Robust estimation for ergodic Markovian processes

Alexandre Lecestre[*]

June 28, 2023

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.

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.* [2] and Baraud & Birgé [4]. They provide a non-asymptotic deviation bound for the Hellinger distance h between P and their ρ-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 μ is any measure that dominates both P and Q, the result being independent of μ. It is shown in those articles that the ρ-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 ρ-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 $n^{-1}\sum\limits_{i=1}^{n} h^2(P_i,P)$ 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.

[*]This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 811017

We only focus on the theoretical aspects and performances of our estimation method. We prove a general result, Theorem 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 dimensions. 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 [18] 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.* [11] provide an overall survey of the different results in the literature. Most theoretical guarantees are either asymptotic or restricted to specific parametric models. Lehéricy [15] 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^*,π^*,Q^*,F^*), where $P_{w,Q,F}$ is defined by

$$P_{w,Q,F} = \sum_{1\leq k_1,\dots,k_L\leq K} w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \bigotimes_{l=1}^{L} F_{k_l}. \tag{1}$$

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 aim 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 choose automatically 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

$$d\left((w,Q,F),(\overline{w},\overline{Q},\overline{F})\right) \leq C\left(\overline{w},\overline{Q},\overline{F}\right) h^2\left(P_{w,Q,F}, P_{\overline{w},\overline{Q},\overline{F}}\right), \forall P_{w,Q,F} \in \mathscr{M}. \tag{2}$$

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 (2) when the densities are bounded. For two probability distributions P, Q dominated by a positive measure μ, we have

$$||p-q||_2^2 \leq 4(||p||_\infty + ||q||_\infty)h^2(P,Q), \tag{3}$$

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 that the statistician selects 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 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. We dedicate Section 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 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 $\{1,2,\dots,k\}$. 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 θ, α and β. 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 $x_+ = x \vee 0$.

2 Construction of the estimator and main result

Let $X_1,\dots,X_n$ 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 ρ-estimators developed by Baraud, Birgé and Sart in [2] and [4] 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.

2.1 Reminders of ρ-estimation

We denote by ψ the function given by

$$\psi : \left| \begin{array}{l} [0,+\infty] \to [-1,1] \\ x \mapsto \frac{x-1}{x+1} \end{array} \right. . \tag{4}$$

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 σ-finite measure μ. For $n \geq 1$, we denote by $\mathbf{T}_n$ and $\mathbf{\Upsilon}_n$ the functions given by

$$\mathbf{T}_n : \left| \begin{array}{l} \mathscr{X}^n \times \mathcal{M} \times \mathcal{M} \to [-1,1] \\ (\mathbf{x},q,q') \mapsto \sum\limits_{k=1}^{n} \psi\left(\sqrt{\frac{q'(x_i)}{q(x_i)}}\right) \end{array} \right. \tag{5}$$

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

$$\mathbf{\Upsilon}_n : \left| \begin{array}{l} \mathscr{X}^n \times \mathcal{M} \\ (\mathbf{x},q) \mapsto \sup_{q' \in \mathcal{M}} \mathbf{T}_n(\mathbf{x},q,q') \end{array} \right. . \tag{6}$$

For $\mathbf{x}$ in $\mathscr{X}^n$, we define the (nonvoid) set $\mathscr{E}_n(\mathbf{x})$ by

$$\mathscr{E}_n(\mathbf{x}) = \left\{ Q = q \cdot \mu \,\middle|\, q \in \mathcal{M}, \mathbf{\Upsilon}_n(\mathbf{x},q) < \inf_{q' \in \mathcal{M}} \mathbf{\Upsilon}_n(\mathbf{x},q') + 11.36 \right\}. \tag{7}$$

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 ρ-estimator on $\mathscr{M}$. The constant 11.36 is given by (7) and (19) in [4] but can be replaced by any smaller positive number.

One of the main results of ρ-estimation is Theorem 1 in [4]. For independent random variables $X_1,\dots,X_n$, any ρ-estimator $\hat{P} = \hat{P}(n,\mathbf{X},\mathscr{M})$ satisfies an inequality of the form

$$\mathbb{P}\left(\frac{C}{n} \sum_{i=1}^{n} h^2(P_i,\hat{P}) \leq \inf_{Q \in \mathscr{M}} n^{-1} \sum_{i=1}^{n} h^2(P_i,Q) + \frac{D_n(\mathscr{M}) + \xi}{n} \right) \geq 1 - e^{-\xi}, \tag{8}$$

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 ρ-dimension. It is an important feature of ρ-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 ρ-dimension is formally introduced in the appendix (Section B).

