UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2023-124
The Faculty of Science, Technology and Medicine

# DISSERTATION

Defence held on 14/12/2023 in Esch-sur-Alzette

to obtain the degree of

# DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

# EN MATHÉMATIQUES

by

Alexandre LECESTRE
Born on 6 February 1996 in Rochester Hills (USA)

# ROBUST ESTIMATION FOR POSSIBLY DEPENDENT OBSERVATIONS: APPLICATION TO MIXTURE AND HIDDEN MARKOV MODELS

## Dissertation defence committee

Dr. Yannick Baraud, dissertation supervisor
*Professor, Université du Luxembourg*

Dr. Lutz Dümbgen
*Professor, Universität Bern*

Dr. Mark Podolskij, Chairman
*Professor, Université du Luxembourg*

Dr. Markus Reiß
*Professor, Humboldt-Universität zu Berlin*

Dr. Elisabeth Gassiat, Vice Chairman
*Professor, Université Paris-Saclay*

**Acknowledgements**

First and foremost, I would like to thank Yannick for giving me the opportunity to do this PhD under his supervision, dealing with very interesting topics. I am extremely grateful for his guidance, advice and patience. I learned a lot about doing research and especially presenting results as clearly and simply as possible which, hopefully, is the case in this thesis.

I would like to mention that my PhD was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017.

I am honored that Lutz Dümbgen and Elisabeth Gassiat accepted to review this manuscript. I would also like to thank the other members of the jury Mark Podolskij and Markus Reiß.

I met many people that made my time at the university so special. Listing all the names would be too long so I would like to thank all the PhD students and postdocs that have made the department an enjoyable and familiar environment.

In particular, thank you to Juntong and Shiwi. Sharing an office with the two of you has been a real pleasure and made the tough times more bearable.

Certain persons from the department and the *Bureau des Études Doctorales* made the administrative tasks in our everyday life much easier. Thank you to Katharina, Marie, Catherine, Suzanne and Émilie.

A big thank you to my roommates and dear friends, Tessa and Paris. My time in Luxembourg will definitely be associated with the moments we shared together.

The pandemic and this PhD has prevented me from seeing my friends as much as I would like to but they all have been very understanding. I hope we can make up for it in the years to come. A special thanks to Fabrice, Antonin, Clémence, Antoine, Dorian, Guillaume, Agnès, and many others, for being great friends, good listeners and sharing your invaluable wisdom ;)

I want to thank my family: my parents and my brothers. They have been supporting me unconditionally from the beginning.

To Angela, I am extremely grateful for all you have done to help me the best you could. I know it was not always easy for you. For your love, patience and support, for the confidence boosts you gave me when I was needing them the most, I cannot thank you enough.

# Contents


**1 Introduction** **1**

1.1 Mixture models . . . 1
1.2 Hidden Markov models . . . 5
1.3 Diffusion processes . . . 9
1.4 Contribution . . . 10
1.4.1 Framework . . . 12
1.4.2 Results of $\rho$-estimation . . . 13
1.4.3 Mixture models (Chapter 2) . . . 15
1.4.4 Dealing with dependent observations and applications (Chapter 3) . . . 19
1.4.5 Model selection for finite state space HMMs (Chapter 4) . . . 23
1.4.6 General HMMs (Chapter 5) . . . 24
1.5 Reminder of Vapnik-Chervonenkis theory . . . 24
1.6 Possible extensions . . . 26
1.7 Organization of the thesis . . . 26

**2 Finite mixture models** **27**

2.1 Introduction . . . 28
2.2 The statistical framework . . . 30
2.3 Estimation on a mixture model based on simple emission families . . . 32
2.3.1 Construction of the estimator on $\mathscr{Q}_K$ . . . 32
2.3.2 The performance of the estimator . . . 33
2.3.3 The case of totally bounded emission models . . . 35
2.3.4 Application to the estimation of a general Gaussian mixture . . . 37
2.3.5 Parameter estimation . . . 38
2.4 Model selection . . . 43
2.4.1 Construction of the estimator . . . 43
2.4.2 Estimation on a mixture model based on composite emission families . . . 44
2.4.3 Selection of the order $K$ . . . 45
2.A Main result . . . 48
2.A.1 Proof of Proposition 2.3 . . . 48
2.B Theorems . . . 57
2.B.1 Proof of Theorem 2.12 . . . 58
2.B.2 Proof of Theorem 2.1 . . . 60
2.B.3 Proof of Theorem 2.2 . . . 60
2.B.4 Proof of Theorem 2.8 . . . 62
2.B.5 Proof of Theorem 2.10 . . . 62
2.C Density estimation . . . 63
2.C.1 Proof of Theorem 2.3 . . . 63
2.C.2 Proof of Proposition 2.1 . . . 63
2.C.3 Proof of Theorem 2.11 . . . 66

2.D Regular parametric models . . . 67
2.D.1 Proof of Theorem 2.4 . . . 67
2.D.2 Proof of Theorem 2.9 . . . 69
2.D.3 Proof of Theorem 2.5 . . . 73
2.E Two-component mixture models . . . 73
2.E.1 Proof of Theorem 2.6 . . . 73
2.E.2 Proof of Theorem 2.7 . . . 74
2.F VC-subgraph classes of functions . . . 79
2.F.1 Proof of Lemma 2.1 . . . 80
2.F.2 Proof of Lemma 2.2 . . . 80

**3 Dependent observations** **83**
3.1 Introduction . . . 84
3.2 Construction of the estimator and main result . . . 86
3.2.1 Reminders of $\rho$-estimation . . . 87
3.2.2 From independent to dependent data . . . 88
3.2.3 Main result . . . 89
3.2.4 Robust properties of our estimator . . . 90
3.2.5 The particular case of Markov chains . . . 90
3.3 Estimation of the invariant distribution of a diffusion process . . . 91
3.3.1 Langevin equation . . . 91
3.3.2 The framework . . . 92
3.3.3 log-concave densities . . . 92
3.4 Hidden Markov models . . . 95
3.4.1 Stationary hidden Markov models . . . 95
3.4.2 The framework . . . 96
3.4.3 Estimation . . . 96
3.5 Selection of the spacing parameter . . . 107
3.5.1 Framework and result . . . 107
3.5.2 Robustness . . . 108
3.5.3 Application to hidden Markov models . . . 109
3.A Auxiliary results . . . 112
3.A.1 Proof of Lemma 3.1 . . . 112
3.B Main results . . . 113
3.B.1 Proof of Theorem 3.1 . . . 114
3.B.2 Proof of Lemma 3.2 . . . 115
3.B.3 Proof of Corollary 3.1 . . . 115
3.B.4 Proof of Lemma 3.3 . . . 116
3.B.5 Proof of Lemma 3.4 . . . 116
3.C Kolmogorov processes . . . 116
3.C.1 Proof of Theorems 3.2 and 3.3 . . . 117
3.C.2 Proof of Lemma 3.5 . . . 117
3.C.3 Proof of Lemma 3.6 . . . 119
3.D Hidden Markov models . . . 126
3.D.1 Proof of Theorem 3.4 . . . 126
3.D.2 Proof of Corollary 3.3 . . . 130
3.D.3 Proof of Theorems 3.5 and 3.6 . . . 130
3.D.4 Proof of Proposition 3.1 . . . 130
3.D.5 Proof of Proposition 3.2 . . . 131
3.D.6 Proof of Theorem 3.9 . . . 135

3.D.7 Proof of Theorem 3.10 . . . 138
3.D.8 Proof of Proposition 3.3 . . . 139
3.E Selection of the spacing parameter . . . 149
3.E.1 Proof of Theorem 3.11 . . . 149
3.E.2 Proof of Lemma 3.7 . . . 154
3.E.3 Proof of Theorem 3.12 . . . 155
3.E.4 Proof of Corollary 3.6 . . . 157

**4 Model selection for HMMs** **159**
4.1 Introduction . . . 160
4.2 The model selection procedure . . . 160
4.2.1 Reminders of $\rho$-estimation . . . 160
4.2.2 The estimator . . . 161
4.3 Application to finite state space hidden Markov models . . . 161
4.3.1 The framework . . . 161
4.3.2 General result of model selection . . . 163
4.3.3 Selection of the order . . . 164
4.3.4 Selection of the emission models . . . 165
4.A General results . . . 168
4.A.1 Proof of Theorem 4.4 . . . 168
4.B Application to HMMs . . . 169
4.B.1 Proof of Theorem 4.1 . . . 169
4.B.2 Proof of Corollary 4.1 . . . 170
4.B.3 Proof of Theorem 4.2 . . . 170
4.B.4 Proof of Theorem 4.3 . . . 171
4.B.5 Exponential families . . . 171

**5 General state space HMMs** **173**
5.1 Introduction . . . 174
5.2 The framework . . . 174
5.2.1 Reminders of $\rho$-estimation . . . 175
5.2.2 Our estimation procedure . . . 176
5.3 Approximation of general HMMs by finite mixtures . . . 177
5.4 Main result . . . 177
5.A General . . . 179
5.B General hidden Markov models . . . 179
5.B.1 Proof of Theorem 5.1 . . . 179
5.B.2 Proof of Theorem 5.2 . . . 179
5.B.3 Proof of Corollary 5.1 . . . 180
5.B.4 Proof of Proposition 5.1 . . . 181
5.B.5 Proof of Lemma 5.1 . . . 184

# Chapter 1

# Introduction

In this dissertation, we consider different problems that all fit within the following framework.

**General Problem.** *We observe $n$ (possibly dependent) random variables $X_1, \ldots, X_n$ on a measurable space $(\mathscr{X}, \mathcal{X})$ which are presumed to have a common distribution $\overline{P}$ and we aim at estimating $\overline{P}$. When $\overline{P}$ belongs to a parametric family of distributions $\{P_\theta; \theta \in \Theta\}$ which is identifiable, we also want to estimate the parameter $\overline{\theta}$ such that $\overline{P} = P_{\overline{\theta}}$.*

This includes the generic situation of probability estimation from independent and identically distributed (i.i.d.) observations but also the estimation of the stationary distribution of discrete time processes. In this context, we denote by $\boldsymbol{\mathscr{P}}_{\mathbf{X}}$ the class of all probability distributions on the measurable space $(\boldsymbol{\mathscr{X}}, \boldsymbol{\mathcal{X}}) = (\mathscr{X}^n, \mathcal{X}^{\otimes n})$.

**Definition 1.1.** *We call* model *any (nonvoid) subset $\boldsymbol{\mathscr{M}}$ of $\boldsymbol{\mathscr{P}}_{\mathbf{X}}$.*

In the case where observations are assumed to be i.i.d., a model $\boldsymbol{\mathscr{M}}$ should naturally be of the form

$$\boldsymbol{\mathscr{M}} = \left\{ P^{\otimes n}; P \in \mathscr{M} \right\}, \tag{1.1}$$

where $\mathscr{M}$ is a subset of $\mathscr{P}_X$, the class of all probability distributions on $(\mathscr{X}, \mathcal{X})$. In that case, we might informally make the abuse of calling $\mathscr{M}$ the model. We consider the estimation of the distribution $\overline{P}$ for different types of models. We first consider the simpler case of independent observations with mixture models. In a second time we consider models for dependent observations, namely hidden Markov models or discretely observed diffusion processes. We present those models with a review of the related literature hereafter.

## 1.1 Mixture models

Mixture distributions are a flexible tool for modeling heterogeneous data in an independent context. We illustrate it with the following toy example.

**Example 1.1.** *In a population, a proportion $w \in (0,1)$ is diseased and therefore a proportion $1 - w$ is healthy. We have access to one of the vital signs of each individual that we denote $X$. This quantity $X$ is distributed according to a distribution $F_0$ for the healthy population and according to a distribution $F_1 \neq F_0$ for the diseased population. Therefore the vital sign $X$ of an individual chosen at random from the overall population is distributed according to the distribution*

$$P = (1 - w)F_0 + wF_1. \tag{1.2}$$

The distribution $P$ is a two-component mixture distribution. Different problems are worth investigating in this situation such as the estimation of the mixture distribution $P$, or also the estimation of the different features of this distribution such as the proportion of diseased $w$ or the distributions $F_0$ and $F_1$ characterizing each health status. In machine learning, people are also interested in clustering. In this example it means guessing if an individual $i$ is healthy or diseased given its vital sign $X_i$.

Example 1.1 can be generalized to model a wide variety of phenomena. For a complete introduction to mixture models and an overview of the different applications we refer to the books of Mclachlan & Peel [71] and Frühwirth-Schnatter [38]. Finite mixture models contain distributions of the form

$$P_{w,F} = \sum_{k=1}^{K} w_k F_k, \tag{1.3}$$

where $K \geq 1$, $w$ belongs to the simplex $\mathcal{W}_K = \{w \in [0,1]^K; w_1 + \cdots + w_k = 1\}$ and $F_1, \ldots, F_K$ are probability distributions on the same measurable space, e.g. $(\mathscr{X}, \mathcal{X})$ in the situation of the **General Problem**. The distribution $P_{w,F}$ is a mixture with $K$ components, each component $k$ is characterized by the proportion $w_k$ called the *weight* and the probability $F_k$ called *emission distribution*.

One can easily see that the different parameters are not identifiable in general. We can always add an arbitrary number of components with null weights or merge components, e.g. $P_{w,F}$ is also a mixture with one component and emission distribution $P_{w,F}$. Therefore one usually considers mixtures with restrictions on the emission distributions in order to avoid this problem. The most common example is Gaussian mixture models (GMMs), where all the emission distributions are Gaussian. In that context, we can define as follows the canonical number of components of a mixture distribution called the order.

**Definition 1.2.** *Given a fixed class of distributions $\mathscr{F}$, we can define the* order *of a finite mixture distribution $\overline{P}$ as the smallest integer $K$ such that $\overline{P} = P_{w,F}$ with $F_1, \ldots, F_K \in \mathscr{F}$. Therefore the order depends on the class of distributions $\mathscr{F}$,*

We can notice that even if the order is identifiable it does not mean that the parameters are identifiable as shown by the following example. For real numbers $a < b$ we denote by $\mathcal{U}(a,b)$ the uniform distribution on the segment $[a,b]$. If $\mathscr{F}$ is the class of uniform distributions given by $\mathscr{F} = \{\mathcal{U}(a,b); a < b\}$, the distribution $\overline{P} = \frac{3}{4}\mathcal{U}(0,3) + \frac{1}{4}\mathcal{U}(1,2)$ is of order 2 but can also be written as

$$\overline{P} = \frac{1}{2}\mathcal{U}(0,2) + \frac{1}{2}\mathcal{U}(1,3).$$

Therefore, restricting the emission distributions to a specific class is enough to establish a canonical number of components. However, it is generally not enough to guarantee that the parameters are identifiable. Also, one should notice that identifiability can only be up to relabeling of the components, i.e. we have

$$P_{w,F} = \sum_{k=1}^{K} w_{\tau(k)} F_{\tau(k)}$$

for any permutation $\tau$ on $\{1,2,\ldots,K\}$. Sufficient conditions for identifiability have been established by Chandra [21], Henna [51] or Atienza *et al.* [7] for instance. Those conditions allowed to prove identifiability for some parametric emission models such as normal distributions, gamma distributions, or Weibull distributions. Gassiat [45] presents a review of the different situations for which we have identifiability, in a nonparametric context. If the parameters are identifiable their estimation is a well defined problem. However, one might have to estimate the order of the target distribution if it is unknown.

The monograph of Titterington *et al.* [83] and the book of Frühwirth-Schnatter *et al.* [39] provide a good overview of the different statistical procedures that have been developed for mixture models. Bayesian and likelihood-based approaches are the most commonly used but methods based on moments or spectral methods have also been considered. As finite mixtures are used to describe heterogeneous data, they are a common tool in model-based clustering. McLachlan & Basford [70] give a good overview of the topic, mainly with GMMs. We are mostly interested in density estimation based on mixtures and the estimation of the different parameters.

One can look at Figueiredo [37] as a general formulation for the problem of the estimation of the parameters, with parametric models for the emission distributions. Their approach is probably the most common one and uses a penalized likelihood criterion to select the order and estimate the parameters. The sole estimation of the order is also a problem itself that has been investigated. Dacunha-Castelle & Gassiat [23] provide a convergence rate for their estimator of the order in parametric models based on moments of the parameters. Kéribin [56] proves the strong consistency of an estimator of the order based on a penalized likelihood approach. They prove that the associated estimators of the other parameters are strongly consistent. This type of result is quite usual, along with results of asymptotic normality for the parameter estimators.

On the other hand, non-asymptotic results are very rare, especially for the estimation of the parameters. We are only aware of the results of Gadat *et al.* [40] which consider the following problem. The true distribution has a density $f^*$ with respect to the Lebesgue measure on $\mathbb{R}^d$ given by

$$f^* = (1-\lambda^*)\phi + \lambda^*\phi(\cdot - z^*), \tag{1.4}$$

where $\phi$ is a known square integrable density and the parameters $(\lambda^*,z^*) \in (0,1)\times\mathbb{R}^d\setminus\{0\}$ are to be estimated. They prove an oracle inequality for their least squares estimator $\hat{f}$ of $f^*$ with respect to the $L_2$-loss. Their estimator being of the form $\hat{f} = (1-\hat{\lambda})\phi + \hat{\lambda}\phi(\cdot - \hat{z})$ it naturally gives estimators $\hat{\lambda}$ and $\hat{z}$ of $\lambda^*$ and $z^*$. Under some regularity conditions on the density $\phi$, they establish non-asymptotic deviation bounds for the parameter estimators which lead to the usual $1/\sqrt{n}$ parametric rate with respect to the Euclidean distance, up to a logarithmic factor. This is for fixed parameters $\lambda^*$ and $z^*$. They also investigate how those rates are deteriorated when those parameters are allowed to go to 0 with $n$ which corresponds to the limit cases for identifiability.

Finite mixtures are also used a lot in density (or distribution) estimation as their flexibility allows them to approximate distributions that are quite complex. We refer to Li & Barron [66] as a good introduction to the subject. They present the generic approach in a general context. Let $\mathcal{G} = \{\phi_\theta; \theta\in\Theta\}$ be a parametric set of densities, typically with respect to the Lebesgue measure on $\mathbb{R}^d$. A density $f$ is said to have a mixture representation if it can be written as

$$f(x) = \phi_Q(x) = \int_\Theta \phi_\theta(x) Q(d\theta), \tag{1.5}$$

where $Q$ is a probability distribution on $\Theta$. Under some conditions on $\mathcal{G}$ and $Q$, such a density can be well approximated by finite mixtures with emission densities in $\mathcal{G}$, given that the number of components is large enough. Therefore we can use standard estimation procedures for finite mixtures to estimate densities with a mixture representation. Different examples have been considered for the family $\mathcal{G}$. Bochkina & Rousseau [18] consider mixtures of Gamma distributions, i.e.

$$\phi_\theta(x) = x^{z-1}e^{-zx/\theta}\left(\frac{z}{\theta}\right)^z \frac{1}{\Gamma(z)}, x,\theta\in(0,\infty),$$

where $z$ is a parameter that they estimate and $\Gamma$ is the Gamma function given by

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt,$$

for all $z$ in $(0,\infty)$. They obtain posterior convergence rate for a Bayesian estimator when the true density satisfies regularity and tail conditions. Kruijer *et al.* [59] and Rousseau [77] obtain similar results with different classes of densities. In [59], they consider location mixtures of densities of the form

$$\phi_\theta(x) = \frac{1}{2\sigma\Gamma\left(1+\frac{1}{p}\right)} e^{-(|x-\theta|/\sigma)^p}, x,\theta \in \mathbb{R}, \tag{1.6}$$

where $\sigma$ is a scale parameter they estimate. In [77], they consider mixture of beta-densities, i.e. for

$$\phi_\theta(x) = x^{a-1}(1-x)^{b-1}\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \theta = (a,b) \in (0,\infty)^2.$$

Maugis-Rabusseau & Michel [68] consider location mixtures of normal distributions, i.e.

$$\phi_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, x,\theta \in \mathbb{R},$$

where the scale parameter $\sigma$ depends on the number components which is selected with a penalized likelihood criterion. They establish non-asymptotic deviation bounds for their estimator with respect to the Hellinger loss and prove it is minimax adaptive to the regularity of the true density, up to a logarithmic factor.

Even though finite mixtures are flexible and approximate wide classes of distributions, they can be restrictive to model real phenomenona. For instance, in Example 1.1 there are only two states which correspond to "sick" or "healthy" and it can appear quite simplistic. One could refine the model with a high number of components. Another approach is to consider general mixtures, such as the density $\phi_Q$ considered in (1.5).

**Example 1.2.** *Let $(\Theta,\mathcal{T})$ be a measurable space and $\mathscr{F} = \{F_\theta; \theta \in \Theta\}$ be a parametric family of distributions on $(\mathscr{X},\mathcal{X})$. Each individual has a health level $H$ in $\Theta$ and given $H = \theta$, $X$ is distributed according to the distribution $F_\theta \in \mathscr{F}$. Therefore the vital sign $X$ of an individual chosen at random from the overall population will follow the distribution $P$ given by*

$$P(A) = \int_\Theta F_\theta(A) P_H(d\theta), \forall A \in \mathcal{X}, \tag{1.7}$$

*where the mixing distribution $P_H$ is the distribution of $H$ within the population.*

Example 1.1 is a special case where the mixing distribution $P_H$ is a discrete distribution with two support points. The most common problem involving such general mixtures is the problem of deconvolution, widely studied in the signal processing community. It corresponds to the model of additive measurement error where observations $X_1,\ldots,X_n$ are given by

$$X_i = H_i + \xi_i,$$

where $H_1,\ldots,H_n$ are i.i.d. random variables, referred to as the signal, and $\xi_1,\ldots,\xi_n$ are i.i.d. random variables independent of $H$, referred to as the noise. If $\xi_i$ has a density $\phi$ and $H_i$ has a distribution $P_H$, then $X$ has a density

$$f(x) = \int_{\mathbb{R}^d} \phi(x-h) P_H(dh), \tag{1.8}$$

which is given by the convolution of $\phi$ and $P_H$. Therefore the problem of estimating $P_H$ from observations $X_1,\ldots,X_n$ is often called the deconvolution problem. We can see from (1.8) that, in this case, the distribution of $X$ is a general location mixture associated with the density $\phi$. We refer to the book Meister [73] for a complete introduction to the subject which includes a review of estimation methods and results, all of which are asymptotic results, i.e. consistency results,

convergence rates or asymptotic normality results. We can mention a few non-asymptotic results. Dedecker *et al.*[26] prove a bound on the $p$-Wasserstein loss, $p \geq 1$, for their estimator of $P_H$ when observations belong to $\mathbb{R}$ and the density $\phi$ is smooth and known. Gassiat *et al.* [47] consider the case where the distribution of the noise is unknown. They provide an identifiability in the multivariate case with conditions on the structure of the noise and the signal. They establish convergence rates for the estimator of the density of $H$ based on Fourier inversion techniques, which is very common for deconvolution problems.

Some results exist for general mixtures that are not based on location families. Genovese & Wasserman [48] and Ghosal & van der Vaart [49] consider the estimation of $P$ given by (1.7) where $\mathscr{F}$ is the location-scale family of univariate normal distribution. They obtain convergence rates for a maximum likelihood estimator based on sieves, i.e. finite mixtures, with respect to the Hellinger loss when the true mixture distribution is compactly supported. They also consider the case where the mixing distribution satisfies light tail conditions.

## 1.2 Hidden Markov models

We can build upon Example 1.1 to introduce hidden Markov models with another toy example. We were only considering the distribution of the vital sign $X$ across the population at a fixed time. When the vital sign is discretely observed over time we can model its behavior as follows.

**Example 1.3.** *We only consider one individual and observe its daily average vital sign over $n$ days. We denote by $X_i$ and $H_i$ their average vital sign and their health status on day $i$ respectively, with the convention $H_i = 0$ when the individual is healthy and $H_i = 1$ when sick. We model $(H_i)_{1\leq i\leq n}$ as a homogeneous Markov chain and we denote by $q_{a,b} = \mathbb{P}(H_{i+1} = b|H_i = a)$, $a,b \in \{0,1\}$, the transition probabilities. In that case, the distribution of $(X_1,\dots,X_n)$ is given by*

$$\begin{aligned} P &= \sum_{h_1,\dots,h_n\in\{0,1\}} \mathbb{P}(H_1 = h_1,\dots,H_n = h_n) \bigotimes_{i=1}^{n} F_{h_i} \\ &= \sum_{h_1,\dots,h_n\in\{0,1\}} \mathbb{P}(H_1 = h_1) q_{h_1,h_2}\dots q_{h_{n-1},h_n} \bigotimes_{i=1}^{n} F_{h_i}. \end{aligned} \tag{1.9}$$

It is called a hidden Markov model as the Markov chain $(H_i)_i$ is not observed and $(X_i)_i$ is the only accessible data. As for Example 1.1, different statistical problems are of interest. One might want to estimate the features of the model, i.e. the transition probabilities $(q_{a,b})_{a,b\in\{0,1\}}$, the initial probabilities $\mathbb{P}(H_1 = 0)$,$\mathbb{P}(H_1 = 1)$ and the distributions $F_0$ and $F_1$ characterizing each health status. Another problem is to find out if the individual is sick at time $t$ given the values of their vital sign up to time $t$. People are also interested in predicting $X_{t+1}$ given past observations $X_1,\dots,X_t$. Those two problems are referred to as filtering and predicting in machine learning.

We can generalize this example to any number of hidden states. Hidden Markov models (HMMs) were formally introduced for the first time by Baum & Petrie [14] in 1966.

**Definition 1.3.** *We say that the pair $(X_t,H_t)_{t\geq 1}$ is a HMM if:*

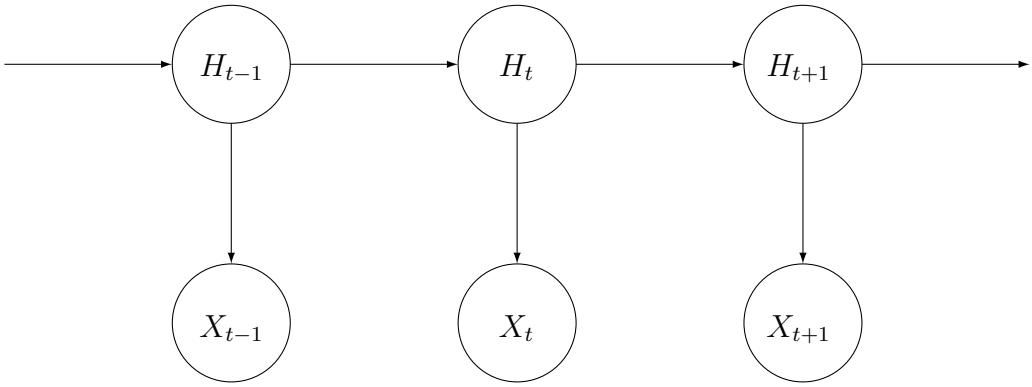
- *$(H_t)_{t\geq 1}$ is a Markov chain,*
- *conditionally on $(H_t)_{t\geq 1}$ the variables $(X_t)_{t\geq 1}$ are independent,*
- *and the distribution of $X_t$ only depends on $H_t$ for each $t \geq 1$.*

*In that case, we shall say that $(X_t)_t$ is generated by a HMM.*

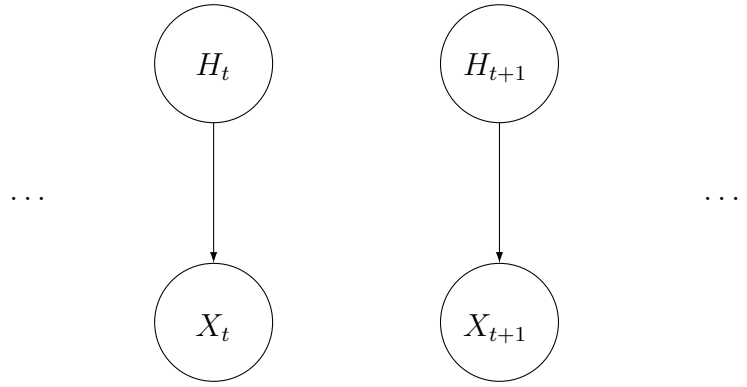
If we denote by $\mathcal{L}(Y)$ the probability distribution of a random variable $Y$, we can write the last two points of the definition as

$$\mathcal{L}\left((X_t)_{t\geq 1}|(H_t)_{t\geq 1}\right) = \bigotimes_{t\geq 1} \mathcal{L}(X_t|H_t). \tag{1.10}$$

HMMs are often represented by dependence graphs of the following form.




Although the process $(H_t,X_t)_{t\geq 1}$ is a Markov chain it is not the case for $(X_t)_{t\geq 1}$ in general. In comparison, mixture models are usually represented by the following graph.




As seen in Example 1.3, homogeneous hidden Markov models can be fully described with the state space $\mathscr{H}$, the Markov kernel $Q$ (defined below) and initial distribution $\pi$ of the hidden Markov chain $(H_t)_{t\geq 1}$, and the conditional distributions $(F_h)_{h\in\mathscr{H}}$ of $X_t$ given $H_t$. The distributions $(F_h)_{h\in\mathscr{H}}$ are called *emission distributions*.

**Definition 1.4.** *A Markov kernel on a measurable space $(\mathscr{H},\mathcal{H})$ is an application $Q : \mathscr{H}\times\mathcal{H}$ such that:*

- *for all $h$ in $\mathscr{H}$, $Q(h,\cdot)$ is a probability measure on $(\mathscr{H},\mathcal{H})$;*
- *for all $A$ in $\mathcal{H}$, the application $h \mapsto Q(x,A)$ is measurable.*

One can see that mixture models are a special case of hidden Markov models. It corresponds to the specific situation where there is a distribution $w$ on $(\mathscr{H},\mathcal{H})$ such that $Q(h,\cdot) = w$ for all $h$ in $\mathscr{H}$. It implies that the variables $H_1,H_2,\ldots,H_n$ are i.i.d. with common distribution $w$.

Finite state space HMMs correspond to the case where $\mathscr{H}$ is finite. Based on (1.10) and (1.9) we can see that $\boldsymbol{\mathscr{M}}$ should contain finite mixture distributions on $(\boldsymbol{\mathscr{X}},\boldsymbol{\mathcal{X}})$ with specific conditions on the weights and the emission distributions. We can always identify the state space $\mathscr{H}$ with $\{1,2,\dots,K\}$ and identify the Markov kernel with the transition matrix $(Q_{h,h'})_{h,h'\in\mathscr{H}} := (\mathbb{P}(H_{t+1}=h'|H_t=h))_{h,h'\in\mathscr{H}}$. If $\tau:\{1,2,\dots,K\}\to\mathscr{H}$ is a one-to-one map, the distribution of $(X_t)_{t\geq 1}$ can equally be described by the initial distribution $\pi'=(\pi_{\tau(k)})_{1\leq k\leq K}$, the transition matrix $Q'=(Q_{\tau(k_1),\tau(k_2)})_{1\leq k_1,k_2\leq K}$ and the emission distributions $F'=(F_{\tau(k)})_{1\leq k\leq K}$. Therefore we shall say that $(X_t)_{t\geq 1}$ is generated by a hidden Markov models with parameters $(K,Q',\pi',F')$.

**Definition 1.5.** *Let $\mathscr{F}\subset\mathscr{P}_X$ be a class of probability distributions on $(\mathscr{X},\mathcal{X})$. We define the* order *of $(X_t)_t$, with respect to $\mathscr{F}$, as the minimal value of $K$ such that $(X_t)_t$ is generated by a HMM with parameters $(K,Q,\pi,F)$ with $F_1,\dots,F_K$ in $\mathscr{F}$.*

One should note that, contrary to mixtures, we cannot arbitrarily merge two components (or states in the context of HMMs) so that restricting emission distributions to a specific class of probabilities is not necessary. For instance in Example 1.1 we can say that the distribution $P$ in (1.2) is a mixture with one component and emission distribution $(1-w)F_0+wF_1$ however we cannot write $(X_t,H_t)_t$ as a HMM with one unique state in Example 1.3 in general. It is only possible if $F_0=F_1$ or if the Markov chain $(H_t)_t$ has an absorbing state $a$ which has initial distribution 1, i.e. if $\mathbb{P}(H_{t+1}=a|H_t=a)=1=\mathbb{P}(H_1=a)$. Abraham *et al.* [2] note that those cases correspond to $X_1,X_1,\dots,X_n$ being independent random variables. They investigate the possibility to learn the parameters in those limit cases.

As for mixtures, it is easy to see the parameters are at best identifiable up to a permutation on the state space$\{1,\dots,K\}$. Gassiat *et al.* [43] and Alexandrovich *et al.* [3] provide general results showing that the parameters are identifiable from the distribution of consecutive observations. Let $\mathbb{P}^{(L)}_{(K,Q,\pi,F)}$ denote the distribution of $(Y_1,\dots,Y_L)$ when $(Y_t)_{t\geq 1}$ is generated by a stationary HMM with parameters $(K,\pi,Q,F)$, i.e.

$$\mathbb{P}^{(L)}_{(K,Q,\pi,F)} = \sum_{1\leq k_1,\dots,k_L\leq K} \pi_{k_1}Q_{k_1,k_2}\dots Q_{k_{L-1},k_L}\bigotimes_{l=1}^{L}F_{k_l}, \tag{1.11}$$

where $\pi$ is stationary with respect to $Q$. Under some assumptions, the equality $\mathbb{P}^{(L)}_{(K,Q,\pi,F)} = \mathbb{P}^{(L)}_{(K,Q',\pi',F')}$ implies $(Q',\pi',F')=(Q,\pi,F)$ up to a permutation on $\{1,2,\dots,K\}$. Taking $L\geq 3$ is enough under the condition that the emission distributions $F_1,\dots,F_K$ are linearly independent. If the emission distributions are only distinct, taking $L=2K+1$ is sufficient to have identifiability.

We refer to the book of Cappé *et al.* [55] for an exhaustive review of the topic of statistical inference for finite state space HMMs, particularly for parametric models. As for mixtures, maximum likelihood and Bayesian approaches are the most popular ones. For a different approach, we can mention Anandkumar *et al.* [6] that proposes a method of moments to estimate the means of the emission distributions. Standard results of consistency and asymptotic normality for the MLE are given in [55], under the assumption that the order is known. Gassiat [44] and Gassiat & Boucheron [42] give results of consistency for penalized maximum likelihood estimators of the order. We refer to Lehéricy [64] for a recent state of the art in finite state space HMMs.

Some recent papers adopted a slightly different approach more adapted to nonparametric estimation which is somewhat similar to the one of Gadat *et al.* [40] for mixture models given by (1.4). Following the identifiability results mentioned above, one might believe that the parameters $(Q,\pi,F)$ and $(Q',\pi',F')$ should be close to each other if the associated distributions $\mathbb{P}^{(L)}_{(K,Q,\pi,F)}$ and $\mathbb{P}^{(L)}_{(K,Q',\pi',F')}$ given by (1.11) are close to each other. Although the reverse is quite

easy to prove, see Proposition 3.6, this implication has only been proved recently. De Castro *et al.* [25] show this is true for the emission densities when the initial distributions and transition matrices are the same, i.e. $\pi' = \pi$ and $Q' = Q$. Lehéricy [63, 65] gets rid of this limitation and obtain results of the following nature. If $F_1, \dots, F_K$ have densities $f_1, \dots, f_K$ with respect to a reference measure $\nu$, we define the associated density

$$p^{(L)}_{(K,Q,\pi,\mathbf{f})}(x_1, \dots, x_L) = \sum_{1 \leq k_1, \dots, k_L \leq K} \pi_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} f_{k_1}(x_1) \dots f_{k_L}(x_L),$$

with respect to $\mu = \nu^{\otimes L}$. For square integrable emission densities and under some technical assumptions, for $(Q,\pi,\mathbf{f})$ in a neighborhood of $(Q^*,\pi^*,\mathbf{f}^*)$ we have

$$d^2\left((Q,\pi,\mathbf{f}),(Q^*,\pi^*,\mathbf{f}^*)\right) \leq c(Q^*,\pi^*,\mathbf{f}^*) \left|\left|p^{(3)}_{(K^*,Q,\pi,\mathbf{f})} - p^{(3)}_{(K^*,Q^*,\pi^*,\mathbf{f}^*)}\right|\right|_2^2, \tag{1.12}$$

where $c(Q^*,\pi^*,\mathbf{f}^*)$ is a positive constant that depends on the parameters $Q^*$, $\pi^*$ and $\mathbf{f}^*$,

$$d^2\left((Q,\pi,\mathbf{f}),(Q^*,\pi^*,\mathbf{f}^*)\right) = \inf_{\tau \in \mathcal{S}_{K^*}} \left\{ \sum_{k=1}^{K^*} (\pi_{\tau(k)} - \pi^*_k)^2 + \sum_{1 \leq k_1,k_2 \leq K^*} (Q_{\tau(k_1),\tau(k_2)} - Q^*_{k_1,k_2})^2 + \sum_{k=1}^{K^*} \left|\left|f_{\tau(k)} - f^*_k\right|\right|^2_{L_2(\nu)} \right\},$$

and $\mathcal{S}_{K^*}$ is the set of all permutations on $\{1,2,\dots,K^*\}$. Therefore, it is possible to deduce deviation bounds for the parameter estimators from a deviation bound on an estimator of $p^* = p^{(3)}_{(K^*,Q^*,\pi^*,\mathbf{f}^*)}$ when the order $K^*$ is known. Lehéricy [63] proposes a penalized least squares estimator that consistently estimates the order. They prove an oracle inequality with respect to the $L_2$-loss for the estimation of $p^*$ and deviation bounds for the parameters, conditioned on the event where the order estimator is exact. They obtain convergence rates for their estimators that are minimax up to logarithmic factors, and adaptive to the regularity of the true densities $f^*_1, \dots, f^*_{K^*}$. It is only adaptive to the worst regularity of the different densities however this limitation is lifted by Lehéricy [65] which achieves state-by-state adaptivity estimation of the emission densities given the transition matrix $Q^*$ and the distribution $\pi^*$, or minimax estimators of them. Abraham *et al.* [1] consider the estimation of the parameters for nonparametric hidden Markov models with two states. They show that estimating the smoother emission density first can improve the estimation of the second emission density.

The same way finite mixtures can be generalized to more complex models, it can be interesting to consider HMMs with state spaces that are not necessarily finite. We can illustrate it with the following example based on Example 1.2.

**Example 1.4.** *Let $(\Theta,\mathcal{T})$ be a measurable space and $\mathcal{F} = \{F_\theta; \theta \in \Theta\}$ be a parametric family of distributions on $(\mathcal{X},\mathcal{X})$ such that $\theta \mapsto F_\theta(A)$ is measurable for all $A \in \mathcal{X}$. Each individual has a health status in $\Theta$. We focus on one individual and denote by $X_i$ and $H_i$ their average vital sign and health status on day $i$. Given $H_i = \theta$, $X_i$ is distributed according to the distribution $F_\theta$. We assume $(H_i)_i$ is a homogeneous Markov chain with initial distribution $w$ and Markov kernel $Q$. In that case, the distribution $P$ of $(X_1, \dots, X_n)$ is given by*

$$P(A_1, \dots, A_n) = \int_{\Theta^n} F_{h_1}(A_1) \dots F_{h_n}(A_n) w(dh_1) Q(h_1,dh_2) \dots Q(h_{n-1},dh_n),$$

*for all $A_1, \dots, A_n \in \mathcal{X}$.*

As for general mixtures, most of the existing literature on the subject comes from signal processing and focuses on the following situation. Observations $X_1, \dots, X_n$ are assumed to be generated by the model

$$X_i = H_i + \epsilon_i, \tag{1.13}$$

where the signal $(H_i)_i$ is a Markov chain and the noise $(\epsilon_i)_i$ is i.i.d. random variables, independent from $(H_i)_i$. This corresponds to a translation hidden Markov model. If $(H_i)_i$ is a stationary Markov chain with initial distribution $\pi$ and Markov kernel $Q$, and $\Phi$ is the distribution of $\epsilon_i$, the distribution $\mathbb{P}^{(L)}_{(\pi,Q,\Phi)}$ of $(X_1,\dots,X_L)$ is given by

$$\mathbb{P}^{(L)}_{(\pi,Q,\Phi)}(A_1,\dots,A_L) = \int \Phi(A_1-h_1)\dots\Phi(A_L-h_L)\pi(dh_1)Q(h_1,dh_2)\dots Q(h_{L-1},dh_L),$$

for all measurable sets $A_1,\dots,A_L$ with $A-h=\{a-h; a\in A\}$. Even though models such as (1.13) have been used a lot for applications, there are very few theoretical results, especially if the distribution of the noise is unknown. Gassiat *et al.* [46] prove that the parameters are identifiable from $\mathbb{P}^{(2)}_{(\pi,Q,\Phi)}$ in a fully nonparametric setting. They also prove consistency results for a least squares estimator and a maximum likelihood estimator of the associated density.

Cases of general HMMs that do not fit within the framework of (1.13) have been investigated by Douc & Matias [31] and Douc *et al.* [32]. In [31], they consider a parametric setting for the emission densities and the Markov kernels and prove the convergence in probability of the maximum likelihood parameter estimator. They make assumptions that are equivalent to the standard assumptions for finite state space HMMs however they do not require the true initial distribution to be stationary but only the Markov kernel to be ergodic. Some assumptions are relaxed in [32] which contains similar results.

## 1.3 Diffusion processes

Stochastic differential equations are used to model a wide variety of processes with behaviors apparently random. Applications are numerous in many different fields and inferring features of the models used in practice is of great interest. A common framework is to assume that the observations come from the stationary solution of a stochastic differential equation (SDE) of the form

$$dY_t = b(Y_t)dt + a(Y_t)dB_t, \tag{1.14}$$

where $(B_t)_{t\geq 0}$ is a $d$-dimensional Brownian motion and the functions $b:\mathbb{R}^d\to\mathbb{R}^d$, $a:\mathbb{R}^d\to\mathbb{R}^{d\times d}$ are such that everything is well-defined. The features to estimate can be the function $a$, called the *volatility* or *diffusion coefficient*, the function $b$, called the *drift*, and the invariant distribution associated with equation (1.14).

We mention some standard references on the subject of estimation for diffusion processes to give a quick overview of the topic. Kessler [57] considers a parametric model for one dimensional ergodic diffusion. They prove the asymptotic normality of their parameter estimator, obtained through the minimization of a contrast function, based on discrete observations $Y_{t_1},\dots,Y_{t_n}$, $0\leq t_1<\dots<t_{n-1}<t_n<\infty$ of a stationary solution $(Y_t)_t$ of (1.14). Comte *et al.* [22] consider a similar situation but in the nonparametric context. They prove non-asymptotic deviation bounds for their estimators of the drift and the volatility based on least squares. They also show adaptivity properties for a penalized criterion. The most common approach when it comes to diffusion processes is based on kernels, using regularity properties of such processes. For instance, Dalalyan & Reiß [24] consider a nonparametric framework with constant volatility for a continuous multidimensional observation $(Y_t)_{0\leq t\leq T}\in(\mathbb{R}^d)^{[0,T]}$ and use kernel estimators for the drift function and the invariant density. They obtain asymptotic rates with respect to the $L_2$-loss when the regularity of the drift function is known.

We can note the following similarity with hidden Markov models when considering discrete observations of a stationary solution. In some cases, the features of the considered diffusion model are not accessible through the stationary distribution of the process. For instance, this is the case in Nickl [75] and Hoffmann & Ray [52] which deduce the feature of interest from the distribution of consecutive observations.

## 1.4 Contribution

This thesis proposes a generic approach to the **General Problem** with applications to mixture models, hidden Markov models and diffusion processes. In particular, we want to provide results for those models under the weakest assumptions possible on the true distribution of the observations as they can never be checked in practice. We provide a list of standard assumptions that people usually make and that are problematic/unrealistic/too restrictive. This is illustrated with examples.

Before we do so, we introduce the Hellinger distance $h$ between probability distributions defined as follows. For two probability distributions $P$ and $Q$ on a measurable space $(\mathscr{X},\mathcal{X})$,

$$h(P,Q) = \sqrt{\frac{1}{2}\int_{\mathscr{X}} \left(\sqrt{\frac{dP}{d\mu}(x)} - \sqrt{\frac{dQ}{d\mu}(x)}\right)^2 \mu(dx)} \in [0,1], \tag{1.15}$$

where $\mu$ is any positive measure dominating both $P$ and $Q$, the result being independent of the choice of $\mu$.

- It is common to assume that the observations $X_1,\dots,X_n$ are i.i.d. with common distribution $\overline{P}$ and in addition to assume this distribution belongs to the model. This assumption is quite important for classical estimators such as the MLE as shown by the following example.

  **Example 1.5.** *The model contains mixtures of two uniform distributions on segments of length* 1*, i.e. the distributions of the form*

  $$P_{w,a_1,a_2} = w\mathcal{U}(a_1,a_1+1) + (1-w)\mathcal{U}(a_2,a_2+1), w \in (0,1), a_1 \neq a_2 \in \mathbb{R}.$$

  *Now assume that the true distribution is given by*

  $$\overline{P} = \frac{1}{2}\mathcal{U}(.45 - n^{-1},1.45 + n^{-1}) + \mathcal{U}(3.1 - n^{-1},4.1 + n^{-1}). \tag{1.16}$$

  *The model is a good approximation of the true distribution as* $h^2(\overline{P}, P_{1/2,.45,3.1}) \leq 2/n$*. However, the likelihood is null for any distribution in the model as soon as there are observations* $X_{i_1},X_{i_2},X_{i_3}$ *such that* $X_{i_1}+1 < X_{i_2} < X_{i_3} - 1$*. In particular, it is implied by the event*

  $$\left\{\exists i_1,i_2,i_3, X_{i_1} \in [.45 - n^{-1},.45), X_{i_2} \in (1.45,1.45 + n^{-1}] \text{ and } X_{i_3} \in [3.1 - n^{-1},4.1 + n^{-1}]\right\},$$

  *which has probability at least* $1 - e^{-3/8}(2 - e^{-4/10}) - (1/2)^6 > 0.07$ *for* $n \geq 6$*. Therefore, there is at least a* 7% *chance that the MLE is not defined.*

  We see that even if the true distribution is very close to the model but not in it, the maximum likelihood approach fails.

- It seems normal to put conditions on the model that the statistician chooses, usually because of the loss or the estimation method considered. On the other hand, it is quite restrictive to also put those conditions on the true distributions. For example, it is very common to assume that the true distribution admits a density with respect to some reference measure, usually the Lebesgue measure. In particular, when people use a least squares approach or when they simply consider the $L_2$-loss, they assume this density to be square integrable. For instance, this is the case of Lehéricy [63] which assumes the true distribution is a stationary hidden Markov model with emission densities that are square integrable and uniformly bounded. If one makes those assumptions they ban some models from their framework such as the following example.

**Example 1.6.** *Consider the location mixture model containing distributions of the form*

$$(1-w)\mathbb{S}_{\alpha,0} + w\mathbb{S}_{\alpha,z}, \tag{1.17}$$

*where* $\alpha \in (0,1)$ *and* $\mathbb{S}_{\alpha,z}$ *is the distribution defined by the density* $s_{\alpha,z}$ *with respect to the Lebesgue measure given by*

$$s_{\alpha,z} : x \in \mathbb{R} \mapsto \frac{1-\alpha}{2|x-z|^{\alpha}} \mathbb{1}_{|x-z|\in(0,1]}. \tag{1.18}$$

One can notice that such densities are unbounded, and they are not square integrable when $\alpha \geq 1/2$. This implies that both the maximum likelihood and the least squares approaches fail to apply here. We want to propose an approach that would have more flexibility and allow us to consider such models. To do so, it is natural to work with a loss that does not require any assumption on the true distribution such as the $L_1$-loss (equivalent to the total variation distance) or the Hellinger loss.

- Another restriction of standard frameworks is to assume that the observations are equally distributed. This can easily be wrong in practice. Following Example 1.1, it may happen that some of the values of the vital sign were erroneously reported. It might also happen that a measuring instrument is defective such that the observations gathered by this instrument would not be distributed according to the same distribution as the values measured by the other instruments. If the proportion of corrupted data is small enough it should still be possible to estimate the distribution of interest. Similarly, it is quite common to assume that the observation process is stationary in the case of finite state space HMMs (see Lehéricy [63] or Abraham *et al.* [1]) but it appears restrictive and does not take into account the possibility to have corrupted observations.

Those different points motivate us to work with a statistical framework that is slightly different from the one we introduced at the beginning of this chapter. We build upon the work of Baraud *et al.* [9] and Baraud & Birgé [11] that developed $\rho$-estimators to solve some of the problems we just raised among others. However, they did not consider departures from the assumption that the observations are independent. We provide a way to obtain theoretical guarantees for $\rho$-estimators without the independence assumption. Assuming that the observations are independent is the same as assuming that the joint distribution

$$\mathbf{P}^* = \mathcal{L}(X_1, \dots, X_n)$$

is equal to the product of the marginal distributions

$$\mathbf{P}^{ind} = \mathcal{L}(X_1) \otimes \cdots \otimes \mathcal{L}(X_n) = P_1 \otimes \cdots \otimes P_n. \tag{1.19}$$

We quantify the dependence within the observations through the Kullback-Leibler divergence $\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)$ of $\mathbf{P}^*$ from $\mathbf{P}^{ind}$, defined as follows. For two probability distributions $P$ and $Q$ on the same measurable space $(\mathscr{A},\mathcal{A})$, the Kullback-Leibler divergence of $P$ from $Q$ is given by

$$\mathbf{K}(P||Q) = \begin{cases} \int_{\mathscr{A}} \log\left(\frac{dP}{dQ}(x)\right) P(dx) \text{ if } P \ll Q, \\ +\infty \text{ otherwise.} \end{cases} \tag{1.20}$$

We show that in general we can do as if the observations were independent when $\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)$ is small enough (see Section 1.4.4). We illustrate that with the following example.

**Example 1.7.** *The observations $X_1,\dots,X_n$ are assumed to be i.i.d. with common distribution the univariate normal distribution $\mathcal{N}(z,1)$ and we try to estimate the location parameter $z$ however the random variable $\mathbf{X}=(X_1,\dots,X_n)$ actually follows a multivariate normal distribution $\mathcal{N}(\mathbf{z},\Sigma_\varepsilon)$ where*

$$\mathbf{z}=(z,\dots,z)\in\mathbb{R}^n \text{ and } \Sigma_\varepsilon=\begin{pmatrix} 1 & \varepsilon & \varepsilon^2 & \dots & \varepsilon^{n-1} \\ \varepsilon & \ddots & \ddots & \ddots & \vdots \\ \varepsilon^2 & \ddots & \ddots & \ddots & \varepsilon^2 \\ \vdots & \ddots & \ddots & \ddots & \varepsilon \\ \varepsilon^{n-1} & \dots & \varepsilon^2 & \varepsilon & 1 \end{pmatrix}\in\mathbb{R}^{n\times n}, \tag{1.21}$$

*where $z\in\mathbb{R}$ and $\varepsilon\in(-1,1)$. The assumption that the observations are independent only allows us to consider the case $\varepsilon=0$. It might happen that results obtained under this assumption completely fall apart even for very small values of $\varepsilon$. In this example, we have $\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)=\frac{n-1}{2}\ln\left(\frac{1}{1-\varepsilon^2}\right)$ and the Theorem 1.4 stated later shows that the performance of the $\rho$-estimator is not significantly worse for $\epsilon$ of order at most $n^{-1/2}$.*

We can use this flexibility with respect to the independence assumption to consider the estimation of the stationary distribution of processes such as hidden Markov models or diffusion processes for instance. But there is no obvious reason for the dependence $\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)$ to be small in general. We propose a way to obtain a smaller dependence for mixing processes by selecting a subset of the observations. Intuitively, if $(X_k)_{k\in\mathbb{Z}}$ is a mixing process the variables $X_{k_1}$ and $X_{k_2}$ are "almost independent" for $|k_1-k_2|$ "large enough". We refer to Bradley [19] for a review of the different notions of mixing. Based on this idea, for an integer $s$ we build the subset $\mathbf{X}^{(s)}$ of observations as follows

$$\mathbf{X}^{(s)}=\left(X_1,X_{2+s},X_{1+2(s+1)},\dots,X_{1+n(s)(s+1)}\right), \tag{1.22}$$

with $n(s)=\lfloor(n-1)/(s+1)\rfloor$. It means we take observations separated by blocks of $s$ consecutive observations. A large value of $s$ gives a smaller dependence term but it also makes the set of observations used for the estimation smaller. We show that this strategy is efficient for HMMs and some diffusion processes. The key point is that those processes satisfy strong mixing properties.

### 1.4.1 Framework

Let $\mathbf{X}=(X_1,X_2,\dots,X_n)$ be a random variable on $(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{X}})=(\mathcal{X}^{\otimes n},\mathcal{X}^{\otimes n})$, where $(\mathcal{X},\mathcal{X})$ is a measurable space. We denote by $\mathbf{P}^*$ the distribution of $\mathbf{X}$ and by $P_i=\mathcal{L}(X_i)$ the marginal distribution of $X_i$ for $i\in\{1,\dots,n\}$. We do as if the observations were independent and identically distributed with common distribution $\overline{P}$, i.e. as if the distribution $\mathbf{P}^*$ were of the form $\overline{P}^n$. Therefore we take models $\boldsymbol{\mathscr{M}}$ of the form given by (1.1). As mentioned earlier, we measure how far are the observations from being independent through the Kullback-Leibler divergence of $\mathbf{P}^*$ from the product distribution of the marginals $\mathbf{P}^{ind}$, given by (1.19). For an estimator $\hat{P}\in\mathscr{M}$ of $\overline{P}$, we measure its accuracy with the Hellinger-type loss

$$\frac{1}{n}\sum_{i=1}^n h^2\left(P_i,\hat{P}\right),$$

where $h$ denotes the Hellinger distance defined by (1.15). In the ideal situation where the observations are identically distributed with common distribution $\overline{P}$, this loss becomes $h^2(\overline{P},\hat{P})$.

### 1.4.2 Results of $\rho$-estimation

The $\rho$-estimators developed by Baraud *et al.* [9, 11] are based on robust tests and can be seen as a refinement of $T$-estimation developed by Birgé [16]. We illustrate the performance of $\rho$-estimators in the context of the **General Problem** under the independence assumption. Let $\mathscr{M}$ be a class of probability distribution on $(\mathscr{X},\mathcal{X})$ and $\boldsymbol{\mathscr{M}}$ be the associated model on $(\boldsymbol{\mathscr{X}},\boldsymbol{\mathcal{X}})$ given by (1.1).

**Theorem 1.1.** *(Theorem 1 [11])*
*For independent random variables $X_1,\ldots,X_n$ with arbitrary distributions $P_1,\ldots,P_n$, the $\rho$-estimator $\hat{P}=\hat{P}(\mathbf{X},\boldsymbol{\mathscr{M}})\in\mathscr{M}$ satisfies*

$$C\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}h^2\left(\hat{P},P_i\right)\right]\leq\inf_{Q\in\mathscr{M}}\frac{1}{n}\sum_{i=1}^{n}h^2\left(Q,P_i\right)+\frac{D(\boldsymbol{\mathscr{M}})}{n}, \tag{1.23}$$

*where $C$ is a positive universal constant and $D(\boldsymbol{\mathscr{M}})\geq 1$ is a bound on the $\rho$-dimension of the model $\boldsymbol{\mathscr{M}}$ which depends on $n$.*

We mention that there is also a similar result of model selection for $\rho$-estimator (see Theorem 2 [11]). The $\rho$-dimension function is formally introduced in [11]. Let us explain the result above and illustrate the different properties of $\rho$-estimators. The upper bound in (1.23) is composed of two terms, the approximation or bias term $\inf_{Q\in\mathscr{M}}n^{-1}\sum_{i=1}^{n}h^2\left(Q,P_i\right)$ and the dimension term $D(\boldsymbol{\mathscr{M}})/n$. In the simple situation where the observations are actually i.i.d. with distribution $\overline{P}$ in $\mathscr{M}$, the approximation term vanishes and we have

$$C\mathbb{E}\left[h^2\left(\hat{P},\overline{P}\right)\right]\leq\frac{D(\boldsymbol{\mathscr{M}})}{n}.$$

One can see that the dimension term determines an upper bound on the convergence rate of the estimator over $\mathscr{M}$. In most of the applications of $\rho$-estimation, the rate $D(\boldsymbol{\mathscr{M}})/n$ is optimal up to a logarithmic factor.

The $\rho$-dimension can be related to more common notions of dimension such as the VC-dimension that we briefly introduce in Section 1.5, with a few results and references on the subject. Let $\mathcal{M}$ be a class of density functions associated with $\mathscr{M}$, with respect to a $\sigma$-finite positive measure on $(\mathscr{X},\mathcal{X})$. If $\mathcal{M}$ is VC-subgraph with VC-index $V(\mathscr{M})$, we can take

$$D(\boldsymbol{\mathscr{M}})=CV(\mathscr{M})\log n, \tag{1.24}$$

where $\boldsymbol{\mathscr{M}}$ is the model given by (1.1) and $C$ is a universal constant.

As we said earlier, assuming that the observations are i.i.d. with common distribution $\overline{P}$ in $\mathscr{M}$ could be debatable. The bias or approximation term measures how far is the model from the truth. It accounts for the robustness of the estimator $\hat{P}$. We can illustrate it considering specific types of departures from this assumption.

- If the observations are i.i.d. with common distribution $\overline{P}$ that does not belong to $\mathscr{M}$, from (1.23) we get

$$C\mathbb{E}\left[h^2\left(\hat{P},\overline{P}\right)\right]\leq h^2\left(\overline{P},\mathscr{M}\right)+\frac{D(\boldsymbol{\mathscr{M}})}{n},$$

where the notation $h^2(\overline{P},\mathscr{M})$ is defined as follows. For a distribution $P\in\mathscr{P}_X$ and a class of distributions $\mathscr{Q}\subset\mathscr{P}_X$, we write

$$h\left(P,\mathscr{F}\right)=\inf_{Q\in\mathscr{Q}}h(P,Q).$$

We can see that the estimator is robust to misspecification, i.e. the deviation bound is not significantly worse as long as the distance $h\left(\overline{P},\mathscr{M}\right)$ is of order not larger than $\sqrt{D_n(\mathscr{M})/n}$.

• We can also consider the situation where the observations are only obtained after an alteration of a sample from a distribution $\overline{P}$ in $\mathscr{M}$, that we model as follows. The observations $X_1,\dots,X_n$ are given by

$$X_i = E_i\overline{X}_i + (1-E_i)\underline{X}_i, \tag{1.25}$$

where the variables $\overline{X} = (\overline{X}_i)_{1\le i\le n}$ are i.i.d. with common distribution $\overline{P}$, the $(E_i)_i$ are independent Bernoulli variables and $(\underline{X}_i)_i$ are independent random variables with arbitrary distributions. For $i$ in $\{1,2,\dots,n\}$, the probability of observing $\underline{X}_i$ instead of $\overline{X}_i$ is $q_i = \mathbb{P}(E_i = 0)$. This includes the Hüber $\varepsilon$-contamination model (see [53]) where observations are i.i.d. with common distribution $P$ of the form

$$P = (1-\varepsilon)\overline{P} + \varepsilon Q. \tag{1.26}$$

It corresponds to the case where $\mathbb{P}(E_i = 0) = \varepsilon$ and $\mathcal{L}(\underline{X}_i) = Q$ for all $i \in \{1,\dots,n\}$. It also includes adversarial contamination inspired by adversarial machine learning. Before the statistician is given access to an i.i.d. sample $\overline{X}_1,\dots,\overline{X}_n$, an adversary can select any subset $(\overline{X}_i)_{i\in I}$ of observations and replace them with any arbitrary values $(\underline{X}_i)_{i\in I}$. In this thesis we will rather not use the term adversarial contamination and rather say that the observations $(\underline{X}_i)_{i\in I}$ are *outliers*. We can be even more general and not assume that $\overline{P}$ is necessarily in $\mathscr{M}$ in the situation described by (1.25). In that case we get

$$C'\mathbb{E}\left[h^2\left(\hat{P},\overline{P}\right)\right] \le h^2\left(\overline{P},\mathscr{M}\right) + \frac{1}{n}\sum_{i=1}^{n} q_i + \frac{D(\mathscr{M})}{n},$$

where $C' = C/(1+C)$ is a universal constant and $C$ comes from (1.23). Basically we have split the approximation term in two terms using the convexity of the squared Hellinger distance. In this thesis, we repeatedly use the inequality

$$h^2(P,\lambda Q_1 + (1-\lambda)Q_2) \le \lambda h^2(P,Q_1) + (1-\lambda)h^2(P,Q_2), \tag{1.27}$$

for all distributions $P$, $Q_1$ and $Q_2$ on $(\mathscr{X},\mathcal{X})$ and all $\lambda \in [0,1]$. The first term $h^2\left(\overline{P},\mathscr{M}\right)$ accounts for the misspecification, it quantifies how far is the model from the distribution of interest. The second term $n^{-1}\sum_{i=1}^{n} q_i$ accounts for the alteration of the data, it quantifies how far are the observations from being an i.i.d. sample from the distribution $\overline{P}$. On average, the number of contaminated observations is $q_1+\dots+q_n$, which corresponds to $n\varepsilon$ (respectively $|I|$) in the case of contamination (respectively outliers). As long as those terms are of order not larger than $D(\mathscr{M})/n$, the upper bound for the risk of the estimator is not significantly worse. Therefore we shall say that the estimator $\hat{P}$ is robust to misspecification and to alteration of the data.

An estimator being robust to misspecification is interesting in itself but also because it allows us to consider approximation models. This approach is quite popular in nonparametric estimation when a class of densities or regression functions is approximated by a simpler class of functions. It is the case in the situation we described earlier with (1.5) for instance. In our framework, we will consider nets with respect to the Hellinger distance.

**Definition 1.6.** *We say that $\mathscr{M} \subset \mathscr{P}_X$ is an $\eta$-net of $\overline{\mathscr{M}} \subset \mathscr{P}_X$ if, for all $P$ in $\overline{\mathscr{M}}$, there exists $Q$ in $\mathscr{M}$ such that $h(P,Q) \le \eta$.*

We can easily see with Theorem 1.1 what we would obtain with approximate models. Let $\mathscr{M}[\eta]$ be an $\eta$-net of $\overline{\mathscr{M}}$ with respect to the Hellinger distance. If $\hat{\mathscr{M}}[\eta]$ is the associated model

given by (1.1), the $\rho$-estimator $\hat{P} = \hat{P}(\mathbf{X},\mathscr{M}[\eta])$ satisfies

$$C\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} h^2\left(\hat{P},P_i\right)\right] \leq \inf_{Q\in\overline{\mathscr{M}}} \frac{1}{n}\sum_{i=1}^{n} h^2\left(Q,P_i\right) + \eta^2 + \frac{D(\mathscr{M}[\eta])}{n}. \tag{1.28}$$

In particular, for i.i.d. observations with distribution $\overline{P}$, we get

$$C\mathbb{E}\left[h^2\left(\hat{P},\overline{P}\right)\right] \leq h^2\left(\overline{P},\overline{\mathscr{M}}\right) + \eta^2 + \frac{D(\mathscr{M}[\eta])}{n}.$$

Therefore, we can obtain a uniform bound on the convergence rate over $\overline{\mathscr{M}}$ by balancing the terms $\eta^2$ and $D(\mathscr{M}[\eta])/n$.

### 1.4.3 Mixture models (Chapter 2)

In Chapter 2, we fill a gap in the literature providing non asymptotic guarantees in a very general framework, along with robustness results. For classes of distributions $\mathscr{F}_1,\ldots,\mathscr{F}_K \in \mathscr{P}_X$, we define the associated $K$-component mixture models

$$\mathscr{M}(K,\mathscr{F}_1,\ldots,\mathscr{F}_K) = \left\{\sum_{k=1}^{K} w_k F_k; w \in \mathcal{W}_K, F_k \in \mathscr{F}_k, \forall k \in [K]\right\}.$$

We will call the classes $\mathscr{F}_1,\ldots,\mathscr{F}_K$ *emission models*. We assume that the emission models are dominated by a $\sigma$-finite measure $\mu$, typically the Lebesgue measure, and there are classes of density functions $\mathcal{F}_1,\ldots,\mathcal{F}_K$ with respect to $\mu$ associated with $\mathscr{F}_1,\ldots,\mathscr{F}_K$. If we denote by $\mathscr{M}$ the model given by (1.1) with $\mathscr{M} \subset \mathscr{M}(K,\mathscr{F}_1,\ldots,\mathscr{F}_K)$, we prove a bound on the $\rho$-dimension that depends on the VC-dimension of the emission density models. We can take

$$D(\mathscr{M}) = C(V_1 + \cdots + V_K)\log n, \tag{1.29}$$

where $C$ is a universal constant and $V_K$ is the VC-dimension of $\mathcal{F}_k$. We can deduce the following deviation bound from this result and the general inequality (1.23).

**Theorem 1.2.** *(Theorem 2.1) For independent random variables $X_1,\ldots,X_n$ with arbitrary distributions $P_1,\ldots,P_n$, the $\rho$-estimator $\hat{P} = \hat{P}(\mathbf{X},\mathscr{M}) \in \mathscr{M}$ satisfies*

$$C\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} h^2\left(P_i,\hat{P}\right)\right] \leq \inf_{Q\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^{n} h^2\left(P_i,Q\right) + \frac{(V_1 + \cdots + V_K)\log n}{n}, \tag{1.30}$$

*where $C$ is a positive universal constant.*

To our knowledge, there is no similar result of robustness in a general framework for mixture models. We can obtain risk bounds and convergence rates for the considered models determining the VC-dimensions $V_1,\ldots,V_K$ of the corresponding emission models. For instance, if we consider Example 1.5 we can show that the VC-dimension of the class of uniform densities is 2, and therefore the $\rho$-estimator satisfies

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}\right)\right] \leq h^2\left(\overline{P},\mathscr{M}\right) + \frac{4\log n}{n} \leq \frac{2 + 4\log n}{n}.$$

The fact that the true distribution $\overline{P}$ does not belong to the model does not deteriorate the performance of the estimator, contrary to the MLE.

We consider the cases of multivariate location mixtures and multivariate location-scale mixtures of normal distributions. Let $\mathrm{Cov}_{+*}(d)$ denote the class of $d \times d$ symmetric and positive-definite matrices.

**Theorem 1.3.** *(Corollary 2.1)*
*We assume the observations $X_1,\dots,X_n$ are i.i.d. with common distribution $\overline{P}$.*

- *Let $\mathscr{M}_{ls}$ be the $d$-dimensional Gaussian location-scale mixture model with $K$ component, i.e.*
$$\mathscr{F}_1 = \cdots = \mathscr{F}_K = \left\{\mathcal{N}(\mu,\Sigma); \mu \in \mathbb{R}^d, \Sigma \in Cov_{+*}(d)\right\}.$$
*There is a positive universal constant $C > 0$ such that the $\rho$-estimator $\hat{P}$ on $\mathscr{M}_{ls}$ satisfies*
$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}\right)\right] \leq \frac{Kd^2\left[1+\log\left(\frac{n}{d^2}\vee K\right)\right]}{n},$$
*for all $\overline{P} \in \mathscr{M}_{ls}$.*

- *Let $\mathscr{M}_{loc}(\Sigma)$ be the $d$-dimensional Gaussian location mixture model associated to the covariance matrix $\Sigma \in Cov_{+*}(d)$, i.e.*
$$\mathscr{F}_1 = \cdots = \mathscr{F}_K = \left\{\mathcal{N}(\mu,\Sigma); \mu \in \mathbb{R}^d\right\}.$$
*There is a positive universal constant $C > 0$ such that the $\rho$-estimator $\hat{P}$ on $\mathscr{M}_{loc}(\Sigma)$ satisfies*
$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}\right)\right] \leq \frac{Kd\left[1+\log\left(\frac{n}{d}\vee K\right)\right]}{n},$$
*for all $\overline{P} \in \mathscr{M}_{loc}(\Sigma)$.*

Those rates are optimal up to a logarithmic factor. Doss *et al.* [30] obtain the optimal rate with no logarithmic factor for Gaussian location mixtures with known isotropic covariance matrix. However, the dependency in the number of components $K$ of their bound is worse than exponential when it is just linear for our estimator.

We also consider nonparametric settings for $s$-concave and log-concave emission densities. Let $\mathcal{C}$ be the class of concave functions $\mathbb{R} \to [-\infty,\infty)$. We say that a function $f : \mathscr{X} \to \mathbb{R}$ is $s$-concave if there exists $g$ in $\mathcal{C}$ such that
$$\begin{cases} f = g_+^{1/s} \text{ for } s > 0, \\ f = (-g)_+^{1/s} \text{ for } s \in (-1,0). \end{cases}$$
Similarly we say that such a function $f$ is log-concave if there exists $g$ in $\mathcal{C}$ such that
$$f = \exp g. \tag{1.31}$$
The class of log-concave densities includes many usual parametric densities such as Gaussian, exponential, logistic or Laplace densities. It is possible to use finite nets to approximate the class of log-concave (or $s$-concave) densities that are upper bounded by a uniform constant $M$. We follow the approach exposed earlier to obtain (1.28) and deduce a deviation bound for our estimator based on mixtures of log-concave (or $s$-concave) densities (see Corollary 2.2). Let $\mathscr{M}$ be the model of all mixtures with $K$ emission densities upper bounded by $M$ and $s$-concave, the case $s = 0$ corresponding to log-concave emission densities. There exists a constant $C(M,s)$ depending on $M$ and $s$ such that the $\rho$-estimator $\hat{P}$ on $\mathscr{M}$ satisfies
$$C(M,s)h^2\left(\overline{P},\hat{P}\right) \leq \frac{K\left[1+\log\left(Kn\right)\right]}{n^{4/5}},$$

where the observations are i.i.d. with common distribution $\overline{P}$ in $\mathscr{M}$. The uniform bound on the convergence rate we obtain is similar to the one obtained by Doss & Wellner [29] for a single component, i.e. in the case $K = 1$, up to a logarithmic factor.

We also consider the estimation of a general mixture of normal distributions. It corresponds to Example 1.2 in the case where the class $\mathscr{F}$ is the family of univariate normal distributions. We consider location-scale mixtures with mixing distribution supported on a compact set, i.e. distributions with a density with respect to the Lebesgue measure of the form

$$p_\eta(x) = \int_{\mathbb{R}\times(0,\infty)} \frac{e^{-(x-z)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \eta(d(z,\sigma^2)),$$

where $\eta$ is the mixing distribution and there are (finite) positive constants $M,\sigma_-,\sigma_+$ such that $\eta([-M,M]\times[\sigma_-^2,\sigma_+^2]) = 1$. Our estimator achieves the same rate as the one given by Ghosal & van der Vaart [49] with less assumptions on the true distribution.

We can also deduce results for the estimation of the parameters. Our aim is to obtain similar results as for the estimation of the mixture distribution, in particular we want non-asymptotic deviation bounds that also allows us to exhibit the robustness of the parameter estimators. Although identifiability is necessary to ensure that the problem of estimating the parameters is not ill posed, it is not enough to obtain non-asymptotic results. Our approach is to show that when two distributions in the model are close to each other then the associated parameters are also close to each other. Ideally, for a parametric model $\mathscr{M} = \{P_\theta,\theta\in\Theta\},\Theta\subset\mathbb{R}^d$, we would have an inequality of the form

$$C(\overline{\theta})\left(1\wedge||\theta-\overline{\theta}||_2^2\right) \leq h^2\left(P_\theta,P_{\overline{\theta}}\right),\forall\theta\in\Theta, \tag{1.32}$$

for each $\overline{\theta}\in\Theta$, where $C(\overline{\theta})$ is a positive constant that depends on $\overline{\theta}$. In that case, we can deduce the following result from Theorem 1.2 and the convexity inequality (1.27). For all $\overline{\theta}$ in $\Theta$, the $\rho$-estimator $\hat{P} = P_{\hat{\theta}}$ on $\mathscr{M}$ satisfies

$$C(\overline{\theta})\mathbb{E}\left[1\wedge||\overline{\theta}-\hat{\theta}||_2^2\right] \leq \frac{1}{n}\sum_{i=1}^n h^2\left(P_i,P_{\overline{\theta}}\right) + \frac{(V_1+\cdots+V_K)\log n}{n},$$

where $C(\overline{\theta})$ is a positive constant that depends on $\overline{\theta}$. We can deduce a convergence rate for the parameter estimators when the model is well specified. We can see that the parameter estimators are robust for similar reasons as for the distribution estimator, as long as the parametric model we consider satisfies an inequality similar to (1.32). Under this assumption in particular, we do not worry whether the model is exact or not before considering the estimation of the parameters. True parameters might not exist but we can always aim for the best approximation within the model, if the associated distribution is not too far from the true distribution.

We use the theory of Ibragimov & Has'minskiĭ [54] to prove inequalities similar to (1.32) for regular parametric emission models (see Theorem 2.4). We can also use existing results that relate to the $L_2$-distance between densities instead of the Hellinger distance between distributions when the considered densities are bounded. Let $P = p\cdot\mu$ and $Q = q\cdot\mu$ be two probability distributions with bounded densities $p$ and $q$ with respect to a positive measure $\mu$. We have the inequality

$$||p-q||_2^2 \leq 4\left(||p||_\infty+||q||_\infty\right)h^2\left(P,Q\right). \tag{1.33}$$

We consider the specific case of two-component location mixtures with one known location parameter given by (1.4) investigated by Gadat *et al.* [40]. We show that our method applies to the different location families they consider, i.e. location families based on a Cauchy, Gaussian, Laplace or skew Gaussian distribution. We obtain similar results with respect to the Hellinger distance with weaker assumptions.

We also consider parameter estimation for the mixture model

$$\mathscr{M} = \{P_{w,z} = (1-w)\mathbb{S}_{\alpha,0} + w\mathbb{S}_{\alpha,z}; w \in (0,1], z \in (-\infty,0) \cup (0,\infty)\},$$

defined earlier in Example 1.6. It does not fit in common frameworks as the emission densities are unbounded with a singularity. However, we can still obtain an interesting lower bound on the Hellinger distance. For all $w^* \in (0,1]$ and all $z^* \neq 0$, there is a positive constant $C(\alpha,w^*,z^*)$ such that for all $z \in \mathbb{R}$ and all $w \in [0,1]$, we have

$$h^2\left(P_{w^*,z^*}, P_{w,z}\right) \geq C(\alpha,z^*,w^*)\left[(w^*)^{1/\alpha}\left(1 \wedge |z-z^*|^{1-\alpha}\right) + (w^*-w)^2\left(1 \wedge |z^*|\right)\right].$$

This inequality allows us to deduce convergence rates for the estimation of $w^*$ and $z^*$ from Theorem 1.2 (see Theorem 2.7). We obtain the usual $1/\sqrt{n}$ parametric rate, with respect to the Euclidean distance, for the estimation of the weight $w^*$, up to a logarithmic factor. The singularity of the emission densities allows us to obtain a faster convergence rate for the location parameter $z^*$. Up to a logarithmic factor, this rate is of order $n^{-1/(1-\alpha)}$ which is minimax and faster than the parametric rate.

All the results mentioned so far were obtained by considering one mixture model with a fixed number of components $K$ and fixed emission models $\mathscr{F}_1,\ldots,\mathscr{F}_K$. We can lift this restriction using model selection. We provide a general result of model selection and focus on two cases: the selection of the order for a fixed emission model and the selection of the emission models for a fixed number of components.

The first situation is illustrated with the selection of the order of an estimator based on mixtures of univariate Gaussian distributions. We prove a uniform bound on the convergence rate of our estimator over a class of distributions with densities satisfying regularity and tail conditions. There exists a positive constant $C_{\underline{\beta},\overline{\beta}}$ such that if the observations are i.i.d. with distribution $P \in \mathscr{H}_\beta$ with $0 < \underline{\beta} \leq \beta \leq \overline{\beta}$, our estimator $\hat{P}$ satisfies

$$C_{\underline{\beta},\overline{\beta}}\mathbb{E}\left[h^2(P,\hat{P})\right] \leq \frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}},$$

where $\mathscr{H}_\beta$ is a class of distributions with associated densities having regularity index $\beta$ (see Theorem 2.11). We obtain the same rate as Maugis-Rabusseau & Michel [68] which is minimax up to a logarithmic factor and our estimator is adaptive to the regularity of the target density. However, we do not need to know bounds $\underline{\beta}$ and $\overline{\beta}$ on $\beta$ to construct our estimator which is the case in [68].

In the second situation, the number of components $K$ is fixed but the emission distributions can belong to different emission models. We consider an application for emission distributions that are either Gaussian or Cauchy, i.e. distributions of the form

$$P_{w,j,z,\sigma} = \sum_{i=1}^{j} w_i \mathcal{N}(z_i,\sigma_i^2) + \sum_{i=j+1}^{K} w_i \text{Cauchy}(z_i,\sigma_i). \tag{1.34}$$

We prove that, when the sample is large enough, with high probability we can identify the number of components corresponding to each type of distribution, i.e. the integer $j$ in the example above, and the parameter estimators satisfy the inequality

$$C(w,j,z,\sigma)\left(||\overline{w}-\hat{w}||^2 + \sum_{k=1}^{j^*}\left|\left|(\overline{z}_k,\overline{\sigma}_k^2) - (\hat{z}_k,\hat{\sigma}_k^2)\right|\right|^2 \wedge 1 + \sum_{k=j^*+1}^{K}||(\overline{z}_k,\overline{\sigma}_k) - (\hat{z}_k,\hat{\sigma}_k)||^2 \wedge 1\right) \leq \frac{K\log n}{n},$$

where $C(w,j,z,\sigma)$ is a positive constant depending on $P_{w,j,z,\sigma}$ (see Theorem 2.9). To our knowledge, this is the first result of this nature.

### 1.4.4 Dealing with dependent observations and applications (Chapter 3)

**Robustness to the independence assumption**

In Chapter 2 we make the assumption that observations are independent. However, in Chapter 3 we show that our estimator is robust to small deviations from this assumption. We have the following result for general models $\mathscr{M}$ of the form (1.1), extending Theorem 1.1.

**Theorem 1.4.** *Theorem 3.1*
*For (possibly dependent) random variables $X_1,\dots,X_n$ with respective distributions $P_1,\dots,P_n$, the $\rho$-estimator $\hat{P} = \hat{P}(\mathbf{X},\mathscr{M})$ satisfies*

$$C\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} h^2\left(\hat{P},P_i\right)\right] \leq \inf_{Q\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^{n} h^2\left(Q,P_i\right) + \frac{D(\mathscr{M})}{n} + \frac{\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)}{n}, \tag{1.35}$$

*where $C$ is a positive universal constant, $\mathbf{P}^*$ is the joint distribution $\mathcal{L}(X_1,\dots,X_n)$ and $\mathbf{P}^{ind}$ is the product of the marginal distributions given by (1.19).*

One should notice that this result is assumption-free. It is similar to (1.23) with an additional term that accounts for the dependence within the observations. We can see that this deviation bound is not significantly worse than the one we would have if the observations were independent, as long as the quantity $\mathbf{K}\left(\overline{\mathbf{P}}^*||\overline{\mathbf{P}}^{ind}\right)$ is of order not larger than $D(\mathscr{M})$.

We show that the robustness to misspecification and alteration of the data is not damaged by removing the independence assumption. Let $X_1,\dots,X_n$ be observations given by

$$X_i = E_i\overline{X}_i + (1-E_i)\underline{X}_i, \tag{1.36}$$

where $\overline{\mathbf{X}} = (\overline{X}_i)_i$ is the process of interest, $E_1,\dots,E_n$ are Bernoulli random variables and $\underline{\mathbf{X}} = (\underline{X}_i)_i$ is the contamination process. For $i$ in $\{1,2,\dots,n\}$, the probability of observing $\underline{X}_i$ instead of $\overline{X}_i$ is denote by $q_i = \mathbb{P}(E_i = 0)$.

**Definition 1.7.** *Independent contamination.*
*We talk about independent contamination of the data if the observations $X_1,\dots,X_n$ are given by (1.36) and the variables $E_1,\dots,E_n,\underline{X}_1,\dots,\underline{X}_n$ and $\overline{\mathbf{X}}$ are mutually independent.*

We show that, in the case of independent contamination, the dependence term is not bigger than without any contamination, i.e.

$$\mathbf{K}\left(\mathcal{L}(X_1,\dots,X_n)||\mathcal{L}(X_1)\otimes\cdots\otimes\mathcal{L}(X_n)\right) \leq \mathbf{K}\left(\mathcal{L}(\overline{X}_1,\dots,\overline{X}_n)||\mathcal{L}(\overline{X}_1)\otimes\cdots\otimes\mathcal{L}(\overline{X}_n)\right).$$

Therefore our estimator is robust to independent contamination.

**Proposition 1.1.** *(Corollary 3.1)*
*Let the observations $X_1,X_2,\dots,X_n$ be given by the contamination described by (1.36). If the contamination is independent, the $\rho$-estimator $\hat{P} = \hat{P}(\mathbf{X},\mathscr{M})$ satisfies*

$$C\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} h^2\left(\hat{P},\overline{P}_i\right)\right] \leq \inf_{Q\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^{n} h^2\left(Q,\overline{P}_i\right) + \frac{1}{n}\sum_{i=1}^{n} q_i + \frac{D(\mathscr{M})}{n} + \frac{\mathbf{K}\left(\overline{\mathbf{P}}^*||\overline{\mathbf{P}}^{ind}\right)}{n},$$

*where $C$ is a positive universal constant, $\overline{\mathbf{P}}^* = \mathcal{L}(\overline{X}_1,\dots,\overline{X}_n)$ and $\overline{\mathbf{P}}^{ind} = \bigotimes_{i=1}^n \overline{P}_i = \bigotimes_{i=1}^n \mathcal{L}(\overline{X}_i)$.*

We can see that the deviation bound is not significantly worse as long as the average contamination rate $n^{-1}\sum_{i=1}^{n} q_i$ is of order not larger than $D(\mathscr{M})/n$.

### Estimation strategy for mixing processes

Our estimation strategy basically follows the path initiated in (1.22) but is more elaborate and we use all the observations for a better robustness to independent contamination. We can still use it to simply explain what type of results we can obtain. In the ideal situation where $X_1,\dots,X_n$ are (possibly dependent) random variables with common distribution $\overline{P}$ in $\mathscr{M}$, the $\rho$-estimator $\hat{P}_s = \hat{P}(\mathbf{X}^{(s)},\mathscr{M}^{(s)})$ satisfies

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq \frac{D(\mathscr{M}^{(s)})}{n(s)} + \frac{\mathbf{K}\left(\mathbf{P}_s^*||\mathbf{P}_s^{ind}\right)}{n(s)},$$

where $C$ is a positive universal constant, $\mathscr{M}^{(s)} = \left\{P^{\otimes n(s)}; P\in\mathscr{M}\right\}$, $\mathbf{P}_s^* = \mathcal{L}(\mathbf{X}^{(s)})$ and $\mathbf{P}_s^{ind} = \overline{P}^{\otimes n(s)}$. Based on this deviation bound, if one knows how the quantity $\mathbf{K}\left(\mathbf{P}_s^*||\mathbf{P}_s^{ind}\right)$ behaves with respect to $s$, they can choose a value of $s$ that balances the two terms, as we already know that $n(s)$ behaves roughly like $n/(s+1)$ and the dimension $D(\mathscr{M}^{(s)})$ depends on the choice of the model and should be known.

We prove that the term $\mathbf{K}\left(\mathbf{P}_s^*||\mathbf{P}_s^{ind}\right)/n(s)$ decreases geometrically with respect to $s$ for finite state space HMMs and for a class of diffusion processes observed at regular time steps. It is possible to deduce a bound on the converge rate of our estimator of $\overline{P}$. In that case, we obtain bounds on the convergence rate for the estimation of $\overline{P}$ by taking $s$ of order $\log^2 n$, or $c\log n$ where $c$ is a constant depending on the true distribution. It shows that the number of observations used for the estimation is of order $n/\log^2 n$, in the latter case, and up to a logarithmic factor the convergence rate is the same as if the observations were independent.

### Hidden Markov models

We follow the strategy discussed earlier relying on the fact that the parameters of a stationary finite state space hidden Markov model can be deduced from the distribution of consecutive observations. For integers $L$ and $s$, and observations $\mathbf{X} = (X_1,\dots,X_n)$ we define a new set of variables

$$\mathbf{Y}^{(s)} = (Y_1,Y_{2+s},Y_{1+2(s+1)},\dots,Y_{1+n(s)(s+1)}), \tag{1.37}$$

where $n(s) = \lfloor (n-L)/(s+1)\rfloor$ and

$$Y_i = (X_i,\dots,Y_{i+L-1}), \tag{1.38}$$

for $i$ in $\{1,\dots,n+1-L\}$. One can see that distributions of the form $\mathbb{P}^{(L)}_{(K,Q,\pi,F)}$ given by (1.11) are finite mixtures of product distributions therefore we can rely on the work established in Chapter 2. For classes of distributions $\mathscr{F}_1,\dots,\mathscr{F}_K$, the associated model for the distribution of $L$ consecutive observations of a finite state space HMM is given by

$$\mathscr{H}(K,\mathscr{F}_1,\dots,\mathscr{F}_K) = \left\{\mathbb{P}^{(L)}_{(K,Q,\pi,F)}; w\in\mathcal{W}_K, Q\in\mathcal{T}_K, F_k\in\mathscr{F}_k, \forall k\in[K]\right\}.$$

We use (1.29) to obtain a bound on the $\rho$-dimension of models based on $\mathscr{H}(K,\mathscr{F}_1,\dots,\mathscr{F}_K)$. Assume $\mathcal{F}_1,\dots,\mathcal{F}_K$ are classes of density functions (with respect to a common $\sigma$-finite positive measure) associated with the emission models $\mathscr{F}_1,\dots,\mathscr{F}_K$. If we denote by $\mathscr{M}$ the model given by (1.1) with $\mathscr{M}\subset\mathscr{H}(K,\mathscr{F}_1,\dots,\mathscr{F}_K)$, we can take

$$D(\mathscr{M}) = CL\left(\sum_{1\leq k_1,\dots,k_L\leq K} V_{k_1,\dots,k_L}\right)\log n,$$

where $C$ is a universal positive constant and $V_{k_1,\dots,k_L}$ is the VC-dimension of the product of density models

$$\mathcal{F}_{k_1,\dots,k_L} := \{\mathbf{x} \mapsto f_{k_1}(x_1)\dots f_{k_L}(x_L); f_{k_l} \in \mathcal{F}_{k_l}, \forall l \in [L]\}, k_1,\dots,k_L \in [K],$$

see Proposition 3.5.

If $\mathbf{X}$ is an ergodic finite state space HMM with parameters $(K^*,Q^*,\pi,F^*)$, i.e. $Q^*$ is irreducible and aperiodic, then $Q^*$ admits exactly one invariant distribution that we denote $\pi^*$ and we define the associated distribution $P^*$ given by

$$P^* = \mathbb{P}^{(L)}_{(K^*,Q^*,\pi^*,F^*)}, \tag{1.39}$$

where $\mathbb{P}^{(L)}_{(K,Q,\pi,F)}$ is given by (1.11). Ergodicity also allows us to bound the dependence term and to show that most of the distributions $(\mathcal{L}(Y_i))_i$ lie in a small neighborhood of $P^*$, making the stationarity assumption unnecessary. In that case, there exist positive constants $C(Q^*)$ and $c(Q^*)$ depending on $Q^*$ such that for $s \geq c(Q^*)\log n$, the $\rho$-estimator $\hat{P}_s = \hat{P}(\mathbf{Y}^s,\mathscr{M}^{(s)}) \in \mathscr{M}$ satisfies

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\mathscr{M}\right) + L\left(\sum_{1\leq k_1,\dots,k_L\leq K} V_{k_1,\dots,k_L}\right)\frac{(s+1)\log n}{n},$$

see Theorem 3.4 and (3.40) in particular. If we know the constant $c(Q^*)$, or eventually an upper bound $c^+$ on it, we can take $s = \lceil c^+ \log n\rceil$ which gives

$$C(Q^*,c^+)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\mathscr{M}\right) + L\left(\sum_{1\leq k_1,\dots,k_L\leq K} V_{k_1,\dots,k_L}\right)\frac{\log^2 n}{n},$$

and if $P^*$ belongs to $\mathscr{M}$ we obtain a convergence rate of order $n^{-1}\log^2 n$ with respect to the squared Hellinger distance, which is optimal up to a logarithmic factor. If it is not the case we can take $s$ of order $\log^2$ which establishes a rate with only an additional logarithmic factor. We obtain a bound of order $n^{-1}\log^3$ for all $P^*$ in the subset $\mathscr{M}^* \subset \mathscr{M}$ of ergodic HMMs given by

$$\mathscr{M}^* := \left\{ \mathbb{P}^{(L)}_{w,Q,F} \in \mathscr{M};\ \begin{array}{c} Q \text{ irreducible }, \\ Q \text{ aperiodic}, \\ \text{and } w = Qw \end{array} \right\}.$$

However, this bound is not uniform as some of the constants depend on the transition matrix of the true HMM.

We show a bound for the VC-dimensions $V_{k_1,\dots,k_L}$ when all the emission models are exponential families, i.e. for all $k$ the emission model has a associated class of densities $\mathcal{F}_k$ given by

$$\mathcal{F}_k := \left\{ f_\theta : x \mapsto e^{\langle \eta_k(\theta),T_k(x)\rangle + A_k(\theta) + B_k(x)}; \theta \in \overline{\Theta}_k \right\},$$

where $\overline{\Theta}_k$ is a non-empty set, $T : \mathscr{X} \to \mathbb{R}^{d_k}$ and $B : \mathscr{X} \to \mathbb{R}$ are measurable functions, $\eta : \overline{\Theta} \to \mathbb{R}^{d_k}$ is such that $A_k(\theta) := -\log \int_{\mathscr{X}} e^{\langle \eta_k(\theta),T_k(x)\rangle + B_k(x)}\nu(dx)$ is well-defined. In that case, we have $V_{k_1,\dots,k_L} \leq 3 + d_{k_1} + \dots + d_{k_L}$ for all $k_1,\dots,k_L$ such that

$$D(\mathscr{M}) \leq C\left[3K^L + LK^{L-1}(d_1 + \dots + d_K)\right],$$

where $C$ is a positive universal constant (see Proposition 3.1). Exponential families include usual parametric models such as normal, exponential, gamma or beta distributions for example. We consider the cases of location and location-scale families of multivariate normal distributions and provide bounds on the convergence rate for the estimation of $P^*$ in Theorem 3.7.

