable and usually finite) collection of models $(S_m)_{m\in\mathcal{M}}$, let us represent each model S_m by the minimum contrast estimator $\widehat{s}_m$ related to γ_n. The purpose is to select the "best" estimator among the collection $(\widehat{s}_m)_{m\in\mathcal{M}}$. Ideally, one would like to consider $m(s)$ minimizing the risk $\mathbb{E}[\ell(s,\widehat{s}_m)]$ with respect to $m\in\mathcal{M}$. The minimum contrast estimator $\widehat{s}_{m(s)}$ on the corresponding model $S_{m(s)}$ is called an *oracle* (according to the terminology introduced by Donoho and Johnstone, see [47] for instance). Unfortunately, since the risk depends on the unknown parameter s, so does $m(s)$ and the oracle is *not* an estimator of s. However, the risk of an oracle can serve as a benchmark which will be useful in order to evaluate the performance of any data driven selection procedure among the collection of estimators $(\widehat{s}_m)_{m\in\mathcal{M}}$. Note that this notion is different from the notion of true model. In other words, if s belongs to some model S_{m_0}, this does not necessarily imply that $\widehat{s}_{m_0}$ is an oracle. The idea is now to consider data-driven criteria to select an estimator which tends to mimic an oracle, i.e., one would like the risk of the selected estimator $\widehat{s}_{\widehat{m}}$ to be as close as possible to the risk of an oracle.

1.1.3 Model Selection via Penalization

Let us describe the method. The *model selection via penalization* procedure consists in considering some proper *penalty function* pen: $\mathcal{M}\to\mathbb{R}_+$ and take $\widehat{m}$ minimizing the *penalized criterion*

$$\gamma_n(\widehat{s}_m)+\mathrm{pen}(m)$$

over $\mathcal{M}$. We can then define the selected model $S_{\widehat{m}}$ and the selected estimator $\widehat{s}_{\widehat{m}}$.

This method is definitely not new. Penalized criteria have been proposed in the early 1970s by Akaike (see [2]) for penalized log-likelihood in the density estimation framework and Mallows for penalized least squares regression (see [41] and [84]), where the variance σ^2 of the errors of the regression framework is assumed to be known for the sake of simplicity. In both cases the penalty functions are proportional to the number of parameters D_m of the corresponding model S_m

- Akaike : D_m/n
- Mallows' C_p : $2D_m\sigma^2/n$.

Akaike's heuristics leading to the choice of the penalty function D_m/n heavily relies on the assumption that the dimensions and the number of the models are bounded with respect to n and n tends to infinity.

Let us give a simple motivating example for which those assumptions are clearly not satisfied.

A Case Example: Change Points Detection

Change points detection on the mean is indeed a typical example for which these criteria are known to fail. A noisy signal ξ_j is observed at each time j/n on $[0,1]$. We consider the fixed design regression framework

$$\xi_j = s\,(j/n) + \varepsilon_j,\ 1 \le j \le n$$

where the errors are i.i.d. centered random variables. Detecting change points on the mean amounts to select the "best" piecewise constant estimator of the true signal s on some arbitrary partition m with endpoints on the regular grid $\{j/n,\ 0 \le j \le n\}$. Defining S_m as the linear space of piecewise constant functions on partition m, this means that we have to select a model among the family $(S_m)_{m\in\mathcal{M}}$, where $\mathcal{M}$ denotes the collection of all possible partitions by intervals with end points on the grid. Then, the number of models with dimension D, i.e., the number of partitions with D pieces is equal to $\binom{n-1}{D-1}$ which grows polynomially with respect to n.

The Nonasymptotic Approach

The approach to model selection via penalization that we have developed (see for instance the seminal papers [20] and [12]) differs from the usual parametric asymptotic approach in the sense that:

- The number as well as the dimensions of the models may depend on n.
- One can choose a list of models because of its approximation properties: wavelet expansions, trigonometric or piecewise polynomials, artificial neural networks etc.