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 A.1.

Lemma 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 $\mathbf{P}$*

$$\mathbb{P}_{\mathbf{X}\sim\mathbf{P}}\left(l\left(\hat{\theta}(\mathbf{X}),\theta\right)\geq A+\frac{B+\xi^{\beta}}{n}\right)\leq e^{-\xi},\forall\xi>0,\tag{9}$$

then, when $\mathbf{X}$ has distribution $\mathbf{Q}$

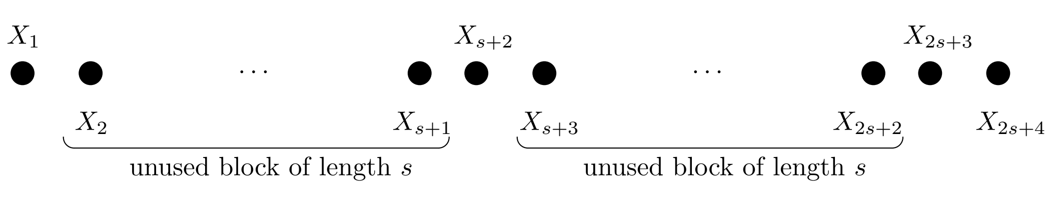
$$\mathbb{E}_{\mathbf{X}\sim\mathbf{Q}}\left[l\left(\hat{\theta}(\mathbf{X}),\theta\right)\right]\leq A+\frac{B+\left(2+\frac{3}{2}\mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right)\right)^{\beta}}{n},$$

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 ρ-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 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 $i \in [n(s,b)]$

$$X_i^{(s,b)} := X_{b+(i-1)(s+1)} \in \mathscr{X}, \forall i \in [n(s,b)], \tag{10}$$

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

$$\mathbf{P}_{s,b}^* := \mathcal{L}\left(\mathbf{X}^{(s,b)}\right) \text{ and } \mathbf{P}_{s,b}^{ind} := \bigotimes_{i=1}^{n(s,b)} \mathcal{L}\left(X_i^{(s,b)}\right). \tag{11}$$

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 the 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 ρ-estimator $\hat{P}\left(n(s,b), \mathbf{X}^{(s,b)}, \mathscr{M}\right)$ is defined in Section 2.1.

2. We denote by $\hat{P}_s = \hat{P}_s(\mathbf{X}, \mathscr{M})$ any element of $\mathscr{M}$ that satisfies

$$\sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b}, \hat{P}_s\right) \leq \inf_{Q \in \mathscr{M}} \sum_{b=1}^{s+1} n(s,b) h^2\left(\hat{P}_{s,b}, Q\right) + \iota, \tag{12}$$

   where ι is any fixed constant in $(0,1273]$.

2.3 Main result

We assume that the ρ-dimension function (see Section B) is uniformly bounded by a function $m \mapsto D_m(\mathscr{M}) \geq 1$ which is non-decreasing.

Theorem 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 (12) satisfies*

$$\mathbb{E}_{\mathbf{P}^*}\left[n^{-1} \sum_{i=1}^n h^2\left(P_i, \hat{P}_s\right)\right] \leq \frac{c_0}{n} \inf_{Q \in \mathscr{M}} \sum_{i=1}^n h^2(P_i, Q) \tag{13}$$
$$+ c_1 \frac{(s+1)}{n}\left[17 + D_{n(s,1)}(\mathscr{M})\right] + \frac{c_2}{n} \sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind}\right),$$

where $c_0 = 602$, $c_1 = 20056/4.7$ and $c_2 = 30084$.

The proof of this result is postponed to Section 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 (13) the following

$$C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq \frac{(s+1)D_{n(s,1)}(\mathscr{M})}{n} + n^{-1}\sum_{i=1}^{n} h^2\left(P_i,\overline{P}\right) \tag{14}$$
$$+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right),$$

where C is a universal positive constant. Up to the factor $(s+1)$, the first term in 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. We expect that for a choice of s that is sufficiently large the observations

$$\mathbf{X}^{(s,b)} := \left(X_b,X_{b+(s+1)},\ldots,X_{b+n(s,b)(s+1)}\right) \quad \text{with } b \in [s+1]$$

be nearly independent and consequently that the quantity $n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ be small compared to the first term.