As for mixture models, we can deduce deviation bounds for the parameter estimators when the Hellinger distance between distributions is lower bounded by a distance between the corresponding parameters (see (1.32)). We use the theory of Ibragimov & Has'minskiĭ [54] to prove such an inequality for models with exponential families as emission models and satisfying some regularity conditions. We obtain the expected parametric rate for the parameter estimators, up to a logarithmic factor. We illustrate this result for finite state space HMMs with emission distributions that are exponential distributions in Theorem 3.9. If the observations come from a HMM with an ergodic transition matrix $\overline{Q}$ and emission distributions $\mathcal{E}(\overline{\theta}_1),\dots,\mathcal{E}(\overline{\theta}_K)$, for $s=\lceil \log^2 n\rceil$ our parameter estimators satisfy

$$C(\overline{Q},\overline{\theta})\mathbb{E}\left[||\overline{w}-\hat{w}||^2+||\overline{Q}-\hat{Q}||^2+\sum_{k=1}^{K}(\overline{\theta}_k-\hat{\theta}_k)^2\wedge 1\right]\leq K^3\frac{\log^3 n}{n},$$

where $\overline{w}$ is the stationary distribution with respect to $\overline{Q}$, $C(\overline{w},\overline{Q},\overline{\theta})$ is a positive constant depending on the true parameters and $\mathcal{E}(\theta)$ is the exponential distribution with parameter $\theta$.

We can also consider nonparametric estimation with finite nets as an approximation, following the approach of (1.28). We illustrate it for emission models containing distributions with densities that are log-concave. Under the ergodicity assumption and if the emission distributions have a log-concave density with respect to the Lebesgue measure, there exist a positive constants $C(P^*)$ such that for $s=\lceil \log^2 n\rceil$ we have

$$C(P^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right]\leq L^2K^L\frac{\log^{a_d} n}{n^{b_d}},$$

where the constants $a_d$ and $b_d$ depend on the dimension $d$ and are given in Theorem 3.5. The convergence rates are similar to the ones obtained for density estimation of a log-concave density from i.i.d. observation (see Kim & Samworth [58] and Kur *et al.* [60]). We only obtain worse logarithmic terms due to the higher complexity of the model and the dependence within the observations. It is possible to deduce deviation bounds for the parameter estimators under additional assumptions, using inequality (1.33) and the results of Lehéricy [63], such as (1.12).

We also consider an atypical example inspired by Example 1.6. We take $L=2$ and believe that $P^*$ is of the form

$$P_{w,z,q}=(1-w)\mathbb{S}_{\alpha,0}\otimes(q_{0,1}\mathbb{S}_{\alpha,z}+(1-q_{0,1})\mathbb{S}_{\alpha,0})+w\mathbb{S}_{\alpha,z}\otimes(q_{1,0}\mathbb{S}_{\alpha,0}+(1-q_{1,0})\mathbb{S}_{\alpha,z}).$$

This corresponds to a translation hidden Markov model with two states, one of which is known. If the true distribution is an ergodic finite state space HMM and $P^*$ is actually of the form $P_{\overline{w},\overline{z},\overline{q}}$, for $s=\lceil \log^2 n\rceil$ our parameter estimators satisfy

$$C\left(P^*\right)\mathbb{E}\left[(\overline{w}-\hat{w})^2+(\overline{q}_{12}-\hat{q}_{12})^2+(\overline{q}_{21}-\hat{q}_{21})^2+(|\overline{z}-\hat{z}|\wedge 1)^{1-\alpha}\right]\leq\frac{\log^3 n}{n},$$

where $C(P^*)$ is a positive constant that depends on $P^*$. We obtain the usual $1/\sqrt{n}$ parametric rate, with respect to the Euclidean distance, for the estimation of the transition probabilities $\overline{q}_{0,1},\overline{q}_{1,0}$ and of the stationary distribution $\overline{w}$, up to a logarithmic factor. As in the case of mixtures, we obtain the faster rate $n^{-1/(1-\alpha)}$ for the location parameter $\overline{z}$, up to a logarithmic factor.

### Diffusion processes

In Section 3.3, we consider the problem of estimating the invariant distribution of the stochastic differential equation

$$dY_t=dB_t-\nabla U(Y_t)dt, \tag{1.40}$$

where $(B_t)_{t\geq 0}$ is a $d$-dimensional Brownian motion, $d \geq 1$. Under some conditions on $U : \mathbb{R}^d \to \mathbb{R}$ this equation is well defined and the associated invariant distribution $\overline{P}$ has a density $\overline{p}$ with respect to the Lebesgue measure given by

$$\overline{p}(y) = \frac{e^{-2U(y)}}{\int_{\mathbb{R}^d} e^{-2U(y)} dy}.$$

We consider the estimation of $\overline{P}$ from discrete observations $X_1, \ldots, X_n$ assuming they are given by $X_i = Y_{t_i}$, where $(Y_t)_t$ is a stationary solution of (1.40) and $0 \leq t_1 < \cdots < t_{n-1} < t_n < \infty$. For the sake of simplicity we consider a constant time step $\Delta_t > 0$ such that $t_{i+1} = t_i + \Delta_t$

Strict convexity of $U$ is a sufficient condition for everything to be well-defined and it has two interesting consequences. In that case, the distribution $\overline{P}$ has a log-concave density with respect to the Lebesgue measure. There is a rich literature on log-concave density estimation in the i.i.d. context, particularly for the maximum likelihood estimator. Dümbgen & Rufibach [34] and Schuhmacher & Dümbgen [81] established consistency results for the maximum likelihood. Dümbgen *et al.* [35] investigated the approximation properties of log-concave densities. More recently, Kim & Samworth [58] and Kur *et al.* [60] proved non-asymptotic results for the maximum likelihood estimator with respect to the squared Hellinger distance. We can rely on some of the different results contained in those papers to consider the estimation of $\overline{P}$.

We obtain a risk bound for our estimator $\hat{P}$ of $\overline{P}$ in any dimension. There is a positive constant $c(U)$ such that for $s \geq c(U) \log n$ we have

$$C(U,\Delta_t)\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq \frac{(s+1)\log^{a_d} n}{n^{b_d}},$$

where $a_d$ and $b_d$ are constants given in Theorem 3.2 that depend on the dimension $d$. We obtain convergence rates that are similar to the ones of Kim & Samworth [58] and Kur *et al.* [60] for i.i.d. observations. We only obtain worse powers of $\log n$ except in the 3-dimensional case. To our knowledge, the approach we propose is quite new in the context of diffusion processes. In particular, there are no comparable results of robust estimation in a similar framework.

#### Selection of the spacing parameter

The procedure described by (1.22) requires the statistician to specify the spacing parameter $s$ giving the subset $\mathbf{X}^{(s)}$ of observations to be used for the estimation. One needs some knowledge on the true distribution in order to choose a satisfactory value of $s$. This is restrictive as we want to avoid making any assumption on the true distribution of the observations. In Section 3.5, we propose a strategy to automatically select a value of $s$ from a second set of observations. For a subset $\mathscr{M}$ of $\mathscr{P}_X$ and independent sets of observations

$$\mathbf{X}^{(1)} := \left(X_1^{(1)}, \ldots, X_{n_1}^{(1)}\right) \text{ and } \mathbf{X}^{(2)} := \left(X_1^{(2)}, \ldots, X_{n_2}^{(2)}\right)$$

we consider the following procedure. We use the first set $\mathbf{X}^{(1)}$ to get an estimator $\hat{P}_s = \hat{P}\left(\mathbf{X}^{(1,s)}, \mathscr{M}^{(s)}\right)$ for different values of $s$ in a set $S$. In a second time, we use $\mathbf{X}^{(2)}$ to select a value $\hat{s}$ in $S$ and our final estimator is $\hat{P} = \hat{P}_{\hat{s}}$. We show under minimal assumptions that our estimator performs almost as good as if we knew the optimal value for $s$. We also show that the estimator is still robust to independent contamination and we present an application to finite state space hidden Markov models.

### 1.4.5 Model selection for finite state space HMMs (Chapter 4)

For the sake of simplicity, we do not consider the possibility to use model selection in Chapter 3. We extend this framework in Chapter 4 with model selection for finite state space HMMs,

allowing us to consider different possible values for the order $K$ and/or different emission models. For the sake of simplicity, we consider the two situations separately. For the selection of the order, we can show that if the true distribution belongs to the model with order $K^*$, we do not underestimate $K^*$ for $n$ large enough. For the estimation of the distribution $P^*$ given by (1.39), the estimator achieves the same rate as if $K^*$ was known. We illustrate this with an application with Poisson emission distributions. For the selection of the emission models, we take the example of multivariate Gaussian emission models. We define classes of covariance matrices fixing some of the coefficients to be null and associate a class of multivariate normal distributions to each class of covariance matrices. We consider those classes as potential emission models and use model selection with a penalty related to the number of zeros in the covariance matrices. We show that the dependence on the dimension in the risk bound can be improved if the true covariance matrices of the emission distributions are sparse, which is very interesting in high dimensions.

### 1.4.6 General HMMs (Chapter 5)

We present another extension in Chapter 5 considering hidden Markov models with a general state space. We consider the case of univariate normal emission distributions. The model assumes that observations $X_1,\dots,X_n$ are given by

$$X_i = z_i + \sigma_i^2 \varepsilon_i,$$

where $(z_i,\sigma_i^2)_i$ is a Markov chain on $\mathbb{R}\times(0,\infty)$ and $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. standard normal distributions. We prove a result of identifiability from the distribution of two consecutive observations. If $(h_i)_i = (z_i,\sigma_i^2)_i$ is a Markov chain with Markov kernel $Q$ and initial distribution $\pi$ stationary with respect to $Q$, the distribution $P_{\pi,Q}$ of $(X_i,X_{i+1})$ has a density with respect to the Lebesgue measure given by

$$p_{\pi,Q}(x_1,x_2) = (2\pi)^{-1} \int \sigma_1^{-1} e^{-\frac{(x_1-z_1)^2}{2\sigma_1^2}} \sigma_2^{-1} e^{-\frac{(x_2-z_2)^2}{2\sigma_2^2}} Q(h_1,dh_2)\pi(dh_1).$$

We prove that such distributions can be well approximated by finite mixtures of multivariate normal distributions when $\pi$ and $Q$ are supported on a compact set. We build an estimator of $P_{\pi,Q}$ based on such mixtures and prove a bound on its convergence rate. To our knowledge, this is the first non-asymptotic result for a general HMMs that are not translation HMMs. In our case, it corresponds to the situation where the variance $\sigma_i^2$ is constant.

## 1.5 Reminder of Vapnik-Chervonenkis theory

For a more detailed introduction to VC-subgraph classes we refer the reader to Van der Vaart & Wellner [84] (Section 2.6.5) and Baraud *et al.* [9] (Section 8). The VC-dimension was originally introduced by Vapnik & Chervonenkis [85] to measure the complexity of a model in binary classification. Let $\mathcal{C}$ be a collection of subsets of a set $\mathscr{X}$. We say that $\mathcal{C}$ *shatters* a set $S = \{x_1,\dots,x_k\} \subset \mathscr{X}$ if each subset of $S$ can be obtained by taking its intersection $C \cap S$ with some set $C \in \mathcal{C}$. This means that a classification algorithm based on $\mathcal{C}$ can learn a perfect classifier for the sample $S$.

**Definition 1.8.** *The VC-dimension of $\mathcal{C}$ is the largest cardinality $|S|$ of a set $S \subset \mathscr{X}$ that $\mathcal{C}$ shatters. We say that $\mathcal{C}$ is a VC-class (of sets) if its VC-dimension $V(\mathcal{C})$ is finite. We can also use the VC-index $\overline{V}(\mathcal{C})$ given by $\overline{V}(\mathcal{C}) = V(\mathcal{C}) + 1$.*

This notion can be extended to classes of real-valued functions through the classes of sets given by their subgraph. Let $\mathcal{F}$ be a collection of functions $\mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. For $f$ in $\mathcal{F}$ we define its subgraph by $C_f = \{(x,t) \in \mathcal{X} \times \mathbb{R}; f(x) > t\}$.

**Definition 1.9.** *The VC-dimension of the class of functions $\mathcal{F}$ is the VC-dimension of the class of sets $\mathcal{C}_{\mathcal{F}} = \{C_f; f \in \mathcal{F}\}$. We say that $\mathcal{F}$ is a VC-class or a VC-subgraph class if its VC-dimension $V(\mathcal{F})$ is finite.*

We can also use the notion of VC-index for classes of functions. The remarks below immediately follow from this definition.

- If $\mathscr{F}$ is VC-subgraph with dimension $V$, then any subset $\mathscr{G} \subset \mathscr{F}$ is VC-subgraph with dimension at most $V$.
- If $\mathscr{F}$ is a finite set, $\mathscr{F}$ is VC-subgraph and its dimension is not larger than $V = \log_2(|\mathscr{F}|) \vee 1$.

We can state some results that give an idea of the VC-dimension in some cases. The following one relates the dimension of a finite-dimensional vector space of functions to its VC-dimension.

**Proposition 1.2.** *(Lemma 2.6.15 [84])*
*Any finite-dimensional vector space $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ with dimension $d(\mathcal{F})$ is VC-subgraph with VC-dimension smaller than or equal to $d(\mathcal{F}) + 1$.*

The result below allows us to establish bounds on the VC-dimension that are more complicated than simple vector spaces.

**Proposition 1.3.** *(Proposition 42 [9])*
*Let $\mathcal{F}$ be VC-subgraph with VC-dimension $V$ on a set $\mathcal{X}$.*

1. *For all function $g : \mathcal{X} \to \mathbb{R}$, the class of functions $\mathcal{F} + g = \{f + g, f \in \mathcal{F}\}$ is VC-subgraph with VC-dimension not larger than $V$.*
2. *For all monotone function $\varphi$ on $\mathbb{R}$, the class of functions $\varphi(\mathcal{F}) = \{\varphi \circ f, f \in \mathcal{F}\}$ is VC-subgraph with VC-dimension not larger than $V$.*

With the two propositions above, we can already prove that any exponential family of densities is VC-subgraph. For multivariate normal densities for instance, we can see that we recover the usual notion of dimension in parametric models, up to an additive constant.

**Proposition 1.4.** *(Lemma 2.1)*
*Let $d \geq 1$. Let $Cov_{+*}(d)$ be the set of $d \times d$ symmetric and positive-definite matrices. For $\mu \in \mathbb{R}^d$ and $\Sigma \in Cov_{+*}(d)$, we denote by $g_{\mu,\Sigma}$ the density function of $\mathcal{N}(\mu,\Sigma)$ with respect to the Lebesgue measure given by*

$$g_{\mu,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

*Let $\mathcal{G}_d$ be the location-scale family of densities given by $\mathcal{G}_d := \left\{g_{\mu,\Sigma}; \mu \in \mathbb{R}^d, \Sigma \in Cov_{+*}\right\}$. For a fixed $\Sigma$, we denote by $\mathcal{G}_{loc}(\Sigma)$ the associated location family given by $\mathcal{G}_{loc}(\Sigma) := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d\right\}$. The sets $\mathcal{G}_d$ and $\mathcal{G}_{loc}(\Sigma)$ are VC-subgraph with VC-index bounded by $3 + \frac{d(d+3)}{2}$ and $3 + d$ respectively.*

## 1.6 Possible extensions

We can identify different directions in which the work of this thesis can be extended.

- We considered continuous mixtures of Gaussian distributions and HMMs with general state space with Gaussian emission distributions. This could be generalized to other location or location-scale parametric families for the emission distribution and eventually to a nonparametric framework.
- Our framework focuses on cases where observations are identically distributed or close to it. This does not allow us to investigate time series with trends or cycles for example.
- We provided some results similar to (1.12) to lower bound the Hellinger distance between distributions by a distance on the parameters for specific cases of mixture models and hidden Markov models. It would be interesting to investigate further the problem of deducing the parameters from the distribution of interest in cases that have not been treated or eventually other types of latent variable models.
- We considered a specific class of stochastic differential equations. The approach could be easily extended to other diffusion processes. The main difficulty is to bound the dependence term. To do so we need to investigate the quantity

$$\mathbf{K}\left(\mathcal{L}(X_t,X_{t+s})||\mathcal{L}(X_t)\otimes\mathcal{L}(X_{t+s})\right),t,s>0.$$

## 1.7 Organization of the thesis

Chapter 2 is based on the paper Lecestre [62] published in *ESAIM: PS* and is dedicated to mixture models. We consider observations that are not necessarily independent in Chapter 3 and present applications to finite state space HMMs and a type of diffusion processes. This chapter is based on the arXiv paper Lecestre [61]. The last two chapters are ongoing projects that extend the applications to hidden Markov models presented in Chapter 3. Chapter 4 presents a general result of model selection for finite state space HMMs with applications to the selection of the order and the selection of the emission models. Chapter 5 deals with estimation for general state space hidden Markov models for the particular case of normal emission distributions.

# Chapter 2

# Robust estimation in finite mixture models

**Abstract**

We observe a $n$-sample, the distribution of which is assumed to belong, or at least to be close enough, to a given mixture model. We propose an estimator of this distribution that belongs to our model and possesses some robustness properties with respect to a possible misspecification of it. We establish a non-asymptotic deviation bound for the Hellinger distance between the target distribution and its estimator when the model consists of a mixture of densities that belong to VC-subgraph classes. Under suitable assumptions and when the mixture model is well-specified, we derive risk bounds for the parameters of the mixture. Finally, we design a statistical procedure that allows us to select from the data the number of components as well as suitable models for each of the densities that are involved in the mixture. These models are chosen among a collection of candidate ones and we show that our selection rule combined with our estimation strategy result in an estimator which satisfies an oracle-type inequality.

## 2.1 Introduction

Mixture models are a flexible tool for modeling heterogeneous data, e.g. from a population consisting of multiple hidden homogeneous subpopulations. Finite mixture models are models containing distribution of the form

$$P_{w,F} = \sum_{k=1}^{K} w_k F_k, \tag{2.1}$$

where $K \geq 2$, each $F_k$ belongs to a specific class of probability distributions (e.g. normal distributions in the case of Gaussian mixture models) and $w$ belongs to the simplex $\mathcal{W}_K = \left\{w \in [0,1]^K; w_1 + \cdots + w_k = 1\right\}$. For a complete introduction to mixture models and an overview of the different applications we refer to the books of Mclachlan & Peel [71] and Frühwirth-Schnatter [38].

Assume we have a sample $\mathbf{X} := (X_1, \ldots, X_n)$ of i.i.d. data, each coordinate following the probability distribution $P^*$. The majority of the statistical methods based on finite mixture models aim to solve one of the following problems: density estimation (estimation of $P^*$), parameter estimation (estimation of $w^*$ and/or $F^*$ assuming $P^* = P_{w^*,F^*}$) and clustering. The monographs of Everitt & Hand [36] or Titterington *et al.* [83] provide a good overview of the different estimation methods that have been developed for mixture models such as maximum likelihood, minimum chi-square, moments method and Bayesian approaches. Although algorithms are numerous, theoretical guarantees are mostly asymptotic and restricted to very specific situations. To our knowledge, only a few non-asymptotic results have been established in the case of density estimation based on Gaussian Mixture Models (GMMs). The approximation and entropy properties of Gaussian mixture sieves have been investigated by Kruijer *et al.* [59], Ghosal & van der Vaart [49] and Genovese & Wasserman [48] where bounds on the convergence rate are given for the MLE and Bayesian estimators. Similarly, Maugis & Michel [69] use a penalized version of the MLE to build a Gaussian mixture estimator with non asymptotic adaptive properties proven in [68]. However, those results rely on relatively strong assumptions and estimators are not proved to be robust to small departures from those assumptions.

This paper aims to provide non-asymptotic results in a very general setting. In our framework, the data are assumed to be independent but not necessarily i.i.d. Our mixture model consists of probabilities of the form (2.1) where the $F_k$ admit densities, called *emission densities*, that belong to classes of function that are VC-subgraph. We investigate the performances of $\rho$-estimators, as defined by Baraud and Birgé [11], on finite mixture models. This paper only focuses on the theoretical aspects and performances. We do not consider here the problem of

computing estimators in practice. Our main result, Theorem 2.1, is an exponential deviation inequality for the risk of the estimator $\hat{P}$, which is measured with an Hellinger-type loss. We get an upper bound on the risk that is the sum of two terms. The first one is an approximation term which provides a measure of the distance between the true distribution of the data and our mixture model. The second term is a complexity term that depends on the classes containing the emission densities and which is proportional to the sum of their VC-indices. We deduce from this deviation bound that the estimator is not only robust with respect to model misspecification but also to contamination and the presence of outliers among the data set. Dealing with models that may be approximate allows us to build estimators that possess properties over wider classes of distribution. Ghosal & Van der Vaart [49] used finite location-scale Gaussian mixtures to approximate general Gaussian mixtures with compactly supported mixing distribution. They consider mixtures with scale parameters lying between two constants that depend on the true distribution. By using a similar approximation (see Proposition 2.1), we show in Theorem 2.3 that our estimator achieves the same rate of convergence but without any restriction on the scale parameters so that the model we consider does not depend on the true mixing distribution. In particular, our result is insensitive to translation or rescaling.

Under suitable identifiability assumptions and when the distribution of the data belongs to our model, hence is of the form (2.1), we also analyze the performance of our estimators of the parameters $w_1, \ldots, w_K$ and $F_1, \ldots, F_K$. In order to establish convergence rates, we relate the Hellinger distance between the distribution of the data and its estimator to a suitable distance between the corresponding parameters. A general technique is using Fisher's information and results of Ibragimov & Has'minskiĭ [54] for regular parametric models. We can also use other results specific to parameter estimation in mixture models such as what Gadat *et al.* [41] proved in the context of two component mixtures with one known component. In both situations, we obtain, up to a logarithmic parameter, the usual $1/\sqrt{n}$-rate of convergence for regular parametric models. We also provide with Theorem 2.7 the example of a parametric model for which our techniques allow us to establish faster convergence rates while classical methods based on the likelihood or the least-squares fail to apply and hence give nothing.

In many applications, starting with a single mixture model may be restrictive and a more reasonable approach is to consider candidate ones for estimating the number of components of the mixture and proposing suitable models for the emission densities. To tackle this problem, we design a model selection procedure from which we establish, under suitable assumptions, an oracle-type inequality. We consider several illustrations of this strategy. For example, we use a penalized estimator to select the number of components of a Gaussian mixture estimator and obtain similar adaptivity results as Maugis-Rabusseau & Michel [68]. We also consider a model with a fixed number of components but each emission density can either belong to the Gaussian or to the Cauchy location-scale family. We prove that if we know the number of components, we can estimate consistently the proportions of Gaussian and Cauchy components as well as their location and scale parameters. To our knowledge, this result is the first of its kind.

The extension of the theory of $\rho$-estimation to mixture models is based on Proposition 2.3 below. The proof of this result relies on an upper bound for the expectation of the supremum of an empirical process over a mixture of VC-subgraph classes. It generalizes the result that was previously established for a single VC-subgraph class. The key argument in the proof is the uniform entropy property of VC-subgraph classes that still holds for the overall density mixture model with lower bounded weights.

The paper is organized as follows. We describe our statistical framework in Section 2.2. In Section 2.3, we present the construction of the estimator on a single mixture model. We state the general result for density estimation on a single model and illustrate the performance of the estimator on the specific example of GMMs. The problem of estimating the parameters of the mixture is addressed in the subsection 2.3.5. Finally, Section 2.4 is devoted to model selection

criterion and the properties of the estimator on the selected model. The appendix contains all the proofs that are gathered in the same sections when they are related. Those sections include the main results, density estimation, the parametric estimation in regular parametric models, the case of two-component mixtures with one known component and the lemmas.

## 2.2 The statistical framework

We observe $n$ independent random variables $X_1,X_2,\ldots,X_n$ with respective marginal distributions $P_1^*,P_2^*,\ldots,P_n^*$ on the measurable space $(\mathcal{X},\mathcal{X})$. We model the joint distribution $\mathbf{P}^* = P_1^* \otimes P_2^* \otimes \cdots \otimes P_n^*$ of $\mathbf{X} = (X_1,X_2,\ldots,X_n)$ by a probability of the form $\overline{P}^{\otimes n}$ doing as if the observations were i.i.d. with common distribution $\overline{P}$. We assume that $\overline{P}$ is a mixture of the form (2.1) where $K$ is a positive integer, the $w_k$ some positive weights that satisfy $\sum_{k=1}^K w_k = 1$, and $F_k$ probability distributions. In order to model each of these probabilities we introduce a collection $\left\{\overline{\mathscr{F}}_{k,\lambda}; k \geq 1, \lambda \in \Lambda_k\right\}$ of possible models and assume that for each $k \in \{1,\ldots,K\}$, $F_k$ belongs to $\cup_{\lambda\in\Lambda_k}\overline{\mathscr{F}}_{k,\lambda}$. We denote by $\mathscr{Q}_K$ the family of distributions of the previous form. For each $k \geq 1$, we call $F_k$ an emission probability, $\overline{\mathscr{F}}_{k,\lambda}$ an emission model, and $\mathscr{E}_k = \left\{\overline{\mathscr{F}}_{k,\lambda}; \lambda \in \Lambda_k\right\}$ an emission family. Based on the observation of $\mathbf{X}$, our aim is to design an estimator $\hat{P}$ of $\overline{P}$ of the form

$$\hat{P} = \sum_{k=1}^{\hat{K}} \hat{w}_k \hat{F}_k \in \bigcup_{K\geq 1} \mathscr{Q}_K \tag{2.2}$$

where $\hat{K}$, $(\hat{w}_k)_{1\leq k\leq \hat{K}}$ and $(\hat{F}_k)_k$ are estimators of $K$, $(w_k)_k$ and $(F_k)_k$ respectively. There are a lot of possibilities for the collections $\Lambda_k$, depending on the estimation strategy (nonparametric, polynomial basis, wavelets, ...). We illustrate it in detail with the following example of usual parametric models on $\mathbb{R}$.

**Example 2.1.** *Let us take* $\Lambda_k = \{1,2,3\}$ *with*

- *the Gaussian location-scale family,*

$$\overline{\mathscr{F}}_{k,1} = \mathscr{G} = \{\mathcal{N}(\mu,\sigma); \mu \in \mathbb{R}, \sigma > 0\}; \tag{2.3}$$

- *the Cauchy location-scale family,*

$$\overline{\mathscr{F}}_{k,2} = \mathscr{C} = \{Cauchy(\mu,\sigma); \mu \in \mathbb{R}, \sigma > 0\};$$

- *and the Laplace location-scale family,*

$$\overline{\mathscr{F}}_{k,3} = \mathscr{L} = \{Laplace(\mu,\sigma); \mu \in \mathbb{R}, \sigma > 0\}.$$

The classical situation that has been considered in the literature corresponds to the case where the collection $\left\{\overline{\mathscr{F}}_{k,\lambda}; k \geq 1, \lambda \in \Lambda_k\right\}$ reduces to a single emission model $\mathscr{F}$, for example the family of Gaussian distributions, and the problem is to estimate $K$ and the emission probabilities $F_k$ under the assumption that they all belong to $\mathscr{F}$. This assumption is quite restrictive and we rather consider a collection $\mathscr{E}_k$ of candidate models for $F_k$ that may even depend on $k$. We say that $\mathscr{E}_k$ is simple when it reduces to a single emission model $\overline{\mathscr{F}}_k$ and composite otherwise.

In order to evaluate the performance of the estimator $\hat{P}$, we introduce on the set $\mathscr{P}$ of all product probabilities on $(\mathscr{X}^n,\mathcal{X}^{\otimes n})$ the Hellinger-type distance $\mathbf{h}$ defined by

$$\mathbf{h}(\mathbf{Q},\mathbf{Q}') = \sqrt{\sum_{i=1}^{n} h^2(Q_i,Q_i')}, \quad \text{for } \mathbf{Q} = \bigotimes_{i=1}^{n} Q_i, \mathbf{Q}' = \bigotimes_{i=1}^{n} Q_i' \in \mathscr{P}, \tag{2.4}$$

where $h$ is the Hellinger distance on the set $\mathscr{P}$ of probability distributions on $(\mathscr{X},\mathcal{X})$. We recall that for $Q$, $Q'$ in $\mathscr{P}$

$$h^2(Q,Q') = \frac{1}{2}\int \left(\sqrt{\frac{dQ}{d\mu}} - \sqrt{\frac{dQ'}{d\mu}}\right)^2 d\mu,$$

where $\mu$ is a measure that dominates both $Q$ and $Q'$, the result being independent of $\mu$.

**Assumption 2.1.** *For all $k \geq 1$, the set $\Lambda_k$ is at most countable (which means finite or countable) and such that for all $\lambda$ in $\Lambda_k$, $\overline{\mathscr{F}}_{k,\lambda}$ contains an at most countable subset $\mathscr{F}_{k,\lambda}$ which is dense in $\overline{\mathscr{F}}_{k,\lambda}$ with respect to the Hellinger distance $h$.*

This assumption is only made for technical reasons, i.e. it ensures the measurability of the different objects considered in the proofs. But it is not really restrictive as, from a very practical point of view, one would only deal with rational numbers which already restrict to countable models. Moreover, one can check that $\mathscr{F}_{k,1} = \{\mathcal{N}(\mu,\sigma); \mu \in \mathbb{Q}, \sigma \in \mathbb{Q} \cap (0,\infty)\}$ satisfy our assumption in the context of Example 2.1. It holds as well for $\mathscr{F}_{k,2}$, $\mathscr{F}_{k,3}$ and $\mathscr{F}_{k,4}$ with the same construction. Given Assumption 2.1 we can fix some notation. The countability condition implies that there exists a $\sigma$-finite measure $\mu$ that dominates all the $\overline{\mathscr{F}}_{k,\lambda}$ for $k \geq 1$ and $\lambda \in \Lambda_k$. Throughout this paper, we fix such a measure $\mu$ and associate to each emission model $\overline{\mathscr{F}}_{k,\lambda}$ a family of density distributions $\overline{\mathcal{F}}_{k,\lambda}$ such that $\overline{\mathscr{F}}_{k,\lambda} = \left\{f \cdot \mu; f \in \overline{\mathcal{F}}_{k,\lambda}\right\}$. In all the different examples considered in the rest of the paper $\mu$ is the Lebesgue measure. As explained, Assumption 2.1 is necessary for very technical reasons. The next assumption allows us to bound the "dimension" of the model (see the introduction or Proposition 2.3).

**Assumption 2.2.** *For all $k \geq 1$ and $\lambda \in \Lambda_k$, the family of density distributions $\overline{\mathcal{F}}_{k,\lambda}$ is VC-subgraph with VC-index smaller than or equal to $V_{k,\lambda} \geq 1$.*

In order to avoid too much technicality in the core of this paper, we dedicated Section 2.F to VC-subgraph classes of functions with the definition and proofs of the different results. The next lemma shows that the VC-index corresponds to what we expect as the "dimension" of the model in the case of multivariate for normal distributions.

**Lemma 2.1.** *Let $d \geq 1$. Let $Cov_{+*}(d)$ be the set of $d \times d$ symmetric and positive-definite matrices. For $\mu \in \mathbb{R}^d$ and $\Sigma \in Cov_{+*}(d)$, we denote by $g_{\mu,\Sigma}$ the density function of $\mathcal{N}(\mu,\Sigma)$ with respect to the Lebesgue measure given by*

$$g_{\mu,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

*Let $\mathcal{G}_d$ be the location-scale family of densities given by $\mathcal{G}_d := \left\{g_{\mu,\Sigma}; \mu \in \mathbb{R}^d, \Sigma \in Cov_{+*}\right\}$. For a fixed $\Sigma$, we denote by $\mathcal{G}_{loc}(\Sigma)$ the associated location family given by $\mathcal{G}_{loc}(\Sigma) := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d\right\}$. The sets $\mathcal{G}_d$ and $\mathcal{G}_{loc}(\Sigma)$ are VC-subgraph with VC-index bounded by $3 + \frac{d(d+3)}{2}$ and $3+d$ respectively.*

The dependence in $d$ is linear and quadratic for the location family and location-scale family respectively, as for the number of parameters needed to describe each class. Throughout this paper we shall use the following notation. For $\mathbf{P} = P_1 \otimes \cdots \otimes P_n \in \mathscr{P}$ and $\mathscr{A} \subset \mathscr{P}$, we write

$$\mathbf{h}^2(\mathbf{P},\mathscr{A}) = \inf_{Q\in\mathscr{A}} \mathbf{h}^2\left(\mathbf{P},Q^{\otimes n}\right) = \inf_{Q\in\mathscr{A}} \sum_{i=1}^{n} h^2(P_i,Q).$$

For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the only integer satisfying $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$ and similarly $\lceil x \rceil$ denotes the integer satisfying $\lceil x \rceil - 1 < x \leq \lceil x \rceil$. Moreover, if $x > 0$ we write $\log_+(x) = \log(x) \vee 0$. If $A$ is a finite set, we denote its cardinal by $|A|$ and if $A$ is infinite, we write $|A| = \infty$. For $k$ in $\mathbb{N}^*$, we denote by $[k]$ the set $\{1,2,\ldots,k\}$. The notation $C(\theta)$ will mean that the constant $C = C(\theta)$ depends on the parameter or set of parameters $\theta$.

## 2.3 Estimation on a mixture model based on simple emission families

In this section, we assume that the $\mathscr{E}_k = \left\{\overline{\mathscr{F}}_k\right\}$ are simple for all $k \geq 1$ and that $\overline{P}$ belongs to $\mathscr{Q}_K$ for some known value of $K \geq 1$. This means that we know that $\overline{P}$ is a mixture of at most $K$ emission probabilities $F_1,\ldots,F_K$ and that $F_k$ belongs to $\overline{\mathscr{F}}_k$ for all $k \in [K]$. Under Assumption 2.2, we denote by $V_k$ the VC-index of $\overline{\mathscr{F}}_k$.

### 2.3.1 Construction of the estimator on $\mathscr{Q}_K$

For $\delta$ in $(0,1/K]$, we define the subset $\mathscr{Q}_{K,\delta}$ of $\mathscr{Q}_K$ by

$$\mathscr{Q}_{K,\delta} := \left\{ \sum_{k=1}^{K} w_k F_k \in \mathscr{Q}_K; w \in \mathcal{W}_K \cap ([\delta,1] \cap \mathbb{Q})^K, F_k \in \mathscr{F}_k \right\} \tag{2.5}$$

where the $\mathscr{F}_k$ are the countable and dense subsets of $\overline{\mathscr{F}}_k$ provided by Assumption 2.1. We associate to $\mathscr{Q}_{K,\delta}$ the family $\mathcal{Q}_{K,\delta}$ of densities with respect to $\mu$ and the $\rho$-estimator $\hat{P}_\delta$ of $\overline{P}$ based on the family $\mathcal{Q}_{K,\delta}$. We recall that $\hat{P}_\delta$ is defined as follows. Given

$$\psi : \left| \begin{array}{ccc} [0,+\infty] & \to & [-1,1] \\ x & \mapsto & \frac{x-1}{x+1} \end{array} \right. , \tag{2.6}$$

we set for $\mathbf{x} = (x_1,\ldots,x_n) \in \mathscr{X}^n$ and $q,q' \in \mathcal{Q}_{K,\delta}$

$$\mathbf{T}(\mathbf{x},q,q') := \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q'(x_i)}{q(x_i)}}\right), \tag{2.7}$$

with the convention $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$, and

$$\boldsymbol{\Upsilon}(\mathbf{X},q) := \sup_{q'\in\mathcal{Q}_{K,\delta}} \mathbf{T}(\mathbf{X},q,q'). \tag{2.8}$$

The $\rho$-estimator $\hat{P}_\delta$ is any measurable element of the closure (with respect to the Hellinger distance) of the set

$$\boldsymbol{\mathscr{E}}(\psi,\mathbf{X}) := \left\{ Q = q\cdot\mu; q \in \mathcal{Q}_{K,\delta}, \boldsymbol{\Upsilon}(\mathbf{X},q) < \inf_{q'\in\mathcal{Q}_{K,\delta}} \boldsymbol{\Upsilon}(\mathbf{X},q') + 11.36 \right\}. \tag{2.9}$$

This construction follows [11] and the constant 11.36 is given by (7) in [11]. This constant does not play an essential role and can be replaced by any smaller positive number. Ideally, one would take an estimator that achieves the infimum but it might happen that no minimizer exists. Using (2.9) allows us to avoid this problem without significantly deteriorating the deviation bounds we obtain for our estimator.

As explained earlier, we only focus on the theoretical aspects in this paper. Although $\rho$-estimators have been developed to obtain theoretical rather than computational properties, it is possible to actually compute the estimators in practice for some models and to run simulations, as in Baraud & Chen [12] (Section 5).

### 2.3.2 The performance of the estimator

The following result holds.

**Theorem 2.1.** *Let $\delta \in (0,1/K]$ and $\xi > 0$. Assume that Assumptions 2.1 and 2.2 hold and set $\overline{V} = V_1 + \cdots + V_K$. Any $\rho$-estimator $\hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}$ satisfies with probability at least $1 - e^{-\xi}$,*

$$\begin{aligned}\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq c_0 \left[\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + n(K-1)\delta\right] \\ + c_1 116.1 \overline{V}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}}\right)\right] \\ + c_1(1.49 + \xi),\end{aligned} \tag{2.10}$$

*where $c_0 = 300$, $c_1 = 5014$. In particular, for the choice $\delta = \frac{\overline{V}}{n(K-1)} \wedge \frac{1}{K}$, the resulting estimator $\hat{P} = \hat{P}_\delta$ satisfies*

$$C\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) \leq \mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}_K) + \overline{V}\left[1 + \log\left(\frac{Kn}{\overline{V} \wedge n}\right)\right] + \xi, \tag{2.11}$$

*with probability at least $1 - e^{-\xi}$, where $C$ is a positive universal constant.*

The proof of the theorem is postponed to Section 2.B.2. One can notice that the bound we obtain does not depend on the space $\mathscr{X}$, e.g. on the dimension $d$ in the case $\mathscr{X} = \mathbb{R}^d$, but only on the VC-indices $V_1, \ldots, V_K$ and on $\delta$. Inequality (2.10) shows the influence of the choice of the parameter $\delta$ on the performance of the estimator $\hat{P}_\delta$. Hereafter, we shall choose $\delta$ as in the second part of Theorem 2.1 and therefore only comment on inequality (2.11). Given $\overline{P}$ in $\mathscr{Q}_K$, it follows from the triangle inequality and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for all non-negative numbers $a$ and $b$, that

$$nh^2\left(\overline{P}, \hat{P}\right) = \mathbf{h}^2\left(\overline{P}^{\otimes n}, \hat{P}^{\otimes n}\right) \leq 2\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) + 2\mathbf{h}^2\left(\mathbf{P}^*, \overline{P}^{\otimes n}\right).$$

We immediately derive from (2.11) that on a set of probability at least $1 - e^{-\xi}$

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \frac{1}{n}\sum_{i=1}^{n} h^2(P_i^*, \overline{P}) + \frac{\overline{V}\log\left(Kn/\overline{V}\right) + \xi}{n}. \tag{2.12}$$

In the ideal situation where the observations are i.i.d. with common distribution $\overline{P} \in \mathscr{Q}_K$, we obtain that

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \frac{\overline{V}\log\left(Kn/\overline{V}\right) + \xi}{n}.$$

Integrating this result with respect to $\xi$ and the fact that $\overline{P}$ is arbitrary in $\mathscr{Q}_K$ leads to the uniform risk bound

$$\sup_{\overline{P}\in\mathscr{Q}_K} \mathbb{E}\left[h^2\left(\overline{P},\hat{P}\right)\right] \leq C'\frac{\overline{V}\log\left(Kn/\overline{V}\right)}{n}, \tag{2.13}$$

where $C'$ is a positive universal constant. This means that up to a logarithmic factor, the estimator $\hat{P}$ uniformly converges over $\mathscr{Q}_K$ at the rate $1/\sqrt{n}$ with respect to the Hellinger distance. One knows that when working with the Hellinger distance, no estimator can do better that this $1/\sqrt{n}$ rate (see (1.1) in [15]).

We can see that we only need to bound the quantity $\overline{V}$ to deduce deviation inequalities in specific cases. Therefore, we can already get a bound on the convergence rate for Gaussian mixtures with Lemma 2.1.

**Corollary 2.1.**

- *Let $\mathscr{Q}_K$ be the Gaussian location-scale mixture model, i.e. $\overline{\mathscr{F}}_1 = \cdots = \overline{\mathscr{F}}_K = \left\{\mathcal{N}(\mu,\Sigma);\mu\in\mathbb{R}^d,\Sigma\in Cov_{+*}(d)\right\}$. There is a positive universal constant $C>0$ such that, for any $\rho$-estimator $\hat{P}=\hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}$, for all $\overline{P}\in\mathscr{Q}_K$ and for all $\xi>0$, we have*

$$Ch^2\left(\overline{P},\hat{P}\right) \leq \frac{Kd^2\left[1+\log\left(\frac{n}{d^2}\vee K\right)\right]+\xi}{n},$$

*with probability at least $1-e^{-\xi}$.*

- *Let $\mathscr{Q}_K$ be the Gaussian location mixture model associated to a fixed covariance matrix $\Sigma\in Cov_{+*}(d)$, i.e. $\overline{\mathscr{F}}_1 = \cdots = \overline{\mathscr{F}}_K = \left\{\mathcal{N}(\mu,\Sigma);\mu\in\mathbb{R}^d\right\}$. There is a positive universal constant $C>0$ such that, for any $\rho$-estimator $\hat{P}=\hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}$, for all $\overline{P}\in\mathscr{Q}_K$ and for all $\xi>0$, we have*

$$Ch^2\left(\overline{P},\hat{P}\right) \leq \frac{Kd\left[1+\log\left(\frac{n}{d}\vee K\right)\right]+\xi}{n},$$

*with probability at least $1-e^{-\xi}$.*

Those rates would be optimal if the logarithmic factor was necessary. Doss *et al.* [30] proved it is not the case for Gaussian location mixtures with known isotropic covariance matrix. They provide an estimator that achieves the minimax rate $\sqrt{d/n}$ with respect to the Hellinger distance. However, the dependency in $K$ of their bound in (1.12) is worse than exponential when it is just linear for our estimator.

Our assumption that the families of density functions $\overline{\mathcal{F}}_k$ are VC-subgraph is actually weak since it includes situations where these models consist of unbounded densities or densities which are not in $L_2$ which to our knowledge have never been considered in the literature. A concrete example of such situations is the following one. Let $g$ be some non-increasing function on $(0,+\infty)$ which is unbounded, nonnegative and satisfies $\int_0^{+\infty} g(x)dx = \frac{1}{2}$ and $\overline{\mathscr{F}}_k$ is the translation model associated to the family of densities $\left\{x\mapsto g(|x-\theta|)\mathbb{1}_{|x-\theta|>0};\theta\in\mathbb{R}\right\}$ for all $k\in\{1,\ldots,K\}$. It follows from Proposition 42-(vi) of Baraud *et al.* [9] that the VC-index of $\overline{\mathcal{F}}_k$ is not larger than 10.

When the data are independent but not i.i.d., we derive from inequality (2.12) that the estimator $\hat{P}$ performs almost as well as in the i.i.d. case as long as the marginals $P_1^*,\ldots,P_n^*$ are close enough to $\overline{P}$. This means that the estimator is robust with respect to a possible misspecification of the model and the departure from the assumption that the data are i.i.d. In particular, this includes the situations where the dataset contains some outliers or has been contaminated. Consider Hüber's contamination model where a proportion $\epsilon$ of the data is contaminated, i.e. we have $P^* = (1-\epsilon)\overline{P}+\epsilon Q$, where $\overline{P}$ is the probability distribution we want to estimate and $Q$ is the distribution of the contaminated data. In this situation, for any

probability distribution $Q$, using (2.12) and the convexity property of the Hellinger distance we get

$$Ch^2\left(\overline{P},\hat{P}\right) \leq \epsilon + \frac{\overline{V}\log(n)+\xi}{n}. \tag{2.14}$$

We can see that there is no perturbation of the convergence rate as long as the contamination rate $\epsilon$ remains small as compared to $\overline{V}\log(n)/n$. Contrary to other loss functions, the Hellinger distance does not allow to obtain a better rate than $\sqrt{\epsilon}$ in the general case (see Birgé [17]). Inequality (2.18), stated later, also allows to consider misspecification for the emission models for example.

### 2.3.3 The case of totally bounded emission models

We might also consider emission models for which we do not have any bound on the VC-index. For a subset $\mathscr{N}$ of $\mathscr{P}$ and $\eta \in [0,1]$, the $\eta$-covering number $N(\eta,\mathscr{N},h)$ of $\mathscr{N}$, with respect to the Hellinger distance, is the minimum number of balls $\mathcal{B}_h(P_i,\eta)$, $i=1,\ldots,N$, necessary to cover $\mathscr{N}$. In that case, the set $\mathscr{N}[\eta]=\{P_i; i=1,\ldots,N\}$ constitutes a finite approximation of $\mathscr{N}$, i.e. for all $Q$ in $\mathscr{N}$ there exists $i\in\{1,\ldots,N\}$ such that $h(Q,P_i)\leq\eta$. We say that $\mathcal{N}$ is totally bounded (for the Hellinger distance) if its $\eta$-covering number is finite for all $\eta\in(0,1]$. A direct consequence of the definition of VC-subgraph classes is that any finite set $\mathcal{F}$ of real-valued functions is VC-subgraph with VC-index at most $V(\mathcal{F})\leq\log_2(|\mathcal{F}|)$. Consequently, we can still use $\rho$-estimation for models that are not proven to satisfy Assumption 2.2 but still are such that emission models are totally bounded.

**Theorem 2.2.** *Let $\overline{\mathscr{F}}_k$ be a totally bounded class of distributions for all $k\in\{1,\ldots,K\}$ with $K\geq 2$. Let $\mathscr{Q}_K$ be the mixture model defined by*

$$\mathscr{Q}_K=\left\{\sum_{k=1}^K w_k F_k; w\in\mathcal{W}_K, F_k\in\overline{\mathscr{F}}_k, \forall k\in\{1,\ldots,K\}\right\}.$$

*Assume there are constants $A_k\geq 1$ and $\alpha_k$ such that $\log_2 N(\eta,\mathscr{F}_k,h)\leq\left(\frac{A_k}{\eta}\right)^{\alpha_k}$ for all $k$ in $[K]$ and for all $\eta\in(0,1)$. Let $\epsilon$ be in $(0,1)$. For $k$ in $[K]$, let $\mathscr{F}_k[\epsilon]$ be a minimal $\epsilon$-net of $\overline{\mathscr{F}}_k$ such that $|\mathscr{F}_k[\epsilon]|=N(\epsilon,\mathscr{F}_k,h)$. Let $\mathscr{Q}_{K,\delta}[\epsilon]$ be the countable model defined by*

$$\mathscr{Q}_{K,\delta}[\epsilon]=\{P_{w,F}; w\in\mathcal{W}_K, w_k\geq\delta, w_k\in\mathbb{Q}, F_k\in\mathscr{F}_k[\epsilon], \forall k\in\{1,\ldots,K\}\}.$$

*Take $\delta=\frac{\overline{V}}{n(K-1)}\wedge\frac{1}{K}$ with*

$$\overline{V}=\sum_{k=1}^K\log_2(|\mathscr{F}_k[\epsilon]|)\leq\sum_{k=1}^K\left(\frac{A_k}{\epsilon}\right)^{\alpha_k},$$

*where $\epsilon=n^{-\frac{1}{\alpha_{\max}+2}}$ and $\alpha_{\max}=\max_{1\leq k\leq K}\alpha_k$. There exists a positive constant $C$ such that for any $\rho$-estimator $\hat{P}=\hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}[\epsilon]$, for all $\xi>0$, we have*

$$C\mathbf{h}^2\left(\mathbf{P}^*,\left(\hat{P}_\delta\right)^{\otimes n}\right)\leq\mathbf{h}^2\left(\mathbf{P}^*,\mathscr{Q}_K\right)+n^{\frac{\alpha_{\max}}{\alpha_{\max}+2}}\sum_{k=1}^K A_k^{\alpha_k}\left[1+\log(Kn)\right]+\xi,$$

*with probability at least $1-e^{-\xi}$. In particular, if the observations are i.i.d. with common distribution $P^*\in\mathscr{P}$ we have*

$$Ch^2\left(P^*,\hat{P}_\delta\right)\leq h^2\left(\overline{P},\mathscr{Q}_K\right)+n^{-\frac{2}{\alpha_{\max}+2}}\sum_{k=1}^K A_k^{\alpha_k}\left[1+\log(Kn)\right]+\frac{\xi}{n},$$

This theorem is proved in Section 2.B.3 (page 60) and we illustrate it with the following example. Doss & Wellner [29] provide a bound on the entropy for classes of log-concave and $s$-concave densities. Let $\mathcal{C} = \{\varphi : \mathbb{R} \to [-\infty,\infty); \varphi \text{ is a closed, proper concave function}\}$ where *proper* and *closed* are defined in [76] (Sections 4 and 7). For $0 < M < \infty$ and $s > -1$, let $\mathcal{P}_{M,s}$ be the class of densities defined by

$$\mathcal{P}_{M,s} = \left\{ p \in \mathcal{P}_s; \sup_{x\in\mathbb{R}} p(x) \leq M, 1/M \leq p(x) \text{ for all } |x| \leq 1 \right\},$$

where $\mathcal{P}_s = \{p : \int p d\lambda = 1\} \bigcap h_s \circ \mathcal{C}$, $\lambda$ is the Lebesgue measure on $\mathbb{R}$ and $h_s : \mathbb{R} \to \mathbb{R}$ is defined by

$$h_s(y) = \begin{cases} e^y, & s = 0 \\ (-y)_+^{1/s}, & s \in (-1,0), \\ y_+^{1/s}, & s > 0. \end{cases}$$

We fix such values of $M$ and $s$. Let $\mathcal{Q}_K$ be the density model of mixtures of $s$-concave densities (or log-concave for $s = 0$) defined by

$$\mathcal{Q}_K = \left\{ \sum_{k=1}^{K} w_k f_k; w \in \mathcal{W}_K, f_k \in \mathcal{P}_{M,s} \right\},$$

with $K \geq 2$. Let $\mathscr{Q}_K$ be the class of distributions associated to $\mathcal{Q}_K$. The class $\mathcal{P}_{M,s}$ is not proven to be VC-subgraph but it is totally bounded. As a direct consequence of Theorem 3.1 of Doss & Wellner [29], there exists a positive constant $A$, depending only on $M$ and $s$, such that for all $\epsilon$ in (0,1], we have

$$\log_2 N(\epsilon,\mathcal{P}_{M,s},h) \leq A\epsilon^{-1/2}.$$

In particular, it means there exists a $\epsilon$-net $\mathcal{P}_{M,s}[\epsilon]$ such that $\log_2 (|\mathcal{P}_{M,s}[\epsilon]|) \leq (A^2/\epsilon)^{1/2}$. Let $\mathcal{Q}_{K,\delta}[\epsilon]$ be the countable density model given by

$$\mathcal{Q}_{K,\delta}[\epsilon] = \left\{ \sum_{k=1}^{K} w_k f_k; w \in \mathcal{W}_K, w_k \geq \delta, w_k \in \mathbb{Q}, f_k \in \mathcal{P}_{M,s}[\epsilon] \right\}.$$

One can check that $\mathscr{Q}_{K,\delta}[\epsilon]$ is also a $\epsilon$-net of $\mathscr{Q}_{K,\delta}$ with respect to the Hellinger distance using inequality (2.18) hereafter page 38. The application of Theorem 2.2 on this example gives the following result.

**Corollary 2.2.** *Assume there exists $P^*$ in $\mathscr{P}$ such that $\mathbf{P}^* = (P^*)^{\otimes n}$. Take $\epsilon = n^{-2/5}$ and $\delta = n^{-4/5} \wedge K^{-1}$. Let $\hat{P} = \hat{P}_\delta$ be a $\rho$-estimator on $\mathscr{Q}_{K,\delta}[\epsilon]$. There exists a constant $C(M,s)$ such that for all $\xi > 0$, we have*

$$C(M,s)h^2\left(P^*,\hat{P}\right) \leq h^2\left(P^*,\mathscr{Q}_K\right) + \frac{K}{n^{4/5}}\left[1 + \log\left(Kn\right)\right] + \frac{\xi}{n},$$

*with probability at least $1 - e^{-\xi}$.*

This corollary provides a risk bound over the class of distributions associated to mixtures of $s$-concave densities. Up to a logarithmic factor, the estimator $\hat{P}$ uniformly converges over $\mathscr{Q}_K$ at the rate $n^{-2/5}$ with respect to the Hellinger distance, which is the same rate given in Theorem 3.2 of Doss & Wellner [29] for the MLE over the model $\mathcal{P}_{M,s}$, i.e. for $K = 1$.

### 2.3.4 Application to the estimation of a general Gaussian mixture

We denote by $\phi_\sigma$ the density function of the normal distribution (with respect to the Lebesgue measure on $\mathbb{R}$) with mean 0 and variance $\sigma^2 > 0$, i.e.

$$\phi_\sigma : x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{2.15}$$

We assume $P^*$ is of the following form or is close enough to a distribution of the form

$$p_H(x) = \int \phi_\sigma(x-z) dH(z,\sigma), \forall x \in \mathbb{R}.$$

We say that $p_H$ is the Gaussian mixture density with mixing distribution $H$. We want to approximate any distribution of this form with finite Gaussian mixtures, i.e. distribution with densities of the same form with mixing distribution supported on a finite set. For a mixing measure $H$ on $\mathbb{R} \times \mathbb{R}^{+*}$, we denote by $\mathrm{supp}(H)$ its support. To obtain an approximation result, we need to consider mixing measures $H$ that are supported on a compact set, i.e. there exist $A \geq 0$ and $R \geq 1$ such that $\mathrm{supp}(H) \subset [-A,A] \times [1,R]$. The Hellinger distance being invariant to translation and rescaling, we consider the following class of densities. For $A > 0$ and $R \geq 1$ we define

$$\mathcal{C}(A,R) = \left\{ p_H ; \exists l \in \mathbb{R}, \exists s > 0, \mathrm{supp}(H) \subset [l - sA, l + sA] \times [s, sR] \right\}$$

and we denote by $\mathscr{C}(A,R)$ the associated class of distributions. We denote by $\mathscr{G}_{mix,K}$ the Gaussian mixture model with $K$ components associated to the class of densities $\mathcal{G}_{mix,K}$ defined by

$$\mathcal{G}_{mix,K} := \left\{ \sum_{k=1}^{K} w_k \phi_{\sigma_k}(\cdot - z_k) ; w \in \mathcal{W}_K, \sigma_k \in (0, +\infty), z_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\} \right\}. \tag{2.16}$$

This situation corresponds to $\overline{\mathcal{F}}_k = \mathcal{G}_1$ for all $k \in \{1, \ldots, K\}$. We can approximate the class $\mathscr{C}(A,R)$ with the model $\mathscr{G}_{mix,K}$ as indicated by the following result.

**Proposition 2.1.** *For $K \geq 2\left(24A^2 + 1\right)^2$, we have*

$$\sup_{P_H \in \mathscr{C}(A,R)} h^2\left(P_H, \mathscr{G}_{mix,K}\right) \leq \frac{1}{2} \exp\left( -\frac{K^{1/2}}{12\sqrt{6}R^2} \right) \left[ K^{1/4} \frac{3\sqrt{2}}{\sqrt{e\pi} 7^{1/4}} + R \right].$$

This proposition allows us to obtain a deviation bound on the estimation over $\mathscr{C}(A,R)$, with Theorem 2.1. Its proof is postponed to Section 2.C.2.

**Theorem 2.3.** *For $R \geq 1$ and $n \geq e$, we take $K = K(R,n) := \lceil 864 R^4 \log^2(n) \rceil$. Let $\hat{P}$ be a $\rho$-estimator on $\mathscr{G}_{K,\delta}$ with $\delta$ as in (2.11) and assume the true distribution is i.i.d., i.e. $\mathbf{P}^* = (P^*)^{\otimes n}$. There exists a numeric constant $C > 0$, hence not depending on $R$, such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have*

$$C h^2\left(P^*, \hat{P}\right) \leq h^2\left(P^*, \mathscr{C}(A,R)\right) + \frac{R^4 \log^3(n) + \xi}{n}, \tag{2.17}$$

*for $A = A(R,n) := \sqrt{\frac{12\sqrt{3}-1}{24}} R \log^{1/2}(n)$.*

This result is proven in Section 2.C.1. Therefore, for a fixed $R$, we obtain a rate of $\log^{3/2}(n)/\sqrt{n}$ over $\mathscr{C}(\infty,R) := \bigcup_{A>0} \mathscr{C}(A,R)$ with respect to the Hellinger distance. We can also consider larger classes of distributions, with $R$ increasing as $n$ increases but it would deteriorate this rate. Our result is still an improvement of Theorem 4.2 from [49] as it requires weaker assumptions. Their result is sensitive to translation or scaling and they have to specify bounds $0 < \underline{\sigma} < \overline{\sigma}$ in the model such that $H^*$ is supported on a compact set $[-a,a] \times [\underline{\sigma},\overline{\sigma}]$. Moreover, our estimator is robust, to contamination for instance. Assume we have an $\epsilon$ contamination rate of our data, i.e. $P^*$ is of the form $P^* = (1-\epsilon)\overline{P} + \epsilon Q$ with $\epsilon \in (0,1)$, $\overline{P} \in \mathscr{C}(A(R,n),R)$ and $Q$ is any probability distribution. Then, our estimator satisfies $Ch^2(P^*,\hat{P}) \leq \epsilon + \frac{R^4 \log^3(n)+\xi}{n}$ on an event of probability $1 - e^{-\xi}$. As long as $\epsilon$ remains small as compared to $R^4 \log^3(n)/n$, the rate is not deteriorated by the contamination.

### 2.3.5 Parameter estimation

We say that $\hat{w}$ and $\hat{F}$ are $\rho$-estimators if the resulting mixture distribution $\hat{P}$ given by

$$\hat{P} = \sum_{k=1}^{K} \hat{w}_k \hat{F}_k$$

is a $\rho$-estimator. We have a general result for the performance of $\hat{P}$ but not for $\hat{w}$ and $\hat{F}$. Hopefully we those parameter estimators would inherit the properties of $\hat{P}$ under additional assumptions. Some results about the robust estimation of parameters exist in the machine learning community, see Diakonikolas *et al.* [28] for instance. As before, the available results are all restricted to specific cases such as Gaussian mixture models. Convexity properties ensure that we always have the upper bound

$$h\left(P_{w,F}, P_{v,G}\right) \leq \inf_{\tau \in \mathscr{S}_K} \left\{ h(w, v \circ \tau) + \max_{k \in [K]} h\left(F_k, G_{\tau(k)}\right) \right\}, \tag{2.18}$$

for all mixing weights and emission distributions (see Lemma 2.8), where $\mathscr{S}_K$ denotes the set of all permutations of $[K]$ and $\mathcal{W}_K$ is seen as the set of probability distributions on $[K]$ and justify the notation $h(w,v \circ \tau)$. Therefore, a good estimation of the mixing weights $w = (w_1,\ldots,w_K)$ and of the emission distributions $F = (F_1,\ldots,F_K)$ ensures a good estimation of the mixture distributions $P_{w,F}$. However the converse is not true as the parameters are not even identifiable in general.

**Example 2.2.** *Let $\overline{\mathscr{F}}$ be the set of uniform distributions $\mathcal{U}(a,b)$ the uniform distribution on the interval $(a,b)$ of positive lengths. Then the parameters $w$ and $F$ in the mixture model*

$$\mathscr{Q}_2 = \left\{ w_1 F_1 + (1-w_1)F_2; w_1 \in (0,1), F_1,F_2 \in \overline{F} \right\}$$

*are not identifiable since*

$$\frac{3}{4}\mathcal{U}(0,1) + \frac{1}{4}\mathcal{U}(1/3,2/3) = \frac{1}{2}\mathcal{U}(0,2/3) + \frac{1}{2}\mathcal{U}(1/3,1).$$

We shall say that $P = P_{w,F}$ is identifiable (with respect to the model) if for all $v$ in $\mathcal{W}_K$ and all $G$ in $\mathscr{F}_1 \times \cdots \times \mathscr{F}_K$, we have

$$P_{w,F} = P_{v,G} \Rightarrow \exists \tau \in \mathscr{S}_K, \forall k \in [K], w_k = v_{\tau(k)} \text{ and } F_k = G_{\tau(k)},$$

There is a wide literature about identifiability that includes the works of Teicher [82] and Sapatinas [79] for example. Allman *et al.* [4] provides identifiability conditions in a nonparametric

framework but this is quite unusual. In this section, we will consider a unique parametric model for the emission models, i.e. we have $\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K \subset \{F_\theta;\theta\in\Theta\}$ with $\Theta\subset\mathbb{R}^d$ and assume $P_{w^*,F_{\theta^*}}$ is the true distribution or the best approximation within the model. Identifiability is a minimum requirement for the parameter estimators to be meaningful but we can hardly get more than consistency with it.

There is one approach that allows not to consider the identifiability issue is to consider the estimation of the mixing distribution instead of the parameters themselves, i.e. $w_1^*\delta_{\theta_1^*}+\cdots+w_K^*\delta_{\theta_K^*}$ where $\delta_x$ is the Dirac measure in $x$. Most results are given for the $L_1$-Wassertein metric $W_1$ which can be defined as follow for $\Theta\subset\mathbb{R}$. For probability distributions $G_1$,$G_2$ on $\Theta$, we have

$$W_1(G_1,G_2) := \sup_{f\in Lip(1)}\int_\Theta f(dG_1-dG_2), \tag{2.19}$$

where $Lip(1)$ is the class of Lipschitz functions with Lipschitz constant at most 1. Heinrich & Kahn [50] establish minimax rates of estimation for mixing distribution under some regularity and strong identifiability conditions. Wu & Yang [87] prove that their denoised method of moments for univariate Gaussian mixtures provides an estimator of the mixing distribution that reaches the optimal rate with respect to $W_1$. They also prove an oracle bound for density estimation in the case of misspecification similar to (2.14), for the total variation distance instead of the Hellinger distance. However, they only consider misspecified distributions that are sub-Gaussian and in dimension one.

Our approach is to derive bounds on the convergence rates for the parameter estimators from (2.11). Typically, we are looking for an inequality of the form

$$h(P_{w^*,F_{\theta^*}},P_{w,F_\theta}) \geq C(w^*,\theta^*)\left[\sum_{k=1}^K d_\Theta(\theta_k^*,\theta_k)+d_{\mathcal{W}}(w^*,w)\right],\forall w\in\mathcal{W}_K,\forall\theta\in\Theta, \tag{2.20}$$

where $C(w^*,\theta^*)$ is positive, $d_\Theta$ is a distance on $\Theta$ and $d_{\mathcal{W}}$ is a distance on $\mathcal{W}_K$. Intuitively, if we can estimate each parameter individually we should be able to estimate the mixing distribution as well. Formally, for $\Theta\subset\mathbb{R}$, we have

$$W_1\left(\sum_{k=1}^K w_k^*\delta_{\theta_k^*},\sum_{k=1}^K w_k\delta_{\theta_k}\right)\leq\sum_{k=1}^K|\theta_k^*-\theta_k|+\max_i|\theta_i^*|\cdot\sum_{k=1}^K|w_k^*-w_k|,\forall w\in\mathcal{W}_K,\forall\theta\in\Theta^K,$$

which is a direct consequence of (2.19). One can see that when $d_\Theta$ and $d_{\mathcal{W}_K}$ in (2.20) are the $L_1$ distance we can deduce a bound for the estimation of the mixing distribution. The main difficulty remains to obtain a lower bound on the Hellinger distance between mixtures. There are still some situations where we do have such a lower bound.

### Regular parametric model

Let $K$ be an integer larger than 1. We consider parametric emission models associated to density models $\overline{\mathcal{F}}_k=\{f_k(\cdot;\alpha),\alpha\in A_k\}$, where $A_k$ is a subset of $\mathbb{R}^{d_k}$ for all $k\in\{1,\ldots,K\}$. It is always possible to find a countable dense subset of $A_k$ with respect to the Euclidean distance on $\mathbb{R}^{d_k}$. We assume there is a reasonably good connection between the Hellinger distance on the emission models and the Euclidean distances on the parameter spaces such that a dense subset of $A_k$ would translate into a dense subset of the emission model with respect to the Hellinger distance. This assumption is very weak and does not seem to be restrictive in any way. In the different examples we consider we can always consider $A_k\cap\mathbb{Q}^{d_k}$ as a dense subset of $A_k$. Therefore Assumption 2.1 is satisfied with $\mathcal{F}_k=\{f_k(\cdot;\alpha),\alpha\in B_k\}$. We denote by $\mathscr{Q}_K$ the distribution model associated to the mixture density model

$$\mathcal{Q}_K=\left\{p(\cdot;\theta)=\sum_{k=1}^{K-1}w_kf_k(\cdot;z_k)+(1-w_1-\cdots-w_{K-1})f_K(\cdot;\alpha_K);\theta=(w,\alpha)\in\Theta\right\},$$

where $\Theta$ is an open convex subset of $\left\{w \in (0,1)^{K-1}; \sum_{k=1}^{K-1} w_k < 1\right\} \times A_1 \times \cdots \times A_K$. To be in the context of regular parametric models consider by Ibragimov & Has'minskiĭ [54] we need to make some assumptions.

**Assumption 2.3.** *The classes of functions $\overline{\mathcal{F}}_1, \ldots, \overline{\mathcal{F}}_K$ satisfy the following regularity conditions.*

*a) The function $z \mapsto f_k(x;z)$ is continuous on $A_k$ (with respect to the Euclidean distance) for $\mu$-almost all $x \in \mathcal{X}$, for all $k \in \{1, \ldots, K\}$.*

*b) For all $k \in \{1, \ldots, K\}$, for $\mu$-almost all $x \in \mathcal{X}$ the function $u \mapsto f_k(x;u)$ is differentiable at the point $u = \alpha$ and for all $j \in \{1, \ldots, d_k\}$, we have*

$$\int_{\mathcal{X}} \left| \frac{\partial f_k(x;\alpha)}{\partial \alpha_j} \right|^2 \frac{\mu(dx)}{f_k(x;\alpha)} < \infty.$$

*c) The function $\theta \mapsto \psi(\cdot;\theta) = \frac{\partial}{\partial \theta} p^{1/2}(\cdot;\theta)$ is continuous in the space $L_2(\mu)$.*

*d) The class of densities $\overline{\mathcal{F}}_k$ is VC-subgraph with VC-index not larger than $V_k$ for all $k \in \{1, \ldots, K\}$. We write $\overline{V} = V_1 = \cdots + V_k$.*

The work of Ibragimov & Has'minskiĭ [54] allows to derive a deviation inequality on the Euclidean distance between parameters using Fisher's information.

**Theorem 2.4.** *Let $\overline{\theta}$ be in $\Theta$. Assume the Fisher's information matrix*

$$I\left(\overline{\theta}\right) = \int_{\mathcal{X}} \frac{\partial p\left(x;\overline{\theta}\right)}{\partial \theta} \left( \frac{\partial p\left(x;\overline{\theta}\right)}{\partial \theta} \right)^T \frac{\mu(dx)}{p\left(x;\overline{\theta}\right)}$$

*is definite positive and $\inf_{\substack{||\overline{\theta}-\theta|| \geq a \\ \theta \in \Theta}} h^2\left(P_{\overline{\theta}}, P_\theta\right) > 0$ for all $a > 0$. Let $\hat{P} = P_{\hat{w},\hat{F}}$ be a $\rho$-estimator on $\mathcal{Q}_{K,\delta}$, with $\delta$ as in (2.11). There exists a positive constant $C\left(\overline{\theta}\right)$ such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have*

$$C\left(\overline{\theta}\right)\left( ||\overline{w} - \hat{w}||^2 + \sum_{k=1}^{K} 1 \wedge ||\overline{\alpha}_k - \hat{\alpha}_k||^2 \right) \leq \frac{1}{n}\left[ \mathbf{h}^2\left(\mathbf{P}^*, P_{\overline{\theta}}^{\otimes n}\right) + \overline{V}\log(n) + \xi \right]. \tag{2.21}$$

*And assuming $P^* = P_{\overline{\theta}}$, we obtain the usual parametric convergence rate up to a logarithmic factor for the parameter estimators.*

This result is proven in Section 2.D.1. Following the proof and Theorem 7.6 [54], the best constant $C\left(\overline{\theta}\right)$ depends on the smallest eigenvalue of the Fisher's information matrix $I\left(\overline{\theta}\right)$ and the geometry induced by the Hellinger distance around $\overline{\theta}$ in $\Theta$. Inequality (2.21) proves that even if "true parameters" might not exist the parameter estimators can be meaningful as long as $\mathbf{P}^*$ is relatively close to the model. The Gaussian mixture model is the most common mixture model and it is a regular parametric model. Let $K \geq 2$ and for all $k$ in $[K]$ take $\mathcal{F}_k = \mathcal{G}_1$, where $\mathcal{G}_1$ is given in Lemma 2.1. We define a binary relation on $\mathbb{R} \times (0,\infty)$ by

$$(z_1,\sigma_1) > (z_2,\sigma_2) \Leftrightarrow \begin{cases} \sigma_1 > \sigma_2; \\ \text{or } \sigma_1 = \sigma_2 \text{ and } z_1 > z_2. \end{cases} \tag{2.22}$$

We consider the parameters $\theta = (w_1, \ldots, w_{K-1}, z_1, \sigma_1^2, \ldots, z_K, \sigma_K^2)$ belonging to the set

$$\Theta = \left\{ \theta \in (0,1)^{K-1} \times \left(\mathbb{R} \times \mathbb{R}^*\right)^K ; \sum_{k=1}^{K-1} w_k < 1, (z_1,\sigma_1) > \cdots > (z_K,\sigma_K) \right\}.$$

**Theorem 2.5.** *Assume* $P^* = P_{\overline{\theta}} = \sum_{k=1}^{K} \overline{w}_k \mathcal{N}(\overline{z}_k, \overline{\sigma}_k^2)$ *such that* $(\overline{z}_1, \overline{\sigma}_1) > \cdots > (z_K, \sigma_K)$ *are all distinct and* $\inf_{1 \le k \le K} \overline{w}_k > 0$. *Let* $\hat{P}$ *be a* $\rho$*-estimator on* $\mathcal{S}_{K,\delta}$, *with* $\delta$ *as in (2.11). There exists a positive constant* $C\left(\overline{\theta}\right)$ *such that, for all* $\xi > 0$, *we have*

$$C\left(\overline{\theta}\right)\left(\sum_{k=1}^{K-1} ||\overline{w}_k - \hat{w}_k||^2 + \sum_{k=1}^{K} \left|\left|\left(\overline{z}_k, \overline{\sigma}_k^2\right) - \left(\hat{z}_k, \hat{\sigma}_k^2\right)\right|\right|^2 \wedge 1\right) \le \frac{5K\log(n) + \xi}{n}, \tag{2.23}$$

*with probability at least* $1 - e^{-\xi}$.

This result is proven in Section 2.D.3. Our estimator reaches the optimal rate of convergence up to a logarithmic factor. One can notice that the assumption of ordered couples of parameters $(z_j, \sigma_j^2)$ can be replaced by considering distinct couples only and taking the infimum over permutation of the hidden states in (2.23).

### Connection with the $L_2$-distance

We can use results from the literature that do not apply to the Hellinger distance but to other ones such as the $L_2$-distance between densities. There is a general inequality between the $L_2$ and Hellinger distances when the density functions are bounded, i.e.

$$||p - q||_2^2 \le 4\left(||p||_\infty + ||q||_\infty\right) h^2(P,Q). \tag{2.24}$$

Assume one can prove an inequality of the following type. For any $w, v$ in $\mathcal{W}_K$ and any $f_k, g_k$ in $\overline{\mathcal{F}}_k$ for all $k \in \{1, \ldots, K\}$ such that the resulting mixtures belong to our model, we have

$$\underline{c}\left(d_{\mathcal{W}}^2(w,v) + \max_{k \in [K]} d_F^2(f_k, g_k)\right) \le \left|\left|\sum_{k=1}^{K} w_k f_k - \sum_{k=1}^{K} v_k g_k\right|\right|_2^2, \tag{2.25}$$

where $d_{\mathcal{W}}$ is a distance on $\mathcal{W}_K$ and $d_F$ is a distance on $\bigcup_{1 \le k \le K} \overline{\mathcal{F}}_k$. Moreover, assuming the density models $\overline{\mathcal{F}}_k$ are uniformly bounded, i.e.

$$\sup_{k \in [K]} \sup_{f \in \overline{\mathcal{F}}_k} ||f||_\infty =: U < \infty, \tag{2.26}$$

we get

$$d_{\mathcal{W}}^2(w,v) + \max_{k \in [K]} d_F^2(f_k, g_{\tau(k)}) \le \frac{8U}{\underline{c}} h^2\left(\sum_{k=1}^{K} w_k F_k, \sum_{k=1}^{K} v_k G_k\right).$$

Here again, a density estimation result implies a result for the parameter estimation. We can apply this method to the special case of a two-component mixture model with one known component. Let $\phi$ be a density function on $\mathbb{R}^d$ with respect to the Lebesgue measure. We consider the 2-component mixture model $\mathcal{Q}$ associated to the class of densities

$$\mathcal{Q} = \left\{x \mapsto p_{w,z}(x) = (1-w)\phi(x) + w\phi(x-z); w \in [0,1], z \in \mathbb{R}^d\right\}, \tag{2.27}$$

with $\overline{\mathcal{F}}_1 = \{\phi\}$ and $\overline{\mathcal{F}}_2 = \left\{x \mapsto \phi(x-z); z \in \mathbb{R}^d\right\}$. Gadat *et al.* [41] proved an inequality such as (2.25) in this situation. They still require the following assumptions on $\phi$.

**Assumption 2.4.** *The function* $\phi$ *belongs to* $\mathcal{C}^3\left(\mathbb{R}^d\right) \cap \mathbb{L}^2\left(\mathbb{R}^d\right)$. *For any* $M > 0$, *there exists a function* $g$ *in* $\mathbb{L}^2\left(\mathbb{R}^d\right)$ *such that*

$$\forall x \in \mathbb{R}^d, \forall z \in [-M, M]^d, |\phi(x) - \phi(x-z)| \le ||z|| g(x)$$

*and*

$$\int g^2(x)\phi^{-1}(x)dx < +\infty.$$

In this context, we have the desired inequality with respect to the $L^2$-distance.

**Proposition 2.2.** *(inequality (7.11) [41])*
*Under Assumption 2.4, for all $M > 0$, there exists a positive constant $c(\phi,M)$ such that for all $z_1,z_2 \in [-M,M]^d$ and $w_1,w_2 \in (0,1)$,*

$$c(\phi,M)||z_1||^2 \left(||z_2||^2 (w_1 - w_2)^2 + w_1^2 ||z_1 - z_2||^2\right) \leq ||p_{w_1,z_1} - p_{w_2,z_2}||_2^2.$$

One can notice that Assumption 2.4 implies that $\phi$ is bounded (see Assumption $(\mathbf{H}_{\mathcal{S}})$ in [41]). Hence, we can deduce a deviation inequality for $\rho$-estimators of parameters.

**Theorem 2.6.** *We assume $\overline{\mathcal{F}}_2$ has a finite VC-index $V$, $w^* \in (0,1]$ and $z^* \neq 0$. For $\delta$ as in (2.11), there exists a positive constant $C(\phi,w^*,z^*)$ and an integer $n_0 = n_0(\phi,w^*,z^*)$ such that for any $\rho$-estimator $\hat{P} = P_{\hat{w},\hat{z}}$ on $\mathscr{Q}_\delta$, $n \geq n_0$ and for all $\xi \in (0,\xi_n)$, we have*

$$C(\phi,z^*,w^*)\left((w^* - \hat{w})^2 + \left(||z^* - \hat{z}||^2 \wedge 1\right)\right) \leq \frac{\xi + (V+1)\log(n)}{n},$$

*with probability at least $1 - e^{-\xi}$, where $\xi_n = (1+V)[1 + \log(2n/(1+V))])$.*

This result is proven in Section 2.E.1. It implies the consistency of $\hat{z}$ and consequently the consistency of $\hat{w}$ if $z^* \neq 0$, the parameter $w^*$ being ill defined if $z^* = 0$. We can deduce a bound on the convergence rate for $\hat{z}$ and also for $\hat{\lambda}$ but only for $n$ large enough. It is similar to Theorem 3.1 of Gadat *et al.* [41] with a smaller power for the logarithmic term. This slight improvement is allowed by the VC assumption. Furthermore, we do not need to know a value of $M$ such that $z^* \in [-M,M]$ or to specify it in the model. The examples of translation families taken by Gadat *et al.* [41] (Section 6) all satisfy the VC assumption.

**Lemma 2.2.** *We have the following VC-subgraph classes of density functions.*

- *The Cauchy location-scale family $\mathcal{C}$ of density functions, given hereafter by (2.30), is VC-subgraph with VC-index $V(\mathcal{C}) \leq 5$.*
- *As a consequence of Lemma 2.1, the univariate normal location-scale family $\mathcal{G}_1$ is VC-subgraph with VC-index at most 5.*
- *The Laplace location family $\mathcal{L}$ of density functions defined by*

$$\mathcal{L} = \left\{x \mapsto \frac{1}{2}e^{-|x-z|}; z \in \mathbb{R}\right\}$$

  *is VC-subgraph with VC-index $V(\mathcal{L}) \leq 29$.*
- *The location family of densities $\mathcal{SG}_\alpha$ associated to the skew Gaussian density defined by*

$$\mathcal{SG}_\alpha = \left\{x \mapsto 2\phi_1(x-z)\int_{-\infty}^{x-z}\phi_1(\alpha t)dt; z \in \mathbb{R}\right\}$$

  *is VC-subgraph with VC-index $V(\mathcal{SG}_\alpha) \leq 10$ for all $\alpha \in \mathbb{R}$, where $\phi_1$ is given by (2.15).*

This lemma is proven in Section 2.F. By inclusion, if the bound holds for the location-scale family it also holds for the location family with fixed scale parameter.

**Proving a lower bound for a specific example**

In some specific situations, it is relatively easy to prove a lower bound on the Hellinger distance. This is what we do in the following example and it allows us to obtain faster rates than the usual parametric one. Let $\alpha$ be in (0,1). We denote by $s_\alpha$ the probability density function with respect to the Lebesgue measure on $\mathbb{R}$ defined by

$$s_\alpha : x \in \mathbb{R} \mapsto \frac{1-\alpha}{2|x|^\alpha} \mathbb{1}_{|x|\in(0,1]}.$$

We consider $\mathcal{Q}$ as in (2.27) with $\phi = s_\alpha$ and for $w \in [0,1]$ and $z \in \mathbb{R}$, we write

$$p_{w,z} = (1-w)s_\alpha + w s_\alpha(\cdot - z).$$

We can prove that the Hellinger distance $h(P_{w,z},P_{w',z'})$ is lower bounded by some distance between the parameters which leads to the following theorem.

**Theorem 2.7.** *For $w^* > 0$ and $z^* \neq 0$, there is a positive constant $C(\alpha,z^*,w^*)$ such that, for any $\rho$-estimator $\hat{P} = P_{\hat{w},\hat{z}}$ on $\mathcal{Q}_\delta$ with $\delta = 10/n$ and $n \geq 20$, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$ we have*

$$C(\alpha,z^*,w^*)\left[1 \wedge |\hat{z} - z^*|^{1-\alpha} + (w^* - \hat{w})^2\right] \leq \frac{\log(n) + \xi}{n}.$$

This result is proven in Section 2.E.2. It implies rather directly that our estimators $\hat{w}$ and $\hat{z}$ estimate $w^*$ and $z^*$ at a rate which is at worst $\sqrt{(\log n)/n}$ and $(n^{-1}\log n)^{1/(1-\alpha)}$ respectively. This latter rate is faster than the usual $1/\sqrt{n}$-rate for all $\alpha \in (0,1)$. Up to the logarithmic factors, these rates are optimal. For $\hat{z}$, it a consequence of Theorem 1.1 in [54] (Chapter VI), noticing that $s_\alpha$ has a singularity of order $-\alpha$ in 0, and with the fact that we cannot do better than $1/\sqrt{n}$ for the Hellinger distance. One can notice that both maximum likelihood and least squares approaches do not apply here since we consider density functions that are unbounded, and not even square integrable for $\alpha \in [1/2,1)$.

## 2.4 Model selection

In Section 2.3 we consider estimation on a model with a fixed order $K$ and simple emission families. We use model selection to overcome this restriction in this section and consider composite emission families and/or models with different orders.

### 2.4.1 Construction of the estimator

Let $\Theta$ be a subset of

$$\bigcup_{K\geq 1} \{K\} \times \prod_{k=1}^{K} \Lambda_k.$$

Let $\delta : \Theta \to (0,1]$ be such that for $\theta = (K,\lambda_1,\ldots,\lambda_K) \in \Theta$, $\delta(\theta) \in (0,1/K]$. We write

$$\mathcal{Q}_\delta(\theta) = \left\{ \sum_{k=1}^{K} w_k F_k ; w \in \mathcal{W}_K, w_k \geq \delta, w_k \in \mathbb{Q}, F_k \in \mathcal{F}_k, \forall k \in [K] \right\}.$$

We define $\mathcal{Q}_\delta$ by

$$\mathcal{Q}_\delta = \bigcup_{\theta\in\Theta} \mathcal{Q}_\delta(\theta).$$

We associate to $\mathscr{Q}_\delta$ the family $\mathcal{Q}_\delta$ of densities with respect to $\mu$ and the $\rho$-estimator $\hat{P}_\delta$ of $\overline{P}$ based on the family $\mathcal{Q}_\delta$. Assuming we have a penalty function $\mathbf{pen} : \mathcal{Q}_\delta \to \mathbb{R}$, we set

$$\boldsymbol{\Upsilon}(\mathbf{X},q) = \sup_{q' \in \mathcal{Q}_\delta} [\mathbf{T}(\mathbf{X},q,q') - \mathbf{pen}(q')] + \mathbf{pen}(q), \tag{2.28}$$

for all $q \in \mathcal{Q}_\delta$. The $\rho$-estimator $\hat{P}_\delta$ is any measurable element of the closure (with respect to the Hellinger distance) of the set $\boldsymbol{\mathscr{E}}(\psi,\mathbf{X})$, as defined by (2.9). One can notice that a constant penalty function does not have any impact on the definition of $\boldsymbol{\Upsilon}$ and brings us back to the previous situation.

### 2.4.2 Estimation on a mixture model based on composite emission families

Let $K$ be larger than or equal to 2. Let $L$ be a subset of $\prod_{k=1}^K \Lambda_k$ and define $\Theta$ by $\Theta = \{K\} \times L$, i.e. $K$ is fixed. For $\lambda = (\lambda_1, \ldots, \lambda_K) \in L$, the model $\mathscr{Q}(\lambda)$ is a subset of

$$\left\{ \sum_{k=1}^K w_k F_k ; w \in \mathcal{W}_K, F_k \in \overline{\mathscr{F}}_{\lambda_k}, \forall k \in [K] \right\}$$

and we define its countable subset $\mathscr{Q}_\delta(\lambda)$ by

$$\mathscr{Q}_\delta = \left\{ \sum_{k=1}^K w_k F_k \in \mathscr{Q}(\lambda); w \in \mathcal{W}_K, w_k \geq \delta(\lambda), w_k \in \mathbb{Q}, F_k \in \mathscr{F}_{\lambda_k}, \forall k \in [K] \right\},$$

where $\delta$ is any function $L \to (0,1/K]$, and $\mathscr{Q}_\delta = \bigcup_{\lambda \in L} \mathscr{Q}_\delta(\lambda)$. Under Assumption 2.2, we write $\overline{V}(\lambda) = V(\lambda_1) + \cdots + V(\lambda_K)$.

**Theorem 2.8.** *Let $\Delta$ be a mapping $L \to \mathbb{R}^+$ such that $\sum_{\lambda \in L} e^{-\Delta(\lambda)} \leq 1$. Let $\mathbf{pen}$ be the penalty function defined by*

$$\mathbf{pen}(q) = \kappa \inf_{\lambda \in L | Q \in \mathscr{Q}(\lambda)} \left[ 116.1 \overline{V}(\lambda) \left[ 5.82 + \log\left( \frac{(K+1)^2}{\delta(\lambda)} \right) + \log_+\left( \frac{n}{\overline{V}(\lambda)} \right) \right] + \Delta(\lambda) \right], \tag{2.29}$$

*where $\kappa$ is given by (19) in [9]. Assume there is $P^*$ in $\mathscr{P}$ such that $\mathbf{P}^* = (P^*)^{\otimes n}$. For the choice $\delta(\lambda) = \frac{\overline{V}(\lambda)}{n(K-1)} \wedge \frac{1}{K}$, there is a positive constant $C$ such that the resulting estimator $\hat{P} = \hat{P}_\delta$ satisfies the following. For all $\xi > 0$, with probability at least $1 - e^{-\xi}$ we have*

$$Ch^2(P^*,\hat{P}) \leq \inf_{\lambda \in L} \left\{ h^2(P^*, \mathscr{Q}(\lambda)) + \frac{1}{n} \left( \overline{V}(\lambda) \left[ 1 + \log\left( \frac{Kn}{\overline{V}(\lambda) \wedge n} \right) \right] + \Delta(\lambda) + \xi \right) \right\}.$$

*The constant $C$ is universal, in particular it does not depend on $K$ or on the choice of the model.*

This proof of this theorem is postponed to Section 2.B.4. It is a general result for the situation where you know the number $K$ of subpopulations, or at least want to fix it for the estimation, but are hesitating on the models for the emission distributions. For instance, let us consider Gaussian and Cauchy location-scale families for the composite emission families, an example simpler than Example 2.1. For all $k \in \{1, \ldots, K\}$, we take $\Lambda_k = \{1,2\}$ with $\overline{\mathscr{F}}_1 = \mathscr{G}$ and $\overline{\mathscr{F}}_2 = \mathscr{C}$, where $\mathscr{C}$ is the Cauchy location-scale family of distributions associated to the density class

$$\mathcal{C} = \left\{ x \mapsto \frac{1}{\pi\sigma} \frac{1}{1 + \left(\frac{x-z}{\sigma}\right)^2} ; z \in \mathbb{R}, \sigma > 0 \right\}. \tag{2.30}$$

We consider the model $\mathscr{Q} = \cup_{0\le j\le K}\mathscr{Q}_j$ with

$$\mathscr{Q}_j = \left\{ \sum_{k=1}^{j} w_k \mathcal{N}(z_k,\sigma_k^2) + \sum_{k=j+1}^{K} w_k \text{Cauchy}(z_k,\sigma_k);\ \begin{array}{l}(z_1,\sigma_1) > \cdots > (z_j,\sigma_j),\\ (z_{j+1},\sigma_{j+1}) > \cdots > (z_K,\sigma_K)\end{array} \right\},$$

where the order $>$ on the parameters $(z_k,\sigma_k)$ is defined by (2.22) and allows to have identifiability properties again here. Lemma 2.2 gives the same bound on the VC-indices of $\mathcal{G}_1$ and $\mathcal{C}$ therefore (2.29) provides a constant penalty function, hence we will consider a null penalty function.

**Theorem 2.9.** *Assume* $P^* = \sum_{k=1}^{j^*} \overline{w}_k \mathcal{N}(\overline{z}_k,\overline{\sigma}_k^2) + \sum_{k=j^*+1}^{K} \overline{w}_k \mathit{Cauchy}(\overline{z}_k,\overline{\sigma}_k) \in \mathscr{Q}_{j^*}$ *with* $(\overline{z}_1,\overline{\sigma}_1) > \cdots > (\overline{z}_{l^*},\overline{\sigma}_{l^*})$ *and* $(\overline{z}_{l^*+1},\overline{\sigma}_{l^*+1}) > \cdots > (\overline{z}_K,\overline{\sigma}_K)$. *Let* $\hat{P}$ *be a* $\rho$*-estimator on* $\mathscr{Q}_\delta$ *with* $\delta = \frac{5}{n} \bigwedge \frac{1}{K}$ *and a null penalty. There exists an integer* $n_0(P^*)$ *and a positive constant* $C(P^*)$ *such that for* $n \ge n_0(P^*)$ *there exists an event of probability* $1-(n(K+1))^{-K}$ *on which* $\hat{P} \in \mathscr{Q}_{j^*}$ *and*

$$C(P^*)\left( ||\overline{w}-\hat{w}||^2 + \sum_{k=1}^{j^*} \left|\left|(\overline{z}_k,\overline{\sigma}_k^2) - (\hat{z}_k,\hat{\sigma}_k^2)\right|\right|^2 \wedge 1 + \sum_{k=j^*+1}^{K} ||(\overline{z}_k,\overline{\sigma}_k) - (\hat{z}_k,\hat{\sigma}_k)||^2 \wedge 1 \right) \le \frac{K\log(n(K+1))}{n}.$$

This result is proven in Section 2.D.2. Following the proof, the constant $C(P^*)$ depends both on the distance between $P^*$ and the "wrong models" $\mathscr{Q}_j, j \ne j^*$ and on the smallest eigen value of the Fisher's information matrix (within the regular parametric model $\mathscr{Q}_{j^*}$). Theorem 2.9 shows that it is possible to identify the true emission models for $n$ large enough and if this identification is established we can also estimate the different parameters. This seems to be somehow original as we did not find any result of this kind in the literature.

### 2.4.3 Selection of the order $K$

We consider $\Theta$ of the form $\Theta = \bigcup_{K\in\mathscr{K}} \{K\} \times \{\lambda\}^K$, where $\mathscr{K}$ is a subset of $\{1,\dots,n\}$. For $K \in \mathscr{K}$, we write $\overline{\mathscr{F}} = \overline{\mathscr{F}}_\lambda$ and $\mathscr{F} = \mathscr{F}_\lambda$ its countable and dense subset given by Assumption 2.1. For $K \in \mathscr{K}$, the model $\mathscr{Q}(K)$ is a subset of

$$\left\{ \sum_{k=1}^{K} w_k F_k; w \in \mathcal{W}_K, F_k \in \overline{\mathscr{F}}, \forall k \in [K] \right\}.$$

We define $\mathscr{Q}_\delta(K) := \left\{ \sum_{k=1}^{K} w_k F_k \in \mathscr{Q}(K); w \in \mathcal{W}_K, w_k \ge \delta, w_k \in \mathbb{Q}, F_k \in \mathscr{F}, \forall k \in [K] \right\}$ and $\mathscr{Q}_\delta = \bigcup_{K\in\mathscr{K}} \mathscr{Q}_\delta(K)$, where $\delta : \mathscr{K} \to (0,1]$ satisfies $\delta(K) \le 1/K$. Under Assumption 2.2, we denote by $V$ the VC-index of $\overline{\mathcal{F}}$, therefore $\overline{V}(K) = K \times V$. If $\hat{P} = \hat{P}_\delta$ is a $\rho$-estimator on $\mathscr{Q}_\delta$, we denote by $\hat{K}$ the smallest integer $K$ in $\mathscr{K}$ such that $\hat{P} \in \mathscr{Q}_\delta(K)$.