It may perfectly happen that many models in the list have the same dimension and in our view, the "complexity" of the list of models is typically taken into account via the choice of the penalty function of the form

$$(C_1 + C_2 L_m)\,\frac{D_m}{n}$$

where the weights L_m satisfy the restriction

$$\sum_{m\in\mathcal{M}} e^{-L_m D_m} \le 1$$

and C_1 and C_2 do not depend on n.

As we shall see, concentration inequalities are deeply involved both in the construction of the penalized criteria and in the study of the performance of the resulting *penalized estimator* $\widehat{s}_{\widehat{m}}$.

The Role of Concentration Inequalities

Our approach can be described as follows. We take as a loss function the nonnegative quantity $\ell(s,t)$ and recall that our aim is to mimic the oracle, i.e., minimize $\mathbb{E}[\ell(s,\widehat{s}_m)]$ over $m \in \mathcal{M}$.

Let us introduce the centered *empirical process*

$$\overline{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma_n(t)].$$

By definition a penalized estimator $\widehat{s}_{\widehat{m}}$ satisfies for every $m \in \mathcal{M}$ and any point $s_m \in S_m$

$$\begin{aligned}\gamma_n\left(\widehat{s}_{\widehat{m}}\right) + \mathrm{pen}(\widehat{m}) &\leq \gamma_n(\widehat{s}_m) + \mathrm{pen}(m)\\ &\leq \gamma_n(s_m) + \mathrm{pen}(m)\end{aligned}$$

or, equivalently if we substitute $\overline{\gamma}_n(t) + \mathbb{E}[\gamma_n(t)]$ to $\gamma_n(t)$

$$\overline{\gamma}_n\left(\widehat{s}_{\widehat{m}}\right) + \mathrm{pen}(\widehat{m}) + \mathbb{E}\left[\gamma_n\left(\widehat{s}_{\widehat{m}}\right)\right] \leq \overline{\gamma}_n(s_m) + \mathrm{pen}(m) + \mathbb{E}[\gamma_n(s_m)].$$

Subtracting $\mathbb{E}[\gamma_n(s)]$ on each side of this inequality finally leads to the following important bound

$$\begin{aligned}\ell\left(s,\widehat{s}_{\widehat{m}}\right) \leq \ell(s,s_m) &+ \mathrm{pen}(m)\\ &+ \overline{\gamma}_n(s_m) - \overline{\gamma}_n\left(\widehat{s}_{\widehat{m}}\right) - \mathrm{pen}(\widehat{m}).\end{aligned}$$

Hence, the penalty should be

- heavy enough to annihilate the fluctuations of $\overline{\gamma}_n(s_m) - \overline{\gamma}_n\left(\widehat{s}_{\widehat{m}}\right)$;
- but not too large since ideally we would like that $\ell(s,s_m) + \mathrm{pen}(m) \leq \mathbb{E}[\ell(s,\widehat{s}_m)]$.

Therefore, we see that an accurate calibration of the penalty should rely on a sharp evaluation of the fluctuations of $\overline{\gamma}_n(s_m) - \overline{\gamma}_n\left(\widehat{s}_{\widehat{m}}\right)$. This is precisely why we need local concentration inequalities in order to analyze the uniform deviation of $\overline{\gamma}_n(u) - \overline{\gamma}_n(t)$ when t is close to u and belongs to a given model. In other words, the key is to get a good control of the supremum of some conveniently weighted empirical process

$$\frac{\overline{\gamma}_n(u) - \overline{\gamma}_n(t)}{a(u,t)},\ t \in S_{m'}.$$

The prototype of such bounds is by now the classical Gaussian concentration inequality to be proved in Chapter 3 and Talagrand's inequality for empirical processes to be proved in Chapter 5 in the non-Gaussian case.

1.2 Concentration Inequalities

More generally, the problem that we shall deal with is the following. Given independent random variables $X_1, \ldots, X_n$ taking their values in $\mathcal{X}^n$ and some functional $\zeta : \mathcal{X}^n \to \mathbb{R}$, we want to study the concentration property of $Z = \zeta(X_1, \ldots, X_n)$ around its expectation. In the applications that we have in view the useful results are sub-Gaussian inequalities. We have in mind to prove inequalities of the following type

$$\mathbb{P}[Z - \mathbb{E}[Z] \geq x] \leq \exp\left(-\frac{x^2}{2v}\right), \text{ for } 0 \leq x \leq x_0, \tag{1.2}$$

and analogous bounds on the left tail.