2.4 Robust properties of our estimator

The robustness properties of ρ-estimators in the independent context are illustrated in Section 5 [4]. 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

$$Y_i = E_iX_i + (1-E_i)Z_i,\forall i \in [n]. \tag{15}$$

The next result shows that the mixing regime is not altered by independent contamination/outliers. It is proven in Section B.2.

Lemma 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 1 from this. We define p_i by $\mathbb{P}(E_i = 1) = p_i$ for $i \in [n]$.

Corollary 1. *Let $\hat{P}_s = \hat{P}_s\left(\mathbf{Y},\mathscr{M}\right)$ be the estimator given by (12). There is a positive universal constant C such that in the situation of Lemma 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 (11).

This result is proven in Section 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 ϵ_{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.

2.5 The particular case of Markov chains

Under the assumption that $X_1,\dots,X_n$ 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. *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

$$I(\sigma(X_i),\sigma(X_{i+1})) := \mathbf{K}\left(\mathcal{L}(X_i,X_{i+1})||\mathcal{L}(X_i)\otimes\mathcal{L}(X_{i+1})\right). \tag{16}$$

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 (10).*

This result is proven in Section 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 *coefficient of information* by Bradley [8]. This result also extends to hidden Markov models.

Lemma 4. *If* $(X_i,H_i)_{1\leq i\leq n}$ *is a HMM, we have*

$$\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right) \leq \sum_{i=2}^{n} I(\sigma(H_{i-1}),\sigma(H_i)).$$

In particular for all s *in* $\{0,1,\dots,s_{\max}\}$ *and all* b *in* $[s+1]$,

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{(s,b)}_{ind}\right) \leq \sum_{i=1}^{n(s,b)-1} I\left(\sigma(H_{b+(i-1)(s+1)}),\sigma(H_{b+i(s+1)})\right).$$

The proof of this result is postponed to Section 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

$$n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \leq Ce^{-rs},$$

for all s in $\{0,1,\dots,s_{\max}\}$. For $s \geq r^{-1}\log n$ 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 (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 Estimation of the invariant distribution of a diffusion process

We consider some diffusion processes that have been investigated by Royer [19] and use the same vocabulary that they introduced.

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

$$dY_t = dB_t - \nabla U(Y_t)dt, \tag{17}$$

where $B = (B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Its solution are called Kolmogorov processes in Royer [19]. We assume that U satisfies the following.

Assumption 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*

$$\inf_{x\in\mathbb{R}^d} \left\{ ||\nabla U(x)||_2^2 - \mathrm{Tr}\left(U''(x)\right) \right\} > -\infty, \tag{18}$$

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

$$\overline{p}(x) = Z^{-1}\exp(-2U(x)) \text{ with } Z = \int_{\mathbb{R}^d} e^{-2U(x)}dx. \tag{19}$$

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 [19]).

Lemma 5. *Let $(Y_t)_{t\geq 0}$ be a stationary solution of the Langevin equation associated to a convex function U that satisfies Assumption 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 C.2. We aim to estimate $\overline{P}$ from discrete observations of a stationary Kolmogorov process.

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 (17) for some unknown convex function U that satisfies Assumption 1 and $t_{i+1} = t_i + \Delta_t$ for all $i \in [n-1]$. As a consequence of (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 log-concave densities

We refer to Kim & Samworth [12] for the problem of estimating of log-concave densities from i.i.d. observations in low dimensions ($d \in [3]$). Kur *et al.* [13] 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 σ-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 Σ, we denote by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ the smallest and largest eigenvalues respectively of Σ. 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

$$\mathrm{Sym}(\lambda_-,\lambda_+) = \left\{ \Sigma \text{ covariance matrix}, \lambda_- \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq \lambda_+ \right\}.$$

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 ϵ-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 C.3 and based on the work of Kim & Samworth [12] for $d \in [3]$ and Kur *et al.* [13] for $d \geq 4$.