**Theorem 2.10.** *Let* $\Delta$ *be a function* $\mathscr{K} \to \mathbb{R}^+$ *satisfying* $\sum_{K\in\mathscr{K}} e^{-\Delta(K)} \le 1$. *We consider the penalty function defined by*

$$\mathbf{pen}(q) = \kappa \inf_{K\in\mathscr{K}|Q\in\mathscr{Q}(K)} \left[ 116.1KV \left[ 5.82 + \log\left(\frac{(K+1)^2}{\delta(K)}\right) + \log_+\left(\frac{n}{KV}\right)\right] + \Delta(K)\right], \quad (2.31)$$

*where* $\kappa$ *is given by (19) in [9]. Assume there exists* $P^*$ *in* $\mathscr{P}$ *such that* $\mathbf{P}^* = (P^*)^{\otimes n}$. *For the choice* $\delta(1) = 1$ *and* $\delta(K) = \frac{V}{n} \bigwedge \frac{1}{K}$ *for* $K \ge 2$, *there is a positive constant* $C$ *such that any*

*$\rho$-estimator $\hat{P} = \hat{P}_\delta$ on $\mathscr{Q}_\delta$ satisfies the following. For all $\xi > 0$, with probability at least $1 - e^{-\xi}$ we have*

$$Ch^2(P^*,\hat{P}) \leq \inf_{K\in\mathscr{K}} \left\{ h^2(P^*, \mathscr{Q}(K)) + \frac{KV\log(n) + \xi + \Delta(K)}{n} \right\}. \tag{2.32}$$

*The constant $C$ is universal, in particular it does not depend on $\mathscr{F}$ and therefore neither on $V$.*

This result is proven in Section 2.B.5. It gives an oracle inequality and it provides a way to determine the number of clusters if one wants to use mixture models in order to do clustering. It is also interesting in the context of density estimation. Once again, we take advantage of the approximation properties of GMMs to use our estimator for density estimation on a wider class. We use the approximation result proven by Maugis & Michel [68]. Let $\beta > 0$, $r = \lfloor\beta\rfloor$ and $k \in \mathbb{N}$ such that $\beta \in (2k, 2k+2]$. Let also $\mathcal{P}$ be the 8-tuple of parameters $(\gamma, l^+, L, \epsilon, C, \alpha, \xi, M)$ where $L$ is a polynomial function on $\mathbb{R}$ and the other parameters are positive constants. We define the density class $\mathcal{H}(\beta, \mathcal{P})$ of all densities $p$ satisfying the following conditions.

- For all $x$ and $y$ such that $|y - x| \leq \gamma$,

  $$(\log p)^{(r)}(x) - (\log p)^{(r)}(y) \leq r!L(x)|y - x|^{\beta-r}.$$

  Furthermore for all $j \in \{0, \dots, r\}$,

  $$|(\log p)^{(j)}(0)| \leq l^+.$$

- We have

  $$\max_{1\leq j\leq r} \int_{\mathbb{R}} \left|(\log p)^{(j)}(x)\right|^{\frac{2\beta+\epsilon}{j}} p(x)dx \vee \int_{\mathbb{R}} |L(x)|^{2+\frac{\epsilon}{\beta}} p(x)dx \leq C.$$

- For all $x \in \mathbb{R}, p(x) \leq M\psi(x)$.

- The function $f$ is strictly positive, non-decreasing on $(-\infty, -\alpha)$ and non-increasing on $(\alpha,\infty)$. For all $x \in [-\alpha, \alpha]$ we have $p(x) \geq \xi$ .

This class of functions can be approximated by Gaussian mixture models, the quality of the approximation depending on the regularity parameter $\beta$.

**Lemma 2.3.** *(Lemma 6.1 in [68])*
*For $0 < \underline{\beta} < \overline{\beta}$, there exists a set of parameters $\mathcal{P}(\underline{\beta},\overline{\beta})$ and a positive constant $c_{\underline{\beta},\overline{\beta}}$ such that for all $\beta \in \left[\underline{\beta},\overline{\beta}\right]$, all $p \in \mathcal{H}\left(\beta,\mathcal{P}(\underline{\beta},\overline{\beta})\right)$ and all $K \geq 2$, we have*

$$h^2\left(P, \mathscr{G}_{mix,K}\right) \leq c_{\underline{\beta},\overline{\beta}} \frac{(\log K)^{3\beta}}{K^{2\beta}},$$

*where $\mathscr{G}_{mix,K}$ is given by (2.16).*

We consider $\mathscr{K} = \{2, \dots, n\}$, $\Delta(K) = K$ and the penalty function **pen** as in (2.31).

**Theorem 2.11.** *Let $\hat{P} = \hat{P}_\delta$ be a $\rho$-estimator on $\mathscr{Q}_\delta$ with $\delta$ as in (2.32). For $0 < \underline{\beta} < \overline{\beta}$, there exist a positive constant $C_{\underline{\beta},\overline{\beta}}$ such that for any $p^*$ in $\mathcal{H}\left(\beta,\mathcal{P}(\underline{\beta},\overline{\beta})\right)$ with $\beta \in \left[\underline{\beta},\overline{\beta}\right]$, for all $\xi > 0$, we have*

$$h^2(P^*,\hat{P}) \leq C_{\underline{\beta},\overline{\beta}} \left( \frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}} + \frac{\xi}{n} \right),$$

*with probability at least $1 - e^{-\xi}$.*

This theorem is proven in Section 2.C.3. It provides an upper bound on the convergence rate of our estimator of order $(\log n)^{5\beta/(4\beta+2)} n^{-\beta/(2\beta+1)}$. It is the same rate obtained in Theorem 2.9 of Maugis & Michel [68] and therefore our estimator as well is minimax adaptive to the regularity $\beta$, up to a power of $\log(n)$. Moreover, in our setting there is no need to specify $\underline{\beta}$ nor $\overline{\beta}$ in our model i.e. there is no condition on the location and scale parameters of each component. Intuitively, this would allow us to obtain a better approximation bound but we did not have time to look into that direction.

# Appendix

## 2.A Main result

In this section we prove the main result of this paper, Proposition 2.3, which gives an upper bound on the $\rho$-dimension for finite mixture models. The $\rho$-dimension function is properly introduced in [11]. Bounding the $\rho$-dimension is the key element as it allows to obtain the general result Theorem 2.12 as a direct application of Theorem 2 [11]. We recall definitions from [11] that we adapt to our context, in particular the function $\psi$ defined by (2.6) satisfies Assumption 2 of Baraud & Birgé [11] with $a_0 = 4, a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see Proposition 3 [11]) which gives the different constants. Let $\mathscr{M}$ be a countable subset of $\mathscr{P}$. For $y > 0$ and $\overline{P} \in \mathscr{P}$ we write

$$\mathscr{B}^{\mathscr{M}}\left(\overline{P},y\right) = \left\{Q \in \mathscr{M}; \mathbf{h}^2\left(\mathbf{P}^*,\overline{P}^{\otimes n}\right) + \mathbf{h}^2\left(\mathbf{P}^*,Q^{\otimes n}\right) < y^2\right\}.$$

If $\mathcal{Q}$ is a set of probability density functions with respect to a $\sigma$-finite measure $\nu$ such that $\mathscr{M} \cup \{\overline{P}\} = \{q \cdot \nu; q \in \mathcal{M}\}$, we write

$$w\left(\nu,\mathcal{M},\mathscr{M},\overline{P},y\right) = \left[\sup_{Q \in \mathscr{B}^{\mathscr{M}}(\overline{P},y)} \left|\mathbf{T}\left(\mathbf{X},\overline{p},q\right) - \mathbb{E}_{\mathbf{P}^*}\left[\mathbf{T}\left(\mathbf{X},\overline{p},q\right)\right]\right|\right].$$

Similarly, we define $\mathbf{w}^{\mathscr{M}}\left(\overline{P},y\right) = \inf_{(\nu,\mathcal{M})} w\left(\nu,\mathcal{M},\mathscr{M},\overline{P},y\right)$, where the infimum is taken over all couples $(\nu,\mathcal{M})$ such that $\mathcal{M}$ is the class of density functions associated to $\mathscr{M}$ with respect to $\nu$, $\sigma$-finite measure. We can now define the $\rho$-dimension function of $\mathscr{M}$ by

$$D^{\mathscr{M}}\left(\mathbf{P}^*,\overline{P}^{\otimes n}\right) = \left[\beta^2 \sup\left\{y^2; \mathbf{w}^{\mathscr{Q}}\left(\overline{P},y\right) > \frac{3y^2}{64}\right\}\right] \bigvee 1,$$

with $\beta = \frac{\sqrt{3}}{2^{5+1/4}}$. Following the notation established in Section 2.4, we need to bound the $\rho$-dimension function over each $\mathscr{Q}_\delta(\theta)$ in order to apply Theorem 2 [11].

**Proposition 2.3.** *Under Assumption 2.2, for $\theta = (K,\lambda_1,\ldots,\lambda_K) \in \Theta$, we write*

$$\overline{V}(\theta) = V_{1,\lambda_1} + \cdots + V_{K,\lambda_K},$$

*where $V_{k,\lambda_k}$ is an upper bound on the VC-index of $\overline{\mathcal{F}}_{k,\lambda_k}$. For all $\mathbf{P} \in \mathscr{P}$ and $\overline{P} \in \mathscr{Q}_\delta$, we have the following bound*

$$D^{\mathscr{Q}_\delta(\theta)}\left(\mathbf{P},\overline{P}^{\otimes n}\right) \leq D_n(\delta,\theta) := 545.3\overline{V}(\theta)\left[5.82 + \log\left(\frac{(K+1)^2}{\delta(\theta)}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right)\right]. \quad (2.33)$$

### 2.A.1 Proof of Proposition 2.3

The strategy of the proof is based on the following remark. One can notice that if for some pair $(\nu,\mathcal{Q})$ there is $y_0$ such that $w\left(\nu,\mathcal{Q},\mathscr{Q},\overline{P},y\right) \leq \frac{3y^2}{64}$ for all $y \geq y_0$, then we have

$$D^{\mathscr{Q}}\left(\mathbf{P}^*,\overline{P}^{\otimes n}\right) \leq (\beta y_0)^2 \bigvee 1. \quad (2.34)$$

Let $\theta'$ be an element of $\Theta$ such that $\overline{P}$ belongs to $\mathscr{Q}_\delta(\theta')$. Following notation of Section 2.4, we prove such an inequality for the pair $(\mu,\mathcal{Q}_\delta(\theta) \cup \{\overline{p}\})$ where $\overline{p}$ is the density function in $\mathcal{Q}_\delta(\theta')$ associated to $\overline{P}$. To bound $w\left(\nu,\mathcal{Q},\mathscr{Q},\overline{P},y\right)$, we are going to bound the entropy of $\mathscr{B}^{\mathscr{Q}_\delta(\theta)}\left(\overline{P},y\right)$ which is possible since each emission model is associated to VC-subgraph classes of density functions (see Assumption 2.2). For a metric space $(\mathscr{A},d)$ and $\epsilon > 0$, we denote by $N(\epsilon,\mathscr{A},d)$ the minimal number of balls of radius $\epsilon$ needed to cover $\mathscr{A}$. The next lemma provides a bound on the covering number for our model, up to some modification.

**Lemma 2.4.** *For $\theta = (K,\lambda_1,\ldots,\lambda_K)$, we write $\overline{V}(\theta) = V_{1,\lambda_1} + \cdots + V_{K,\lambda_K}$ and we define*

$$\mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right) := \left\{\psi\left(\sqrt{\frac{q}{\overline{p}}}\right); q \in \mathcal{Q}_\delta(\theta)\right\}.$$

*For any probability distribution $R$, we have*

$$\forall \epsilon \leq 2, \log N\left(\epsilon,\mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right), ||\cdot||_{L_2(R)}\right) \leq \overline{V}(\theta)\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta(\theta)}\right) + 2\overline{V}(\theta)\log(1/\epsilon). \quad (2.35)$$

The next lemma is an intermediate result in the proof of Theorem 2 [12]. It allows to bound the expectation of the supremum of an empirical process from a bound on the covering number on the considered space of functions.

**Lemma 2.5.** *Let $\mathcal{F}$ be an at most countable set of measurable functions $\mathcal{X} \to \mathbb{R}$ such that for any probability distribution $P$ on $(\mathcal{X},\mathcal{X})$, we have*

$$\log(N(\epsilon,\mathcal{F},||\cdot||_{L_2(P)})) \leq a + b\log(1/\epsilon).$$

*Let $X_1,\ldots,X_n$ be $n$ independent random variables with values in $(\mathcal{X},\mathcal{X})$. We define $Z(\mathcal{F})$ by*

$$Z(\mathcal{F}) = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right)\right|$$

*and assume $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f^2(X_i)\right] \leq \sigma^2 \leq 1$. Let $q \in (0,1)$. We have*

$$\mathbb{E}\left[Z(\mathcal{F})\right] \leq 32A^2 + A2\sqrt{2n\sigma^2},$$

*with $A = \frac{1+q}{1-q}\left(1 + \frac{b}{\log 2 + 2a + b\log(1/q)}\right)\sqrt{\log 2 + 2a + b\log(1/q) + 2b\log(1/\sigma)}$.*

Let $y$ be a positive real number. We set

$$\mathcal{F}_{\delta,\theta,y}\left(\overline{P}\right) = \left\{\psi\left(\sqrt{\frac{q}{\overline{p}}}\right); Q = q\cdot\mu \in \mathcal{Q}_\delta(\theta), \mathbf{h}^2\left(\mathbf{P}^*,\overline{\mathbf{P}}\right) + \mathbf{h}^2\left(\mathbf{P}^*,\hat{\mathbf{P}}\right) < y^2\right\} \subset \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right).$$

Since $\psi$ satisfies Assumption 2 [11] and given Lemma 2.4, we can apply Lemma 2.5 with $\sigma^2 = (3\sqrt{2}y^2/n) \wedge 1$,

$$a = \overline{V}(\theta)\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta(\theta)}\right) \text{ and } b = 2\overline{V}(\theta).$$

We get

$$w^{\mathcal{Q}_\delta(\theta)}\left(\mu,\mathcal{Q}_\delta(\theta)\cup\{\overline{p}\},\mathcal{Q}_\delta(\theta),\overline{P},y\right) \leq \mathbb{E}\left[Z\left(\mathcal{F}_{\delta,\theta,y}\right)\right] \leq 32A^2 + A2\sqrt{2n\sigma^2},$$

with A given in Lemma 2.5. Let us try to find a simple upper bound for it. In our situation, dropping the dependency on $\theta$ in the notation, we have

$$\begin{aligned}\frac{b}{\log 2 + 2a + b\log(1/q)} &= \frac{2\overline{V}}{\log 2 + 2\overline{V}\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta}\right) + 2\overline{V}\log(1/q)} \\ &\leq \frac{1}{\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta q}\right)} \\ &\leq \frac{1}{\log\left(\frac{e^{1+1/e}8K(K+1)^2}{q}\right)} \leq \frac{1}{\log\left(\frac{e^{1+1/e}2^4\times 3^2}{q}\right)},\end{aligned}$$

hence

$$A \leq \frac{1+q}{1-q}\left(1+\frac{1}{\log\left(\frac{e^{1+1/e}2^4\times 3^2}{q}\right)}\right)\sqrt{2\overline{V}\left[\log\left(\frac{e^{1+1/e}2^{13/4}}{q}\right)+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]}.$$

For $q = 1/9$, we have

$$\begin{aligned}
A &\leq \frac{5}{4}\left(1+\frac{1}{1+\frac{1}{e}+4\log(6)}\right)\sqrt{2\overline{V}\left[\frac{1}{e}+1+\log\left(2^{13/4}\times 9\right)+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]} \\
&\leq \frac{5}{4}\times 1.12\sqrt{2\overline{V}\left[5.82+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]} \\
&= 2.8\sqrt{2\overline{V}\left[5.82+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]}.
\end{aligned}$$

Finally,

$$w^{\mathscr{Q}_\delta(\theta)}\left(\mu,\mathcal{Q}_\delta(\theta)\cup\{\overline{p}\},\overline{P},y\right) \leq C_0\sqrt{n\overline{V}\sigma^2\mathscr{L}(\sigma,\delta,\theta)}+C_1\overline{V}\mathscr{L}(\sigma,\delta,\theta) \tag{2.36}$$

with $\mathscr{L}(\sigma,\delta,\theta) = 5.82+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)$, $C_0 = 2.8\times 4 = 11.2$ and $C_1 = 2^6\times 2.8^2$. Then we follow the proof of Proposition 6 [12]. For $D \geq \frac{\beta^2}{3\sqrt{2}}\overline{V} = 2^{-11}\overline{V}$ and $y \geq \beta^{-1}\sqrt{D}$, we have

$$\begin{aligned}
\mathscr{L}(\sigma,\delta,\theta) &= 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{3\sqrt{2}y^2}\right) \\
&\leq 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{\beta^2 n}{3\sqrt{2}D}\right) \\
&= 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{2^{11}D}\right) \\
&\leq 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{\overline{V}}\right) = L.
\end{aligned}$$

We combine it with (2.36) and since $y \geq \beta^{-1}\sqrt{D}$ we get

$$\begin{aligned}
w^{\mathscr{Q}_\delta(\theta)}(\mu,\mathcal{Q}_\delta(\theta)\cup\{\overline{p}\},\overline{P},y) &\leq 11.2\times\sqrt{3\sqrt{2}}y\sqrt{\overline{V}L}+2^6\times 2.8^2\overline{V}L \\
&= \frac{3y^2}{64}\left[\frac{64\times 11.2\times 2^{1/4}\sqrt{\overline{V}L}}{\sqrt{3}y}+\frac{2^{12}\times 2.8^2\overline{V}L}{3y^2}\right] \\
&\leq \frac{3y^2}{64}\left[\frac{64\times 11.2\times 2^{1/4}\sqrt{\overline{V}L}}{\sqrt{3}\beta^{-1}\sqrt{D}}+\frac{2^{12}\times 2.8^2\overline{V}L}{3\beta^{-2}D}\right] \\
&= \frac{3y^2}{64}\left[2\times 11.2\frac{\sqrt{\overline{V}L}}{\sqrt{D}}+2\sqrt{2}\times 2.8^2\frac{\overline{V}L}{D}\right].
\end{aligned}$$

For $D = 545.3\overline{V}L \geq \overline{V}L\left[\sqrt{11.2^2+2\sqrt{2}\times 2.8^2}+11.2\right]^2$ we have $D \geq 2^{-11}\overline{V}$ since $L \geq 5.82$. Moreover, for all $y \geq y_0 = \beta^{-1}\sqrt{D}$, we have $w^{\mathscr{Q}_\delta(\theta)}(\mu,\mathcal{Q}_\delta(\theta)\cup\{\overline{p}\},\overline{P},y) \leq \frac{3y^2}{64}$ which allows to conclude with (2.34). We now turn to the proofs of the two lemmas.

## Proof of Lemma 2.5

The lemma is an intermediate result in the proof of Theorem 2 of Baraud & Chen [12]. We write $\overline{Z}(f) = \sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{n} \epsilon_i f(X_i)\right|$ where $\epsilon_1,\dots,\epsilon_n$ are i.i.d. Rademacher random variables. We follow the proof with $h(x) = a+b\log(1/x)$ in (39) and everything stays the same up to equation (42). We get

$$\mathbb{E}\left[\overline{Z}(\mathcal{F})\right] \leq \sqrt{2n}\frac{1+q}{1-q}\int_0^B \sqrt{\log 2 + 2a + b\log(1/q) + 2b\log(1/u)}du,$$

with $B = \sqrt{\sigma^2 + \frac{8\mathbb{E}\left[\overline{Z}(\mathcal{F})\right]}{n}} \wedge 1$. With Lemma 2 [12], we have

$$\mathbb{E}\left[\overline{Z}(\mathcal{F})\right] \leq 16A^2 + A\sqrt{2n\sigma^2},$$

with $A = \frac{1+q}{1-q}\left(1 + \frac{b}{\log 2+2a+b\log(1/q)}\right)\sqrt{\log 2 + 2a + b\log(1/q) + 2b\log(1/\sigma)}$. Classical symmetrization arguments imply

$$\mathbb{E}\left[Z(\mathcal{F})\right] \leq 2\mathbb{E}\left[\overline{Z}(\mathcal{F})\right] \leq 32A^2 + A\sqrt{2n\sigma^2}. \quad \square$$

## Proof of Lemma 2.4

We write $\phi = \psi\left(\sqrt{\cdot/\overline{p}}\right)$. We drop the dependency on $\theta$ in this proof.

**Lemma 2.6.** *For any probability distribution $R$ on $(\mathscr{X},\mathcal{X})$, for $w,v \in \mathcal{W}_K$ such that $w_k,v_k \geq \delta$ for $k = 1,\dots,K$ and for any probability densities $q_1,\dots,q_K,r_1,\dots,r_K$, we have*

$$\begin{aligned}&||\phi\circ(w_1q_1+\cdots+w_kq_K) - \phi\circ(v_1r_1+\cdots+v_Kr_K)||_{L_2(R)}\\ &\leq \frac{1}{\sqrt{\delta}}\sum_{k=1}^{K}||\phi\circ q_k - \phi\circ r_k||_{L_2(R)} + \frac{2}{\delta}||w-v||_\infty,\end{aligned} \tag{2.37}$$

*where $||w-v||_\infty = \max_{k\in[K]}|w_k - v_k|$.*

This lemma implies that for any probability distribution $R$ on $(\mathscr{X},\mathcal{X})$, we have

$$\begin{aligned}\log N\left(\epsilon,\mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right),||\cdot||_{L_2(R)}\right) &\leq \log N\left(\epsilon_{K+1},\mathcal{W}_K,||\cdot||_\infty\right)\\ &\quad + \sum_{k=1}^{K}\log N\left(\epsilon_k,\phi\circ\mathcal{F}_k,||\cdot||_{L_2(R)}\right),\end{aligned} \tag{2.38}$$

where $\phi\circ\mathcal{F}_k := \left\{\phi\circ f \middle| F\in\mathscr{F}_k\right\}$ for $k = 1,\dots,K$ and $\epsilon = \frac{\epsilon_1+\cdots+\epsilon_K}{\sqrt{\delta}} + \frac{2\epsilon_{K+1}}{\delta}$. Let us bound the covering numbers involved in the latter inequality. From Proposition 42 in [9] and Lemma 1 in [12], we have the following bound . For any probability measure $R$ on $(\mathscr{X},\mathcal{X})$ and for all $\epsilon_k \in (0,2)$, we have

$$\log N\left(\epsilon_k,\phi\circ\mathcal{F}_k,||\cdot||_{L_2(R)}\right) \leq \log\left(eV_k(8e)^{V_k-1}\right) + 2(V_k-1)\log(1/\epsilon_k). \tag{2.39}$$

We also need a bound on the covering number of $\mathcal{W}_K$. For $\epsilon_{K+1} > 0$, we have

$$\log N\left(\epsilon_{K+1},\mathcal{W}_K,||\cdot||_\infty\right) \leq K\log\left(\frac{3}{\epsilon_{K+1}}\right). \tag{2.40}$$

The proof comes at the end on page 53. We can now combine (2.38), (2.39) and (2.40). For $\epsilon \in (0,2)$ and $\delta \in (0,1/K]$, we take

$$\epsilon_{K+1} = \epsilon\frac{\delta}{2}\frac{K}{K+\sum_{k=1}^{K} 2(V_k-1)} \text{ and } \epsilon_j = \epsilon\sqrt{\delta}\frac{2(V_j-1)}{K+\sum_{k=1}^{K} 2(V_k-1)}, j = 1,\ldots,K.$$

We get

$$\begin{aligned}
\log N\left(\epsilon, \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right), ||\cdot||_{L_2(R)}\right) &\leq K\log\left(\frac{3}{\epsilon_{K+1}}\right) + \log\left(e^K\left(\prod_k V_k\right)(8e)^{\sum_k (V_k-1)}\right) \\
&\quad + \sum_{k=1}^{K} 2(V_k-1)\log(1/\epsilon_k) \\
&= K\log\left(\frac{6}{\epsilon\delta}\frac{K+\sum_{k=1}^{K} 2(V_k-1)}{K}\right) \\
&\quad + \log\left(e^K\left(\prod_k V_k\right)(8e)^{\overline{V}-K}\right) \\
&\quad + \sum_{k=1}^{K} 2(V_k-1)\log\left(\frac{1}{\epsilon\sqrt{\delta}}\frac{K+\sum_{j=1}^{K} 2(V_j-1)}{2(V_k-1)}\right) \\
&= \log\left(\frac{\left[K+\sum_{j=1}^{K} 2(V_j-1)\right]^{K+\sum_{j=1}^{K} 2(V_j-1)}}{K^K \times \prod_{k=1}^{K}[2(V_k-1)]^{2(V_k-1)}}\right) \\
&\quad + \overline{V}\log\left(\left[\prod_k V_k\right]^{1/\overline{V}}\right) \\
&\quad + \log\left(\frac{6^K e^{\overline{V}} 8^{\overline{V}-K}}{\delta^{\overline{V}}}\right) + (2\overline{V}-K)\log(1/\epsilon).
\end{aligned}$$

The following inequalities allow to simplify this. For all $x_1,\ldots,x_n \geq 0$ such that $x_1+\cdots+x_n > 0$, we have

$$\log\left(\frac{(x_1+\cdots+x_n)^{x_1+\cdots+x_n}}{x_1^{x_1}\ldots x_n^{x_n}}\right) \leq (x_1+\cdots+x_n)\log(n), \tag{2.41}$$

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{x_1+\cdots+x_n}} \leq \left(e^{\frac{1}{e}}\right)^{\frac{1}{n^n}} \leq e^{1/e}. \tag{2.42}$$

Then, we get

$$\begin{aligned}
\log N\left(\epsilon, \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right), ||\cdot||_{L_2(R)}\right) &\leq \left[K+\sum_{j=1}^{K} 2(V_j-1)\right]\log(K+1) + \overline{V}\log\left(e^{1/e}\right) \\
&\quad + \log\left(\frac{e^{\overline{V}} 8^{\overline{V}}}{\delta^{\overline{V}}}\right) + (2\overline{V}-K)\log(1/\epsilon) \\
&\leq \overline{V}\log\left(\frac{e^{1+1/e} 8(K+1)^2}{\delta}\right) + 2\overline{V}\log(1/\epsilon).
\end{aligned}$$

To conclude we need to prove (2.41), (2.42) and (2.40).

**Proof of (2.41) and (2.42)**

• In a first time, we assume $x_1 + \cdots + x_n = 1$, i.e. $x \in \mathcal{W}_n$. Then

$$\log\left(\frac{(x_1+\cdots+x_n)^{x_1+\cdots+x_n}}{x_1^{x_1}\dots x_n^{x_n}}\right) = -\sum_{i=1}^n x_i \log(x_i) \text{ and } \left(\prod_{i=1}^n x_i\right)^{\frac{1}{x_1+\cdots+x_n}} = \prod_{i=1}^n x_i.$$

Both functions $x \mapsto -\sum_{i=1}^n x_i \log(x_i)$ and $x \mapsto \prod_{i=1}^n x_i$ are bounded and attains a maximum on $\mathcal{W}_n$ for $x_1 = \cdots = x_n = 1/n$, such that

$$-\sum_{i=1}^n x_i \log(x_i) \leq \log(n) \text{ and } \prod_{i=1}^n x_i \leq \left(\frac{1}{n}\right)^n.$$

• In the generic case, we define $s(x) := x_1 + \cdots + x_n > 0$ and $y$ in $\mathcal{W}_n$ by $y_i = x_i/s(x)$ for $i = 1, \dots, n$. We have

$$\log\left(\frac{(x_1+\cdots+x_n)^{x_1+\cdots+x_n}}{x_1^{x_1}\dots x_n^{x_n}}\right) = s(x) \times \left[-\sum_{i=1}^n y_i \log(y_i)\right] \qquad \leq (x_1+\cdots+x_n)\log(n)$$

and

$$\begin{aligned}\left(\prod_{i=1}^n x_i\right)^{\frac{1}{x_1+\cdots+x_n}} &= s(x)^{1/s(x)} \times \left(\prod_{i=1}^n y_i\right)^{1/s(x)} \\ &\leq s(x)^{1/s(x)} \times \left(\frac{1}{n^n}\right)^{1/s(x)} \\ &\leq \left(e^{1/e}\right)^{\frac{1}{n^n}}.\end{aligned}$$

The last inequality comes from $\forall x > 0, x^{1/x} \leq e^{1/e}$ and we get (2.42) with $e \geq 1$ and $n^n \geq 1$. □

**Proof of (2.40)**

Let $\epsilon \in (0,1)$. Let $N$ be an integer greater than $\frac{1}{\epsilon}$. We define

$$\mathcal{W}_{K,N} := \left\{ w \in \mathcal{W}_K \middle| \forall k \in [K], \exists d_k \in \mathbb{N}, w_k = \frac{d_k}{N} \right\}.$$

- One can easily see that there is a bijection between $\mathcal{M}_{K,N}$ and the set

$$\mathcal{D}_{K,N} := \left\{ d_1, \dots, d_K \in \mathbb{N} \middle| \sum_{k=1}^K d_k = N \right\}.$$

  We have the following bound $|\mathcal{D}_{K,N}| = \binom{N+K-1}{N} \leq (N+1)^K$.

- Let $w$ be in $\mathcal{W}_K$. For $k \in [K]$, we write $a_k = \lfloor N w_k \rfloor$. We define $s(a) \in \mathbb{N}$ and $d \in \mathcal{D}_{K,N}$ by $s(a) := a_1 + \cdots + a_K \leq N$ and

$$\forall k \in [K], d_k := a_k + \mathbb{1}_{s(a)+k \leq N} \in [\lfloor N w_k \rfloor, \lfloor N w_k \rfloor + 1].$$

  Therefore, we have $v \in \mathcal{W}_{K,N}$ defined by $v_k = \frac{d_k}{N}$, such that

$$\forall k \in [K], |w_k - v_k| \leq 1/N,$$

  i.e. $||w - v||_\infty \leq 1/N \leq \epsilon$.

Therefore $\mathcal{W}_{K,N}$ is a $\epsilon$-net of $\mathcal{W}_K$ with respect to $||\cdot||_\infty$ and for $N = \lceil 1/\epsilon \rceil \geq 1/\epsilon$ we have

$$\begin{aligned}\log(N(\epsilon,\mathcal{W}_K,d)) &\leq \log\left(|\mathcal{W}_{K,N}|\right) = \log\left(|\mathcal{D}_{K,N}|\right) \\ &\leq K\log(1+N) \leq K\log\left(\frac{3}{\epsilon}\right). \quad \square\end{aligned}$$

This concludes the proof of Lemma 2.4.

## Proof of Lemma 2.6

The result is just the combination of the two following claims and the triangle inequality.

- First claim: For any probability distribution $R$, any nonnegative measurable functions $q_1$,$q_2$,$g$ and any $w \in (0,1)$ we have

$$||\phi \circ (wq_1 + (1-w)g) - \phi \circ (wq_2 + (1-w)g)||_{L_2(R)} \leq \frac{1}{\sqrt{w}} ||\phi \circ q_1 - \phi \circ q_2||_{L_2(R)}. \tag{2.43}$$

- Second claim: Let $g_1,\ldots,g_K$ be $K$ densities. For $w$,$v \in \mathcal{W}_{K,\delta}$, we have

$$||\phi \circ (w_1 g_1 + \cdots + w_K g_K) - \phi \circ (v_1 g_1 + \cdots + v_K g_K)||_{L_2(R)} \leq \frac{2}{\delta}||w - v||_\infty. \tag{2.44}$$

Combining those inequalities, we have

$$\begin{aligned}\left|\left|\phi \circ \left(\sum_{k=1}^K w_k f_k\right) - \phi \circ \left(\sum_{k=1}^K v_k g_k\right)\right|\right|_{L_2(R)} &\leq \left|\left|\phi \circ \left(\sum_{k=1}^K w_k f_k\right) - \phi \circ \left(\sum_{k=1}^K v_k f_k\right)\right|\right|_{L_2(R)} \\ &\quad + \sum_{k=1}^K ||\phi \circ (h_{k-1}) - \phi \circ (h_k)||_{L_2(R)} \\ &\leq \frac{2}{\delta}||w - v||_\infty + \sum_{k=1}^K \frac{1}{\sqrt{\nu_k}} ||\phi \circ (g_k) - \phi \circ (f_k)||_{L_2(R)} \\ &\leq \frac{2}{\delta}||w - v||_\infty + \frac{1}{\sqrt{\delta}} \sum_{k=1}^K ||\phi \circ (g_k) - \phi \circ (f_k)||_{L_2(R)},\end{aligned}$$

with $h_k = \sum_{j=1}^k v_j g_j + \sum_{j=k+1}^K v_j f_j$.

- Proof of (2.43).
  For two probability densities $f_1$ and $f_2$, for $x$ such that $\overline{p}(x) > 0$ and $f_1(x) + f_2(x) > 0$, computation gives

$$\begin{aligned}|\phi \circ f_1(x) - \phi \circ f_2(x)| &= \left|\psi\left(\sqrt{f_1/\overline{p}}(x)\right) - \psi\left(\sqrt{f_2/\overline{p}}(x)\right)\right| \\ &= \left|\frac{\sqrt{\frac{f_1}{\overline{p}}}(x) - 1}{\sqrt{\frac{f_1}{\overline{p}}}(x) + 1} - \frac{\sqrt{\frac{f_2}{\overline{p}}}(x) - 1}{\sqrt{\frac{f_2}{\overline{p}}}(x) + 1}\right| \\ &= \frac{2\left|\sqrt{\frac{f_1}{\overline{p}}}(x) - \sqrt{\frac{f_2}{\overline{p}}}(x)\right|}{\left(\sqrt{\frac{f_1}{\overline{p}}}(x) + 1\right)\left(\sqrt{\frac{f_2}{\overline{p}}}(x) + 1\right)} \\ &= \frac{2\left|\frac{f_1}{\overline{p}}(x) - \frac{f_2}{\overline{p}}(x)\right|}{\left(\sqrt{\frac{f_1}{\overline{p}}}(x) + 1\right)\left(\sqrt{\frac{f_2}{\overline{p}}}(x) + 1\right)\left(\sqrt{\frac{f_1}{\overline{p}}}(x) + \sqrt{\frac{f_2}{\overline{p}}}(x)\right)}.\end{aligned} \tag{2.45}$$

For $f_1 = wq_1 + (1-w)g$ and $f_2 = wq_2 + (1-w)g$, dropping $x$ in the notation, we get

$$\begin{aligned}
&|\phi \circ (wq_1 + (1-w)g) - \phi \circ (wq_2 + (1-w)g)| \\
&= \frac{2w\left|\frac{q_1-q_2}{\overline{p}}\right|}{\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+1\right)\left(\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}+1\right)\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}\right)} \\
&= \frac{2\left|\frac{q_1-q_2}{\overline{p}}\right|}{\left(\sqrt{\frac{q_1}{\overline{p}}}+1\right)\left(\sqrt{\frac{q_2}{\overline{p}}}+1\right)\left(\sqrt{\frac{q_1}{\overline{p}}}+\sqrt{\frac{q_2}{\overline{p}}}\right)} \\
&\times \frac{w\left(\sqrt{\frac{q_1}{\overline{p}}}+1\right)\left(\sqrt{\frac{q_2}{\overline{p}}}+1\right)\left(\sqrt{\frac{q_1}{\overline{p}}}+\sqrt{\frac{q_2}{\overline{p}}}\right)}{\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+1\right)\left(\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}+1\right)\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}\right)} \\
&= |\phi \circ q_1 - \phi \circ q_2| \times \frac{\sqrt{w}\left(\sqrt{\frac{q_1}{\overline{p}}}+1\right)}{\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+1\right)} \times \frac{\sqrt{w}\left(\sqrt{\frac{q_2}{\overline{p}}}+1\right)}{\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}+1} \\
&\times \frac{\sqrt{\frac{q_1}{\overline{p}}}+\sqrt{\frac{q_2}{\overline{p}}}}{\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}}.
\end{aligned}$$

For $w \in (0,1)$ and any $y_1, y_2, z \geq 0$ such that $y_1 + y_2 + z > 0$, we have

$$\begin{aligned}
&\frac{\sqrt{y_1}+\sqrt{y_2}}{\sqrt{wy_1+(1-w)z}+\sqrt{wy_2+(1-w)z}} \times \frac{\sqrt{w}\left(\sqrt{y_1}+1\right)}{\sqrt{wy_1+(1-w)z}+1} \\
&\times \frac{\sqrt{w}\left(\sqrt{y_2}+1\right)}{\sqrt{wy_2+(1-w)z}+1} \\
&\leq \frac{\sqrt{y_1}+\sqrt{y_2}}{\sqrt{wy_1}+\sqrt{wy_2}} \times \frac{\sqrt{w}\left(\sqrt{y_1}+1\right)}{\sqrt{wy_1}+1} \times \frac{\sqrt{w}\left(\sqrt{y_2}+1\right)}{\sqrt{wy_2}+1} \leq \frac{1}{\sqrt{w}}.
\end{aligned}$$

Finally, for $x$ such that $\overline{p}(x) > 0$ and $q_1(x) + q_2(x) + g(x) > 0$, we have

$$|\phi \circ (wq_1 + (1-w)g)(x) - \phi \circ (wq_2 + (1-w)g)(x)| \leq \frac{1}{\sqrt{w}} |\phi \circ q_1(x) - \phi \circ q_2(x)|. \tag{2.46}$$

We now considered the atypical cases given the convention established in section 2.3.1. If $q_1(x) = q_2(x) = r(x) = 0$, we have

$$|\phi \circ (wq_1 + (1-w)g)(x) - \phi \circ (wq_2 + (1-w)g)(x)| = 0$$

whether $\overline{p}(x)$ is positive or not. This equality is also true when $\overline{p}(x) = 0$, $q_1(x) + g(x) > 0$ and $q_2(x) + g(x) > 0$. The last case is for $\overline{p}(x) = q_1(x) = g(x) = 0$ and $q_2(x) > 0$ ($q_1$ and $q_2$ being interchangeable). We have

$$\begin{aligned}
&|\phi \circ (wq_1 + (1-w)g)(x) - \phi \circ (wq_2 + (1-w)g)(x)| \\
&= 1 = |\phi \circ q_1(x) - \phi \circ q_2(x)| \leq \frac{1}{\sqrt{w}} |\phi \circ q_1(x) - \phi \circ q_2(x)|.
\end{aligned}$$

Therefore, inequality (2.46) is always valid and taking the $L_2(R)$ norm provides the desired result. □

- Proof of (2.44).
–We first prove an inequality for mixtures with fixed emission densities. Let $r$ and $q$ be any probability densities on $(\mathscr{X},\mathcal{X})$. Let $w$ and $v$ be in (0,1). Using (2.45) and dropping $x$ in the notation, for $r \neq q$ we have

$$
\begin{aligned}
&|\phi \circ (wr + (1-w)q) - \phi \circ (vr + (1-v)q)| \\
&= \frac{2\,|w-v|\left|\frac{r-q}{\overline{p}}\right|}{\left(\sqrt{\frac{wr+(1-w)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{vr+(1-v)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{wr+(1-w)q}{\overline{p}}}+\sqrt{\frac{vr+(1-v)q}{\overline{p}}}\right)} \\
&\leq \begin{cases} \dfrac{2|w-v|\left|\frac{r-q}{\overline{p}}\right|}{\left(\sqrt{\frac{w|r-q|+(1-w)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{v|r-q|+(1-v)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{w|r-q|+(1-w)q}{\overline{p}}}+\sqrt{\frac{v|r-q|+(1-v)q}{\overline{p}}}\right)}, & \text{if } r>q \\ \dfrac{2|w-v|\left|\frac{r-q}{\overline{p}}\right|}{\left(\sqrt{\frac{(1-w)|q-r|+wr}{\overline{p}}}+1\right)\left(\sqrt{\frac{(1-v)|q-r|+vr}{\overline{p}}}+1\right)\left(\sqrt{\frac{(1-w)|q-r|+wr}{\overline{p}}}+\sqrt{\frac{(1-v)|q-r|+vr}{\overline{p}}}\right)}, & \text{if } r<q. \end{cases} \\
&\leq \begin{cases} \dfrac{2\left|\sqrt{w}-\sqrt{v}\right|\sqrt{\left|\frac{r-q}{\overline{p}}\right|}}{\left(\sqrt{\frac{w|r-q|}{\overline{p}}}+1\right)\left(\sqrt{\frac{v|r-q|}{\overline{p}}}+1\right)}, & \text{if } r>q \\ \dfrac{2\left|\sqrt{1-w}-\sqrt{1-v}\right|\sqrt{\left|\frac{r-q}{\overline{p}}\right|}}{\left(\sqrt{\frac{(1-w)|q-r|}{\overline{p}}}+1\right)\left(\sqrt{\frac{(1-v)|q-r|}{\overline{p}}}+1\right)}, & \text{if } r<q. \end{cases}
\end{aligned}
$$

One can easily check that the function $x \mapsto \frac{\sqrt{x}}{\left(\sqrt{\alpha x}+1\right)\left(\sqrt{\beta x}+1\right)}$ is bounded above by $\left(\alpha^{1/4}+\beta^{1/4}\right)^{-2}$. Therefore, we get

$$
\begin{aligned}
&|\phi \circ (wr + (1-w)q) - \phi \circ (vr + (1-v)q)| \\
&\leq 2\left(\frac{|\sqrt{w}-\sqrt{v}|}{\left(w^{1/4}+v^{1/4}\right)^2} \bigvee \frac{\left|\sqrt{1-w}-\sqrt{1-v}\right|}{\left((1-w)^{1/4}+(1-v)^{1/4}\right)^2}\right) \\
&= 2\left(\frac{\left|w^{1/4}-v^{1/4}\right|}{w^{1/4}+v^{1/4}} \bigvee \frac{\left|(1-w)^{1/4}-(1-v)^{1/4}\right|}{(1-w)^{1/4}+(1-v)^{1/4}}\right).
\end{aligned}
$$

The inequality obviously stands for $x$ such that $r(x) = q(x)$. Therefore we can take the $L_2(R)$-norm and get

$$
\begin{aligned}
&||\phi \circ (wr + (1-w)q) - \phi \circ (vr + (1-v)q)|| \\
&\leq 2\left(\frac{\left|w^{1/4}-v^{1/4}\right|}{w^{1/4}+v^{1/4}} \bigvee \frac{\left|(1-w)^{1/4}-(1-v)^{1/4}\right|}{(1-w)^{1/4}+(1-v)^{1/4}}\right).
\end{aligned} \tag{2.47}
$$

–We can now prove (2.44). Let $g_1,\ldots,g_K$ be $K$ probability densities. Let $w,v \in \mathcal{W}_{K,\delta}$. If $w = v$ the proof is obvious. Therefore we consider $w \neq v$. The idea is to rewrite $w_1g_1+\cdots+w_Kg_K$ and $v_1g_1+\cdots+v_Kg_K$ as 2 component mixtures with the same emission densities, allowing us to use (2.47). We define

$$
t_1 := \max_{1\leq k\leq K} \frac{w_k - v_k}{\mathbb{1}_{w_k>v_k} - v_k} \in [0,1] \text{ and } t_2 := \max_{1\leq k\leq K} \frac{v_k - w_k}{\mathbb{1}_{v_k>w_k} - w_k} \in [0,1].
$$

Since $w \neq v$, we have $t_1,t_2 > 0$. We define two probability densities $f_1$ and $f_2$ by

$$
f_1 := \sum_{k=1}^{K}\left[v_k + \frac{w_k - v_k}{t_1}\right]g_k \text{ and } f_2 := \sum_{k=1}^{K}\left[w_k + \frac{v_k - w_k}{t_2}\right]g_k.
$$

One can check that we have

$$\sum_{k=1}^{K} w_k g_k = \frac{t_1}{t_1 + t_2(1-t_1)} f_1 + \frac{t_2(1-t_1)}{t_1 + t_2(1-t_1)} f_2,$$
$$\sum_{k=1}^{K} \nu_k g_k = \frac{t_1(1-t_2)}{t_2 + t_1(1-t_2)} f_1 + \frac{t_2}{t_2 + t_1(1-t_2)} f_2.$$

We get straight from (2.47) that

$$\begin{aligned}
&||\phi \circ (w_1 g_1 + \cdots + w_k g_K) - \phi \circ (v_1 g_1 + \cdots + v_K g_K)||_{L_2(Q)} \\
&= \left|\left| \phi \circ \left( \frac{t_1}{t_1 + t_2(1-t_1)} f_1 + \frac{t_2(1-t_1)}{t_1 + t_2(1-t_1)} f_2 \right) \right.\right. \\
&\left.\left. - \ \phi \circ \left( \frac{t_1(1-t_2)}{t_2 + t_1(1-t_2)} f_1 + \frac{t_2}{t_2 + t_1(1-t_2)} f_2 \right) \right|\right|_{L_2(Q)} \\
&\leq 2 \left( \frac{\left| \left(\frac{t_2(1-t_1)}{t_1+t_2(1-t_1)}\right)^{1/4} - \left(\frac{t_2}{t_2+t_1(1-t_2)}\right)^{1/4} \right|}{\left(\frac{t_2(1-t_1)}{t_1+t_2(1-t_1)}\right)^{1/4} + \left(\frac{t_2}{t_2+t_1(1-t_2)}\right)^{1/4}} \bigvee \frac{\left| \left(\frac{t_1}{t_1+t_2(1-t_1)}\right)^{1/4} - \left(\frac{t_1(1-t_2)}{t_2+t_1(1-t_2)}\right)^{1/4} \right|}{\left(\frac{t_1}{t_1+t_2(1-t_1)}\right)^{1/4} + \left(\frac{t_1(1-t_2)}{t_2+t_1(1-t_2)}\right)^{1/4}} \right) \\
&= 2 \left( \frac{\left| (t_2(1-t_1))^{1/4} - (t_2)^{1/4} \right|}{(t_2(1-t_1))^{1/4} + (t_2)^{1/4}} \bigvee \frac{\left| (t_1)^{1/4} - (t_1(1-t_2))^{1/4} \right|}{(t_1)^{1/4} + (t_1(1-t_2))^{1/4}} \right) \\
&= 2 \left( \frac{\left| (1-t_1)^{1/4} - 1 \right|}{(1-t_1)^{1/4} + 1} \bigvee \frac{\left| 1 - (1-t_2)^{1/4} \right|}{1 + (1-t_2)^{1/4}} \right) \\
&= 2 \left( \frac{t_1}{\left((1-t_1)^{1/4} + 1\right)^2 \left((1-t_1)^{1/2} + 1\right)} \bigvee \frac{t_2}{\left((1-t_2)^{1/4} + 1\right)^2 \left((1-t_2)^{1/2} + 1\right)} \right) \\
&\leq 2(t_1 \vee t_2).
\end{aligned}$$

We end the proof of (2.44) with the following upper bound on $t_1 \vee t_2$. We have

$$\begin{aligned}
t_1 \vee t_2 &= \max_{1 \leq k \leq K} \left( \frac{w_k - v_k}{\mathbb{1}_{w_k > v_k} - v_k} \vee \frac{v_k - w_k}{\mathbb{1}_{v_k > w_k} - w_k} \right) \\
&= \max_{1 \leq k \leq K} \left\{ |w_k - v_k| \times \max\left( (1-v_k)^{-1}, (1-w_k)^{-1}, v_k, w_k \right) \right\} \\
&\leq \delta^{-1} \, ||w - v||_\infty . \quad \square
\end{aligned}$$

The proof of Lemma 2.6 is now complete.

## 2.B Theorems

In this section we provide a very general result from which we will derive Theorems 2.1, 2.2, 2.8 and 2.10.

**Theorem 2.12.** *Any $\rho$-estimator $\hat{P}$ on $\mathscr{Q}_\delta$ satisfies, with probability at least $1 - e^{-\xi}$,*

$$\begin{aligned}
\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) \leq \inf_{\theta \in \Theta} \Bigg\{ & c_0 \left( \mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}(\theta)) + n(K-1)\delta(\theta) \right) \\
& + c_1 \left( 116.1 \overline{V}(\theta) \left[ 5.82 + \log\left( \frac{(K+1)^2}{\delta(\theta)} \right) + \log_+ \left( \frac{n}{\overline{V}(\theta)} \right) \right] + \Delta(\theta) \right) \Bigg\} \\
& + c_1(1.49 + \xi).
\end{aligned} \tag{2.48}$$

*with* $c_0 = 300$ *and* $c_1 = 5014$. *Moreover, for* $K \geq 2$ *and* $\delta(\theta) = \frac{\overline{V}(\theta)}{n(K-1)} \wedge \frac{1}{K}$, *we have*

$$\log\left(\frac{(K+1)^2}{\delta(\theta)}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right) \leq (2 + \log_2(9)) \log\left(\frac{Kn}{\overline{V}(\theta) \wedge n}\right) \tag{2.49}$$

*and* $n(K-1)\delta(\theta) \leq n \wedge \overline{V}(\theta)$.

### 2.B.1 Proof of Theorem 2.12

We recall that the function $\psi$ defined by (2.6) satisfies Assumption 2 of Baraud and Birgé [11] with $a_0 = 4, a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see Proposition 3 [11]). Using Proposition 2.3, we can apply Theorem 2 [11] with

$$D_n(\delta,\theta) = 545.3\overline{V}(\theta)\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right)\right].$$

There exist constants $\gamma$ and $\kappa$ (given by (19) in [11]) such that, with probability $\geq 1 - e^{-\xi}$, we have

$$\mathbf{h}^2\left(\mathbf{P}^*, \hat{\mathbf{P}}\right) \leq \inf_{\theta\in\Theta}\left[\gamma \mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}_\delta(\theta)) + \frac{4\kappa}{a_1}\left(\frac{D_n(\delta,\theta)}{4.7} + \Delta(\theta)\right)\right] + \frac{4\kappa}{a_1}(1.49 + \xi).$$

**Lemma 2.7.** *For all* $K \geq 2$ *and* $\theta \in \Theta$, *we have*

$$\forall P \in \mathscr{P}, h\left(P, \mathscr{Q}_\delta(\theta)\right) \leq \sqrt{(K-1)\delta(\theta)} + h\left(P, \mathscr{Q}(\theta)\right). \tag{2.50}$$

Using this inequality, we get

$$\mathbf{h}^2\left(\mathbf{P}^*, \hat{\mathbf{P}}\right) \leq \inf_{\theta\in\Theta}\left[2\gamma\left(\mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}(\theta)) + n(K(\theta)-1)\delta(\theta)\right) + \frac{4\kappa}{a_1}\left(116.1\overline{V}(\theta)\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right)\right] + \Delta(\theta)\right)\right] + \frac{4\kappa}{a_1}(1.49 + \xi).$$

From Baraud & Chen [12] (see proof of Theorem 1), we get that $\gamma < 150$ and $4\kappa/a_1 < 5014$. Let us now prove (2.49). We consider $\theta$ such that $K \geq 2$ and we take $\delta(\theta) = \frac{\overline{V}(\theta)}{n(K-1)} \wedge \frac{1}{K}$.

- If $\overline{V}(\theta) \leq n(K-1)/K$, then

$$\begin{aligned}
\log\left(\frac{(K+1)^2}{\delta(\theta)}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right) &= \log\left(\frac{(K^2-1)(K+1)n^2}{\overline{V}(\theta)^2}\right) \\
&= 3\log\left(\frac{Kn}{\overline{V}(\theta)}\right) + \log\left(\frac{(K^2-1)(K+1)\overline{V}(\theta)}{K^3 n}\right) \\
&\leq 3\log\left(\frac{Kn}{\overline{V}(\theta)}\right) + \log\left(\frac{(K^2-1)^2}{K^4}\right) \\
&\leq 3\log\left(\frac{Kn}{\overline{V}(\theta) \wedge n}\right).
\end{aligned}$$

- Otherwise $\overline{V}(\theta) > n(K-1)/K$ and

$$\begin{aligned}\log\left(\frac{(K+1)^2}{\delta(\theta)}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right) &\leq \log\left(\frac{(K+1)^2K^2}{K-1}\right)\\ &= 3\log(K) + \log\left(\frac{K^2+2K+1}{K(K-1)}\right)\\ &\leq 3\log(K) + \log(9/2)\\ &\leq \left[3 + \frac{\log(9)-\log(2)}{\log(2)}\right]\log(K)\\ &\leq (2+\log_2(9))\log\left(\frac{Kn}{\overline{V}(\theta)\wedge n}\right).\end{aligned}$$

Finally, one can check that $n(K-1)\delta(\theta) \leq n \wedge \overline{V}(\theta)$.

**Proof of Lemma 2.7**

For $K \geq 2$ and $\delta \in (0,1/K]$, we define $\mathcal{W}_{K,\delta}$ by

$$\mathcal{W}_{K,\delta} = \mathcal{W}_K \cap [\delta,1]^K. \tag{2.51}$$

We prove by induction that

$$\forall \delta \in (0,1/K],\ \sup_{w\in\mathcal{W}_K} h^2(w,\mathcal{W}_{K,\delta}) \leq 1 - \sqrt{1-(K-1)\delta}. \tag{2.52}$$

• Assume (2.52) holds true for $K \geq 2$. Let $\delta$ be in $(0,1/(K+1))$ and $w$ be in $\mathcal{W}_{K+1}$. Without loss of generality we consider $w_1 \leq w_2 \leq \cdots \leq w_K \leq w_{K+1}$. We define the function $r$ by

$$r : \left|\begin{array}{lll} \mathcal{W}_{K+1} & \to \mathcal{W}_K & \\ w & \mapsto \begin{cases} \left(\frac{w_2}{1-w_1}, \frac{w_3}{1-w_1}, \ldots, \frac{w_K}{1-w_1}\right) & \text{for } w_1 \neq 0,\\ \left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right) & \text{for } w_1 = 1, \end{cases}\end{array}\right.$$

and informally $r^{-1}$ by

$$r^{-1} : \left|\begin{array}{ll} \mathcal{W}_K \times [0,1) & \to \mathcal{W}_{K+1}\\ (w',a) & \mapsto (a,(1-a)w'_1,\ldots,(1-a)w'_K). \end{array}\right.$$

- If $w_1 \geq \delta$ then $w \in \mathcal{M}_{K+1,\delta}$ and $h(w,\mathcal{W}_{K+1,\delta}) = 0$.
- Otherwise $w_1 < \delta$ and we build a distribution $v \in \mathcal{W}_{K+1,\delta}$ to approximate $w$. Take $\eta = \delta/(1-\delta) \in (0,1/K]$. From (2.52), there exists $v' \in \mathcal{M}_{K,\eta}$ such that $h^2(r(w),v') \leq 1 - \sqrt{1-(K-1)\eta}$. Now take $v = r^{-1}(\delta,v')$. We have $v_1 = \delta$ and for $j \geq 2$, $v_j = (1-\delta)v'_{j-1} \geq (1-\delta)\eta = \delta$. Therefore $v$ belongs to $\mathcal{W}_{K+1,\delta}$. We also have

$$\begin{aligned} h^2(w,v) &= \frac{1}{2}\left[\left(\sqrt{w_1}-\sqrt{\delta}\right)^2 + \left(\sqrt{1-w_1}-\sqrt{1-\delta}\right)^2\right]\\ &\quad + \sqrt{1-w_1}\sqrt{1-\delta}h^2(r(w),v')\\ &\leq \left[1-\sqrt{1-\delta}\right] + \sqrt{1-\delta}\left[1-\sqrt{1-(K-1)\eta}\right]\\ &= 1 - \sqrt{1-\delta}\sqrt{1-(K-1)\delta/(1-\delta)}\\ &= 1-\sqrt{1-K\delta}. \end{aligned}$$

• We now prove (2.52) for $K = 2$. Let $w$ be in $\mathcal{W}_2$ and without loss of generality assume that $w_1 \leq 1/2 \leq w_2$. Once again we only need to consider $w_1 < \delta$. Then we take $v = (\delta, 1-\delta)$ and we get

$$\begin{aligned} h^2(w,\mathcal{W}_{2,\delta}) &\leq h^2(w,v) \\ &= \frac{1}{2}\left[\left(\sqrt{w_1}-\sqrt{\delta}\right)^2 + \left(\sqrt{1-w_1}-\sqrt{1-\delta}\right)^2\right] \\ &\leq 1-\sqrt{1-\delta}. \end{aligned}$$

This ends the proof of (2.52). We can now prove Lemma 2.7. Let $P \in \mathscr{P}$ and $P_{w,F} \in \mathscr{Q}(\theta)$. There is $v \in \mathcal{W}_{K,\delta}$ such that $P_{v,F} \in \mathscr{Q}_\delta(\theta)$ and

$$h^2\left(w,v\right) \leq 1-\sqrt{1-(K-1)\delta} \leq (K-1)\delta.$$

By a density argument we can assume that $v \in \mathbb{Q}^K$. Therefore,

$$\begin{aligned} h\left(P,\mathscr{Q}_\delta(\theta)\right) &\leq h\left(P,P_{v,F}\right) \\ &\leq h\left(P_{v,F},P_{w,F}\right) + h\left(P,P_{w,F}\right) \\ &\leq \sqrt{(K-1)\delta} + h\left(P,P_{w,F}\right) \end{aligned}$$

where the last inequality comes from Lemma 2.8. Then, taking the infimum over $\mathscr{Q}(\theta)$ ends the proof. □

### 2.B.2 Proof of Theorem 2.1

It is a direct application of Theorem 2.12 in the specific situation where

$$\Theta = \{\theta = (K,\lambda_1,\lambda_2,\ldots,\lambda_K)\}.$$

Then, taking $\Delta(\theta) = 0$, inequality (2.48) becomes

$$\begin{aligned} \mathbf{h}^2\left(\mathbf{P}^*,\left(\hat{P}_\delta\right)^{\otimes n}\right) &\leq c_0\left(\mathbf{h}^2(\mathbf{P}^*,\mathscr{Q}_K) + n(K-1)\delta\right) \\ &\quad + c_1 116.1 \overline{V}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}}\right)\right] \\ &\quad + c_1(1.49+\xi). \end{aligned}$$

With (2.49), we have

$$\begin{aligned} \mathbf{h}^2\left(\mathbf{P}^*,\hat{P}^{\otimes n}\right) &\leq c_0\left(\mathbf{h}^2(\mathbf{P}^*,\mathscr{Q}) + n \wedge \overline{V}\right) \\ &\quad + c_1 116.1\left(2+\log_2(9)\right)\overline{V}\left[5.82 + \log\left(\frac{Kn}{\overline{V}\wedge n}\right)\right] \\ &\quad + c_1(1.49+\xi), \end{aligned}$$

for $K \geq 2$. One can easily check that it still holds for $K = 1$ (see [11]). Therefore (2.11) is proven.

### 2.B.3 Proof of Theorem 2.2

Let $\mathscr{Q}_K[\epsilon]$ be the model defined by

$$\mathscr{Q}_K[\epsilon] = \left\{\sum_{k=1}^K w_k F_k; w \in \mathcal{W}_K, F_k \in \mathscr{F}_k[\epsilon], \forall k \in [K]\right\}.$$

Since the class $\overline{\mathscr{F}}_k$ is totally bounded, the set $\mathscr{F}_k[\epsilon]$ is finite for all $k \in [K]$. We satisfy Assumptions 2.1 and 2.2 and therefore can apply Theorem 2.1 with

$$\overline{V} = \sum_{k=1}^{K} \log_2(|\mathscr{F}_k[\epsilon]|) \leq \sum_{k=1}^{K} \left(\frac{A_k}{\epsilon}\right)^{\alpha_k}.$$

Let $\hat{P} = \hat{P}_\delta$ be a $\rho$-estimator on $\mathscr{Q}_{K,\delta}[\epsilon]$. For all $\xi > 0$, we have

$$\begin{aligned}\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq\ & c_0\left[\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K[\epsilon]\right) + n(K-1)\delta\right] \\ & + c_1 116.1 \overline{V}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}}\right)\right] \\ & + c_1(1.49 + \xi),\end{aligned}$$

with probability at least $1 - e^{-\xi}$.

**Lemma 2.8.** *Let $w$ and $v$ be in $\mathcal{W}_K$. Let $F_k$ and $G_k$ be in $\mathscr{P}$ for all $k \in \{1, \ldots, K\}$. We have*

$$h\left(\sum_{k=1}^{K} w_k F_k, \sum_{k=1}^{K} v_k G_k\right) \leq h(w, v) + \max_{k \in [K]} h\left(F_k, G_k\right).$$

This lemma implies that $\mathscr{Q}_K[\epsilon]$ is a $\epsilon$-net of $\mathscr{Q}_K$ with respect to the Hellinger distance, and in particular

$$\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K[\epsilon]\right) \leq 2\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + 2n\epsilon^2.$$

Therefore, if we use (2.11) with $\overline{V} = \sum_{k=1}^{K} \left(\frac{A_k}{\epsilon}\right)^{\alpha_k}$ we get

$$C\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq 2\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + 2n\epsilon^2 + \epsilon^{-\alpha_{\max}} \sum_{k=1}^{K} A_k^{\alpha_k}\left[1 + \log(Kn)\right] + \xi.$$

Finally, for $\epsilon = n^{-\frac{1}{\alpha_{\max}+2}}$, there exists a positive constant $C$ such that for all $\xi > 0$, we have

$$C\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq \mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + n^{\frac{\alpha_{\max}}{\alpha_{\max}+2}} \sum_{k=1}^{K} A_k^{\alpha_k}\left[1 + \log\left(Kn\right)\right] + \xi,$$

with probability at least $1 - e^{-\xi}$.

**Proof of Lemma 2.8**

With Young's inequality, we can easily prove the following inequality

$$\forall x, y, z \in \mathbb{R}_+^K, \left(\sqrt{\sum_{k \in [K]} x_k z_k} - \sqrt{\sum_{k \in [K]} x_k y_k}\right)^2 \leq \sum_{k \in [K]} x_k (\sqrt{z_k} - \sqrt{y_k})^2.$$

Therefore, we get an upper bound on the Hellinger distance between mixture distributions. For $w, v \in \mathcal{W}_K$ and $F_k, G_k \in \mathscr{P}$ for all $k \in [K]$, we have

$$\begin{aligned} h\left(\sum_{k \in [K]} w_k F_k, \sum_{k \in [K]} v_k G_k\right) &\leq h\left(\sum_{k \in [K]} w_k F_k, \sum_{k \in [K]} w_k G_k\right) + h\left(\sum_{k \in [K]} w_k G_k, \sum_{k \in [K]} v_k G_k\right) \\ &\leq \sqrt{\sum_{k \in [K]} w_k h^2\left(F_k, G_k\right)} + h\left(w, v\right) \\ &\leq \max_{k \in [K]} h(F_k, G_k) + h(w, v). \end{aligned}$$

□

### 2.B.4 Proof of Theorem 2.8

Applying Theorem 2.12 in the described setting, we get

$$h^2\left(P^*, \hat{P}\right) \leq \inf_{\lambda \in L}\Bigg[c_0\left(h^2(P^*, \mathscr{Q}(\lambda)) + (K-1)\delta(\lambda)\right) + c_2\left\{\frac{116.1\overline{V}(\lambda)}{n}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta(\lambda)}\right) + \log_+\left(\frac{n}{\overline{V}(\lambda)}\right)\right] + \Delta(\lambda)\right\}\Bigg] + c_2\frac{1.49+\xi}{n},$$

with probability at least $1-e^{-\xi}$. As $K \geq 2$ and $\delta(\lambda) = \frac{\overline{V}(\lambda)}{n(K-1)} \wedge \frac{1}{K}$ we have the following with (2.49). and finally we have

$$\begin{aligned}\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) &\leq \inf_{\lambda \in L}\Big\{c_0\left(\mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}(\lambda)) + n \wedge \overline{V}(\lambda)\right) \\ &\quad + c_1\left(116.1\overline{V}(\lambda)\left[5.82 + (2+\log_2(9))\log\left(\frac{Kn}{\overline{V}(\lambda)\wedge n}\right)\right] + \Delta(\lambda)\right)\Big\} \\ &\quad + c_1(1.49+\xi) \\ &\leq C\inf_{\lambda \in L}\left\{\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}(\lambda)\right) + \overline{V}(\lambda)\left[1 + \log\left(\frac{Kn}{\overline{V}(\lambda)\wedge n}\right)\right] + \Delta(\lambda)\right\} + \xi,\end{aligned}$$

where $C$ is a positive numeric constant that does not depend on $L$.

### 2.B.5 Proof of Theorem 2.10

Applying Theorem 2.12, we get

$$h^2\left(P^*, \hat{P}\right) \leq \inf_{K \in \mathscr{K}}\Bigg[c_0\left(h^2(P^*, \mathscr{Q}(K)) + (K-1)\delta(K)\right) + c_2\left\{\frac{116.1KV}{n}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta(K)}\right) + \log_+\left(\frac{n}{KV}\right)\right] + \Delta(K)\right\}\Bigg] + c_2\frac{1.49+\xi}{n},$$

with probability at least $1-e^{-\xi}$. For $K=1$ and $\delta(K)=1$ we have $(K-1)\delta(K) = 0 \leq KV \wedge n$ and

$$\log\left(\frac{(K+1)^2}{\delta(K)}\right) + \log_+\left(\frac{n}{KV}\right) = 2\log(2) + \log\left(\frac{n}{KV\wedge n}\right).$$

Combining this inequality with (2.49), we have

$$5.82 + \log\left(\frac{(K+1)^2}{\delta(K)}\right) + \log_+\left(\frac{n}{KV}\right) \leq (5.82 + 2\log(2)) + (2+\log_2(9))\log\left(\frac{Kn}{KV\wedge n}\right)$$

for all $K \geq 1$. Finally, there is a numeric constant $C > 0$ that is universal, such that for all $\xi > 0$ we have

$$Ch^2\left(P^*, \hat{P}\right) \leq \inf_{K \in \mathscr{K}}\left[h^2(P^*, \mathscr{Q}(K)) + \frac{1}{n}\left\{KV\left[1 + \log\left(\frac{Kn}{KV\wedge n}\right)\right] + \Delta(K)\right\}\right] + \frac{\xi}{n},$$

with probability at least $1-e^{-\xi}$.

## 2.C Density estimation

This section gathers the proofs of density estimation results, namely Theorems 2.3 and 2.11.

### 2.C.1 Proof of Theorem 2.3

The Gaussian location-scale family of density functions is VC-subgraph with VC-index $V(\mathcal{C}) \leq 5$ (see Lemma 2.2). Proposition 2.1 provides an approximation bound for $\mathscr{C}(A,R)$. The proof can be found on page 63. We can now apply Theorem 2.1 with those two propositions. With (2.11), there exists a universal constant $C$ such that for $\mathbf{P}^* = (P^*)^{\otimes n}$, $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$\begin{aligned} Ch^2\left(P^*,\hat{P}\right) &\leq h^2\left(P^*,\mathscr{C}(A,R)\right) + \exp\left(-\frac{K^{1/2}}{12\sqrt{6}R^2}\right)\left[K^{1/4}\frac{3\sqrt{2}}{\sqrt{e\pi}7^{1/4}} + R\right] \\ &\quad + \frac{K\log(n) + \xi}{n} \\ &\leq h^2\left(P^*,\mathscr{C}(A,R)\right) + \frac{1}{n}\left[\frac{3\sqrt{2}}{(e\pi)^{1/2}7^{1/4}}\left(864R^4\log^2(n)+1\right)^{1/4} + R\right] \\ &\quad + \frac{\left(864R^4\log^2(n)+1\right)\log(n) + \xi}{n} \end{aligned}$$

One can check that the assumptions ensure that $\log(n) \geq 1$ and therefore

$$Ch^2\left(P^*,\hat{P}\right) \leq h^2\left(P^*,\mathscr{C}(A,R)\right) + \frac{R\log^{1/2}(n)}{n}\left[\frac{3\sqrt{2}865^{1/4}}{(e\pi)^{1/2}7^{1/4}} + 1\right] + \frac{865R^4\log^3(n)+\xi}{n}.$$

Finally, there exists a numeric constant $C > 0$ such that, for $K = \left\lceil 864R^4\log^2(n)\right\rceil \geq 2\left(24A^2+1\right)^2$, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$Ch^2\left(P^*,\hat{P}\right) \leq h^2(P^*,\mathscr{C}(A,R)) + \frac{R^4\log^3(n)+\xi}{n}.$$

The different conditions are satisfied for $n \geq \exp\left(\frac{A^2}{R^2}\frac{25}{12\sqrt{3}}\right)$.

### 2.C.2 Proof of Proposition 2.1

We first need the following result.

**Lemma 2.9.** *Let $k$ be a positive integer. For any probability distribution $H$ on $[-a,a]\times[\underline{\sigma},\overline{\sigma}]$, there is a discrete probability distribution $H'$ supported by $k(2k-1)+1$ points in $[-a,a]\times[\underline{\sigma},\overline{\sigma}]$ such that*

$$d_{TV}\left(P_H,P_{H'}\right) \leq \inf_{m>1}\left\{\frac{\sqrt{2/\pi}}{\underline{\sigma}}am\left(\frac{ea^2(1+m)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{2\underline{\sigma}}\exp\left(-\frac{(m-1)^2a^2}{2\overline{\sigma}^2}\right)\right\}.$$

The proof is postponed at the end of this one. Let $A$ and $R$ be two real numbers respectively greater than 0 and 1. As a direct consequence of this lemma, for any $l \in \mathbb{R}$, any probability distribution $H$ on $[l \pm \underline{\sigma}A]\times[\underline{\sigma},R\underline{\sigma}]$ and for $K \geq k(2k-1)+1$, we have

$$h^2(P_H,\mathscr{G}_K) \leq \inf_{m>0}\left\{\sqrt{2/\pi}A(1+m)\left(\frac{eA^2(2+m)^2}{2k}\right)^k + \frac{R}{2}\exp\left(-\frac{m^2A^2}{2R^2}\right)\right\}.$$

The goal is to have an upper bound without an infimum. For that we are going to take a value of $m$ given by the parameters $A$ and $R$. Now

$$h^2(P_H, \mathscr{G}_K) \leq \inf_{m\geq 2} \left\{ \sqrt{2/\pi} A \frac{3}{2} m \left( \frac{eA^2 4m^2}{2k} \right)^k + \frac{R}{2} \exp\left( -\frac{m^2 A^2}{2R^2} \right) \right\}$$
$$= \inf_{m\geq 2} \left\{ \frac{3}{\sqrt{2\pi}} Am \left( \frac{2eA^2 m^2}{k} \right)^k + \frac{R}{2} \exp\left( -\frac{m^2 A^2}{2R^2} \right) \right\}.$$

Let $W$ denote the Lambert $W$ function restricted to $(0;\infty)$ such that $W(x)$ is the only positive number such that $W(x)e^{W(x)} = x$. For $m = \frac{\sqrt{2W(1/4eR^2)}R}{A} k^{1/2}$ and $k \geq \frac{2A^2}{W(1/4eR^2)R^2}$, to ensure that $m \geq 2$, we get

$$h^2(P_H, \mathscr{G}_K) \leq \frac{3}{\sqrt{2\pi}} \sqrt{2W(1/4eR^2)} R k^{1/2} \left( 4eR^2 W(1/4eR^2) \right)^k + \frac{R}{2} \exp\left( -kW(1/4eR^2) \right)$$
$$= R \exp\left( -kW(1/4eR^2) \right) \left[ k^{1/2} 3\sqrt{W(1/4eR^2)/\pi} + 1/2 \right].$$

Let us simplify this bound using simple properties of the function $W$.

- For all $x > 0$, $0 < W(x) < x$.
- For all $x \in (0,1), x(1-x) < W(x)$. Therefore,

$$W(1/4eR^2) \geq \frac{1}{4eR^2} \left( 1 - \frac{1}{4eR^2} \right)$$
$$\geq \frac{(1 - 1/4e)}{4eR^2} = \frac{4e-1}{16e^2R^2} \geq \frac{1}{12R^2}.$$

Therefore, we have

$$h^2(P_H, \mathscr{G}_K) \leq R \exp\left( -\frac{k}{12R^2} \right) \left[ k^{1/2} \frac{3}{2R\sqrt{e\pi}} + 1/2 \right].$$

Since $K \geq 2\left(24A^2 + 1\right)^2$, one can check that the set

$$B = \left\{ k \in \mathbb{N} : K \geq k(2k-1) + 1 \text{ and } k \geq \frac{2A^2}{R^2 W(1/4eR^2)} \right\}$$

is not empty, e.g. $\lceil 24A^2 \rceil \in B$. We set $k = \max B \geq 1$, i.e. $k = \left\lfloor \frac{1}{4} + \sqrt{(K - 7/8)/2} \right\rfloor \leq \sqrt{K} \frac{2}{\sqrt{7}}$, we have

$$K \in \{n(2n-1)+1, \ldots, (2n+1)(n+1)\} \Rightarrow k = n \geq \sqrt{K} \frac{n}{\sqrt{(2n+1)(n+1)}}.$$

Since $x \mapsto \frac{x}{\sqrt{(2x+1)(x+1)}}$ is non-decreasing on $[1, +\infty)$, we have $k \geq \sqrt{K}/\sqrt{6}$ for all $K \geq 2$. Finally, we have

$$h^2(P_H, \mathscr{G}_K) \leq R \exp\left( -\frac{k}{12R^2} \right) \left[ k^{1/2} \frac{3}{2R\sqrt{e\pi}} + 1/2 \right]$$
$$\leq R \exp\left( -\frac{K^{1/2}}{12\sqrt{6}R^2} \right) \left[ K^{1/4} \frac{3\sqrt{2}}{2R\sqrt{e\pi}7^{1/4}} + 1/2 \right]$$
$$= \frac{1}{2} \exp\left( -\frac{K^{1/2}}{12\sqrt{6}R^2} \right) \left[ K^{1/4} \frac{3\sqrt{2}}{\sqrt{e\pi}7^{1/4}} + R \right].$$

One can see that $\underline{\sigma}$ does not play a role here and is equivalent to $s$ in the definition of $\mathcal{C}(A,R)$.

**Proof of Lemma 2.9**

The bound is obtained following the proofs of lemmas in Ghosal & van der Vaart [49]

- 1st step:
  For $|x| > a$ we have,

$$\begin{aligned} p_H(x) &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) dH(z,\sigma) \\ &\leq \frac{1}{\sqrt{2\pi\underline{\sigma}^2}} \exp\left(-\frac{(|x|-a)^2}{2\overline{\sigma}^2}\right). \end{aligned} \tag{2.53}$$

- 2nd step:
  See Lemma A.1 in Ghosal & van der Vaart [49]. Take $N = k(2k-1)+1$. There is a discrete distribution $H'$ with at most $K$ support points in $[-a,a] \times [\underline{\sigma},\overline{\sigma}]$ such that

$$\int z^l \sigma^{-(2j+1)} dH(z,\sigma) = \int z^l \sigma^{(2j+1)} dH'(z,\sigma) \tag{2.54}$$

  for $l = 0,\ldots,2k-2$ and $j = 0,\ldots,k-1$. Because of (2.54) we get

$$\int \sum_{j=0}^{k-1} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} dH(z,\sigma) = \int \sum_{j=0}^{k-1} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} dH'(z,\sigma),$$

  for $x \in \mathbb{R}$. Taylor's expansion of the exponential function ([49]),

$$\left| \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right| \leq \left(\frac{e(x-z)^2}{k2\sigma^2}\right)^k.$$

  Therefore,

$$\begin{aligned} &\sqrt{2\pi} \sup_{|x|\leq M} |p_H(x) - p_{H'}(x)| \\ &\quad = \sup_{|x|\leq M} \left| \int \frac{1}{\sigma} \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) dH(z,\sigma) \right. \\ &\quad \left. - \int \frac{1}{\sigma} \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) dH'(z,\sigma) \right| \\ &\quad = \sup_{|x|\leq M} \left| \int \frac{1}{\sigma} \left[ \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right] dH(z,\sigma) \right. \\ &\quad \left. - \int \frac{1}{\sigma} \left[ \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right] dH'(z,\sigma) \right| \\ &\quad \leq 2 \sup_{\substack{|x|\leq M \\ |z|\leq a \\ \underline{\sigma}\leq\sigma\leq\overline{\sigma}}} \frac{1}{\sigma} \left| \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right| \\ &\quad \leq 2 \sup_{\substack{|x|\leq M \\ |z|\leq a \\ \underline{\sigma}\leq\sigma\leq\overline{\sigma}}} \frac{1}{\sigma} \left(\frac{e(x-z)^2}{k2\sigma^2}\right)^k \\ &\quad \leq \frac{2}{\underline{\sigma}} \left(\frac{e(M+a)^2}{k2\underline{\sigma}^2}\right)^k. \end{aligned}$$

Obviously, the inequality (2.53) holds also for $p_{H'}$. We combine it with the last one we obtained in order to bound the total variation distance. Therefore, for $M = ma$, $m > 1$, we have

$$\begin{aligned} d_{TV}\left(P_H, P_{H'}\right) &= \frac{1}{2}\int |p_H(x) - p_{H'}(x)|dx \\ &\leq M \sup_{|x|\leq M} |p_H(x) - p_{H'}(x)| + \frac{1}{2}\int_{|x|>M} p_H(x) \vee p_{H'}(x)dx \\ &\leq \frac{\sqrt{2/\pi}}{\underline{\sigma}} M \left(\frac{e(M+a)^2}{2k\underline{\sigma}^2}\right)^k + \frac{1}{2}\int_{|x|>M} \frac{1}{\sqrt{2\pi\underline{\sigma}^2}} \exp\left(-\frac{(|x|-a)^2}{2\overline{\sigma}^2}\right) dx \\ &\leq \frac{\sqrt{2/\pi}}{\underline{\sigma}} M \left(\frac{e(M+a)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{\underline{\sigma}}\int_{x>M} \frac{1}{\sqrt{2\pi\overline{\sigma}^2}} \exp\left(-\frac{(x-a)^2}{2\overline{\sigma}^2}\right) dx \\ &\leq \frac{\sqrt{2/\pi}}{\underline{\sigma}} am \left(\frac{ea^2(1+m)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{2\underline{\sigma}} \exp\left(-\frac{(m-1)^2a^2}{2\overline{\sigma}^2}\right). \end{aligned}$$

Finally, writing $A = a/\underline{\sigma}$ and $R = \overline{\sigma}/\underline{\sigma}$, we have

$$d_{TV}\left(P_H, P_{H'}\right) \leq \inf_{m>1} \left\{ \frac{\sqrt{2/\pi}}{\underline{\sigma}} am \left(\frac{ea^2(1+m)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{2\underline{\sigma}} \exp\left(-\frac{(m-1)^2a^2}{2\overline{\sigma}^2}\right)\right\}. \quad \square$$

This concludes the proof of Proposition 2.1.

## 2.C.3 Proof of Theorem 2.11

We firs provide the proof of Lemma 2.3 which provides the necessary bound for the approximation.

### Proof of Lemma 2.3

We will use notation from [68]. With Lemma 7.23 [67] and an inclusion argument, we have

$$h^2\left(P,\mathscr{G}_K\right) \leq h^2\left(P,\mathcal{S}_K\right) \leq \frac{1}{2} D_{KL}\left(P||\mathcal{S}_K\right).$$

Combined with Lemma 6.1 [68], we get

$$\begin{aligned} h^2\left(P,\mathscr{G}_K\right) &\leq \frac{c_{\underline{\beta},\overline{\beta}}}{2}\lambda(K)^{2\beta} \\ &= \frac{c_{\underline{\beta},\overline{\beta}}}{2}\left(a_{\overline{\beta}} K^{-1}\left(\ln K\right)^{3/2}\right)^{2\beta} \\ &\leq C_{\underline{\beta},\overline{\beta}} \frac{\left(\ln K\right)^{3\beta}}{K^{2\beta}}, \end{aligned}$$

with $C_{\underline{\beta},\overline{\beta}} = c_{\underline{\beta},\overline{\beta}} a_{\overline{\beta}}^{2\overline{\beta}}/2$. $\square$

The Gaussian location-scale family of density functions is VC-subgraph (see Lemma 2.2). For $0 < \underline{\beta} < \overline{\beta}$ and $\beta \in [\underline{\beta},\overline{\beta}]$, let $\mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta},\overline{\beta})\right)$ be the class of density functions defined in Maugis-Rabusseau & Michel [68]. One can check that

$$\sum_{k\in\mathscr{K}} e^{-\Delta(K)} \leq 1,$$

for $\Delta(K) = K$. Applying Theorem 2.10, for $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$\begin{aligned} Ch^2(P^*,\hat{P}) &\leq \inf_{K\in\mathscr{K}} \left\{ h^2(P^*,\mathscr{G}_K) + \frac{K(5\log(n)+1)+\xi}{n} \right\} \\ &\leq 2h^2(P^*,\mathcal{H}\left(\beta,\mathcal{P}(\underline{\beta},\overline{\beta})\right)) + \inf_{K\in\mathscr{K}} \left\{ 2c_{\underline{\beta},\overline{\beta}} \frac{(\log K)^{3\beta}}{K^{2\beta}} + \frac{K(5\log(n)+1)}{n} \right\} + \frac{\xi}{n}. \end{aligned}$$

Therefore, following the proof of Theorem 2.9 of Maugis-Rabusseau & Michel [68], we have

$$\begin{aligned} \inf_{K\in\mathscr{K}} \left\{ 2c_{\underline{\beta},\overline{\beta}} \frac{(\log K)^{3\beta}}{K^{2\beta}} + \frac{K(5\log(n)+1)}{n} \right\} &\lesssim c_{\underline{\beta},\overline{\beta}} \inf_{K\in\mathscr{K}} \left\{ \frac{(\log K)^{3\beta}}{K^{2\beta}} + \frac{K\log(n)}{n} \right\} \\ &\lesssim c_{\underline{\beta},\overline{\beta}} \frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}}. \end{aligned}$$

Finally, there exists $C_{\underline{\beta},\overline{\beta}}$ such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$h^2(P^*,\hat{P}) \leq C_{\underline{\beta},\overline{\beta}} \left( \frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}} + \frac{\xi}{n} \right).$$

## 2.D Regular parametric models

This section gathers the proof of Theorems 2.4, 2.9 and 2.5.

### 2.D.1 Proof of Theorem 2.4

We apply the results of Ibragimov & Has'minskiĭ [54] (Chapter 1, Section 7.1 and 7.3) to parametric mixture models. We recall the notation

$$p(\cdot;\theta) = \sum_{k=1}^{K-1} w_k f_k(\cdot;\alpha_k) + (1 - w_1 - \cdots - w_{K-1}) f_K(\cdot;\alpha_K)$$

and $\Theta = \left\{ w \in (0,1)^{K-1}, \sum_{k=1}^{K-1} w_k < 1 \right\} \times A_1 \times \cdots \times A_K$. Obviously, $\Theta$ is an open convex subset of $\mathbb{R}^{K-1} \times \mathbb{R}^d_1 \times \cdots \times \mathbb{R}^{d_K}$. We first check that Assumption 2.3 implies that the model is regular.

- a) $\Rightarrow \theta \mapsto p(x;\theta)$ is continuous on $\Theta$ for $\mu$-almost all $x \in \mathscr{X}$.
- b) $\Rightarrow$ For $\mu$-almost all $x \in \mathscr{X}$ the function $u \mapsto p(x;u)$ is differentiable at the point $u = \theta$. For all $k \in \{1,\ldots,K\}$ and $j \in \{1,\ldots,d_k\}$, we have

$$\begin{aligned} \int_{\mathscr{X}} \left| \frac{\partial p(x;\theta)}{\partial \alpha_{k,j}} \right|^2 \frac{\mu(dx)}{p(x;\theta)} &= \int_{\mathscr{X}} \left| \frac{\partial f_k(x;\alpha_k)}{\partial \alpha_{k,j}} \right|^2 \frac{w_k^2}{p(x;\theta)} \mu(dx) \\ &\leq \int_{\mathscr{X}} \left| \frac{\partial f_k(x;\alpha_k)}{\partial \alpha_{k,j}} \right|^2 \frac{\mu(dx)}{f_k(x;\alpha_k)} < \infty. \end{aligned}$$

It also works with $k = K$ since $w$ is fixed here. For $k \in \{1, \ldots, K-1\}$ we get

$$\begin{aligned}\int_{\mathscr{X}} \left|\frac{\partial p(x;\theta)}{\partial w_k}\right|^2 \frac{\mu(dx)}{p(x;\theta)} &= \int_{\mathscr{X}} \left(f_k(x;\alpha_k) - f_K(x,\alpha_K)\right)^2 \frac{\mu(dx)}{p(x;\theta)} \\ &\leq \frac{2}{w_k}\int_{\mathscr{X}} f_k^2(x;\alpha_k)\frac{\mu(dx)}{f_k(x;\alpha_k)} \\ &+ \frac{2}{1-w_1-\cdots-w_k}\int_{\mathscr{X}} f_K^2(x;\alpha_k)\frac{\mu(dx)}{f_K(x;\alpha_K)} \\ &= \frac{2}{w_k} + \frac{2}{1-w_1-\cdots-w_k} < \infty.\end{aligned}$$

Therefore, we have a regular statistical experiment (see [54]). Since the Fisher's information matrix

$$I\left(\overline{\theta}\right) = \int_{\mathscr{X}} \frac{\partial p\left(x;\overline{\theta}\right)}{\partial\theta}\left(\frac{\partial p\left(x;\overline{\theta}\right)}{\partial\theta}\right)^T \frac{\mu(dx)}{p\left(x;\overline{\theta}\right)}$$

is definite positive. We can apply Theorem 7.6 of Ibragimov & Has'minskiĭ [54] which says that we have

$$\liminf_{t\to 0} ||t||^{-2} h^2(P_{\overline{\theta}}, P_{\overline{\theta}+t}) \geq \lambda(\overline{\theta})/4.$$

where $\lambda(\overline{\theta})$ is the smallest eigen value of the Fisher's information matrix $I(\overline{\theta})$. Therefore there exists $a > 0$ such that

$$\inf_{\theta\in\Theta:\left|\left|\overline{\theta}-\theta\right|\right|<a} \left|\left|\overline{\theta}-\theta\right|\right|^{-2} h^2\left(P_{\overline{\theta}},P_\theta\right) \geq \lambda\left(\overline{\theta}\right)/8.$$

Finally, there exists a positive constant $C\left(\overline{\theta}\right) = \frac{\lambda(\overline{\theta})}{8} \wedge \inf_{\substack{||\theta-\overline{\theta}||\geq a \\ \theta\in\Theta}} h^2\left(P_{\overline{\theta}},P_\theta\right) > 0$ such that

$$\forall\theta\in\Theta, \left(1+\left|\left|\overline{\theta}-\theta\right|\right|^{-2}\right) h^2\left(P_{\overline{\theta}},P_\theta\right) \geq C\left(\overline{\theta}\right).$$

We apply Theorem 2.1 so that with probability at least $1-e^{-\xi}$ we have

$$\begin{aligned}\frac{1}{n}\left[\mathbf{h}^2\left(\mathbf{P}^*,P_{\overline{\theta}}^{\otimes n}\right) + \overline{V}\log(n) + \xi\right] \geq Ch^2\left(P_{\overline{\theta}},P_{\hat{\theta}}\right) &\geq \frac{\left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2}{1+\left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2} C\times C\left(\overline{\theta}\right) \\ &\geq \frac{\left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2\wedge b}{1+b} C\times C(\overline{\theta}),\end{aligned}$$

for any $b \geq 0$. Since $||\overline{w}-\hat{w}||^2 \leq K\sum_{k=1}^{K-1}(\overline{w}_k-\hat{w}_k)^2$ and

$$\begin{aligned}\sum_{k=1}^{K-1}\left(\overline{w}_k-\hat{w}_k\right)^2 + \sum_{k=1}^{K}\left[||\overline{\alpha}_k-\hat{\alpha}_K||^2\wedge 1\right] &\leq \sum_{k=1}^{K-1}\left(\overline{w}_k-\hat{w}_k\right)^2 + \left[\sum_{k=1}^{K}||\overline{\alpha}_k-\hat{\alpha}_K||^2\right]\wedge K \\ &\leq \left[\sum_{k=1}^{K-1}\left(\overline{w}_k-\hat{w}_k\right)^2 + \sum_{k=1}^{K}||\overline{\alpha}_k-\hat{\alpha}_K||^2\right]\wedge(K+1) \\ &= \left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2\wedge(K+1),\end{aligned}$$

we get, with $b = K+1$,

$$\frac{1}{n}\left[\mathbf{h}^2\left(\mathbf{P}^*,P_{\overline{\theta}}^{\otimes n}\right) + \overline{V}\log(n) + \xi\right] \geq \left[\frac{1}{K}||\overline{w}-\hat{w}||^2 + \sum_{k=1}^{K}||\overline{\alpha}_k-\hat{\alpha}_k||^2\wedge 1\right]\frac{C\times C(\overline{\theta})}{K+2},$$

with probability at least $1-e^{-\xi}$.

### 2.D.2 Proof of Theorem 2.9

Assumption 2.2 is satisfied with Lemma 2.2. For all $j$ in $\{0,\dots,K\}$, we have $\overline{V}_j = 5K$. We apply Theorem 2.8 with $\Delta_j = \log(K+1)$ for all $j \in \{0,\dots,K\}$. This induces a constant penalty function and one can check that this does not modify the definition of $\rho$-estimators compared to a null penalty function. Therefore, the estimator can be computed with a null penalty. There exists a positive constant that does not depend on $P^*$ such that for $n \geq 5K$, any $\rho$-estimator $\hat{P}_\delta$ on $\mathscr{Q}_\delta$ satisfies, with probability at least $1-e^{-\xi}$,

$$Ch^2(P^*,\hat{P}) \leq \frac{K\log\left(n(K+1)\right)+\xi}{n}.$$

The following lemma allows to prove that for $n$ large enough, the estimator $\hat{P}$ belongs to the true model $\mathscr{Q}_{j^*}$ with high probability.