Ideally, one would like that $v = \mathrm{Var}(Z)$ and $x_0 = \infty$. More reasonably, we shall content ourselves with bounds for which v is a "good" upper bound for $\mathrm{Var}(Z)$ and x_0 is an explicit function of n and v.

1.2.1 The Gaussian Concentration Inequality

In the Gaussian case, this program can be fruitfully completed. We shall indeed see in Chapter 3 that whenever $\mathcal{X}^n = \mathbb{R}^n$ is equipped with the canonical Euclidean norm, $X_1, \ldots, X_n$ are i.i.d. standard normal and ζ is assumed to be Lipschitz, i.e.,

$$|\zeta(y) - \zeta(y')| \leq L\|y - y'\|, \text{ for every } y, y' \text{ in } \mathbb{R}^n$$

then, on the one hand $\mathrm{Var}(Z) \leq L^2$ and on the other hand the Cirelson–Ibragimov–Sudakov inequality ensures that

$$\mathbb{P}[Z - \mathbb{E}[Z] \geq x] \leq \exp\left(-\frac{x^2}{2L^2}\right), \text{ for all } x \geq 0.$$

The remarkable feature of this inequality is that its dependency with respect to the dimension n is entirely contained in the expectation $\mathbb{E}[Z]$. Extending this result to more general situations is not so easy. It is in particular unclear to know what kind of regularity conditions should be required on the functional ζ. A Lipschitz type condition with respect to the Hamming distance could seem to be a rather natural and attractive candidate. It indeed leads to interesting results as we shall see in Chapter 5. More precisely, if d denotes Hamming distance on $\mathcal{X}^n$ defined by

$$d(y, y') = \sum_{i=1}^{n} \mathbb{1}_{y_i \neq y'_i}, \text{ for all } y, y' \text{ in } \mathcal{X}^n$$

and ζ is assumed to be Lipschitz with respect to d

$$|\zeta(y) - \zeta(y')| \leq Ld(y, y'), \text{ for all } y, y' \text{ in } \mathcal{X}^n \tag{1.3}$$

then it can be proved that

$$\mathbb{P}\left[Z - \mathbb{E}\left[Z\right] \geq x\right] \leq \exp\left(-\frac{2x^2}{nL^2}\right), \text{ for all } x \geq 0.$$

Let us now come back to the functional which naturally emerges from the study of penalized model selection criteria.

1.2.2 Suprema of Empirical Processes

Let us assume T to be countable in order to avoid any measurability problem. The supremum of an empirical process of the form

$$Z = \sup_{t \in T} \sum_{i=1}^{n} f_t\left(X_i\right)$$

provides an important example of a functional of independent variables both for theory and applications. Assuming that $\sup_{t \in T} \left\|f_t\right\|_\infty \leq 1$ ensures that the mapping

$$\zeta : y \to \sup_{t \in T} \sum_{i=1}^{n} f_t\left(y_i\right)$$

satisfies the Lipschitz condition (1.3) with respect to the Hamming distance d with $L = 2$ and therefore

$$\mathbb{P}\left[Z - \mathbb{E}\left[Z\right] \geq x\right] \leq \exp\left(-\frac{x^2}{2n}\right), \text{ for all } x \geq 0. \tag{1.4}$$

However, it may happen that the variables $f_t\left(X_i\right)$ have a "small" variance uniformly with respect to t and i. In this case, one would expect a better variance factor in the exponential bound but obviously Lipschitz's condition with respect to Hamming distance alone cannot lead to such an improvement.

In other words Lipschitz property is not sharp enough to capture the *local* behavior of empirical processes which lies at the heart of our analysis of penalized criteria for model selection. It is the merit of Talagrand's inequality for empirical processes to provide an improved version of (1.4) which will turn to be an efficient tool for analyzing the uniform increments of an empirical process as expected.