Lemma 6. *For all positive ϵ there exists an ϵ-net $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ such that*

$$|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]| \leq \begin{cases} \dfrac{9}{\eta_1} \dfrac{M(\lambda_+ - \lambda_-)}{\lambda_-^{3/2}} e^{\overline{K_1} \epsilon^{-1/2}} \textit{ for } d = 1, \\ \dfrac{3^8 \pi}{\eta_2^3} \dfrac{M^2 (\lambda_+ - \lambda_-)^2 \lambda_+}{\lambda_-^4} e^{\overline{K_2} \epsilon^{-1} \log_{++}^{3/2}(1/\epsilon)} \textit{ for } d = 2, \\ \dfrac{2^7 3^{27/2} \pi^3}{\eta_3^6} \dfrac{M^3 (\lambda_+ - \lambda_-)^3 \lambda_+^3}{\lambda_-^{15/2}} e^{\overline{K_3} \epsilon^{-2}} \textit{ for } d = 3, \end{cases}$$

where η_d and $\overline{K}_d$ are constants given in Theorem 4 [12] that only depend on d, and with $\log_{++}(x) = \max(1, \log x)$. *For $d \geq 4$ and all positive ϵ there exists an ϵ-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 η_d and $\overline{K}_d$ are constants that only depend on d.

3.3.1 The case $d \in \{1,2,3\}$

Let $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ be a ϵ-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ that satisfies the bound given in Lemma 6 for

$$\lambda_+ = \lambda_-^{-1} = M := \begin{cases} \exp\left(\overline{K}_1 (n/\log n)^{1/5} \right) \text{ for } d = 1, \\ \exp\left(\overline{K}_2 n^{1/3} \log^{2/3} n \right) \text{ for } d = 2, \\ \exp\left(\overline{K}_3 (n/\log n)^{1/2} \right) \text{ for } d = 3, \end{cases} \tag{20}$$

and

$$\epsilon := \begin{cases} n^{-2/5} \log^{2/5} n \text{ for } d = 1, \\ n^{-1/3} \log^{5/6} n \text{ for } d = 2, \\ n^{-1/4} \log^{1/4} n \text{ for } d = 3. \end{cases} \tag{21}$$

The following result holds and its proof can be found in Section C.1.

Theorem 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 ρ-estimator $\hat{P}_s$ given by (12) with $\mathscr{M} = \mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ satisfies for all $\overline{P} \in \mathscr{P}_X$*

$$C_d\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\mathscr{F}_{\lambda_-,\lambda_+,M}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(P_i,\overline{P}\right) \tag{22}$$

$$+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$$

$$+\begin{cases} n^{-4/5}\left(\log^{4/5} n + s\log^{-1/5} n\right) \textit{ for } d=1, \\ n^{-2/3}\left(\log^{5/3} n + s\log^{2/3} n\right) \textit{ for } d=2, \\ n^{-1/2}\left(\log^{1/2} n + s\log^{-1/2} n\right) \textit{ for } d=3, \end{cases}$$

for positive constants C_1,C_2,C_3. In particular if the model described in Section 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

$$C(U,d,\Delta_t)\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq \begin{cases} n^{-4/5}\left(\log^{4/5} n + s\log^{-1/5} n\right) \textit{ for } d=1, \\ n^{-2/3}\left(\log^{5/3} n + s\log^{2/3} n\right) \textit{ for } d=2, \\ n^{-1/2}\left(\log^{1/2} n + s\log^{-1/2} n\right) \textit{ for } d=3, \end{cases}$$

where $\overline{P}$ is the invariant distribution given by (19).

Inequality (22) is a consequence of Theorem 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 ϵ given by (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 [12] in the i.i.d. case. Our power of $\log n$ is even better for $d=3$. As mentioned in Section 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 (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 2.4. 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 $i \in [n]$,

$$X_i = E_iY_{t_1+(i-1)\Delta_t} + (1-E_i)Z_i,$$

where $(Y_t)_t$ is a stationary solution of the Langevin equation (17) for some unknown convex function U that satisfies Assumption 1.

Corollary 2. *Let $\hat{P}_s$ be the estimator given by (12) with $\mathscr{M} = \mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$. If $E_1,\ldots,E_n,Z_1,\ldots,Z_n$ 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

$$C(U,d,\Delta_t)\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq n^{-1}\sum_{i=1}^{n}(1-p_i) \tag{23}$$

$$+\begin{cases} n^{-4/5}\left(\log^{4/5} n + s\log^{-1/5} n\right) \text{ for } d=1, \\ n^{-2/3}\left(\log^{5/3} n + s\log^{2/3} n\right) \text{ for } d=2, \\ n^{-1/2}\left(\log^{1/2} n + s\log^{-1/2} n\right) \text{ for } d=3, \end{cases},$$

where $p_i = \mathbb{P}(E_i = 1)$ 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 (23).