**Lemma 2.10.** *Let $j \in \{0,\dots,K\}$ and assume there is a sequence*

$$(P_n)_n = \left(\sum_{k=1}^{j} w_{k,n}\mathcal{N}(z_{k,n},\sigma_{k,n}^2) + \sum_{k=j+1}^{K} w_{k,n}\,Cauchy(z_{k,n},\sigma_{k,n})\right)_n \in \mathscr{Q}_j^{\mathbb{N}}$$

*such that $\lim_{n\to\infty} h(P_n,P^*) = 0$. Then, $j = j^*$ and there is a subsequence $\left(P_{\psi(n)}\right)_n$ such that $\lim_{n\to\infty}(z_{k,\psi(n)},\sigma_{k,\psi(n)})_{1\leq k\leq K} = (\overline{z}_k,\overline{\sigma}_k)_{1\leq k\leq K}$.*

This implies that $\alpha = \inf_{j\neq j^*} h\left(P^*,\mathscr{Q}_j\right) > 0$. For $n \geq n_0 = \inf\{n \geq 1 : C^{-1}\alpha^{-1}K < n/\log(n(K+1))\}$ and $0 < \xi < \frac{Cn\alpha}{K\log(n(K+1))}$, there is an event $\Omega_{\xi,n}$ of probability $1-e^{-\xi}$ such that

$$Ch^2(P^*,\hat{P}) \leq \frac{K\log\left(n(K+1)\right)+\xi}{n} \text{ and } \hat{P} \in \mathscr{Q}_{j^*}.$$

From now, we follow the proof of Theorem 2.3 to prove a lower bound on the Hellinger distance $h(P^*,P)$ for $P \in \mathscr{Q}_{j^*}$.

**Lemma 2.11.** *There exists a positive constant $\overline{a}$ such that for all $P_\theta = \sum_{k=1}^{j^*} w_k\mathcal{N}(z_k,\sigma_k^2) + \sum_{k=j^*+1}^{K} w_k\,Cauchy(z_k,\sigma_k) \in \mathscr{Q}_{j^*}$,*

$$h^2\left(P^*,P_\theta\right) \geq \overline{a}\Bigg(||w-\overline{w}||^2 + \sum_{k=1}^{j^*}\left|\left|\left(z_k,\sigma_k^2\right)-\left(\overline{z}_k,\overline{\sigma}_k^2\right)\right|\right|_2^2 \wedge 1 + \sum_{k=j^*+1}^{K}||(z_k,\sigma_k)-(\overline{z}_k,\overline{\sigma}_k)||_2^2 \wedge 1\Bigg).$$

Finally, there is a constant $\overline{C}$ such that for $\xi$ and $n$, on the event $\Omega_{\xi,n}$, we have

$$\overline{C}\left(||\hat{w}-\overline{w}||^2 + \sum_{k=1}^{j^*}\left|\left|\left(\hat{z}_k,\hat{\sigma}_k^2\right)-\left(\overline{z}_k,\overline{\sigma}_k^2\right)\right|\right|_2^2 \wedge 1 + \sum_{k=j^*+1}^{K}||(\hat{z}_k,\hat{\sigma}_k)-(\overline{z}_k,\overline{\sigma}_k)||_2^2 \wedge 1\right) \leq \frac{K\log(n(K+1))+\xi}{n}.$$

We still have to prove Lemmas 2.10 and 2.11.

**Proof of Lemma 2.10**

Let $j \in \{0,\dots,K\}$ and assume there is a sequence

$$(P_n)_n = \left(\sum_{k=1}^{j} w_{k,n}\mathcal{N}(z_{k,n},\sigma_{k,n}^2) + \sum_{k=j+1}^{K} w_{k,n}\text{Cauchy}(z_{k,n},\sigma_{k,n})\right) \in \mathscr{Q}_j^{\mathbb{N}}$$

such that $\lim_{n\to\infty} h(P_n,P^*) = 0$. The mixing weights are bounded so we can assume we are already considering a sequence such that $w_{k,n} \xrightarrow[n\to\infty]{} w_{k,\infty}$ for all $k \in \{1,\dots,K\}$. For the other parameters, it is always possible to extract a subsequence $P_{\psi(n)}$ such that for all $k$

$$z_{k,\psi(n)} \xrightarrow[n\to\infty]{} \begin{cases} z_{k,\infty} \in \mathbb{R}, \\ \text{or } \pm\infty, \end{cases} \quad \text{and } \sigma_{k,\psi(n)} \xrightarrow[n\to\infty]{} \begin{cases} \sigma_{k,\infty} \in \mathbb{R}^+, \\ \text{or } +\infty. \end{cases}$$

We now consider the different cases possible (dropping the dependency on $\psi$ in the notation).

- If $z_{k,n} \xrightarrow[n\to\infty]{} \pm\infty$ (without loss of generality we consider $+\infty$ in the proof), for $b \in \mathbb{R}$, we have

$$\begin{aligned} P_n([b,+\infty[) &\geq w_{k,n}\Big[\mathbb{1}_{k\leq j}\mathcal{N}(z_{k,n},\sigma_{k,n}^2)([b,+\infty[) \\ &\quad + \mathbb{1}_{k>j}\text{Cauchy}(z_{k,n},\sigma_{k,n})([b,+\infty[)\Big] \\ &\geq \frac{w_{k,n}}{2} \text{ for } n \text{ large enough.} \end{aligned}$$

  Assume $w_{k,\infty} > 0$. Since $P^*([b,+\infty[) \xrightarrow[b\to\infty]{} 0$, there exists $b$ such that $P^*([b,+\infty[) \leq w_{j,\infty}/4$. On the other hand we have $P^*([b,+\infty[) = \lim_{n\to\infty} P_{\theta_n}([b,+\infty[) \geq w_{k,\infty}/2$. Therefore, it means that $w_{k,\infty} = 0$ and it also holds for $z_{k,n} \to -\infty$.

- If $z_{k,n} \xrightarrow[n\to\infty]{} z_{k,\infty} \in \mathbb{R}$ and $\sigma_{k,n} \xrightarrow[n\to\infty]{} 0$, for $b > 0$ we have

$$\begin{aligned} P_n([z_{k,\infty}-b,z_{k,\infty}+b]) &\geq w_{k,n}\Big(\mathbb{1}_{k\leq j}\mathcal{N}(z_{k,n},\sigma_{k,n}^2)([b,+\infty[) \\ &\quad + \mathbb{1}_{k>j}\text{Cauchy}(z_{k,n},\sigma_{k,n})([b,+\infty[)\Big) \to w_{k,\infty}. \end{aligned}$$

  Assume $w_{k,\infty} > 0$. Since $P^*([z_{k,\infty}-b,z_{k,\infty}+b]) \xrightarrow[b\to 0]{} 0$, there exists $b > 0$ such that $P^*([z_{k,\infty}-b,z_{k,\infty}+b]) \leq w_{j,\infty}/2$. On the other hand we have $P^*([z_{k,\infty}-b,z_{k,\infty}+b]) = \lim_{n\to\infty} P_n([z_{k,\infty}-b,z_{k,\infty}+b]) \geq w_{k,\infty}$. Therefore, it means that $w_{k,\infty} = 0$.

- If $z_{k,n} \to z_{k,\infty} \in \mathbb{R}$ and $\sigma_{k,n} \to \infty$, for $a > 0$ we have

$$\begin{aligned} P_n([-a,a]) &\leq (1-w_{k,n}) \\ &\quad + w_{k,n}\Big(\mathbb{1}_{k\leq j}\mathcal{N}(z_{k,n},\sigma_{k,n}^2)([-a,a]) + \mathbb{1}_{k>j}\text{Cauchy}(z_{k,n},\sigma_{k,n})([-a,a])\Big) \\ &\xrightarrow[n\to\infty]{} (1-w_{k,\infty}). \end{aligned}$$

  Since $P^*([-a,+a]) \xrightarrow[a\to+\infty]{} 1$, we get $w_{k,\infty} = 0$

This proves that $P_n$ converges to

$$P_\infty = \sum_{\substack{k\leq j(\lambda)\\ w_{k,\infty}>0}} w_{k,\infty}\mathcal{N}(z_{k,\infty},\sigma_{k,\infty}^2) + \sum_{\substack{k>j(\lambda)\\ w_{k,\infty}>0}} w_{k,\infty}\text{Cauchy}(z_{k,\infty},\sigma_{k,\infty}),$$

and necessarily $P^* = P_\infty$. Lemma 2.10 with the assumptions on $P^*$ implies $j = j^*$ and there exist two permutations $\tau_g,\tau_c$ respectively on $\{1,\dots,j^*\}$ and $\{j^*+1,\dots,K\}$ such that $(\overline{\pi}_k,\overline{z}_k,\overline{\sigma}_k) = (w_{\tau_g(k)},z_{\tau_g(k)},\sigma_{\tau_g(k)})$ for $k$ in $\{1,\dots,j^*\}$ and $(\overline{\pi}_k,\overline{z}_k,\overline{\sigma}_k) = (w_{\tau_c(k)},z_{\tau_c(k)},\sigma_{\tau_c(k)})$ for $k$ in $\{j^*+1,\dots,K\}$. □

**Proof of Lemma 2.11**

- The map $(z,\sigma^2) \mapsto g(x;z,\sigma^2) = \phi_\sigma(x-z)$ is continuous and differentiable on $\mathbb{R} \times \mathbb{R}^{+*}$ with

$$\partial_z \phi_\sigma(x-z) = \phi_\sigma(x-z)\frac{(x-z)}{\sigma^2}$$
$$\partial_{\sigma^2} \phi_\sigma(x-z) = \phi_\sigma(x-z)\left[\frac{(x-z)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right].$$

  Similarly $(z,\sigma) \mapsto f(x;z,\sigma) = \frac{1}{\pi\sigma}\frac{1}{c(x;z,\sigma)}$ is continuous and differentiable on $\mathbb{R} \times \mathbb{R}^{+*}$ with

$$\partial_z f(x;z,\sigma) = \frac{1}{\pi\sigma^3}\frac{x-z}{c^2(x;z,\sigma)}$$
$$\partial_\sigma f(x;z,\sigma) = \frac{1}{\pi\sigma^2 c(x;z,\sigma)}\left[1 - \frac{2}{c(x;z,\sigma)}\right].$$

  Moreover, on can check that we have

$$\int_{\mathbb{R}} \left|\partial_z g(x;z,\sigma^2)\right|^2 \frac{dx}{g(x;z,\sigma^2)} = \int_{\mathbb{R}} \frac{(x-z)^2}{\sigma^4}\phi_\sigma(x-z)dx < \infty$$
$$\int_{\mathbb{R}} \left|\partial_{\sigma^2} g(x;z,\sigma^2)\right|^2 \frac{dx}{g(x;z,\sigma^2)} = \int_{\mathbb{R}} \left[\frac{(x-z)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right]^2 \phi_\sigma(x-z)dx < \infty$$
$$\int_{\mathbb{R}} \left|\partial_z f(x;z,\sigma)\right|^2 \frac{dx}{f(x;z,\sigma)} = \int_{\mathbb{R}} \frac{(x-z)^2}{\pi\sigma^5 c^3(x;z,\sigma)}dx < \infty$$
$$\int_{\mathbb{R}} \left|\partial_{\sigma^2} f(x;z,\sigma^2)\right|^2 \frac{dx}{f(x;z,\sigma)} = \int_{\mathbb{R}} \frac{1}{\pi\sigma^3 c(x;z,\sigma)}\left[1 - \frac{2}{c(x;z,\sigma)}\right]^2 dx < \infty.$$

- The function $\theta \mapsto \psi(\cdot;\theta) = \frac{\partial}{\partial\theta} p^{1/2}(\cdot;\theta)$, where

$$p(x;\theta) = \sum_{k=1}^{j^*} w_k \phi_{\sigma_k}(x-z_k) + \sum_{k=j^*+1}^{K} \frac{1}{\pi\sigma c(x;z,\sigma)}$$

  and

$$\theta = (w_1,\ldots,w_{K-1},z_1,\ldots,z_K,\sigma_1^2,\ldots,\sigma_{j^*}^2,\sigma_{j^*+1},\ldots,\sigma_K),$$

  is continuous in the space $L_2(\mu)$.

- We apply Theorem 1 of Meijer & Ypma [72]. For $j^* < K$,

$$\begin{aligned}
&\det(I(\theta)) = 0 \\
&\Rightarrow \exists \lambda \neq 0, \sum_{k=1}^{j^*} \phi_{\sigma_k}(x-z_k)\left(\frac{w_k\lambda_{z_k}(x-z_k)}{\sigma_k^2} + w_k\lambda_{\sigma_k^2}\left[\frac{(x-z)^2}{2\sigma_k^4} - \frac{1}{2\sigma_k^2}\right] + \lambda_{w_k}\right) \\
&+ \sum_{k=j^*+1}^{K-1}\left(\frac{w_k\lambda_{z_k}(x-z_k)}{\pi\sigma^3 c^2(x;z_k,\sigma_k)} + \frac{w_k\lambda_{\sigma_k}}{\pi\sigma_k^2}\left[\frac{1}{c(x;z_k,\sigma_k)} - \frac{2}{c^2(x;z_k,\sigma_k)}\right] + \frac{\lambda_{w_k}}{\pi\sigma_k c(x;z_k,\sigma_k)}\right) \\
&+ (1-w_1-\cdots-w_{K-1})\left(\frac{\lambda_{z_K}(x-z_K)}{\pi\sigma_K^3 c^2(x;z_K,\sigma_K)} + \frac{\lambda_{\sigma_K}}{\pi\sigma_K^2}\left[\frac{1}{c(x;z_K,\sigma_K)} - \frac{2}{c^2(x;z_K,\sigma_K)}\right]\right) \\
&- \frac{1}{\pi\sigma_K c(x;z_K,\sigma_K)}\sum_{k=1}^{K-1}\lambda_{w_k} = 0 \text{ for } \mu\text{-almost all } x.
\end{aligned}$$

For $j^* = K$,

$$\begin{aligned}
&\det(I(\theta)) = 0 \\
&\Rightarrow \exists \lambda \neq 0, \sum_{k=1}^{K-1} \phi_{\sigma_k^2}(x - z_k) \left( w_k \lambda_{z_k} \frac{(x - z_k)}{\sigma_k^2} + w_k \lambda_{\sigma_k^2} \left[ \frac{(x-z)^2}{2\sigma_k^4} - \frac{1}{2\sigma_k^2} \right] + \lambda_{w_k} \right) \\
&+ \phi_{\sigma_K}(x - z_K) \left\{ (1 - w_1 - \cdots - w_{K-1}) \left( \lambda_{z_K} \frac{(x - z_K)}{\sigma_K^2} + \lambda_{\sigma_K^2} \left[ \frac{(x-z)^2}{2\sigma_K^4} - \frac{1}{2\sigma_K^2} \right] \right) \right. \\
&\left. - \sum_{k=1}^{K-1} \lambda_{w_k} \right\} = 0 \text{ for } \mu\text{-almost all } x.
\end{aligned}$$

**Lemma 2.12.** *Let* $(z_1,\sigma_1),\ldots,(z_K,\sigma_K)$ *be distinct elements of* $\mathbb{R} \times \mathbb{R}^{+*}$. *For any integer* $n$, *the families*

$$A = \left\{ x \mapsto x^j \phi_{\sigma_i}(x - z_i); i \in \{1,\ldots,K\}, j \in \{0,\ldots,n\} \right\}$$

*and*

$$B = \left\{ x \mapsto \frac{x^j}{c^l(x; z_i,\sigma_i)}; i \in \{1,\ldots,K\}, l \in \{1,2\}, j \in \{0,1\} \right\}$$

*are linearly independent. Moreover, the linear spaces* $\boldsymbol{Span}_{\mathbb{R}}(A)$ *and* $\boldsymbol{Span}_{\mathbb{R}}(B)$ *are orthogonal.*

This proves that $I(\bar{\theta})$ is non singular.

- We now check $\inf_{\substack{||\bar{\theta}-\theta|| \geq a \\ P_\theta \in \mathscr{Q}_{j^*}}} h^2(P_{\bar{\theta}},P_\theta) > 0, \forall a > 0$. It is a direct consequence of Lemma 2.10.

- $\mathscr{Q}(\lambda^*)$ is a regular parametric model. We consider the parameter to be $\sigma$ for the Cauchy distribution and $\sigma^2$ for the Gaussian distribution. Obviously, $(z,\sigma) \mapsto g(x; z,\sigma) = \frac{1}{\pi\sigma} \frac{1}{c(x;z,\sigma)}$, with $c(x; z,\sigma) = 1 + \left(\frac{x-z}{\sigma}\right)^2$ is continuous and differentiable on $\mathbb{R} \times \mathbb{R}^{+*}$ with

$$\begin{aligned}
\partial_z g(x; z,\sigma) &= \frac{2(x - z)}{\pi\sigma^3 c^2(x; z,\sigma)} \\
\partial_\sigma g(x; z,\sigma) &= \frac{1}{\pi\sigma^2 c(x; z,\sigma)} - \frac{2}{\pi\sigma^2 c^2(x; z,\sigma)}.
\end{aligned}$$

Moreover, on can check that we have

$$\int_{\mathbb{R}} |\partial_z g(x; z,\sigma)|^2 \frac{dx}{g(x; z,\sigma)} = \int_{\mathbb{R}} \frac{4(x - z)^2}{\pi\sigma^3 c^3(x; z,\sigma)} dx < \infty$$

and

$$\int_{\mathbb{R}} |\partial_\sigma g(x; z,\sigma)|^2 \frac{dx}{g(x; z,\sigma)} = \int_{\mathbb{R}} \frac{1}{\pi\sigma^3 c(x; z,\sigma)} \left[ 1 - \frac{2}{c(x; z,\sigma)} \right]^2 dx < \infty.$$

- With the results of [54], we get that there is a constant $a^* > 0$ such that

$$\forall P_\theta \in \mathscr{Q}(\lambda^*), a^* \frac{||\theta - \bar{\theta}||^2}{1 + ||\theta - \bar{\theta}||^2} \leq h^2(P^*,P_\theta). \quad \square$$

**Proof of Lemma 2.12**

- Let $f$ be any function in $\mathbf{Span}_{\mathbb{R}}(A) \cap \mathbf{Span}_{\mathbb{R}}(B)$. Therefore there are constants $(\lambda_{g,i,j})_{\substack{1\leq i\leq K,\\ 0\leq j\leq n}}$ and $(\lambda_{c,i,l,j})_{\substack{1\leq i\leq K,\\ 0\leq j\leq 1\leq l\leq 2}}$ such that

$$f(x) = \sum_{i=1}^{K}\sum_{j=0}^{n} \lambda_{g,i,j} x^j \phi_{\sigma_i}(x - z_i) = \sum_{i=1}^{K}\sum_{l=1}^{2}\sum_{j=0}^{1} \lambda_{c,i,l,j} \frac{x^j}{c^l(x; z_i, \sigma_i)}.$$

  Since $f \in \mathbf{Span}_{\mathbb{R}}(A)$, we have $f(x) = o_{\pm\infty}(x^{-k}), \forall k \in \mathbb{N}$. Therefore $\lambda_{c,i,l,j} = 0$ for all $i,j,l$ and $f = 0$. This proves $\mathbf{Span}_{\mathbb{R}}(A) \cap \mathbf{Span}_{\mathbb{R}}(B) = \{0\}$.

- One can check that $>$ is a strict total order such that

$$(z_1,\sigma_1) > (z_2,\sigma_2) \Rightarrow x^j \phi_{\sigma_2}(x - z_2)/\phi_{\sigma_1}(x - z_1) \xrightarrow[x\to+\infty]{} 0,$$

  for any $j \in \mathbb{N}$. Let $\lambda$ be such that $\sum\limits_{i,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i) = 0$ for all $x$. Without loss of generality, we assume $(z_1,\sigma_1) > \cdots > (z_K,\sigma_K)$. Therefore,

$$\begin{aligned} 0 &= \sum_{i,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i) \\ &= \sum_{i,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i)/\phi_{\sigma_1}(x - z_1) + \sum_j \lambda_{1,j} x^j \\ &= \sum_j \lambda_{1,j} x^j + o_{+\infty}(1). \end{aligned}$$

  It implies that $\lambda_{1,j} = 0$ for all $j$. Then, we have $\sum\limits_{i\geq 2,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i) = 0$. By induction, we get that $\lambda = 0$ which proves that the family is indeed linearly independent.

- The partial fraction decomposition theorem implies that $B$ is linearly independent. □

This concludes the proof of Theorem 2.9.

### 2.D.3 Proof of Theorem 2.5

We apply Theorem 2.1 and Lemma 2.11 (see page 69) with $j^* = K$.

## 2.E Two-component mixture models

This section gathers the proofs of the results for the two-component mixture model with one known component, namely Theorems 2.6 and 2.7.

### 2.E.1 Proof of Theorem 2.6

We take $M = ||z^*||_\infty + 1$ to have (2.55). With Proposition 2.2, there exists a positive constant $C$ (depending on $\phi$ and $M$) such that for all $z \in [-M,M]^d$, and all $\lambda \in [0,1]$, we have

$$C(\phi,M)||z^*||^2 \left( ||z||^2 (\lambda^* - \lambda)^2 + (\lambda^*)^2 ||z^* - z||^2 \right) \leq ||p_{\lambda^*,z^*} - p_{\lambda,z}||^2.$$

One can prove (using Proposition 2.1 in [41] and $\lambda^* \neq 0$) that we have

$$\inf_{\substack{z\notin[-M,M]^d,\\ \lambda\in[0,1]}} ||p_{\lambda^*,z^*} - p_{\lambda,z}||^2 > 0. \tag{2.55}$$

Therefore, there is a constant $C(\phi,\lambda^*,z^*)$ such that for all $z \in \mathbb{R}^d$ and all $\lambda \in [0,1]$,

$$C(\phi,\lambda^*,z^*)\left(\left(||z||^2 \wedge 1\right)(\lambda^*-\lambda)^2 + (\lambda^*)^2\left(||z^*-z||^2 \wedge 1\right)\right) \leq ||p_{\lambda^*,z^*} - p_{\lambda,z}||_2^2.$$

Since $\phi$ is bounded, with inequality (2.24), there is another constant $C(\phi,\lambda^*,z^*)$ such that for all $z \in \mathbb{R}^d$ and $\lambda \in [0,1]$ we have

$$C(\phi,\lambda^*,z^*)\left(\left(||z||^2 \wedge 1\right)(\lambda^*-\lambda)^2 + (\lambda^*)^2\left(||z^*-z||^2 \wedge 1\right)\right) \leq h^2\left(P_{\lambda^*,z^*},P_{\lambda,z}\right).$$

One can check the following

$$\begin{aligned} h^2(P_{\lambda^*,z^*},P_{\hat{\lambda},\hat{z}}) \leq C(\phi,\lambda^*,z^*)(\lambda^*)^2\left(||z^*||^2 \wedge 1\right)/2 &\Rightarrow ||z^*-\hat{z}||^2 \wedge 1 \leq (||z^*||^2 \wedge 1)/4 \\ &\Rightarrow ||\hat{z}|| \wedge 1 \geq \frac{||z^*||}{2} \wedge 1. \end{aligned}$$

We use Theorem 2.1 for an upper bound on $h^2(P_{\lambda^*,z^*},P_{\hat{\lambda},\hat{z}})$. For $n \geq n_0(\phi,\lambda^*,z^*)$, with

$$n_0(\phi,\lambda^*,z^*) := \inf\left\{n \geq 1+V \middle| \frac{4(1+V)[1+\log(2n/(1+V))]}{nC(\lambda^*)^2\left(||z^*||^2 \wedge 1\right)} \leq C(\phi,\lambda^*,z^*)\right\},$$

for $0 < \xi \leq \xi_n = (1+V)[1+\log(2n/(1+V))]$, with probability at least $1-e^{-\xi}$ we have

$$\begin{aligned} Ch^2\left(P_{\lambda^*,z^*},P_{\hat{\lambda},\hat{z}}\right) &\leq \frac{1}{n}\left\{(1+V)\left[1+\log\left(\frac{2n}{(V+1)}\right)\right]+\xi\right\} \\ &\leq C \times C(\phi,\lambda^*,z^*)(\lambda^*)^2\left(||z^*||^2 \wedge 1\right)/2, \end{aligned}$$

where $C$ is the constant given in Theorem 2.1. Therefore, there is a new constant $C(\phi,\lambda^*,z^*)$ such that for $n \geq n_0$ and $\xi \in (0,\xi_n)$, with probability at least $1-e^{-\xi}$ we have

$$C(\phi,\lambda^*,z^*)\left((\lambda^*-\lambda)^2 + \left(||z^*-z||^2 \wedge 1\right)\right) \leq \frac{(1+V)\left[1+\log(2n/(1+V))\right]+\xi}{n}.$$

### 2.E.2 Proof of Theorem 2.7

We need some preliminary results before applying Theorem 2.1.

**Proposition 2.4.** *For $\lambda^* \in (0,1]$ and $z^* \neq 0$, there is a positive constant $C(\alpha,\lambda^*,z^*)$ such that for all $z \in \mathbb{R}$ and all $\lambda \in [0,1]$, we have*

$$h^2\left(P_{\lambda^*,z^*},P_{\lambda,z}\right) \geq C(\alpha,z^*,\lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1 \wedge |z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2(1 \wedge |z^*|)\right].$$

Since $s_\alpha$ is unimodal, the class of densities $\{x \mapsto s_\alpha(x-z), z \in \mathbb{R}\}$ is VC-subgraph with VC-dimension not larger than 10 (see Section 2.3.2). With Theorem 2.1 and Proposition 2.4, there exists a positive constant $C(\alpha,\lambda^*,z^*)$ such that for all $\xi > 0$, we have

$$C(\alpha,z^*,\lambda^*)\left[1 \wedge |\hat{z}-z^*|^{1-\alpha} + \left(\lambda^*-\hat{\lambda}\right)^2\right] \leq \frac{\log(n)+\xi}{n},$$

with probability at least $1-e^{-\xi}$.

**Proof of Proposition 2.4**

We write

$$f_z(x) = s_\alpha(x-z) = \frac{1-\alpha}{2|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}.$$

We define $g$ by

$$g(x) = \frac{2}{1-\alpha}\left(\sqrt{(1-\lambda^*)f_0(x)+\lambda^* f_{z^*}(x)} - \sqrt{(1-\lambda)f_0(x)+\lambda f_z(x)}\right)^2$$

such that

$$2h^2\left(P_{\lambda^*,z^*}, P_{\lambda,z}\right) = \frac{1-\alpha}{2}\int_{-\infty}^{+\infty} g(x)dx.$$

**Lemma 2.13.** *Assuming $z\cdot z^* > 0$ and $|z^*-z| \le \frac{1}{(1-\alpha)^{2/\alpha}}$. There exists $C(\alpha,z^*,\lambda^*) > 0$ such that*

$$\int g(x)dx \ge C(\alpha,z^*,\lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1\wedge|z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2\left(1\wedge|z^*|\right)\right].$$

**Lemma 2.14.** *For $z\cdot z^* \le 0$, we have*

$$\int g(x)dx \ge \lambda^*\alpha^2\frac{1\wedge\left[(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}\right]}{1-\alpha}.$$

**Lemma 2.15.** *For $|z-z^*| > \frac{1}{(1-\alpha)^{2/\alpha}}$ and $z^*\cdot z > 0$, we have*

$$\int g(x)dx = \lambda^*(1\wedge|z^*|).$$

Combining those three lemmas, there exists a positive constant $C(\alpha,z^*,\lambda^*)$ such that

$$h^2\left(P_{\lambda^*,z^*}, P_{\lambda,z}\right) \ge C'(\alpha,z^*,\lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1\wedge|z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2\left(1\wedge|z^*|\right)\right],$$

for all $\lambda$ in $[0,1]$ and $z$ in $\mathbb{R}$. Without loss of generality, we assume $z^* > 0$ through the proof of the lemmas. □

**Proof of Lemma 2.13**

Without loss of generality, we consider $z^* > 0$ for now.
• For $x \in ]-1,0[$, we have

$$g(x) = \frac{1}{|x|^\alpha}\left(\sqrt{1-\lambda^*+\lambda^*\frac{|x|^\alpha}{|x-z^*|^\alpha}\mathbb{1}_{|x-z^*|\in(0,1]}} - \sqrt{1-\lambda+\lambda\frac{|x|^\alpha}{|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}}\right)^2.$$

If $z^*\wedge z \ge 1$ then,

$$g(x) = \frac{1}{|x|^\alpha}\left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2$$

and

$$\int_{-1}^0 g(x)dx \ge \left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2\frac{1}{1-\alpha}.$$

Otherwise $z^*\wedge z \in (0,1)$ then for $x\in]-1, z^*\wedge z-1[$,

$$\int_{-1}^{z^*\wedge z-1} g(x)dx \ge \left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2\frac{1-(1-z\wedge z^*)^{1-\alpha}}{1-\alpha}.$$

Finally,

$$\int_{-1}^{0} g(x)dx \geq \left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2 \frac{1-(1-z\wedge z^*)_+^{1-\alpha}}{1-\alpha}.$$

• For $x \in ]z^* \vee z, z^* \vee z + 1[$, we have

$$g(x) = \frac{1}{|x - z^* \vee z|^\alpha} \left( \sqrt{(1-\lambda^*)\frac{|x-z^*\vee z|^\alpha}{|x|^\alpha}\mathbb{1}_{|x|\in(0,1]} + \lambda^* \frac{|x-z^*\vee z|^\alpha}{|x-z^*|^\alpha}\mathbb{1}_{|x-z^*|\in(0,1]}} \right.$$
$$\left. - \sqrt{(1-\lambda)\frac{|x-z^*\vee z|^\alpha}{|x|^\alpha}\mathbb{1}_{|x|\in(0,1]} + \lambda \frac{|x-z^*\vee z|^\alpha}{|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}} \right)^2.$$

- If $z < z^*$, with $V < \frac{1}{|z-z^*|}$, for $x \in ]z^*, z^* + V|z - z^*|[$, we have

$$\bullet \frac{|x-z^*|}{|x|} \leq V\frac{|z^*-z|}{z^*} \leq V,$$
$$\bullet \frac{|x-z^*|}{|x-z|} \leq \frac{V|z^*-z|}{(1+V)|z^*-z|} \leq V.$$

  We get

$$\int_{z^*}^{z^*+V|z-z^*|} g(x)dx \geq \left(\sqrt{\lambda^*} - \sqrt{V^\alpha}\right)^2 \int_{z^*}^{z^*+V|z-z^*|} \frac{dx}{|x-z^*\vee z|^\alpha}$$
$$= \left(\sqrt{\lambda^*} - \sqrt{V^\alpha}\right)^2 \frac{(V|z^*-z|)^{1-\alpha}}{1-\alpha}.$$

  We take $V = (\lambda^*)^{1/\alpha}(1-\alpha)^{2/\alpha} \leq \frac{(\lambda^*)^{1/\alpha}}{|z^*-z|} \leq \frac{1}{|z^*-z|}$, and we have

$$\int_{z^*}^{z^*+V|z-z^*|} g(x)dx \geq \lambda^*\alpha^2 \frac{(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}$$
$$= \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}.$$

- If $z \geq z^*$, we obtain the same way

$$\int_{z}^{z+1} g(x)dx \geq \frac{\lambda^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}.$$

Finally, for any $z^*$ in $\mathbb{R}$, using the following inequalities

$$\forall x,y \in [0,1], 1-(1-|x|)_+^{1-\alpha} \geq (1-\alpha)(1\wedge|x|) \text{ and } \left(\sqrt{x}-\sqrt{y}\right)^2 \geq (x-y)^2/4, \tag{2.56}$$

we get

$$\int g(x)dx \geq \mathbb{1}_{|z|\geq|z^*|}\left[\frac{(\lambda)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha} + (\lambda^*-\lambda)^2(1\wedge|z^*|)\right]$$
$$\mathbb{1}_{|z|<|z^*|}\left[\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha} + (\lambda^*-\lambda)^2(1\wedge|z|)\right].$$

- If $|z| \geq |z^*|$:

– if $\lambda > c\lambda^*$, then

$$\begin{aligned}\int g(x)dx &\geq \frac{(\lambda^*)^{1/\alpha}c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha} + (\lambda^*-\lambda)^2\,(1\wedge|z^*|)\\ &\geq C_1(\alpha,c)\left[(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha} + (1\wedge|z^*|)\,(\lambda^*-\lambda)^2\right]\end{aligned}$$

with $C_1(\alpha,c) = 1\bigwedge \frac{c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}$;

– otherwise $\int g(x)dx \geq (\lambda^*)^2(1-c)^2(1\wedge|z^*|)$,

$$(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha} + (1\wedge|z^*|)\,(\lambda^*-\lambda)^2 \leq (\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)$$

and finally

$$\begin{aligned}\int g(x)dx \geq &\frac{(\lambda^*)^2(1-c)^2(1\wedge|z^*|)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)}\\ &\times\left[(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha} + (1\wedge|z^*|)\,(\lambda^*-\lambda)^2\right].\end{aligned}$$

- If $|z| < |z^*|$:

– if $|z| \geq d|z^*|$, then

$$\begin{aligned}\int g(x)dx &\geq \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha} + (\lambda^*-\lambda)^2\,d(1\wedge|z^*|)\\ &\geq C_2(\alpha,d)\left[(\lambda^*)^{1/\alpha}\,|z-z^*|^{1-\alpha} + (\lambda^*-\lambda)^2\,(1\wedge|z^*|)\right],\end{aligned}$$

with $C_2(\alpha,d) = d\wedge\frac{\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}$;

– otherwise $\int g(x)dx \geq \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}}{1-\alpha}$ and

$$(\lambda^*)^{1/\alpha}\,|z-z^*|^{1-\alpha} + (\lambda^*-\lambda)^2\,(1\wedge|z^*|) \leq (\lambda^*)^{1/\alpha}\,\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)$$

and finally

$$\begin{aligned}\int g(x)dx \geq &\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}/(1-\alpha)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)}\\ &\times\left[(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha} + (1\wedge|z^*|)\,(\lambda^*-\lambda)^2\right].\end{aligned}$$

Finally,

$$\int g(x)dx \geq C(\alpha,z^*,\lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1\wedge|z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2\,(1\wedge|z^*|)\right],$$

with

$$C(\alpha,z^*,\lambda^*) = \min\left(1, \frac{c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}\right.$$
$$\frac{(\lambda^*)^2(1-c)^2(1\wedge|z^*|)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}}+(1\wedge|z^*|)},$$
$$d, \frac{\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha},$$
$$\left.\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}/(1-\alpha)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}}+(1\wedge|z^*|)}\right)$$
$$= \min\left(1, \frac{c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}, d,\right.$$
$$\frac{(\lambda^*)^2(1-c)^2(1\wedge|z^*|)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}}+(1\wedge|z^*|)},$$
$$\left.\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}/(1-\alpha)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}}+(1\wedge|z^*|)}\right). \quad \square$$

**Proof of Lemma 2.14**

Without loss of generality, we take $z^* > 0$.
• For $x \in ]z^*, z^*(1+a)[$, $a < (z^*)^{-1}$ we have

$$g(x) = \frac{1}{|x-z^*|^\alpha}\left(\sqrt{(1-\lambda^*)\frac{|x-z^*|^\alpha}{|x|^\alpha}\mathbb{1}_{|x|\in(0,1]}+\lambda^*}\right.$$
$$\left.-\sqrt{(1-\lambda)\frac{|x-z^*|^\alpha}{|x|^\alpha}\mathbb{1}_{|x|\in(0,1]}+\lambda\frac{|x-z^*|^\alpha}{|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}}\right)^2.$$

and

$$\frac{|x-z^*|}{|x-z|} \le \frac{|x-z^*|}{|x|} \le \frac{a}{1+a} \le a.$$

We get

$$\int_{z^*}^{z^*+a} g(x)dx \ge \left(\sqrt{\lambda^*}-\sqrt{a^\alpha}\right)^2 \int_{z^*}^{z^*+a}\frac{dx}{|x-z^*|^\alpha}$$
$$= \left(\sqrt{\lambda^*}-\sqrt{a^\alpha}\right)^2 \frac{(az^*)^{1-\alpha}}{1-\alpha}.$$

We take $a = (\lambda^*)^{1/\alpha}(1-\alpha)^{2/\alpha} \le \frac{1}{z^*}$, and we have

$$\int_{z^*}^{z^*+a} g(x)dx \ge \lambda^*\alpha^2\frac{(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}(z^*)^{1-\alpha}}{1-\alpha}.$$

Otherwise $a = 1/z^* \le (\lambda^*)^{1/\alpha}(1-\alpha)^{2/\alpha}$ and

$$\int_{z^*}^{z^*+a} g(x)dx \ge \lambda^*\alpha^2\frac{1}{1-\alpha}.$$

Finally,

$$\int_{z^*}^{z^*+1} g(x)dx \ge \lambda^*\alpha^2\frac{1\wedge\left[(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}(z^*)^{1-\alpha}\right]}{1-\alpha}. \quad \square$$

**Proof of Lemma 2.15**

Without loss of generality, we take $z^* \geq 0$.

- If $z \geq z^* + \frac{1}{(1-\alpha)^{2/\alpha}}$. For $x \in ]z^* \vee 1, (z^*+1) \wedge (z-1)[$, we have

$$g(x) = \frac{\lambda^*}{|x - z^*|^\alpha}.$$

One can prove that

$$|z - z^*| - 1 \geq \frac{1}{(1-\alpha)^{2/\alpha}} - 1 \geq 1.$$

  - If $z^* \geq 1$, then We get

$$\begin{aligned}\int_{z^*}^{z^*+1} g(x)dx &\geq \frac{\lambda^*}{1-\alpha}\left[1 \bigwedge |z - z^*| - 1\right]^{1-\alpha} \\ &\geq \frac{\lambda^*}{1-\alpha}.\end{aligned}$$

  - If $z^* \leq 1$, then

$$\begin{aligned}\int_{1}^{(z^*+1)\wedge(z-1)} g(x)dx &\geq \frac{\lambda^*}{1-\alpha}\left[1 \bigwedge (|z - z^*| - 1)^{1-\alpha} - (1 - z^*)^{1-\alpha}\right] \\ &\geq \frac{\lambda^*}{1-\alpha}\left[1 - (1 - z^*)^{1-\alpha}\right].\end{aligned}$$

- If $z^* \geq z + \frac{1}{(1-\alpha)^{2/\alpha}}$, we get

$$\int_{z^*}^{z^*+1} g(x)dx = \frac{\lambda^*}{1-\alpha}.$$

Finally,

$$\int_{z^*}^{z^*+1} g(x)dx = \frac{\lambda^*}{1-\alpha}\left[1 - (1 - z^*)_+^{1-\alpha}\right] \geq \lambda^*(1 \wedge z^*). \quad \square$$

## 2.F VC-subgraph classes of functions

For more detailed introductions to VC-subgraph classes we refer the reader to Van der Vaart & Wellner [84] (Section 2.6.5) and Baraud *et al.*[9] (Section 8).

**Definition 2.1.** *Definition 41 [9]*
*Let $\mathscr{C}$ be a non-empty class of subsets of a set $\Xi$. If $A \subset \Xi$ with $|A| = n$, then*

$$\Delta_n(\mathscr{C}, A) = |\{A \cap B, B \in \mathscr{C}\}| \text{ and } \Delta_n(\mathscr{C}) = \max_{A \subset \Xi, |A|=n} \Delta_n(\mathscr{C}, A).$$

*If $V = \sup\{n \in N | \Delta_n(\mathscr{C}) = 2n\} < +\infty$, then $\mathscr{C}$ is a VC-class with VC-dimension $V$ and VC-index $\overline{V} = \inf\{n \in N | \Delta_n(\mathscr{C}) < 2n\} = V + 1$. A class $\mathscr{F}$ of functions from a set $\mathscr{X}$ with values in $(-\infty, +\infty]$ is VC-subgraph with dimension $V$ and index $V$ if the class of subgraphs $\{(x, u) \in \mathscr{X} \times \mathbb{R}, f(x) > u\}$ as $f$ varies among $\mathscr{F}$ is a VC-class of sets in $\mathscr{X} \times \mathbb{R}$ with dimension $V$ and index $\overline{V}$.*

It immediately follows from this definition the following:

- if $\mathscr{F}$ is VC-subgraph with dimension $V$, then any subset $\mathscr{G} \subset \mathscr{F}$ is VC-subgraph with dimension at most $V$,

- if $\mathscr{F}$ is a finite set, $\mathscr{F}$ is VC-subgraph and its dimension is not larger than $V = \log_2(|\mathscr{F}|) \vee 1$.

The main reason for using VC-subgraph theory is the uniform entropy property. Namely, if $\mathscr{F}$ is a VC-subgraph set of measurable functions on $(\mathscr{X},\mathcal{X})$ with VC-dimension $V$ and $||f||_\infty \leq 1$ for all $f \in \mathscr{F}$, it follows from Lemma 1 in Baraud & Chen [12] that, for any probability $P$ on $(\mathscr{X},\mathcal{X})$ we have

$$N\left(\epsilon,\mathscr{F},L_r(P)\right) \leq e(V+1)(2e)^V \left(\frac{2}{\epsilon}\right)^{rV}.$$

## 2.F.1 Proof of Lemma 2.1

Let $\mathrm{Cov}_{+*}(d)$ be the set of $d \times d$ symmetric and positive-definite matrices. The normal distributions on $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathrm{Cov}_{+*}(d)$ admits $g_{\mu,\Sigma}$, defined by

$$g_{\mu,\Sigma} : x \mapsto \frac{\exp\left(-\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)\right)}{\sqrt{(2\pi)^k\left|\Sigma\right|}},$$

as a density with respect to the Lebesgue measure on $\mathbb{R}^d$, where $|\Sigma|$ denotes the determinant of $|\Sigma|$. We have

$$\begin{aligned}\log(g_{\mu,\Sigma}(x)) &= -\frac{1}{2}\log\left((2\pi)^k\left|\Sigma\right|\right) - \frac{1}{2}\left(x-\mu\right)^T\Sigma^{-1}\left(x-\mu\right)\\ &= -\frac{1}{2}\log\left((2\pi)^k\left|\Sigma\right|\right) - \frac{1}{2}\mu^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma x.\end{aligned}$$

For the location-scale family $\mathcal{G}_d := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d, \Sigma \in \mathrm{Cov}_{+*}\right\}$, we have

$$\mathcal{G}_d \subset \exp \circ \left\{x \mapsto a + \sum_{i\leq j} b_{i,j}x_ix_j + \sum_{i=1}^d c_ix_i; a \in \mathbb{R}, (b_{ij})_{i\leq j} \in \mathbb{R}^{d(d+1)/2}, c \in \mathbb{R}^d\right\}.$$

Since $\left\{x \mapsto a + \sum\limits_{i\leq j} b_{i,j}x_ix_j + \sum\limits_{i=1}^d c_ix_i; a \in \mathbb{R}, (b_{ij}) \in \mathbb{R}^{d(d+1)/2}, c \in \mathbb{R}^d\right\}$ is a vector space of dimension $1 + d(d+3)/2$ and exp is monotone, we get that $V(\mathscr{G}_d) \leq 3 + \frac{d(d+3)}{2}$. For $\Sigma \in \mathrm{Cov}_{+*}(d)$ fixed, the location family $\mathcal{G}_{loc}(\Sigma) := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d\right\}$, we have

$$\mathcal{G}_{loc}(\Sigma) \subset \exp \circ \left(x \mapsto -\frac{x^T\Sigma x}{2} + \left\{x \mapsto a + \sum_{i=1}^d b_ix_i; a \in \mathbb{R}, b \in \mathbb{R}^d\right\}\right).$$

With similar arguments and the fact that $x \mapsto -\frac{x^T\Sigma x}{2}$ is a fixed function, we have $V\left(\mathcal{G}_{loc}(\Sigma)\right) \leq 3 + d$.

## 2.F.2 Proof of Lemma 2.2

The different arguments used in this proof are from Proposition 42 of Baraud *et al.*[9] and Lemmas 2.6.15 and 2.6.16 from van der Vaart & Wellner [84]. We remind the reader that the VC-index is the VC-dimension plus 1.

- For the Cauchy location-scale family, we have

$$\mathcal{C} = \square^{-1} \circ \left\{x \mapsto \pi\sigma\left[1 + \left(\frac{x-z}{\sigma}\right)^2\right]; \sigma > 0, z \in \mathbb{R}\right\},$$

where $\square^{-1}$ is the inverse function on $(0,+\infty)$. Since

$$\left\{x \mapsto \pi\sigma\left[1+\left(\frac{x-z}{\sigma}\right)^2\right]; \sigma>0, z\in\mathbb{R}\right\} \subset \mathbb{R}_2[x] = \left\{x\mapsto ax^2+bx+c; (a,b,c)\in\mathbb{R}^3\right\}$$

and $\square^{-1}$ is monotone, we get that $V(\mathcal{C}) \leq 3+2$.

- For univariate normal distribution, it is a direct consequence of Lemma 2.1.
- We have

$$\begin{aligned}\mathcal{L} &= \left\{x\mapsto \frac{1}{2b}\exp\left(-\frac{|x-z|}{b}\right); z\in\mathbb{R}, b>0\right\}\\ &= \exp\circ\left\{x\mapsto -\log(2b)+b^{-1}[(x-z)\wedge(z-x)]; z\in\mathbb{R}, b>0\right\}\\ &\subset \exp\circ\left(\{x\mapsto ax+b; a,b\in\mathbb{R}\}\wedge\{x\mapsto ax+b; a,b\in\mathbb{R}\}\right).\end{aligned}$$

  Since exp is monotone and $\{x\mapsto ax+b; a,b\in\mathbb{R}\}$ is a vector space of dimension 2, we get that $\mathcal{L}$ is VC-subgraph with VC-index not larger than $V(\mathcal{L}) \leq 4.701\times 2(2+1)+1 = 29.206$.

- Azzalini & Capitanio [8] proved that the probability density function of the skew-normal distribution is unimodal, therefore the translation family $\mathcal{SG}_\alpha$ is VC-subgraph with VC-index at most 10 (see Section 2.3.2).

# Chapter 3

# Robust estimation for dependent observations and applications

**Abstract**

We observe $n$ possibly dependent random variables, the distribution of which is presumed to be stationary even though this might not be true, and we aim at estimating the stationary distribution. We establish a non-asymptotic deviation bound for the Hellinger distance between the target distribution and our estimator. If the dependence within the observations is small, the estimator performs as good as if the data were independent and identically distributed. In addition our estimator is robust to misspecification and contamination. If the dependence is too high but the observed process is mixing, we can select a subset of observations that is almost independent and retrieve results similar to what we have in the i.i.d. case. We apply our procedure to the estimation of the invariant distribution of a diffusion process and to finite state space hidden Markov models.

## 3.1 Introduction

We observe $n$ random variables $X_1,\dots,X_n$ with common distribution $P$ which is assumed to belong, or at least to be close enough, to a given model $\mathscr{M}$. Our aim is to estimate $P$ with an estimator $\hat{P}$ taking values in $\mathscr{M}$. These random variables are not necessarily independent however we assume that for indices $i \neq j$ with $|i-j|$ large enough, the distribution of the couple $(X_i,X_j)$ is close to $P \otimes P$. We also want our estimator to be robust to contamination and outliers.

When we actually dispose of an independent sample, this problem has already been investigated in Baraud *et al.*[9] and Baraud & Birgé [11]. They provide a non-asymptotic deviation bound for the Hellinger distance $h$ between $P$ and their $\rho$-estimator. For two probability distributions $P$ and $Q$ on the same measurable space, the Hellinger distance $h(P,Q)$ between $P$ and $Q$ is given by

$$h^2(P,Q) = \frac{1}{2}\int \left(\sqrt{dP/d\mu} - \sqrt{dQ/d\mu}\right)^2 d\mu,$$

where $\mu$ is any measure that dominates both $P$ and $Q$, the result being independent of $\mu$. It is shown in those articles that the $\rho$-estimator is robust in the following sense. Even if the variables $X_i$ do not have a common distribution $P$ but marginals $P_i$ such that most of them are relatively close to a distribution $P \in \mathscr{M}$, then the $\rho$-estimator is almost as efficient as when the data is i.i.d.with common distribution $P$. The obtained risk bounds are minimax, up to a logarithmic factor, when the model is well-specified and are not significantly deteriorated as long as the approximation term $n^{-1}\sum_{i=1}^{n} h^2(P_i,P)$ is relatively small in the misspecified case.

We want to obtain similar results when we do not satisfy the independence assumption but the observations are almost independent. This can happen for processes with mixing properties. We only focus on the theoretical aspects and performances of our estimation method. We prove a general result, Theorem 3.1, which gives a bound in expectation for the risk of our estimator $\hat{P}$ with respect to an Hellinger-type loss. This result is free of any assumption on the data and the risk bound is the sum of three terms: the approximation term mentioned above, a dimension which measures the complexity of the model $\mathscr{M}$, and a dependence term which measures how far the observations are from being independent. We quantify the dependence within the sample using Kullback-Leibler divergence of the joint distribution from the product of the marginal distributions. Our risk bound is as good as when the data is independent as long as the dependence term is not bigger than the other terms. We have the following approach for when the dependence term is too big. We split our data in order to get a subset of the original observations for which the dependence term is small enough.

We apply this method for the estimation of an invariant distribution of a discretely observed diffusion process. Under some condition the stationary solution of a Langevin equation is mixing

and its invariant distribution has a log-concave density with respect to the Lebesgue measure. We can refer to the literature on the estimation of a log-concave density in the i.i.d. context and adapt our procedure to this situation. We obtain convergence rates for our estimator in any dimension. Those rates are similar to the minimax rates for i.i.d. estimation, with a worse logarithmic power.

Our main application is hidden Markov models (HMMs). These models are widely applied to model state dependent processes where the state process is Markovian but is not observed. We refer the interested reader to Mor, Garhwal and Kumar [74] for a review of applications of HMMs. Let $Y_1,\dots,Y_N,H_1,\dots,H_N$ be random variables. We say that $(Y_i,H_i)_{1\leq i\leq N}$ is a hidden Markov model (HMM) if $(H_i)_i$ is a Markov chain and each variable $Y_i$ only depends on the associated $H_i$. In particular the variables $Y_1,\dots,Y_N$ are independent conditionally on $(H_i)_i$. It is called a hidden Markov model as the Markov chain $(H_i)_i$ is typically not observed and $(Y_i)_i$ is the only accessible data.

We focus on homogeneous finite state space HMMs. Such processes can be completely described by the number $K$ of hidden states $h_1,\dots,h_K$, the initial distribution $w$ and the transition matrix $Q$ of the hidden Markov chain, and the set of emission distributions $F = (F_1,\dots,F_K)$, where $F_k$ is the conditional distribution of $Y_i$ given $H_i = h_k$. In that case we say that $(Y_i,H_i)_i$ is a HMM with parameters $(K,w,Q,F)$. Because the hidden state space does not have a particular importance, we will always assume it is of the form $\{1,2,\dots,K\}$. For a particular class of distributions $\mathscr{F}$ there is a minimal value of $K$ such that $(Y_i,H_i)_i$ is a HMM with parameters $(K,w,Q,F)$ with $F_1,\dots,F_K \in \mathscr{F}$. This value of $K$ is called the order of the HMM (with respect to $\mathscr{F}$). Typically one aims at estimating these parameters from stationary observations $(Y_i)_{1\leq i\leq N}$.

Numerous estimation methods have been developed to estimate some or all of the parameters. Cappé *et al.*[55] provide an overall survey of the different results in the literature. Most theoretical guarantees are either asymptotic or restricted to specific parametric models. Lehéricy [63] provided non-parametric and non-asymptotic results for a penalized least squares estimator with the following approach. They first estimate the distribution $P_L = P_{\pi^*,Q^*,F^*}$ of $L$ consecutive observations $Y_i,Y_{i+1},\dots,Y_{i+L-1}$ of a stationary ergodic HMM with parameters $(K^*,\pi^*,Q^*,F^*)$, where $P_{w,Q,F}$ is defined by

$$P_{w,Q,F} = \sum_{1\leq k_1,\dots,k_L\leq K} w_{k_1} Q_{k_1,k_2}\dots Q_{k_{L-1},k_L} \bigotimes_{l=1}^{L} F_{k_l}. \tag{3.1}$$

They use model selection to consistently estimate the order $K^*$. When the estimation of the order is correct, it is possible to deduce the different parameters from $P_L$ for $L$ large enough. They show that $L \geq 3$ is enough for linearly independent emission densities. They lower bound the $L^2$-distance between densities by a distance on the parameters. Therefore a risk bound for the estimation of $P_L$ is enough to obtain risk bounds for the parameter estimators.

However their estimator is not robust to misspecification nor to contamination and there is no estimator that tackles this problem for general finite state space HMMs. The estimation method we propose aims at solving this problem. For the sake of simplicity we do not aim at estimating the order $K^*$. We do not look into this particular aspect in this paper however model selection can be considered to automatically choose an order from the data. This is to be treated in a subsequent paper.

We use the tools we develop in the first part with $\mathscr{M}$ containing distributions of the form $P_{w,Q,F}$ to obtain a robust estimator $\hat{P}$ of $P_L$, hence $\hat{P}$ being of the form $\hat{P} = P_{\hat{w},\hat{Q},\hat{F}}$. We have a general risk bound for $\hat{P}$ which is free of any assumption on the data from which we obtain convergence rates when we assume that the observations come from an ergodic finite state space HMM. In particular the stationarity of the observations is not necessary. We show that the performance of our estimator is not significantly worsened when the model is misspecified as

long as the distance to the true distribution is small compared to the rate we have in the well-specified case. Similarly the performance of our estimator is not deteriorated by contamination as long as the contamination rate is not too big.

We can deduce risk bounds for the parameter estimators $\hat{w},\hat{Q},\hat{F}$ under some conditions on the model $\mathscr{M}$. We need an inequality of the form

$$d\left((w,Q,F),(\overline{w},\overline{Q},\overline{F})\right) \leq C\left(\overline{w},\overline{Q},\overline{F}\right) h^2\left(P_{w,Q,F}, P_{\overline{w},\overline{Q},\overline{F}}\right), \forall P_{w,Q,F} \in \mathscr{M}. \tag{3.2}$$

We obtain convergence rates for the estimation of the parameters when the model is well specified. If the model is misspecified but $\overline{P} = P_{\overline{w},\overline{Q},\overline{F}}$ is the best approximation of $P_L$ within our model our estimators $\hat{w},\hat{Q},\hat{F}$ should be close to $\overline{w},\overline{Q},\overline{F}$ when this approximation is relatively good.

It is possible to use the results that already exist for the $L_2$-norm to obtain an inequality like (3.2) when the densities are bounded. For two probability distributions $P$, $Q$ dominated by a positive measure $\mu$, we have

$$||p-q||_2^2 \leq 4(||p||_\infty + ||q||_\infty)h^2(P,Q), \tag{3.3}$$

where $p = dP/d\mu$ and $q = dQ/d\mu$. It is also possible to prove inequalities directly for the Hellinger distance in some cases. We do so for models with emission densities that belong to exponential families with some regularity. We also consider an example with classes of emission densities that are unbounded and not even square integrable in some cases. For this example we obtain rates that are faster than the parametric rate for one of the parameters. Classical estimators such as the maximum likelihood or least-squares estimators do not apply as the considered densities are unbounded.

Our estimation method requires the statistician to select themself a subset of the observations that should be almost independent. This is not possible without any knowledge on the distribution of the data. We propose to overcome this restriction and provide a way to automatically select an almost independent subset of observations when we dispose of a second set of observations independent from the first one. We obtain a general risk bound and show that for ergodic HMMs we retrieve the same rate of convergence as when the optimal way of selecting observations is known. This method is still robust to misspecification and contamination.

The paper is organized as follows. In Section 3.2, we present our estimation procedure and our main result in a general framework. We consider the application to the estimation of the invariant distribution of a diffusion process in Section 3.3. We dedicate Section 3.4 to finite state space hidden Markov models. Finally, we propose a complete procedure for situations in which we do not know the mixing regime in Section 3.5. The proofs of all the different results can be found in the appendix.

**Notation.** For a set $A$, we denote by $|A|$ its cardinal which can be infinite. For an integer $k$, we denote by $[k]$ the set $\{1,2,\ldots,k\}$. We denote by $\mathbb{R}_+$ the set of non-negative real numbers. For a real number $x$, we denote by $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) the only integer $k$ satisfying $k-1 < x \leq k$ (resp. $k \leq x < k+1$). For a random variable $X$ we denote by $\mathcal{L}(X)$ its probability distribution. The notation $C(\theta,\alpha,\beta)$ means that $C(\theta,\alpha,\beta)$ is a constant that depends on the parameters $\theta$, $\alpha$ and $\beta$. It can change from one inequality to the other. On the other hand a constant written $C$ will be universal. For a real number $x$ we denote by $x_+$ its positive part given by $x_+ = x \vee 0$.

## 3.2 Construction of the estimator and main result

Let $X_1,\ldots,X_n$ be $n$ possibly dependent random variables on the measurable space $(\mathscr{X},\mathcal{X})$. Our aim is to estimate their marginal distribution $P^*$ doing as if they were identically distributed, even though this might not be exactly the case. We denote by $\mathscr{P}_X$ the class of all probability

distribution on $(\mathscr{X},\mathcal{X})$ and for $i \in [n]$ by $P_i = \mathcal{L}(X_i) \in \mathscr{P}_X$ the true marginal distribution of $X_i$. We also want our estimator of $P^*$ to be robust to misspecification, contamination and outliers. The $\rho$-estimators developed by Baraud, Birgé and Sart in [9] and [11] are perfectly adapted to this task when the observations are independent. We prove that their performances remain almost as good when the observations are close to being independent.

### 3.2.1 Reminders of $\rho$-estimation

We denote by $\psi$ the function given by

$$\psi : \left| \begin{array}{l} [0,+\infty] \to [-1,1] \\ x \mapsto \frac{x-1}{x+1} \end{array} \right. . \tag{3.4}$$

Let $\mathscr{M}$ be a countable subset of $\mathscr{P}_X$ such that there is an associated set of density functions $\mathcal{M}$ with respect to a $\sigma$-finite measure $\mu$. For $n \geq 1$, we denote by $\mathbf{T}_n$ and $\boldsymbol{\Upsilon}_n$ the functions given by

$$\mathbf{T}_n : \left| \begin{array}{l} \mathscr{X}^n \times \mathcal{M} \times \mathcal{M} \to [-1,1] \\ (\mathbf{x},q,q') \mapsto \sum_{k=1}^{n} \psi\left(\sqrt{\frac{q'(x_i)}{q(x_i)}}\right) \end{array} \right. \tag{3.5}$$

with the convention $0/0 = 1$, $a/0 = +\infty$ for all $a > 0$, and

$$\boldsymbol{\Upsilon}_n : \left| \begin{array}{l} \mathscr{X}^n \times \mathcal{M} \\ (\mathbf{x},q) \mapsto \sup_{q' \in \mathcal{M}} \mathbf{T}_n(\mathbf{x},q,q') \end{array} \right. . \tag{3.6}$$

For $\mathbf{x}$ in $\mathscr{X}^n$, we define the (nonvoid) set $\mathscr{E}_n(\mathbf{x})$ by

$$\mathscr{E}_n(\mathbf{x}) = \left\{ Q = q \cdot \mu \middle| q \in \mathcal{M}, \boldsymbol{\Upsilon}_n(\mathbf{x},q) < \inf_{q' \in \mathcal{M}} \boldsymbol{\Upsilon}_n(\mathbf{x},q') + 11.36 \right\}. \tag{3.7}$$

We denote by $\hat{P}(n,\mathbf{X},\mathscr{M})$ any measurable element of the closure of $\mathscr{E}_n(\mathbf{X})$ with respect to the Hellinger distance and we call it a $\rho$-estimator on $\mathscr{M}$. The constant 11.36 is given by (7) and (19) in [11] but can be replaced by any smaller positive number.

One of the main results of $\rho$-estimation is Theorem 1 in [11]. For independent random variables $X_1,\ldots,X_n$, any $\rho$-estimator $\hat{P} = \hat{P}(n,\mathbf{X},\mathscr{M})$ satisfies an inequality of the form

$$\mathbb{P}\left( \frac{C}{n} \sum_{i=1}^{n} h^2(P_i,\hat{P}) \leq \inf_{Q \in \mathscr{M}} n^{-1} \sum_{i=1}^{n} h^2(P_i,Q) + \frac{D_n(\mathscr{M}) + \xi}{n} \right) \geq 1 - e^{-\xi}, \tag{3.8}$$

where $C$ is a positive numeric constant and $D_n(\mathscr{M}) \geq 1$ is a dimension term that measures the complexity of the model $\mathscr{M}$. This dimension term corresponds to a bound on the $\rho$-dimension. It is an important feature of $\rho$-estimation as it determines the bound on the convergence rate of the estimator. If we actually dispose of i.i.d. observations with common distribution $\overline{P}$ in $\mathscr{M}$, we get

$$\mathbb{P}\left( Ch^2(\overline{P},\hat{P}) \leq \frac{D_n(\mathscr{M}) + \xi}{n} \right) \geq 1 - e^{-\xi},$$

which leads to the bound $D_n(\mathscr{M})/n$ on the convergence rate, up to a multiplicative constant. The notion of $\rho$-dimension is formally introduced in the appendix (Section 3.B).

### 3.2.2 From independent to dependent data

To extend the previous result to non-independent samples, we use the following idea which is not specific to our framework. We state this basic principle in a general context. Let $\hat{\theta} : \mathscr{X}^n \to \Theta$ be an estimator of some quantity $\theta \in \Theta$. The next result is proven in Section 3.A.1.

**Lemma 3.1.** *Let $l : \Theta \times \Theta \to \mathbb{R}_+$ be a loss function, $\mathbf{P}$,$\mathbf{Q}$ two distributions on a measurable space $(\mathscr{Y},\mathcal{X})$ and $\beta \in (0,1]$. Assume that when $\mathbf{Y}$ has distribution $\mathbf{P}$*

$$\mathbb{P}_{\mathbf{X}\sim\mathbf{P}}\left(l\left(\hat{\theta}(\mathbf{X}),\theta\right) \geq A + \frac{B+\xi^\beta}{n}\right) \leq e^{-\xi}, \forall \xi > 0, \tag{3.9}$$

*then, when $\mathbf{X}$ has distribution $\mathbf{Q}$*

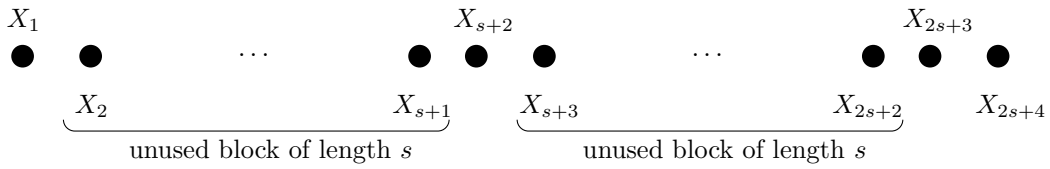
$$\mathbb{E}_{\mathbf{X}\sim\mathbf{Q}}\left[l\left(\hat{\theta}(\mathbf{X}),\theta\right)\right] \leq A + \frac{B + \left(2 + \frac{3}{2}\mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right)\right)^\beta}{n},$$

*where $\mathbf{K}$ is the Kullback-Leibler divergence given by*

$$\mathbf{K}(Q||P) = \begin{cases} \int \log\left(\frac{dQ}{dP}\right) dQ \text{ if } Q \ll P, \\ +\infty \text{ otherwise.} \end{cases}$$

Deviation inequalities for $\rho$-estimators $\widehat{\theta}$ have been established under the assumption that one observes independent random variables $X_1,\ldots,X_n$, hence when the distribution of $\mathbf{X} = (X_1,\ldots,X_N)$ is $\mathbf{P} = \mathcal{L}(X_1)\otimes\cdots\otimes\mathcal{L}(X_n)$. Our idea is to apply Lemma 3.1 with a distribution $\mathbf{Q} \ll \mathbf{P}$, which is not a product probability, in order to establish a risk bound for the estimator $\widehat{\theta}$ when the observations $X_1,\ldots,X_n$ are possibly dependent. The quantity $\mathbf{K}(\mathbf{Q}||\mathbf{P})$ measures thus a departure from independence. We consider subsets of the original data $X_1,\ldots,X_n$ when this quantity is too big.

Let $n$ be larger than 2. We build subsets of observations by taking them separated by blocks of length $s \in \mathbb{N}$, as described in the diagram below.




Formally, for $s \in \{0,1,\ldots,s_{\max}\}$,$s_{\max} := \lfloor (n-2)/2 \rfloor$ and $b \in [s+1]$, we define

$$n(s,b) := \left\lfloor \frac{n+s+1-b}{1+s} \right\rfloor \geq 2,$$

for $i \in [n(s,b)]$

$$X_i^{(s,b)} := X_{b+(i-1)(s+1)} \in \mathscr{X}, \forall i \in [n(s,b)], \tag{3.10}$$

and

$$\mathbf{X}^{(s,b)} := \left(X_i^{(s,b)}, i \in [n(s,b)]\right).$$

We obtain $s+1$ subsets $\mathbf{X}^{(s,1)},\ldots,\mathbf{X}^{(s,s+1)}$ with sizes $n(s,1),\ldots,n(s,s+1)$ respectively. For each block $b \in [s+1]$, we consider the probabilities $\mathbf{P}_{s,b}^*$ and $\mathbf{P}_{s,b}^{ind}$ which are defined by

$$\mathbf{P}_{s,b}^* := \mathcal{L}\left(\mathbf{X}^{(s,b)}\right) \text{ and } \mathbf{P}_{s,b}^{ind} := \bigotimes_{i=1}^{n(s,b)} \mathcal{L}\left(X_i^{(s,b)}\right). \tag{3.11}$$

We denote for short $\mathbf{P}^* := \mathbf{P}_{0,1}^*$ the distribution of $\mathbf{X} = (X_1,\ldots,X_n)$ and $\mathbf{P}^{ind} := \mathbf{P}_{0,1}^{ind} = \mathcal{L}(X_1)\otimes\cdots\otimes\mathcal{L}(X_n)$. Our estimator is obtained with the following statistical procedure.

1. Let $s$ be in $\{0,1,\ldots,s_{\max}\}$. For $b$ in $[s+1]$, we denote by $\hat{P}_{s,b}$ the estimators given by
$$\hat{P}_{s,b} := \hat{P}\left(n(s,b),\mathbf{X}^{(s,b)},\mathscr{M}\right),$$
where the $\rho$-estimator $\hat{P}\left(n(s,b),\mathbf{X}^{(s,b)},\mathscr{M}\right)$ is defined in Section 3.2.1.

2. We denote by $\hat{P}_s = \hat{P}_s(\mathbf{X},\mathscr{M})$ any element of $\mathscr{M}$ that satisfies
$$\sum_{b=1}^{s+1} n(s,b)h^2\left(\hat{P}_{s,b},\hat{P}_s\right) \leq \inf_{Q\in\mathscr{M}} \sum_{b=1}^{s+1} n(s,b)h^2\left(\hat{P}_{s,b},Q\right) + \iota, \tag{3.12}$$
where $\iota$ is any fixed constant in $(0,1273]$.

### 3.2.3 Main result

We assume that the $\rho$-dimension function (see Section 3.B) is uniformly bounded by a function $m \mapsto D_m(\mathscr{M}) \geq 1$ which is non-decreasing.

**Theorem 3.1.** *For any random variables $X_1,\ldots,X_n$ on $(\mathscr{X},\mathcal{X})$, the estimator $\hat{P}_s = \hat{P}_s(\mathbf{X},\mathscr{M})$ given by (3.12) satisfies*
$$\mathbb{E}_{\mathbf{P}^*}\left[n^{-1}\sum_{i=1}^n h^2\left(P_i,\hat{P}_s\right)\right] \leq \frac{c_0}{n}\inf_{Q\in\mathscr{M}}\sum_{i=1}^n h^2(P_i,Q) \tag{3.13}$$
$$+c_1\frac{(s+1)}{n}\left[17+D_{n(s,1)}(\mathscr{M})\right] + \frac{c_2}{n}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right),$$
*where $c_0 = 602$, $c_1 = 20056/4.7$ and $c_2 = 30084$.*

The proof of this result is postponed to Section 3.B.1. One can check that we do not need any assumption on the data to obtain this result. We only need a condition on the model $\mathscr{M}$ which is chosen by the statistician. However a posteriori assumptions are necessary to make this bound meaningful. It follows from the triangle inequality and $(a+b)^2 \leq 2a^2+2b^2$ for all non-negative numbers $a$ and $b$ that for any $\overline{P}\in\mathscr{M}$,
$$nh^2\left(\overline{P},\hat{P}_s\right) \leq 2\sum_{i=1}^n h^2\left(P_i,\hat{P}_s\right) + 2\sum_{i=1}^n h^2\left(P_i,\overline{P}\right).$$
We derive from (3.13) the following
$$C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq \frac{(s+1)D_{n(s,1)}(\mathscr{M})}{n} + n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right) \tag{3.14}$$
$$+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right),$$
where $C$ is a universal positive constant. Up to the factor $(s+1)$, the first term on the right-hand side of this inequality corresponds to the bound we would get if the data were truly i.i.d. with distribution $\overline{P}\in\mathscr{M}$. In this ideal situation, both the second and third term vanish. When the data are not identically distributed, the second term is not zero but its size remains small when most of the true marginal distributions $P_1,\ldots,P_n$ lie close enough to an element $\overline{P}\in\mathscr{M}$. The third term accounts for the fact that the data are possibly dependent. For a sufficiently large value of $s$, we expect the observations
$$\mathbf{X}^{(s,b)} := \left(X_b, X_{b+(s+1)},\ldots,X_{b+n(s,b)(s+1)}\right) \quad \text{with } b\in[s+1]$$
to be nearly independent, and consequently the quantity $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ to be small compared to the first term.

### 3.2.4 Robust properties of our estimator

The robustness properties of $\rho$-estimators in the independent context are illustrated in Section 5 [11]. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be the true process of interest such that $\mathcal{L}(X_i) = \overline{P}$ for all $i$ in $[n]$. We actually observe a contaminated version of it. Let $Z_1, \ldots, Z_n$ be random variables with any distributions. Let $E_1, \ldots, E_n$ be Bernoulli random variables such that

$$Y_i = E_i X_i + (1 - E_i) Z_i, \forall i \in [n]. \tag{3.15}$$

The next result shows that the mixing regime is not altered by independent contamination/outliers. It is proven in Section 3.B.2.

**Lemma 3.2.** *If $E_1, \ldots, E_n, Z_1, \ldots, Z_n$ and $\mathbf{X}$ are mutually independent, we have*

$$\mathbf{K}\left(\mathcal{L}\left(\mathbf{Y}\right) || \mathcal{L}(Y_1) \otimes \cdots \otimes \mathcal{L}(Y_n)\right) \leq \mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right) || \mathcal{L}(X_1) \otimes \cdots \otimes \mathcal{L}(X_n)\right).$$

We can deduce a corollary of Theorem 3.1 from this. We define $p_i$ by $\mathbb{P}(E_i = 1) = p_i$ for $i \in [n]$.

**Corollary 3.1.** *Let $\hat{P}_s = \hat{P}_s\left(\mathbf{Y}, \mathscr{M}\right)$ be the estimator given by (3.12). There is a positive universal constant $C$ such that in the situation of Lemma 3.2, we have*

$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \mathscr{M}\right) + n^{-1}\sum_{i=1}^{n}(1 - p_i) + \frac{(s+1)D_{n(s,1)}(\mathscr{M})}{n} + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right),$$

*where $\mathbf{P}^*_{s,b}$ and $\mathbf{P}^{ind}_{s,b}$ are given by (3.11).*

This result is proven in Section 3.B.3. Inspired by Hüber's contamination model, we consider the situation $\overline{P} \in \mathscr{M}$ and $p_i = 1 - \epsilon_{cont}$ for all $i \in [n]$. We get

$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq \epsilon_{cont} + \frac{(s+1)D_{n(s,1)}(\mathscr{M})}{n} + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right).$$

Our bound on the convergence rate is not deteriorated as long as the contamination rate $\epsilon_{cont}$ is small compared to the other terms. Equally, we can consider the case where the $E_i$ are deterministic, i.e. there is a subset $I \subset [n]$ such that $\mathbb{P}(E_i = 0) = \mathbb{1}_{i \in I}$. We get

$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq \frac{|I|}{n} + \frac{(s+1)D_{n(s,1)}(\mathscr{M})}{n} + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right).$$

As before, our bound on the convergence rate is not deteriorated as long as the proportion of outliers $|I|/n$ is small compared to the other terms on the right hand side.

### 3.2.5 The particular case of Markov chains

Under the assumption that $X_1, \ldots, X_n$ is a Markov chain, the quantity $\mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{(s,b)}_{ind}\right)$ can be written in a form given in the lemma below.

**Lemma 3.3.** *If $\mathbf{X}$ is a Markov chain,*

$$\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right) || \mathcal{L}\left(X_1\right) \otimes \cdots \otimes \mathcal{L}\left(X_n\right)\right) = \sum_{i=2}^{n} I(\sigma(X_i), \sigma(X_{i+1})),$$

*where*

$$I(\sigma(X_i),\sigma(X_{i+1})) := \mathbf{K}\left(\mathcal{L}(X_i,X_{i+1})||\mathcal{L}(X_i)\otimes\mathcal{L}(X_{i+1})\right). \tag{3.16}$$

*In particular for all $s$ in $\{0,1,\dots,s_{\max}\}$ and all $b$ in $[s+1]$,*

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{(s,b)}_{ind}\right) = \sum_{i=2}^{n(s,b)} I\left(\sigma(X_i^{(s,b)}),\sigma(X_{i+1}^{(s,b)})\right),$$

*where the $X_i^{(s,b)}$ are given by (3.10).*

This result is proven in Section 3.B.4. It tells us that for Markov chains we only need to consider the simpler quantities $I(\sigma(X_i),\sigma(X_{i+s+1}))$ referred to as *coefficients of information* by Bradley [19]. This result also extends to hidden Markov models.

**Lemma 3.4.** *If $(X_i,H_i)_{1\le i\le n}$ is a HMM, we have*

$$\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right) \le \sum_{i=2}^{n} I(\sigma(H_{i-1}),\sigma(H_i)).$$

*In particular for all $s$ in $\{0,1,\dots,s_{\max}\}$ and all $b$ in $[s+1]$,*

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{(s,b)}_{ind}\right) \le \sum_{i=1}^{n(s,b)-1} I\left(\sigma(H_{b+(i-1)(s+1)}),\sigma(H_{b+i(s+1)})\right).$$

The proof of this result is postponed to Section 3.B.5. This means that for HMMs we only need to consider the coefficients of information of the hidden chain. In what follows we consider different processes for which the coefficient of information has an exponential decay. In that case there exist positive constants $C$ and $r$ such that

$$n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \le Ce^{-rs},$$

for all $s$ in $\{0,1,\dots,s_{\max}\}$. For $s \ge r^{-1}\log n$ the quantity $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ is small compared to the first term on the right hand side in (3.14), as it cannot be of order smaller than $1/n$. Such a constant $r$ is usually not known in practice but taking $s$ of order $\log^2 n$ ensures that for $n$ large enough the quantity we consider remains small compared to the term $(s+1)D_{n(s,1)}(\mathscr{M})/n$. We pay the price of not knowing the constant $r$ with a worse logarithmic term in the latter quantity.

## 3.3 Estimation of the invariant distribution of a diffusion process

We consider some diffusion processes that have been investigated by Royer [78] and use the same vocabulary that they introduced.

### 3.3.1 Langevin equation

Let $d$ be a positive integer and $U:\mathbb{R}^d\to\mathbb{R}$ be a function of class $\mathcal{C}^2$. The Langevin equation is the following stochastic differential equation

$$dY_t = dB_t - \nabla U(Y_t)dt, \tag{3.17}$$

where $B=(B_t)_{t\ge0}$ is a $d$-dimensional Brownian motion. Its solution are called Kolmogorov processes in Royer [78]. We assume that $U$ satisfies the following.

**Assumption 3.1.** *The function $U$ is convex on $\mathbb{R}^d$ and there exists a positive constant $\underline{\lambda}(U)$ such that the smallest eigenvalue of the Hessian matrix $U''(x)$ at $x \in \mathbb{R}^d$ is not smaller than $\underline{\lambda}(U)$ for all $x$ in $\mathbb{R}^d$. Besides we have*

$$\inf_{x\in\mathbb{R}^d} \left\{ ||\nabla U(x)||_2^2 - \mathrm{Tr}\left(U''(x)\right) \right\} > -\infty, \tag{3.18}$$

*where $\mathrm{Tr}(A)$ is the trace of the matrix $A$.*

Under our assumption on the eigenvalues of $U''$, $\int_{\mathbb{R}^d} e^{-\alpha U(x)} dx$ is finite for all $\alpha > 0$ and we may define the probability measure $\overline{P}$ with density $\overline{p}$ with respect to the Lebesgue measure on $\mathbb{R}^d$ given by

$$\overline{p}(x) = Z^{-1} \exp(-2U(x)) \text{ with } Z = \int_{\mathbb{R}^d} e^{-2U(x)} dx. \tag{3.19}$$

The probability $\overline{P}$ is the invariant probability distribution with respect to the semi-group associated to the Langevin equation (see Lemma 2.2.23 [78]).

**Lemma 3.5.** *Let $(Y_t)_{t\geq 0}$ be a stationary solution of the Langevin equation associated to a convex function $U$ that satisfies Assumption 3.1. For all $s_0 > 0$, there exists a positive constant $C(U,s_0)$ such that for all $t > 0$ and $s \geq s_0$, we have*

$$I(\sigma(Y_t), \sigma(Y_{t+s})) \leq C(U,s_0) \exp(-2\underline{\lambda}(U)s).$$

This result is proven in Section 3.C.2. We aim to estimate $\overline{P}$ from discrete observations of a stationary Kolmogorov process.

### 3.3.2 The framework

We consider the following statistical model for the observations $X_1, X_2, \ldots, X_n$. For all $i \in [n]$, $X_i = Y_{t_i}$ where $\mathbf{Y} = (Y_t)_{t\geq 0}$ is a stationary solution of the Langevin equation (3.17) for some unknown convex function $U$ that satisfies Assumption 3.1 and $t_{i+1} = t_i + \Delta_t$ for all $i \in [n-1]$. As a consequence of (3.19), the $X_i$ are distributed according to the invariant measure $\overline{P}$ which has a log-concave density $\overline{p} : x \mapsto Z^{-1} \exp(-2U(x))$ with respect to the Lebesgue measure. We therefore consider the set of distributions that admit a log-concave density on $\mathbb{R}^d$ with respect to the Lebesgue measure. As usual, this describes our statistical model but we do not want to assume that it perfectly describes reality. In the following section we recall some results about the problem of estimating a log-concave density from i.i.d. observations.

### 3.3.3 log-concave densities

We refer to Kim & Samworth [58] for the problem of estimating log-concave densities from i.i.d. observations in low dimensions ($d \in [3]$). Kur *et al.* [60] investigated the same problem in higher dimensions ($d \geq 4$). We denote by $\mathcal{F}_d$ the set of upper semi-continuous, log-concave probability densities with respect to the Lebesgue measure, equipped with the $\sigma$-algebra it inherits as a subset of $L_1(\mathbb{R}^d)$. We denote by $\mathscr{F}_d$ the associated set of probability distributions on $\mathbb{R}^d$. For $f \in \mathcal{F}_d$, we define

$$\overline{x}_f := \int_{\mathbb{R}^d} x f(x) dx \in \mathbb{R}^d \text{ and } \Sigma_f := \int_{\mathbb{R}^d} (x - \mu_f)(x - \mu_f)^T f(x) dx \in \mathbb{R}^{d\times d}.$$

For a symmetric, positive-definite $d \times d$ matrix $\Sigma$, we denote by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ the smallest and largest eigenvalues respectively of $\Sigma$. For $0 < \lambda_- < \lambda_+ < \infty$ and $M > 0$, we define

$$\mathcal{F}_{\lambda_-,\lambda_+,M} := \left\{ f \in \mathcal{F}_d; ||\overline{x}_f|| \leq M, \Sigma \in \mathrm{Sym}(\lambda_-,\lambda_+) \right\},$$

where

$$\mathrm{Sym}(\lambda_-,\lambda_+) = \{\Sigma \text{ covariance matrix}, \lambda_- \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \lambda_+\}\,.$$

We denote by $\mathscr{F}_{\lambda_-,\lambda_+,M}$ the class of probability distributions associated to $\mathcal{F}_{\lambda_-,\lambda_+,M}$.

Given a subset $\mathscr{A}$ of a class $\mathscr{P}$ of probability distributions and $\epsilon \geq 0$, we say that $\mathscr{A}[\epsilon]$ is an $\epsilon$-net of $\mathscr{A}$ if $\mathscr{A}[\epsilon] \subset \mathscr{P}$ and for all $Q$ in $\mathscr{A}$ there exists $R$ in $\mathscr{A}[\epsilon]$ such that $h\,(Q,R) \leq \epsilon$. The case $\epsilon = 0$ corresponds to $\mathscr{A}[\epsilon]$ being dense in $\mathscr{A}$. The following result is proven in Section 3.C.3 and based on the work of Kim & Samworth [58] for $d \in [3]$ and Kur *et al.* [60] for $d \geq 4$.

**Lemma 3.6.** *For all positive $\epsilon$ there exists an $\epsilon$-net $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ such that*

$$|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]| \leq \begin{cases} \dfrac{9}{\eta_1}\dfrac{M(\lambda_+-\lambda_-)}{\lambda_-^{3/2}} e^{\overline{K_1}\epsilon^{-1/2}} & \text{for } d=1, \\ \dfrac{3^8\pi}{\eta_2^3}\dfrac{M^2(\lambda_+-\lambda_-)^2\lambda_+}{\lambda_-^4} e^{\overline{K_2}\epsilon^{-1}\log_{++}^{3/2}(1/\epsilon)} & \text{for } d=2, \\ \dfrac{2^7 3^{27/2}\pi^3}{\eta_3^6}\dfrac{M^3(\lambda_+-\lambda_-)^3\lambda_+^3}{\lambda_-^{15/2}} e^{\overline{K_3}\epsilon^{-2}} & \text{for } d=3, \end{cases}$$

*where $\eta_d$ and $\overline{K}_d$ are constants given in Theorem 4 [58] that only depend on $d$, and with $\log_{++}(x) = \max(1,\log x)$. For $d \geq 4$ and all positive $\epsilon$ there exists an $\epsilon$-net $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ such that*

$$|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]| \leq C_d \frac{\lambda_+^{d(d-1)/2} M^d (\lambda_+-\lambda_-)^d}{\lambda_-^{d(d+1)/2}} \exp\left(\overline{K}_d \epsilon^{-(d-1)} \log^{(d+1)(d+2)/2}(\epsilon^{-1})\right),$$

*where $\eta_d$ and $\overline{K}_d$ are constants that only depend on $d$.*

**The case $d \in \{1,2,3\}$**

Let $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ be a $\epsilon$-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ that satisfies the bound given in Lemma 3.6 for

$$\lambda_+ = \lambda_-^{-1} = M := \begin{cases} \exp\left(\overline{K}_1\,(n/\log n)^{1/5}\right) & \text{for } d=1, \\ \exp\left(\overline{K}_2 n^{1/3}\log^{2/3} n\right) & \text{for } d=2, \\ \exp\left(\overline{K}_3\,(n/\log n)^{1/2}\right) & \text{for } d=3, \end{cases} \tag{3.20}$$

and

$$\epsilon := \begin{cases} n^{-2/5}\log^{2/5} n & \text{for } d=1, \\ n^{-1/3}\log^{5/6} n & \text{for } d=2, \\ n^{-1/4}\log^{1/4} n & \text{for } d=3. \end{cases} \tag{3.21}$$

The following result holds and its proof can be found in Section 3.C.1.

**Theorem 3.2.** *Let $n \geq 3$ and $X_1,X_2,\ldots,X_n$ be arbitrary random variables with marginal distributions $P_1,\ldots,P_n$. The $\rho$-estimator $\hat{P}_s$ given by (3.12) with $\mathscr{M} = \mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ satisfies for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C_d\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq\; & h^2\left(\overline{P},\mathscr{F}_{\lambda_-,\lambda_+,M}\right) + n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right) \\ & + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ & + \begin{cases} n^{-4/5}\left(\log^{4/5} n + s\log^{-1/5} n\right) & \text{for } d=1, \\ n^{-2/3}\left(\log^{5/3} n + s\log^{2/3} n\right) & \text{for } d=2, \\ n^{-1/2}\left(\log^{1/2} n + s\log^{-1/2} n\right) & \text{for } d=3, \end{cases} \end{aligned} \tag{3.22}$$

*for positive constants $C_1$,$C_2$,$C_3$. In particular if the model described in Section 3.3.2 is exact and $s \geq (2\underline{\lambda}(U))^{-1} \log n$, there exists a positive constant $C(U,d,\Delta_t)$ such that for $n$ large enough*

$$C(U,d,\Delta_t)\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq \begin{cases} n^{-4/5}\left(\log^{4/5} n + s\log^{-1/5} n\right) \text{ for } d=1, \\ n^{-2/3}\left(\log^{5/3} n + s\log^{2/3} n\right) \text{ for } d=2, \\ n^{-1/2}\left(\log^{1/2} n + s\log^{-1/2} n\right) \text{ for } d=3, \end{cases}$$

*where $\overline{P}$ is the invariant distribution given by (3.19).*

Inequality (3.22) is a consequence of Theorem 3.1 and does not require any assumption on the data. The last term comes from the control of the dimension of the net $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ and the choice of $\epsilon$ given by (3.21). Ideally, most of the distributions $P_i$ lie in a small neighborhood of a distribution $\overline{P}$ in $\mathscr{F}_{\lambda_-,\lambda_+,M}$ so that the first two terms in the bound remain small compared to the last term. Those two terms vanish when the model is exact and a good choice of $s$ guarantees the term $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ is negligible with respect to the last one.

We can derive convergence rates for the optimal choice of $s$ given $\underline{\lambda}(U)$. One can check that up to a logarithmic factor, we obtain the same rates as Theorem 5 [58] in the i.i.d. case. Our power of $\log n$ is even better for $d=3$. As mentioned in Section 3.2.5, the knowledge of $\underline{\lambda}(U)$ is not necessary to obtain convergence rates. We obtain slightly worse powers of $\log n$ in the convergence rates for $s$ of order $\log^2 n$. We can also derive results for i.i.d. observations from (3.22) by taking the term $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ down to 0 which provides a result for the robust estimation of a log-concave density from i.i.d. observations.

In order to illustrate the robustness of our estimators we consider the situation of Section 3.2.4. Let $Z_1,\dots,Z_n$ be random variables with any distributions and $E_1,\dots,E_n$ be Bernoulli random variables such that for all $i \in [n]$,

$$X_i = E_i Y_{t_1+(i-1)\Delta_t} + (1-E_i)Z_i,$$

where $(Y_t)_t$ is a stationary solution of the Langevin equation (3.17) for some unknown convex function $U$ that satisfies Assumption 3.1.

**Corollary 3.2.** *Let $\hat{P}_s$ be the estimator given by (3.12) with $\mathscr{M} = \mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$. If $E_1,\dots,E_n,Z_1,\dots,Z_n$ and $\mathbf{X}$ are mutually independent, there exists a positive constant $C(U,d,\Delta_t)$ such that for $s \geq (2\underline{\lambda}(U))^{-1}\log n$ we have*

$$C(U,d,\Delta_t)\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-1}\sum_{i=1}^{n}(1-p_i) + \begin{cases} n^{-4/5}\left(\log^{4/5} n + s\log^{-1/5} n\right) \text{ for } d=1, \\ n^{-2/3}\left(\log^{5/3} n + s\log^{2/3} n\right) \text{ for } d=2, \\ n^{-1/2}\left(\log^{1/2} n + s\log^{-1/2} n\right) \text{ for } d=3, \end{cases} \quad (3.23)$$

*where $p_i = \mathbb{P}(E_i = 1)$ for all $i \in [n]$.*

One can see that our deviation bound is not significantly worse as long as the average proportion of contamination $n^{-1}\sum_{i=1}^{n}(1-p_i)$ remains small compared to the last term on the right hand side of (3.23).

**The case $d \geq 4$**

Let $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ be an $\epsilon$-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ that satisfies the bound given in Lemma 3.5 with

$$\lambda_+ = \lambda_-^{-1} = \exp\left(\frac{\epsilon^{-(d-1)}\log^{(d+1)(d+2)/2}(\epsilon^{-1})}{d^2}\right) \tag{3.24}$$

$$M = \exp\left(\frac{\epsilon^{-(d-1)}\log^{(d+1)(d+2)/2}(\epsilon^{-1})}{d}\right), \tag{3.25}$$

with

$$\epsilon = n^{-\frac{1}{d+1}}\log^{\frac{1}{d+1}+\frac{d+2}{2}} n. \tag{3.26}$$

The following result holds and its proof can be found in Section 3.C.1.

**Theorem 3.3.** *Let $n \geq 3$ and $X_1,X_2,\ldots,X_n$ be arbitrary random variables with marginal distributions $P_1,\ldots,P_n$. The $\rho$-estimator $\hat{P}_s$ given by (3.12) with $\mathscr{M} = \mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ satisfies for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C_d\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq\ & h^2\left(\overline{P},\mathscr{F}_{\lambda_-,\lambda_+,M}\right) + n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right) \\ & + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ & + n^{-\frac{2}{d+1}}\left(\log^{d+2+\frac{2}{d+1}} n + s\log^{d+1+\frac{2}{d+1}} n\right). \end{aligned}$$

*In particular if the model described in Section 3.3.2 is exact and $s \geq (2\underline{\lambda}(U))^{-1}\log n$, there exists a positive constant $C(U,d,\Delta_t)$ such that for $n$ large enough*

$$C(U,d,\Delta_t)\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-\frac{2}{d+1}}\left(\log^{d+2+\frac{1}{d+1}} n + s\log^{d+1+\frac{2}{d+1}} n\right),$$

*where $\overline{P}$ is the invariant distribution given by (3.19).*

This result is equivalent to Theorem 3.2 and the comments that applied to it also apply now. Our estimator is also robust and tolerates a higher contamination rate as the convergence rate is slower. One can check that up to a logarithmic factor, we have the same rate that Kur *et al.* [60] obtain for the estimation of log-concave estimation from i.i.d. observations. We can derive a result equivalent to Corollary 3.2 for $d \geq 4$. Our estimator can tolerate an average proportion of contamination of order not larger than $n^{-\frac{2}{d+1}}\log^{d+2+\frac{2}{d+1}} n$ without its performance being significantly deteriorated.