It will be one of the main goals of Chapter 5 to prove the following version of Talagrand's inequality. Under the assumption that $\sup_{t \in T} \left\|f_t\right\|_\infty \leq 1$, there exists some absolute positive constant η such that

$$\mathbb{P}\left[Z - \mathbb{E}\left[Z\right] \geq x\right] \leq \exp\left(-\eta\left(\frac{x^2}{\mathbb{E}\left[W\right]} \wedge x\right)\right), \tag{1.5}$$

where $W = \sup_{t \in T} \sum_{i=1}^{n} f_t^2\left(X_i\right)$. Note that (1.5) a fortiori implies some sub-Gaussian inequality of type (1.2) with $v = \mathbb{E}\left[W\right] / \left(2\eta\right)$ and $x_0 = \sqrt{\mathbb{E}\left[W\right]}$.

1.2.3 The Entropy Method

Building upon the pioneering works of Marton (see [87]) on the one hand and Ledoux (see [77]) on the other hand, we shall systematically derive concentration inequalities from information theoretic arguments. The elements of information theory that we shall need will be presented in Chapter 2 and used in Chapter 5. One of the main tools that we shall use is the duality formula for entropy. Interestingly, we shall see how this formula also leads to statistical minimax lower bounds. Our goal will be to provide a simple proof of Talagrand's inequality for empirical processes and extend it to more general functional of independent variables. The starting point for our analysis is Efron–Stein's inequality. Let $X' = X_1^{'}, \ldots, X_n^{'}$ be some independent copy of $X = X_1, \ldots, X_n$ and define

$$Z_i' = \zeta\left(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n\right).$$

Setting

$$V^+ = \sum_{i=1}^n \mathbb{E}\left[\left(Z - Z_i'\right)_+^2 \mid X\right],$$

Efron–Stein's inequality (see [59]) ensures that

$$\mathrm{Var}\left(Z\right) \leq \mathbb{E}\left[V^+\right]. \tag{1.6}$$

Let us come back to empirical processes and focus on centered empirical processes for the sake of simplicity. This means that we assume the variables X_i to be i.i.d. and $\mathbb{E}\left[f_t\left(X_1\right)\right] = 0$ for every $t \in T$. We also assume T to be finite and consider the supremum of the empirical process

$$Z = \sup_{t \in T} \sum_{i=1}^n f_t\left(X_i\right),$$

so that for every i

$$Z_i' = \sup_{t \in T}\left[\left(\sum_{j \neq i}^n f_t\left(X_j\right)\right) + f_t\left(X_i'\right)\right].$$

Taking t^* such that $\sup_{t \in T} \sum_{j=1}^n f_t\left(X_j\right) = \sum_{j=1}^n f_{t^*}\left(X_j\right)$, we have for every $i \in [1, n]$

$$Z - Z_i' \leq f_{t^*}\left(X_i\right) - f_{t^*}\left(X_i'\right)$$

which yields

$$\left(Z - Z_i'\right)_+^2 \leq \left(f_{t^*}\left(X_i\right) - f_{t^*}\left(X_i'\right)\right)^2$$

and therefore by independence of X_i' from X we derive from the centering assumption $\mathbb{E}\left[f_t\left(X_i^{'}\right)\right] = 0$ that

$$\mathbb{E}\left[(Z - Z_i')_+^2 \mid X\right] \leq f_{t^*}^2(X_i) + \mathbb{E}\left[f_{t^*}^2(X_i')\right].$$

Hence, we deduce from Efron–Stein's inequality that

$$\mathrm{Var}(Z) \leq 2\mathbb{E}[W],$$

where $W = \sup_{t\in T} \sum_{i=1}^n f_t^2(X_i)$.

The conclusion is therefore that the variance factor appearing in Talagrand's inequality turns out to be the upper bound which derives from Efron–Stein's inequality. The main guideline that we shall follow in Chapter 5 is that, more generally, the adequate variance factor v to be considered in (1.2) is (up to some absolute constant) the upper bound for the variance of Z provided by Efron–Stein's inequality.