3.3.2 The case $d \geq 4$

Let $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ be an ϵ-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ that satisfies the bound given in Lemma 5 with

$$\lambda_+ = \lambda_-^{-1} = \exp\left(\frac{\epsilon^{-(d-1)}\log^{(d+1)(d+2)/2}(\epsilon^{-1})}{d^2}\right) \tag{24}$$

$$M = \exp\left(\frac{\epsilon^{-(d-1)}\log^{(d+1)(d+2)/2}(\epsilon^{-1})}{d}\right), \tag{25}$$

with

$$\epsilon = n^{-\frac{1}{d+1}}\log^{\frac{1}{d+1}+\frac{d+2}{2}} n. \tag{26}$$

The following result holds and its proof can be found in Section C.1.

Theorem 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 ρ-estimator $\hat{P}_s$ given by (12) with $\mathscr{M} = \mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ satisfies for all $\overline{P} \in \mathscr{P}_X$*

$$C_d\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\mathscr{F}_{\lambda_-,\lambda_+,M}\right) + n^{-1}\sum_{i=1}^{n}h^2\left(P_i,\overline{P}\right)$$
$$+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$$
$$+ n^{-\frac{2}{d+1}}\left(\log^{d+2+\frac{2}{d+1}} n + s\log^{d+1+\frac{2}{d+1}} n\right).$$

In particular if the model described in Section 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 (19).

This result is equivalent to Theorem 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.* [13] obtain for the estimation of log-concave estimation from i.i.d. observations. We can derive a result equivalent to Corollary 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.

4 Hidden Markov models

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 (1), and we have $\mathcal{L}(Y_i,Y_{i+1},\dots,Y_{i+L-1}) = P_L$ 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 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 $\mathcal{L}\left(Y_i,\dots,Y_{i+L-1}\right)$ converges exponentially fast to the distribution

$$P^* = P_{\pi^*,Q^*,F^*}, \tag{27}$$

where π^* is the only invariant distribution with respect to Q^*.

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

$$X_i = \left(Y_i,Y_{i+1},\dots,Y_{i+L-1}\right),i=1,\dots,n, \tag{28}$$

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 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

$$\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right) := \left\{P_{w,Q,F};\ \begin{array}{c}\forall k \in [K], w \in \mathcal{W}_K,\\ Q \in \mathcal{T}_K, F_k \in \overline{\mathscr{F}}_k\end{array}\right\} \subset \mathscr{P}_X, \tag{29}$$

where $P_{w,Q,F}$ is given by (1),

$$\mathcal{T}_K = \left\{Q \in [0,1]^{K\times K}; \sum_{j=1}^{K} Q_{ij} = 1, \forall i \in \{1,\dots,K\}\right\}, \tag{30}$$

$$\text{and } \mathcal{W}_K = \left\{w \in [0,1]^K; w_1+\dots+w_K = 1\right\}. \tag{31}$$

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 $\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$.

4.3 Estimation

Let ν be a σ-finite measure on $(\mathscr{Y},\mathcal{Y})$ and we denote by μ the associated σ-finite measure on $(\mathscr{X},\mathcal{X})$ given by $\mu := \nu^{\otimes L}$. We consider emission models that satisfy the following.

Assumption 3. *We dispose of countable sets $\mathcal{F}_i,i=1,\dots,K$ of probability density functions (with respect to ν) 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 ϵ-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
$$\overline{V} := \sum_{1 \le k_1, \dots, k_L \le K} V_{k_1,\dots,k_L}. \tag{32}$$

We refer to van der Vaart & Wellner [21] (Section 2.6.5) and Baraud *et al.* [2] (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
$$V(\mathcal{F}) \le 1 + \log_2(|\mathcal{F}|). \tag{33}$$
Therefore we can consider finite ϵ-nets as we did in Section 3. We also show in Section 4.3.2 that exponential families satisfy our assumption.