## 3.4 Hidden Markov models

### 3.4.1 Stationary hidden Markov models

Let $(Y_i,H_i)_i$ be a finite state space HMM with parameters $(K^*,w^*,Q^*,F^*)$. If $w^*$ is invariant with respect to $Q^*$, then the process $(Y_i,H_i)_i$ is stationary. As explained in the introduction, we aim at estimating the different parameters through the distribution of consecutive observations. For $L \geq 2$ we define $P_L = P_{w^*,Q^*,F^*}$ with $P_{W^*,Q^*,F^*}$ defined by (3.1), and we have $\mathcal{L}(Y_i,Y_{i+1},\ldots,Y_{i+L-1}) = P_L$ for all $i$. We have identically distributed but dependent random variables from which we can estimate $P_L$. It is possible to relax the stationary assumption.

**Assumption 3.2.** *Let $(Y_i,H_i)_i$ be a finite state space HMM with parameters $(K^*,w^*,Q^*,F^*)$ such that $Q^*$ is irreducible and aperiodic.*

In this case we do not have identically distributed observations anymore. However the distribution $\mathcal{L}(Y_i,\dots,Y_{i+L-1})$ converges exponentially fast to the distribution

$$P^* = P_{\pi^*,Q^*,F^*}, \tag{3.27}$$

where $\pi^*$ is the only invariant distribution with respect to $Q^*$.

### 3.4.2 The framework

Let $Y_1,Y_2,\dots,Y_N$ be random variables taking values in a measurable space $(\mathscr{Y},\mathcal{Y})$. Let $L$ be in $\{2,3,\dots,\lfloor N/2\rfloor\}$ and $n$ be the integer given by $n = N+1-L$. We define the new random variables

$$X_i = (Y_i,Y_{i+1},\dots,Y_{i+L-1}),i = 1,\dots,n, \tag{3.28}$$

taking values in the measurable space $(\mathscr{X},\mathcal{X}) = \left(\mathscr{Y}^L,\mathcal{Y}^{\otimes L}\right)$. We follow the notation established in Section 3.2.

We denote $\mathscr{P}_Y$ the class of all probability distributions on $(\mathscr{Y},\mathcal{Y})$. For $K \geq 2$ and subsets $\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K$ of $\mathscr{P}_Y$, we denote by $\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$ the set of distributions defined by

$$\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right) := \left\{P_{w,Q,F};\ \begin{array}{c}\forall k \in [K], w \in \mathcal{W}_K,\\ Q \in \mathcal{T}_K, F_k \in \overline{\mathscr{F}}_k\end{array}\right\} \subset \mathscr{P}_X, \tag{3.29}$$

where $P_{w,Q,F}$ is given by (3.1),

$$\mathcal{T}_K = \left\{Q \in [0,1]^{K\times K}; \sum_{j=1}^{K} Q_{ij} = 1, \forall i \in \{1,\dots,K\}\right\}, \tag{3.30}$$

$$\text{and } \mathcal{W}_K = \left\{w \in [0,1]^K; w_1+\dots+w_K = 1\right\}. \tag{3.31}$$

We call *emission models* the sets $\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K$. Let $\overline{\mathscr{M}}$ be a non-empty subset of $\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$.

### 3.4.3 Estimation

Let $\nu$ be a $\sigma$-finite measure on $(\mathscr{Y},\mathcal{Y})$ and we denote by $\mu$ the associated $\sigma$-finite measure on $(\mathscr{X},\mathcal{X})$ given by $\mu := \nu^{\otimes L}$. We consider emission models that satisfy the following.

**Assumption 3.3.** *We dispose of countable sets $\mathcal{F}_i, i = 1,\dots,K$ of probability density functions (with respect to $\nu$) such that*

1. *for all $k$ in $[K]$, the set of distributions $\mathscr{F}_i := \{f\cdot\nu; f \in \mathcal{F}_i\}$ is an $\epsilon$-net of $\overline{\mathscr{F}}_i$ with respect to the Hellinger distance;*

2. *for any $k_1,\dots,k_L \in [K]$, the class of functions*

$$\mathcal{F}_{k_1,\dots,k_L} = \left\{\mathbf{x} \in \mathscr{Y}^L \mapsto f_1(x_1)\dots f_L(x_L); f_l \in \mathcal{F}_{k_l}, \forall l \in [L]\right\}$$

*is VC-subgraph with VC-index not larger than $V_{k_1,\dots,k_L}$. Then we write*

$$\overline{V} := \sum_{1\leq k_1,\dots,k_L\leq K} V_{k_1,\dots,k_L}. \tag{3.32}$$

We refer to van der Vaart & Wellner [84] (Section 2.6.5) and Baraud *et al.* [9] (Section 8) as an introduction to VC-subgraph classes of functions. We just mention the following example. Any finite set $\mathcal{F}$ of real-valued functions is VC-subgraph with VC-index $V(\mathcal{F})$ that satisfies

$$V(\mathcal{F}) \leq 1 + \log_2(|\mathcal{F}|). \tag{3.33}$$

Therefore we can consider finite $\epsilon$-nets as we did in Section 3.3. We also show in Section 3.4.3 that exponential families satisfy our assumption.

We consider countable approximations of $\mathcal{W}_K$ and $\mathcal{T}_K$ given by

$$\mathcal{W}_{\delta,K} := \mathcal{W}_K \cap ([\delta,1] \cap \mathbb{Q})^K \text{ and } \mathcal{T}_{\delta,K} := \mathcal{T}_K \cap ([\delta,1] \cap \mathbb{Q})^{K\times K}, \tag{3.34}$$

for $0 < \delta \leq 1/K$. We define $\mathscr{H}_\delta$ by

$$\mathscr{H}_\delta := \{P_{w,Q,f}; w \in \mathcal{W}_{\delta,K}, Q \in \mathcal{T}_{\delta,K}, f_k \in \mathscr{F}_k, \forall i \in [K]\}, \tag{3.35}$$

where the sets $(\mathscr{F}_k)_{1\leq k\leq K}$ are given in Assumption 3.3. This lower bound $\delta$ is a technicality for bounding the dimension of our model. We define the countable set of distributions

$$\mathscr{M}_\delta := \left\{ P_{w,Q,F} \in \mathscr{H}_\delta; \exists P_{w',Q',F'} \in \overline{\mathscr{M}}, \begin{array}{c} h^2(Q_{k\cdot},Q'_{k\cdot}) \leq (K-1)\delta \\ h(F_k,F'_k) \leq \epsilon, \forall k \in [K], \\ h^2(w,w') \leq (K-1)\delta, \end{array} \right\}, \tag{3.36}$$

which is a good approximation of $\overline{\mathscr{M}}$ for small values of $\delta$ and $\epsilon$. We denote by $\hat{P}_{s,\delta}$ the estimator

$$\hat{P}_{s,\delta} := \hat{P}_s(\mathscr{M}_\delta, \mathbf{X}), \tag{3.37}$$

as defined by (3.12). The following theorem is proven in Section 3.D.1.

**Theorem 3.4.** *Let $N \geq K+L$ and $Y_1,\dots,Y_N$ be arbitrary random variables. Under Assumption 3.3, let $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with*

$$\delta = \frac{\overline{V}}{n(s,1)(K-1)} \wedge \frac{1}{K}. \tag{3.38}$$

*There exists a positive constant $C$ such that for all $\overline{P} \in \mathscr{P}_X$,*

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^n h^2\left(\overline{P},P_i\right) + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) + L\epsilon^2 + (s+1)L\overline{V}\frac{\log n}{n}. \tag{3.39}$$

*In particular under Assumption 3.2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n \vee (L-1)$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + L\epsilon^2 + L\overline{V}\frac{s\log n}{n}, \tag{3.40}$$

*where $P^*$ is given by (3.27).*

Inequality (3.39) is a consequence of Theorem 3.1 and does not require any assumption on the data. The last two terms come respectively from the approximation of $\overline{\mathscr{M}}$ by $\mathscr{M}$ and the control of the dimension of $\mathscr{M}$. Ideally, we can take $\overline{P}$ in $\overline{\mathscr{M}}$ such that most of the distributions $P_i$ lie in a small neighborhood of $\overline{P}$ so that the first two terms in the bound remain small compared to the last term. Under Assumption 3.2 the quantity $\sum_{i=1}^n h^2(P^*,P_i)$

is bounded and a good choice of $s$ guarantees the term $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b})$ to be negligible with respect to the last one. The optimal choice of $s$ depends on a constant $c(Q^*)$ which relates to the spectral gap of $Q^*$. We distinguish two cases in order to obtain convergence rates over the class

$$\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right) := \left\{P_{w,Q,F}\in\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right);\ \begin{array}{c}Q \text{ irreducible ,}\\ Q \text{ aperiodic,}\\ \text{and } w=Qw\end{array}\right\}. \tag{3.41}$$

The first case is when we satisfy Assumption 3.3 with $\epsilon=0$. In that situation and for $P^*$ in $\overline{\mathscr{M}}=\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$ the first two terms in (3.40) vanish. For the optimal choice of $s$ our estimator achieves the convergence rate $n^{-1}\log^2 n$ with respect to the squared Hellinger distance over $\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$. This means that up to a logarithmic term we achieve the optimal rate $1/n$ in the independent context (see Birgé [15]). As mentioned in Section 3.2.5, the knowledge of $c(Q^*)$ is not necessary to obtain convergence rates. We only obtain slightly worse powers of $\log n$ in the convergence rates for $s=\log^2 n$.
The second case is when we cannot take $\epsilon=0$. In that situation the term $\overline{V}$ depends on $\epsilon$ and we proceed as in Section 3.3. We obtain a convergence rate taking $\epsilon$ that goes to 0 with $n$ at a rate that balances the last two terms in (3.40). This happens when $\epsilon^2/\overline{V}$ is of order $n^{-1}$ up to a logarithmic term. We put it in application in Section 3.4.3.

In order to illustrate the robustness of our estimators we consider the situation of Section 3.2.4. Let $Z_1,\dots,Z_N$ be random variables with any distributions and $E_1,\dots,E_N$ be Bernoulli random variables such that for all $i\in[N]$,

$$Y_i = E_iY_i' + (1-E_i)Z_i,$$

where $\mathbf{Y}'$ satisfies Assumption 3.2. The following result is proven in Section 3.D.2.

**Corollary 3.3.** *Let $N\geq K+L$ and $\hat{P}_s=\hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\delta$ given by (3.38). If $E_1,\dots,E_N,Z_1,\dots,Z_N$ and $\mathbf{Y}'$ are mutually independent, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s\geq c(Q^*)\log n$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + \frac{L}{N}\sum_{i=1}^N(1-p_i) + L\epsilon^2 + L\overline{V}\frac{s\log n}{n}, \tag{3.42}$$

*where $p_i=\mathbb{P}(E_i=1)$ for all $i\in[N]$ and $\delta$ is given by (3.38).*

One can see that our deviation bound is not significantly worse as long as the average proportion of contamination $\frac{L}{N}\sum_{i=1}^N(1-p_i)$ remains small compared to the last two terms. One would typically look at the following situation. We assume that the model is well specified, i.e. $P^*\in\overline{\mathscr{M}}$. For Hüber's contamination model, i.e. $p_i=1-\alpha_{cont}$ for all $i\in[N]$, we get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\left[\alpha_{cont}+\epsilon^2+\overline{V}\frac{s\log n}{n}\right], \tag{3.43}$$

for $s\geq c(Q^*)\log n$. The bound on the convergence rate is not deteriorated as long as the contamination rate $\alpha_{cont}$ is small compared to $\epsilon^2+\overline{V}\frac{s\log n}{n}$. We can also consider the situation where $\mathbb{P}(E_i=0)=\mathbb{1}_{i\in I}$ for some subset $I\subset[N]$. We get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\left[\frac{|I|}{N}+\epsilon^2+\overline{V}\frac{s\log n}{n}\right], \tag{3.44}$$

for $s\geq c(Q^*)\log n$. As before, our bound on the convergence rate is not deteriorated as long as the proportion of outliers $|I|/N$ is small compared to $\epsilon^2+\overline{V}\frac{s\log n}{n}$.

### log-concave emission densities

We use results and notation given in Section 3.3. Let $d$ be a positive integer and $\epsilon \in (0,1)$. Let $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ be an $\epsilon$-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ that satisfies the bound given in Lemma 3.5. We take $\overline{\mathscr{F}}_k = \mathscr{F}_{\lambda_-,\lambda_+,M}$ for all $k \in [K]$ and satisfy Assumption 3.3 with

$$\overline{V} = K^L \left(1 + L \log_2 \left(\left|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]\right|\right)\right). \tag{3.45}$$

We take $\overline{\mathscr{M}} = \mathscr{H}\left(K, \mathscr{F}_{\lambda_-,\lambda_+,M}, \ldots, \mathscr{F}_{\lambda_-,\lambda_+,M}\right)$. We distinguish the two cases $d \in \{1,2,3\}$ and $d \geq 4$.

For $d \in \{1,2,3\}$ we take $\lambda_+, \lambda_-, M$ as in (3.20) and $\epsilon$ as in (3.21). The following result holds and its proof can be found in Section 3.D.3.

**Theorem 3.5.** *Let $N \geq K + L$ and $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\delta$ given by (3.38). There exist positive constants $C_1, C_2, C_3$ such that for all $\overline{P} \in \mathscr{P}_X$,*

$$\begin{aligned} C_d \mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \overline{\mathscr{M}}\right) + n^{-1} \sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) \\ + n^{-1} \sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) \\ + (s+1) L^2 K^L \times \begin{cases} n^{-4/5} \log^{4/5} n \text{ for } d = 1, \\ n^{-2/3} \log^{5/3} n \text{ for } d = 2, \\ n^{-1/2} \log^{1/2} n \text{ for } d = 3. \end{cases} \end{aligned} \tag{3.46}$$

*In particular under Assumption 3.2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*) \log n$ we have*

$$C(Q^*) \mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq h^2\left(P^*, \overline{\mathscr{M}}\right) + s L^2 K^L \times \begin{cases} n^{-4/5} \log^{4/5} n \text{ for } d = 1, \\ n^{-2/3} \log^{5/3} n \text{ for } d = 2, \\ n^{-1/2} \log^{1/2} n \text{ for } d = 3, \end{cases}$$

*where $P^*$ is given by (3.27).*

Inequality (3.46) is a consequence of Theorem 3.4 and does not require any assumption on the data. We can deduce convergence rates over the class $\mathscr{H}^*\left(K, \mathscr{F}_d, \ldots, \mathscr{F}_d\right)$, where $\mathscr{F}_d$ is the set of distributions with log-concave densities defined in Section 3.3. For the optimal choice of $s$, we have

$$C(Q^*) \mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq L^2 K^L \times \begin{cases} n^{-4/5} \log^{9/5} n \text{ for } d = 1, \\ n^{-2/3} \log^{8/3} n \text{ for } d = 2, \\ n^{-1/2} \log^{3/2} n \text{ for } d = 3, \end{cases} \tag{3.47}$$

for all $P^*$ in $\mathscr{H}^*\left(K, \mathscr{F}_d, \ldots, \mathscr{F}_d\right)$. We see that we have a worse power of $\log n$ compared to Theorem 3.2. It comes from an additional logarithmic factor in the dimension term for HMMs. Corollary 3.3 tells us our estimator is also robust to contamination and outliers. Let us illustrate it for $d = 1$. We can see from (3.43) that our bound is not significantly worse as long as the contamination rate $\alpha_{cont}$ is of order not larger than $n^{-4/5} \log^{9/5} n$. Similarly (3.44) tells us that a number $|I|$ of outliers of order not larger than $n^{1/5} \log^{9/5} n$ does not significantly deteriorate our bound on the convergence rate of our estimator. We can follow the same train of thought for $d = 2$ and $d = 3$ and deduce the level of contamination or outliers our estimator can tolerate before its performance significantly worsens.

For $d \geq 4$ we take $\lambda_+, \lambda_-^{-1}$ as in (3.24), $M$ as in (3.25) and $\epsilon$ as in (3.26). The following result holds and its proof can be found in Section 3.D.3.

**Theorem 3.6.** *Let $N \geq K + L$ and $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\delta$ given by (3.38). There exist a positive constant $C_d$ such that for all $\overline{P} \in \mathscr{P}_X$,*

$$\begin{aligned} C_d \mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) \\ + n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) \\ + (s+1)L^2 K^L n^{-\frac{2}{d+1}} \log^{d+2+\frac{2}{d+1}} n. \end{aligned}$$

*In particular under Assumption 3.2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*) \log n$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq h^2\left(P^*, \overline{\mathscr{M}}\right) + sL^2K^L n^{-\frac{2}{d+1}} \log^{d+2+\frac{2}{d+1}} n, \tag{3.48}$$

*where $P^*$ is given by (3.27).*

Inequality (3.46) does not require any assumption on the data. We can deduce convergence rates over the class $\mathscr{H}^*\left(K, \mathscr{F}_d, \ldots, \mathscr{F}_d\right)$. For the optimal choice of $s$, we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq L^2K^L n^{-\frac{2}{d+1}} \log^{d+3+\frac{2}{d+1}} n$$

for all $P^* \in \mathscr{H}^*\left(K, \mathscr{F}_d, \ldots, \mathscr{F}_d\right)$. As for $d \leq 3$, we have the same rate as in Section 3.3 with a worse power of $\log n$ due to the higher complexity of HMMs. Our estimator is also robust to contamination and outliers. We can see from (3.43) that our bound is not significantly worse as long as the contamination rate $\alpha_{cont}$ is of order not larger than $n^{-\frac{2}{d+1}} \log^{d+3+\frac{2}{d+1}} n$. Similarly (3.44) tells us that a number of outliers of order not larger than $n^{\frac{d-1}{d+1}} \log^{d+3+\frac{2}{d+1}} n$ does not significantly deteriorate our bound on the convergence rate of our estimator.

### Exponential families as emission models

We introduce exponential families as follow. Let $d$ be a positive integer and $\eta : \overline{\Theta} \to \mathbb{R}^d$ be a function over a non-empty set $\overline{\Theta}$. Let $T : \mathscr{Y} \to \mathbb{R}^d$ and $B : \mathscr{Y} \to \mathbb{R}$ be measurable functions such that

$$\int_{\mathscr{Y}} e^{\langle \eta(\theta), T(x)\rangle + B(x)} \nu(dx) < \infty, \forall \theta \in \overline{\Theta},$$

we denote by $\mathcal{E}\left(\overline{\Theta}, \eta, T, d, B\right)$ the exponential family defined by

$$\mathcal{E}\left(\overline{\Theta}, \eta, T, d, B\right) := \left\{f_\theta : x \mapsto e^{\langle \eta(\theta), T(x)\rangle + A(\theta) + B(x)}; \theta \in \overline{\Theta}\right\}, \tag{3.49}$$

where

$$A(\theta) := -\log\left(\int_{\mathscr{Y}} e^{\langle \eta(\theta), T(x)\rangle + B(x)} \nu(dx)\right).$$

It is a set of probability density functions with respect to $\nu$.

**Assumption 3.4.** *For all $k \in \{1, \ldots, K\}$,*

1. *$\overline{\mathscr{F}}_k$ is of the form*

$$\overline{\mathscr{F}}_k = \left\{q \cdot \nu; q \in \mathcal{E}\left(\overline{\Theta}_k, \eta_k, T_k, d_k, B_k\right)\right\}, \tag{3.50}$$

2. *$\Theta_k$ is a countable subset of $\overline{\Theta}_k$ such that*

$$\mathscr{F}_k = \left\{q \cdot \nu; q \in \mathcal{E}\left(\Theta_k, \eta_{k|\Theta_k}, T_k, d_k, B_k\right)\right\}$$

*is a dense subset of $\overline{\mathscr{F}}_k$.*

The next result is proven in Section 3.D.4 and shows that the last assumption is sufficient to satisfy our main assumption.

**Proposition 3.1.** *Under Assumption 3.4, we satisfy Assumption 3.3 with* $\epsilon = 0$ *and* $V_{k_1,\dots,k_L} = 3 + \sum_{k_l=1}^{L} d_{k_l}$. *Therefore we have*

$$\overline{V} = 3K^L + LK^{L-1}\left(d_1 + \cdots + d_K\right). \tag{3.51}$$

We can see that the constant $\overline{V}$ does not depend on $\mathscr{X}$ but on the dimensions $d_1,\dots,d_K$ which is the actual measure of the complexity of the exponential families. To our knowledge, the existence of a countable dense subset is satisfied for all the common exponential families. We obtain the following result for $\overline{\mathscr{M}} \subset \mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$.

**Corollary 3.4.** *Let* $N \geq K + L$ *and* $\hat{P}_s = \hat{P}_{s,\delta}$ *be the estimator given by (3.37) with* $\delta$ *given by (3.38). There exists a positive constant* $C$ *such that for all* $\overline{P} \in \mathscr{P}_X$, *we have*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{s,\delta}\right)\right] \leq\ & h^2\left(\overline{P},\overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(\overline{P},P_i\right) \\ & + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ & + (s+1)LK^{L-1}\left(K + L(d_1+\cdots+d_K)\right)\log n. \end{aligned}$$

*In particular under Assumption 3.2, there exist positive constants* $C(Q^*)$ *and* $c(Q^*)$ *such that for* $s \geq c(Q^*)\log n$ *we have*

$$\begin{aligned} C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq\ & h^2\left(P^*,\overline{\mathscr{M}}\right) \\ & + LK^{L-1}\left(K + L(d_1+\cdots+d_K)\right)\frac{s\log n}{n}, \end{aligned} \tag{3.52}$$

*where* $P^*$ *is given by (3.27).*

This result is a direct consequence of Theorem 3.4 and Proposition 3.1. We can deduce a bound on the convergence rate over $\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$. For the optimal choice of $s$, we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq LK^{L-1}\left(K + L(d_1+\cdots+d_K)\right)\frac{\log^2 n}{n},$$

for all $P^*$ in $\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$. We obtain the optimal $1/n$ rate with respect to the squared Hellinger distance, up to a logarithmic factor. Corollary 3.3 shows that our estimator is also robust to contamination and outliers. From (3.43) we see that our bound is not significantly worse as long as the contamination rate $\alpha_{cont}$ is of order not larger than $n^{-1}\log^2 n$. Similarly, we get from (3.44) that the performance of our estimator is not altered as long as the number of outliers $|I|$ is of order not larger than $\log^2 n$.

Let us illustrate how Corollary 3.4 applies with the following example. Let $d$ be a positive integer and $\mathrm{Cov}_{+*}(d)$ be the set of $d \times d$ symmetric and positive-definite matrices. For $z \in \mathbb{R}^d$ and $\Sigma \in \mathrm{Cov}_{+*}(d)$, we denote by $g_{z,\Sigma}$ the density function of the normal distribution $\mathcal{N}(z,\Sigma)$ with respect to the Lebesgue measure given by

$$g_{z,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{(z-m)^T\Sigma^{-1}(z-m)}{2}\right), \tag{3.53}$$

where $|\Sigma|$ denotes the determinant of $\Sigma$. Let $\mathcal{G}_d$ be the location-scale family of densities given by $\mathcal{G}_d := \{g_{z,\Sigma}; z \in \mathbb{R}^d, \Sigma \in \mathrm{Cov}_{+*}(d)\}$. One can check it is an exponential family with $\mathcal{G}_d = \mathcal{E}\left(\mathbb{R}^d \times \mathrm{Cov}_{+*}(d), \eta, T, \frac{d(d+3)}{2}, 0\right)$ where

$$T(x) = \left(x, \left(x_i^2\right)_{1\le i\le d}, (x_i x_j)_{1\le i<j\le d}\right) \text{ and}$$
$$\eta(z,\Sigma) = \left(\Sigma^{-1}z, -\frac{1}{2}\left(\Sigma^{-1}_{ii}\right)_{1\le i\le d}, -\left(\Sigma^{-1}_{ij}\right)_{1\le i<j\le d}\right).$$

For a fixed $\Sigma$ we denote by $\mathcal{G}_{loc}(\Sigma)$ the associated location family given by $\mathcal{G}_{loc}(\Sigma) := \{g_{z,\Sigma}; z \in \mathbb{R}^d\}$. It is also an exponential family with $\mathcal{G}_{loc}(\Sigma) = \mathcal{E}\left(\mathbb{R}^d \times \mathrm{Cov}_{+*}(d), \eta, T, d, B\right)$, where

$$\eta(z) = \Sigma^{-1}z, T(x) = x \text{ and } B(x) = -\frac{x^T\Sigma^{-1}x}{2}.$$

We denote by $\mathscr{G}_d$ and $\mathscr{G}_{loc}(\Sigma)$ respectively, the sets of probability distributions associated to $\mathcal{G}_d$ and $\mathcal{G}_{loc}(\Sigma)$. The next result is a consequence of Corollary 3.4.

**Theorem 3.7.** *Let $N \ge K + L$ and $Y_1, \ldots, Y_N$ be arbitrary random variables.*

- *Let $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\overline{\mathscr{M}} = \mathscr{H}(K, \mathscr{G}_d, \ldots, \mathscr{G}_d)$ and $\delta$ given by (3.38). There exists a positive constant $C$ such that for all $\overline{P} \in \mathscr{P}_X$*

$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \le h^2\left(\overline{P}, \overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(\overline{P}, P_i\right)$$
$$n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right)$$
$$+ (s+1)L^2K^L d(d+3)\frac{\log n}{n}. \tag{3.54}$$

  *In particular under Assumption 3.2 there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \ge c(Q^*)\log n$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \le h^2\left(P^*, \overline{\mathscr{M}}\right) + (s+1)L^2K^L d(d+3)\frac{\log n}{n},$$

  *where $P^*$ is given by (3.27).*

- *Let $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\overline{\mathscr{M}} = \mathscr{H}(K, \mathscr{G}_{loc}(\Sigma), \ldots, \mathscr{G}_{loc}(\Sigma))$ and $\delta$ given by (3.38). There exists a positive constant $C$ such that for all $\overline{P} \in \mathscr{P}_X$*

$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \le h^2\left(\overline{P}, \overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(\overline{P}, P_i\right)$$
$$+ n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right)$$
$$+ (s+1)L^2K^L d\frac{\log n}{n}, \tag{3.55}$$

  *for any $\Sigma$ in $Cov_{+*}(d)$. In particular under Assumption 3.2 there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \ge c(Q^*)\log n$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \le h^2\left(P^*, \overline{\mathscr{M}}\right) + (s+1)L^2K^L d\frac{\log n}{n},$$

  *where $P^*$ is given by (3.27).*

Inequalities (3.54) and (3.55) are consequences of Corollary 3.4 and do not require any assumption on the data. We deduce bounds on the convergence rate of our estimator over $\mathscr{H}^*(K,\mathscr{G}_d,\ldots,\mathscr{G}_d)$ and $\mathscr{H}^*(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma))$. For the optimal choice of $s$ we obtain the rate $n^{-1}\log^2 n$ with respect to the squared Hellinger distance both for $P^* \in \mathscr{H}^*(K,\mathscr{G}_d,\ldots,\mathscr{G}_d)$ and $P^* \in \mathscr{H}^*(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma))$. This rate is optimal up to a logarithmic factor. We can see that the dependence on the dimension $d$ is linear for the model $\mathscr{H}(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma))$ while its quadratic for $\mathscr{H}^*(K,\mathscr{G}_d,\ldots,\mathscr{G}_d)$.

We can obtain similar results for any exponential family. It is also possible to consider hidden Markov models with different exponential families as emission models. The next section investigates the estimation of the parameters.

### Estimation of the parameters with emission exponential families

We say that $\hat{\pi}$, $\hat{Q}$ and $\hat{F}$ are $\rho$-estimators of $\pi^*$, $Q^*$ and $F^*$ if $P_{\hat{w},\hat{Q},\hat{F}} = \hat{P}_{s,\delta}$ is an estimator of $P^*$ given by (3.37). If we consider models of densities that are uniformly bounded, we can use (3.3) and Theorem 9 of Lehéricy [63] to deduce risk bounds for the parameter estimators. It is also possible to use the results of Ibragimov and Has'minskiĭ [54] for regular parametric models.

We consider that Assumption 3.4 is satisfied with $\overline{\Theta}_k \subset \mathbb{R}^{e_k}$ for all $k \in [K]$. For $k \in [K]$ we denote by $F_{\theta_k}$ the probability distribution given by the parameter $\theta_k \in \overline{\Theta}_k$, i.e. $F_{\theta_k} = f_{\theta_k} \cdot \nu$ with $f_\theta$ given by (3.49). Let $\overline{\Phi}$ be an open convex subset of $O_K^{K+1} \times \overline{\Theta}_1 \times \cdots \times \overline{\Theta}_K$, where

$$O_K = \left\{ \mathbf{a} \in (0,1)^{K-1}, a_1 + \cdots + a_{K-1} < 1 \right\}.$$

For $\phi$ in $\overline{\Phi}$, we can define $w \in \mathcal{W}_K$, $Q \in \mathcal{T}_K$ and $\theta \in \overline{\Theta}_1 \times \cdots \times \overline{\Theta}_K$ by $\phi = (\phi_w, \phi_{Q,1}, \ldots, \phi_{Q,K}, \phi_\theta)$ with

$$\begin{aligned}(w_1,\ldots,w_{K-1}) &= \phi_w \in O_K,\\ (Q_{k,1},\ldots,Q_{K-1,1}) &= \phi_{Q,k} \in O_K,\\ (\theta_1,\ldots,\theta_K) &= \phi_\theta \in \overline{\Theta}_1 \times \cdots \times \overline{\Theta}_K.\end{aligned}$$

We denote by $\overline{\mathscr{M}}$ the model given by

$$\overline{\mathscr{M}} := \left\{ P_\phi = p(\cdot;\phi)\cdot\mu; \phi \in \overline{\Phi} \right\} \tag{3.56}$$

and

$$p(\mathbf{x};\phi) = \sum_{1 \le k_1,\ldots,k_L \le K} w_{k_1} Q(k_2|k_1)\ldots Q(k_L|k_{L-1}) \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l).$$

We need the following assumption to make sure we can deduce $\phi$ from $P_\phi$.

**Assumption 3.5.** *For all $k$ in $[K]$,*

- *the map $\theta_k \mapsto F_{\theta_k}$ is continuous on $\overline{\Theta}_k$ with respect to the Hellinger distance;*
- *the functions $\eta_k$ and $A_k$ are of class $\mathcal{C}^1$ on $\overline{\Theta}_k$;*
- *for all $\theta_k$ in $\overline{\Theta}_k$, we have $\int ||T_k(x)||^2 f_{\theta_k}(x)\nu(dx) < \infty$ and*

$$\int ||T_k(x)||^2 \left| f_{\theta_k}(x) - f_{\theta'_k}(x) \right| \nu(dx) \xrightarrow[||\theta_k - \theta'_k|| \to 0]{} 0.$$

The next result is proven in Section 3.D.5 and shows that under some conditions we can deduce the parameters from the distribution $P_\phi$.

**Proposition 3.2.** *Under Assumption 3.5 the information matrix $I$ function given by*

$$I_{ij} : \phi \mapsto I(\phi)_{ij} = \int_{\mathscr{X}^L} \partial_{\phi_i} p(\mathbf{x};\phi) \partial_{\phi_j} p(\mathbf{x};\phi) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}$$

*is well-defined and continuous on $\overline{\Phi}$. We define the subset $\Phi^* \subset \overline{\Phi}$ by*

$$\Phi^* := \left\{ \overline{\phi} \in \overline{\Phi}; \begin{array}{c} I\left(\overline{\phi}\right) \text{ is definite positive and} \\ \inf\limits_{\substack{||\overline{\phi}-\phi||\geq a \\ \phi\in\overline{\Phi}}} h^2\left(P_{\overline{\phi}}, P_\phi\right) > 0, \forall a > 0, \end{array} \right\} \tag{3.57}$$

*For all $\phi^* \in \Phi^*$, there exists a positive constant $C(\phi^*)$ such that*

$$C(\phi^*)\left[||w^* - w||_2^2 + ||Q^* - Q||_2^2 + \sum_{k=1}^K ||\theta_k^* - \theta_k||_2^2 \wedge 1\right] \leq h^2\left(P_{\phi^*}, P_\phi\right), \tag{3.58}$$

*for all $\phi$ in $\overline{\Phi}$.*

The constant $C(\phi^*)$ depends on the inverse of the smallest eigenvalue of $I(\phi^*)$ and the geometry of $\overline{\Phi}$ around $\phi^*$ induced by the Hellinger distance on $\overline{\mathscr{M}}$. The next result is a consequence of Proposition 3.2 and Corollary 3.4.

**Theorem 3.8.** *Let $N \geq K + L$ and $Y_1, \ldots, Y_N$ be arbitrary random variables. Let $P_{\hat{\phi}} = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\delta$ given by (3.38). Under Assumption 3.5, for all $\overline{\phi} \in \Phi^*$ there exists a positive constant $C(\overline{\phi})$ such that*

$$\begin{aligned} C\left(\overline{\phi}\right)\mathbb{E}\left[||\overline{w} - \hat{w}||_2^2 + \left|\left|\overline{Q} - \hat{Q}\right|\right|_2^2 + \sum_{k=1}^K \left|\left|\overline{\theta}_k - \hat{\theta}_k\right|\right|_2^2 \wedge 1\right] \\ \leq n^{-1}\sum_{i=1}^n h^2\left(P_{\overline{\phi}}, P_i\right) + n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind}\right) \\ + (s+1)LK^{L-1}\left(K + L(d_1 + \cdots + d_K)\right)\frac{\log n}{n}. \end{aligned} \tag{3.59}$$

*In particular under Assumption 3.2, there exist positive constants $C(\overline{\phi},Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have*

$$\begin{aligned} C\left(\overline{\phi},Q^*\right)\mathbb{E}\left[||\overline{w} - \hat{w}||_2^2 + \left|\left|\overline{Q} - \hat{Q}\right|\right|_2^2 + \sum_{k=1}^K \left|\left|\overline{\theta}_k - \hat{\theta}_k\right|\right|_2^2 \wedge 1\right] \\ \leq h^2\left(P^*, P_{\overline{\phi}}\right) + LK^{L-1}\left(K + L(d_1 + \cdots + d_K)\right)\frac{s\log n}{n}, \end{aligned} \tag{3.60}$$

*where $P^*$ is given by (3.27).*

Inequality (3.59) is a consequence of Proposition 3.2 and Corollary 3.4. It does not require any assumption on the data and shows that the estimators of the parameters can be meaningful even if the model is misspecified. Ideally there exists $\overline{\phi}$ in $\Phi^*$ such that most of the distributions $P_i$ lie in a small neighborhood of $P_{\overline{\phi}}$ so that the first term of our bound is small compared to the last term. In that case the estimators $\hat{w}, \hat{Q}, \hat{\theta}_1, \ldots, \hat{\theta}_K$ converge to a small neighborhood around $\overline{w}, \overline{Q}, \overline{\theta}_1, \ldots, \overline{\theta}_K$, where $P_{\overline{\phi}}$ should be seen as the best approximation of the true distribution in the model. We can deduce bounds on the convergence rate of our parameter estimators in the well-specified case from (3.60). For $P^* = P_{\phi^*} \in \mathscr{H}^*\left(K, \overline{\mathscr{F}}_1, \ldots, \overline{\mathscr{F}}_K\right)$ with $\phi^* \in \Phi^*$ and for the

optimal choice of $s$, we retrieve the usual parametric rate for each parameter estimator, up to a logarithmic factor. Let us illustrate this with the following example.

We consider exponential distributions for the emission models, i.e. we have $\overline{\mathscr{F}}_i = \overline{\mathscr{E}}$ for all $i$ in $[K]$ with

$$\overline{\mathscr{E}} := \left\{ f_\theta \cdot \nu; f_\theta \in \mathcal{E}\left(\overline{\Theta}, \mathrm{id}_\Theta, -\mathrm{id}_{\mathscr{X}}, 1, 0\right)\right\} \tag{3.61}$$

where $\overline{\Theta} = (0,\infty)$, $\mathscr{X} = [0,\infty)$, $\nu$ is the Lebesgue measure on $\mathscr{X}$, and we can deduce $A : \theta \mapsto \log\theta$. This means we have $f_\theta : x \mapsto \theta e^{-\theta x}\mathbb{1}_{x\geq 0}$ for any $\theta > 0$. One can easily check that we satisfy Assumption 3.5, the last condition being a direct consequence of the dominated convergence theorem. We define $\overline{\Phi}$ by

$$\overline{\Phi} = O_K^{K+1} \times \left\{\theta \in \Theta^K; \theta_1 > \theta_2 > \cdots > \theta_K\right\}, \tag{3.62}$$

and $\overline{\mathscr{M}}$ as in (3.56). The condition on the parameters $\theta$ ensures identifiability over $\overline{\Phi}$ and $\overline{\Phi}^* = \overline{\Phi}$. The choice $L = 3$ is enough to obtain the result of Proposition 3.2. The next theorem is proven in Section 3.D.6.

**Theorem 3.9.** *Let $N \geq K+3$ and $Y_1, \ldots, Y_N$ be arbitrary random variables. Let $P_{\hat{\phi}} = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\delta$ given by (3.38). For any $\overline{\phi}$ in $\overline{\Phi}$ there exists a positive constant $C(\overline{\phi})$ such that we have*

$$\begin{aligned} &C\left(\overline{\phi}\right)\mathbb{E}\left[||\overline{w} - \hat{w}||_2^2 + \left|\left|\overline{Q} - \hat{Q}\right|\right|_2^2 + \sum_{k=1}^K \left(\overline{\theta}_k - \hat{\theta}_k\right)^2 \wedge 1\right] \\ &\leq n^{-1}\sum_{i=1}^n h^2\left(P_{\overline{\phi}}, P_i\right) + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) + (s+1)K^3\frac{\log n}{n}. \end{aligned}$$

*In particular under Assumption 3.2, there exist positive constants $C(\overline{\phi},Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have*

$$\begin{aligned} &C\left(\overline{\phi},Q^*\right)\mathbb{E}\left[||\overline{w} - \hat{w}||^2 + \left|\left|\overline{Q} - \hat{Q}\right|\right|^2 + \sum_{k=1}^K \left(\overline{\theta}_k - \hat{\theta}_k\right)^2 \wedge 1\right] \\ &\leq h^2\left(P^*, P_{\overline{\phi}}\right) + sK^3\frac{\log n}{n}, \end{aligned} \tag{3.63}$$

*where $P^*$ is given by (3.27).*

Our different parameter estimators all reach the usual parametric rate up to a logarithmic factor. One can notice that the ordering of the $\theta_k$ in (3.62) can be replaced by considering only distinct values and taking the infimum over permutation of the hidden states.

It is possible to follow the same scheme to obtain similar results for other exponential families, including HMMs with different exponential families as emission models. The difficulty relies in determining the set $\Phi^*$ given by (3.57).

### Another example

In this section we consider a relatively simple example that does not fit any framework already investigated but for which we can obtain risk bounds for the estimation of the parameters. Let $\nu$ be the Lebesgue measure on $\mathbb{R}$ and $\alpha$ be in $(0,1)$. We denote by $f_\alpha$ the probability density function with respect to $\nu$ defined by

$$f_\alpha : x \in \mathbb{R} \mapsto \frac{1-\alpha}{2}\frac{\mathbb{1}_{|x|\in[0,1]}}{|x|^\alpha},$$

with the convention $1/0 = +\infty$. For $z$ in $\mathbb{R}$, we denote by $F_{\alpha,z}$ the probability distribution associated to the density $x \mapsto f_\alpha(x-z)$. We fix $L=2$ and consider the model $\overline{\mathscr{M}}$ defined by

$$\overline{\mathscr{M}} = \{P_{w,q,z}; w,q_{12},q_{21} \in [0,1], z \in \mathbb{R}\},$$

where

$$\begin{aligned} P_{w,q,z} = {} & wF_{\alpha,0} \otimes [(1-q_{12})F_{\alpha,0} + q_{12}F_{\alpha,z}] \\ & + (1-w)F_{\alpha,z} \otimes [q_{21}F_{\alpha,0} + (1-q_{21})F_{\alpha,z}]. \end{aligned}$$

The distributions $P_{w,q,z}$ correspond to translation hidden Markov models with one known location parameter. The following result is proven in Section 3.D.8 and shows that we can deduce the parameters from the distribution $P_{w,q,z}$.

**Proposition 3.3.** *For $z^* \neq 0$, $w^* < 1$ and $q^*_{21} < 1$, there is a constant $C(\alpha,z^*,w^*,q^*)$ such that we have*

$$\begin{aligned} C(\alpha,z^*,w^*,q^*)h^2\left(P_{w,q,z}, P_{w^*,q^*,z^*}\right) \geq {} & \left(|z-z^*| \wedge 1\right)^{1-\alpha} + (w^*)^2\left(q_{12}-q^*_{12}\right)^2 \\ & + (1-w^*)^2\left(q_{12}-q^*_{12}\right)^2 + \left(w-w^*\right)^2, \end{aligned}$$

*for all $w,q_{12},q_{21} \in [0,1]$ and all $z \in \mathbb{R}$.*

We can deduce a deviation bound for the parameter estimators. The model $\overline{\mathscr{M}}$ is a subset of $\mathscr{H}(2,\overline{\mathscr{F}}_\alpha,\overline{\mathscr{F}}_\alpha)$, with $\overline{\mathscr{F}}_\alpha = \{F_{\alpha,z}; z \in \mathbb{R}\}$. We satisfy Assumption 3.3 with $\epsilon = 0$, $\mathcal{F}_\alpha = \{f_\alpha(\cdot - z); z \in \mathbb{Q}\}$ and $\overline{V} = 784$. The next result is proven in Section 3.D.7.

**Theorem 3.10.** *Let $N \geq K+2$ and $P_{\hat{w},\hat{q},\hat{z}} = \hat{P}_{s,\delta}$ be the estimator given by (3.37) with $\delta$ given by (3.38). For all $\overline{z} \neq 0$, $\overline{w} < 1$, $\overline{q}_{12} \in [0,1]$ and $\overline{q}_{21} < 1$, there exists a positive constant $C(\alpha,\overline{z},\overline{w},\overline{q})$ such that we have*

$$\begin{aligned} & C\left(\alpha,\overline{z},\overline{w},\overline{q}\right)\mathbb{E}\left[(\overline{w}-\hat{w})^2 + (\overline{q}_{12}-\hat{q}_{12})^2 + (\overline{q}_{12}-\hat{q}_{12})^2 + \left(|\overline{z}-\hat{z}| \wedge 1\right)^2\right] \\ & \leq \frac{1}{n}\sum_{i=1}^{n} h^2\left(P_{\overline{w},\overline{q},\overline{z}}, P_i\right) + \frac{1}{n}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) + (s+1)\frac{\log n}{n}. \end{aligned} \tag{3.64}$$

*In particular under Assumption 3.2, there exist positive constants $C(\overline{\phi},Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have*

$$\begin{aligned} & C\left(\overline{\phi},Q^*\right)\mathbb{E}\left[(\overline{w}-\hat{w})^2 + (\overline{q}_{12}-\hat{q}_{12})^2 + (\overline{q}_{21}-\hat{q}_{21})^2 + \left(|\overline{z}-\hat{z}| \wedge 1\right)^{1-\alpha}\right] \\ & \leq h^2\left(P^*, P_{\overline{w},\overline{q},\overline{z}}\right) + \frac{s\log n}{n}, \end{aligned} \tag{3.65}$$

*where $P^*$ is given by (3.27).*

Inequality (3.64) does not require any assumption on the data. It is a consequence of Proposition 3.3 and Theorem 3.4. We can deduce convergence rates for our parameter estimators from (3.65) for $P^* = P_{\pi^*,q^*,z^*}$ with $z^* \neq 0$, $w^* < 1$ and $q^*_{21} < 1$. The estimators $\hat{w}$ and $\hat{q}$ achieve the usual parametric rate up to a logarithmic factor. However the location estimator $\hat{z}$ reaches the faster rate $(n^{-1}\log^2 n)^{1/(1-\alpha)}$. This rate is optimal up the logarithmic factor. It is a consequence of Theorem 1.1 in [54] (Chapter VI), noticing that $f_\alpha$ has a singularity of order $-\alpha$ in 0, and with the fact that we cannot do better than $1/n$ for the Hellinger distance. One should notice that $f_\alpha$ is unbounded for all $\alpha \in (0,1)$. Therefore the maximum likelihood and the least squares estimators are undefined and those methods do not apply on $\overline{\mathscr{M}}$. In addition, we can see that $f_\alpha$ is not square integrable for $\alpha \in [1/2,1)$.

## 3.5 Selection of the spacing parameter

Until now we gave results that required a good choice of the spacing parameter $s$, given some bound on the dependence term $\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$. This section propose a way to automatically select a value of $s$ from the data, assuming that we dispose of two independent sets of observations. We use the first set to produce an estimator $\hat{P}_s$ for different values of $s$. We then use the second set to produce an estimator $\hat{s}$ of the optimal value of $s$.

### 3.5.1 Framework and result

Let $X_1^{(1)},\dots,X_{n_1}^{(1)},X_1^{(2)},\dots,X_{n_2}^{(2)}$ be $n_1+n_2$ random variables on the measurable space $(\mathscr{X},\mathcal{X})$. We define $P_i^{(j)}$ by $P_i^{(j)} := \mathcal{L}(X_i^{(j)})$ for all $j$ in $[2]$ and all $i$ in $[n_j]$. We also write

$$\mathbf{P}^*_{s,b} = \mathcal{L}\left(X_b^{(1)},\dots,X_{b+n_1(s,b)(s+1)}\right) \text{ and } \mathbf{P}^{ind}_{s,b} = \bigotimes_{i=1}^{n_1(s,b)} \mathcal{L}\left(X^{(1)}_{b+(i-1)(s+1)}\right),$$

with

$$n_1(s,b) = \left\lfloor \frac{n_1+s+1-b}{1+s} \right\rfloor. \tag{3.66}$$

Let $S$ be a subset of $\{0,1,\dots,s_{\max}\}$, $s_{\max} = \lfloor (n_1-2)/2 \rfloor$. Let $(\mathscr{M}_s)_{s\in S}$ be countable subsets of $\mathscr{P}_X$ such that the $\rho$-dimension function (see Section 3.B) is uniformly bounded over $\mathscr{M}_s$ by a non-decreasing function $m \mapsto D_m(\mathscr{M}_s) \geq 1$ for all $s \in S$. We follow the procedure below.

1. For $s$ in $S$, let $\hat{P}_s = \hat{P}_s\left(\mathscr{M}_s,\mathbf{X}^{(1)}\right)$ be the estimator given by (3.12). Conditionally on $\mathbf{X}^{(1)}$, we define the finite model
$$\widehat{\mathscr{M}_S} = \widehat{\mathscr{M}_S}\left(\mathbf{X}^{(1)}\right) := \left\{\hat{P}_s : s \in S\right\}.$$
2. Let $\hat{P}$ be the $\rho$-estimator $\hat{P} = \hat{P}\left(n_2,\mathbf{X}^{(2)},\widehat{\mathscr{M}_S}\right)$ given by (3.7). We denote by $\hat{s}$ the value of $s$ such that $\hat{P} = \hat{P}_{\hat{s}}$ and we write
$$\hat{P} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right). \tag{3.67}$$

We make the following assumption.

**Assumption 3.6.** *The random variables*

$$\mathbf{X}^{(1)} := \left(X_1^{(1)},\dots,X_{n_1}^{(1)}\right) \text{ and } \mathbf{X}^{(2)} := \left(X_1^{(2)},\dots,X_{n_2}^{(2)}\right)$$

*are independent.*

The following result is proven in Section 3.E.1.

**Theorem 3.11.** *Let $n_1,n_2 \geq 3$ and $\hat{P} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)$ be the estimator given by (3.67). Under Assumption 3.6, there exists a positive constant $C > 0$ such that for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] \leq\ & n_1^{-1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right) + n_2^{-1}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) \\ & + \inf_{t\in[n_2]}\left\{\frac{t}{n_2}\left(1+\log(|S|)\right) + \lceil n_2/t\rceil \beta_t\left(\mathbf{X}^{(2)}\right)\right\} \\ & + \inf_{s\in S}\left\{h^2\left(\overline{P},\mathscr{M}_s\right) + \frac{(s+1)D_{n_1(s,1)}(\mathscr{M}_s)}{n_1} + n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right\}, \end{aligned} \tag{3.68}$$

*where the mixing coefficient $\beta_t\left(\mathbf{X}^{(2)}\right)$ is given by (1.2.5) in Dedecker* et al. *[27].*

One can check that we do not need any assumption other than Assumption 3.6 to obtain this result. We need to make additional assumptions a posteriori to make this bound meaningful. Let us interpret this inequality in simpler cases. We consider there is $\mathscr{M}$ such that $\mathscr{M}_s = \mathscr{M}$ for all $s \in S$. If the data were truly i.i.d. with distribution $\overline{P} \in \mathscr{M}$, we would get

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}\right)\right] \leq \frac{(s+1)D_{n_1(s,1)}(\mathscr{M})}{n_1} + \frac{(1+\log(|S|))}{n_2}.$$

The second term is the bound we get for i.i.d. estimation from a $n_2$-sample over a finite model of cardinal $|S|$. When the data are not identically distributed, the quantity

$$n_2^{-1}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + n_1^{-1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)$$

is not zero but it remains small when most of the true marginal distributions $P_i^{(j)}$ lie close enough to some distribution $\overline{P}$ in $\mathscr{M}$. The terms $n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}_{s,b}^*||\mathbf{P}_{s,b}^{ind}\right)$ and $\lceil n_2/t\rceil\beta_t(\mathbf{X}^{(2)})$ account for the possible dependence within $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ respectively. They vanish if the observations $X_1^{(1)},\ldots,X_{n_1}^{(1)},X_1^{(2)},\ldots,X_{n_2}^{(2)}$ are all independent. Contrary to Theorem 3.4 we do not have to choose a good value of $s$ as the method automatically select a reasonable $s$ in $S$ as long as the $P_i^{(j)}$ can be well approximated by a distribution $\overline{P} \in \mathscr{M}$.

### 3.5.2 Robustness

Let $\overline{\mathbf{X}}^{(1)} = \left(\overline{X}_1^{(1)},\ldots,\overline{X}_{n_1}^{(1)}\right)$ and $\overline{\mathbf{X}}^{(2)} = \left(\overline{X}_1^{(2)},\ldots,\overline{X}_{n_2}^{(2)}\right)$ be the true processes of interest such that $P_i^{(j)} = \overline{P}$ for all $j \in [2]$ and $i \in [N_j]$. We actually observe a contaminated version of it. Let $Z_1^{(1)},\ldots,Z_{N_1}^{(1)},Z_1^{(2)},\ldots,Z_{N_2}^{(2)}$ be random variables with any distributions and $E_1^{(1)},\ldots,E_{N_1}^{(1)},E_1^{(2)},\ldots,E_{N_2}^{(2)}$ be Bernoulli random variables such that for all $j \in [2]$ and all $i \in [N_j]$,

$$X_i^{(j)} = E_i\overline{X}_i^{(j)} + (1-E_i^{(j)})Z_i^{(j)}. \tag{3.69}$$

For $s \in \{0,1,\ldots,s_{\max}\}$ and $b \in [s+1]$, we define the distributions

$$\overline{\mathbf{P}}_{s,b}^* = \mathcal{L}\left(\overline{X}_b^{(1)},\ldots,\overline{X}_{b+n_1(s,b)(s+1)}^{(1)}\right) \text{ and } \overline{\mathbf{P}}_{s,b}^{ind} = \bigotimes_{i=1}^{n_1(s,b)}\mathcal{L}\left(\overline{X}_{b+(i-1)(s+1)}^{(1)}\right).$$

The next result is a complement of Lemma 3.2 and is proven in Section 3.E.2.

**Lemma 3.7.** *If $E_1^{(1)},Z_1^{(1)},\ldots,E_{n_1}^{(1)},Z_{n_1}^{(1)},E_1^{(2)},Z_1^{(2)},\ldots,E_{n_2}^{(2)},Z_{n_2}^{(2)},\overline{\mathbf{X}}^{(1)}$ and $\overline{\mathbf{X}}^{(2)}$ are mutually independent, we have*

$$\mathbf{K}\left(\mathbf{P}_{s,b}^*||\mathbf{P}_{s,b}^{ind}\right) \leq \mathbf{K}\left(\overline{\mathbf{P}}_{s,b}^*||\overline{\mathbf{P}}_{s,b}^{ind}\right),\forall s \in \{0,1,\ldots,s_{\max}\},\forall b \in [s+1], \tag{3.70}$$

*and*

$$\beta_t\left(\mathbf{X}^{(2)}\right) \leq \beta_t\left(\overline{\mathbf{X}}^{(2)}\right),\forall t \geq 1.$$

We define $p_i^{(j)}$ by $\mathbb{P}\left(E_i^{(j)}=1\right) = p_i^{(j)}$ for $j \in [2]$ and $i \in [N_j]$.

**Corollary 3.5.** *Let $n_1,n_2 \geq 3$ and $\hat{P} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)$ be the estimator given by (3.67). There exists a positive constant $C$ such that in the situation of Lemma 3.7 and for all $\overline{P} \in \mathscr{P}_X$,*

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] \leq n_1^{-1}\sum_{i=1}^{n_1}(1-p_i^{(1)}) + n_2^{-1}\sum_{i=1}^{n_2}(1-p_i^{(2)}) + \inf_{t\in[n_2]}\left\{\frac{t}{n_2}\left(1+\log(|S|)\right) + \lceil n_2/t\rceil \beta_t\left(\overline{\mathbf{X}}^{(2)}\right)\right\} + \inf_{s\in S}\left\{h^2\left(\overline{P},\mathscr{M}_s\right) + \frac{(s+1)D_{n_1(s,1)}(\mathscr{M}_s)}{n_1} + n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\overline{\mathbf{P}}^*_{s,b}||\overline{\mathbf{P}}^{ind}_{s,b}\right)\right\}.$$

This result is a direct consequence of Theorem 3.11 and Lemma 3.7. We illustrate the performance of our estimator with hidden Markov models.

### 3.5.3 Application to hidden Markov models

Let $Y_1^{(1)},\ldots,Y_{N_1}^{(1)},Y_1^{(2)},\ldots,Y_{N_2}^{(2)}$ be random variables taking values in the measurable space $(\mathscr{Y},\mathcal{Y})$. Let $L$ be in $\{2,3,\ldots,\lfloor(N_1 \wedge N_2)/2\rfloor\}$ and $n_j = N_j + 1 - L$ for $j \in [2]$. We define the new random variables

$$X_i^{(j)} = \left(Y_i^{(j)},Y_{i+1}^{(j)},\ldots,Y_{i+L-1}^{(j)}\right), i \in [n_j], j \in [2],$$

taking values in the measurable space $(\mathscr{X},\mathcal{X}) = \left(\mathscr{Y}^L,\mathcal{Y}^{\otimes L}\right)$. We adapt Assumption 3.2 to this context.

**Assumption 3.7.** *Let $\left(Y_i^{(1)},H_i^{(1)}\right)_i$ and $\left(Y_i^{(2)},H_i^{(2)}\right)_i$ be finite state space HMM with parameters $(K^*,w_1^*,Q^*,F^*)$ and $(K^*,w_2^*,Q^*,F^*)$ such that $Q^*$ is irreducible and aperiodic.*

Under this assumption $Q^*$ has only one invariant distribution $\pi^*$ and we define the distribution $P^*$ by (3.27). Let $\tau \geq e$ and $J = \lfloor \log_\tau(\lfloor(n_1-2)/2\rfloor)\rfloor$. Let $S$ be the set given by

$$S = \{0\} \cup \left\{\left\lceil \tau^j \right\rceil ; j \in \{0,1,\ldots,J\}\right\}. \tag{3.71}$$

Let $\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K$ be subsets of $\mathscr{P}_Y$ such that Assumption 3.3 is satisfied. Let $\overline{\mathscr{M}}$ be a non-empty subset of the model $\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K\right)$ defined by (3.29). For $s$ in $S$, we take $\mathscr{M}_s = \mathscr{M}_{\delta(s)}$ with

$$\delta(s) = \frac{\overline{V}}{n_1(s,1)(K-1)} \bigwedge \frac{1}{K},$$

where $\mathscr{M}_\delta$ is given by (3.36) and $n_1(s,1)$ given by (3.66). The following result is proven in Section 3.E.3.

**Theorem 3.12.** *Let $N_1,N_2 \geq K+L$ and $\hat{P} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)$ be the estimator given by (3.67). Under Assumption 3.6, there is a numeric constant $C > 0$ such that for all $\overline{P} \in \mathscr{P}_X$*

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] \leq h^2\left(\overline{P},\overline{\mathscr{M}}\right) + n_1^{-1}\sum_{i=1}^{n_1}h^2\left(P_i^{(1)},\overline{P}\right) + n_2^{-1}\sum_{i=1}^{n_2}h^2\left(P_i^{(2)},\overline{P}\right) + L\epsilon^2 + \inf_{t\in[n_2]}\left\{\frac{t\log\log n_1}{n_2} + \lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right)\right\} + \inf_{s\in S}\left\{\frac{(s+1)L\overline{V}\log n_1}{n_1} + n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right\}. \tag{3.72}$$

*In particular under Assumption 3.7, there exists a positive constant $C(Q^*)$ such that*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_{\hat{s}}\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + L\epsilon^2 + \tau L\overline{V}\frac{\log^2 n_1}{n_1} + \frac{\log n_2 \log\log n_1}{n_2}, \tag{3.73}$$

*where $P^*$ is given by (3.27).*

Inequality (3.72) is a consequence of Theorem 3.11 and only requires Assumption 3.6. Under Assumption 3.7 we can control the different terms and obtain (3.73). If $\epsilon = 0$, the ideal situation is to have the same number of observations in each set, i.e. $n_1 = n_2 = n$. In this case we have

$$C\left(Q^*\right)\mathbb{E}\left[h^2\left(P^*,\hat{P}\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + L\tau\overline{V}\frac{\log^2 n}{n},$$

and the first vanishes when the model is well specified which gives the rate $n^{-1}\log^2 n$ with respect to the squared Hellinger distance over $\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K\right)$. When $\epsilon > 0$ the quantity $\overline{V}$ depends on $\epsilon$ and we need to balance the second and third term in (3.73), i.e. $\epsilon^2/\overline{V}$ is of order $n_1^{-1}$ up to a logarithmic term. Then the ideal situation only requires $n_2$ to be of order $\epsilon^{-2}$ up to logarithmic term and the bound on the convergence rate is of order $\epsilon^2$. For example, we would have $\epsilon^{-2} = n_1^{\frac{2}{d+1}}\log^{-\frac{2}{d+1}-(d+2)} n_1$ in the situation of Theorem 3.6. In both cases, it shows that we recover a value of $s$ that allows us to obtain the same rate as when the optimal value is known. This is especially interesting for the robustness aspect of our estimator.

Let us consider a situation similar to Section 3.5.2. Let $Z_1^{(1)},\ldots,Z_{N_1}^{(1)},Z_1^{(2)},\ldots,Z_{N_2}^{(2)}$ be random variables with any distributions and $E_1^{(1)},\ldots,E_{N_1}^{(1)},E_1^{(2)},\ldots,E_{N_2}^{(2)}$ be Bernoulli random variables such that for all $j \in [2]$ and all $i \in [N_j]$,

$$Y_i^{(j)} = E_i\overline{Y}_i^{(j)} + (1 - E_i^{(j)})Z_i^{(j)}.$$

The following result is proven in Section 3.E.4.

**Corollary 3.6.** *Let $\hat{P}_{\hat{s}} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)$ be the estimator given by (3.67). If $E_1^{(1)},Z_1^{(1)},\ldots,E_{n_1}^{(1)},Z_{n_1}^{(1)}$, $E_1^{(2)},Z_1^{(2)},\ldots,E_{n_2}^{(2)},Z_{n_2}^{(2)},\overline{\mathbf{X}}^{(1)}$ and $\overline{\mathbf{X}}^{(2)}$ are mutually independent, and if $\overline{\mathbf{Y}}^{(1)}$ and $\overline{\mathbf{Y}}^{(2)}$ satisfy Assumption 3.7, there exists a positive constant $C(Q^*)$ such that*

$$\begin{aligned} C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_{\hat{s}}\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) &+ \frac{L}{N_1}\sum_{i=1}^{N_1}\left(1 - p_i^{(1)}\right) + \frac{L}{N_2}\sum_{i=1}^{N_2}\left(1 - p_i^{(2)}\right) \\ &+ L\epsilon^2 + \tau L\overline{V}\frac{\log^2 n_1}{n_1} + \frac{\log n_2 \log\log n_1}{n_2}, \end{aligned}$$

*where $P^*$ is given by (3.27) and $p_i^{(j)} = \mathbb{P}\left(E_i^{(j)} = 1\right)$ for all $j \in [2]$ and $i \in [N_j]$.*

One can see that our deviation bound is not significantly worse as long as the average proportions of contamination $N_1^{-1}\sum_{i=1}^{N_1}(1 - p_i^{(1)})$ and $N_2^{-1}\sum_{i=1}^{N_2}(1 - p_i^{(2)})$ are small compared to $\epsilon^2 + \tau\overline{V}\frac{\log^2 n_1}{n_1}$ and $\frac{\log n_2 \log\log n_1}{n_1}$ respectively. We interpret this result further for $P^* \in \overline{\mathscr{M}}$, $\epsilon = 0$ and $n_1 = n_2 = n$. Let us consider Hüber's contamination model with $p_i^{(j)} = 1 - \alpha_{cont}$ for all $j \in [2]$ and $i \in [N]$. In this situation we get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\left[\alpha_{cont} + \frac{\tau\overline{V}\log^2 n}{n}\right].$$

Our bound on the convergence rate is not deteriorated as long as the contamination rate $\alpha_{cont}$ is small compared to $\frac{\tau\overline{V}\log^2 n}{n}$. We can also consider the situation $\mathbb{P}(E_i^{(j)} = 0) = \mathbb{1}_{i\in I_j}$ for some subsets $I_1 \subset [N]$ and $I_2 \subset [N]$. We get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq L\left[\frac{|I_1| + |I_2|}{N} + \frac{\tau\overline{V}\log^2 n}{n}\right].$$

Our bound on the convergence rate is not deteriorated as long as the proportions of outliers $|I_1|/N$,$|I_2|/N$ are small compared to the other term $\frac{\tau\overline{V}\log^2 n}{n}$.

# Appendix

## 3.A Auxiliary results

We denote by $\mathrm{C}(\mathscr{X})$ the set given by

$$\mathrm{C}(\mathscr{X}) = \bigcup_{n\geq 1} \{n\} \times \mathscr{X}^n.$$

Let $d : \mathscr{A} \times \mathscr{A} \to \mathbb{R}$ be a loss function where $\mathscr{A} \subset \mathscr{P}_X$ denotes a set of admissible probability distributions. Let $\mathscr{M}$ be a subset of $\mathscr{A}$. Let $\hat{P} : C(\mathscr{X}) \to \mathscr{M}$ be an estimation method.

**Assumption 3.8.** *There exist constants $C_0 > 0, \beta \in (0,1]$ and non decreasing functions $f,g$ such that for all independent random variables $X_1,\ldots,X_n$ with distributions $P_1,\ldots,P_n \in \mathscr{A}$ and for all $\xi > 0$*

$$\mathbb{P}\left(\sum_{i=1}^{n} d\left(P_i,\hat{P}(n,\mathbf{X})\right) \leq C_0 \inf_{Q\in\mathscr{M}} \sum_{i=1}^{n} d(P_i,Q) + f(n) + g(n)\xi^{\beta}\right) \geq 1 - e^{-\xi}.$$

Many estimators satisfy such an assumption, see for instance mean discrepancy estimators [5], $T$-estimators [16] or $l$-estimators [10]. We can get rid of the independence assumption with the following result.

**Proposition 3.4.** *Under Assumption 3.8, for all random variables $X_1,\ldots,X_n$ with distributions $P_1,\ldots,P_n \in \mathscr{A}$ we have*

$$\mathbb{E}\left[\sum_{i=1}^{n} d\left(P_i,\hat{P}(n,\mathbf{X})\right)\right] \leq C_0 \inf_{Q\in\mathscr{Q}} \sum_{i=1}^{n} d(P_i,Q) + f(n) + g(n)\left[2 + \frac{3}{2}\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)\right]^{\beta},$$

*where*

$$\mathbf{P}^* = \mathcal{L}(X_1,\ldots,X_n) \text{ and } \mathbf{P}^{ind} = \mathcal{L}(X_1)\otimes\ldots\otimes\mathcal{L}(X_n).$$

This result is obtained by applying Lemma 3.1 that we prove hereafter, with $\mathbf{P} = \mathbf{P}^{ind}$ and $\mathbf{Q} = \mathbf{P}^*$.

### 3.A.1 Proof of Lemma 3.1

We use Lemma 48 in [9]. For $\lambda \in (0,a^{-1/\beta})$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Q}}\left[\lambda\left(nl\left(\hat{\theta}(\mathbf{X}),\theta\right) - nA - B\right)_+^{1/\beta}\right] \\
&\leq \log\left(1 + \int_0^{+\infty} e^{\xi}\mathbf{P}\left(l\left(\hat{\theta}(\mathbf{X}),\theta\right) > A + \frac{B + (\xi/\lambda)^{\beta}}{n}\right)d\xi\right) + \mathbf{K}(\mathbf{Q}||\mathbf{P}) \\
&\leq \log\left(1 + \int_0^{+\infty} e^{\xi}e^{-\xi/\lambda}d\xi\right) + \mathbf{K}(\mathbf{Q}||\mathbf{P}) = \log\left(\frac{1}{1-\lambda}\right) + \mathbf{K}(\mathbf{Q}||\mathbf{P}).
\end{aligned}$$

We have

$$\mathbb{E}_{\mathbf{Q}}\left[\left(nl\left(\hat{\theta}(\mathbf{X}),\theta\right) - nA - B\right)_+^{1/\beta}\right] \leq \lambda^{-1}\left[\log\left(\frac{1}{1-\lambda}\right) + \mathbf{K}(\mathbf{Q}||\mathbf{P})\right].$$

Assuming $\mathbf{K}(\mathbf{Q}||\mathbf{P}) < \infty$, minimization over $\lambda$ demands

$$\log(1-\lambda) - \mathbf{K}(\mathbf{Q}||\mathbf{P}) + \frac{\lambda}{1-\lambda} = 0.$$

Let $\lambda^*$ be such a number. In that case

$$(\lambda^*)^{-1}\left[\log\left(\frac{1}{1-\lambda^*}\right) + \mathbf{K}(\mathbf{Q}||\mathbf{P})\right] = \frac{1}{1-\lambda^*}.$$

We set $a(x) = x - \log(1+x)$ for $x$ in $(0,+\infty)$. Following the proof of Proposition 5 [9], $a$ is increasing and

$$\forall x > 0, a^{-1}(x) \leq x + \sqrt{2x}.$$

Since $\frac{\lambda^*}{1-\lambda^*} = a^{-1}(\mathbf{K}(\mathbf{Q}||\mathbf{P}))$, we get

$$\begin{aligned}\frac{1}{1-\lambda^*} = 1 + \frac{\lambda^*}{1-\lambda^*} &\leq 1 + \mathbf{K}(\mathbf{Q}||\mathbf{P}) + \sqrt{2\mathbf{K}(\mathbf{Q}||\mathbf{P})} \\ &\leq 2 + \frac{3}{2}\mathbf{K}(\mathbf{Q}||\mathbf{P}).\end{aligned}$$

Finally, with Jensen's inequality we get

$$\mathbb{E}_{\mathbf{Q}}\left[l\left(\hat{\theta}(\mathbf{X}),\theta\right)\right] \leq A + \frac{B + \left(2 + \frac{3}{2}\mathbf{K}(P||Q)\right)^{\beta}}{n}.$$

## 3.B Main results

This section gathers the proofs of Theorem 3.1, Corollary 3.1 and Lemmas 3.2, 3.3, 3.4. We first give a formal definition of the $\rho$-dimension function that is originally introduced in Baraud & Birgé [11]. We slightly modify some notation to adapt it to our context. The function $\psi$ defined by (3.4) satisfies Assumption 2 [11] with $a_0 = 4, a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see Proposition 3 [11]). Let $n$ be a positive integer and $\mathscr{M}$ be a countable subset of $\mathscr{P}_X$. For $y > 0$, $\mathbf{P}^{ind} = \bigotimes_{i=1}^n P_1^{ind} \in \mathscr{P}_X^{\otimes n}$ and $P \in \mathscr{M}$ we write

$$\mathscr{B}^{\mathscr{M}}\left(\mathbf{P}^{ind},\overline{P},y\right) := \left\{Q \in \mathscr{M}; \sum_{i=1}^n h^2\left(P_i^{ind},P\right) + h^2\left(P_i^{ind},Q\right) < y^2\right\}.$$

If $\mathcal{M}$ is a countable set of probability density functions with respect to a $\sigma$-finite measure $\nu$ such that $\mathscr{M} = \{Q = q\cdot\nu; q \in \mathcal{M}\}$, we write

$$w\left(\nu,\mathcal{M},\mathscr{M},\mathbf{P}^{ind},P,y\right) = \mathbb{E}_{\mathbf{X}\sim\mathbf{P}^{ind}}\left[\sup_{Q\in\mathscr{B}^{\mathscr{M}}(\mathbf{P}^{ind},P,y)} |\mathbf{Z}_n(\mathbf{X},p,q)|\right],$$

where

$$\mathbf{Z}_n(\mathbf{X},q,q') := \mathbf{T}_n(\mathbf{X},q,q') - \mathbb{E}_{\mathbf{P}^{ind}}\mathbf{T}_n(\mathbf{X},q,q'),$$

and $\mathbf{T}_n$ is given by (3.5). We define $\mathbf{w}^{\mathscr{M}}\left(\mathbf{P}^{ind},P,y\right) = \inf_{(\nu,\mathcal{M})} w\left(\nu,\mathcal{M},\mathscr{M},\mathbf{P}^{ind},P,y\right)$, where the infimum is taken over all couples $(\nu,\mathcal{M})$ such that $\mathcal{M}$ is the class of density functions associated to $\mathscr{M}$ with respect to a $\sigma$-finite measure $\nu$. We define the $\rho$-dimension function by

$$D^{\mathscr{M}}\left(\mathbf{P}^{ind},P^{\otimes n}\right) = \left[\frac{3}{2^{21/2}}\sup\left\{y^2; \mathbf{w}^{\mathscr{M}}\left(\mathbf{P}^{ind},P,y\right) > \frac{3y^2}{64}\right\}\right] \bigvee 1.$$

As mentioned at the beginning of Section 3.2 we consider cases for which we have a uniform bound over the $\rho$-dimension function. More precisely we assume there is a non-increasing function $m \mapsto D_m(\mathscr{M})$ such that

$$D^{\mathscr{M}}\left(\mathbf{P}^{ind},P^{\otimes m}\right) \leq D_m(\mathscr{M}), \forall \mathbf{P}^{ind} \in \mathscr{P}_X^{\otimes m}, \forall P \in \mathscr{M}.$$

### 3.B.1 Proof of Theorem 3.1

From Theorem 1 of Baraud & Birgé [11], we have that for all independent random variables $X_1,\ldots,X_n$ with respective distributions $P_1,\ldots,P_n$, for all $Q \in \mathscr{M}$ and for all $\xi > 0$, we have

$$\sum_{i=1}^{n} h^2\left(P_i, \hat{P}(n,\mathbf{X},\mathscr{M})\right) \leq \gamma \sum_{i=1}^{n} h^2\left(P_i, Q\right) + \frac{4\kappa}{a_1}\left(\frac{D_n(\mathscr{M})}{4.7} + 1.49 + \xi\right),$$

with probability at least $1 - e^{-\xi}$, where $\gamma$ and $\kappa$ are given in [11] and satisfy $\gamma \leq 150$ and $\frac{4\kappa}{a_1} \leq 5014$ (see proof of Theorem 1 [13], page 32). We can take the infimum for $Q$ over $\mathscr{M}$ and it shows we satisfy Assumption 3.8 with $C_0 = 150$, $f(n) = 5014\left(\frac{D_n(\mathscr{M}))}{4.7} + 1.49\right)$, $g(n) = 5014$ and $\beta = 1$. From Proposition 3.4, we have

$$\mathbb{E}\left[\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_s\right)\right] \leq 150 \inf_{Q\in\mathscr{Q}} \sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, Q\right) + 5014\left(\frac{D_{n(s,b)}(\mathscr{M})}{4.7} + 3.49 + \frac{3}{2}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right),$$

for all $b \in [s+1]$. From (3.12), we have

$$\begin{aligned}
\sum_{i=1}^{n} h^2\left(P_i, \hat{P}_s\right) &= \sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_s\right) \\
&\leq 2\sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right) + 2\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b}, \hat{P}_s\right) \\
&\leq 2\sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right) + 2\inf_{Q\in\mathscr{M}}\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b}, Q\right) + 2\iota \\
&\leq 4\sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right) + 2\inf_{Q\in\mathscr{M}}\sum_{i=1}^{N} h^2\left(P_i, Q\right) + 2\iota.
\end{aligned}$$

Combining the inequalities above, we obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} h^2\left(P_i, \hat{P}_s\right)\right] &\leq 600\sum_{b=1}^{s+1}\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n(s,b)} h^2(P_{b+(i-1)(s+1)}, Q) + 2\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} l\left(P_i, Q\right) \\
&\quad + 20056\sum_{b=1}^{s+1}\left(\frac{D_{n(s,b)}(\mathscr{M})}{4.7} + 3.49 + \frac{3}{2}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right) + 2\iota \\
&\leq 602\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2(P_i, Q) + 20056(s+1)\left(\frac{D_{n(s,1)}(\mathscr{M})}{4.7} + 3.49\right) \\
&\quad + 30084\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) + 2\iota.
\end{aligned}$$

Since $\iota \leq 2546 < 20056 \times \frac{0.597}{4.7}$, we get

$$\mathbb{E}_{\mathbf{P}^*}\left[\sum_{i=1}^{n} h^2\left(P_i, \hat{P}_s\right)\right] \leq 602\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + \frac{20056}{4.7}(s+1)\left[D_{n(s,1)}(\mathscr{M}) + 17\right] + 30084\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right).$$

### 3.B.2 Proof of Lemma 3.2

For $\mathbf{e} \in \{0,1\}^n$, we denote by $I(\mathbf{e})$ the set given by $I(\mathbf{e}) = \{i \in [n]; e_i = 1\}$. From the convexity property of the Kullback-Leibler divergence, we have

$$\begin{aligned}
&\mathbf{K}\left(\mathcal{L}\left(\mathbf{Y}\right)||\mathcal{L}(Y_1) \otimes \cdots \otimes \mathcal{L}(Y_n)\right) \\
&\quad \leq \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}(\mathbf{E}=\mathbf{e})\mathbf{K}\left(\mathcal{L}\left(\mathbf{Y}|\mathbf{E}=\mathbf{e}\right)||\mathcal{L}(Y_1|E_1=e_1) \otimes \cdots \otimes \mathcal{L}(Y_n|E_n=e_N)\right) \\
&\quad = \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}(\mathbf{E}=\mathbf{e})\mathbf{K}\left(\mathcal{L}\left((X_i)_{i\in I(\mathbf{e})}\right) \otimes \bigotimes_{i\notin I(\mathbf{e})} \mathcal{L}\left(Z_i\right) || \bigotimes_{i\in I(\mathbf{e})} \mathcal{L}(X_i) \otimes \bigotimes_{i\notin I(\mathbf{e})} \mathcal{L}(Z_i)\right) \\
&\quad = \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}(\mathbf{E}=\mathbf{e})\mathbf{K}\left(\mathcal{L}\left((X_i)_{i\in I(\mathbf{e})}\right) || \bigotimes_{i\in I(\mathbf{e})} \mathcal{L}(X_i)\right).
\end{aligned}$$

We need an auxiliary result before ending the proof.

**Lemma 3.8.** *For random variables $A,B,C$ such that $\mathcal{L}(A) \ll \mathcal{L}(B)$, we have*

$$\mathbf{K}\left(\mathcal{L}(A)||\mathcal{L}(B)\right) \leq \mathbf{K}\left(\mathcal{L}\left(A,C\right)||\mathcal{L}(B) \otimes \mathcal{L}(C)\right). \tag{3.74}$$

With this result we have

$$\mathbf{K}\left(\mathcal{L}\left((X_i)_{i\in I(\mathbf{e})}\right) || \bigotimes_{i\in I(\mathbf{e})} \mathcal{L}(X_i)\right) \leq \mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}(X_1) \otimes \cdots \otimes \mathcal{L}(X_n)\right),$$

which allows us to conclude.

#### Proof of Lemma 3.8

Let $\mu_1$ and $\mu_2$ be measures dominating $\mathcal{L}(B)$ and $\mathcal{L}(C)$ respectively. We write

$$p_{B,C} = \frac{d\mathcal{L}(B,C)}{d\mu_1 \otimes \mu_2}, p_{A,C} = \frac{d\mathcal{L}(A,C)}{d\mu_1 \otimes \mu_2}, p_A = \frac{d\mathcal{L}(A)}{d\mu_1}, p_B = \frac{d\mathcal{L}(B)}{d\mu_1}, p_C = \frac{d\mathcal{L}(C)}{d\mu_2}.$$

We have

$$\begin{aligned}
\mathbf{K}\left(\mathcal{L}(A,C)||\mathcal{L}(B) \otimes \mathcal{L}(C)\right) &= \int p_{A,C}(x,z) \log\left(\frac{p_{A,C}(x,z)}{p_B(x)p_C(z)}\right) \mu_1(dx)\mu_2(dz) \\
&= \int p_{A,C}(x,z) \log\left(\frac{p_{A,C}(x,z)}{p_A(x)p_C(z)}\right) \mu_1(dx)\mu_2(dz) \\
&+ \int p_{A,C}(x,z) \log\left(\frac{p_A(x)}{p_B(x)}\right) \mu_1(dx)\mu_2(dz) \\
&= \mathbf{K}\left(\mathcal{L}(A,C)||\mathcal{L}(A) \otimes \mathcal{L}(C)\right) + \mathbf{K}\left(\mathcal{L}(A)||\mathcal{L}(B)\right).
\end{aligned}$$

The non-negativity of the Kullback-Leibler divergence concludes the proof.

### 3.B.3 Proof of Corollary 3.1

One can check that we have

$$\begin{aligned}
h^2\left(\overline{P},\hat{P}_s\right) &\leq 2n^{-1}\sum_{i=1}^n h^2\left(\mathcal{L}(Y_i),\overline{P}\right) + 2n^{-1}\sum_{i=1}^n h^2\left(\mathcal{L}(Y_i),\hat{P}_s\right) \\
&\leq 2n^{-1}\sum_{i=1}^n (1-p_i) + 2n^{-1}\sum_{i=1}^n h^2\left(\mathcal{L}(Y_i),\hat{P}_s\right),
\end{aligned}$$

and for $Q$ in $\mathscr{M}$

$$\begin{aligned}\sum_{i=1}^{n} h^2\left(\mathcal{L}(Y_i),Q\right) &\leq 2\sum_{i=1}^{n} h^2\left(\mathcal{L}(Y_i),\overline{P}\right) + 2\sum_{i=1}^{n} h^2\left(\overline{P},Q\right)\\ &\leq 2\sum_{i=1}^{n}(1-p_i) + 2nh^2\left(\overline{P},Q\right).\end{aligned}$$

We can conclude with Theorem 3.1 and Lemma 3.2.

### 3.B.4 Proof of Lemma 3.3

We have

$$\begin{aligned}\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right) &= \mathbb{E}\left[\mathbf{K}\left(\mathcal{L}\left(X_n|X_1,\ldots,X_{n-1}\right)||\mathcal{L}\left(X_n\right)\right)\right]\\ &\quad+\mathbf{K}\left(\mathcal{L}\left(X_1,\ldots,X_{n-1}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_{n-1}\right)\right),\end{aligned}$$

and with the Markov property

$$\begin{aligned}\mathbb{E}\left[\mathbf{K}\left(\mathcal{L}\left(X_n|X_1,\ldots,X_{n-1}\right)||\mathcal{L}\left(X_n\right)\right)\right] &= \mathbb{E}\left[\mathbf{K}\left(\mathcal{L}\left(X_n|X_{n-1}\right)||\mathcal{L}\left(X_n\right)\right)\right]\\ &= \mathbf{K}\left(\mathcal{L}\left(X_{n-1},X_n\right)||\mathcal{L}\left(X_{n-1}\right)\otimes\mathcal{L}\left(X_n\right)\right).\end{aligned}$$

Therefore

$$\begin{aligned}\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right) &= \mathbf{K}\left(\mathcal{L}\left(X_1,\ldots,X_{n-1}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_{n-1}\right)\right)\\ &\quad+\mathbf{K}\left(\mathcal{L}\left(X_{n-1},X_n\right)||\mathcal{L}\left(X_{n-1}\right)\otimes\mathcal{L}\left(X_n\right)\right),\end{aligned}$$

and we can conclude by induction.

### 3.B.5 Proof of Lemma 3.4

If $(\mathbf{X},\mathbf{H})$ a hidden Markov chain, with Lemma 3.3 we have

$$\begin{aligned}&\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right)\\ &\leq\sum_{i=2}^{n}\mathbf{K}\left(\mathcal{L}\left(X_{i-1},H_{i-1},X_i,H_i\right)||\mathcal{L}\left(X_{i-1},H_{i-1}\right)\otimes\mathcal{L}\left(X_i,H_i\right)\right).\end{aligned}$$

We need the following result. For random variables $A_1,A_2,B_1,B_2$, we have

$$\begin{aligned}&\mathbf{K}\left(\mathcal{L}\left(A_1,B_1,A_2,B_2\right)||\mathcal{L}\left(A_1,B_1\right)\otimes\mathcal{L}\left(A_2,B_2\right)\right)\\ &=\mathbf{K}\left(\mathcal{L}\left(A_1,A_2\right)||\mathcal{L}\left(A_1\right)\otimes\mathcal{L}\left(A_2\right)\right)\\ &+\mathbb{E}\left[\mathbf{K}\left(\mathcal{L}\left(B_1,B_2|A_1,A_2\right)||\mathcal{L}\left(B_1|A_1\right)\otimes\mathcal{L}\left(B_2|A_2\right)\right)\right].\end{aligned}$$

With the non-negativity of the Kullback-Leibler divergence we get

$$\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right)\leq\sum_{i=2}^{n}\mathbf{K}\left(\mathcal{L}\left(H_{i-1},H_i\right)||\mathcal{L}\left(H_{i-1}\right)\otimes\mathcal{L}\left(H_i\right)\right).$$

## 3.C Kolmogorov processes

This section gathers the proofs of Theorems 3.2, 3.3 and Lemmas 3.5, 3.6.

### 3.C.1 Proof of Theorems 3.2 and 3.3

From Proposition 6 [11], we can take $D_n(\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]) = 9\log(2|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]|)$. From Theorem 3.1 there exists a positive constant $C$ such that

$$\begin{aligned} C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\mathscr{F}_{\lambda_-,\lambda_+,M}\right) + \epsilon^2 + n^{-1}\sum_{i=1} h^2\left(P_i,\overline{P}\right) \\ + n^{-1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ + \frac{s+1}{n}\left[1+\log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|)\right]. \end{aligned}$$

Given the bounds on $\log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|)$ given by Lemma 3.6, we obtain the following inequalities.

- For $d=1$ we have $\epsilon^2 = n^{-4/5}\log^{4/5} n$ and

$$\begin{aligned} \log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|) &\leq \log(9/\eta_1) + \frac{7}{2}\log M + \overline{K}_1\epsilon^{-1/2} \\ &= \log(9/\eta_1) + \frac{9}{2}\overline{K}_1 n^{1/5}\log^{-1/5} n. \end{aligned}$$

- For $d=2$ we have $\epsilon^2 = n^{-2/3}\log^{5/3} n$ and

$$\begin{aligned} \log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|) &\leq \log\left(\frac{3^8\pi}{\eta_2^3}\right) + 9\log M + \overline{K}_2\epsilon^{-1}\log_{++}^{3/2}(1/\epsilon) \\ &\leq \log\left(\frac{3^8\pi}{\eta_2^3}\right) + \frac{28}{3}\overline{K}_2 n^{1/3}\log^{2/3} n. \end{aligned}$$

- For $d=3$ we have $\epsilon^2 = n^{-1/4}\log^{1/4} n$ and

$$\begin{aligned} \log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\delta]|) &\leq \log\left(\frac{2^7 3^{27/2}\pi^3}{\eta_3^6}\right) + \frac{33}{2}\log M + \overline{K}_3\epsilon^{-2} \\ &= \log\left(\frac{2^7 3^{27/2}\pi^3}{\eta_3^6}\right) + \frac{33}{2}\overline{K}_3 n^{1/2}\log^{-1/2} n. \end{aligned}$$

This proves the bound (3.22). Lemma 3.5 allows us to conclude the proof of Theorem 3.2.

For $d \geq 4$ we have $\epsilon^2 = n^{-\frac{2}{d+1}}\log^{d+2+\frac{2}{d+1}} n$ and

$$\begin{aligned} \log(|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|) &\leq \log C_d + \left(\overline{K}_d + 2 + \frac{1}{d} + \frac{1}{d^2}\right)\epsilon^{-(d-1)}\log^{(d+1)(d+2)/2}(\epsilon^{-1}) \\ &\leq \log C_d + \frac{1}{d+1}\left(\overline{K}_d + 2 + \frac{1}{d} + \frac{1}{d^2}\right) n^{\frac{d-1}{d+1}}\log^{\frac{2}{d+1}+d+1} n. \end{aligned}$$

Lemma 3.5 allows us to conclude the proof of Theorem 3.3.

### 3.C.2 Proof of Lemma 3.5

We have

$$I\left(\sigma(Y_t),\sigma(Y_{t+s})\right) = \mathbf{K}\left(\mathcal{L}\left(Y_t,Y_{t+s}\right)||\mathcal{L}\left(Y_t\right)\otimes\mathcal{L}\left(Y_{t+s}\right)\right) = \mathbb{E}\left[\mathbf{K}\left(\mathcal{L}\left(Y_{t+s}|Y_t\right)||\mathcal{L}\left(Y_{t+s}\right)\right)\right].$$

Since $(Y_t)_{t\geq 0}$ is stationary we have $\mathcal{L}(Y_{t+s}) = \overline{P}$. For $x \in \mathbb{R}^d$ fixed, we write

$$A_x(s) = \mathbf{K}\left(\mathcal{L}(Y_s^x)||\overline{P}\right),$$

where $Y_t^x$ is the solution of (3.17) satisfying $Y_0^x = x$. We follow the proof of Theorem 3.2.7 [78] with their notation. From (44) therein we have

$$A_x(s) \leq \mathbb{E}\left[\left(\log(Z) + U(x) + U(W_s) - 2v(W_s) - \frac{1}{2}\int_0^s [|\nabla U|^2 - \Delta U](W_t)dt\right) F\right], \tag{3.75}$$

where

- $W$ is the Brownian motion starting from $x$,
- $F$ is the density of the distribution of $X^x$ over $\mathcal{C}([0,s])$ with respect to the distribution $P$ of $W$ given by

$$F = \exp\left(U(x) - U(W_s) - \frac{1}{2}\int_0^s [|\nabla U|^2 - \Delta U](W_t)dt\right),$$

- $v$ is such that $\exp(-2v)$ is the Gaussian density of $\mathcal{L}(W_s)$ with respect to the Lebesgue measure, i.e.

$$\exp(-2v(y)) = (2\pi t)^{-d/2}\exp\left(-\frac{(x-y)^2}{2s}\right), \forall y \in \mathbb{R}^d. \tag{3.76}$$

Let us check that the right-hand side of (3.75) is finite. From (3.76), we have $-2v(y) \leq -\frac{d}{2}\log(2\pi s)$. Also

$$-\frac{1}{2}\int_0^s \left[|\nabla U|^2 - \Delta U\right](W_t)dt \leq -\frac{Cs}{2},$$

where $C$ is given by (3.18). Since $\mathbb{E}F = 1$, we get

$$A_x(s) \leq \log(Z) + U(x) - \frac{d}{2}\log(2\pi s) - \frac{Cs}{2} + \mathbb{E}\left[U(W_s)F\right].$$

We only need to consider the last term $\mathbb{E}\left[U(W_s)F\right]$. We have

$$\begin{aligned}
\mathbb{E}\left[U(W_s)F\right] &= \mathbb{E}\left[U(W_s)\exp\left(U(x) - U(W_s) - \frac{1}{2}\int_0^s [|\nabla U|^2 - \Delta U](W_t)dt\right)\right] \\
&= e^{U(x)}\mathbb{E}\left[U(W_s)\exp\left(-U(W_s) - \frac{1}{2}\int_0^s [|\nabla U|^2 - \Delta U](W_t)dt\right)\right] \\
&\leq e^{U(x)-\frac{Cs}{2}}\mathbb{E}\left[U(W_s)\exp\left(-U(W_s)\right)\right] \\
&\leq e^{U(x)-\frac{Cs}{2}}\mathbb{E}\left[U^+(W_s)\exp\left(-U(W_s)\right)\right] \\
&\leq e^{U(x)-\frac{Cs}{2}}||g||_\infty,
\end{aligned}$$

where $g$ is defined on $\mathbb{R}^+$ by $g(x) = x\exp(-x)$. We end up with

$$\begin{aligned}
A_x(s) &\leq \log(Z) + U(x) - \frac{d}{2}\log(2\pi s) - \frac{Cs}{2} + e^{U(x)-\frac{Cs}{2}}||g||_\infty \\
&\leq \log(Z) - \frac{d}{2}\log(2\pi s) - \frac{Cs}{2} + e^{U(x)}||g||_\infty\left(1 + e^{-\frac{Cs}{2}}\right).
\end{aligned} \tag{3.77}$$

Therefore, $A_x(s)$ is finite for all $s > 0$ and all $x \in \mathbb{R}^d$. From Theorem 3.1.29 and Theorem 3.2.5 of Royer [78], for all $s_0 > 0$, we have

$$A_x(s) \leq A_x(s_0)\exp\left(-2m(s - s_0)\right), \forall s > s_0. \tag{3.78}$$

Therefore with (3.77) and (3.78), we have

$$\begin{aligned} I\left(\sigma(Y_t),\sigma(Y_{t+s})\right) &= \mathbb{E}\left[A_{Y_t}(s)\right] \\ &\leq \exp\left(-2m(s-s_0)\right)\mathbb{E}\left[A_{Y_t}(s_0)\right] \\ &\leq e^{-2m(s-s_0)}\left[\log(Z)-\frac{d}{2}\log(2\pi s_0)-\frac{Cs_0}{2}+\mathbb{E}\left[e^{U(Y_t)}\right]||g||_\infty(1+e^{-\frac{Cs_0}{2}})\right] \\ &= e^{-2m(s-s_0)}\left[\log(Z)-\frac{d}{2}\log(2\pi s_0)-\frac{Cs_0}{2}+||g||_\infty(1+e^{-\frac{Cs_0}{2}})Z^{-1}\int_{\mathbb{R}^d}e^{-U(x)}dx\right] \\ &=: C(s_0)e^{-2ms}, \end{aligned}$$

for $s \geq s_0 > 0$ with $C(s_0) < \infty$ since $\int_{\mathbb{R}^d} e^{-\alpha U(x)}dx < \infty$ for all $\alpha$.

### 3.C.3 Proof of Lemma 3.6

We divide the proof in two parts, first the case $d \leq 3$ and the case $d \geq 4$ in a second time.

**Case** $d \in \{1,2,3\}$. For $\xi > 0$ and $\nu \in (0,1)$, let

$$\tilde{\mathcal{F}}_d^{\xi,\nu} = \left\{\tilde{f} \in \mathcal{F}_d : ||\overline{x}_{\tilde{f}}||_2 \leq \xi \text{ and } 1-\nu < \lambda_{\min}(\Sigma_{\tilde{f}}) \leq \lambda_{\max}(\Sigma_{\tilde{f}}) \leq 1+\nu\right\}.$$

We first state the classic bound

$$N(B_2(M), ||\cdot||_2, \epsilon) \leq \left(\frac{3M}{\epsilon}\right)^d, \tag{3.79}$$

where $B_2(M)$ is the ball of radius $M$ in $\mathbb{R}^d$ with respect to the Euclidean distance $||\cdot||_2$. Let $B_2(M)\left[\sqrt{\lambda_-}\right]$ be a $\sqrt{\lambda_-}$-net of $B_2(M)$ with respect to the Euclidean distance $||\cdot||_2$, with $\left|B_2(M)\left[\sqrt{\lambda_-}\right]\right| \leq (3M/\lambda_-)^d$. Let $\mathrm{Sym}(\lambda_-,\lambda_+)[\eta_d\lambda_-]$ be a $\eta_d\lambda_-$-net of $\mathrm{Sym}(\lambda_-,\lambda_+)$ with respect to the operator norm $||\cdot||_{op}$, with $|\mathrm{Sym}(\lambda_-,\lambda_+)[\eta_d\lambda_-]| \leq N_\Sigma(\lambda_+,\lambda_-,d,\eta_d\lambda_-)$. Let $\tilde{F}_d^{1,\eta_d}[\epsilon]$ be an $\epsilon$-net of $\tilde{F}_d^{1,\eta_d}$ with respect to the Hellinger distance. We define

$$\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon] := \left\{ (\det\Sigma)^{-1/2} g\left(\Sigma^{-1/2}(\cdot-\overline{x})\right); \begin{array}{l} \overline{x} \in B_2(M)\left[\sqrt{\lambda_-}\right], \\ \Sigma \in \mathrm{Sym}(\lambda_-,\lambda_+)[\eta_d\lambda_-], \\ g \in \tilde{F}_d^{1,\eta_d}[\epsilon] \end{array} \right\}$$

and we show it is an $\epsilon$-net of $\mathcal{F}_{\lambda_-,\lambda_+,M}$ with respect to the Hellinger distance. For $f \in \mathcal{F}_{\lambda_-,\lambda_+,M}$, there is $\Sigma$ in $\mathrm{Sym}(\lambda_-,\lambda_+)[\eta_d\lambda_-]$ and $\overline{x}$ in $B_2(M)[\sqrt{\lambda_-}]$ such that

$$||\overline{x}_f - \overline{x}||_2 \leq \sqrt{\lambda_-} \text{ and } ||\Sigma_f - \Sigma||_{op} \leq \lambda_-\eta_d.$$

We write $\tilde{f} = (\det\Sigma)^{1/2} f\left(\Sigma^{1/2}\cdot + \overline{x}\right)$. Let us check that $\tilde{f}$ belongs to $\tilde{F}_d^{1,\eta_d}$. We have

$$||\overline{x}_{\tilde{f}}||_2 = ||\Sigma^{-1/2}(\overline{x}_f - \overline{x})||_2 \leq \frac{||\overline{x}_f - \overline{x}||_2}{\sqrt{\lambda_-}} \leq 1,$$

and

$$||\Sigma_{\tilde{f}} - I||_{op} = ||\Sigma^{-1/2}\Sigma_f\Sigma^{-1/2} - I||_{op} = ||\Sigma^{-1/2}(\Sigma_f - \Sigma)\Sigma^{-1/2}||_{op} \leq \frac{||\Sigma_f - \Sigma||_{op}}{\lambda_-} \leq \eta_d.$$

Therefore $\tilde{f} \in \tilde{F}_d^{1,\eta_d}$ and there is $g \in \tilde{F}_d^{1,\eta_d}[\epsilon]$ such that $h\left(\tilde{f},g\right) \leq \epsilon$. Since the Hellinger distance is invariant by translation and scaling, we have

$$h\left(f,(\det\Sigma)^{-1/2}g\left(\Sigma^{-1/2}(\cdot-\mu)\right)\right) = h\left(\tilde{f},g\right) \leq \epsilon,$$

which proves that $\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ is an $\epsilon$-net of $\mathcal{F}_{\lambda_-,\lambda_+,M}$. Therefore

$$|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]| \leq \left(\frac{3M}{\sqrt{\lambda_-}}\right)^d \times N_\Sigma(\lambda_+,\lambda_-,d,\eta_d\lambda_-) \times |\tilde{F}_d^{1,\eta_d}[\epsilon]|.$$

We need to bound the different entropy numbers now. For a metric space $(\mathscr{A},d)$ and $\epsilon > 0$, we denote by $N(\epsilon,\mathscr{A},d)$ the minimal number of balls of radius $\epsilon$, with respect to $d$, to cover $\mathscr{A}$.

The next result provides a bound on the entropy for the class of covariance matrices we are considering. Let $||\cdot||_{op}$ denote the operator norm on square matrices induced by the Euclidean distance. For matrices with real-valued eigenvalues, it is equivalent to the largest absolute value of its eigenvalues.

**Lemma 3.9.** *We have*

$$N\left(\epsilon, Sym(\lambda_-,\lambda_+), ||\cdot||_{op}\right) \leq \begin{cases} \frac{3(\lambda_+-\lambda_-)}{\epsilon} & \text{for } d=1, \\ \left(\frac{9}{\epsilon}\right)^3 (\lambda_+-\lambda_-)^2\lambda_+\pi & \text{for } d=2, \\ 2\left(\frac{2\cdot 3^{5/4}\sqrt{\lambda_+(\lambda_+-\lambda_-)\pi}}{\epsilon}\right)^6 & \text{for } d=3. \end{cases} \tag{3.80}$$

*In higher dimensions, we have*

$$N\left(\epsilon, Sym(\lambda_-,\lambda_+), ||\cdot||_{op}\right) \leq C\left(\frac{3}{4}\right)^d \frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}}(2\lambda_+)^{d(d-1)/2}(\lambda_+-\lambda_-)^d \\ \times (d+1)^{d(d+1)/2} d^{(d-1)(d+2)/2}(d-1)^{(d-1)/2}\epsilon^{-d(d+1)/2},$$

*with* $C = \frac{e^{1/2}}{3^{1/2}2^3}$.

Theorem 4 [58] gives a bound on $|\tilde{F}_d^{1,\eta_d}[\epsilon]|$ which allows us to conclude the proof of Theorem 3.2.

**Case** $d \geq 4$. We use Theorem 3 of Kur *et al.* [60]. We follow some of their notation. Let $d \geq 4$. There exist positive constants $\xi_d$ and $\overline{K}_d$ such that

$$\log N(\epsilon,\mathscr{F}_{d,\tilde{I}},h) \leq \overline{K}_d\epsilon^{-(d-1)}\log_{++}(\epsilon^{-1})^{(d+1)(d+2)/2},$$

where $\mathscr{F}_{d,\tilde{I}}$ is the set of distributions associated to

$$\mathcal{F}_{d,\tilde{I}} = \left\{\tilde{f} \in \mathcal{F}_d : ||\overline{x}_{\tilde{f}}||_2 \leq \xi_d \text{ and } 1/2 < \lambda_{\min}(\Sigma_{\tilde{f}}) \leq \lambda_{\max}(\Sigma_{\tilde{f}}) \leq 2\right\}.$$

Let $\mathcal{F}_{d,\tilde{I}}[\epsilon]$ be a set of probability densities with respect to the Lebesgue measure such that $\mathscr{F}_{d,\tilde{I}}[\epsilon] = \{f(x)dx; f \in \mathcal{F}_{d,\tilde{I}}\}$ is an $\epsilon$-net of $\mathscr{F}_{d,\tilde{I}}$ with respect to the Hellinger distance and

$$\log|\mathscr{F}_{d,\tilde{I}}[\epsilon]| \leq \overline{K}_d\epsilon^{-(d-1)}\log_{++}(\epsilon^{-1})^{(d+1)(d+2)/2}.$$

Let $B_2(M)\left[\xi_d\sqrt{\lambda_-}\right]$ be a $\xi_d\sqrt{\lambda_-}$-net of $B_2(M)$ with respect to the Euclidean distance $||\cdot||_2$, with $\left|B_2(M)\left[\xi_d\sqrt{\lambda_-}\right]\right| \leq (3M/\xi_d\sqrt{\lambda_-})^d$. Let $\mathrm{Sym}(\lambda_-,\lambda_+)[\lambda_-/3]$ be a $\lambda_-/3$-net of $\mathrm{Sym}(\lambda_-,\lambda_+)$ with respect to the operator norm $||\cdot||_{op}$, with $|\mathrm{Sym}(\lambda_-,\lambda_+)[\lambda_-/3]| \leq N_\Sigma(\lambda_+,\lambda_-)$. We define

$$\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon] := \left\{ (\det\Sigma)^{-1/2} g\left(\Sigma^{-1/2}(\cdot-\overline{x})\right); \begin{array}{l} \overline{x} \in B_2(M)\left[\xi_d\sqrt{\lambda_-}\right], \\ \Sigma \in \mathrm{Sym}(\lambda_-,\lambda_+)[\lambda_-/3], \\ g \in \mathcal{F}_{d,\tilde{I}}[\epsilon] \end{array} \right\}$$

and we show that $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon] = \{f(x)dx; f \in \mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]\}$ is an $\epsilon$-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ with respect to the Hellinger distance. For $f \in \mathcal{F}_{\lambda_-,\lambda_+,M}$, there is $\Sigma$ in $\mathrm{Sym}(\lambda_-,\lambda_+)[\lambda_-/3]$ and $\overline{x}$ in $B_2(M)[\xi_d\sqrt{\lambda_-}]$ such that

$$||\overline{x}_f - \overline{x}||_2 \leq \xi_d\sqrt{\lambda_-} \text{ and } ||\Sigma_f - \Sigma||_{op} \leq \lambda_-/3.$$

We write $\tilde{f} = (\det\Sigma)^{1/2} f\left(\Sigma^{1/2}\cdot + \overline{x}\right)$. Let us check that $\tilde{f}$ belongs to $\mathcal{F}_{d,\tilde{I}}$. We have

$$||\overline{x}_{\tilde{f}}||_2 = ||\Sigma^{-1/2}(\overline{x}_f - \overline{x})||_2 \leq \frac{||\overline{x}_f - \overline{x}||_2}{\sqrt{\lambda_-}} \leq \xi_d,$$

and

$$||\Sigma_{\tilde{f}} - I|| = ||\Sigma^{-1/2}\Sigma_f\Sigma^{-1/2} - I|| = ||\Sigma^{-1/2}(\Sigma_f - \Sigma)\Sigma^{-1/2}|| \leq \frac{||\Sigma_f - \Sigma||}{\lambda_-} \leq 1/3.$$

Hence

$$\lambda_{\min}(\Sigma_{\tilde{f}}) \geq 2/3 > 1/2 \text{ and } \lambda_{\max}(\Sigma_{\tilde{f}}) \leq 4/3 < 2.$$

Therefore we have $\tilde{f} \in \mathcal{F}_{d,\tilde{I}}$ and there is $g \in \mathcal{F}_{d,\tilde{I}}[\epsilon]$ such that $h\left(\tilde{f}(x)dx, gx)dx\right) \leq \epsilon$. Since the Hellinger distance is invariant by translation and scaling, we have

$$h\left(f(x)dx, (\det\Sigma)^{-1/2} g\left(\Sigma^{-1/2}(x-\overline{x})\right)dx\right) = h\left(\tilde{f}(x)dx, g(x)dx\right) \leq \epsilon,$$

which proves that $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ is an $\epsilon$-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$. Therefore

$$|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]| \leq \left(\frac{3M}{\xi_d\sqrt{\lambda_-}}\right)^d \times N_\Sigma(\lambda_+,\lambda_-,d) \times |\mathscr{F}_{d,\tilde{I}}[\epsilon]|.$$

With Lemma 3.9 we get

$$\begin{aligned}
|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]| &\leq C\left(\frac{3M}{\xi_d\sqrt{\lambda_-}}\right)^d \left(\frac{3}{4}\right)^d \frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}} (2\lambda_+)^{d(d-1)/2} (\lambda_+ - \lambda_-)^d \\
&\quad \times (d+1)^{d(d+1)/2} d^{(d-1)(d+2)/2} (d-1)^{(d-1)/2} \left(\frac{\lambda_-}{3}\right)^{-d(d+1)/2} \\
&\quad \times \exp\left(\overline{K}_d \epsilon^{-(d-1)} \log(\epsilon^{-1})^{(d+1)(d+2)/2}\right) \\
&\leq C_d \frac{\lambda_+^{d(d-1)/2} M^d (\lambda_+ - \lambda_-)^d}{\lambda_-^{d(d+1)/2}} \exp\left(\overline{K}_d \epsilon^{-(d-1)} \log(\epsilon^{-1})^{(d+1)(d+2)/2}\right).
\end{aligned}$$

**Proof of Lemma 3.9**

For $d = 1$, we have $\mathrm{Sym}(\lambda_-,\lambda_+) = [\lambda_-,\lambda_+]$. The result follows from classical entropy bounds. Otherwise, every real valued symmetric matrix $\Sigma$ can be written as $\Sigma = UDU^T$ where $D$ is the diagonal matrix containing the real eigenvalues of $\Sigma$ and $U$ is an orthonormal matrix. For $\Sigma_1 = U_1 \mathrm{diag}(\lambda_{1,1},\ldots,\lambda_{d,1}) U_1^T$ and $\Sigma_2 = U_2 \mathrm{diag}(\lambda_{1,2},\ldots,\lambda_{d,2}) U_2^T$ we have

$$\begin{aligned}
||\Sigma_1 - \Sigma_2|| &\leq ||U_1(D_1 - D_2)U_1^T|| + ||(U_1 - U_2)D_2U_1^T|| + ||U_2D_2(U_1 - U_2)^T|| \\
&\leq ||D_1 - D_2|| + 2\lambda_+||U_1 - U_2|| \\
&= \max_{1\leq i\leq d} |\lambda_{i,1} - \lambda_{i,2}| + 2\lambda_+||U_1 - U_2||.
\end{aligned}$$

Therefore

$$N\left(\mathrm{Sym}(\lambda_-,\lambda_+), ||\cdot||, \epsilon\right) \leq N\left(B((\lambda_+ - \lambda_-)/2), ||\cdot||_\infty, \epsilon_1\right) \times N\left(\mathrm{ON}(d), ||\cdot||, \epsilon_2\right)$$

with $\epsilon = \epsilon_1 + 2\lambda_+\epsilon_2$. We have the classic bound

$$N\left(B((\lambda_+ - \lambda_-)/2), ||\cdot||_\infty, \epsilon_1\right) \leq \left(3\frac{\lambda_+ - \lambda_-}{2\epsilon_1}\right)^d.$$

- For $d = 2$, the orthonormal matrices are of the form

$$U_{\alpha,\theta} = \begin{pmatrix} \cos(\theta) & -\alpha\sin(\theta) \\ \sin(\theta) & \alpha\cos(\theta) \end{pmatrix}, \theta \in [0,2\pi], \alpha \in \{-1,1\}.$$

  We have

$$||U_{\alpha,\theta} - U_{\alpha,\theta'}||^2 = 2[1 - \cos(\theta - \theta')] \leq (\theta - \theta')^2,$$

  and therefore

$$N\left(\mathrm{ON}(2), ||\cdot||, \epsilon\right) \leq 2\frac{3\pi}{\epsilon} = 6\pi/\epsilon,$$

  where the factor 2 comes from the presence of $\epsilon$ for positively and negatively oriented basis. We obtain the final result for $\epsilon_1 = 2\epsilon/3$ and $\epsilon_2 = \epsilon/6\lambda_+$.

- We proceed similarly for $d = 3$. Every orthonormal basis in dimension 3 can be written in the form

$$U_{\epsilon,\theta,\beta,\gamma} := \begin{pmatrix} \cos\theta & \cos\gamma\sin\theta & -\epsilon\sin\gamma\sin\theta \\ \sin\theta\cos\beta & -\cos\gamma\cos\theta\cos\beta + \sin\gamma\sin\beta & \epsilon(\sin\gamma\cos\theta\cos\beta + \cos\gamma\sin\beta) \\ \sin\theta\sin\beta & -\cos\gamma\cos\theta\sin\beta - \sin\gamma\cos\beta & \epsilon(\sin\gamma\cos\theta\sin\beta - \cos\gamma\cos\beta) \end{pmatrix},$$

  $\theta \in [0,2\pi], \beta \in [0,2\pi], \gamma \in [0,2\pi], \epsilon \in \{-1,1\}$. As before, one can check that we have

$$\begin{aligned} ||U_{\epsilon,\theta,\beta,\gamma} - U_{\epsilon,\theta',\beta,\gamma}|| &\leq |\theta - \theta'|^2 \\ ||U_{\epsilon,\theta,\beta,\gamma} - U_{\epsilon,\theta,\beta',\gamma}|| &\leq |\beta - \beta'|^2 \\ ||U_{\epsilon,\theta,\beta,\gamma} - U_{\epsilon,\theta,\beta,\gamma'}|| &\leq |\theta - \theta'|^2. \end{aligned}$$

  Therefore we have

$$N\left(\mathrm{ON}(3), ||\cdot||, \epsilon\right) \leq \left(N\left([0,2\pi], |\cdot|, \epsilon/\sqrt{3}\right)\right)^3 \leq 2\left(\frac{3\sqrt{3}\pi}{\epsilon}\right)^3, \tag{3.81}$$

  where the factor 2 comes from the presence of $\epsilon$ for positively and negatively oriented basis. We obtain the final result for $\epsilon_1 = \epsilon/2$ and $\epsilon_2 = \epsilon/4\lambda_+$.

- For higher dimensions, we have the following lemma.

  **Lemma 3.10.** *For $d \geq 3$, we can build an $\epsilon$-net $ON(d)[\epsilon]$ of $ON(d)$ with respect to the operator norm such that*

$$|ON(d)[\epsilon]| \leq C\frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}} d^{(d-1)(d+2)/2}(d-1)^{(d-1)(d+1)/2}\epsilon^{-d(d-1)/2}, \forall d \geq 1,$$

  *with $C = \frac{e^{1/2}}{3^{1/2}2^3}$.*

  We obtain the final bound with $\epsilon_1 = \frac{2\epsilon}{d+1}$ and $\epsilon_2 = \frac{\epsilon}{2\lambda_+}\frac{d-1}{d+1}$.

**Proof of Lemma 3.10**

We prove this by induction. From (3.81) we have the desired inequality for $d = 3$ with $C_3 = \frac{e^{1/2}}{3^{1/2}2^3}$. Let $\epsilon$ be in $(0,1]$ and $d \geq 3$. Let us now assume that for $\lambda_1 > 0$ we have a $\lambda_1$-net $ON(d)[\lambda_1]$ with

$$|ON(d)[\lambda_1]| \leq C\frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}}d^{(d-1)(d+2)/2}(d-1)^{(d-1)(d+1)/2}\lambda_1^{-d(d-1)/2}.$$

Let $U \in \mathbb{R}^{d+1}$ be a unitary vector, i.e. $U_1^2 + \cdots + U_{d+1}^2 = 1$. There is $\theta \in [0,2\pi]^d$ such that $U = f(\theta)$ with

$$U_i = f_i(\theta) := \cos\theta_i \prod_{j\leq i} \sin\theta_j,$$

with the convention $\theta_{d+1} = 0$ and that a product over an empty set of indices is equal to 1. We define applications $a_1, \ldots, a_d, a_{d+1}$ by $a_1 = id$ and

$$a_i(\theta) = \left(\theta_1 + \frac{\pi}{2}, \ldots, \theta_{i-1} + \frac{\pi}{2}, \theta_i, \ldots, \theta_d\right), \forall i \in \{2, \ldots, d+1\}.$$

One can check that the set of vectors $A_1(\theta), \ldots, A_{d+1}(\theta) \in \mathbb{R}^{d+1}$, given by $A_i(\theta) = f(a_i(\theta))$ for $i$ in $\{1,2,\ldots,d+1\}$, is an orthonormal basis of $\mathbb{R}^d$. We take $n_j = \left\lceil \frac{\sqrt{d+1-j}}{\lambda_2} \right\rceil, \forall j \in \{1,2,\ldots,d\}$ and we take

$$\mathscr{A}_{d+1}[\lambda_2] := \{A(\psi_{i_1,\ldots,i_d}); i_j \in \{1,2,\ldots,n_j\}, j \in \{1,2,\ldots,d\}\} \subset ON(d+1),$$

with

$$\psi_{i_1,\ldots,i_d} = \left(\frac{\pi(2i_j - 1)}{n_j}\right)_{1\leq j\leq d}.$$

**Lemma 3.11.** *The set*

$$O[\lambda_1,\lambda_2] := \left\{A\begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix}; A \in \mathscr{A}_{d+1}[\lambda_2], B \in ON(d)[\lambda_1]\right\},$$

*is a $\lambda_1 + \sqrt{d}\pi\lambda_2$-net of $ON(d+1)$ with respect to the operator norm.*

One can easily check that we have the following bound

$$|\mathscr{A}_{d+1}[\lambda_2]| \leq \left(\frac{2}{\lambda_2}\right)^d \sqrt{d!}.$$

Therefore, we have

$$\begin{aligned} |O[\lambda_1,\lambda_2]| &= |ON(d)[\lambda_1]| \times |\mathscr{A}_{d+1}[\lambda_2]| \\ &\leq C\frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}}d^{(d-1)(d+2)/2}(d-1)^{(d-1)(d+1)/2}\lambda_1^{-d(d-1)/2} \times \left(\frac{2}{\lambda_2}\right)^d \sqrt{d!}. \end{aligned}$$

For $\lambda_1 = \epsilon\frac{d-1}{d+1}$ and $\lambda_2 = \epsilon\frac{2}{\sqrt{d}\pi(d+1)}$, we get

$$\begin{aligned} &|O[\lambda_1,\lambda_2]| \\ &\leq C\frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}}d^{(d-1)(d+2)/2}(d-1)^{(d-1)(d+1)/2}\left(\frac{d+1}{d-1}\right)^{d(d+1)/2}\epsilon^{-d(d-1)/2} \\ &\times \sqrt{d!}\left(\sqrt{d}\pi(d+1)\right)^d\epsilon^{-d} \\ &= C(d-1)^{-(d+1)/2}d^{-1}\sqrt{d!}e^{(d-1)/2}\frac{\pi^{d(d+1)/2}}{e^{(d-1)(d-2)/2}}(d+1)^{d(d+3)/2}d^{d(d+2)/2}\epsilon^{-d(d+1)/2}. \end{aligned}$$

We use the bound $n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n}}$ and we get

$$\begin{aligned}
&|O[\lambda_1,\lambda_2]| \\
&\leq C(d-1)^{-(d+1)/2} d^{-1/2} \sqrt{(d-1)!} e^{(d-1)/2} \frac{\pi^{d(d+1)/2}}{e^{(d-1)(d-2)/2}} (d+1)^{d(d+3)/2} d^{d(d+2)/2} \epsilon^{-d(d+1)/2} \\
&\leq C(d-1)^{-3/4} d^{-1/2} (2\pi)^{1/4} e^{\frac{1}{24(d-1)}} \frac{\pi^{d(d+1)/2}}{e^{(d-1)(d-2)/2}} (d+1)^{d(d+3)/2} d^{d(d+2)/2} \epsilon^{-d(d+1)/2}.
\end{aligned}$$

We have

$$(d-1)^{-3/4} d^{-1/2} (2\pi)^{1/4} e^{\frac{1}{24(d-1)}} \leq 1$$

for all $d \geq 3$. Therefore, we satisfy the desired property for $d+1$ with $ON[\epsilon] = O[\lambda_1,\lambda_2]$.

**Proof of Lemma 3.11**

Let $C = (C_1 \dots C_{d+1})$ be in $ON(d+1)$. There is $\theta$ in $[0,2\pi]^d$ such that $C_1 = A_1(\theta)$. Let $B$ be the matrix in $ON(d)$ given by

$$A(\theta)^T C = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix}.$$

There exists $\psi_{i_1,\dots,i_d}$ such that

$$\left| \theta_i - \frac{\pi(2i_j - 1)}{n_j} \right| \leq \frac{\pi}{n_j} \leq \frac{\pi\lambda_2}{\sqrt{d+1-j}}, \forall j \in \{1,\dots,d\}.$$

**Lemma 3.12.** *We have*

$$||A(\theta) - A(\theta+h)||_{op} \leq \sqrt{\sum_{k=0}^{d-1} (d-k) h_{k+1}^2}.$$

Therefore we have

$$||A(\theta) - A(\psi_{i_1,\dots,i_d})||_{op} \leq d^{1/2} \pi \lambda_2.$$

There exists $B'$ in $ON(d)[\lambda_1]$ such that $||B - B'||_{op} \leq \lambda_1$. We define $C' \in ON(d+1)$ by

$$C' = A(\psi_{i_1,\dots,i_d}) \begin{pmatrix} 1 & 0 \\ 0 & B' \end{pmatrix} \in ON[\lambda_1,\lambda_2].$$

Then we have

$$\begin{aligned}
||C - C'||_{op} &\leq \left|\left| A(\theta) \begin{pmatrix} 0 & 0 \\ 0 & B - B' \end{pmatrix} \right|\right|_{op} + \left|\left| (A(\theta) - A(\psi_{i_1,\dots,i_d})) \begin{pmatrix} 1 & 0 \\ 0 & B' \end{pmatrix} \right|\right|_{op} \\
&\leq ||B - B'||_{op} + ||A(\theta) - A(\psi_{i_1,\dots,i_d})||_{op} \\
&\leq \lambda_1 + d^{1/2} \pi \lambda_2.
\end{aligned}$$

**Proof of Lemma 3.12**

For $\theta \in \mathbb{R}^d$ and $h \in \mathbb{R}^d$, we define $U_0 = f(\theta)$ and

$$U_i = f(\theta_1 + h_1, \dots, \theta_i + h_i, \theta_{i+1}, \dots, \theta_d), i \in \{1,\dots,d\}.$$

Similarly, we write $A^{(i)} = A(\theta^{(h,i)})$ with

$$\theta^{(h,i)} = (\theta_1 + h_1, \dots, \theta_i + h_i, \theta_{i+1}, \dots, \theta_d),$$

for $i \in \{0,1,\ldots,d\}$ and $j \in \{1,\ldots,d+1\}$. It implies $A_1^{(0)} = U_0$ and $A_1^{(d)} = U_d$. We have

$$\begin{aligned} A_{ij}^{(k)} &= f_i(a_j(\theta^{(h,k)})) = \cos\left(a_j(\theta^{(h,k)})\right) \prod_{l\leq i} \sin\left(a_j(\theta^{(h,k)})\right) \\ &= \cos\left(\theta_i + \mathbb{1}_{i<j}\frac{\pi}{2} + \mathbb{1}_{l\leq i}h_i\right) \prod_{l\leq i} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + \mathbb{1}_{l\leq k}h_l\right), \end{aligned}$$

and therefore

$$A_{ij}^{(k+1)} - A_{ij}^{(k)} = \begin{cases} 0 \text{ if } i \leq k \\ \prod\limits_{l\leq k} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right) \\ \times \left[\cos\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + h_{k+1}\right) - \cos\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2}\right)\right] \text{ if } i = k+1 \\ \prod\limits_{\substack{l<i \\ l\neq k+1}} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + \mathbb{1}_{l\leq k}\right) \times \cos\left(\theta_i + \mathbb{1}_{i<j}\frac{\pi}{2}\right) \\ \times \left[\sin\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + h_{k+1}\right) - \sin\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2}\right)\right] \text{ if } i > k+1, \end{cases}$$

$$= 2\sin\left(\frac{h_{k+1}}{2}\right) \prod_{l\leq k} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right)$$

$$\times \begin{cases} 0 \text{ if } i \leq k \\ -\sin\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + \frac{h_{k+1}}{2}\right) \text{ if } i = k+1 \\ \cos\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + \frac{h_{k+1}}{2}\right) \prod\limits_{k+1<l<i} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2}\right) \\ \times \cos\left(\theta_i + \mathbb{1}_{i<j}\frac{\pi}{2}\right) \text{ if } i > k+1. \end{cases}$$

We have ($k+1 \leq d$, $k \geq 0$)

$$\begin{aligned} &||A^{(k+1)} - A^{(k)}||_F^2 = \sum_{i,j} \left(A_{ij}^{(k+1)} - A_{ij}^{(k)}\right)^2 \\ &= 4\sin^2\left(\frac{h_{k+1}}{2}\right) \sum_{1\leq j\leq d+1} \prod_{l\leq k} \sin^2\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right) \left[\sin^2\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + \frac{h_{k+1}}{2}\right)\right. \\ &\left. + \cos^2\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + \frac{h_{k+1}}{2}\right) \sum_{i=k+2}^{d+1} \prod_{k+1<l<i} \sin^2\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2}\right) \cos^2\left(\theta_i + \mathbb{1}_{i<j}\frac{\pi}{2}\right)\right] \\ &= 4\sin^2\left(\frac{h_{k+1}}{2}\right) \sum_{1\leq j\leq d+1} \prod_{l\leq k} \sin^2\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right) \\ &= 4\sin^2\left(\frac{h_{k+1}}{2}\right) \left[(d+1-k) \prod_{l\leq k} \sin^2\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right)\right. \\ &\left. + \sum_{1\leq j\leq k} \prod_{l\leq k} \sin^2\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right)\right] \\ &\leq 4\sin^2\left(\frac{h_{k+1}}{2}\right) \left[(d+1-k) \prod_{l\leq k} \cos^2\left(\theta_l + h_l\right) + 1 - \prod_{l\leq k} \cos^2\left(\theta_l + h_l\right)\right] \\ &\leq 4\sin^2\left(\frac{h_{k+1}}{2}\right) (d-k) \prod_{l\leq k} \cos^2\left(\theta_l + h_l\right) \\ &\leq (d-k)h_{k+1}^2. \end{aligned}$$

Finally, with $||\cdot||_{op} \leq ||\cdot||_F$ we get

$$\begin{aligned}||A^{(d)} - A^{(0)}||_{op} &\leq \sum_{k=0}^{d-1} ||(A^{(k+1)} - A^{(k)})^T||_{op} \\ &\leq \sum_{k=0}^{d-1} (d-k)h_{k+1}^2.\end{aligned}$$

# 3.D Hidden Markov models

This section gathers the proof of Theorems 3.4, 3.5, 3.6, 3.9, 3.10, Corollary 3.3 and Proposition 3.1, 3.2, 3.3.

## 3.D.1 Proof of Theorem 3.4

The next result is proven in Section 3.D.1 and gives a bound on the $\rho$-dimension function.

**Proposition 3.5.** *Under Assumption 3.3 and with $\delta(s)$ given by (3.38), we can take*

$$D_{n(s,1)}\left(\mathscr{M}_{\delta(s)}\right) = CL\overline{V}\left[1 + \log\left(\frac{Kn(s,1)}{\overline{V} \wedge n(s,1)}\right)\right],$$

*with $C = 3930$.*

With Theorem 3.1 we have

$$\begin{aligned}C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \mathscr{M}_\delta\right) &+ n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) \\ &+ n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind}\right) + (s+1)L\overline{V}\frac{\log n}{n},\end{aligned}$$

for some positive constant $C$. The following result is proven in Proposition 3.D.1 and tells us how well $\mathscr{M}_\delta$ approximates $\mathscr{M}$.

**Proposition 3.6.** *For $K \geq 2$, $w$,$v$ in $\mathcal{W}_K$, $Q$,$R$ in $\mathcal{T}_K$ and probability distributions $F_1, \ldots, F_K$,$G_1, \ldots, G_K$ on $(\mathscr{Y}, \mathcal{Y})$, we have*

$$\begin{aligned}h^2\left(P_{w,Q,F}, P_{v,R,G}\right) \leq h^2(w,v) &+ (L-1)\max_{k\in[K]} h^2\left(Q_{k\cdot}, R_{k\cdot}\right) \\ &+ L\max_{k\in[K]} h^2\left(F_k, G_k\right).\end{aligned}$$

With Proposition 3.6 and inequality (B.5) in Lecestre [62] we have

$$h^2\left(P, \mathscr{M}_\delta\right) \leq (K-1)L\delta + L\epsilon^2, \forall P \in \mathscr{M}. \tag{3.82}$$

With the choice of $\delta$ given in (3.38) we get

$$\begin{aligned}C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \mathscr{M}\right) &+ n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) + n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind}\right) \\ &+ L\epsilon^2 + (s+1)L\overline{V}\frac{\log n}{n},\end{aligned}$$

for some positive constant $C$. We now turn to the second bound in Theorem 3.4. The next result is proven later in Section 3.D.1.

**Lemma 3.13.** *Under Assumption 3.2, there are positive constants $C(Q^*)$ and $r(Q^*)$ that only depend on $Q^*$ such that*

$$n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\leq C(Q^*)e^{-r(Q^*)s},\forall s\geq L-1,\forall b\in[s+1],$$

*and $h^2\left(P^*,P_i\right)\leq C(Q^*)e^{-r(Q^*)i}$ for all $i\in[n]$.*

In this situation, for $\overline{P}=P^*$ and $s\geq L-1$ we have

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right]\leq h^2\left(\overline{P},\mathscr{M}\right)+\frac{C(Q^*)}{n(e^{r(Q^*)}-1)}+C(Q^*)e^{-r(Q^*)s}$$
$$+L\epsilon^2+(s+1)L\overline{V}\frac{\log n}{n},$$

for some positive constant. The condition on $s$ leads to the desired inequality.

**Proof of Proposition 3.5**

From Proposition A.1. [62], we have

$$D^{\mathscr{H}_{\delta(s)}}\left(\bigotimes_{i=1}^{n(s,b)}P_i,Q^{\otimes n(s,b)}\right)\leq 545.3\overline{V}\left[5.82+\log\left(\frac{(K^L+1)^2}{\delta(s)^L}\right)+\log_+\left(\frac{n(s,b)}{\overline{V}}\right)\right].$$

- If $\overline{V}\leq n(s,1)(K-1)/K$, we have

$$\begin{aligned}\log\left(\frac{(K^L+1)^2}{\delta(s,b)^L}\right)+\log_+\left(\frac{n(s,b)}{\overline{V}}\right)&\leq\log\left(\frac{(K^L+1)^2n(s,1)^L(K-1)^L}{\overline{V}^L}\frac{n(s,1)}{\overline{V}}\right)\\&=\log\left(\frac{(K^L+1)^2(K-1)^L}{K^{L+1}}\right)+\log\left(\frac{K^{L+1}n(s,1)^{L+1}}{\overline{V}^{L+1}}\right)\\&=\log\left(\frac{(K^L+1)^2(K^2-1)^L}{K^{L+1}(K+1)^L}\right)+(L+1)\log\left(\frac{Kn(s,1)}{\overline{V}}\right).\end{aligned}$$

  One can check that for $L\geq 2$, we have $\frac{(K^L+1)^2(K^2-1)^L}{K^{L+1}(K+1)^L}\leq K^{2L-1}$ for all $K\geq 1$. Therefore,

$$\begin{aligned}\log\left(\frac{(K^L+1)^2}{\delta(s)^L}\right)+\log_+\left(\frac{n(s,b)}{\overline{V}}\right)&\leq(2L-1)\log K+(L+1)\log\left(\frac{Kn(s,1)}{\overline{V}}\right)\\&\leq 3L\log\left(\frac{KN}{\overline{V}}\right)=3L\log\left(\frac{KN}{\overline{V}\wedge N}\right).\end{aligned}$$

- Otherwise $\overline{V}>n(s,1)(K-1)/K$ and $\log\left(\frac{Kn(s,1)}{\overline{V}\wedge n(s,1)}\right)=\log K$. We have

$$\begin{aligned}\log\left(\frac{(K^L+1)^2}{\delta(s)^L}\right)+\log_+\left(\frac{n(s,b)}{\overline{V}}\right)&\leq\log\left(\frac{(K^L+1)^2K^Ln(s,1)}{\overline{V}}\right)\\&=\log\left(\frac{Kn(s,1)}{\overline{V}}\right)+(L-1)\log K+2\log\left(1+K^L\right)\\&\leq 3L\log\left(\frac{Kn(s,1)}{\overline{V}\wedge n(s,1)}\right)+2\log(1+K^{-L})\\&\leq 2\log 2+3L\log\left(\frac{Kn(s,1)}{\overline{V}\wedge n(s,1)}\right).\end{aligned}$$

**Proof of Proposition 3.6**

With Lemma B.3 [62], we have

$$h\left(P_{w,Q,F},P_{v,R,G}\right) \leq h\left(wQ^{\bigcirc L},vR^{\bigcirc L}\right) + \max_{k_1,\ldots,k_L\in[K]^L} h\left(\bigotimes_{l=1}^{L} F_{k_l},\bigotimes_{l=1}^{L} G_{k_l}\right),$$

with

$$wQ^{\bigcirc L}(k_1,\ldots,k_L) = w_{k_1}Q_{k_1,k_2}\ldots Q_{k_{L-1},k_L},\forall k_1,\ldots,k_L \in [K]. \tag{3.83}$$

Let $\rho$ denote the Hellinger affinity defined by $\rho = 1-h^2$ For $\rho_- = \min_{k\in[K]}\rho\left(Q_{k,\cdot},R_{k,\cdot}\right)$, we have

$$\begin{aligned}
h^2\left(wQ^{\bigcirc L},vR^{\bigcirc L}\right) &= 1-\rho\left(wQ^{\bigcirc L},vR^{\bigcirc L}\right)\\
&= 1-\sum_{k_1,\ldots,k_L}\sqrt{w_{k_1}v_{k_1}Q_{k_1,k_2}R_{k_1,k_2}\ldots Q_{k_{L-1},k_L}R_{k_{L-1},k_L}}\\
&= 1-\sum_{k_1,\ldots,k_{L-1}}\sqrt{w_{k_1}v_{k_1}Q_{k_1,k_2}R_{k_1,k_2}\ldots Q_{k_{L-2},k_{L-1}}R_{k_{L-2},k_{L-1}}}\rho\left(Q_{k_{L-1},\cdot},R_{k_{L-1},\cdot}\right)\\
&\leq 1-\rho_-\sum_{k_1,\ldots,k_{L-1}}\sqrt{w_{k_1}v_{k_1}Q_{k_1,k_2}R_{k_1,k_2}\ldots Q_{k_{L-2},k_{L-1}}R_{k_{L-2},k_{L-1}}}.
\end{aligned}$$

By induction we get

$$h^2\left(wQ^{\bigcirc L},vR^{\bigcirc L}\right) \leq 1-\rho_-^{L-1}\rho\left(w,v\right) \leq h^2(w,v) + (L-1)\max_{k\in[K]} h^2\left(Q_{k,\cdot},R_{k,\cdot}\right).$$

We also have

$$\begin{aligned}
h^2\left(\bigotimes_{l=1}^{L} F_{k_l},\bigotimes_{l=1}^{L} G_{k_l}\right) &= 1-\rho\left(\bigotimes_{l=1}^{L} F_{k_l},\bigotimes_{l=1}^{L} G_{k_l}\right)\\
&= 1-\prod_{l=1}^{L}\rho\left(F_{k_l},G_{k_l}\right) \leq \sum_{l=1}^{L} h^2\left(F_{k_l},G_{k_l}\right),
\end{aligned}$$

which allows us to conclude the proof.

**Proof of Lemma 3.13**

Let $s$ be not smaller than $L-1$ and $b$ be in $[s+1]$. Since $(Y_i,H_i)_{1\leq i\leq N}$ is a hidden Markov model, we have that

$$\left(X_i^{(s,b)},H_i^{(L,s,b)}\right)_{1\leq i\leq n}$$

is also a hidden Markov model, with

$$X_i^{(s,b)} = X_{b+(i-1)(s+1)} \text{ and } H_i^{(L,s,b)} = \left(H_{b+(i-1)(s+1)},\ldots,H_{b+(i-1)(s+1)+L-1}\right).$$

From Lemma 3.4, we have

$$\mathbf{K}\left(\mathbf{P}_{s,b}^{*}||\mathbf{P}_{s,b}^{ind}\right) \leq \sum_{i=1}^{n(s,b)-1}\mathbf{K}\left(\mathcal{L}\left(H_i^{(L,s,b)},H_{i+1}^{(L,s,b)}\right)||\mathcal{L}\left(H_i^{(L,s,b)}\right)\otimes\mathcal{L}\left(H_{i+1}^{(L,s,b)}\right)\right).$$

We can use the following result to bound the terms in the sum on the right-hand side of the inequality.

**Lemma 3.14.** *Let $A$ and $B$ be random variables taking values in the finite sets $\mathscr{A}$ and $\mathscr{B}$ respectively. We have*

$$\mathbf{K}\left(\mathcal{L}(A,B)||\mathcal{L}(A)\otimes\mathcal{L}(B)\right)\leq 2\sum_{a\in\mathscr{A}} d_{TV}\left(\mathcal{L}(B|A=a),\mathcal{L}(B)\right).$$

For $k_1,\ldots,k_{2L}\in[K^*]$, we have

$$\begin{aligned}&\mathbb{P}\left(H_{i+1}^{(L,s,b)}=(k_{L+1},\ldots,k_{2L})|H_i^{(L,s,b)}=(k_1,\ldots,k_L)\right)\\&=Q^*_{k_{2L-1},k_{2L}}\ldots Q^*_{k_{L+1},k_{L+2}}(Q^*)^{s+2-L}_{k_L,k_{L+1}}\end{aligned}$$

Therefore, we have

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\leq 2\sum_{i=1}^{n(s,b)-1}\sum_{k\in[K^*]} d_{TV}\left((Q^*)^{s+2-L}_{k,\cdot},\nu_i Q^{s+2-L}\right),$$

where $\nu_i=w^*(Q^*)^{b+(i-1)(s+1)+L-2}$ is the distribution of $H_{b+(i-1)(s+1)+L-1}$. Since $Q^*$ is irreducible and aperiodic, there exists a unique invariant probability $\pi^*$ and there are positive constants $C(Q^*)$ and $r(Q^*)$ such that

$$d_{TV}\left((Q^*)^t_{k,\cdot},\pi^*\right)\leq C(Q^*)e^{-r(Q^*)t},\forall k\in[K^*],\forall t\geq 1.$$

Combining the different inequalities we get

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\leq 4K^*(n(s,b)-1)C(Q^*)e^{-r(Q^*)(s+1)}.$$

We have

$$h^2\left(P^*,P_i\right)\leq d_{TV}\left(P^*,P_i\right)=d_{TV}\left(\pi^*,w^*(Q^*)^{i-1}\right)\leq C(Q^*)e^{-r(Q^*)(i-1)}.$$

**Proof of Lemma 3.14**

We denote by $(\mathscr{A}\times\mathscr{B})^+$ the set $\{(a,b)\in\mathscr{A}\times\mathscr{B};\mathbb{P}(A=a,B=b)>0\}$. We have

$$\begin{aligned}\mathbf{K}\left(\mathcal{L}(A,B)||\mathcal{L}(A)\otimes\mathcal{L}(B)\right)&=\sum_{(a,b)\in(\mathscr{A}\times\mathscr{B})^+}\mathbb{P}\left(A=a,B=b\right)\log\left(\frac{\mathbb{P}\left(A=a,B=b\right)}{\mathbb{P}\left(A=a\right)\mathbb{P}\left(B=b\right)}\right)\\&\leq\sum_{(a,b)\in(\mathscr{A}\times\mathscr{B})^+}\mathbb{P}\left(A=a,B=b\right)\left(\frac{\mathbb{P}\left(A=a,B=b\right)}{\mathbb{P}\left(A=a\right)\mathbb{P}\left(B=b\right)}-1\right)\\&=\sum_{(a,b)\in(\mathscr{A}\times\mathscr{B})^+}\frac{\left(\mathbb{P}\left(A=a,B=b\right)-\mathbb{P}\left(A=a\right)\mathbb{P}\left(B=b\right)\right)^2}{\mathbb{P}\left(A=a\right)\mathbb{P}\left(B=b\right)}.\end{aligned}$$

For $(a,b)\in(\mathscr{A}\times\mathscr{B})^+$,

$$\begin{aligned}&\frac{\left(\mathbb{P}\left(A=a,B=b\right)-\mathbb{P}\left(A=a\right)\mathbb{P}\left(B=b\right)\right)^2}{\mathbb{P}\left(A=a\right)\mathbb{P}\left(B=b\right)}\\&=\left|\mathbb{P}\left(A=a|B=b\right)-\mathbb{P}\left(A=a\right)\right|\times\left|\mathbb{P}\left(B=b|A=a\right)-\mathbb{P}\left(B=b\right)\right|\\&\leq\left|\mathbb{P}\left(B=b|A=a\right)-\mathbb{P}\left(B=b\right)\right|.\end{aligned}$$

Finally, we get

$$\mathbf{K}\left(\mathcal{L}(A,B)||\mathcal{L}(A)\otimes\mathcal{L}(B)\right)\leq\sum_{a\in\mathscr{A}}2d_{TV}\left(\mathcal{L}\left(B|A=a\right),\mathcal{L}\left(B\right)\right).$$

### 3.D.2 Proof of Corollary 3.3

We have

$$\mathbb{P}\left(X_i = (Y'_i, \dots, Y'_{i+L-1})\right) \geq \mathbb{P}\left(E_i = \cdots = E_{i+L-1} = 1\right) = p_i p_{i+1} \dots p_{i+L-1},$$

and with the convexity of the squared Hellinger distance

$$\begin{aligned} h^2\left(P_i, P^*\right) &\leq p_i p_{i+1} \dots p_{i+L-1} h^2\left(P'_i, P^*\right) + \left(1 - p_i p_{i+1} \dots p_{i+L-1}\right) \\ &\leq h^2\left(P'_i, P^*\right) + \left(1 - p_i\right) + \cdots + \left(1 - p_{i+L-1}\right), \end{aligned}$$

where $P'_i = \mathcal{L}(Y'_i, \dots, Y'_{i+L-1})$. One can check that $n \geq 1 + N/2$ with our conditions on $L$. With Theorem 3.4, Lemma 3.2 and Lemma 3.13 we have

$$\begin{aligned} C\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq h^2\left(P^*, \mathscr{M}\right) &+ \frac{C(Q^*)}{n(e^{r(Q^*)} - 1)} + \frac{L}{N}\sum_{i=1}^{N}(1 - p_i) \\ &+ e^{-r(Q^*)s} + L\epsilon^2 + (s+1)L\overline{V}\frac{\log n}{n}, \end{aligned}$$

for some positive constant $C$ and $s \geq L - 1$.

### 3.D.3 Proof of Theorems 3.5 and 3.6

With (3.45) and Theorem 3.4, we have

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \mathscr{M}\right) &+ n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) \\ &+ n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) \\ &+ L\epsilon^2 + (s+1)L^2K^L \log(2|\mathscr{F}_{\lambda_-,\lambda_+,M}|[\epsilon])\frac{\log n}{n}. \end{aligned}$$

We can simply follow the proof of Theorems 3.2 and 3.3 to conclude.

### 3.D.4 Proof of Proposition 3.1

The proof relies on the following lemma.

**Lemma 3.15.** *The set $\mathcal{A}$ of probability density functions, defined by*

$$\mathcal{A} = \left\{(x_1, \dots, x_L) \mapsto q_1(x_1) \dots q_L(x_L); q_i \in \mathcal{E}\left(\overline{\Theta}_i, \eta_i, T_i, d_i, B_i\right), \forall i \in \{1, \dots, L\}\right\},$$

*is VC-subgraph with VC-index $3 + d_1 + \cdots + d_L$.*

As $L \geq 2$ and $\max_{1 \leq k \leq K} d_k \geq 2$, Assumption 3.3 is met with

$$\overline{V} = 3K^L + K^{L-1}L\sum_{k=1}^{K} d_k \leq K^L\left(3 + L\max_{1 \leq k \leq K} d_k\right).$$

**Proof of Lemma 3.15**

We have

$$\begin{aligned}\mathcal{A} &= \left\{(x_1,\dots,x_L) \mapsto f_{\theta_1}(x_1)\dots f_{\theta_L}(x_L); \theta_i \in \overline{\Theta}_i, \forall i \in \{1,\dots,L\}\right\} \\ &= \exp\circ\left\{(x_1,\dots,x_L) \mapsto \sum_{i=1}^{L}\langle \eta_i(\theta_i), T_i(x_i)\rangle + A_i(\theta_i) + B_i(x_i), \forall i \in \{1,\dots,L\}\right\} \\ &\subset \exp\circ(V+B)\end{aligned}$$

with $B : (x_1,\dots,x_L) \mapsto B_i(x_1) + \dots + B_i(x_L)$ and

$$V = \left\{(x_1,\dots,x_L) \mapsto A + \sum_{i=1}^{K}\langle \eta_i, T_i(x_i)\rangle; \eta_i \in \mathbb{R}^d, \forall i \in \{1,\dots,L\}, A \in \mathbb{R}\right\}.$$

The set $V$ is a vector space of dimension $1 + d_1 + \dots + d_L$ and exp is monotone, therefore, from Proposition 42-(i,ii) [9] and Lemma 2.6.15 [84] and Lemma 2.6.18-(v) [84], the class of functions $\mathcal{A}$ is VC-subgraph with $VC$-index $V(\mathcal{A}) \leq 3 + d_1 + \dots + d_L$.

## 3.D.5 Proof of Proposition 3.2

We first need the following lemma to apply results of regular parametric models.

**Lemma 3.16.** *Under Assumption 3.5, our model is regular, i.e.*

- $\phi \mapsto p(\mathbf{x};\phi)$ *is continuous for all* $\mathbf{x}$,
- *it is differentiable for all* $\mathbf{x}$,
- *and the information matrix function*

$$I : \phi \mapsto I(\phi) = \int_{\mathcal{X}^L} \partial_\phi p(\mathbf{x};\phi)\left(\partial_\phi p(\mathbf{x};\phi)\right)^T \frac{\mu(\mathbf{x})}{p(\mathbf{x};\phi)}$$

*is well-defined and continuous.*

We can now apply results of Ibragimov and Has'minskiĭ [54], in particular (7.20) which is a consequence of Theorem 7.6. Let $\kappa$ be a compact subset of $\overline{\Phi}$ such that $\overline{\Phi}$ belongs to the interior of $\kappa$. There is a positive constants $a(\kappa)$ such that

$$\forall \phi \in \kappa, h^2\left(P_\phi, P_{\overline{\phi}}\right) \geq a(\kappa)\frac{||\phi-\overline{\phi}||^2}{1+||\phi-\overline{\phi}||^2} \geq \frac{a(\kappa)}{1+b(\kappa)}||\phi-\overline{\phi}||^2,$$

with $b(\kappa) = \max_{\phi\in\kappa}||\phi-\overline{\phi}||^2$. We know that $c(\kappa) := \inf_{\phi\in\overline{\Phi}\setminus\kappa} h^2\left(P_\phi, P_{\overline{\phi}}\right)$ is positive. Therefore, there exist a positive constant $C(\overline{\phi})$ such that

$$\begin{aligned}\forall \phi \in \overline{\Phi}, h^2\left(P_\phi, P_{\overline{\phi}}\right) &\geq \mathbb{1}_{\phi\in\kappa}\frac{a(\kappa)}{1+b(\kappa)}||\phi-\overline{\phi}||^2 + \mathbb{1}_{\phi\in\overline{\Phi}\setminus\kappa}c(\kappa) \\ &\geq C(\overline{\phi})\left[||\overline{w}-w||^2 + \left|\left|\overline{Q}-Q\right|\right|^2 + \sum_{k=1}^{K}\left|\left|\overline{\theta}-\theta\right|\right|^2 \wedge 1\right].\end{aligned}$$

**Proof of Lemma 3.16**

For $k_1,\dots,k_L \in [K]$ we have

$$p(\mathbf{x};\phi) \geq w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l). \tag{3.84}$$

- Since $\eta_k$ and $A_k$ are continuous for all $k$ in $[K]$, then the applications $\theta_k \mapsto f_{\theta_k}(x)$ are continuous for all $x \in \mathscr{X}$ and so is $\phi \mapsto p(\mathbf{x};\phi)$ for all $\mathbf{x} \in \mathscr{Y}^L$.
- The function $u \mapsto p(\mathbf{x};u)$ is differentiable at the point $u = \phi$ for all $\mathbf{x} \in \mathscr{Y}^L$ since $A_k$ and $\eta_k$ are differentiable for all $k \in [K]$. For all $\overline{k} \in [K]$ and $j \in [e_k]$,

$$\begin{aligned}
\partial_{\theta_{\overline{k},j}} p(\mathbf{x};\phi) &= \sum_{k_1,\dots,k_L} w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \sum_{l=1}^{L} \mathbb{1}_{k_l=\overline{k}} \left( \prod_{i \neq l} f_{\theta_{k_j}}(x_j) \right) \partial_{\theta_{\overline{k},j}} f_{\theta_{\overline{k}}}(x_l) \\
&= \sum_{k_1,\dots,k_L} w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \prod_{i=1}^{L} f_{\theta_{k_i}}(x_i) \\
&\times \sum_{l=1}^{L} \mathbb{1}_{k_l=\overline{k}} \left[ \langle \partial_{\theta_{\overline{k},j}} \eta_{\overline{k}}(\theta_{\overline{k}}), T_{\overline{k}}(x_l) \rangle + \partial_{\theta_{\overline{k},j}} A_{\overline{k}}(\theta_{\overline{k}}) \right].
\end{aligned} \tag{3.85}$$

For $\overline{k} \in [K-1]$ and $k' \in [K]$ we have

$$\begin{aligned}
\partial_{w_{\overline{k}}} p(\mathbf{x};\phi) &= \sum_{k_2,\dots,k_L} Q_{\overline{k},k_2} \dots Q_{k_{L-1},k_L} f_{\theta_{\overline{k}}}(x_1) \prod_{l=2}^{L} f_{\theta_{k_l}}(x_l) \\
&- \sum_{k_2,\dots,k_L} Q_{K,k_2} \dots Q_{k_{L-1},k_L} f_{\theta_K}(x_1) \prod_{l=2}^{L} f_{\theta_{k_l}}(x_l)
\end{aligned} \tag{3.86}$$

and

$$\begin{aligned}
\partial_{Q_{k',\overline{k}}} p(\mathbf{x};\phi) &= \sum_{k_1,k_2,\dots,k_L} w_{k_1} \partial_{Q_{k',\overline{k}}} \left[ Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \right] \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) \\
&= \sum_{k_1,k_2,\dots,k_L} w_{k_1} \prod_{i=1}^{L} f_{k_i,\theta_{k_i}}(x_i) \sum_{l=2}^{L} \left[ \mathbb{1}_{(k',\overline{k})=(k_{l-1},k_l)} - \mathbb{1}_{(k',K)=(k_{l-1},k_l)} \right] \prod_{\substack{2 \leq j \leq L, \\ j \neq l}} Q_{k_{j-1},k_j}.
\end{aligned} \tag{3.87}$$

Since $A_k$ and $\eta_k$ are $\mathcal{C}^1$, we just need to check that the functions

$$\phi \mapsto \int_{\mathscr{Y}^L} T_{\overline{k},j}(x_i) T_{\overline{k}',j'}(x_{i'}) \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}, \tag{3.88}$$

$$\phi \mapsto \int_{\mathscr{Y}^L} T_{\overline{k},j}(x_i) \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}, \tag{3.89}$$

$$\phi \mapsto \int_{\mathscr{Y}^L} \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}, \tag{3.90}$$

are well-defined and continuous for all $k_1,k'_1,\dots,k_L,k'_L,\overline{k},\overline{k} \in [K], j \in [d_{\overline{k}}], j' \in [d_{\overline{k}'}], i,i' \in [L]$, where

$$T_k(x) = (T_{k,1}(x),\dots,T_{k,d_k}(x)) \in \mathbb{R}^{d_k}, \forall x \in \mathscr{Y}.$$

We deal with integrability in the first time and then look at continuity, using (3.84) repeatedly.

- We have

$$\begin{aligned}
0 &\leq \int_{\mathscr{Y}^L} \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)} \\
&\leq \left(w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L}\right)^{-1} \int_{\mathscr{Y}^L} \prod_{l=1}^{L} f_{\theta_{k'_j}}(x_j) \mu(d\mathbf{x}) \\
&= \left(w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L}\right)^{-1} < \infty,
\end{aligned}$$

and (3.90) is well defined. Similarly

$$\begin{aligned}
0 &\leq \int_{\mathscr{Y}^L} \left|T_{\overline{k},j}(x_i)\right| \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)} \\
&\leq \left(w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}\right)^{-1} \int_{\mathscr{Y}} \left|T_{\overline{k},j}(x_i)\right| f_{\theta_{k_i}}(x_i) \nu(dx_i) \\
&\leq \left(w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}\right)^{-1} \sqrt{\int_{\mathscr{Y}} \left|T_{\overline{k},j}(x_i)\right|^2 f_{\theta_{k_i}}(x_i) \nu(dx_i)} < \infty,
\end{aligned}$$

and (3.89) is well defined. Finally

$$\begin{aligned}
0 &\leq \int_{\mathscr{Y}^L} \left|T_{\overline{k},j}(x_i) T_{\overline{k}',j'}(x_{i'})\right| \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)} \\
&\leq \left(w_{k_1} w_{k'_1} Q_{k_1,k_2} Q_{k'_1,k'_2} \dots Q_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}\right)^{-1/2} \\
&\times \int_{\mathscr{Y}^L} \left|T_{\overline{k},j}(x_i) T_{\overline{k}',j'}(x_{i'})\right| \sqrt{\prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \mu(d\mathbf{x})} \\
&\leq \left(w_{k_1} w_{k'_1} Q_{k_1,k_2} Q_{k'_1,k'_2} \dots Q_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}\right)^{-1/2} \\
&\times \sqrt{\int_{\mathscr{Y}} \left|T_{\overline{k},j}(x_i)\right|^2 f_{\theta_{k_i}}(x_i) \nu(dx_i)} \sqrt{\int_{\mathscr{Y}} \left|T_{\overline{k}',j'}(x_{i'})\right|^2 f_{\theta_{k'_{i'}}}(x_{i'}) \nu(dx_{i'})} < \infty,
\end{aligned}$$

and (3.88) is well defined. The Fisher information matrix $I(\phi)$ is well-defined for all $\phi$. We now turn to continuity.

- We have

$$\begin{aligned}
&\left|\frac{\prod_{l=1}^L f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)} - \frac{\prod_{l=1}^L f_{k_l,\theta'_{k_l}}(x_l) f_{k'_l,\theta'_{k'_l}}(x_l)}{p(\mathbf{x};\phi')}\right| \\
&\leq \frac{\prod_{l=1}^L f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)}\left|\prod_{l=1}^L f_{\theta_{k_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\right| \\
&+ \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\prod_{l=1}^L f_{\theta_{k'_l}}(x_l)\left|\frac{1}{p(\mathbf{x};\phi)} - \frac{1}{p(\mathbf{x};\phi')}\right| \\
&+ \frac{\prod_{l=1}^L f_{\theta'_{k_l}}(x_l)}{p(\mathbf{x};\phi')}\left|\prod_{l=1}^L f_{\theta_{k'_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k'_l}}(x_l)\right| \\
&\leq \frac{\left|\prod_{l=1}^L f_{\theta_{k_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\right|}{w_{k'_1} Q_{k'_1,k'_2}\ldots Q_{k'_{L-1},k'_L}} \\
&+ \frac{|p(\mathbf{x};\phi) - p(\mathbf{x};\phi')|}{w'_{k_1} w_{k'_1} Q'_{k_1,k_2} Q_{k'_1,k'_2}\ldots Q'_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}} \\
&+ \frac{\left|\prod_{l=1}^L f_{\theta_{k'_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k'_l}}(x_l)\right|}{w'_{k_1} Q'_{k_1,k_2}\ldots Q'_{k_{L-1},k_L}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\int_{\mathscr{Y}^L} \frac{\prod_{l=1}^L f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)}\mu(d\mathbf{x}) - \int_{\mathscr{Y}^L} \frac{\prod_{l=1}^L f_{\theta'_{k_l}}(x_l) f_{\theta'_{k'_l}}(x_l)}{p(\mathbf{x};\phi')}\mu(d\mathbf{x})\right| \\
&\leq \frac{2d_{TV}\left(\otimes_{l=1}^L F_{\theta_{k_l}}, \otimes_{l=1}^L F_{\theta'_{k_l}}\right)}{w_{k'_1} Q_{k'_1,k'_2}\ldots Q_{k'_{L-1},k'_L}} \\
&+ \frac{2d_{TV}\left(P_\phi, P_{\phi'}\right)}{w'_{k_1} w_{k'_1} Q'_{k_1,k_2} Q_{k'_1,k'_2}\ldots Q'_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}} \\
&+ \frac{2d_{TV}\left(\otimes_{l=1}^L F_{\theta_{k'_l}}, \otimes_{l=1}^L F_{\theta'_{k'_l}}\right)}{w'_{k_1} Q'_{k_1,k_2}\ldots Q'_{k_{L-1},k_L}}.
\end{aligned}$$

Since convergence with respect to the total variation distance and to the Hellinger distance are equivalent, we get continuity of (3.90) with Proposition 3.6. Similarly, we have

$$\begin{aligned}
&\left|\int_{\mathscr{Y}^L} \frac{T_{\overline{k},j}(x_i)\prod_{l=1}^L f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)}\mu(d\mathbf{x}) - \int_{\mathscr{Y}^L} \frac{T_{\overline{k},j}(x_i)\prod_{l=1}^L f_{\theta'_{k_l}}(x_l) f_{\theta'_{k'_l}}(x_l)}{p(\mathbf{x};\phi')}\mu(d\mathbf{x})\right| \\
&\leq \frac{\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)|\left|\prod_{l=1}^L f_{\theta_{k_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\right|\mu(d\mathbf{x})}{w_{k'_1} Q_{k'_1,k'_2}\ldots Q_{k'_{L-1},k'_L}} \\
&+ \frac{\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)|\,|p(\mathbf{x};\phi) - p(\mathbf{x};\phi')|\,\mu(d\mathbf{x})}{w'_{k_1} w_{k'_1} Q'_{k_1,k_2} Q_{k'_1,k'_2}\ldots Q'_{k_{L-1};k_L} Q_{k'_{L-1},k'_L}} \\
&+ \frac{\int |T_{k_l}(x_l)|\left|\prod_{i=l}^L f_{\theta_{k'_l}}(x_l) - \prod_{i=1}^L f_{k'_l,\theta'_{k'_l}}(x_l)\right|\mu(d\mathbf{x})}{w'_{k_1} Q'_{k_1,k_2}\ldots Q'_{k_{L-1},k_L}}.
\end{aligned}$$

We have

$$\begin{aligned}&\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \, |p(\mathbf{x};\phi) - p(\mathbf{x};\phi')| \, \mu(d\mathbf{x}) \\ &\leq \sum_{1\leq k_1,\dots,k_L\leq K} \int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \left| \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) - \prod_{l=1}^{L} f_{\theta'_{k_l}}(x_l) \right| \mu(d\mathbf{x})\end{aligned}$$

and

$$\begin{aligned}&\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \left| \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) - \prod_{l=1}^{L} f_{\theta'_{k_l}}(x_l) \right| \mu(d\mathbf{x}) \\ &\leq \int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \left| f_{\theta_{k_i}}(x_i) - f_{\theta'_{k_i}}(x_i) \right| \nu(dx_i) \\ &+ 2 \int_{\mathscr{Y}} |T_{\overline{k},j}(x_i)| f_{\theta_{k_i}}(x_i) \nu(dx_i) \times \sum_{l<i} d_{TV}\left(F_{\theta_{k_l}}, F_{\theta'_{k_l}}\right) \\ &+ 2 \int_{\mathscr{Y}} |T_{\overline{k},j}(x_i)| f_{\theta'_{k_i}}(x_i) \nu(dx_i) \times \sum_{l>i} d_{TV}\left(F_{\theta_{k_l}}, F_{\theta'_{k_l}}\right).\end{aligned}$$

As

$$\begin{aligned}&\int_{\mathscr{Y}} |T_{\overline{k},j}(x)| \left| f_{\theta_k}(x) - f_{\theta'_k}(x) \right| \nu(dx) \\ &\leq \sqrt{\int_{\mathscr{Y}} |T_{\overline{k},j}(x)|^2 \left| f_{\theta_k}(x) - f_{\theta'_k}(x) \right| \nu(dx)} \times \sqrt{2 d_{TV}\left(F_{\theta_k}, F_{\theta'_k}\right)} \xrightarrow[\theta'_k \to \theta_k]{} 0.\end{aligned}$$

for all $k \in [K]$ and $\theta_k \in \Theta_k$, we get continuity of (3.89). Similarly, we only need

$$\int_{\mathscr{Y}} |T_{\overline{k},j}(x)|^2 \left| f_{\theta_k}(x) - f_{\theta'_k}(x) \right| \nu(dx) \xrightarrow[\theta'_k \to \theta_k]{} 0$$

to obtain the continuity of (3.88).

### 3.D.6 Proof of Theorem 3.9

We start the proof with two lemmas that ensure we fit into the framework of Proposition 3.2.

**Lemma 3.17.** *The information matrix $I(\phi)$ is definite positive for all $\phi$ in $\overline{\Phi}$.*

**Lemma 3.18.** *Let $(\phi_n)_{n\in\mathbb{N}}$ be a sequence in $\overline{\Phi}$. If $\lim_{n\to\infty} h\left(P_{\phi_n}, P_{\overline{\phi}}\right) = 0$, then we have $\lim_{n\to\infty} \phi_n = \overline{\phi}$.*

One can see that Lemma 3.18 implies that $\inf_{\substack{||\phi-\overline{\phi}||\geq a \\ \phi\in\overline{\Phi}}} h^2\left(P_{\phi}, P_{\overline{\phi}}\right) > 0$ for all $a > 0$. Therefore we can apply Proposition 3.2. From Proposition 3.1, we get $\overline{V} \leq (3+L)K^L = 5K^3$.

#### Proof of Lemma 3.17

For $\mathbf{k} = (k_1, \dots, k_L) \in [K]^L$, the notation $wQ^{\bigcirc L}(\mathbf{k})$ is defined by (3.83). Following Theorem 1 of Meijer & Ypma [72], we have

$$\det(I(\phi)) = 0 \Leftrightarrow \exists \lambda \neq 0, \sum_i \lambda_i \partial_{\phi_i} p(\mathbf{x};\phi) = 0 \text{ for } \mu\text{-almost all } \mathbf{x}.$$

We can use (3.85), (3.86) and (3.87) to get

$$
\begin{aligned}
0 = &\sum_{\mathbf{k}\in[K]^L} wQ^{\bigcirc L}(\mathbf{k}) \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) \sum_{l=1}^{L}\sum_{j=1}^{e_{k_l}} \lambda_{\theta_{k_l,j}} \left[\langle \partial_{\theta_{k_l,j}} \eta_{k_l}(\theta_{k_l}), T_{k_l}(x_l)\rangle + \partial_{\theta_{k_l,j}} A_{k_l}(\theta_{k_l})\right] \\
&+ \sum_{k_1=1}^{K-1} \lambda_{w_{k_1}} \left[f_{\theta_{k_1}}(x_1) - f_{\theta_K}(x_1)\right] \sum_{k_2,\dots,k_L} \frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}} \prod_{i=2}^{L} f_{\theta_{k_i}}(x_i) \\
&+ \sum_{l=2}^{L}\sum_{k_l=1}^{K-1} \sum_{k_1,\dots,k_{l-1},k_{l+1},\dots,k_L} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}} \left[f_{\theta_{k_l}}(x_l) - f_{\theta_K}(x_l)\right] \prod_{i\neq l} f_{\theta_k}(x_i),
\end{aligned}
$$

for almost all $x$. If we apply it to exponential distributions, we get

$$
\begin{aligned}
0 = &- \sum_{\mathbf{k}\in[K]^L} wQ^{\bigcirc L}(\mathbf{k})\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L} \left(\sum_{l=1}^{L}\lambda_{\theta_{k_l}} x_l\right) \\
&+ \sum_{k_1=1}^{K-1} \lambda_{w_{k_1}} \sum_{k_2,\dots,k_L} \frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}} \theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L} \\
&- \sum_{k_1=1}^{K-1} \lambda_{w_{k_1}} \sum_{k_2,\dots,k_L} \frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}} \theta_K\theta_{k_2}\dots\theta_{k_L} e^{-\theta_K x_1-\dots-\theta_{k_L}x_L} \\
&+ \sum_{l=2}^{L}\sum_{k_l=1}^{K-1} \sum_{k_i;i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}} \theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_l}x_l} \\
&- \sum_{l=2}^{L}\sum_{k_l=1}^{K-1} \sum_{k_i;i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}} \theta_{k_1}\dots\theta_{k_{l-1}}\theta_K\theta_{k_{l+1}}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_K x_l-\dots-\theta_{k_l}x_l}.
\end{aligned} \tag{3.91}
$$

As $\theta_1 > \dots > \theta_K$, we can identify the coefficients for each $x \mapsto e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L}$. For $\mathbf{k} \in [K-1]^L$, we get

$$
\begin{aligned}
0 &= -wQ^{\bigcirc L}(\mathbf{k})\theta_{k_1}\dots\theta_{k_L}\left(\sum_{l=1}^{L}\lambda_{\theta_{k_l}}x_l\right) + \lambda_{w_{k_1}}\frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}}\theta_{k_1}\dots\theta_{k_L} \\
&\quad + \sum_{l=2}^{L}\lambda_{Q_{k_{l-1},k_l}}\frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}}\theta_{k_1}\dots\theta_{k_L} \text{ for almost all } \mathbf{x} \\
&\Rightarrow 0 = \lambda_{\theta_{k_1}} = \dots = \lambda_{\theta_{k_L}} = \frac{\lambda_{w_{k_1}}}{w_{k_1}} + \sum_{l=2}^{L}\frac{\lambda_{Q_{k_{l-1},k_l}}}{Q_{k_{l-1},k_l}}.
\end{aligned}
$$

This implies $\lambda_{\theta_k} = 0$ for all $k \in [K-1]$ and there are quantities $\lambda_w^*$ and $\lambda_Q^*$ such that $\frac{\lambda_{w_k}}{w_k} = \lambda_k^*$ for all $k \in [K-1]$ and $\frac{\lambda_{Q_{k_1,k_2}}}{Q_{k_1,k_2}} = \lambda_Q^*$ for $k_1$,$k_2 \in [K-1]$ and $\lambda_w^* + (L-1)\lambda_Q^* = 0$. Therefore, (3.91) becomes

$$
\begin{aligned}
0 = &\lambda_w^* \sum_{k_1=1}^{K-1} \sum_{k_2,\dots,k_L} wQ^{\bigcirc L}(\mathbf{k})\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L} \\
&- \lambda_w^* \sum_{k_2,\dots,k_L} \left(\sum_{k_1=1}^{K-1} wQ^{\bigcirc L}(\mathbf{k})\right) \theta_K\theta_{k_2}\dots\theta_{k_L} e^{-\theta_K x_1-\dots-\theta_{k_L}x_L} \\
&+ \sum_{l=2}^{L}\sum_{k_l=1}^{K-1} \sum_{k_i:i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}} \theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_l}x_l} \\
&- \sum_{l=2}^{L}\sum_{k_l=1}^{K-1} \sum_{k_i:i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}} \theta_{k_1}\dots\theta_K\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_K x_l-\dots-\theta_{k_l}x_l}.
\end{aligned} \tag{3.92}
$$

For $k_2,\dots,k_L \in [K-1]^{L-1}$, we write $\mathbf{k}' = (K,k_2,\dots,k_L)$ and with identification with respect to $\mathbf{x} \mapsto e^{-\theta_K x_1 - \theta_{k_2} x_2 - \cdots - \theta_{k_L} x_L}$ we have

$$\begin{aligned} 0 &= -\lambda_w^* \left( \sum_{k_1=1}^{K-1} wQ^{\bigcirc L}(\mathbf{k}) \right) \theta_K \theta_{k_2} \dots \theta_{k_L} + \lambda_{Q_{K,k_2}} \frac{wQ^{\bigcirc L}(\mathbf{k}')}{Q_{K,k_2}} \theta_K \theta_{k_2} \dots \theta_{k_L} \\ \Rightarrow \lambda_w^* &\left( \sum_{k_1=1}^{K-1} w_{k_1} Q_{k_1,k_2} \right) = \frac{\lambda_{Q_{K,k_2}}}{Q_{K,k_2}} w_K Q_{K,k_2}. \end{aligned}$$

For $k \in [K-1]$,

$$\frac{\lambda_{Q_{K,k}}}{Q_{K,k}} = \lambda_w^* \beta_k \text{ with } \beta_k = \frac{\sum_{k'=1}^{K-1} w_{k'} Q_{k',k}}{w_K Q_{K,k}}. \tag{3.93}$$

Finally (3.92) becomes

$$\begin{aligned} 0 &= \lambda_w^* \sum_{k_1=1}^{K-1} \sum_{k_2,\dots,k_L} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \cdots - \theta_{k_L} x_L} \\ &- \lambda_w^* \sum_{k_2,\dots,k_L} \left( \sum_{k_1=1}^{K-1} wQ^{\bigcirc L}(\mathbf{k}) \right) \theta_K \theta_{k_2} \dots \theta_{k_L} e^{-\theta_K x_1 - \cdots - \theta_{k_L} x_L} \\ &+ \lambda_Q^* \sum_{l=2}^{L} \sum_{k_{l-1},k_l \in [K-1]} \sum_{\substack{k_i \in [K]; \\ i \notin \{l-1,l\}}} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \cdots - \theta_{k_L} x_L} \\ &+ \lambda_w^* \sum_{l=2}^{L} \sum_{k_l \in [K-1]} \sum_{\substack{k_i \in [K]; \\ i \neq l}} \beta_{k_l} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_{l-2}} \theta_K \theta_{k_l} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \cdots - \theta_K x_{l-1} - \theta_{k_l} x_l - \cdots - \theta_{k_L} x_L} \\ &- \lambda_Q^* \sum_{l=2}^{L} \sum_{k_{l-1},k_l \in [K-1]} \sum_{\substack{k_i \in [K]; \\ i \notin \{l-1,l\}}} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_{l-1}} \theta_K \theta_{k_{l+1}} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \cdots - \theta_{k_{l-1}} x_{l-1} - \theta_K x_l - \cdots - \theta_{k_L} x_L} \\ &- \lambda_w^* \sum_{l=2}^{L} \sum_{k_l=1}^{K-1} \sum_{\substack{k_i \in [K]; \\ i \neq l}} \beta_{k_l} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_{l-2}} \theta_K \theta_K \theta_{k_{l+1}} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \cdots - \theta_K x_{l-1} - \theta_K x_l - \cdots - \theta_{k_L} x_L}. \end{aligned}$$

Identification with respect to $\mathbf{x} \mapsto e^{-\theta_K x_1 \cdots - \theta_K x_K}$ gives

$$\begin{aligned} 0 &= -\lambda_w^* \left( \sum_{k=1}^{K-1} w_{k_1} \right) Q_{K,K}^{L-1} - \lambda_w^* \sum_{l=2}^{L-1} \sum_{k_l=1}^{K-1} \beta_{k_l} w_K Q_{K,K}^{L-3} Q_{k_l,K} Q_{K,k_l}) - \lambda_w^* \sum_{k_L=1}^{K-1} \beta_{k_L} w_K Q_{K,K}^{L-2} Q_{K,k_L} \\ \Rightarrow 0 &= \lambda_w^* \left[ (1-w_K) Q_{K,K}^2 + (L-2) \sum_{k_2=1}^{K-1} w_K \beta_{k_2} Q_{k_2,K} Q_{K,k_2} + Q_{K,K} \sum_{k_2=1}^{K-1} w_K \beta_{k_2} Q_{K,k_2} \right] \\ \Rightarrow 0 &= \lambda_w^* \left[ (1-w_K) Q_{K,K}^2 + (L-2) \sum_{k_2=1}^{K-1} \left( \sum_{k_1}^{K-1} w_{k_1} Q_{k_1,k_2} \right) Q_{k_2,K} + Q_{K,K} \sum_{k_2=1}^{K-1} \sum_{k_1=1}^{K-1} w_{k_1} Q_{k_1,k_2} \right], \end{aligned}$$

where the last inequality comes from the definition of $\beta_k$. One can notice the quantity between the brackets is positive as a consequence of the definition of $O_K$. Therefore, we necessarily have $\lambda_w^* = 0$ and consequently $\lambda_Q^* = \lambda K,1 = \cdots = \lambda_{K,K-1} = 0$ which means $\lambda = 0$ and therefore the information matrix is definite positive.

**Proof of Lemma 3.18**

The parameters $w_k$ and $Q_{k,k'}$ are bounded so we can assume the sequences $w_{k,n}$ and $Q_{k,k'_n}$ are converging, with respective limits $w_k^*$ and $Q_{k,k'}^*$, even if it means extracting a subsequence. For other parameters, it is always possible to extract a subsequence $\phi_{\psi(n)}$ such that for all $k$ in $[K]$, we have $\theta_{k,\psi(n)} \xrightarrow[n\to\infty]{} \theta_k^* \in [0,\infty]$. We can deduce from the definition of $\overline{\Phi}$ that $\theta_1^* \geq \theta_2^* \geq \cdots \geq \theta_K^*$. Let us consider the following cases, dropping the dependency on $\psi$ in the notation.

- If $\theta_k^* = +\infty$, we have $\theta_{k,n} e^{-\theta_{k,n}x} \cdot dx \xrightarrow[n\to\infty]{\mathbb{P}} \text{Dirac}(0)$. Since $\lim_{n\to\infty} h\left(P_{\phi,n}, P_{\overline{\phi}}\right)$, we get that $w_{k_1}^* Q_{k_1,k_2}^* \dots Q_{k_{L-1},L} = 0$ if $k$ appears in $k_1, k_2, \dots, k_L$.
- If $\theta_k^* = 0$. We have
$$P_{\overline{\phi}}\left([\theta_{k,n}^{-1}, +\infty)^L\right) \leq (e^{-\overline{\theta}_K/\theta_{k,n}})^L \xrightarrow[n\to\infty]{} 0,$$
and
$$P_{\phi_n}\left([\theta_{k,n}, +\infty)^L\right) \geq w_{k_n} Q_{k_n,k_n}^{L-1} e^{-L}.$$
Since $\lim_{n\to\infty} h\left(P_{\phi_n}, P_{\overline{\phi}}\right) = 0$, we must have $w_k^* (Q_{k,k}^*)^{L-1} = 0$.

This proves that $P_{\phi_n}$ converges to

$$P_\infty(d\mathbf{x}) = \sum_{k_1,\dots,k_L \in [K]^+} w_{k_1}^* Q_{k_1,k_2}^* \dots Q_{k_{L-1},k_L}^* \theta_{k_l}^* \prod_{l=1}^{L} e^{-\theta_{k_l}^* x_l} dx_1 \dots dx_L,$$

with $[K]^+ = \{k \in [K]; \theta_k^* \in (0,\infty)\}$, and necessarily $P_\infty = P_{\overline{\phi}}$. We can easily identify the different parameters which implies that $(w^*, Q^*, \theta^*)$ and $(\overline{w}, \overline{Q}, \overline{\theta})$ are equal up to a permutation $\sigma$ on $[K]$. The ordering of the $\overline{\theta}_k$ and the $\theta_k^*$ ensures that this equality is true, not even up to a permutation.

## 3.D.7 Proof of Theorem 3.10

We just need to check that we satisfy Assumption 3.3. Then we can combine Proposition 3.3 and Theorem 3.4. We use Definition 41 [9] that allows us to consider functions taking values in $(-\infty, +\infty]$. From Lemma 2.6.15 [84], we have that

$$\{\mathbf{x} \mapsto (x_1 - z_1)(x_2 - z_2); z_1, z_2 \in \mathbb{R}\} \subset \{\mathbf{x} \mapsto a x_1 + b x_2 + x_1 x_2 + c; a, b, c \in \mathbb{R}\}$$

is VC-subgraph with VC-dimension smaller than or equal to 4. With Proposition 42-$(v)$ [9], we get that $\{\mathbf{x} \mapsto |x_1 - z_1| \cdot |x_2 - z_2|; z_1, z_2 \in \mathbb{R}\}$ is VC-subgraph with VC-dimension not larger than 37.608. We now need the following result.

**Lemma 3.19.** *If $\mathscr{A} \subset \mathcal{P}(\mathscr{X})$ is a VC-class with dimension $V$, then $\mathscr{F}_{\mathscr{A},a} := \{p_{A,a}; A \in \mathscr{A}\}$ is VC-subgraph with dimension $V$ for any $a$ in $\mathbb{R}$ where*

$$p_{A,a}(x) := \begin{cases} a \text{ if } x \in A, \\ +\infty \text{ otherwise.} \end{cases}$$

Since $\mathscr{C} := \{C_{z_1,z_2} := [z_1 \pm 1] \times [z_2 \pm 1]; z_1, z_2 \in \mathbb{R}\}$ is VC with VC-dimension 4, we get that $\mathscr{F}_{\mathscr{C},0}$ is VC-subgraph with VC-dimension 4. We can apply Proposition 42-$(v)$ [9] one more time which implies that $\mathscr{G} = \{\mathbf{x} \mapsto g_{z_1,z_2}(\mathbf{x}); z_1, z_2 \in \mathbb{R}\}$ is VC-subgraph with dimension at most $4.701(37.608 + 4) \leq 196$, with

$$\begin{aligned} g_{z_1,z_2}(\mathbf{x}) &:= p_{C_{z_1,z_2},0} \vee |x_1 - z_1| \cdot |x_2 - z_2| \\ &= \begin{cases} |x_1 - z_1| \cdot |x_2 - z_2| \text{ if } x \in [z_1 \pm 1] \times [z_2 \pm 1], \\ +\infty \text{ otherwise.} \end{cases} \end{aligned}$$

We need another lemma before we have a bound on the VC-dimension of

$$\mathscr{S}_{\alpha,2} := \left\{ \mathbf{x} \mapsto f_\alpha(x_1 - z_1) f_\alpha(x_2 - z_2) = \frac{(1-\alpha)^2}{4} \frac{1}{g^\alpha_{z_1,z_2}(\mathbf{x})}; z_1, z_2 \in \mathbb{R} \right\}.$$

**Lemma 3.20.** *Let $\mathscr{G}$ be a set of functions $\mathscr{X} \to [0,\infty]$. If $\mathscr{G}$ is VC-subgraph with VC-dimension at most $V$, then $\mathscr{G}^{-1} := \left\{ \frac{1}{g}; g \in \mathscr{G} \right\}$ is VC-subgraph with VC-dimension at most $V$, with the convention $1/0 = +\infty$ and $1/+\infty = 0$.*

Combining this lemma with Proposition 42-$(ii)$ [9], we get that $\mathscr{S}_{\alpha,2}$ is VC-subgraph with VC-dimension at most 196. This proves that we satisfy Assumption 3.3 with

$$\overline{V} = 4 \times 196 = 784.$$

**Proof of Lemma 3.19**

Assume that $\mathscr{F}_{\mathscr{A}}$ has VC-dimension larger than $V$. Therefore, there is $(x_i,u_i)_{i\in[V+1]} \in (\mathscr{X} \times \mathbb{R})^{[V+1]}$ such that for each $I \subset [V+1]$ we can find $A_I$ in $\mathscr{A}$ such that $i \in I \Leftrightarrow f_{A_I}(x_i) > u_i$. Necessarily, we have $u_i \geq a$ for all $i \in [V+1]$ and therefore $i \in I \Leftrightarrow x_i \notin A_I$. Therefore, $\mathscr{A}$ can shatter $(u_i)_{i_i n[V+1]}$ which contradicts the fact that its VC-dimension is at most $V$.

**Proof of Lemma 3.20**

We adapt the proof of Lemma 2.6.18 [84]. Let $(x_i,u_i)_{i\in[n]} \in (\mathscr{X} \times \mathbb{R})^n$ be such that for each $I \subset [n]$, we have $g_I \in \mathscr{G}$ such that

$$i \in I \Leftrightarrow \frac{1}{g_I(x_i)} > u_i.$$

For all $i \in [n]$, we necessarily have $u_i \geq 0$ and we define $a_i := \max\{g_J(x_i); \frac{1}{g_J(x_i)} > u_i\}$. One can check that we have

$$g_I(x_i) > a_i \Leftrightarrow \frac{1}{g_I(x_i)} \leq u_i.$$

Therefore $\mathscr{G}$ shatters $(x_i,a_i)_{i\in[n]} \in (\mathscr{X} \times \mathbb{R})^n$ which implies $n \leq V$.

## 3.D.8 Proof of Proposition 3.3

For $\pi = (\pi_{11},\pi_{12},\pi_{21},\pi_{22}) \in \mathcal{W}_4$ and $z \in \mathbb{R}$ we write

$$p_{\pi,z} := \pi_{11} f_\alpha \otimes f_\alpha + \pi_{12} f_\alpha \otimes f_\alpha(\cdot - z) + \pi_{21} f_\alpha(\cdot - z) \otimes f_\alpha + \pi_{22} f_\alpha(\cdot - z) \otimes f_\alpha(\cdot - z).$$

We define $\pi^* \in \mathcal{W}_4$ by $\pi^*_{11} = w^*(1 - q^*_{12})$, $\pi^*_{12} = w^* q^*_{12}$ and $\pi^*_{21} = (1 - w^*) q^*_{21}$. We also define $g : \mathcal{W}_4 \times \mathbb{R} \to \mathbb{R}$ by

$$g(\pi,z) = 2h^2\left(P_{\pi^*,z^*}, P_{\pi,z}\right) = \int_{\mathbb{R}^2} a^2_{\pi,z}(x_1,x_2)dx,$$

with $a_{\pi,z} : \mathbb{R}^2 \to \mathbb{R}$ defined by $a_{\pi,z}(x_1,x_2) = |\sqrt{p_{\pi,z}} - \sqrt{p_{\pi^*,z^*}}|$. We will drop the dependence on $\pi$ and $z$, and just write $a = a_{\pi,z}$. Without loss of generality we can assume $z^* > 0$ as we have $h^2(P_{\pi,-z},P_{\pi^*,-z^*}) = h^2(P_{\pi,z},P_{\pi^*,z^*})$. We define the set of parameters

$$\mathscr{Y} = \left\{ (\pi,z) \in \mathcal{W}_4 \times \mathbb{R}; z \in \left( \frac{z^*}{2} \vee z^* - \beta^*, z^* + \beta^* \right) \right\},$$

where $\beta^* \in (0,1]$ is set in the proof of Lemma 3.21 which proves the desired inequality on $\mathscr{Y}$.

**Lemma 3.21.** *There is a positive constant $C(\alpha,z^*,\pi^*)$ such that*

$$g(\pi,z) \geq C(\alpha,z^*,\pi^*)\left[(\pi_{11}^*-\pi_{11})^2+(\pi_{12}^*-\pi_{12})^2+(\pi_{21}^*-\pi_{21})^2+|z-z^*|^{1-\alpha}\right],$$

*for all $(\pi,z)$ in $\mathscr{Y}$.*

We also get that $g$ is lower bounded out of $\mathscr{Y}$ with the following lemma.

**Lemma 3.22.** *There is a positive constant $C(\alpha,z^*,\pi_{22}^*)$ such that*

$$g(\pi,z) \geq C(\alpha,z^*,\pi_{22}^*), \forall (\pi,z) \notin \mathscr{Y}.$$

One can check that we have $|z-z^*|^{1-\alpha} = (|z-z^*| \wedge 1)^{1-\alpha}$ for $(\pi,z) \in \mathscr{Y}$. And since $(\pi_{11}^*-\pi_{11})^2+(\pi_{12}^*-\pi_{12})^2+(\pi_{21}^*-\pi_{21})^2+(|z-z^*|\wedge 1)^{1-\alpha} \leq 3$ for all $\pi$ and all $z$, there is a positive constant $C(\alpha,z^*,\pi^*)$ such that

$$g(\pi,z) \geq C(\alpha,z^*,\pi^*)\left[(\pi_{11}^*-\pi_{11})^2+(\pi_{12}^*-\pi_{12})^2+(\pi_{21}^*-\pi_{21})^2+(|z-z^*|\wedge 1)^{1-\alpha}\right],$$

for all $\pi,z$. We now relate the distance to $\pi^*$ to the distance to $(w^*,q^*)$ with the following result.

**Lemma 3.23.** *For $w,q_{12},q_{21} \in [0,1]$ we have*

$$\begin{aligned}&(\pi_{11}-\pi_{11}^*)^2+(\pi_{12}-\pi_{12}^*)^2+(\pi_{21}-\pi_{21}^*)^2\\ &\geq \max\left(\frac{1}{2}(w-w^*)^2, \frac{(1-w^*)^2}{3}\left(q_{21}^*-q_{21}\right)^2, (w^*)^2\left(q_{12}-q_{12}^*\right)^2\right).\end{aligned}$$

This last result allows us to conclude the proof of Proposition 3.3.

**Proof of Lemma 3.21**

We will repeatedly use the following inequality

$$\forall x,y>0, \left|x^{1-\gamma}-y^{1-\gamma}\right| \geq \frac{(1-\gamma)|x-y|}{(x\vee y)^\gamma}. \tag{3.94}$$

Let $(\pi,z)$ be in $\mathscr{Y}$. Our goal is to lower bound $a$ on subsets of $\mathscr{Y}$ by a quantity related to the difference between some parameters. Inequalities (3.95), (3.96), (3.98) and (3.99) will be proved later.