We consider countable approximations of $\mathcal{W}_K$ and $\mathcal{T}_K$ given by
$$\mathcal{W}_{\delta,K} := \mathcal{W}_K \cap ([\delta,1] \cap \mathbb{Q})^K \text{ and } \mathcal{T}_{\delta,K} := \mathcal{T}_K \cap ([\delta,1] \cap \mathbb{Q})^{K \times K}, \tag{34}$$
for $0 < \delta \le 1/K$. We define $\mathscr{H}_\delta$ by
$$\mathscr{H}_\delta := \{P_{w,Q,f}; w \in \mathcal{W}_{\delta,K}, Q \in \mathcal{T}_{\delta,K}, f_k \in \mathscr{F}_k, \forall i \in [K]\}, \tag{35}$$
where the sets $(\mathscr{F}_k)_{1 \le k \le K}$ are given in Assumption 3. This lower bound δ is a technicality for bounding the dimension of our model. We define the countable set of distributions
$$\mathscr{M}_\delta := \left\{ P_{w,Q,F} \in \mathscr{H}_\delta; \exists P_{w',Q',F'} \in \overline{\mathscr{M}}, \begin{array}{r} h^2(Q_{k\cdot}, Q'_{k\cdot}) \le (K-1)\delta \\ h(F_k, F'_k) \le \epsilon, \forall k \in [K], \\ h^2(w, w') \le (K-1)\delta, \end{array} \right\}, \tag{36}$$
which is a good approximation of $\overline{\mathscr{M}}$ for small values of δ and ϵ. We denote by $\hat{P}_{s,\delta}$ the estimator
$$\hat{P}_{s,\delta} := \hat{P}_s(\mathscr{M}_\delta, \mathbf{X}), \tag{37}$$
as defined by (12). The following theorem is proven in Section D.1.

Theorem 4. *Let $N \ge K + L$ and $Y_1, \dots, Y_N$ be arbitrary random variables. Under Assumption 3, let $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with*
$$\delta = \frac{\overline{V}}{n(s,1)(K-1)} \wedge \frac{1}{K}. \tag{38}$$
There exists a positive constant C such that for all $\overline{P} \in \mathscr{P}_X$,
$$C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \le h^2\left(\overline{P}, \overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(\overline{P}, P_i\right) + n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right)$$
$$+ L\epsilon^2 + (s+1)L\overline{V}\frac{\log n}{n}. \tag{39}$$
In particular under Assumption 2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \ge c(Q^*)\log n \vee (L-1)$ we have
$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \le h^2\left(P^*, \overline{\mathscr{M}}\right) + L\epsilon^2 + L\overline{V}\frac{s\log n}{n}, \tag{40}$$
where P^* is given by (27).

Inequality (39) is a consequence of Theorem 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 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

$$\begin{aligned} &\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K\right) \\ &:= \left\{ P_{w,Q,F} \in \mathscr{H}\left(K,\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K\right);\ \begin{array}{c} Q \text{ irreducible },\\ Q \text{ aperiodic},\\ \text{and } w = Qw \end{array} \right\}. \end{aligned} \tag{41}$$

The first case is when we satisfy Assumption 3 with $\epsilon = 0$. In that situation and for P^* in $\overline{\mathscr{M}} = \mathscr{H}\left(K,\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K\right)$ the first two terms in (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 $\mathscr{H}^*\left(K,\overline{\mathscr{F}}_1,\ldots,\overline{\mathscr{F}}_K\right)$. This means that up to a logarithmic term we achieve the optimal rate $1/n$ in the independent context (see Birgé [6]). As mentioned in Section 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 ϵ and we proceed as in Section 3. We obtain a convergence rate taking ϵ that goes to 0 with n at a rate that balances the last two terms in (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 4.3.1.

In order to illustrate the robustness of our estimators we consider the situation of Section 2.4. 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 $\mathbf{Y}'$ satisfy Assumption 2. The following result is proven in Section D.2.

Corollary 3. *Let $N \geq K + L$ and $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with δ given by (38). If $E_1,\ldots,E_N,Z_1,\ldots,Z_N$ 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*

$$\begin{aligned} C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + \frac{L}{N}\sum_{i=1}^{N}(1-p_i) \\ + L\epsilon^2 + L\overline{V}\frac{s\log n}{n}, \end{aligned} \tag{42}$$

where $p_i = \mathbb{P}(E_i = 1)$ for all $i \in [N]$ and δ is given by (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

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L\left[\alpha_{cont} + \epsilon^2 + \overline{V}\frac{s\log n}{n}\right], \tag{43}$$

for $s \geq c(Q^*) \log n$. The bound on the convergence rate is not deteriorated as long as the contamination rate α_{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

$$C(Q^*) \mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq L\left[\frac{|I|}{N} + \epsilon^2 + \overline{V}\frac{s \log n}{n}\right], \tag{44}$$

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 $\epsilon^2 + \overline{V}\frac{s \log n}{n}$.