- For $I_{11} = [-1,b)^2$ with $b = (z^* \wedge z \wedge 1) - 1$, we have

$$\int_{I_{11}} a(x_1,x_2)^2 dx_1 dx_2 \geq \frac{(1-\alpha)\left(1\wedge|z^*|/2\right)^2}{16}\left(\pi_{11}^*-\pi_{11}\right)^2. \tag{3.95}$$

- For

$$I_{22} = \begin{cases} (z^*,z^*+1)\times\left(z^*,z^*+(1-\alpha)^{2/\alpha}(\pi_{22}^*)^{1/\alpha}|z-z^*|\right) & \text{if } z^* \geq z,\\ \left(\frac{z^*}{2}\vee(z^*-1),z^*\right)\times\left(z^*,z^*+\frac{(1-\alpha)(\pi_{22}^*)^{1/\alpha}}{(1-\alpha)\left(2(\pi_{22}^*)^{1/\alpha}+1\right)+2}|z-z^*|\right) & \text{otherwise,}\end{cases}$$

we have

$$\int_{I_{22}} a^2(x_1,x_2)dx \geq \frac{\alpha^2}{4^3}\left(\frac{3-\alpha}{2-\alpha}\right)^2\left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}(\pi_{22}^*)^{1/\alpha}\left(1\wedge|z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}. \tag{3.96}$$

- Let $\beta \in (0,1]$. For

$$I_{12} := (-1, -(1 - z \wedge z^*)_+) \times (z \vee z^* + b_-, z \vee z^* + b_+),$$
$$I_{21} := (z \vee z^* + b_-, z \vee z^* + b_+) \times (-1, -(1 - z \wedge z^*)_+),$$

with

$$\begin{aligned} b_+ &= \mathbb{1}_{z \vee z^* \geq \beta}(1 - |z - z^*|) + \mathbb{1}_{z \vee z^* < \beta} \frac{z \vee z^*(1-\beta)}{\beta} \\ &\geq \mathbb{1}_{z^* \geq \beta}(1-\beta) + \mathbb{1}_{z^* < \beta} \frac{z^*(1-\beta)}{\beta} = (1 \wedge |z^*|/\beta)(1-\beta) \end{aligned} \tag{3.97}$$

and $b_- = b_+\delta$, $\delta \in (0,1)$. We have

$$\int_{I_{12}} a^2(x_1,x_2)dx \geq (\pi^*_{12} - \pi_{12})^2 \frac{(1-\alpha)^2 (1 \wedge |z^*|/2)}{8^2} (b_+)^{1-\alpha} (1-\delta) \mathbb{1}_{\Omega_{12}}, \tag{3.98}$$
$$\int_{I_{21}} a^2(x_1,x_2)dx \geq (\pi^*_{21} - \pi_{21})^2 \frac{(1-\alpha)^2 (1 \wedge |z^*|/2)}{8^2} (b_+)^{1-\alpha} (1-\delta) \mathbb{1}_{\Omega_{21}}, \tag{3.99}$$

with

$$I_{12} := \left\{ |\pi^*_{12} - \pi_{12}| \geq 2 \left[ \frac{\alpha |z - z^*|}{\delta b_+} + |\pi_{11} - \pi^*_{11}|(1-\beta)^\alpha \right] \right\},$$
$$I_{21} := \left\{ |\pi^*_{21} - \pi_{21}| \geq 2 \left[ \frac{\alpha |z - z^*|}{\delta b_+} + |\pi_{11} - \pi^*_{11}|(1-\beta)^\alpha \right] \right\}.$$

Combining (3.95), (3.96), (3.98) and (3.99), we have

$$\begin{aligned} \int a^2(x_1,x_2)dx \geq{}& (\pi^*_{11} - \pi_{11})^2 \frac{(1-\alpha)^2 (1 \wedge |z^*|/2)^2}{16} \\ &+ |z - z^*|^{1-\alpha} \frac{\alpha^2}{4^3} \left( \frac{3-\alpha}{2-\alpha} \right)^2 \left( \frac{1-\alpha}{5-3\alpha} \right)^{1-\alpha} (\pi^*_{22})^{1/\alpha} (1 \wedge |z^*|/2)^{1-\alpha} \\ &+ (\pi^*_{12} - \pi_{12})^2 \frac{(1-\alpha)^2 (1 \wedge |z^*|/2)}{8^2} (b_+)^{1-\alpha} (1-\delta) \mathbb{1}_{\Omega_{12}} \\ &+ (\pi^*_{21} - \pi_{21})^2 \frac{(1-\alpha) (1 \wedge |z^*|/2)}{8^2} (b_+)^{1-\alpha} (1-\delta) \mathbb{1}_{\Omega_{21}}, \end{aligned}$$

for $(\pi,z) \in \mathscr{Y}$. Then we can apply the following lemma.

**Lemma 3.24.** *Let $g, A_1, A_2, A_3, B$ be functions $\Theta \to \mathbb{R}$ and $D_1, D_{2,3}, D_B, C_A, C_B$ be positive constants such that*

$$\forall \theta \in \Theta, g(\theta) \geq D_1 A_1^2(\theta) + D_{2,3} \left( A_2^2(\theta) \mathbb{1}_{\Omega_2} + A_3^2(\theta) \mathbb{1}_{\Omega_3} \right) + D_B(\theta) B^{1-\alpha},$$

*where $\Omega_2$ and $\Omega_3$ are subsets of $\Theta$ given by*

$$\Omega_i := \{ \theta \in \Theta; A_i(\theta) \geq C_A A_1(\theta) + C_B B(\theta) \}.$$

*Then we have*

$$g(\theta) \geq \min \left( \frac{D_B}{1 + 4C_B^2}, \frac{D_1}{1 + 4C_A^2}, D_{2,3} \right) \left[ A_1^2(\theta) + A_2^2(\theta) + A_3^2(\theta) + B^{1-\alpha}(\theta) \right],$$

*for all $\theta$ in $\Theta$.*

In our situation, we get

$$\int a^2(x_1,x_2)dx \geq C(\alpha,z^*,\pi^*)\left[(\pi^*_{11}-\pi_{11})^2+(\pi^*_{12}-\pi_{12})^2+(\pi^*_{21}-\pi_{21})^2+|z-z^*|^{1-\alpha}\right]$$

with

$$\begin{aligned}C(\alpha,z^*,\pi^*) &= \min\left(\frac{\frac{\alpha^2}{4^3}\left(\frac{3-\alpha}{2-\alpha}\right)^2\left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}(\pi^*_{22})^{1/\alpha}(1\wedge|z^*|/2)^{1-\alpha}}{1+4^2\frac{\alpha^2}{\delta^2 b_+^2}}, \frac{\frac{(1-\alpha)^2(1\wedge|z^*|/2)^2}{4^2}}{1+4(1-\beta)^{2\alpha}},\right.\\ &\qquad \left.\frac{(1-\alpha)(1\wedge|z^*|/2)}{8^2}(b_+)^{1-\alpha}(1-\delta)\right)\\ &\geq \min\left(\frac{\frac{\alpha^2}{4^3}\left(\frac{3-\alpha}{2-\alpha}\right)^2\left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}(\pi^*_{22})^{1/\alpha}(1\wedge|z^*|/2)^{1-\alpha}}{1+4^2\frac{\alpha^2}{\delta^2(1\wedge|z^*|/2)^2(1-\beta)^2}}, \frac{(1-\alpha)^2(1\wedge|z^*|/2)^2}{4^2\left(1+4(1-\beta)^{2\alpha}\right)},\right.\\ &\qquad \left.\frac{(1-\alpha)(1\wedge|z^*|/2)}{8^2}(1\wedge|z^*|/2)^{1-\alpha}(1-\beta)^{1-\alpha}(1-\delta)\right)>0.\end{aligned}$$

We can optimize this bound with respect to $\beta$ and $\delta$, which gives $\beta^*$ depending only $z^*$, $\alpha$ and $\pi^*$. This concludes the proof of Lemma 3.21. We now prove the different inequalities.

*Proof of (3.95).* For $x_1,x_2\in[-1,0)^2$, we have

$$\begin{aligned}a(x_1,x_2) &= \frac{1-\alpha}{2|x_1|^{\alpha/2}|x_2|^{\alpha/2}}\\ &\times\left|\sqrt{\pi^*_{11}+\pi^*_{12}\frac{\mathbb{1}_{|x_2-z^*|\in(0,1]}|x_2|^\alpha}{|x_2-z^*|^\alpha}+\pi^*_{22}\frac{\mathbb{1}_{|x_1-z^*|\in(0,1]}\mathbb{1}_{|x_2-z^*|\in(0,1]}|x_1|^\alpha|x_2|^\alpha}{|x_1-z^*|^\alpha|x_2-z^*|^\alpha}+\pi^*_{21}\frac{\mathbb{1}_{|x_1-z^*|\in(0,1]}|x_1|^\alpha}{|x_1-z^*|^\alpha}}\right.\\ &\left.-\sqrt{\pi_{11}+\pi_{12}\frac{\mathbb{1}_{|x_2-z|\in(0,1]}|x_2|^\alpha}{|x_2-z|^\alpha}+\pi_{22}\frac{\mathbb{1}_{|x_1-z|\in(0,1]}\mathbb{1}_{|x_2-z|\in(0,1]}|x_1|^\alpha|x_2|^\alpha}{|x_1-z|^\alpha|x_2-z|^\alpha}+\pi_{21}\frac{\mathbb{1}_{|x_1-z|\in(0,1]}|x_1|^\alpha}{|x_1-z|^\alpha}}\right|.\end{aligned}$$

We set $b=\min(z^*,z,1)-1$. For $x_1,x_2\in[-1,b)^2$, we have

$$a(x_1,x_2)=\frac{1-\alpha}{2|x_1|^{\alpha/2}|x_2|^{\alpha/2}}\left|\sqrt{\pi^*_{11}}-\sqrt{\pi_{11}}\right|$$

and

$$\int_{[-1,b)^2}a(x_1,x_2)^2dx_1dx_2\geq\frac{\left[1-(-)^{1-\alpha}\right]^2}{4}\left|\sqrt{\pi^*_{11}}-\sqrt{\pi_{11}}\right|^2.$$

Finally, with (3.94) we always have

$$\begin{aligned}\int_{[-1,b)^2}a(x_1,x_2)^2dx_1dx_2 &\geq \frac{\left[1-(1-z\wedge z^*)_+^{1-\alpha}\right]^2}{4}\left(\sqrt{\pi^*_{11}}-\sqrt{\pi_{11}}\right)^2\\ &\geq \frac{(1-\alpha)\left(1\wedge|z^*|/2\right)^2}{4^2}\left(\pi^*_{11}-\pi_{11}\right)^2.\end{aligned}$$

*Proof of (3.96).* We need to consider two different cases.

- *First case $z^*\geq z$.* For $x\in I_{22}=(z^*,z^*+1)\times(z^*,z^*+V|z-z^*|)$ with $V<\frac{1}{|z-z^*|}$, we have $\frac{|x_2-z^*|}{|x_2-z|}\leq V$, $\frac{|x_2-z^*|}{|x_2|}\leq V$, $\frac{|x_1-z^*|}{|x_1-z|}\leq 1-|z-z^*|\leq 1$ and $\frac{|x_1-z^*|}{|x_1|}\leq\frac{1}{1+z^*}\leq 1$. Therefore, for $x\in I_{22}$, we have

$$a(x_1,x_2)\geq\frac{1-\alpha}{2|x_1-z^*|^{\alpha/2}|x_2-z^*|^{\alpha/2}}\left(\sqrt{\pi^*_{22}}-V^{\alpha/2}\right)_+.$$

For $V = (1-\alpha)^{2/\alpha}(\pi^*_{22})^{1/\alpha} < \frac{1}{|z-z^*|}$, we have

$$\int_{I_{22}} a^2(x_1,x_2)dx = \frac{\left(\sqrt{\pi^*_{22}} - V^{\alpha/2}\right)_+^2}{4} \left(V|z-z^*| \wedge 1\right)^{1-\alpha}$$
$$\geq \frac{(\pi^*_{22})^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z-z^*|^{1-\alpha}}{4}.$$

- *Second case $z^* < z$.* For $x \in \left(\frac{z^*}{2} \vee (z^*-1), z^*\right) \times (z^*, z^* + a|z-z^*|)$, $b \leq 1/2$ we have

$$a(x_1,x_2) \geq \frac{1-\alpha}{2|x_1-z^*|^\alpha |x_2-z^*|^\alpha} \left(\sqrt{\pi^*_{22}} - \left(\frac{b}{1-b}\right)^{\alpha/2}\right)_+.$$

For $b = (\pi^*_{22})^{1/\alpha} b'$ we have

$$\int_{I_{22}} a^2(x_1,x_2)dx \geq \frac{\left(\sqrt{\pi^*_{22}} - \left(\frac{b}{1-b}\right)^{\alpha/2}\right)_+^2}{4} \left(1 \wedge |z^*|/2\right)^{1-\alpha} b^{1-\alpha}|z-z^*|^{1-\alpha}$$
$$\geq \frac{(\pi^*_{22})^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}}{4} \left(1 - \left(\frac{b'}{1-b'(\pi^*_{22})^{1/\alpha}}\right)^{\alpha/2}\right)_+^2 (b')^{1-\alpha}$$
$$\geq \frac{(\pi^*_{22})^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}}{4} \frac{\alpha^2}{4} \left(\frac{1 - b'\left(1 + (\pi^*_{22})^{1/\alpha}\right)}{1-b'(\pi^*_{22})^{1/\alpha}}\right)^2 (b')^{1-\alpha}.$$

With $b' = \frac{1-\alpha}{(1-\alpha)(2\pi+1)+2}$ we have

$$\int_{I_{22}} a^2(x_1,x_2)dx \geq \frac{\alpha^2 (\pi^*_{22})^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}}{4^2}$$
$$\times \left(\frac{2 + (1-\alpha)(\pi^*_{22})^{1/\alpha}}{2 + (1-\alpha)\left(1 + (\pi^*_{22})^{1/\alpha}\right)}\right)^2 \left(\frac{1-\alpha}{(1-\alpha)(2\pi+1)+2}\right)^{1-\alpha}$$
$$\geq \frac{\alpha^2 (\pi^*_{22})^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}}{4^2}$$
$$\times \left(\frac{3-\alpha}{2+2(1-\alpha)}\right)^2 \left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}$$
$$= \frac{\alpha^2(3-\alpha)^2}{4^3(2-\alpha)^2} \left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha} (\pi^*_{22})^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}.$$

Finally, we always have have

$$\int_{I_{22}} a^2(x_1,x_2)dx \geq \frac{\alpha^2}{4^3}\left(\frac{3-\alpha}{2-\alpha}\right)^2 \left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha} (\pi^*_{22})^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}.$$

*Proof of (3.98).* We prove it for $I_{12}$ assuming $z^* \leq z$. The proof is similar for $I_{21}$ and for $z \leq z^*$. For $b = 0 \wedge (z^*-1)$ and $0 < c_- < c_+ < 1 - |z-z^*|$, we set $I_{12} = (-1,b) \times (z + c_-, z^* + 1)$. For $x_1, x_2 \in I_{12}$, we have

$$\frac{2|x_1|^{\alpha/2}|x_2-z|^{\alpha/2}}{1-\alpha} a(x_1,x_2) = \left| \frac{(\pi^*_{12} - \pi_{12}) + \pi^*_{12}\left(\frac{|x_2-z|^\alpha}{|x_2-z^*|^\alpha} - 1\right) + (\pi^*_{11} - \pi_{11})\frac{|x_2-z|^\alpha \mathbb{1}_{x_2 \leq 1}}{|x_2|^\alpha}}{\sqrt{\pi^*_{12}\frac{|x_2-z|^\alpha}{|x_2-z^*|^\alpha} + \pi^*_{11}\frac{|x_2-z|^\alpha \mathbb{1}_{x_2\leq 1}}{|x_2|^\alpha}} + \sqrt{\pi^*_{12} + \pi_{11}\frac{|x_2-z|^\alpha \mathbb{1}_{x_2\leq 1}}{|x_2|^\alpha}}} \right|.$$

We also have

$$\mathbb{1}_{x_2\le 1}\frac{|x_2-z|}{|x_2|}\le U(z,c_-,c_+):=\begin{cases}\frac{c_+}{z+c_+} & \text{if } z+c_+\le 1,\\ 1-z & \text{if } z+c_-<1<z+c_+,\\ 0 & \text{if } 1\le z+c_-.\end{cases}$$

For $c_+=\mathbb{1}_{z\ge\beta^*}(1-|z-z^*|)+\mathbb{1}_{z<\beta^*}\frac{z(1-\beta^*)}{\beta^*}$ we have $U(z,c_-,c_+)\le 1-\beta^*$. We also have

$$\begin{aligned}1-\frac{|x_2-z|^\alpha}{|x_2-z^*|^\alpha}&\le 1-\left(\frac{c_-}{c_-+|z-z^*|}\right)^\alpha\\ &\le \alpha\frac{\frac{|z-z^*|}{c_-+|z-z^*|}}{\left(\frac{c_-}{c_-+|z-z^*|}\right)^{1-\alpha}}=\alpha\frac{|z-z^*|}{c_-}\left(\frac{c_-}{c_-+|z-z^*|}\right)^\alpha\\ &\le\frac{\alpha|z-z^*|}{c_-}.\end{aligned}$$

Therefore, with $c_-=c_+\delta$, $\delta\in(0,1)$, on $I_{12}$ we have

$$\frac{2|x_1|^{\alpha/2}|x_2-z|^{\alpha/2}}{1-\alpha}a(x_1,x_2)\ge\frac{\left[|\pi^*_{12}-\pi_{12}|-\frac{\alpha|z^*-z|}{b_-}-(\pi^*_{11}-\pi_{11})\,(1-\beta^*)^\alpha\right]_+}{2}.$$

If $|\pi^*_{12}-\pi_{12}|\ge 2\left[|\pi^*_{11}-\pi_{11}|(1-\beta^*)^\alpha+\alpha|z-z^*|/c_-\right]$ then

$$\frac{2|x_1|^{\alpha/2}|x_2-z|^{\alpha/2}}{1-\alpha}a(x_1,x_2)\ge\frac{|\pi^*_{12}-\pi_{12}|}{4}$$

and

$$\begin{aligned}\int_{I_{12}}a^2(x_1,x_2)dx&\ge\frac{(\pi^*_{12}-\pi_{12})^2}{8^2}\left[1-(1-z\wedge z^*)_+^{1-\alpha}\right](c_+)^{1-\alpha}\left[1-\delta^{1-\alpha}\right]\\ &\ge(\pi^*_{12}-\pi_{12})^2\frac{(1-\alpha)^2\,(1\wedge|z|\wedge|z^*|)\,(c_+)^{1-\alpha}\,(1-\delta)}{8^2}.\end{aligned}$$

Otherwise we have $|\pi^*_{12}-\pi_{12}|<2\left[|\pi^*_{11}-\pi_{11}|(1-\beta^*)^\alpha+\alpha|z-z^*|/b_-\right]$.

**Proof of Lemma 3.22**

We need to go through numerous cases and subcases. Let $\beta^*$ be given in Lemma 3.21. Without loss of generality we are going to assume that $z^*>0$.
*Case 1*: $z\ge 0$ and $|z-z^*|\ge\beta^*$. Let $c$ be a positive constant.

- *Subcase 1.1*: $z^*>z$ or ($z^*<z$ and $\pi_{22}\ge c^2\pi^*_{22}$). For $x\in I=(z\vee z^*+\beta^*,z\vee z^*+1)^2$, we have

  $$a(x_1,x_2)=\frac{(1-\alpha)\,(\mathbb{1}_{z>z^*}\pi_{22}+\mathbb{1}_{z^*>z}\pi^*_{22})}{2|x_1-z\vee z^*|^{\alpha/2}|x_2-z\vee z^*|^{\alpha/2}},$$

  and therefore

  $$\begin{aligned}\int_I a^2(x_1,x_2)dx&=\frac{\mathbb{1}_{z>z^*}\pi_{22}+\mathbb{1}_{z^*>z}\pi^*_{22}}{4}\left(1-(\beta^*)^{1-\alpha}\right)^2\\ &\ge\frac{c^2\pi^*_{22}(1-\alpha)^2}{4}\,(1-\beta^*)^2.\end{aligned}$$

- *Subcase 1.2*: $1 \leq z^* < z$ and $\pi_{22} < c^2\pi_{22}^*$. For $x \in (z^*,z^* + 1 \wedge (|z - z^*|/2))^2$, we have

$$\frac{|x_1 - z^*|}{|x_1 - z|} \leq \frac{1 \wedge |z - z^*|/2}{z - z^* - 1 \wedge |z - z^*|/2} \leq 1.$$

We have

$$\begin{aligned} a(x_1,x_2) &\geq \frac{1-\alpha}{2|x_1 - z^*|^{\alpha/2}|x_2 - z^*|^{\alpha/2}} \left(\sqrt{\pi_{22}^*} - \sqrt{\pi_{22}}\right) \\ &\geq \frac{(1-\alpha)\sqrt{\pi_{22}^*}}{2|x_1 - z^*|^{\alpha/2}|x_2 - z^*|^{\alpha/2}} (1-c), \end{aligned}$$

and therefore

$$\begin{aligned} \int_I a^2(x_1,x_2)dx &\geq \frac{\pi_{22}^*}{4} (1-c)^2 \left(1 \wedge \frac{|z - z^*|}{2}\right)^{2(1-\alpha)} \\ &\geq \frac{\pi_{22}^* (1-c)^2}{2^{2(2-\alpha)}} (\beta^*)^{2(1-\alpha)}. \end{aligned}$$

- *Subcase 1.3*: $z^* \in (0,1-\beta^*]$ and $z^* < z$. Let $b$ be in $(0,1)$. For $x \in I = (z^* - bz^*,z^*)^2$ we have

$$\begin{aligned} \frac{|x_1 - z^*|}{|x_1|} &\leq \frac{bz^*}{z^* - bz^*} = \frac{b}{1-b} \\ \frac{|x_1 - z^*|}{|x_1 - z|} &\leq \frac{bz^*}{z - z^* + bz^*} \leq \frac{b\beta^*}{\beta^* + b\beta^*} \leq \frac{b}{1-b}. \end{aligned}$$

It implies

$$a(x_1,x_2) \geq \frac{1-\alpha}{2|x_1 - z^*|^{\alpha/2}|x_2 - z^*|^{\alpha/2}} \left(\sqrt{\pi_{22}^*} - \left(\frac{b}{1-b}\right)^{\alpha}\right)_+,$$

and for $b = b' (\pi_{22}^*)^{1/2\alpha}$ we get

$$\begin{aligned} \int_I a^2(x_1,x_2)dx &\geq \frac{(z^*)^{2(1-\alpha)} (\pi_{22}^*)^{(1-\alpha)/\alpha} (b')^{2(1-\alpha)}}{4} \left(\sqrt{\pi_{22}^*} - \sqrt{\pi_{22}^*} \left(\frac{b'}{1 - (\pi_{22}^*)^{1/2\alpha} b'}\right)^{\alpha}\right)_+^2 \\ &\geq \frac{(z^*)^{2(1-\alpha)} (\pi_{22}^*)^{1/\alpha} (b')^{2(1-\alpha)}}{4} \alpha^2 \left(1 - \frac{b'}{1 - (\pi_{22}^*)^{1/2\alpha} b'}\right)_+^2. \end{aligned}$$

For $b' = \frac{1}{1+2\left(\pi_{22}^*\right)^{1/2\alpha} + \frac{1}{1-\alpha}}$, we have

$$\int_I a^2(x_1,x_2)dx \geq \frac{(z^*)^{2(1-\alpha)} (\pi_{22}^*)^{1/\alpha} \alpha^2}{4\left(1 + 2(\pi_{22}^*)^{1/2\alpha} + \frac{1}{1-\alpha}\right)^{2(1-\alpha)}} \left(1 - \frac{1}{1 + (\pi_{22}^*)^{1/2\alpha} + \frac{1}{1-\alpha}}\right)^2.$$

- *Subcase 1.4*: $z^* < z$ and $z^* \in [1-\beta^*,1]$. Let $b$ be in $(0,1)$. For $x \in I = (z^*,z^* + b\beta^*)^2$ we have

$$\begin{aligned} \frac{|x_1 - z^*|}{|x_1 - z|} &\leq \frac{b\beta^*}{z - z^* - b\beta^*} \leq \frac{b}{1-b}, \\ \frac{|x_1 - z^*|}{|x_1|} &\leq \frac{b\beta^*}{z^* + b\beta^*} \leq \frac{b}{1+b} \leq \frac{b}{1-b}. \end{aligned}$$

It implies

$$a(x_1,x_2) \geq \frac{1-\alpha}{2|x_1-z^*|^{\alpha/2}|x_2-z^*|^{\alpha/2}}\left(\sqrt{\pi^*_{22}}-\left(\frac{b}{1-b}\right)^{\alpha}\right)_+,$$

and for $b = b'\left(\pi^*_{22}\right)^{1/2\alpha}$ we get

$$\begin{aligned}\int_I a^2(x_1,x_2)dx &\geq \frac{(\beta^*)^{2(1-\alpha)}\left(\pi^*_{22}\right)^{(1-\alpha)/\alpha}(b')^{2(1-\alpha)}}{4}\pi^*_{22}\left(1-\left(\frac{b'}{1-b'(\pi^*_{22})^{1/2\alpha}}\right)^{\alpha}\right)^2_+ \\ &\geq \frac{(\beta^*)^{2(1-\alpha)}\left(\pi^*_{22}\right)^{1/\alpha}\alpha^2}{4}(b')^{2(1-\alpha)}\left(1-\frac{b'}{1-b'(\pi^*_{22})^{1/2\alpha}}\right)^2_+.\end{aligned}$$

For $b' = \frac{1}{1+2\left(\pi^*_{22}\right)^{1/2\alpha}+\frac{1}{1-\alpha}}$ we have

$$\int_I a^2(x_1,x_2)dx \geq \frac{(\beta^*)^{2(1-\alpha)}\left(\pi^*_{22}\right)^{1/\alpha}\alpha^2}{4\left(1+2\left(\pi^*_{22}\right)^{1/2\alpha}+\frac{1}{1-\alpha}\right)^{2(1-\alpha)}}\left(1-\frac{1}{1+(\pi^*_{22})^{1/2\alpha}+\frac{1}{1-\alpha}}\right)^2.$$

We can optimize the subcases 1.1 and 1.2 with $c = \frac{(\beta^*/2)^{2(1-\alpha)}}{(\beta^*/2)^{2(1-\alpha)}+(1-\alpha)(1-\beta^*)}$. Gathering the different results, there is a positive constant $C_1(z^*,\pi^*_{22},\alpha)$ such that $\int_{\mathbb{R}^2} a(x_1,x_2)dx \geq C_1(\pi^*_{22},z^*,\alpha)$ for all $z$ satisfying $z \geq 0$ and $|z-z^*| \geq 1-\beta^*$.
*Case 2*: $z < 0$.

- *Subcase 2.1*: $z^* \leq 1$. Let $b$ be in (0,1). For $x \in (z^*,z^*+b)^2$ we have $\frac{|x_1-z^*|}{|x-z|} \leq \frac{|x_1-z^*|}{|x_1|} \leq \frac{b}{z^*}$ and therefore

$$a(x_1,x_2) \geq \frac{1-\alpha}{2|x_1-z^*|^{\alpha/2}|x_2-z^*|^{\alpha/2}}\left[\sqrt{\pi^*_{22}}-\left(\frac{b}{z^*}\right)^{\alpha}\right]_+.$$

  We get $\int_{(z^*,z^*+b)^2} a^2(x_1,x_2)dx \geq \frac{b^{2(1-\alpha)}\left[\sqrt{\pi^*_{22}}-\left(\frac{b}{z^*}\right)^{\alpha}\right]^2_+}{4}$. For $b = z^*(\pi^*_{22})^{1/2\alpha}(1-\alpha)^{1/\alpha} \leq 1$, we have

$$\int_{(z^*,z^*+b)^2} a^2(x_1,x_2)dx \geq \frac{\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}(z^*)^{2(1-\alpha)}(\pi^*_{22})^{1/\alpha}}{4}.$$

- *Subcase 2.2*: $z^* > 1$. For $x \in (z^*,z^*+1)^2$ we have

$$a(x_1,x_2) = \frac{1-\alpha}{2}\sqrt{\frac{\pi^*_{22}}{|x_1-z^*|^{\alpha}|x_2-z^*|^{\alpha}}}.$$

  Therefore we get $\int_{(z^*,z^*+1)^2} a^2(x_1,x_2)dx \geq \frac{\pi^*_{22}}{4}$.

Finally, we have

$$\int_{\mathbb{R}^2} a^2(x_1,x_2)dx \geq \frac{\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}(1\wedge z^*)^{2(1-\alpha)}(\pi^*_{22})^{1/\alpha}}{4}.$$

*Case 3*: $|z-z^*| < \beta^*$ and $z \leq z^*/2$. Let $b$ be in $(0,1/|z-z^*|)$. For $x \in (z^*,z^*+b|z-z^*|)^2$ we have

$$\begin{aligned}\frac{|x_1-z^*|}{|x_1|} &\leq \frac{b|z-z^*|}{z^*+b|z-z^*|} \leq b \\ \frac{|x_1-z^*|}{|x_1-z|} &\leq \frac{b|z-z^*|}{b|z-z^*|+|z^*-z|} \leq b.\end{aligned}$$

Therefore we get

$$a(x_1,x_2) \geq \frac{1-\alpha}{2|x_1-z^*|^{\alpha/2}|x_2-z^*|^{\alpha/2}}\left[\sqrt{\pi^*_{22}}-b^\alpha\right]_+.$$

We get

$$\int_{(z^*,z^*+b|z-z^*|)^2} a^2(x_1,x_2)dx \geq \frac{b^{2(1-\alpha)}|z-z^*|^{2(1-\alpha)}\left[\sqrt{\pi^*_{22}}-b^\alpha\right]_+^2}{4}$$

and for $b=(\pi^*_{22})^{1/2\alpha}(1-\alpha)^{1/\alpha}\leq 1/|z-z^*|$ we have

$$\begin{aligned}\int_{(z^*,z^*+b|z-z^*|)^2} a^2(x_1,x_2)dx &\geq \frac{|z-z^*|^{2(1-\alpha)}(1-\alpha)^{2(1-\alpha)/\alpha}\left(\pi^*_{22}\right)^{(1-\alpha)/\alpha}}{4}\pi^*_{22}\alpha^2\\ &\geq \frac{\alpha^2\left(|z^*|/2\right)^{2(1-\alpha)}(1-\alpha)^{2(1-\alpha)/\alpha}\left(\pi^*_{22}\right)^{1)/\alpha}}{4}.\end{aligned}$$

**Proof of Lemma 3.24.**

- For $\theta$ in $\Omega_2\cap\Omega_3$, we have

$$\begin{aligned}g(\theta) &\geq D_1A_1^2+D_{2,3}\left(A_2^2+A_3^2\right)+D_BB^{1-\alpha}\\ &\geq \min\left(D_1,D_{2,3},D_B\right)\left[A_1^2+A_2^2+A_3^2+B^{1-\alpha}\right].\end{aligned}$$

- For $\theta$ in $\Omega_2\cap\Omega_3^C$, we have

$$g(\theta)\geq D_1A_1^2+D_{2,3}A_2^2+D_BB^{1-\alpha}$$

and

$$A_3^2<\left(C_AA_1+C_BB\right)^2\leq 2C_A^2A_1^2+2C_B^2B^{1-\alpha}.$$

For $b=\frac{D_B}{1+2C_B^2}\bigwedge\frac{D_1}{1+2C_A^2}>0$ we have

$$\begin{aligned}g(\theta) &\geq D_{2,3}A_2^2+\left(D_1-b2C_A^2\right)A_1^2+D_{2,3}A_2^2+(D_B-b2C_B^2)B^{1-\alpha}+bA_3^2\\ &\geq \min\left(\frac{D_B}{1+2C_B^2},\frac{D_1}{1+2C_A^2},D_{2,3}\right)\left[A_1^2+A_2^2+A_3^2+B^{1-\alpha}\right].\end{aligned}$$

- For $\theta$ in $\Omega_2^C\cap\Omega_3^C$, we have

$$g(\theta)\geq D_1A_1^2+D_BB^{1-\alpha}$$

and

$$A_2^2+A_3^2<2\left(C_AA_1+C_BB\right)^2\leq 4C_A^2A_1^2+4C_B^2B^{1-\alpha}.$$

For $b=\frac{D_B}{1+4C_B^2}\bigwedge\frac{D_1}{1+4C_A^2}>0$ we have

$$\begin{aligned}g(\theta) &\geq D_{2,3}A_2^2+\left(D_1-b4C_A^2\right)A_1^2+(D_B-b4C_B^2)B^{1-\alpha}+b\left(A_2^2+A_3^2\right)\\ &\geq \min\left(\frac{D_B}{1+4C_B^2},\frac{D_1}{1+4C_A^2}\right)\left[A_1^2+A_2^2+A_3^2+B^{1-\alpha}\right].\end{aligned}$$

Finally, we always have

$$g(\theta)\geq \min\left(\frac{D_B}{1+4C_B^2},\frac{D_1}{1+4C_A^2},D_{2,3}\right)\left[A_1^2+A_2^2+A_3^2+B^{1-\alpha}\right].$$

**Proof of Lemma 3.23**

We assume there is $w$,$w^*$,$q_{12}$,$q_{12}^2$,$q_{21}$,$q_{21}^*$ in [0,1] such that

$$\pi_{11} = w(1-q_{12}), \pi_{12} = wq_{12}, \pi_{21} = (1-w)q_{21}$$

and

$$\pi_{11}^* = w^*(1-q_{12}^*), \pi_{12}^* = w^*q_{12}^*, \pi_{21}^* = (1-w^*)q_{21}^*.$$

- We have

$$\begin{aligned}(\pi_{11}-\pi_{11}^*)^2 + (\pi_{12}-\pi_{12}^*)^2 &= w^2\left[2\left(q_{12}-\frac{1}{2}\right)^2+\frac{1}{2}\right] \\ &\quad - 2ww^*\left[2\left(q_{12}^*-\frac{1}{2}\right)\left(q_{12}-\frac{1}{2}\right)+\frac{1}{2}\right] \\ &\quad + (w^*)^2\left[2\left(q_{12}^*-\frac{1}{2}\right)^2+\frac{1}{2}\right] \\ &= \frac{1}{2}(w-w^*)^2 + 2\left(w\left(q_{12}-\frac{1}{2}\right) - w^*\left(q_{12}^*-\frac{1}{2}\right)\right)^2 \\ &\geq \frac{1}{2}(w-w^*)^2. \end{aligned} \tag{3.100}$$

Therefore, we also have

$$\begin{aligned}&(\pi_{11}-\pi_{11}^*)^2 + (\pi_{12}-\pi_{12}^*)^2 + (\pi_{21}-\pi_{21}^*)^2 \\ &\geq \frac{1}{2}(w-w^*)^2 + ((1-w)q_{21} - (1-w^*)q_{21}^*)^2 \\ &= (1-w)^2\left[\frac{1}{2}+q_{21}^2\right] + (1-w^*)^2\left[\frac{1}{2}+(q_{21}^*)^2\right] \\ &- (1-w)(1-w^*)\left[1+2q_{21}q_{21}^*\right] \\ &= \left[\frac{1}{2}+q_{21}^2\right]\left((1-w)-(1-w^*)\frac{1+2q_{21}q_{21}^*}{1+2q_{21}^2}\right)^2 \\ &+ (1-w^*)^2\left[\frac{1}{2}+(q_{21}^*)^2\right] - \left[\frac{1}{2}+q_{21}^2\right](1-w^*)^2\left(\frac{1+2q_{21}q_{21}^*}{1+2q_{21}^2}\right)^2 \\ &\geq \frac{(1-w^*)^2}{2\left(1+2q_{21}^2\right)}\left[(1+2(q_{21}^*)^2)(1+2q_{21}^2) - (1+2q_{21}q_{21}^*)^2\right] \\ &= \frac{(1-w^*)^2}{1+2q_{21}^2}\left(q_{21}^*-q_{21}\right)^2 \\ &\geq \frac{(1-w^*)^2}{3}\left(q_{21}^*-q_{21}\right)^2. \end{aligned} \tag{3.101}$$

- Similarly, we have

$$\begin{aligned}(\pi_{11}-\pi_{11}^*)^2+(\pi_{12}-\pi_{12}^*)^2 &= w^2\left[q_{12}^2+(1-q_{12})^2\right]+(w^*)^2\left[(q_{12}^*)^2+(1-q_{12}^*)^2\right]\\ &\quad -2ww^*\left[q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*)\right]\\ &= \left[q_{12}^2+(1-q_{12})^2\right]\left(w-w^*\frac{q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*)}{q_{12}^2+(1-q_{12})^2}\right)^2\\ &\quad +(w^*)^2\left[(q_{12}^*)^2+(1-q_{12}^*)^2-\frac{\left(q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*)\right)^2}{q_{12}^2+(1-q_{12})^2}\right]\\ &\geq \frac{(w^*)^2}{q_{12}^2+(1-q_{12})^2}\left[\left((q_{12}^*)^2+(1-q_{12}^*)^2\right)\left((q_{12})^2+(1-q_{12})^2\right)\right.\\ &\quad \left.-\left(q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*)\right)^2\right]\\ &= (w^*)^2\frac{(q_{12}-q_{12}^*)^2}{q_{12}^2+(1-q_{12})^2}\\ &\geq (w^*)^2(q_{12}-q_{12}^*)^2. \end{aligned} \tag{3.102}$$

Finally, with (3.100),(3.101) and (3.102), we get

$$\begin{aligned}&(\pi_{11}-\pi_{11}^*)^2+(\pi_{12}-\pi_{12}^*)^2+(\pi_{21}-\pi_{21}^*)^2\\ &\geq \max\left(\frac{1}{2}(w-w^*)^2,\frac{(1-w^*)^2}{3}(q_{21}^*-q_{21})^2,(w^*)^2(q_{12}-q_{12}^*)^2\right).\end{aligned}$$

## 3.E Selection of the spacing parameter

This section gathers the proofs of Theorem 3.11, 3.12, Lemma 3.7 and Corollary 3.6.

### 3.E.1 Proof of Theorem 3.11

We first need the following result.

**Lemma 3.25.** *Let $\mathscr{M}$ be a finite set of probability distributions associated to the set of probability density functions $\mathcal{M}$, with respect to the $\sigma$-finite measure $\mu$. Let $\hat{P}=\hat{P}(n,\mathbf{X},\mathcal{M})$ be the $\rho$-estimator given by (3.7). For $t\in[n]$, there is an event $\Omega^*$ such that $\mathbb{P}(\Omega^*)\geq 1-\lceil n/t\rceil\beta_t(\mathbf{X})$ and for all $\xi>0$, with probability at least $1-2|\mathcal{M}|e^{-\xi}$, we have*

$$\begin{aligned}\mathbb{1}_{\Omega^*}\sum_{i=1}^n h^2\left(P_i,\hat{P}\right) &\leq \left(\frac{4a_0}{a_1}+1\right)\inf_{Q\in\mathscr{M}}\sum_{i=1}^n h^2(P_i,Q)\\ &\quad +\frac{8}{3a_1}(\xi+1.47)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right]+\frac{16.48}{a_1},\end{aligned}$$

*with $\alpha_0(t)=\frac{32\times 1.175ta_2^2}{a_1^2}+\frac{8}{3a_1}$, $a_0=4, a_1=3/8$ and $a_2^2=3\sqrt{2}$.*

Consequently, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n h^2\left(P_i,\hat{P}\right)\right] &\leq n\mathbb{P}\left((\Omega^*)^C\right)+\int_0^\infty\mathbb{P}\left(\mathbb{1}_{\Omega^*}\sum_{i=1}^n h^2\left(P_i,\hat{P}\right)\geq u\right)du\\ &\leq n\lceil n/t\rceil\beta_t(\mathbf{X})+\left(\frac{4a_0}{a_1}+1\right)\inf_{Q\in\mathscr{M}}\sum_{i=1}^n h^2(P_i,\mathscr{M})+\frac{16.48}{a_1}\\ &\quad +\frac{8}{3a_1}(2.47+\log(2|\mathcal{M}|))\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right].\end{aligned}$$

We apply this with $\mathcal{M} = \widehat{\mathcal{M}}_S\left(\mathbf{X}^{(1)}\right)$ and conditionally on $\mathbf{X}^{(1)}$. One can check that we have $\sqrt{1+18ta_2^2\alpha_0(t)} \leq 1 + 24\frac{ta_2^2}{a_1}\sqrt{1.175}$. We get

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_{\hat{s}}\right)\Big|\mathbf{X}^{(1)}\right] &\leq c_0' \inf_{s\in S}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_s\left(\mathbf{X}^{(1)}\right)\right)\\ &\quad + c_1'\left(2.47+\log(2|S|)\right)\left[1+96\sqrt{2.35}t\right]\\ &\quad + c_2' + n_2\lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right),\end{aligned}$$

with $c_0' = \frac{4a_0}{a_1}+1 = \frac{131}{3}$, $c_1' = \frac{2\times 8}{3a_1} = \frac{128}{9}$ and $c_2' = \frac{16.48}{a_1} = \frac{131.84}{3}$. As $t$ can be any number in $[n_2]$ we can take the infimum with respect to $t$ in the upper bound. Let $\overline{P}$ be in $\mathscr{P}_X$. We get

$$\begin{aligned}\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] &\leq \frac{2}{n_2}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_{\hat{s}}\right)\right] + \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right)\\ &\leq \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{2c_0'}{n_2}\inf_{s\in S}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_s\right)\right]\\ &\quad + \inf_{t\in[n_2]}\left\{\frac{c_1'}{n_2}\left(2.47+\log(2|S|)\right)\left[1+96\sqrt{2.35}t\right] + 2\lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right)\right\}\\ &\quad + \frac{2c_2'}{n_2}.\end{aligned}$$

From (3.14), for $s$ in $S$, we have

$$\begin{aligned}\frac{1}{n_2}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_s\right)\right] &\leq \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{4}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)\\ &\quad + \frac{4}{n_1}\mathbb{E}\left[\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\hat{P}_s\right)\right]\\ &\leq \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{4}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)\\ &\quad + \frac{4c_0}{n_1}\inf_{Q\in\mathscr{M}_s}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},Q\right) + 4c_1\frac{(s+1)}{n_1}\left[17 + D_{n(s,1)}(\mathscr{M}_s)\right]\\ &\quad + \frac{4c_2}{n_1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}_{s,b}^*||\mathbf{P}_{s,b}^{ind}\right).\end{aligned}$$

We get

$$\begin{aligned}\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] &\leq \frac{2+4c_0'}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{8c_0'}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)\\ &\quad + \inf_{t\in[n_2]}\left\{\frac{c_1'}{n_2}\left(2.47+\log(2|S|)\right)\left[1+96\sqrt{2.35}t\right] + 2\lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right)\right\}\\ &\quad + \frac{2c_2'}{n_2} + \frac{8c_0'}{n_1}\inf_{s\in S}\left\{c_0\inf_{Q\in\mathscr{M}_s}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},Q\right)\right.\\ &\quad \left. + c_1(s+1)\left[D_{n(s,1)}(\mathscr{M}) + 17\right] + c_2\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}_{s,b}^*||\mathbf{P}_{s,b}^{ind}\right)\right\}.\end{aligned}$$

We also have

$$\frac{1}{n_1}\inf_{Q\in\mathscr{M}_s}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},Q\right) \leq 2h^2(\overline{P},\mathscr{M}_s) + \frac{2}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right).$$

**Proof of Lemma 3.25**

For $P_i = \mathcal{L}(X_i), i = 1, \ldots, n$, we write

$$H^2_{Q,Q'} := \sum_{i=1}^{n} h^2\left(P_i, Q\right) + h^2\left(P_i, Q'\right).$$

**Lemma 3.26.** *Let $\delta > 1$ and $\nu > 0$ be such that*

$$e^{-\nu} + \sum_{j \geq 1} e^{-\delta^j \nu} \leq 1.$$

*For $t$ in $\{1, \ldots, n\}$, there is an event $\Omega^*$ satisfying $\mathbb{P}(\Omega^*) \geq 1 - \lceil n/t \rceil \beta_t$ such that for all $p$ in $\mathcal{M}$ and all $\xi > 0$, we have*

$$\mathbf{P}^*\left(\sup_{q \in \mathcal{M}} \left\{ |\mathbf{Z}_n(\mathbf{X}, p, q)| \, \mathbb{1}_{\Omega^*} - \frac{a_1}{2} H^2_{P,Q} \right\} > \frac{2(\upsilon + \xi)}{3} \left[1 + \sqrt{1 + 18 t a_2^2 \alpha}\right]\right) \leq 2|\mathcal{M}| e^{-\xi},$$

*with $\mathbf{P}^* = \mathcal{L}(\mathbf{X})$ and $\alpha \geq \alpha_0(t) = \frac{32 t a_2^2 \delta}{a_1^2} + \frac{8}{3 a_1}$.*

We take $\delta = 1.175$ and $\upsilon = 1.47$ as in [11] Section A.1. Let $\xi > 0$ and $p \in \mathcal{M}$. On the event $\Omega^*$ defined by Lemma 3.26 and with Proposition 3 [11], we have for all $q \in \mathcal{M}$,

$$\begin{aligned}
\mathbf{T}_n(\mathbf{X}, p, q) &\leq \mathbb{E}\mathbf{T}_n(\mathbf{X}, p, q) + |\mathbf{Z}(\mathbf{X}, p, q)| \\
&\leq \sum_{i=1}^{n} \left[ a_0 h^2\left(P_i, P\right) - a_1 h^2\left(P_i, Q\right) \right] \\
&+ \frac{a_1}{2} H^2_{P,Q} + \frac{2(\xi + \upsilon)}{3} \left[1 + \sqrt{1 + 18 t a_2^2 \alpha_0(t)}\right] \\
&= \sum_{i=1}^{n} \left[ \left(a_0 + \frac{a_1}{2}\right) h^2\left(P_i, P\right) - \frac{a_1}{2} h^2\left(P_i, Q\right) \right] \\
&+ \frac{2}{3}(\xi + \upsilon) \left[1 + \sqrt{1 + 18 t a_2^2 \alpha_0(t)}\right].
\end{aligned}$$

Then,

$$\begin{aligned}
\boldsymbol{\Upsilon}_n(\mathbf{X}, p) &= \sup_{q \in \mathcal{M}} \mathbf{T}_n(\mathbf{X}, p, q) \\
&\leq \left(a_0 + \frac{a_1}{2}\right) \sum_{i=1}^{n} h^2\left(\mathbf{P}_i^{ind}, P\right) \\
&- \frac{a_1}{2} \inf_{Q \in \mathcal{M}} \sum_{i=1}^{n} h^2\left(P_i, Q\right) \\
&+ \frac{2}{3}(\xi + \upsilon) \left[1 + \sqrt{1 + 18 t a_2^2 \alpha_0(t)}\right],
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Upsilon}_n(\mathbf{X}, q) &= \sup_{q' \in \mathcal{M}} \mathbf{T}_n(\mathbf{X}, q, p) \\
&\geq \mathbf{T}_n(\mathbf{X}, q, p) = -\mathbf{T}_n(\mathbf{X}, p, q) \\
&\geq -\left(a_0 + \frac{a_1}{2}\right) \sum_{i=1}^{n} h^2\left(P_i, P\right) + \frac{a_1}{2} \sum_{i=1}^{n} h^2\left(P_i, Q\right) \\
&- \frac{2}{3}(\xi + \upsilon) \left[1 + \sqrt{1 + 18 t a_2^2 \alpha_0(t)}\right].
\end{aligned}$$

Since $\boldsymbol{\Upsilon}_n(\mathbf{X},\hat{p}) < \boldsymbol{\Upsilon}_n(\mathbf{X},p) + 8.24$, we have

$$\frac{a_1}{2}\sum_{i=1}^{n} h^2\left(P_i,\hat{P}\right) \leq 2\left(a_0+\frac{a_1}{2}\right)\sum_{i=1}^{n} h^2(P_i,P) - \frac{a_1}{2}\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2(P_i,\mathscr{M}) + \frac{4}{3}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right] + 8.24.$$

Given that $\mathscr{M}$ is finite we can take $P$ such that

$$\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2(P_i,Q) = \sum_{i=1}^{n} h^2(P_i,P).$$

Hence we have

$$\sum_{i=1}^{n} h^2\left(P_i,\hat{P}\right) \leq \left(\frac{4a_0}{a_1}+1\right)\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2(P_i,Q) + \frac{8}{3a_1}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right] + \frac{16.48}{a_1}.$$

**Proof of Lemma 3.26**

**Lemma 3.27.** *For $t$ in $[n]$, there is an event $\Omega^*$ such that $\mathbb{P}(\Omega^*) \geq 1 - \lceil n/t\rceil\beta_t(\mathbf{X})$ and*

$$\forall q,q' \in \mathcal{M}, \forall x>0, \mathbb{P}\left(|\mathbf{Z}_n(\mathbf{X},q,q')|\,\mathbb{1}_{\Omega^*} > \frac{2x}{3}\left[1+\sqrt{1+\frac{18ta_2^2H_{Q,Q'}^2}{x}}\right]\right) \leq 2e^{-x}. \tag{3.103}$$

Let $\xi>0$ and $\alpha>0$. We define $x_0 = \upsilon+\xi$ and for $j\geq 0$,

$$y_{j+1}^2 = \delta y_j^2 = \delta\alpha x_j. \tag{3.104}$$

Let $q,q'$ be in $\mathcal{M}$. We apply Lemma 3.27 according to the value of $H_{Q,Q'}^2$.

- If there is $j\geq 0$ such that $y_j^2 \leq H_{Q,Q'}^2 < y_{j+1}^2$, with probability at least $1-2e^{-x_j}$, we have

$$\begin{aligned}|\mathbf{Z}_n(\mathbf{X},q,q')|\mathbb{1}_{\Omega^*} - \frac{a_1}{2}H_{Q,Q'}^2 &\leq \frac{2x_j}{3}\left[1+\sqrt{1+\frac{18ta_2^2H_{Q,Q'}^2}{x_j}}\right] - \frac{a_1}{2}H_{q,q'}^2\\ &\leq \frac{2x_j}{3}\left[1+\sqrt{1+\frac{18ta_2^2y_{j+1}^2}{x_j}}\right] - \frac{a_1}{2}y_j^2\\ &\leq \frac{2x_j}{3}\left[1+\sqrt{1+18ta_2^2\delta\alpha} - \frac{3a_1\alpha}{4}\right]\\ &\leq 0,\end{aligned}$$

  for

$$\alpha \geq \alpha_0(t) := \frac{32\delta ta_2^2}{a_1} + \frac{8}{3a_1}. \tag{3.105}$$

- If $H_{Q,Q'}^2 < y_0^2$, with probability at least $1-2e^{-x_0}$, we have

$$\begin{aligned}|\mathbf{Z}_n(\mathbf{X},q,q')|\mathbb{1}_{\Omega^*} - \frac{a_1}{2}H_{Q,Q'}^2 &\leq |\mathbf{Z}_n(\mathbf{X},q,q')|\mathbb{1}_{\Omega^*}\\ &\leq \frac{2x_0}{3}\left[1+\sqrt{1+18ta_2^2\alpha}\right].\end{aligned}$$

Let $\overline{p}$ be in $\mathcal{M}$. Finally, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{q\in\mathcal{M}}\left\{|\mathbf{Z}_n(\mathbf{X},\overline{p},q)|\,\mathbb{1}_{\Omega^*}-\frac{a_1}{2}H_{\overline{p},q}^2\right\}>\frac{2x_0}{3}\left[1+\sqrt{1+18ta_2^2\alpha}\right]\right)\\
&\leq\sum_{\substack{q\in\mathcal{M}:\\ H_{\overline{P},Q}^2<y_0^2}}\mathbb{P}\left(|\mathbf{Z}_n(\mathbf{X},\overline{p},q)|\mathbb{1}_{\Omega^*}-\frac{a_1}{2}H_{\overline{p},q}^2>\frac{2x_0}{3}\left[1+\sqrt{1+18ta_2^2\alpha}\right]\right)\\
&+\sum_{j\geq 0}\sum_{\substack{q\in\mathcal{M}:\\ y_j^2\leq H_{\overline{p},q}^2<y_{j+1}^2}}\mathbb{P}\left(|\mathbf{Z}_n(\mathbf{X},\overline{p},q)|\,\mathbb{1}_{\Omega^*}-\frac{a_1}{2}H_{\overline{p},q}^2>0\right)\\
&\leq\sum_{\substack{q\in\mathcal{M}:\\ H_{\overline{p},q}^2<y_0^2}}2e^{-x_0}+\sum_{j\geq 0}\sum_{\substack{q\in\mathcal{M}:\\ y_j^2\leq H_{\overline{p},q}^2<y_{j+1}^2}}2e^{-x_j}\\
&\leq 2|\mathcal{M}|\left(e^{-x_0}+\sum_{j\geq 1}e^{-x_j}\right)=2|\mathcal{M}|\left(e^{-(\upsilon+\xi)}+\sum_{j\geq 1}e^{-\delta^j(\upsilon+\xi)}\right)\\
&\leq 2|\mathcal{M}|e^{-\xi}\left(e^{-\upsilon}+\sum_{j\geq 1}e^{-\delta^j\upsilon}\right)\leq 2|\mathcal{M}|e^{-\xi}.
\end{aligned}$$

**Proof of Lemma 3.27**

We follow the proof of Sart [80] (Proposition B.1). Let $t$ be a positive integer in $[n]$. Let $l$ be the smallest integer larger than $n/2t$. We derive from Berbee's lemma and more precisely from Viennet [36] (page 484) that there exist $B_1^*,\ldots,B_{2lt}^*$ such that

- For $i=1,\ldots,l$, the random vectors

$$B_{i,1}=\left(X_{2(i-1)t+1},\ldots,X_{(2i-1)t}\right)\text{ and }B_{i,1}^*=\left(X_{2(i-1)t+1}^*,\ldots,X_{(2i-1)t}^*\right)\tag{3.106}$$

  have the same distribution, and so have the random vectors

$$B_{i,2}=\left(X_{(2i-1)t+1},\ldots,X_{2it}\right)\text{ and }B_{i,2}^*=\left(X_{(2i-1)t+1}^*,\ldots,X_{2it}^*\right).\tag{3.107}$$

- The random vectors $B_{1,1}^*,\ldots,B*_{l,1}$ are independent. The random vectors $B_{1,2}^*,\ldots,B*_{l,2}$ are also independent.

- The event

$$\Omega^*=\bigcap_{1\leq j\leq l}\left\{B_{j,1}=B_{j,1}^*\right\}\cap\left\{B_{j,2}=B_{j,2}^*\right\}$$

  satisfies $\mathbb{P}\left((\Omega^*)^C\right)\leq 2l\beta_t\left(\mathbf{X}\right)$.

Let $q,q'$ be in $\mathcal{M}$. For simplicity, we write $Z_{q,q'}=\mathbf{Z}(\mathbf{B},q,q')$ and we define

$$\begin{aligned}
Z_{q,q',1}^*&:=\sum_{i=1}^{l}\sum_{j=1}^{t}\left\{\psi\left(\sqrt{\frac{q'}{q}\left(X_{2(i-1)t+j}^*\right)}\right)-\mathbb{E}\left[\psi\left(\sqrt{\frac{q'}{q}\left(X_{2(i-1)t+j}^*\right)}\right)\right]\right\}\mathbb{1}_{2(i-1)t+j\leq n}\\
&=\sum_{i=1}^{l}\sum_{j=1}^{t}z_{2(i-1)t+j}^{q,q'}\mathbb{1}_{2(i-1)t+j\leq n}
\end{aligned}$$

and

$$Z^*_{q,q',2} := \sum_{i=1}^{l}\sum_{j=1}^{t}\left\{\psi\left(\sqrt{\frac{q'}{q}\left(X^*_{(2i-1)t+j}\right)}\right) - \mathbb{E}\left[\psi\left(\sqrt{\frac{q'}{q}\left(X^*_{(2i-1)t+j}\right)}\right)\right]\right\}\mathbb{1}_{(2i-1)t+j\leq n}$$
$$= \sum_{i=1}^{l}\sum_{j=1}^{m} z^{q,q'}_{(2i-1)t+j}\mathbb{1}_{(2i-1)t+j\leq n}.$$

Let $\xi$ be a positive real number. Since

$$|Z_{q,q'}|\mathbb{1}_{\Omega^*} > \xi \Rightarrow |Z^*_{q,q',1}|\mathbb{1}_{\Omega^*} > \xi/2 \text{ or } |Z^*_{q,q',2}|\mathbb{1}_{\Omega^*} > \xi/2, \tag{3.108}$$

we have

$$\mathbb{P}\left(|Z_{q,q'}|\,\mathbb{1}_{\Omega^*} > \xi\right) \leq \mathbb{P}\left(|Z^*_{q,q',1}|\mathbb{1}_{\Omega^*} > \xi/2\right) + \mathbb{P}\left(|Z^*_{q,q',2}|\mathbb{1}_{\Omega^*} > \xi/2\right)$$
$$\leq \mathbb{P}\left(|Z^*_{q,q',1}| > \xi/2\right) + \mathbb{P}\left(|Z^*_{q,q',2}| > \xi/2\right).$$

One can notice that $Z^*_{q,q',1}$ and $Z^*_{q,q',2}$ are sums of $l$ independent variables. Therefore, we can use classic concentration inequalities. First, we can see that

$$V_{q,q',1} = \sum_{i=1}^{l}\mathbb{E}\left[\left(\sum_{j=1}^{t} z^{2(i-1)t+j}_{q,q'}\mathbb{1}_{2(i-1)t+j}\right)^2\right]$$
$$\leq \sum_{i=1}^{l}\sum_{j=1}^{t} t\mathbb{E}\left[\left(z^{2(i-1)t+j}_{q,q'}\right)^2\mathbb{1}_{2(i-1)t+j}\right]$$
$$\leq t\sum_{i=1}^{n}\mathrm{Var}\left(\psi\left(\sqrt{\frac{q'}{q}\left(X^*_i\right)}\right)\right)$$
$$\leq t\sum_{i=1}^{n} a_2^2\left[h^2(P_i,Q) + h^2(P_i,Q')\right] = ta_2^2H^2_{Q,Q'}.$$

The last inequality comes from Proposition 3 in Baraud & Birgé [11] and $a_2^2 = 3\sqrt{2}$. Similarly we have $V_{Q,Q',2} \leq ta_2^2L_{Q,Q'}$. Therefore, Bennett's inequality (see Proposition 2.8 and inequality (2.16) in Massart [67]) guarantees that for all $\xi > 0$ we have

$$\mathbb{P}\left(|Z_{q,q'}|\mathbb{1}_{\Omega^*} > \xi\right) \leq 2\exp\left(-\frac{(\xi/2)^2}{2(ta_2^2H^2_{q,q'} + \xi/6)}\right).$$

For $x > 0$, we take $\xi = \frac{2x}{3}\left[1 + \sqrt{1 + \frac{18ta_2^2H^2_{Q,Q'}}{x}}\right]$ and with probability less than or equal to $2e^{-x}$, we have

$$|Z_{q,q'}|\mathbb{1}_{\Omega^*} > \frac{2x}{3}\left[1 + \sqrt{1 + \frac{18ta_2^2H^2_{Q,Q'}}{x}}\right]. \tag{3.109}$$

## 3.E.2 Proof of Lemma 3.7

We have

$$\beta_t(\mathbf{Y}) = \sup_i \beta\left(\sigma(Y_1,\ldots,Y_i);\sigma(Y_{i+t},\ldots,Y_n)\right)$$
$$= \sup_i d_{TV}\left(\mathcal{L}(Y_1,\ldots,Y_i)\otimes\mathcal{L}(Y_{i+t},\ldots,Y_n),\mathcal{L}(Y_1,\ldots,Y_i,Y_{i+t},\ldots,Y_n)\right).$$

We use the notation $X_a^b = (X_a, \ldots, X_b)$ and similarly for $\mathbf{E}$, $\mathbf{Y}$ and $\mathbf{Z}$. The triangle inequality implies

$$\begin{aligned}
&d_{TV}\left(\mathcal{L}\left(Y_1^i\right) \otimes \mathcal{L}\left(Y_{i+t}^n\right), \mathcal{L}\left(Y_1^n\right)\right) \\
&\leq \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right) d_{TV}\left(\mathcal{L}(Y_1^i|E_1^i=e_1^i) \otimes \mathcal{L}(Y_{i+t}^n|E_{i+t}^N=e_{i+t}^n), \mathcal{L}(Y_1^i,Y_{i+t}^n|E_1^i=e_1^i,E_{i+t}^n=e_{i+t}^n)\right) \\
&= \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right) \beta\left(\sigma((X_j)_{\substack{j\leq i,\\ e_j=1}}), \sigma((X_j)_{\substack{j\geq i+k,\\ e_j=1}})\right).
\end{aligned}$$

We now need the following result to conclude.

**Lemma 3.28.** *For any random variables $A_1$,$A_2$,$B_1$,$B_2$, we have*

$$\beta\left(\sigma(A_1),\sigma(A_2)\right) \leq \beta\left(\sigma(A_1,B_1), \sigma(A_2,B_2)\right).$$

Combining the different inequalities above, we get

$$\begin{aligned}
\beta_t\left(\mathbf{Y}\right) &\leq \sup_i \beta\left(\sigma(Y_1^i); \sigma(Y_{i+t}^n)\right) \\
&= \sup_i \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right) \beta\left(\sigma((X_j)_{\substack{j\leq i,\\ e_j=1}}), \sigma((X_j)_{\substack{j\geq i+t,\\ e_j=1}})\right) \\
&\leq \sup_i \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right) \beta\left(\sigma((X_j)_{j\leq i}), \sigma((X_j)_{j\geq i+t})\right) = \beta_t\left(\mathbf{X}\right).
\end{aligned}$$

**Proof of Lemma 3.28**

Let $\mu_1$, $\mu_2$, $\nu_1$ and $\nu_2$ be measures dominating respectively $\mathcal{L}(A_1)$, $\mathcal{L}(A_2)$, $\mathcal{L}(B_1)$ and $\mathcal{L}(B_2)$. We have

$$\begin{aligned}
&\beta\left(\sigma(A_1),\sigma(A_2)\right) \\
&= \frac{1}{2}\int |p_A(a_1,a_2) - p_{A_1}(a_1)p_{A_2}(a_2)|\mu_1(da_1)\mu_2(da_2) \\
&= \frac{1}{2}\int |\int \left(p_{A,B}(a_1,b_1,a_2,b_2) - p_1(a_1,b_1)p_2(a_2,b_2)\right)\nu_1(db_1)\nu_2(db_2)|\mu_1(da_1)\mu_2(da_2) \\
&\leq \frac{1}{2}\int |p_{A,B}(a_1,b_1,a_2,b_2) - p_1(a_1,b_1)p_2(a_2,b_2)|\nu_1(db_1)\nu_2(db_2)\mu_1(da_1)\mu_2(da_2 \\
&= \beta\left(\sigma(A_1,B_1); \sigma(A_2,B_2)\right),
\end{aligned}$$

with $p_A = \frac{d\mathcal{L}(A_1,A_2)}{d\mu_1\otimes\mu_2}$, $p_{A_1} = \frac{d\mathcal{L}(A_1)}{d\mu_1}$, $p_{A_2} = \frac{d\mathcal{L}(A_2)}{d\mu_2}$, $p_{A,B} = \frac{d\mathcal{L}(A_1,B_1,A_2,B_2)}{d\mu_1\otimes\nu_1\otimes\mu_2\otimes\nu_2}$, $p_1 = \frac{d\mathcal{L}(A_1,B_1)}{d\mu_1\otimes\nu_1}$ and $p_2 = \frac{d\mathcal{L}(A_2,B_2)}{d\mu_2\otimes\mu_2}$.

### 3.E.3 Proof of Theorem 3.12

From (3.82) we have

$$\begin{aligned}
h^2\left(\overline{P}, \mathscr{M}_s\right) &\leq 2L\epsilon^2 + 2L(K-1)\delta(s) + 2h^2\left(\overline{P}, \overline{\mathscr{M}}\right) \\
&\leq 2L\epsilon^2 + 2h^2\left(\overline{P}, \mathscr{M}\right) + 2(s+1)L\frac{\overline{V}}{n_1}.
\end{aligned}$$

From Proposition 3.5 we have $D_{n_1(s,1)}\left(\mathscr{M}_s\right) \leq CL\overline{V}\log n_1$, for a constant $C$. For $S$ defined by (3.71), we have

$$|S| = 2 + \lfloor \log_\tau(\lfloor (n_1-2)/2 \rfloor)\rfloor \leq 2 + \frac{\log n_1}{\log \tau} \leq C\log n_1,$$

for some positive constant $C$. Theorem 3.11 allows us to obtain (3.72).

The following result is proven in Section 3.E.3.

**Lemma 3.29.** *Under Assumption 3.7, there exist positive constants $r(Q^*)$,$C(Q^*) > 0$ such that*

- *for all $j \in [2]$ and all $i \in [n_j]$, we have*

$$h^2\left(P_i^{(j)},P^*\right) \leq C(Q^*)e^{-r(Q^*)i}, \tag{3.110}$$

- *for all $t \in [n_2]$, we have*

$$\beta_t\left(\mathbf{X}^{(2)}\right) \leq C(Q^*)e^{-r(Q^*)t/2}, \tag{3.111}$$

- *for all $s \geq L-1$, all $b$ in $[s+1]$,*

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \leq n(s,b)C(Q^*)e^{-r(Q^*)s}. \tag{3.112}$$

From (3.110) we have

$$\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},P^*\right), \sum_{i=1}^{n_1} h^2\left(P_i^{(1)},P^*\right) \leq \frac{C(Q^*)}{e^{r(Q^*)}-1}.$$

For $t = n_2 \wedge \lceil 4r(Q^*)^{-1}\log n_2\rceil$, with (3.111) we have

$$\begin{aligned}\lceil n_2/t\rceil \beta_t\left(\mathbf{X}^{(2)}\right) &\leq \begin{cases} 1 \text{ for } n_2 \leq r(Q^*)^{-1}4\log n_2, \\ C(Q^*)n_2^{-1} \text{ otherwise,}\end{cases} \\ &\leq n_2^{-1}\left(C(Q^*) \vee r(Q^*)^{-1}4\log n_2\right).\end{aligned}$$

We have the following

$$\begin{aligned}\left\lceil \frac{\log\log n_1 - \log r(Q^*)}{\log\tau}\right\rceil > \left\lfloor \frac{\log\left\lfloor\frac{n_1-2}{2}\right\rfloor}{\log\tau}\right\rfloor &\Rightarrow \frac{\log\log n_1 - \log r(Q^*)}{\log\tau} > \frac{\log\left\lfloor\frac{n_1-2}{2}\right\rfloor}{\log\tau} - 1 \\ &\Rightarrow \tau r(Q^*)^{-1}\log n_1 \geq \left\lfloor\frac{n_1-2}{2}\right\rfloor \\ &\Rightarrow 2\frac{2+\tau r(Q^*)^{-1}\log n_1}{n_1} \geq 1.\end{aligned}$$

For $s = \lceil \tau^j\rceil$ with $j = \left\lceil \frac{\log\log n_1 - \log r(Q^*)}{\log\tau}\right\rceil \wedge \left\lfloor \frac{\log\left\lfloor\frac{n_1-2}{2}\right\rfloor}{\log\tau}\right\rfloor$, we have

$$s \leq \tau^{\frac{\log\log n_1 - \log r(Q^*)}{\log\tau}+1} + 1 = 1 + \tau r(Q^*)^{-1}\log n_1,$$

and inequality (3.112) gives

$$\begin{aligned}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) &\leq C(Q^*)n_1 e^{-r(Q^*)s} \\ &\leq C(Q^*)n_1\left(2\frac{2+\tau r(Q^*)^{-1}\log n_1}{n_1} \vee \frac{1}{n_1}\right) = 2C(Q^*)(2+\tau r(Q^*)^{-1}\log n_1).\end{aligned}$$

These last inequalities give (3.73).

**Proof of Lemma 3.29**

We just have to follow the proof of Lemma 3.13. We already have (3.110) and (3.112). The inequality (3.111) can be deduced from the inequality

$$d_{TV}\left(Q^t_{k,\cdot},\pi\right) \leq Ce^{-rt},$$

and from the definition of $\beta_t$.

### 3.E.4 Proof of Corollary 3.6

We have

$$\mathbb{P}\left(X_i^{(j)} = \left(\overline{Y}_i^{(j)},\dots,\overline{Y}_{i+L-1}^{(j)}\right)\right) \geq \mathbb{P}\left(E_i^{(j)} = \cdots = E_{i+L-1}^{(j)} = 1\right) = p_i^{(j)}p_{i+1}^{(j)}\dots p_{i+L-1}^{(j)},$$

and with the convexity of the squared Hellinger distance

$$\begin{aligned} h^2\left(P_i^{(j)},P^*\right) &\leq p_i^{(j)}p_{i+1}^{(j)}\dots p_{i+L-1}^{(j)}h^2\left(\overline{P}_i^{(j)},P^*\right) + \left(1-p_i^{(j)}p_{i+1}^{(j)}\dots p_{i+L-1}^{(j)}\right) \\ &\leq h^2\left(\overline{P}_i^{(j)},P^*\right) + \left(1-p_i^{(j)}\right) + \cdots + \left(1-p_{i+L-1}^{(j)}\right), \end{aligned}$$

where $\overline{P}_i^{(j)} = \mathcal{L}\left(\overline{Y}_i^{(j)},\dots,\overline{Y}_{i+L-1}^{(j)}\right)$. One can check that $n \geq 1 + N/2$ with our conditions on $L$. With Theorem 3.12, Lemma 3.7 and Lemma 3.29 we have

$$\begin{aligned} C\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq\ & h^2\left(P^*,\mathscr{M}\right) + \frac{C(Q^*)}{n_1(e^{r(Q^*)}-1)} + \frac{C(Q^*)}{n_2(e^{r(Q^*)}-1)} \\ &+ L\epsilon^2 + \frac{L}{N_1}\sum_{i=1}^{N_1}\left(1-p_i^{(1)}\right) + \frac{L}{N_2}\sum_{i=1}^{N_2}\left(1-p_i^{(2)}\right) \\ &+ \inf_{t\in[n_2]}\left\{\frac{t\log\log n_1}{n_2} + \lceil n_2/t\rceil C(Q^*)e^{-r(Q^*)t/2}\right\} \\ &+ \inf_{s\in S}\left\{(s+1)L\overline{V}\frac{\log n_1}{n_1} + e^{-r(Q^*)s}\right\}, \end{aligned}$$

for some positive constant $C$ and $s \geq L-1$. We can control the last terms with reasonable choices of $t$ and $s$ following the proof of Theorem 3.12.

# Chapter 4

# Model selection for finite state space hidden Markov models

**Abstract**

We observe $n$ observations generated that we believe were generated by a finite state space hidden Markov model and we aim at estimating the different parameters, i.e. stationary distribution and transition matrix of the hidden chain and the emission distributions. This chapter is an extension of the work made in Chapter 3. We establish a general result of model selection and focus on two specific situations, the selection of the order for a fixed emission model, and the selection of emission models among a collection for a fixed order.

## 4.1 Introduction

Let $(Y_i,H_i)_i$ be a finite state space HMM with parameters $(K^*,w^*,Q^*,F^*)$. If $w^*$ is invariant with respect to $Q^*$ the process $(Y_i,H_i)_i$ is stationary. As we did in Chapter 3, we aim at estimating the different parameters through the distribution of consecutive observations. For $L \geq 2$, we define the distribution $P_L$ by $P_L = P_{w^*,Q^*,F^*}$, where

$$P_{w,Q,F} = \sum_{1\leq k_1,\dots,k_L\leq K} w_{k_1} Q_{k_1,k_2} \dots Q_{K_{L-1},k_L} \bigotimes_{l=1}^{L} F_{k_l}, \tag{4.1}$$

and we have $\mathcal{L}(Y_i,Y_{i+1},\dots,Y_{i+L-1}) = P_L$ for all $i$. In Chapter 3 we built an estimator of $P_L$ on a fixed model, the order $K$ and the emission models being fixed. This is restrictive as we want to consider different emission models and/or different orders. We use model selection to overcome this problem.

## 4.2 The model selection procedure

Let $X_1,X_2,\dots,X_n$ be random variables taking values in a measurable space $(\mathscr{X},\mathcal{X})$. We denote by $\mathscr{P}_X$ the class of probability distributions on $(\mathscr{X},\mathcal{X})$ and define the distribution $P_i := \mathcal{L}(X_i) \in \mathscr{P}_{\mathcal{X}}$ for all $i \in [n]$.

### 4.2.1 Reminders of $\rho$-estimation

We denote by $\psi$ the function given by

$$\psi : \left| \begin{array}{l} [0,+\infty] \to [-1,1] \\ x \mapsto \frac{x-1}{x+1} \end{array} \right. .$$

Let $\mathscr{M}$ be a countable subset of $\mathscr{P}_X$ such that there is an associated set of density functions $\mathcal{M}$ with respect to a $\sigma$-finite measure $\mu$. Let pen be a penalty function mapping $\mathscr{M}$ to $\mathbb{R}$. For $n \geq 1$, we denote by $\mathbf{T}_n$ and $\boldsymbol{\Upsilon}_n$ the functions given by

$$\mathbf{T}_n : \left| \begin{array}{l} \mathscr{X}^n \times \mathcal{M} \times \mathcal{M} \to [-1,1] \\ (\mathbf{x},q,q') \mapsto \sum_{k=1}^{n} \psi\left(\sqrt{\frac{q'(y_i)}{q(y_i)}}\right) \end{array} \right.$$

with the convention $0/0 = 1$, $a/0 = +\infty$ for all $a > 0$, and

$$\boldsymbol{\Upsilon}_n : \left| \begin{array}{l} \mathscr{X}^n \times \mathcal{M} \\ (\mathbf{x},q) \mapsto \sup_{q'\in\mathcal{M}} \{\mathbf{T}_n(\mathbf{x},q,q') - \mathrm{pen}(Q')\} + \mathrm{pen}(Q) \end{array} \right. .$$

For $\mathbf{x}$ in $\mathscr{X}^n$, we define the (nonvoid) set $\mathscr{E}_n(\mathbf{x})$ by

$$\mathscr{E}_n(\mathbf{x}) = \left\{ Q = q \cdot \mu \middle| q \in \mathcal{M}, \Upsilon_n(\mathbf{x},q) < \inf_{q' \in \mathcal{M}} \Upsilon_n(\mathbf{x},q') + 11.36 \right\}.$$

We denote by $\hat{P}(n,\mathbf{X},\mathscr{M},\text{pen})$ any measurable element of the closure of $\mathscr{E}_n(\mathbf{X})$ with respect to the Hellinger distance and we call it a $\rho$-estimator on $\mathscr{M}$. The constant 11.36 is given by (7) and (19) in [11] but can be replaced by any smaller positive number.

### 4.2.2 The estimator

For $s \in \{0,1,\ldots,s_{\max}\}, s_{\max} = \lfloor (n-2)/2 \rfloor$, we define $s+1$ subsets of observations $\mathbf{X}^{(s,1)}$, $\mathbf{X}^{(s,2)}$, $\ldots$, $\mathbf{X}^{(s,s+1)}$ by

$$X_i^{(s,b)} := X_{b+(i-1)(s+1)} \in \mathscr{X}, \forall i \in [n(s,b)], \tag{4.2}$$

for $b$ in $[s+1]$, where $n(s,b) := \left\lfloor \frac{n+s+1-b}{1+s} \right\rfloor \geq 2$. We define the associated probability distributions $\mathbf{P}_{s,b}^*$ and $\mathbf{P}_{s,b}^{ind}$ by

$$\mathbf{P}_{s,b}^* := \mathcal{L}\left(\mathbf{X}^{(s,b)}\right) \text{ and } \mathbf{P}_{s,b}^{ind} := \bigotimes_{i=1}^{n(s,b)} P_{b+(i-1)(s+1)}. \tag{4.3}$$

We denote for short $\mathbf{P}^* = \mathbf{P}_{0,1}^*$ the distribution of $\mathbf{X} = (X_1,\ldots,X_n)$ and $\mathbf{P}^{ind} = \mathbf{P}_{0,1}^{ind} = \mathcal{L}(X_1) \otimes \cdots \otimes \mathcal{L}(X_n)$. Our estimator is obtained with the following statistical procedure.

1. Let $s$ be in $\{0,\ldots,\lfloor (n-2)/2 \rfloor\}$. For $b$ in $[s+1]$, we denote by $\hat{P}_{s,b}$ the estimators given by
$$\hat{P}_{s,b} := \hat{P}(\mathscr{M},n(s,b),\mathbf{X}^{(s,b)},\text{pen}). \tag{4.4}$$
2. We denote by $\hat{P}_s = \hat{P}_s(\mathscr{M},\mathbf{X},\text{pen})$ any element of $\mathscr{M}$ that satisfies
$$\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b},\hat{P}_s\right) \leq \inf_{Q \in \mathscr{M}} \sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b},Q\right) + \iota, \tag{4.5}$$
where $\iota$ is any fixed constant in (0,7671].

In order to evaluate the performance of our estimator we use the Hellinger distance $h$ defined as follows. For two probability distributions $P$ and $Q$ on the same measurable space,

$$h^2(P,Q) = \frac{1}{2} \int \left( \sqrt{dP/d\mu} - \sqrt{dQ/d\mu} \right)^2 d\mu,$$

where $\mu$ is any measure that dominates both $P$ and $Q$, the result being independent of $\mu$.

## 4.3 Application to finite state space hidden Markov models

### 4.3.1 The framework

Let $Y_1,Y_2,\ldots,Y_N$ be random variables taking values in a measurable space $(\mathscr{Y},\mathcal{Y})$. Let $L$ be $\{2,3,\ldots,\lfloor N/2 \rfloor\}$ and $n$ be the integer given by $n = N+1-L$. We define the new random variables

$$X_i = (Y_i,Y_{i+1},\ldots,Y_{i+L-1}), i = 1,\ldots,n, \tag{4.6}$$

taking values in the measurable space $(\mathscr{X},\mathcal{X}) = \left(\mathscr{Y}^L,\mathcal{Y}^{\otimes L}\right)$. We denote by $\mathscr{P}_Y$ the class of probability distributions on $(\mathscr{Y},\mathcal{Y})$. Following the discussion in the introduction we might make the following assumption.

**Assumption 4.1.** *Let $(Y_i,H_i)_i$ be a finite state space HMM with parameters $(K^*,w^*,Q^*,F^*)$ such that $Q^*$ is irreducible and aperiodic.*

Under this assumption $Q^*$ has only one invariant distribution $\pi^*$ and we define the distribution

$$P^* = \sum_{1\leq k_1,\dots,k_L\leq K} \pi^*_{k_1} Q^*_{k_1,k_2}\dots Q^*_{K_{L-1},k_L} \bigotimes_{l=1}^{L} F^*_{k_l}. \tag{4.7}$$

We do not have identically distributed observations however the distribution $P_i$ converges exponentially fast to $P^*$. For $K\geq 2$ and subsets $\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K$ of $\mathscr{P}_Y$, we denote by $\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$ the set of distributions defined by

$$\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right) := \left\{ P_{w,Q,F};\ \begin{array}{c}\forall k\in[K], w\in\mathcal{W}_K,\\ Q\in\mathcal{T}_K, F_k\in\overline{\mathscr{F}}_k\end{array} \right\} \subset \mathscr{P}_X,$$

where $P_{w,Q,F}$ is given by (4.1), $\mathcal{W}_K = \left\{w\in[0,1]^K; w_1+\dots+w_K=1\right\}$ and

$$\mathcal{T}_K = \left\{ Q\in[0,1]^{K\times K}; \sum_{j=1}^{K} Q_{ij} = 1, \forall i\in\{1,\dots,K\} \right\}.$$

Let $\Lambda$ be a countable set and $\overline{\mathscr{F}}_\lambda\subset\mathscr{P}_Y$ for $\lambda\in\Lambda$. Let $\nu$ be a $\sigma$-finite positive measure on $(\mathscr{Y},\mathcal{Y})$ and we denote by $\mu$ the associated $\sigma$-finite measure on $(\mathscr{X},\mathcal{X})$ given by $\mu:=\nu^{\otimes L}$. Let $\Theta$ be a subset of $\bigcup_{K\geq 2}\{K\}\times\Lambda^K$.

**Assumption 4.2.** *We dispose of countable sets $(\mathcal{F}_\lambda)_{\lambda\in\Lambda}$ of probability density functions (with respect to $\nu$) such that*

1. *for all $\lambda\in\Lambda$, the set of distributions $\mathscr{F}_\lambda := \{f\cdot\nu; f\in\mathcal{F}_\lambda\in\mathcal{F}_\lambda\}$ is an $\epsilon$-net of $\overline{\mathscr{F}}_\lambda$ with respect to the Hellinger distance;*
2. *for all $\theta=(K,\lambda_1,\dots,\lambda_K)\in\Theta$ and all $k_1,\dots,k_L\in[K]$ the class of functions*

$$\mathcal{F}_{\lambda_{k_1},\dots,\lambda_{k_L}} = \left\{ \mathbf{x}\in\mathscr{Y}^L \mapsto f_1(x_1)\dots f_L(x_L); f_l\in\mathcal{F}_{\lambda_{k_l}}, \forall l\in[L] \right\}$$

*is VC-subgraph with VC-index not larger than $V_{\lambda_{k_1},\dots,\lambda_{k_L}}$.*

*Then we write*

$$\overline{V}_\theta := \sum_{1\leq k_1,\dots,k_L\leq K} V_{\lambda_{k_1},\dots,\lambda_{k_L}}.$$

For $\theta=(K,\lambda_1,\dots,\lambda_K)\in\Theta$, let $\overline{\mathscr{M}}_\theta$ be a non-empty subset of $\mathscr{H}_\theta := \mathscr{H}\left(K,\overline{\mathscr{F}}_{\lambda_1},\dots,\overline{\mathscr{F}}_{\lambda_K}\right)$. We use model selection to build an estimator over the collection of models $\left(\overline{\mathscr{M}}\right)_{\theta\in\Theta}$. We denote their union by $\overline{\mathscr{M}}=\bigcup_{\theta\in\Theta}\overline{\mathscr{M}}_\theta$. To perform the procedure described in Section 4.2 we need a countable approximation of $\overline{\mathscr{M}}$.

For $\delta\in(0,1/K]$ we define the sets $\mathcal{W}_{\delta,K} := \mathcal{W}_K\cap([\delta,1]\cap\mathbb{Q})^K$ and $\mathcal{T}_{\delta,K} := \mathcal{T}_K\cap([\delta,1]\cap\mathbb{Q})^{K\times K}$. Let $\delta:\Theta\to(0,1]$ be such that $\delta(\theta)\in(0,1/K]$ for all $\theta=(K,\lambda_1,\dots,\lambda_K)\in\Theta$. We define

$$\mathscr{H}_{\theta,\delta} := \left\{ P_{w,Q,f}; w\in\mathcal{W}_{\delta(\theta),K}, Q\in\mathcal{T}_{\delta(\theta),K}, f_k\in\mathscr{F}_{\lambda_k}, \forall i\in[K] \right\},$$

where the sets $(\mathscr{F}_{\lambda_k})_{1\leq k\leq K}$ are given in Assumption 4.2. We define $\mathscr{M}_{\theta,\delta}$ as the following countable set of distributions

$$\mathscr{M}_{\theta,\delta} := \left\{ P_{w,Q,F}\in\mathscr{H}_{\theta,\delta}; \exists P_{w',Q',F'}\in\overline{\mathscr{M}}_\theta, \begin{array}{c} h^2\left(Q_{k\cdot},Q'_{k\cdot}\right)\leq(K-1)\delta\\ h\left(F_k,F'_k\right)\leq\epsilon,\forall k\in[K],\\ h^2\left(w,w'\right)\leq(K-1)\delta, \end{array} \right\}, \tag{4.8}$$

which is a good approximation of $\overline{\mathscr{M}}_\theta$ for small values of $\delta$ and $\epsilon$ and we take

$$\mathscr{M} := \bigcup_{\theta\in\Theta} \mathscr{M}_{\theta,\delta}. \tag{4.9}$$

### 4.3.2 General result of model selection

Let $\Delta$ be a function $\Theta \to \mathbb{R}_+$ satisfying $\sum_{\theta\in\Theta} e^{-\Delta(\theta)} \leq 1$. The following result is proven in Section 4.B.1.

**Theorem 4.1.** *Let $Y_1,\ldots,Y_N$ be random variables on $(\mathscr{Y},\mathcal{Y})$ and $s$ be in $\{0,1,\ldots,s_{\max}\}$. Under Assumption 4.2, let $\hat{P}_s = \hat{P}_s(\mathscr{M},\mathbf{X},pen)$ be the estimator given by (4.5) with*

$$\delta(\theta) = \frac{\overline{V}_\theta}{n(s,1)(K-1)} \wedge \frac{1}{K}, \tag{4.10}$$

*and*

$$pen(Q) = \kappa \inf_{\theta\in\Theta | Q\in\mathscr{M}_\theta} \left[ 837\left(1+\log\left(\frac{Kn(s,1)}{\overline{V}_\theta \wedge n(s,1)}\right)\right) + \Delta(\theta)\right]. \tag{4.11}$$

*There exists a positive constant $C$ such that*

$$C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right) + L\epsilon^2 + \inf_{\theta\in\Theta}\left[h^2\left(\overline{P},\overline{\mathscr{M}}_\theta\right) + \frac{(s+1)}{n}\left(L\overline{V}_\theta \log n + \Delta(\theta)\right)\right]$$
$$+ n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right). \tag{4.12}$$

*In particular, under Assumption 4.1, there exist positive constants $C(Q^*)$,$c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have*

$$C(Q^*)\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\epsilon^2 + \inf_{\theta\in\Theta}\left[h^2\left(P^*,\overline{\mathscr{M}}_\theta\right) + \frac{s}{n}\left(L\overline{V}_\theta\log n + \Delta(\theta)\right)\right]. \tag{4.13}$$

Inequality (4.12) does not require any assumption on the data. Ideally we can take $\overline{P}$ in $\overline{\mathscr{M}}_\theta$ such that most of the distributions $P_i$ lie in a small neighborhood of $\overline{P}$ so that the first term in the bound remains small compared to $L\epsilon^2 + (s+1)\left(L\overline{V}_\theta\log n + \Delta(\theta)\right)$. In the case where we cannot take $\epsilon = 0$ and the quantity $\overline{V}_\theta$ depends on it, we have to take $\epsilon$ going to 0 with $n$ in a way that balances those two terms.

Under Assumption 4.1 the term $n^{-1}\sum_{i=1}^n h^2(P^*,P_i)$ is negligible and a good choice of $s$ guarantees the term $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ to be negligible as well with respect to the main term $(s+1)\left(L\overline{V}_\theta\log n + \Delta(\theta)\right)$. We define $\mathscr{H}^* = \bigcup_{\theta\in\Theta}\mathscr{H}^*_\theta$ with

$$\mathscr{H}^*_\theta := \left\{ P_{w,Q,F} \in \mathscr{H}(\theta)\,;\ \begin{array}{c} Q \text{ irreducible}, \\ Q \text{ aperiodic}, \\ \text{and } w = Qw \end{array} \right\}. \tag{4.14}$$

If $P^* \in \overline{\mathscr{M}} \cap \mathscr{H}^*$, for $s$ of order $\log^2 n$ and $n$ large enough we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \frac{\log^2 n}{n} \inf_{\substack{\theta\in\Theta \\ P^*\in\overline{\mathscr{M}}_\theta}} \left\{L\overline{V}_\theta\log n + \Delta(\theta)\right\}.$$

As long as $\Delta(\theta^*)$ is of the same order as $L\overline{V}_{\theta^*}\log n$ or smaller, we obtain the same rate as when we only consider $\overline{\mathscr{M}}_{\theta^*}$. We can obtain a better power of $\log n$ if we know $c(Q^*)$.

In addition of having good performances our estimator possesses properties of robustness. In order to illustrate them we consider the following situation. Let $Z_1,\ldots,Z_N$ be random variables with any distributions and $E_1,\ldots,E_N$ be Bernoulli random variables such that for all $i \in [n]$,

$$Y_i = E_iY'_i + (1-E_i)Z_i,$$

where $\mathbf{Y}'$ satisfies Assumption 4.1. The following result is proven in Section 4.B.2.

**Corollary 4.1.** *If $E_1,Z_1,\dots,E_N,Z_N$ and $\mathbf{Y}'$ are mutually independent, there exist positive constant $C(Q^*)$,$c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \frac{L}{N}\sum_{i=1}^{N}(1-p_i) + L\epsilon^2 + \inf_{\theta\in\Theta}\left[h^2\left(P^*,\overline{\mathscr{M}}_\theta\right) + (s+1)\left(L\overline{V}_\theta \log n + \Delta(\theta)\right)\right],$$

*where $p_i = \mathbb{P}(E_i = 1)$ for all $i \in [N]$.*

One can see that our deviation bound is not significantly worse as long as the average proportion of contamination $\frac{L}{N}\sum_{i=1}^{N}(1-p_i)$ remains small compared to the other terms. One would typically look at the following situation. We assume that there exists $\overline{\theta} \in \Theta$ such that $P^* \in \overline{\mathscr{M}}_{\overline{\theta}}$. For Hüber's contamination model, i.e. $p_i = 1 - \alpha_{cont}$ for all $i$, we get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\alpha_{cont} + \frac{s}{n}\left(L\overline{V}_{\overline{\theta}}\log n + \Delta(\overline{\theta})\right), \tag{4.15}$$

for $s \geq c(Q^*)\log n$. The bound on the convergence rate is not deteriorated as long as the contamination rate $\alpha_{cont}$ is small compared to $\frac{s}{n}\left(L\overline{V}_{\overline{\theta}}\log n + \Delta(\overline{\theta})\right)$. We can also consider the situation $\mathbb{P}(E_i = 0) = \mathbb{1}_{i\in I}$ for some subset $I \subset [N]$. We get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \frac{L|I|}{N} + \frac{s}{n}\left(L\overline{V}_{\overline{\theta}}\log n + \Delta(\overline{\theta})\right), \tag{4.16}$$

for $s \geq c(Q^*)\log n$. As before, our bound on the convergence rate is not deteriorated as long as the proportion of outliers $|I|/N$ is small compared to $\frac{s}{n}\left(L\overline{V}_{\overline{\theta}}\log n + \Delta(\overline{\theta})\right)$.

### 4.3.3 Selection of the order

We focus on the specific situation where $\Lambda$ is a single set $\{\lambda\}$. Let $\mathscr{K}$ be a subset of $\{2,\dots,n\}$ and

$$\Theta = \bigcup_{K\in\mathscr{K}} \{K\} \times \{\lambda\}^K.$$

We consider the following case. There is a countable set $\mathcal{F}$ of probability density functions with respect to $\nu$ such that $\mathscr{F} = \{f\cdot\nu; f \in \mathcal{F}\}$ is an $\epsilon$-net of $\overline{\mathscr{F}}_\lambda$ and the class of functions

$$\mathcal{F}^{\otimes L} = \{\mathbf{x} \mapsto f_1(x_1)\dots f_L(x_L); f_l \in \mathcal{F}, \forall l \in [L]\},$$

is VC-subgraph with VC-index not larger than $V$. Then we satisfy Assumption 4.2 and we have $\overline{V}_K = K^L V$. We take $\Theta = \bigcup_{K\in\mathscr{K}}\{K\}\times\{\lambda\}^K$ and drop the dependency on $\lambda$ in the notation. The next result is a consequence of Theorem 4.1.

**Corollary 4.2.** *For pen given by (4.11) with $\Delta(K) = K^L$ and $\delta$ given by (4.10), there exists a positive constant $C$ such that*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-1}\sum_{i=1}^{n} h^2\left(P_i,\overline{P}\right) + L\epsilon^2 + \sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ + \inf_{K\in\mathscr{K}}\left[h^2\left(\overline{P},\overline{\mathscr{M}}_K\right) + (s+1)L^2K^LV\frac{\log n}{n}\right]. \end{aligned} \tag{4.17}$$

*In particular under Assumption 4.1 there exists positive constants $C(Q^*)$,$c(Q^*)$ such that for $s \geq c(Q^*)\log n$,*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\epsilon^2 + \inf_{K\in\mathscr{K}}\left[h^2\left(P^*,\overline{\mathscr{M}}_K\right) + L^2K^LV\frac{s\log n}{n}\right].$$

Inequality (4.17) does not require any assumption on the data. We can see that the model selection procedure allows to recover the performance we would get if we knew the model $\overline{\mathscr{M}}_K$ realizing the best compromise between the distance $h^2\left(P^*,\overline{\mathscr{M}}_K\right)$ and the dimension term $(s+1)L^2K^LVn^{-1}\log n$. If $P^*$ belongs to $\mathscr{M}$, we denote by $K^*$ the smallest integer $K$ in $\mathscr{K}$ such that $P^*\in\overline{\mathscr{M}}_K$. We get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right]\leq L(K^*)^LV\frac{s\log n}{n},$$

for $s\geq c(Q^*)\log n$ and $\epsilon=0$. In particular, if we denote by $\hat{K}$ the smallest integer $K$ such that $\hat{P}_s\in\overline{\mathscr{M}}_K$, we have

$$\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right]\geq h^2\left(P^*,\hat{P}_s\right)\mathbb{P}\left(\hat{K}<K^*-1\right).$$

Since $h^2\left(P^*,\overline{\mathscr{M}}_{K^*-1}\right)>0$, we can deduce a bound on the probability of underestimating the true order. Considering the problem of overestimation would require more work that we did not do for lack of time.

We consider the following example. We take $\mathscr{Y}=\mathbb{N}$ and $\overline{\mathscr{F}}_P=\{Poisson(\lambda);\lambda>0\}$ where $Poisson(\lambda)$ is the Poisson distribution with parameter $\lambda$. It is defined by its density given by $p_\lambda(k)=\frac{\lambda^ke^{-\lambda}}{k!}$. We take $\overline{\mathscr{M}}_K=\mathscr{H}(K,\overline{\mathscr{F}}_P,\ldots,\overline{\mathscr{F}}_P)$ for $K\in\mathscr{K}=\{2,\ldots,N\}$. Let $\mathcal{P}(\mathbb{N})$ denote the power set of $\mathbb{N}$. For $\mathscr{F}_P=\{Poisson(\lambda);\lambda\in\mathbb{Q}\cap(0,\infty)\}$ we satisfy Assumption 4.2 with $\epsilon=0$.

**Theorem 4.2.** *Let $Y_1,\ldots,Y_N$ be random variables on $(\mathbb{N},\mathcal{P}(\mathbb{N}))$ and $s$ be in $\{0,1,\ldots,s_{\max}\}$. Let $\hat{P}_s=\hat{P}_s(\mathscr{M},\mathbf{X},pen)$ be the estimator given by (4.5) with $\delta$,pen and $\Delta$ given by (4.10), (4.11) and $\Delta(K)=K^L$. There exists a positive constant $C$ such that for all $\overline{P}\in\mathscr{P}_X$,*

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right]\leq n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right)+n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$$
$$+\inf_{2\leq K\leq n}\left[h^2\left(\overline{P},\overline{\mathscr{M}}_K\right)+(s+1)L^2K^L\log n\right].$$

*In particular, under Assumption 4.1, there exist positive constants $C(Q^*)$,$c(Q^*)$ such that for $s\geq c(Q^*)\log n$,*

$$C(Q^*)\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(P^*,\hat{P}_s\right)\right]\leq\inf_{2\leq K\leq n}\left[h^2\left(P^*,\overline{\mathscr{M}}_K\right)+(s+1)L^2K^L\log n\right].$$

This result is a consequence of Corollary 4.2 and Proposition 3.1. The comments made earlier apply here as well.

### 4.3.4 Selection of the emission models

We consider another specific situation where the order is fixed and we want to select the emission models. Let $K$ be a fixed integer larger than 1. Let $\mathcal{L}$ be a subset of $\Lambda^K$ and we take $\Theta=\{K\}\times\mathcal{L}$. The next result is a consequence of Theorem 4.1.

**Corollary 4.3.** *Let $\Delta$ be a mapping $\mathcal{L}\to\mathbb{R}_+$ such that $\sum_{\lambda\in\mathscr{L}}e^{-\Delta(\lambda)}\leq 1$. For pens given by (4.11) with $\delta$ given by (4.10), there exists a positive constant $C$ such that*

$$C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right]\leq n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right)+L\epsilon^2+\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$$
$$+\inf_{\lambda\in\mathcal{L}}\left\{h^2\left(\overline{P},\overline{\mathscr{M}}_\lambda\right)+(s+1)L\overline{V}_\lambda\frac{\log n}{n}\right\}.$$

*In particular under Assumption 4.1 there exists positive constants $C(Q^*)$,$c(Q^*)$ such that for $s \geq c(Q^*)\log n$, we get*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\epsilon^2 + \inf_{\lambda\in\mathcal{L}}\left[h^2\left(P^*,\overline{\mathscr{M}}_\lambda\right) + L\overline{V}_\lambda\frac{s\log n}{n}\right].$$

Inequality (4.17) does not require any assumption on the data. We can see that the model selection procedure allows to recover the performance we would get if we knew the model $\overline{\mathscr{M}}_\lambda$ realizing the best compromise between the distance $h^2\left(P^*,\overline{\mathscr{M}}_\lambda\right)$ and the dimension term $(s+1)L\overline{V}_\lambda n^{-1}\log n$. If $P^*$ belongs to $\overline{\mathscr{M}}$, we get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\overline{V}^*\frac{s\log n}{n},$$

for $s \geq c(Q^*)\log n$ and $\overline{V}^* = \min\{V_\lambda;\lambda\in\Lambda \text{ and } P^*\in\overline{\mathscr{M}}_\lambda\}$. We can apply this result for the estimation of hidden Markov models with sparse multivariate normal emission densities.

Let $d$ be a positive integer and $\mathrm{Cov}_{+*}(d)$ be the set of $d\times d$ symmetric and positive-definite matrices. We define $T^+ = \{(i,j);1\leq i\leq j\leq d\}$ and

$$\mathscr{T}^+ = \left\{A\subset T^+;\{(1,1),(2,2),\ldots,(d,d)\}\subset A\right\}.$$

For $A$ in $\mathscr{T}^+$, we denote by $|A|$ its cardinal and we denote by $\mathrm{Cov}_{+*}(d,A)$ the subset of $\mathrm{Cov}_{+*}(d)$ given by

$$\mathrm{Cov}_{+*}(d,A) = \left\{C\in\mathrm{Cov}_{+*}(d);\forall(i,j)\in T^+,(i,j)\notin A\Leftrightarrow C_{i,j}=0\right\}.$$

We denote by $\mathcal{G}_d(A)$ the set of probability density functions (with respect to the Lebesgue measure) given by

$$\mathcal{G}_d(A) = \left\{g_{z,\Sigma}: x\in\mathbb{R}^d\mapsto\frac{\exp\left(-\frac{1}{2}(x-z)^T\Sigma^{-1}(x-z)\right)}{(2\pi)^{d/2}\det(\Sigma)^{1/2}};z\in\mathbb{R}^d,\Sigma^{-1}\in\mathrm{Cov}_{+*}(d,A)\right\},$$

and by $\mathscr{G}_d(A)$ the associated set of probability distribution. We take $\Lambda = \mathscr{T}^+$ and $\Theta = \{K\}\times\Lambda^K$. We take $\overline{\mathscr{M}}_{\mathbf{A}} = \mathscr{H}\left(K,\mathscr{G}_d(A_1),\ldots,\mathscr{G}_d(A_K)\right)$ for $\mathbf{A} = (A_1,\ldots,A_K)\in\Lambda^K$. We satisfy Assumption 4.2 with $\epsilon = 0$,

$$\mathcal{G}_{d,\mathbb{Q}}(A) = \left\{g_{z,\Sigma};z\in\mathbb{Q}^d,\Sigma^{-1}\in\mathrm{Cov}_{+*}(d,A)\cap\mathbb{Q}^{d\times d}\right\},$$

and

$$\overline{V}_{\mathbf{A}} = K^L(3+dL) + LK^{L-1}(|A_1|+\cdots+|A_K|).$$

Let $\sigma(\mathbb{R}^d)$ be the Borel $\sigma$-algebra on $\mathbb{R}^d$. The following result is proven in Section 4.B.4.

**Theorem 4.3.** *Let $Y_1,\ldots,Y_N$ be random variables on $\left(\mathbb{R}^d,\sigma\left(\mathbb{R}^d\right)\right)$ with $d\leq 1+2LK^{L-1}N^{LK^{L-1}}$ and $s$ be in $\{0,1,\ldots,s_{\max}\}$. Let $\hat{P}_s = \hat{P}_s(\mathscr{M},\mathbf{X},pen)$ be the estimator given by (4.5) with $\delta$,$pen$ and $\Delta$ given by (4.10), (4.11) and*

$$\Delta(A_1,\ldots,A_K) := dLK^L + LK^{L-1}\log n\left((|A_1|-d)+\cdots+(|A_K|-d)\right),$$

*respectively. There exists a positive constant $C$ such that for all $\overline{P}\in\mathscr{P}_X$,*

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-1}\sum_{i=1}^n h^2\left(P_i,\overline{P}\right) + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$$
$$+\inf_{\mathbf{A}\in(\mathscr{T}^+)^K}\left[h^2\left(\overline{P},\overline{\mathscr{M}}_{\mathbf{A}}\right) + (s+1)\log n\left(dL^2K^L + LK^{L-1}(|A_1|+\cdots+|A_K|)\right)\right].$$

*In particular, under Assumption 4.1, there exist positive constants $C(Q^*)$,$c(Q^*)$ such that for $s \geq c(Q^*) \log n$,*

$$C(Q^*)\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq \inf_{\mathbf{A}\in(\mathscr{T}^+)^K}\left[h^2\left(P^*,\mathscr{M}_{\mathbf{A}}\right) + (s+1)\log n\left(dL^2K^L + LK^{L-1}(|A_1| + \cdots + |A_K|)\right)\right]. \tag{4.18}$$

We can avoid a quadratic dependence on $d$ and only have a linear one when there is a good approximation of the true distribution that is sparse, i.e. if $P^* \in \mathscr{M}_{\mathbf{A}}$ with $|A_1|,\ldots,|A_K|$ of order $d$.

## 4.A General results

Let $(\mathscr{M}_\theta)_{\theta\in\Theta}$ be countable subsets of $\mathscr{P}_X$, with $\Theta$ countable set. Let $\Delta:\Theta\to\mathbb{R}_+$ be such that

$$\sum_{\theta\in\Theta} e^{-\Delta(\theta)} \leq 1.$$

We assume we have the following bound on the $\rho$-dimension function (see Definition 4 [11]). There exist $D:\Theta\times\mathbb{N}^*\to\mathbb{R}_+$ and $K\geq 0$ such that

$$D^{\mathscr{M}_\theta}\left(\mathbf{P},\overline{\mathbf{P}}\right) \leq D(\theta,n)+TD(\theta',n), \tag{4.19}$$

for all $\mathbf{P}=\bigotimes_{i=1}^n P_i\in\mathscr{P}^{\otimes n}$, all $\overline{\mathbf{P}}=\bigotimes_{i=1}^n \overline{P}_i\in\mathscr{M}_{\theta'}^{\otimes n}$ and all $n\geq 1$. We assume $D$ is nondecreasing in $n$ for all $\theta$ in $\Theta$. For $s$ in $\{0,1,\ldots,s_{\max}\}$, we take the penalty function given by

$$\mathrm{pen}(Q)=\kappa \inf_{\theta\in\Theta|Q\in\mathscr{M}_\theta}\left[\frac{D(\theta,n(s,1))}{4.7}+\Delta(\theta)\right],$$

for all $Q\in\mathscr{M}=\bigcup_{\theta\in\Theta}\mathscr{M}_\theta$ with $\kappa=8\times 35\sqrt{2}+74$.

**Theorem 4.4.** *For any random variables $X_1,\ldots,X_n$ on $(\mathscr{X},\mathcal{X})$, the estimator $\hat{P}_s=\hat{P}_s(\mathscr{M},\mathbf{X},pen)$ satisfies*

$$\mathbb{E}_{\mathbf{P}^*}\left[\sum_{i=1}^n h^2\left(P_i,\hat{P}_s\right)\right] \leq \inf_{\theta\in\Theta}\left[c_1\inf_{Q\in\mathscr{M}_\theta}\sum_{i=1}^n h^2(P_i,Q)+c_2(s+1)\left(\frac{(T+1)D(\theta,n(s,1))}{4.7}+\Delta(\theta)+4\right)\right] + c_2\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right),$$

*with $c_1=304$ and $c_2=30084$.*

### 4.A.1 Proof of Theorem 4.4

Let $b\in[s+1]$ and $\hat{P}_{s,b}$ be the estimator given by (4.4). From Theorem 2 of Baraud & Birgé [11], we have

$$\mathbf{P}^{ind}_{s,b}\left(\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)},\hat{P}_{s,b}\right)\leq \inf_{\theta\in\Theta}\left[\gamma\inf_{Q\in\mathscr{M}_\theta}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)},Q\right) + \frac{4\kappa}{a_1}\left(\frac{(T+1)D(\theta,n(s,b))}{4.7}+\Delta(\theta)+1.49\right)\right]+\frac{4\kappa}{a_1}\xi\right)\geq 1-e^{-\xi},\forall\xi>0,$$

where $\gamma$ and $\kappa$ are given in [11] and satisfy $\gamma\leq 150$ and $\frac{4\kappa}{a_1}\leq 5014$ (see proof of Theorem 1 [13], page 32). Applying Lemma (chapter hmm), we get

$$\mathbb{E}_{\mathbf{P}^*_{s,b}}\left[\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)},\hat{P}_{s,b}\right)\right] \leq \inf_{\theta\in\Theta}\left[\gamma\inf_{Q\in\mathscr{M}_\theta}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)},Q\right) + \frac{4\kappa}{a_1}\left(\frac{(T+1)D(\theta,n(s,b))}{4.7}+\Delta(\theta)\right)\right] + \frac{4\kappa}{a_1}\left[3.49+\mathbf{K}\left(\mathbf{P}^*||\mathbf{P}^{ind}\right)\right].$$

From the definition of $\hat{P}_s$, we have

$$\begin{aligned}
\sum_{i=1}^{n} h^2\left(P_i, \hat{P}_s\right) &= \sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_s\right) \\
&\leq 2\sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right) + 2\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b}, \hat{P}_s\right) \\
&\leq 2\sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right) + 2\inf_{Q\in\mathscr{M}}\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b}, Q\right) + 2\iota \\
&\leq 6\sum_{b=1}^{s+1}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right) + 4\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + 2\iota.
\end{aligned}$$

Since $\iota \leq 7671 \leq 1.53 \times 5014$, $4\kappa/a_1 \leq 5014$ and $\gamma \leq 150$ (see proof of Theorem 1 in [13]), we get

$$\begin{aligned}
\mathbb{E}_{\mathbf{P}^*}\left[\sum_{i=1}^{n} h^2\left(P_i, \hat{P}_s\right)\right] &\leq 6\sum_{b=1}^{s+1}\mathbb{E}_{\mathbf{P}^*_{s,b}}\left[\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_{s,b}\right)\right] + 4\inf_{Q\in\mathscr{M}}\sum_{i=1}^{N} h^2\left(P_i, Q\right) + 2\iota \\
&\leq 6\sum_{b=1}^{s+1}\inf_{\theta\in\Theta}\left[\gamma\inf_{Q\in\mathscr{M}_\theta}\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, Q\right) + \frac{4\kappa}{a_1}\left(\frac{(T+1)D_{n(s,b)}(\theta)}{4.7} + \Delta(\theta)\right)\right] \\
&+ \frac{24\kappa}{a_1}\left[3.49(s+1) + \sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right] + 4\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + 2\iota \\
&\leq \inf_{\theta\in\Theta}\left[(6\gamma+4)\inf_{Q\in\mathscr{M}_\theta}\sum_{i=1}^{n} h^2\left(P_i, Q\right)\right. \\
&\left. + \frac{24\kappa(s+1)}{a_1}\left(\frac{(T+1)D_{n(s,1)}(\theta)}{4.7} + \Delta(\theta) + 3.49\right)\right] + \frac{24\kappa}{a_1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) + 2\iota \\
&\leq \inf_{\theta\in\Theta}\left[304\inf_{Q\in\mathscr{M}_\theta}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + 30084(s+1)\left(\frac{(T+1)D(\theta,n(s,1))}{4.7} + \Delta(\theta) + 4\right)\right] \\
&+ 30084\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right).
\end{aligned}$$

# 4.B Application to HMMs

## 4.B.1 Proof of Theorem 4.1

From Proposition 5 in (chapter hmm) and our choice of $\delta$, we satisfy (4.19) with $T = 0$ and

$$D\left(\theta,n(s,1)\right) = 3930L\overline{V}_\theta\left[1 + \log\left(\frac{Kn(s,1)}{\overline{V}_\theta \wedge n(s,1)}\right)\right].$$

From Proposition 6 and (??), we have

$$\begin{aligned}
\inf_{Q\in\mathscr{M}_{\theta,\delta}}\sum_{i=1}^{n} h^2\left(P_i, Q\right) &\leq 2\inf_{Q\in\overline{\mathscr{M}}_\theta}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + 2(K-1)L\delta(\theta) + 2L\epsilon^2 \\
&\leq 2\inf_{Q\in\overline{\mathscr{M}}_\theta}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + 2(s+1)L\overline{V}_\theta + 2nL\epsilon^2,
\end{aligned}$$

and

$$h^2\left(\overline{P},\hat{P}_s\right) \leq \frac{2}{n}\sum_{i=1}^{n} h^2\left(\overline{P},P_i\right) + \frac{2}{n}\sum_{i=1}^{n} h^2\left(P_i,\hat{P}_s\right)$$
$$\sum_{i=1}^{n} h^2\left(P_i,\hat{P}_s\right) \leq 2h^2\left(\overline{P},\hat{P}_s\right) + \frac{2}{n}\sum_{i=1}^{n} h^2\left(\overline{P},P_i\right).$$

From Theorem 4.4, there exists a positive constant $C$ such that

$$C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-1}\sum_{i=1}^{n} h^2\left(P_i,\overline{P}\right) + \inf_{\theta\in\Theta}\left[h^2\left(\overline{P},\overline{\mathscr{M}}_\theta\right) + nL\epsilon^2 + (s+1)\left(L\overline{V}_\theta \log n + \Delta(\theta)\right)\right] + \sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right).$$

This proves (4.12) and we turn to the second inequality. With Lemma ?? (chapter HMM), we have

$$h^2(P^*,\hat{P}_s) \leq \frac{2}{n}\sum_{i=1}^{n} h^2(P_i,P^*) + \frac{2}{n}\sum_{i=1}^{n}(P_i,\hat{P}_s),$$

and

$$n^{-1}\sum_{i=1}^{n} h^2(P_i,P^*) \leq \frac{2C(Q^*)}{n(e^{r(Q^*)}-1)}.$$

It leads to

$$C(Q^*)\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \inf_{\theta\in\Theta}\left[h^2\left(P^*,\overline{\mathscr{M}}_\theta\right) + \frac{(s+1)}{n}\left(L\overline{V}_\theta \log n + \Delta(\theta)\right)\right] + e^{-r(Q^*)s},$$

and we obtain (4.13) with $s$ large enough.

### 4.B.2 Proof of Corollary 4.1

We have

$$\mathbb{P}\left(X_i = (Y'_i,\ldots,Y'_{i+L-1})\right) \geq \mathbb{P}(E_i = \cdots = E_{i+L-1} = 1) = p_i p_{i+1}\ldots p_{i+L-1},$$

and with the convexity of the squared Hellinger distance

$$\begin{aligned} h^2(P_i,P^*) &\leq p_i p_{i+1}\ldots p_{i+L-1}h^2(P'_i,P^*) + (1 - p_i p_{i+1}\ldots p_{i+L-1}) \\ &\leq h^2(P'_i,P^*) + (1-p_i) + (1-p_{i+1}) + \cdots + (1-p_{i+L-1}), \end{aligned}$$

where $P'_i = \mathcal{L}(Y'_i,\ldots,Y'_{i+L-1})$. One can check that $n \geq 1 + N/2$ with our condition on $L$. With Theorem ?? and Lemma ??, we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \frac{L}{N}\sum_{i=1}^{N}(1-p_i) + L\epsilon^2 + \inf_{\theta\in\Theta}\left[h^2\left(P^*,\overline{\mathscr{M}}_\theta\right) + s\frac{\log n}{n}\left(L\overline{V}_\theta \log n + \Delta(\theta)\right)\right],$$

for our condition on $s$.

### 4.B.3 Proof of Theorem 4.2

One can check that we have $\overline{\mathcal{F}}_P = \mathcal{E}\left((0,\infty), \log, id_{\mathbb{N}}, 1, B\right)$ with $B(k) = -\log(k!)$.

### 4.B.4 Proof of Theorem 4.3

One can check that we have $\mathcal{G}_d(A) = \mathcal{E}\left(\mathbb{R}^d \times \mathrm{Cov}_{+*}(d,A), \eta, T, d+|A|, 0\right)$ with

$$T(x) = \left(x, (x_i^2)_{1\leq i\leq d}, (x_i x_j)_{\substack{i<j \\ (i,j)\in A}}\right)$$
$$\eta(z,C) = \left(Cz, -\frac{1}{2}(C_{ii})_{1\leq i\leq d}, -(C_{ij})_{\substack{i<j \\ (i,j)\in A}}\right)$$

We define $\mathscr{T}_j^+$ by $\mathscr{T}_j^+ = \{A \in \mathscr{T}^+; |A| = j\}$, for $j$ in $\{d, \ldots, d(d+1)/2\}$. We have $|\mathscr{T}_{d+k}^+| = \binom{\frac{d(d-1)}{2}}{k}$. One can check that we have

$$\begin{aligned}
\sum_{A\in\mathscr{T}^+} e^{-\Delta(A_1,\ldots,A_K)} &= \sum_{A\in\mathscr{T}^+} e^{-dLK^L - LK^{L-1}\log n((|A_1|-d)+\cdots+(|A_K|-d))} \\
&= e^{-dLK^L}\left(\sum_{A\in\mathscr{T}^+} e^{-LK^{L-1}\log n(|A|-d)}\right)^K \\
&= e^{-dLK^L}\left(\sum_{j=d}^{d(d+1)/2}\sum_{A\in\mathscr{T}_j^+} e^{-LK^{L-1}\log n(j-d)}\right)^K \\
&= e^{-dLK^L}\left(\sum_{j=0}^{d(d-1)/2}\binom{d(d-1)/2}{j} n^{-LK^{L-1}j}\right)^K \\
&= e^{-dLK^L}\left(1+n^{-LK^{L-1}}\right)^{Kd(d-1)/2} \\
&\leq e^{-dLK^L} e^{\frac{Kd(d-1)}{2}n^{-LK^{L-1}}} \leq 1,
\end{aligned}$$

for $d \leq 1 + 2LK^{L-1}N^{LK^{L-1}}$.

### 4.B.5 Exponential families

Let $d$ be a positive integer and $\eta : \overline{\Phi} \to \mathbb{R}^d$ be a function over a non-empty set $\overline{\Phi}$. Let $T : \mathscr{X} \to \mathbb{R}^d$ and $B : \mathscr{X} \to \mathbb{R}$ be measurable functions such that

$$\int_{\mathscr{X}} e^{\langle\eta(\phi),T(x)\rangle + B(x)}\nu(dx) < \infty, \forall\phi \in \overline{\Phi},$$

we denote by $\mathcal{E}\left(\overline{\Phi},\eta,T,d,B\right)$ the exponential family defined by

$$\mathcal{E}\left(\overline{\Phi},\eta,T,d,B\right) := \left\{f_\phi : x \mapsto e^{\langle\eta(\phi),T(x)\rangle + A(\phi) + B(x)}; \phi \in \overline{\Phi}\right\},$$

where

$$A(\phi) := -\log\left(\int_{\mathscr{X}} e^{\langle\eta(\phi),T(x)\rangle + B(x)}\nu(dx)\right).$$

It is a set of probability density functions with respect to $\nu$. We consider the following situation.

**Assumption 4.3.** *For all $\lambda \in \Lambda$,*

- *$\overline{\mathscr{F}}_\lambda$ is of the form*

$$\overline{\mathscr{F}}_\lambda = \left\{q\cdot\nu; q \in \mathcal{E}\left(\overline{\Phi}_\lambda,\eta_\lambda,T_\lambda,d_\lambda,B_\lambda\right)\right\}, \tag{4.20}$$

- $\Phi_\lambda$ *is a countable subset of* $\overline{\Phi}_\lambda$ *such that* $\mathscr{F}_\lambda = \{q \cdot \nu; q \in \mathcal{F}_\lambda\}$ *is a dense subset of* $\overline{\mathscr{F}}_\lambda$, *with* $\mathcal{F}_\lambda := \mathcal{E}\left(\Phi_\lambda, \eta_{\lambda|\Phi_\lambda}, T_\lambda, d_\lambda, B_\lambda\right)$.

In that situation, we satisfy Assumption 4.2 with $\epsilon = 0$ and

$$V_{\lambda_{k_1},\dots,\lambda_{i_L}} \leq 3 + \sum_{l=1}^{L} d^{\lambda}_{i_l}$$

and therefore

$$\overline{V}_\theta = 3K^L + LK^{L-1}\left(d_{\lambda_1} + \cdots + d_{\lambda_K}\right),$$

for $\theta = (K, \lambda_1, \dots, \lambda_K)$.

# Chapter 5

# General state space hidden Markov models

**Abstract**

We observe $n$ observations generated by a hidden Markov model and aim to estimate the different parameters. We do not assume the state space of the hidden Markov chain to be finite. We consider the case of univariate normal emission densities and establish the identifiability of the stationary distribution and Markov kernel given the distribution of two consecutive observations. We can approximate such a distribution by finite mixtures of multivariate normal distributions and we establish a risk bound for our estimator.

## 5.1 Introduction

Let $(\mathscr{E},\mathcal{E})$ be a measurable space. Let $\pi$ be a probability distribution on $(\mathscr{E},\mathcal{E})$. Let $Q$ be a Markov kernel on $(\mathscr{E},\mathcal{E})$. Let $\overline{\mathscr{F}}$ be a set of emission distributions on a measurable space $(\mathscr{Y},\mathcal{Y})$ of the form

$$\overline{\mathscr{F}} = \{F_h; h \in \mathscr{E}\},$$

and such that the application $h \mapsto F_h(B)$ is measurable for all $B \in \mathcal{Y}$. We say that $(Y_i,H_i)_i$ is a HMM with parameters $\left(\mathscr{E},\pi,Q,\overline{\mathscr{F}}\right)$ if $(H_i)_i$ is a Markov chain with initial distribution $\pi$ on $(\mathscr{E},\mathcal{E})$ and kernel $Q$ on $\mathscr{E} \times \mathcal{E}$, and

$$\mathcal{L}(Y_1,\ldots,Y_N|H_1,\ldots,H_N) = \bigotimes_{i=1}^{N} F_{H_i}.$$

We adopt the same strategy we considered for finite state space HMMs assuming it is possible to deduce the parameters from the distribution of consecutive observations.

Let $(Y_i,H_i)_i$ be a HMM with parameters $(\mathscr{E},\pi^*,Q^*,\overline{\mathscr{F}})$. If $\pi^*$ is invariant with respect to $Q^*$ the process $(Y_i,H_i)_i$ is stationary. In that case, for $L \geq 2$ we have $P_L = \mathcal{L}(Y_iY_{i+1},\ldots,Y_{i+L-1})$ for all $i$, where the distribution $P_L$ is defined by by

$$P_L(B_1,\ldots,B_L) = \int_{\mathscr{E}^L} \left(\prod_{l=1}^{L} F_{h_l}(B_l)\right) \pi^*(dh_1)Q(h_1,dh_2)\ldots Q(h_{L-1},dh_L),$$

for all Borel sets $B_1,\ldots,B_L \in \mathcal{Y}$. We will see that the stationarity assumption is not necessary, only the ergodicity of $Q^*$ is required.

## 5.2 The framework

Let $Y_1,Y_2,\ldots,Y_N$ be random variables taking values in $(\mathbb{R},\mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel $\sigma$-algebra on $\mathbb{R}$. Let $L$ be in $\{2,3,\ldots,\lfloor N/2 \rfloor\}$ and $n$ be the integer given by $n = N + 1 - L$. We define the new random variables

$$X_i = (Y_i,Y_{i+1},\ldots,Y_{i+L-1}), i = 1,\ldots,n, \tag{5.1}$$

taking values in the measurable space $\left(\mathbb{R}^L,\sigma(\mathbb{R}^L)\right)$.

**Assumption 5.1.** *Let $(Y_i,H_i)_i$ be a HMM with parameters $(\mathscr{E},\pi,Q^*,\overline{\mathscr{F}})$ such that $Q^*$ has a density $q^*$ with respect to some measure $\eta$ such that*

$$0 < q_- \leq (q^*)^m(h_1,h_2) \leq q_+ < +\infty. \tag{5.2}$$

Under this assumption $Q^*$ has only one invariant distribution $\pi^*$ and we define the distribution $P^*$ by

$$P^*(B_1,\dots,B_L) = \int_{\mathscr{E}^L} \left(\prod_{l=1}^{L} F_{h_l}(B_l)\right) \pi^*(dh_1)Q^*(h_1,dh_2)\dots Q^*(h_{L-1},dh_L), \tag{5.3}$$

for all Borel sets $B_1,\dots,B_L \in \mathcal{E}$. In that case we do not necessarily have identically distributed observations however the distribution $\mathcal{L}(Y_i,\dots,Y_{i+L-1})$ converges exponentially fast to $P^*$. Our aim is to estimate $P^*$ based on the observations $(Y_i)_i$.

In what follows we believe $P^*$ is of the form (5.3) with Gaussian emission distributions or can be well approximated by such a distribution. We denote by $\overline{\mathscr{N}} = \{\mathcal{N}(z,\sigma^2);(z,\sigma^2) =: h \in \mathscr{H}\}$ the class of univariate normal distributions with $\mathscr{H} = \mathbb{R}\times(0,\infty)$. Let $\mathcal{H}$ be the Borel $\sigma$-algebra on $\mathbb{R}$. Let $\mathscr{P}_H$ and $\mathscr{Q}_H$ be the class of probability distributions and Markov kernels on $(\mathscr{H},\mathcal{H})$ respectively. For $\pi \in \mathscr{P}_H$ and $Q \in \mathscr{Q}_H$, we define the probability distribution $\nu_{\pi,Q}$ on $\left(\mathscr{H}^L,\mathcal{H}^{\otimes L}\right)$ by

$$\nu_{\pi,Q}(dh_1,\dots,dh_L) = \pi(dh_1)Q(h_1,dh_2)\dots Q(h_{L-1},dh_L),$$

and the distribution $P_{\pi,Q}$ on $\left(\mathbb{R}^L,\mathcal{B}(\mathbb{R}^L)\right)$ associated to the probability density function

$$p_{\pi,Q} : (x_1,\dots,x_L) \mapsto \int_{\mathscr{H}^L} \frac{e^{-\frac{(x_1-h_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi\sigma_1^2}} \dots \frac{e^{-\frac{(x_L-h_L)^2}{2\sigma_L^2}}}{\sqrt{2\pi\sigma_L^2}} \nu_{\pi,Q}(dh_1,\dots,dh_L), \tag{5.4}$$

with respect to the Lebesgue measure. We have the following result of identifiability for a compact hidden state space.

**Theorem 5.1.** *Let $K$ be a compact subset of $\mathscr{H}$. For all $\pi,\pi' \in \mathscr{P}_H$ and $Q,Q' \in \mathscr{Q}_H$ such that $\nu_{\pi,Q}(K^L) = \nu_{\pi',Q'}(K^L) = 1$, we have $P_{\pi,Q} = P_{\pi',Q'} \Rightarrow \nu_{\pi,Q} = \nu_{\pi',Q'}$.*

This result is proven in Section 5.B.1. It is similar to the results of Gassiat *et al.*[46] for nonparametric translation HMMs with general state space. Their framework differs from ours but we can see that translation HMMs with Gaussian noise is a specific case of both their framework and ours.

### 5.2.1 Reminders of $\rho$-estimation

We denote by $\mathscr{P}$ the class of probability distributions on $\left(\mathbb{R}^L,\sigma(\mathbb{R}^L)\right)$. We denote by $\psi$ the function given by

$$\psi : \left| \begin{array}{l} [0,+\infty] \to [-1,1] \\ x \mapsto \frac{x-1}{x+1} \end{array} \right. .$$

Let $\mathscr{M}$ be a countable subset of $\mathscr{P}$ with an associated set of density functions $\mathcal{M}$ with respect to the Lebesgue measure on $\mathbb{R}^d$. For $n \geq 1$, we denote by $\mathbf{T}_n$ and $\mathbf{\Upsilon}_n$ the functions given by

$$\mathbf{T}_n : \left| \begin{array}{l} \mathbb{R}^{Ln} \times \mathcal{M} \times \mathcal{M} \to [-1,1] \\ (\mathbf{x},q,q') \mapsto \sum_{k=1}^{n} \psi\left(\sqrt{\frac{q'(x_i)}{q(x_i)}}\right) \end{array} \right.$$

with the convention $0/0 = 1$, $a/0 = +\infty$ for all $a > 0$, and

$$\mathbf{\Upsilon}_n : \left| \begin{array}{l} \mathbb{R}^{Ln} \times \mathcal{M} \\ (\mathbf{x},q) \mapsto \sup_{q' \in \mathcal{M}} \mathbf{T}_n(\mathbf{x},q,q') \end{array} \right. .$$

For $\mathbf{x}$ in $\mathbb{R}^{Ln}$, we define the (nonvoid) set $\mathscr{E}_n(\mathbf{x})$ by

$$\mathscr{E}_n(\mathbf{x}) = \left\{ Q = q \cdot \mu \middle| q \in \mathcal{M}, \boldsymbol{\Upsilon}_n(\mathbf{X},q) < \inf_{q' \in \mathcal{M}} \boldsymbol{\Upsilon}_n(\mathbf{X},q') + 11.36 \right\}. \tag{5.5}$$

We denote by $\hat{P}(n,\mathbf{X},\mathscr{M})$ any measurable element of the closure of $\mathscr{E}_n(\mathbf{X})$ with respect to the Hellinger distance and we call it a $\rho$-estimator on $\mathscr{M}$. The constant 11.36 is given by (7) and (19) in [11] but can be replaced by any smaller positive number.

### 5.2.2 Our estimation procedure

We build a subset of the observations by taking them separated by blocks of length $s$. Formally, for $s \in \{0,1,\ldots,s_{\max}\}, s_{\max} := \lfloor (n-2)/2 \rfloor$ and $b \in [s+1]$, we define

$$n(s,b) := \left\lfloor \frac{n+s+1-b}{1+s} \right\rfloor \geq 2,$$

for $i \in [n(s,b)]$

$$X_i^{(s,b)} := X_{b+(i-1)(s+1)} \in \mathbb{R}^L, \forall i \in [n(s,b)], \tag{5.6}$$

and

$$\mathbf{X}^{(s,b)} := \left( X_i^{(s,b)}, i \in [n(s,b)] \right).$$

We obtain $s+1$ subsets $\mathbf{X}^{(s,1)},\ldots,\mathbf{X}^{(s,s+1)}$ with sizes $n(s,1),\ldots,n(s,s+1)$ respectively. For each block $b \in [s+1]$, we consider the probabilities $\mathbf{P}^*_{s,b}$ and $\mathbf{P}^{ind}_{s,b}$ which are defined by

$$\mathbf{P}^*_{s,b} := \mathcal{L}\left(\mathbf{X}^{(s,b)}\right) \text{ and } \mathbf{P}^{ind}_{s,b} := \bigotimes_{i=1}^{n(s,b)} \mathcal{L}\left(X_i^{(s,b)}\right). \tag{5.7}$$

We denote for short $\mathbf{P}^* := \mathbf{P}^*_{0,1}$ the distribution of $\mathbf{X} = (X_1,\ldots,X_n)$ and

$$\mathbf{P}^{ind} := \mathbf{P}^{ind}_{0,1} = \mathcal{L}(X_1) \otimes \cdots \otimes \mathcal{L}(X_n) = \bigotimes_{i=1}^{n} P_i.$$

In order to measure the dependence within the $X_i^{(s,b)}$ we use the Kullback-Leibler divergence $\mathbf{K}$ defined as follow. For two probability distributions $P$ and $Q$ on the same measurable space,

$$\mathbf{K}(Q||P) = \begin{cases} \int \log\left(\frac{dQ}{dP}\right) dQ \text{ if } Q \ll P, \\ +\infty \text{ otherwise.} \end{cases}$$

Our estimator is obtained with the following statistical procedure.

1. Let $s$ be in $\{0,1,\ldots,s_{\max}\}$. For $b$ in $[s+1]$, we denote by $\hat{P}_{s,b}$ the estimators given by
$$\hat{P}_{s,b} := \hat{P}\left(n(s,b),\mathbf{X}^{(s,b)},\mathscr{M}\right).$$

2. We denote by $\hat{P}_s = \hat{P}_s(\mathbf{Y},\mathscr{M})$ any element of $\mathscr{M}$ that satisfies
$$\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b},\hat{P}_s\right) \leq \inf_{Q \in \mathscr{M}} \sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b},Q\right) + \iota, \tag{5.8}$$
where $\iota$ is any fixed constant in (0,1273].

In order to evaluate the performance of our estimator we use the Hellinger distance $h$ defined by

$$h^2(Q,Q') = \frac{1}{2} \int \left( \sqrt{\frac{dQ}{d\mu}} - \sqrt{\frac{dQ'}{d\mu}} \right)^2 d\mu,$$

where $\mu$ is a measure that dominates both $Q$ and $Q'$, the result being independent of $\mu$.

## 5.3 Approximation of general HMMs by finite mixtures

It is too complicated to consider the class of all distributions $P_{\pi,Q}$ for the estimation but such a probability can be well approximated by a finite mixture when $w$ and $Q$ are supported on compact sets. For $K \geq 2$, we denote by $\overline{\mathscr{M}}_K$ the set of distributions defined by

$$\overline{\mathscr{M}}_K := \left\{ \sum_{k=1}^{K} w_k \bigotimes_{l=1}^{L} F_{k,l}; w \in \mathcal{W}_K, F_{k,l} \in \overline{\mathscr{N}}, \forall k \in [K], \forall l \in [L] \right\}, \tag{5.9}$$

where

$$\mathcal{W}_K = \left\{ w \in [0,1]^K; w_1 + \cdots + w_K = 1 \right\}. \tag{5.10}$$

For $A > 0$ and $R > 1$, we denote by $\mathscr{C}(A,R)$ the set of probability distributions given by

$$\mathscr{C}(A,R) := \left\{ P_{\pi,Q}; \exists l \in \mathbb{R}, \exists s \in (0,\infty), \int_{D(A,R,s,l)} Q(h_{L-1}, dh_L) \dots Q(h_1, dh_2) w(dh_1) = 1 \right\},$$

with $D(A,R,s,l) = ([l \pm sA] \times s[1,R])^L$. We have the following approximation result that is proven in Section 5.B.4.

**Proposition 5.1.** *For $K \geq 1 + (24A^2L+1)^L (48A^2L+1)^L$ and $R \geq 1$, $L \geq 1$, for any $P_{\pi,Q}$ in $\mathscr{C}(A,R)$, we have*

$$h^2\left(P_{\pi,Q}, \overline{\mathscr{M}}_K\right) \leq K^{\frac{L-1}{4L}} \left(\sqrt{\frac{1+\sqrt{2}}{Le\pi}}\right)^{L-1} \exp\left(-\frac{K^{1/2L}}{12LR^2\sqrt{6}}\right) \left[\sqrt{\frac{1+\sqrt{2}}{Le\pi}} K^{1/4L} + R\frac{2^L-1}{2}\right].$$

For a fixed $R$ we can take a well-chosen value of $K$ to obtain the desired approximation guarantee. We build countable subsets of $\overline{\mathscr{M}}_K$ to apply our estimation procedure. For $0 < \delta < 1/K$, we define the subset $\mathscr{M}_{K,\delta}$ by

$$\mathscr{M}_{K,\delta} := \left\{ \sum_{k=1}^{K} w_k \bigotimes_{l=1}^{L} F_{k,l}; w \in \mathcal{W}_{\delta,K}, F_{k,l} \in \mathscr{F}, \forall t \in [T], \forall l \in [L] \right\},$$

with $\mathscr{F} := \{\mathcal{N}(z,\sigma^2); z \in \mathbb{Q}, \sigma \in \mathbb{Q} \cap (0,\infty)\}$ and $\mathcal{W}_{\delta,K} := \mathcal{W}_K \cap ([\delta,1] \cap \mathbb{Q})^K$.

## 5.4 Main result

Our main result is the following theorem and is proven in Section 5.B.2.

**Theorem 5.2.** *Let $\hat{P}_s = \hat{P}_s(\mathbf{Y}, \mathscr{M}_{K,\delta})$ be the estimator given by (5.8) with*

$$\delta = \frac{3+2L}{n(s,1)} \wedge \frac{1}{K} \text{ and } K = \left\lceil (12\sqrt{6}LR^2 \log n)^{2L} \right\rceil. \tag{5.11}$$

*There exists a positive constant $C$ such that for all $R \in [1, (n/12\sqrt{6}L\log n)^{1/4L}]$ and all $\overline{P} \in \mathscr{P}_X$, we have*

$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq h^2\left(\overline{P}, \mathscr{C}(A,R)\right) + n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) + \sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) + (s+1)\left(12\sqrt{6}\right)^{2L} R^{4L} L^{2L+2} \frac{\log^{2L+1} n}{n}, \tag{5.12}$$

*where $A = A(R,n) := \frac{\sqrt{12\sqrt{6}-1}}{4\sqrt{3}} R \log^{1/2} n$. In particular under Assumption 5.1 there exist positive constants $C(Q^*), c(Q^*)$ such that for $s \geq c(Q^*)\log n$,*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq h^2\left(P^*, \mathscr{C}(A,R)\right) + \left(12\sqrt{6}\right)^{2L} L^{2L+2} R^{4L} \frac{s\log^{2L+1} n}{n}. \tag{5.13}$$

Inequality (5.12) does not require any assumption on the data. One should notice that the statistician does not need to specify a value of $A$ and $R$. In fact, we could have in the bound an infimum with respect to $R$ in $[1,(n/12\sqrt{6}L\log n)^{1/4L}]$ of the quantity

$$h^2\left(\overline{P},\mathscr{C}(A,R)\right) + (s+1)\left(12\sqrt{6}\right)^{2L} R^{4L} L^{2L+2}\frac{\log^{2L+1} n}{n}.$$

We can deduce from (5.13) a bound on the convergence rate over the class

$$\mathscr{C}^* = \bigcup_{R>0} \mathscr{C}^*(R) \text{ with } \mathscr{C}^*(R) := \left\{ P_{w,Q} \in \bigcup_{A>0} \mathscr{C}(A,R); Q \text{ satisfies (5.2)} \right\}.$$

For $P^* = P_{w^*,Q^*}$ in $\mathscr{C}^*$, there exist $R^* > 0$ such that $P^* \in \mathscr{C}^*(R^*)$ and for $s$ of order $\log^2 n$ we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \left(12\sqrt{6}\right)^{2L} L^{2L+2}(R^*)^{2L}\frac{\log^{2L+3} n}{n},$$

for $n$ large enough. This result does not require information on the constant $c(Q^*)$. If it were the case we could take an optimal value of $s$ and obtain a slightly better power of $\log n$ in our bound.

To illustrate the robustness of our estimator we consider the following situation. Let $Z_1,\ldots,Z_N$ be random variables with any distributions and $E_1,\ldots,E_N$ be Bernoulli random variables such that for all $i \in [n]$,

$$Y_i = E_i Y_i' + (1-E_i)Z_i,$$

where $\mathbf{Y}'$ satisfies Assumption 5.1. The following result is proven in Section 5.B.3.

**Corollary 5.1.** *Let $R$ be in $[1,(n/12\sqrt{6}L\log n)^{1/4L}]$. Let $\hat{P}_s = \hat{P}_s(\mathbf{Y},\mathscr{M}_{K,\delta})$ be the estimator given by (5.8) with $\delta$ and $K$ given by (5.11). If $E_1,Z_1,\ldots,E_N,Z_N$ and $\mathbf{Y}'$ are mutually independent, there exist positive constants $C(Q^*),c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have*

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\mathscr{C}(A,R)\right) + \frac{L}{N}\sum_{i=1}^{N}(1-p_i) + \left(12\sqrt{6}\right)^{2L} L^{2L+2}R^{4L}\frac{s\log^{2L+1} n}{n}, \tag{5.14}$$

*where $A = A(R,n) := \frac{\sqrt{12\sqrt{6}-1}}{4\sqrt{3}} R\log^{1/2} n$ and $p_i = \mathbb{P}(E_i = 1)$ for all $i \in [N]$.*

One can see that our deviation bound is not significantly worse as long as the average proportion of contamination $\frac{L}{N}\sum_{i=1}^{N}(1-p_i)$ remains small compared to the last term on the right hand side of (5.14). One would typically look at the following situation. We assume that the model is well specified, i.e. $P^* \in \mathscr{C}^*(R^*)$. For Hüber's contamination model, i.e. $p_i = 1-\alpha_{cont}$ for all $i$, and for $n$ large enough we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\alpha_{cont} + \left(12\sqrt{6}\right)^{2L} L^{2L+2}(R^*)^{4L}\frac{s\log^{2L+1} n}{n},$$

for $s \geq c(Q^*)\log n$. The bound on the convergence rate is not deteriorated as long as the contamination rate $\alpha_{cont}$ is of order not larger than $\left(12\sqrt{6}\right)^{2L} L^{2L+2}R^{4L}\frac{\log^{2L+2} n}{n}$. We can also consider the situation where $\mathbb{P}(E_i = 0) = \mathbb{1}_{i\in I}$ for some subset $I \subset [N]$. For $s \geq c(Q^*)\log n$ and $n$ large enough we get

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq \frac{L|I|}{N} + \left(12\sqrt{6}\right)^{2L} L^{2L+2}(R^*)^{4L}\frac{s\log^{2L+1} n}{n}.$$

As before, our bound on the convergence rate is not deteriorated as long as the proportion of outliers $|I|/N$ is small compared to the second term on the right hand side.

# Appendix

## 5.A General

## 5.B General hidden Markov models

### 5.B.1 Proof of Theorem 5.1

We use the extension to the multidimensional case of Theorem 1 of Bruni and Koch [20] (see 4.(b) on page 1352) with $n = 2L, p = L$ and $D = K^L$ which is a compact subset of $\mathbb{R}^{2L}$. We can take

$$\Lambda_2 = \left\{\lambda : (h_1, \ldots, h_L) \in\mapsto \left(z, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_L^2)\right) \in \mathbb{R}^L \times \mathbb{R}^{L\times L}\right\},$$

with the notation $h_i = (z_i, \sigma_i^2)$ and

$$\mathrm{diag}(d_1, \ldots, d_L) = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_L \end{pmatrix}.$$

For a probability distribution $A$ on $D$, we denote by $P_A$ the distribution on $\mathbb{R}^L$ associated to the density

$$p_A : (x_1, \ldots, x_L) \mapsto \int_D \frac{e^{-(x_1-h_1)^2/2\sigma_1^2}}{\sqrt{2\pi\sigma_1^2}} \ldots \frac{e^{-(x_L-h_L)^2/2\sigma_L^2}}{\sqrt{2\pi\sigma_L^2}} A(dh_1, \ldots, dh_L).$$

The results of Bruni and Koch [20] give the identifiability $P_A = P_{A'} \Rightarrow A = A'$, for all distributions $A, A'$ on $D$. We can conclude taking $A = \nu_{\pi,Q}$ and $A' = \nu_{\pi',Q'}$.

### 5.B.2 Proof of Theorem 5.2

From Lemma 3.15 and Section 3.4.3, we have that $\left\{x \mapsto g_{z_1,\sigma_1^2}(x_1) \ldots g_{z_L,\sigma_L^2}(x_L); z_i \in \mathbb{R}, \sigma_i^2 \in (0,\infty)\right\}$ is VC-subgraph with VC-index $3 + L\frac{1(1+3)}{2} = 3 + 2L$. The $\rho$-dimension function is properly defined in Baraud & Birgé [11]. With Proposition A.1 of Lecestre [62], we get

$$D_{n(s,1)}\left(\mathscr{M}_{K,\delta}\right) \leq 545.3K(3+2L)\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n(s,1)}{K(3+2L)}\right)\right].$$

For $\delta$ and $K$ given by (5.11) we have

$$\begin{aligned} D_{n(s,1)}\left(\mathscr{Q}_{K,\delta}\right) &\leq C\left(12\sqrt{6}\right)^{2L} R^{2L} L^{2L+1} \log^{2L} n \log\left(\left(12\sqrt{6}LR^2 \log n\right)^{2L} \vee n\right) \\ &\leq C\left(12\sqrt{6}\right)^{2L} R^{2L} L^{2L+2} \log^{2L+1} n, \end{aligned}$$

where $C$ is a numeric constant that can differ from one line to the next. One can easily check that for $A = A(R,n)$,

$$\left\lceil\left(12\sqrt{6}LR^2 \log n\right)^{2L}\right\rceil \geq 1 + (24A^2L+1)^L(48A^2L+1)^L.$$

With Proposition 5.1 and Lemma B.2 [62] we get

$$\begin{aligned}
h^2\left(Q,\mathscr{M}_{K,\delta}\right) &\leq 2h^2\left(Q,\mathscr{M}_K\right) + 2\frac{(K-1)(3+2L)}{n(s,1)} \\
&\leq 4h^2\left(Q,\mathscr{C}(A,R)\right) + 2\frac{(K-1)(3+2L)}{n(s,1)} \\
&+ 4\left[\left(12\sqrt{6}LR^2\log n\right)^{2L}+1\right]^{\frac{L}{4L}}\left(\sqrt{\frac{1+\sqrt{2}}{Le\pi}}\right)^L\frac{1}{n}\left[1+R\frac{2^L-1}{2}\right] \\
&\leq 4h^2\left(\mu,\mathscr{C}(A,R)\right) + 2\frac{\left(12\sqrt{6}LR^2\log n\right)^{2L}(3+2L)}{n(s,1)} \\
&+ 4\left[1+(12\sqrt{6}LR^2\log n)^{-2L}\right]^{\frac{1}{4}}\left[R^{-1}2^{-L}+\frac{1-2^{-L}}{2}\right] \\
&\times\left(2\sqrt{\frac{12\sqrt{6}(1+\sqrt{2})}{e\pi}}\right)^L\frac{\log^{L/2}n}{n}R^{L+1}.
\end{aligned}$$

We can apply Theorem 3.1 and obtain (5.12). The next lemma is proven in Section 5.B.5.

**Lemma 5.1.** *Under Assumption 5.1 there are positive constants $C(Q^*)$ and $r(Q^*)$ such that*

$$n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \leq C(Q^*)e^{-r(Q^*)s},\forall s\geq L+m-2,\forall b\in[s+1],$$

*and*

$$h^2\left(P^*,P_i\right)\leq C(Q^*)e^{-r(Q^*)i},\forall i\in[n].$$

We can deduce (5.13) from this result and (5.12).

### 5.B.3 Proof of Corollary 5.1

We have

$$\mathbb{P}\left(X_i=(Y'_i,\ldots,Y'_{i+L-1})\right)\geq\mathbb{P}\left(E_i=\cdots=E_{i+L-1}=1\right)=p_ip_{i+1}\ldots p_{i+L-1},$$

and with the convexity of the squared Hellinger distance

$$\begin{aligned}
h^2\left(P_i,P^*\right) &\leq p_ip_{i+1}\ldots p_{i+L-1}h^2\left(P'_i,P^*\right)+\left(1-p_ip_{i+1}\ldots p_{i+L-1}\right) \\
&\leq h^2\left(P'_i,P^*\right)+\left(1-p_i\right)+\cdots+\left(1-p_{i+L-1}\right),
\end{aligned}$$

where $P'_i=\mathcal{L}(Y'_i,\ldots,Y'_{i+L-1})$. One can check that $n\geq 1+N/2$ with our conditions on $L$. With Theorem 5.2, Lemma 3.2 and Lemma 5.1, we have

$$\begin{aligned}
C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] &\leq h^2\left(P^*,\mathscr{C}(A,R)\right)+\frac{L}{N}\sum_{i=1}^{N}(1-p_i) \\
&+(s+1)\left(12\sqrt{6}\right)^{2L}R^{4L}L^{2L+2}\frac{\log^{2L+1}n}{n}+e^{-c(Q^*)s},
\end{aligned}$$

for some positive constant $C(Q^*),r(Q^*)$ and $s\geq L-1$.

### 5.B.4 Proof of Proposition 5.1

For $a > 0$ and $0 < \underline{\sigma} < \overline{\sigma} < \infty$, we denote by $\mathscr{G}_{mix}(T,a,\underline{\sigma},\overline{\sigma})$ the set of mixture probability distributions of the form

$$\sum_{t=1}^{T} w_t g_{z^*_{t,1},\sigma^*_{t,1}}(x_1)\dots\phi_{z^*_{t,L},\sigma^*_{t,L}}(x_L),$$

where $w \in \mathcal{W}_T$, $\sigma^*_{t,l} \in [\underline{\sigma},\overline{\sigma}]$,$z^*_{t,l} \in [-a,a]$ for all $t \in [T]$ and $l \in [L]$. If $H$ is probability distribution on $(\mathbb{R}\times(0,\infty))^L$, we define the density function

$$p_H(x_1,\dots,x_L) = \int g_{h_1}(x_1)\dots g_{h_L}(x_L)H(dh_1,\dots,dh_L),$$

and its associated distribution $P_H(dx) = p_H(x)dx$.

**Lemma 5.2.** *Let $n \geq 1$, $a > 0$ and $0 < \underline{\sigma} < \overline{\sigma} < \infty$. If $H$ is supported on $([-a,a]\times[\underline{\sigma},\overline{\sigma}])^L$, then for $K \geq n^L(2n-1)^L + 1$ and $M \geq a + \overline{\sigma}\sqrt{\log_+(2\pi\overline{\sigma}^2)}$, we have*

$$d_{TV}\left(P_H, \mathscr{G}_{mix}(T,a,\underline{\sigma},\overline{\sigma})\right) \leq \frac{(2M)^L}{(2\pi\underline{\sigma}^2)^{L/2}}\left[\left(\frac{eL\left(M+a\right)^2}{2n\underline{\sigma}^2}\right)^n + \frac{e^{-\frac{(M-a)^2}{2\overline{\sigma}^2}}\frac{\sqrt{2\pi\overline{\sigma}^2}}{2M}}{2}\left(2^L-1\right)\right].$$

As a direct consequence of this lemma and the fact that the Hellinger distance is invariant to translation or scaling we have the following. For any $l \in \mathbb{R}$, any probability distribution $H$ supported on $([l \pm sA]\times s^2[1,R^2])^L$ and for $T \geq n^L(2n-1)^L+1$, with $M = Am$ we have

$$h^2\left(P_H, \mathscr{G}_{mix}(T,a,\underline{\sigma},\overline{\sigma})\right) \leq \frac{(2Am)^L}{(2\pi)^{L/2}}\left[\left(\frac{eLA^2\left(m+1\right)^2}{2n}\right)^n + \frac{e^{-\frac{A^2(m-1)^2}{2R^2}}\frac{R\sqrt{2\pi}}{2Am}}{2}\left(2^L-1\right)\right],$$

for $m \geq 1 + \frac{R}{A}\sqrt{\log_+(2\pi R^2)} = m_- \geq 1$. Therefore,

$$\begin{aligned} h^2\left(P_H, \mathscr{G}_{mix}(T,a,\underline{\sigma},\overline{\sigma})\right) &\leq \inf_{m\geq m_- -1}\frac{(2A(m+1))^L}{(2\pi)^{L/2}}\left[\left(\frac{eLA^2(m+2)^2}{2n}\right)^n + \frac{e^{-\frac{A^2m^2}{2R^2}}\frac{R\sqrt{\pi/2}}{A(m+1)}}{2}\left(2^L-1\right)\right] \\ &\leq \inf_{m\geq m_-+1}\left(2\sqrt{2/\pi}Am\right)^L\left(\frac{2eLA^2m^2}{n}\right)^n + \frac{R\left(2^L-1\right)}{2}\left(2\sqrt{2/\pi}Am\right)^{L-1}e^{-\frac{A^2m^2}{2R^2}}. \end{aligned}$$

We denote by $W$ the Lambert $W$ function restricted to $(0,\infty)$ such that $W(x)$ is the only positive number such that $W(x)e^{W(x)} = x$. For

$$m = \frac{\sqrt{2W\left(1/4eR^2L\right)}R}{A}n^{1/2} \text{ and } n \geq \frac{2A^2}{W\left(1/4eR^2L\right)R^2},$$

we get

$$\begin{aligned}
&\left(2\sqrt{2/\pi}Am\right)^L\left(\frac{2eLA^2m^2}{n}\right)^n+\frac{R\left(2^L-1\right)}{2}\left(2\sqrt{2/\pi}Am\right)^{L-1}e^{-\frac{A^2m^2}{2R^2}}\\
&=\left(4R\sqrt{nW(1/4eR^2L)/\pi}\right)^L\left(eL4R^2W(1/4eR^2L)\right)^n\\
&+\frac{\frac{3}{2}R\left(2^L-1\right)}{2}\left(4R\sqrt{nW(1/4eR^2L)/\pi}\right)^{L-1}e^{-nW(1/4eR^2L)}\\
&=\left(4R\sqrt{\frac{nW(1/4eR^2L)}{\pi}}\right)^{L-1}R\exp\left(-nW(1/4eR^2L)\right)\\
&\times\left[4\sqrt{\frac{nW(1/4eR^2L)}{\pi}}+\frac{2^L-1}{2}\right].
\end{aligned}$$

Following the proof of Proposition 3.5 [62], we have $W(x)<x$ for all $x>0$ and $W(w)\geq x(1-x)$ for all $x\in(0,1)$. Therefore,

$$W(1/4eR^2L)\geq\frac{1}{4eR^2L}\left(1-\frac{1}{4eR^2L}\right)\geq\frac{1}{12LR^2}$$

and

$$\begin{aligned}
h^2\left(P_H,\mathscr{G}_{mix}(T,a,\underline{\sigma},\overline{\sigma})\right)&\leq\left(4R\sqrt{\frac{nW(1/4eR^2L)}{\pi}}\right)^{L-1}R\exp\left(-nW(1/4eR^2L)\right)\\
&\times\left[4\sqrt{\frac{nW(1/4eR^22L)}{\pi}}+\frac{2^L-1}{2}\right]\\
&\leq\left(2\sqrt{\frac{n}{eL\pi}}\right)^{L-1}\exp\left(-\frac{n}{12LR^2}\right)\times\left[2\sqrt{\frac{n}{Le\pi}}+R\frac{2^L-1}{2}\right].
\end{aligned}$$

For $K\geq1+(24A^2L+1)^L(48A^2L+1)^L$, the set

$$B=\left\{n\in\mathbb{N}:K\geq n^L(2n-1)^L+1\text{ and }n\geq\frac{2A^2}{R^2W(1/4eR^2)}\right\}$$

is not empty, e.g. $\lceil24A^2L\rceil\in B$. We set $n=\max B\geq1$, i.e.

$$n=\left\lfloor\frac{1}{4}\left[1+\sqrt{1+8(K-1)^{1/L}}\right]\right\rfloor\leq\frac{1+\sqrt{2}}{4}K^{1/2L},$$

and we have

$$k^L(2k-1)^L+1\leq T<(k+1)^L(2k+1)^L+1\Rightarrow n=k\geq T^{1/2L}\frac{k}{\sqrt{(k+1)(2k+1)}}.$$

As $x\mapsto\frac{x}{\sqrt{(2x+1)(x+1)}}$ is non-decreasing on $[1,+\infty)$, we have $n\geq K^{1/2L}/\sqrt{6}$ since $K\geq2$. Finally, we have

$$\begin{aligned}
h^2\left(P_H,\overline{M}_K\right)&\leq h^2\left(P_H,\mathscr{G}_{mix}(K,a,\underline{\sigma},\overline{\sigma})\right)\\
&\leq\left(2\sqrt{\frac{1+\sqrt{2}}{e4L\pi}K^{1/2L}}\right)^{L-1}\exp\left(-\frac{K^{1/2L}}{12LR^2\sqrt{6}}\right)\times\left[2\sqrt{\frac{1+\sqrt{2}}{4Le\pi}}K^{1/4L}+R\frac{2^L-1}{2}\right]\\
&\leq K^{\frac{L-1}{4L}}\left(\sqrt{\frac{1+\sqrt{2}}{eL\pi}}\right)^{L-1}\exp\left(-\frac{K^{1/2L}}{12LR^2\sqrt{6}}\right)\left[\sqrt{\frac{1+\sqrt{2}}{Le\pi}}K^{1/4L}+R\frac{2^L-1}{2}\right].
\end{aligned}$$

We can conclude by noticing that if $P_{\pi,Q} \in \mathscr{C}(A,R)$ the distribution $H(dh_1,\ldots,dh_L) = \pi(dh_1)Q(h_1,dh_2)\ldots Q(h_{L-1},dh_L)$ is supported on $([l \pm sA] \times s^2[1,R^2])^L$ for some $l \in \mathbb{R}$ and $s > 0$.

**Proof of Lemma 5.2**

We follow the proof of Lemma C.1 in [62]. For any $y \geq 0$ and $n \geq 1$, we have

$$\left| e^{-y} - \sum_{k=0}^{n-1} \frac{(-1)^k y^k}{k!} \right| \leq \frac{(ey)^n}{n^n}.$$

For $y = \sum_{l=1}^{L} \frac{(x_l - z_l)^2}{2\sigma_l^2} = \frac{1}{2}\left|\left|\frac{\mathbf{x}-\mathbf{z}}{\sigma}\right|\right|_2^2$, we get

$$\begin{aligned}
&\left| \prod_{l=1}^{L} g_{z_l,\sigma_l^2}(x_l) - \frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \sum_{k=0}^{n-1} \frac{(-1)^k}{k!2^k} \left( \left|\left| \frac{\mathbf{x}-\mathbf{z}}{\sigma} \right|\right|_2^2 \right)^k \right| \\
&\leq \frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \left( \frac{\frac{e}{2}\left|\left|\frac{\mathbf{x}-\mathbf{z}}{\sigma}\right|\right|_2^2}{n} \right)^n.
\end{aligned}$$

One can see that $\frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \sum_{k=0}^{n-1} \frac{(-1)^k}{k!2^k} \left( \left|\left| \frac{\mathbf{x}-\mathbf{z}}{\sigma} \right|\right|_2^2 \right)^k$ is a polynomial function in $z_1,\ldots,z_L,\sigma_1^{-1},\ldots,\sigma_L^{-1}$ of the form

$$\sum_{i=1}^{L} \sum_{j_i=0}^{2(n-1)} \sum_{l_i=0}^{n-1} c_{j_1,l_1,\ldots,j_L,l_L} z_1^{j_1} \sigma_1^{-(2l_1+1)} \ldots z_L^{j_L} \sigma_L^{-(2l_L+1)}.$$

From Lemma A.1 in Ghosal & van der Vaart [49], there is a discrete distribution $H'$ with at most $n^L(2n-1)^L + 1$ support points such that

$$\int z_1^{j_1} \sigma_1^{-(2l_1+1)} \ldots z_L^{j_L} \sigma_L^{-(2l_L+1)} dH(z_1^L,\sigma_1^L) = \int z_1^{j_1} \sigma_1^{-(2l_1+1)} \ldots z_L^{j_L} \sigma_L^{-(2l_L+1)} dH'(h_1,\ldots,h_L),$$

for all $j_i = 0,\ldots,2(n-1)$ and all $l_i = 0,\ldots,n-1$. Hence, we have

$$\begin{aligned}
&|p_H(\mathbf{x}) - p_{H'}(\mathbf{x})| \\
&\leq \left| p_H(\mathbf{x}) - \int \frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \sum_{k=0}^{n-1} \frac{(-1)^k}{k!2^k} \left( \left|\left| \frac{\mathbf{x}-\mathbf{z}}{\sigma} \right|\right|_2^2 \right)^k dH(h_1,\ldots,h_L) \right| \\
&+ \left| p_{H'}(\mathbf{x}) - \int \frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \sum_{k=0}^{n-1} \frac{(-1)^k}{k!2^k} \left( \left|\left| \frac{\mathbf{x}-\mathbf{z}}{\sigma} \right|\right|_2^2 \right)^k dH'(h_1,\ldots,h_L) \right| \\
&\leq \int \frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \left( \frac{\frac{e}{2}\left|\left|\frac{\mathbf{x}-\mathbf{z}}{\sigma}\right|\right|_2^2}{n} \right)^n dH(h_1,\ldots,h_L) \\
&+ \int \frac{1}{(2\pi)^{L/2}\sigma_1\ldots\sigma_L} \left( \frac{\frac{e}{2}\left|\left|\frac{\mathbf{x}-\mathbf{z}}{\sigma}\right|\right|_2^2}{n} \right)^n dH'(h_1,\ldots,h_L).
\end{aligned}$$

Let $M$ be greater than $a$. For $I \subset [L]$ we define the set

$$A_I := \{\mathbf{x}; |x_i| > M \geq |x_j|, \forall i \in I, \forall j \notin I\}$$

such that $\mathbb{R}^L$ is the disjoint union of the $(A_I)_{I\subset[L]}$.

- For $\mathbf{x} \in A_\emptyset$ i.e. $||\mathbf{x}||_\infty = \max_i |x_i| \leq M$, we have

$$|p_H(\mathbf{x}) - p_{H'}(\mathbf{x})| \leq \frac{2}{(2\pi\underline{\sigma}^2)^{L/2}} \left( \frac{eL\,(M+a)^2}{2n\underline{\sigma}^2} \right)^n .$$

- For $\mathbf{x} \in A_I$ with $I \neq \emptyset$, we have

$$\begin{aligned} p_H(\mathbf{x}) &= \int \prod_{i=1}^{L} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left( -\frac{(x_i - z_i)^2}{2\sigma_i^2} \right) dH(h_1, \ldots, h_L) \\ &\leq \left( \frac{1}{2\pi\underline{\sigma}^2} \right)^{L/2} \prod_{i \in I} \exp\left( -\frac{(|x_i| - a)^2}{2\overline{\sigma}^2} \right), \end{aligned}$$

  and the inequality holds for $p_{H'}$ as well.

We have

$$\begin{aligned} d_{TV}\left(P_H, P_{H'}\right) &= \frac{1}{2} \int |p_H(\mathbf{x}) - p_{H'}(\mathbf{x})| d\mathbf{x} \\ &\leq \frac{1}{2} \sup_{||\mathbf{x}||_\infty \leq M} |p_H(\mathbf{x}) - p_{H'}(\mathbf{x})| \int_{||x||_\infty \leq M} dx \\ &+ \frac{1}{2} \sum_{I \subset [L]: I \neq \emptyset} \int_{A_I} p_H(\mathbf{x}) \vee p_{H'}(\mathbf{x}) dx \\ &\leq \frac{1}{(2\pi\underline{\sigma}^2)^{L/2}} \left( \frac{eL\,(M+a)^2}{2n\underline{\sigma}^2} \right)^n \int_{||x||_\infty \leq M} dx \\ &+ \frac{1}{2(2\pi\underline{\sigma}^2)^{L/2}} \sum_{I \subset [L]: I \neq \emptyset} (2M)^{L-|I|} \left( 2 \int_M^\infty \exp\left( -\frac{(x-a)^2}{2\overline{\sigma}^2} \right) dx \right)^{|I|} \\ &\leq \frac{(2M)^L}{(2\pi\underline{\sigma}^2)^{L/2}} \left( \frac{eL\,(M+a)^2}{2n\underline{\sigma}^2} \right)^n \\ &+ \frac{1}{2(2\pi\underline{\sigma}^2)^{L/2}} \sum_{I \subset [L]: I \neq \emptyset} (2M)^{L-|I|} \left( \exp\left( -\frac{(M-a)^2}{2\overline{\sigma}^2} \right) \sqrt{2\pi\overline{\sigma}^2} \right)^{|I|} . \end{aligned}$$

Finally, for $M \geq a + \overline{\sigma}\sqrt{\log(2\pi\overline{\sigma}^2)}$ and $T \geq n^L(2n-1)^L + 1$, we get

$$\begin{aligned} d_{TV}\left(P_H, \mathscr{G}_{mix}(T, a, \underline{\sigma}, \overline{\sigma})\right) &\leq d_{TV}\left(p_H, p_{H'}\right) \\ &= \frac{(2M)^L}{(2\pi\underline{\sigma}^2)^{L/2}} \left[ \left( \frac{eL\,(M+a)^2}{2n\underline{\sigma}^2} \right)^n + \frac{e^{-\frac{(M-a)^2}{2\overline{\sigma}^2}} \frac{\sqrt{2\pi\overline{\sigma}^2}}{2M}}{2} \left(2^L - 1\right) \right] . \end{aligned}$$

### 5.B.5 Proof of Lemma 5.1

Let $s$ be not be smaller than $L-1$ and $b$ be in $[s+1]$. Following the proof of Lemma 3.13, we have

$$\mathbf{K}\left( \mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind} \right) \leq \sum_{i=1}^{n(s,b)-1} \mathbf{K}\left( \mathcal{L}\left( H_i^{(L,s,b)}, H_{i+1}^{(L,s,b)} \right) || \mathcal{L}\left( H_i^{(L,s,b)} \right) \otimes \mathcal{L}\left( H_{i+1}^{(L,s,b)} \right) \right) .$$

For $s_{\max} \geq s \geq m + L - 2$, we have

$$\frac{d\mathcal{L}(H_i^{(L,s,b)}, H_{i+1}^{(L,s,b)})}{d\mathcal{L}(H_i^{(L,s,b)}) \otimes \mathcal{L}(H_{i+1}^{(L,s,b)})} \leq \frac{q_+}{q_-} < \infty.$$

With Theorem 7 of Verdú [86], there is a constant $C$ such that

$$\begin{aligned}
&\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\\
&\leq C\sqrt{\frac{q_+}{q_-}}\sum_{i=1}^{n(s,b)-1} d_{TV}\left(\mathcal{L}\left(H_i^{(L,s,b)},H_{i+1}^{(L,s,b)}\right),\mathcal{L}\left(H_i^{(L,s,b)}\right)\otimes\mathcal{L}\left(H_{i+1}^{(L,s,b)}\right)\right)\\
&= C\sqrt{\frac{q_+}{q_-}}\sum_{i=1}^{n(s,b)-1} d_{TV}\left(\mathcal{L}\left(H_{b+(i-1)(s+1)+L-1},H_{b+i(s+1)}\right),\mathcal{L}\left(H_{b+(i-1)(s+1)+L-1}\right)\otimes\mathcal{L}\left(H_{b+i(s+1)}\right)\right)\\
&= C\sqrt{\frac{q_+}{q_-}}\sum_{i=1}^{n(s,b)-1} \mathbb{E}\left[d_{TV}\left(\mathcal{L}\left(H_{b+i(s+1)}|H_{b+(i-1)(s+1)+L-1}\right),\mathcal{L}\left(H_{b+i(s+1)}\right)\right)\right].
\end{aligned}$$

We have that $\mathscr{E}$ is an accessible $(m,Q_-)$ small set, with $Q_-$ given by $Q_-(dx)=q_-\eta(dx)$. From Theorem 11.4.2 of Douc *et al.*[33], the kernel $Q^*$ has a unique invariant probability distribution $\pi^*$ and there exist positive constants $C(Q^*)$ and $r(Q^*)$ such that

$$d_{TV}\left(\xi(Q^*)^t,\pi\right)\leq C(Q^*)e^{-r(Q^*)t},$$

for any probability distribution $\xi$ on $(\mathscr{E},\mathcal{E})$. We get

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\leq 2C\sqrt{\frac{q_+}{q_-}}C(Q^*)(n(s,b)-1)e^{-r(Q^*)(s+1)}.$$

We also have

$$h^2\left(P^*,P_i\right)\leq d_{TV}\left(P^*,P_i\right)\leq d_{TV}(\pi^*,\mathcal{L}(H_i))=d_{TV}(\pi^*,\pi Q^{i-1})\leq Ce^{-r(Q^*)(i-1)}.$$

# Bibliography


[1] Kweku Abraham, Elisabeth Gassiat, and Zacharie Naulet. *Frontiers to the learning of nonparametric hidden Markov models*. 2023. arXiv: 2306.16293 [math.ST].

[2] Kweku Abraham, Elisabeth Gassiat, and Zacharie Naulet. "Fundamental Limits for Learning Hidden Markov Model Parameters". In: *IEEE Transactions on Information Theory* 69.3 (2023), pp. 1777–1794. DOI: 10.1109/TIT.2022.3213429.

[3] G. Alexandrovich, H. Holzmann, and A. Leister. "Nonparametric identification and maximum likelihood estimation for hidden Markov models". In: *Biometrika* 103.2 (2016), pp. 423–434. DOI: 10.1093/biomet/asw001.

[4] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. "Identifiability of parameters in latent structure models with many observed variables". In: *The Annals of Statistics* 37.6A (2009), pp. 3099–3132. DOI: 10.1214/09-AOS689.

[5] P Alquier and M Gerber. "Universal Robust Regression via Maximum Mean Discrepancy". In: *Biometrika* (May 2023), asad031. DOI: 10.1093/biomet/asad031.

[6] Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. "A Method of Moments for Mixture Models and Hidden Markov Models". In: *Proceedings of the 25th Annual Conference on Learning Theory*. Vol. 23. Proceedings of Machine Learning Research. Edinburgh, Scotland: PMLR, June 2012, pp. 33.1–33.34.

[7] N Atienza, J Garcia-Heras, and J M Muñoz-Pichardo. "A new condition for identifiability of finite mixture distributions". In: *Metrika.* 63.2 (2006-4). DOI: 10.1007/s00184-005-0013-z.

[8] A. Azzalini and A. Capitanio. *The Skew-Normal and Related Families*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2013. DOI: 10.1017/CBO9781139248891.

[9] Y Baraud, L Birgé, and M Sart. "A new method for estimation and model selection: rho-estimation". In: *Inventiones mathematicae* 207.2 (Feb. 2017), pp. 425–517. DOI: 10.1007/s00222-016-0673-5.

[10] Yannick Baraud. "Tests and estimation strategies associated to some loss functions". In: *Probability Theory and Related Fields* (Aug. 2021), pp. 799–846. DOI: 10.1007/s00440-021-01065-1.

[11] Yannick Baraud and Lucien Birgé. "Rho-estimators revisited: General theory and applications". In: *Ann. Statist.* 46.6B (Dec. 2018), pp. 3767–3804. DOI: 10.1214/17-AOS1675.

[12] Yannick Baraud and Juntong Chen. *Robust estimation of a regression function in exponential families*. 2020. arXiv: 2011.01657 [math.ST].

[13] Yannick Baraud and Juntong Chen. *Robust estimation of a regression function in exponential families*. 2022. arXiv: 2011.01657 [math.ST].

[14] Leonard E. Baum and Ted Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". In: *The Annals of Mathematical Statistics* 37.6 (1966), pp. 1554–1563. DOI: 10.1214/aoms/1177699147.

[15] Lucien Birgé. "Approximation dans les espaces métriques et théorie de l'estimation". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 65 (1983). DOI: 10.1007/BF00532480.

[16] Lucien Birgé. "Model selection via testing: an alternative to (penalized) maximum likelihood estimators". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 42.3 (2006), pp. 273–325. DOI: 10.1016/j.anihpb.2005.04.004.

[17] Lucien Birgé. "On estimating a density using Hellinger distance and some other strange facts". In: *Probability Theory and Related Fields* 71 (1986). DOI: 10.1007/BF00332312.

[18] Natalia Bochkina and Judith Rousseau. "Adaptive density estimation based on a mixture of Gammas". In: *Electronic Journal of Statistics* 11.1 (2017), pp. 916–962. DOI: 10.1214/17-EJS1247.

[19] Richard C. Bradley. "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions". In: *Probability Surveys* 2.none (2005), pp. 107–144. DOI: 10.1214/154957805100000104.

[20] C. Bruni and G. Koch. "Identifiability of Continuous Mixtures of Unknown Gaussian Distributions". In: *The Annals of Probability* 13.4 (1985), pp. 1341–1357. DOI: 10.1214/aop/1176992817.

[21] Satish Chandra. "On the mixtures of probability distributions". In: *Scandinavian Journal of Statistics* (1977), pp. 105–112.

[22] Fabienne Comte, Valentine Genon-Catalot, and Yves Rozenholc. "Penalized nonparametric mean square estimation of the coefficients of diffusion processes". In: *Bernoulli* 13.2 (2007), pp. 514–543. DOI: 10.3150/07-BEJ5173.

[23] Didier Dacunha-Castelle and Elisabeth Gassiat. "The Estimation of the Order of a Mixture Model". In: *Bernoulli* 3.3 (1997), pp. 279–299.

[24] Arnak Dalalyan and Markus Reiß. "Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case". In: *Probability Theory and Related Fields* 137.1 (2007), pp. 25–47. DOI: 10.1007/s00440-006-0502-7.

[25] Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. "Minimax Adaptive Estimation of Nonparametric Hidden Markov Models". In: *J. Mach. Learn. Res.* 17.1 (Jan. 2016), pp. 3842–3884.

[26] J. Dedecker, A. Fischer, and B. Michel. "Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one". In: *Electronic Journal of Statistics* 9.1 (2015), pp. 234–265. DOI: 10.1214/15-EJS997.

[27] J. Dedecker et al. *Weak Dependence: With Examples and Applications*. Lecture Notes in Statistics. Springer New York, 2007. DOI: 10.1007/978-0-387-69952-3.

[28] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. "List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2018. Los Angeles, CA, USA: Association for Computing Machinery, 2018, pp. 1047–1060. DOI: 10.1145/3188745.3188758.

[29] Charles R. Doss and Jon A. Wellner. "Global rates of convergence of the MLEs of log-concave and $s$-concave densities". In: *The Annals of Statistics* 44.3 (2016), pp. 954–981. DOI: 10.1214/15-AOS1394.

[30] Natalie Doss et al. *Optimal estimation of high-dimensional location Gaussian mixtures*. 2021. arXiv: 2002.05818 [math.ST].

[31] Randal Douc and Catherine Matias. "Asymptotics of the maximum likelihood estimator for general hidden Markov models". In: *Bernoulli* 7.3 (2001), pp. 381–420.

[32] Randal Douc et al. "Consistency of the maximum likelihood estimator for general hidden Markov models". In: *The Annals of Statistics* 39.1 (2011), pp. 474–513. DOI: 10.1214/10-AOS834.

[33] Randal Douc et al. "Splitting Construction and Invariant Measures". In: *Markov Chains*. Cham: Springer International Publishing, 2018, pp. 241–264. DOI: 10.1007/978-3-319-97704-1_11.

[34] Lutz Dümbgen and Kaspar Rufibach. "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency". In: *Bernoulli* 15.1 (2009), pp. 40–68. DOI: 10.3150/08-BEJ141.

[35] Lutz Dümbgen, Richard Samworth, and Dominic Schuhmacher. "Approximation by log-concave distributions, with applications to regression". In: *The Annals of Statistics* 39.2 (2011), pp. 702–730. DOI: 10.1214/10-AOS853.

[36] Brian. Everitt and D. J. Hand. *Finite mixture distributions*. English. Chapman and Hall London ; New York, 1981, ix, 143 p. :

[37] Mário A. T. Figueiredo and Anil K. Jain. "Unsupervised Learning of Finite Mixture Models". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 24 (2002), pp. 381–396. DOI: 10.1109/34.990138.

[38] Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer New York, 2006. DOI: 10.1007/978-0-387-35768-3.

[39] Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert. *Handbook of Mixture Analysis*. New York: Chapman and Hall/CRC, 2018. DOI: 10.1201/9780429055911.

[40] Sébastien Gadat et al. "Parameter recovery in two-component contamination mixtures: The $L^2$ strategy". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 56.2 (2020), pp. 1391–1418. DOI: 10.1214/19-AIHP1007.

[41] Sébastien Gadat et al. *Parameter recovery in two-component contamination mixtures: the $\mathbb{L}^2$ strategy*. 2018. arXiv: 1604.00306 [math.ST].

[42] E. Gassiat and S. Boucheron. "Optimal Error Exponents in Hidden Markov Models Order Estimation". In: *IEEE Trans. Inf. Theor.* 49.4 (Sept. 2006), pp. 964–980. DOI: 10.1109/TIT.2003.809574.

[43] E. Gassiat, A. Cleynen, and S. Robin. "Inference in finite state space non parametric Hidden Markov Models and applications". In: *Statistics and Computing* 26 (2016), pp. 61–71. DOI: 10.1007/s11222-014-9523-8.

[44] Elisabeth Gassiat. "Likelihood ratio inequalities with applications to various mixtures". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 38.6 (2002), pp. 897–906. DOI: 10.1016/S0246-0203(02)01125-1.

[45] Elisabeth Gassiat. "Mixtures of Nonparametric Components and Hidden Markov Models". In: *Handbook of Mixture Analysis*. Chapman and Hall/CRC, 2018. DOI: 10.1201/9780429055911-14.

[46] Elisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. "Identifiability and Consistent Estimation of Nonparametric Translation Hidden Markov Models with General State Space". In: *Journal of Machine Learning Research* 21.115 (2020), pp. 1–40.

[47] Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. "Deconvolution with unknown noise distribution is possible for multivariate signals". In: *The Annals of Statistics* 50.1 (2022), pp. 303–323. DOI: 10.1214/21-AOS2106.

[48] Christopher Genovese and Larry Wasserman. "Convergence rates for the Gaussian mixture sieve". In: *Ann. Statist.* 28 (Aug. 2000). DOI: 10.1214/aos/1015956709.

[49] Subhashis Ghosal and Aad W. van der Vaart. "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". In: *The Annals of Statistics* 29.5 (2001), pp. 1233–1263. DOI: 10.1214/aos/1013203452.

[50] Philippe Heinrich and Jonas Kahn. "Strong identifiability and optimal minimax rates for finite mixture estimation". In: *The Annals of Statistics* 46.6A (2018), pp. 2844–2870. DOI: 10.1214/17-AOS1641.

[51] Jogi Henna. "Examples of identifiable mixture". In: *Journal of the Japan Statistical Society* 24.2 (1994), pp. 193–200. DOI: 10.11329/jjss1970.24.193.

[52] Marc Hoffmann and Kolyan Ray. *Bayesian estimation in a multidimensional diffusion model with high frequency data*. 2022. arXiv: 2211.12267 [math.ST].

[53] Peter J. Huber. "Robust Estimation of a Location Parameter". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101. DOI: 10.1214/aoms/1177703732.

[54] I. A. Ibragimov and Has'minskii R. Z. *Statistical Estimation. Asymptotic Theory*. Springer New York, 1981. DOI: 10.1007/978-1-4899-0027-2.

[55] *Inference in Hidden Markov Models*. New York, NY: Springer New York, 2005. DOI: 10.1007/0-387-28982-8.

[56] C. Keribin. "Consistent Estimation of the Order of Mixture Models". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 62.1 (2000), pp. 49–66.

[57] M. Kessler. "Estimation of an Ergodic Diffusion from Discrete Observations." In: *Scandinavian Journal of Statistics* 24.2 (1997), pp. 211–229. DOI: 10.1111/1467-9469.00059.

[58] Arlene K. H. Kim and Richard J. Samworth. "Global rates of convergence in log-concave density estimation". In: *The Annals of Statistics* 44.6 (2016), pp. 2756–2779. DOI: 10.1214/16-AOS1480.

[59] W.T. Kruijer, J. Rousseau, and A.W. van der Vaart. "Adaptive Bayesian density estimation with location-scale mixtures". English. In: *Electronic Journal of Statistics* 4 (2010), pp. 1225–1257. DOI: 10.1214/10-EJS584.

[60] Gil Kur, Yuval Dagan, and Alexander Rakhlin. *Optimality of Maximum Likelihood for Log-Concave Density Estimation and Bounded Convex Regression*. 2020. arXiv: 1903.05315 [math.ST].

[61] Alexandre Lecestre. *Robust estimation for ergodic Markovian processes*. 2023. arXiv: 2307.03666 [math.ST].

[62] Lecestre, Alexandre. "Robust Estimation in Finite Mixture Models*". In: *ESAIM: PS* 27 (2023), pp. 402–460. DOI: 10.1051/ps/2023004.

[63] Luc Lehéricy. "Consistent order estimation for nonparametric hidden Markov models". In: *Bernoulli* 25.1 (2019), pp. 464–498. DOI: 10.3150/17-BEJ993.

[64] Luc Lehéricy. "Estimation adaptative pour les modèles de Markov cachés non paramétriques". Theses. Université Paris-Saclay, Dec. 2018.

[65] Luc Lehéricy. "State-by-state Minimax Adaptive Estimation for Nonparametric Hidden Markov Models". In: *Journal of Machine Learning Research* 19.39 (2018), pp. 1–46.

[66] Jonathan Li and Andrew Barron. "Mixture Density Estimation". In: *Advances in Neural Information Processing Systems*. Vol. 12. MIT Press, 1999.

[67] Pascal Massart. "Concentration inequalities and model selection. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003". In: *Lecture Notes in Mathematics -Springer-verlag-* 1896 (Jan. 2007). DOI: 10.1007/978-3-540-48503-2.

[68] Cathy Maugis and Bertrand Michel. "A non asymptotic penalized criterion for Gaussian mixture model selection". In: *ESAIM: Probability and Statistics* 15 (2011), pp. 41–68. DOI: 10.1051/ps/2009004.

[69] Cathy Maugis-Rabusseau and Bertrand Michel. "Adaptive density estimation for clustering with Gaussian mixtures". In: *ESAIM Probability and Statistics* 17 (2013), pp. 698–724. DOI: 10.1051/ps/2012018.

[70] G. Mclachlan and K. Basford. *Mixture Models: Inference and Applications to Clustering*. Vol. 84. New York: Marcel Dekker, 1988.

[71] G. Mclachlan and David Peel. *Finite Mixture Model*. Vol. 44. Wiley, 2000. DOI: 10.1002/0471721182.

[72] Erik Meijer and Jelmer Y. Ypma. "A Simple Identification Proof for a Mixture of Two Univariate Normal Distributions". In: *J. Classif.* 25.1 (2008), pp. 113–123. DOI: 10.1007/s00357-008-9008-6.

[73] Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*. Springer Berlin, 2009. DOI: 10.1007/978-3-540-87556-7.

[74] Bhavya Mor, Sunita Garhwal, and Ajay Kumar. "A Systematic Review of Hidden Markov Models and Their Applications". In: *Archives of Computational Methods in Engineering* 28 (May 2021), pp. 1429–1448. DOI: 10.1007/s11831-020-09422-4.

[75] Richard Nickl. *Inference for diffusions from low frequency measurements*. 2022. arXiv: 2210.13008 [math.ST].

[76] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 2015. DOI: doi:10.1515/9781400873173.

[77] Judith Rousseau. "Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparamatric estimation of the density". In: *Annals of Statistics* 38.1 (2010), pp. 146–180. DOI: 10.1214/09-AOS703.

[78] Gilles Royer. *Une initiation aux inégalités de Sobolev logarithmiques*. Collection SMF. Cours spécialisés ; Paris: Société mathématique de France, 1999, p. 114.

[79] Theofanis Sapatinas. "Identifiability of mixtures of power-series distributions and related characterizations". In: *Annals of the Institute of Statistical Mathematics* 47 (1995), pp. 447–459. DOI: 10.1007/BF00773394.

[80] Mathieu Sart. "Estimation of the transition dens ity of a Markov chain". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 50.3 (2014), pp. 1028–1068. DOI: 10.1214/13-AIHP551.

[81] Dominic Schuhmacher and Lutz Dümbgen. "Consistency of multivariate log-concave density estimators". In: *Statistics & Probability Letters* 80.5 (2010), pp. 376–380. DOI: https://doi.org/10.1016/j.spl.2009.11.013.

[82] Henry Teicher. "Identifiability of Mixtures". In: *The Annals of Mathematical Statistics* 32.1 (1961), pp. 244–248.

[83] D.M. Titterington et al. *Statistical Analysis of Finite Mixture Distributions*. Applied section. Wiley, 1985.

[84] van der Vaart A.W. & Wellner J.A. *Weak Convergence and Empirical Processes*. Springer New York, 1996. DOI: 10.1007/978-1-4757-2545-2.

[85] V. N. Vapnik and A. Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". In: *Theory of Probability & Its Applications* 16.2 (1971), pp. 264–280. DOI: 10.1137/1116025.

[86] Sergio Verdú. "Total variation distance and the distribution of relative information". In: *2014 Information Theory and Applications Workshop (ITA)*. 2014, pp. 1–3. DOI: 10.1109/ITA.2014.6804281.

[87] Yihong Wu and Pengkun Yang. "Optimal estimation of Gaussian mixtures via denoised method of moments". In: *The Annals of Statistics* 48.4 (2020), pp. 1981–2007. DOI: 10.1214/19-AOS1873.