4.3.1 log-concave emission densities

We use results and notation given in Section 3. Let d be a positive integer and $\epsilon \in (0,1)$. Let $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]$ be an ϵ-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ that satisfies the bound given in Lemma 5. We take $\overline{\mathscr{F}}_k = \mathscr{F}_{\lambda_-,\lambda_+,M}$ for all $k \in [K]$ and satisfy Assumption 3 with

$$\overline{V} = K^L \left(1 + L \log_2\left(|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|\right)\right). \tag{45}$$

We take $\overline{\mathscr{M}} = \mathscr{H}\left(K, \mathscr{F}_{\lambda_-,\lambda_+,M}, \dots, \mathscr{F}_{\lambda_-,\lambda_+,M}\right)$. We distinguish the two cases $d \in \{1,2,3\}$ and $d \geq 4$.

For $d \in \{1,2,3\}$ we take λ_+, λ_-, M as in (20) and ϵ as in (21). The following result holds and its proof can be found in Section D.3.

Theorem 5. *Let $N \geq K + L$ and $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with δ given by (38). There exist positive constants C_1, C_2, C_3 such that for all $\overline{P} \in \mathscr{P}_X$,*

$$\begin{aligned} C_d \mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq\ & h^2\left(\overline{P}, \overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) \\ & + n^{-1}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) \\ & + (s+1) L^2 K^L \times \begin{cases} n^{-4/5} \log^{4/5} n \textit{ for } d = 1, \\ n^{-2/3} \log^{5/3} n \textit{ for } d = 2, \\ n^{-1/2} \log^{1/2} n \textit{ for } d = 3. \end{cases} \end{aligned} \tag{46}$$

In particular under Assumption 2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*) \log n$ we have

$$C(Q^*) \mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq h^2\left(P^*, \overline{\mathscr{M}}\right) + s L^2 K^L \times \begin{cases} n^{-4/5} \log^{4/5} n \textit{ for } d = 1, \\ n^{-2/3} \log^{5/3} n \textit{ for } d = 2, \\ n^{-1/2} \log^{1/2} n \textit{ for } d = 3, \end{cases}$$

where P^* is given by (27).

Inequality (46) is a consequence of Theorem 4 and does not require any assumption on the data. We can deduce convergence rates over the class $\mathscr{H}^*\left(K, \mathscr{F}_d, \dots, \mathscr{F}_d\right)$, where $\mathscr{F}_d$ is the set of distributions with log-concave densities defined in Section 3. For the optimal choice of s, we have

$$C(Q^*) \mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq L^2 K^L \times \begin{cases} n^{-4/5} \log^{9/5} n \text{ for } d = 1, \\ n^{-2/3} \log^{8/3} n \text{ for } d = 2, \\ n^{-1/2} \log^{3/2} n \text{ for } d = 3, \end{cases} \tag{47}$$

for all P^* in $\mathscr{H}^*\left(K,\mathscr{F}_d,\dots,\mathscr{F}_d\right)$. We see that we have a worse power of $\log n$ compared to Theorem 2. It comes from an additional logarithmic factor in the dimension term for HMMs. Corollary 3 tells us our estimator is also robust to contamination and outliers. Let us illustrate it for $d = 1$. We can see from (43) that our bound is not significantly worse as long as the contamination rate α_{cont} is of order not larger than $n^{-4/5}\log^{9/5} n$. Similarly (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 λ_+,λ_-^{-1} as in (24), M as in (25) and ϵ as in (26). The following result holds and its proof can be found in Section D.3.

Theorem 6. *Let $N \geq K + L$ and $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with δ given by (38). There exist a positive constant C_d such that for all $\overline{P} \in \mathscr{P}_X$,*

$$C_d\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\overline{\mathscr{M}}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(P_i,\overline{P}\right)$$
$$+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$$
$$+ (s+1)L^2K^Ln^{-\frac{2}{d+1}}\log^{d+2+\frac{2}{d+1}} n.$$

In particular under Assumption 2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + sL^2K^Ln^{-\frac{2}{d+1}}\log^{d+2+\frac{2}{d+1}} n, \tag{48}$$

where P^* is given by (27).

Inequality (46) does not require any assumption on the data. We can deduce convergence rates over the class $\mathscr{H}^*\left(K,\mathscr{F}_d,\dots,\mathscr{F}_d\right)$. For the optimal choice of s, we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq L^2K^Ln^{-\frac{2}{d+1}}\log^{d+3+\frac{2}{d+1}} n$$

for all $P^* \in \mathscr{H}^*\left(K,\mathscr{F}_d,\dots,\mathscr{F}_d\right)$. As for $d \leq 3$, we have the same rate as in Section 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 (43) that our bound is not significantly worse as long as the contamination rate α_{cont} is of order not larger than $n^{-\frac{2}{d+1}}\log^{d+3+\frac{2}{d+1}} n$. Similarly (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.

4.3.2 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

$$\mathcal{E}\left(\overline{\Theta},\eta,T,d,B\right) := \left\{f_\theta : x \mapsto e^{\langle\eta(\theta),T(x)\rangle+A(\theta)+B(x)}; \theta \in \overline{\Theta}\right\}, \tag{49}$$

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 ν.

Assumption 4. *For all $k \in \{1,\ldots,K\}$,*

1. *$\overline{\mathscr{F}}_k$ is of the form*
$$\overline{\mathscr{F}}_k = \left\{ q \cdot \nu ; q \in \mathcal{E}\left(\overline{\Theta}_k, \eta_k, T_k, d_k, B_k\right)\right\}, \tag{50}$$

2. *Θ_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 D.4 and shows that the last assumption is sufficient to satisfy our main assumption.

Proposition 1. *Under Assumption 4, we satisfy Assumption 3 with $\epsilon = 0$ and $V_{k_1,\ldots,k_L} = 3 + \sum_{k_l=1}^{L} d_{k_l}$. Therefore we have*
$$\overline{V} = 3K^L + LK^{L-1}\left(d_1 + \cdots + d_K\right). \tag{51}$$

We can see that the constant $\overline{V}$ does not depend on $\mathscr{X}$ but on the dimensions $d_1,\ldots,d_K$ 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 $\overline{\mathscr{M}} \subset \mathscr{H}\left(K, \overline{\mathscr{F}}_1, \ldots, \overline{\mathscr{F}}_K\right)$.

Corollary 4. *Let $N \geq K + L$ and $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with δ given by (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 2, there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have
$$\begin{aligned} C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \leq{}& h^2\left(P^*, \overline{\mathscr{M}}\right) \\ &+ LK^{L-1}\left(K + L(d_1 + \cdots + d_K)\right)\frac{s\log n}{n}, \end{aligned} \tag{52}$$

where P^* is given by (27).

This result is a direct consequence of Theorem 4 and Proposition 1. We can deduce a bound on the convergence rate over $\mathscr{H}^*\left(K, \overline{\mathscr{F}}_1, \ldots, \overline{\mathscr{F}}_K\right)$. 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 $\mathscr{H}^*\left(K, \overline{\mathscr{F}}_1, \ldots, \overline{\mathscr{F}}_K\right)$. We obtain the optimal $1/n$ rate with respect to the squared Hellinger distance, up to a logarithmic factor. Corollary 3 shows that our estimator is also robust to contamination and outliers. From (43) we see that our bound is not significantly worse as long as the contamination rate α_{cont} is of order not larger than $n^{-1}\log^2 n$. Similarly, we get from (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 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

$$g_{z,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{(z-m)^T\Sigma^{-1}(z-m)}{2}\right), \tag{53}$$

where $|\Sigma|$ denotes the determinant of Σ. 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

$$\begin{aligned} T(x) &= \left(x, \left(x_i^2\right)_{1\leq i\leq d}, \left(x_ix_j\right)_{1\leq i<j\leq d}\right) \text{ and} \\ \eta(z,\Sigma) &= \left(\Sigma^{-1}z, -\frac{1}{2}\left(\Sigma_{ii}^{-1}\right)_{1\leq i\leq d}, -\left(\Sigma_{ij}^{-1}\right)_{1\leq i<j\leq d}\right). \end{aligned}$$

For a fixed Σ 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 4.

Theorem 7. *Let $N \geq K + L$ and $Y_1, \ldots, Y_N$ be arbitrary random variables.*

- *Let $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with $\overline{\mathscr{M}} = \mathscr{H}\left(K,\mathscr{G}_d,\ldots,\mathscr{G}_d\right)$ and δ given by (38). There exists a positive constant C such that for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\overline{\mathscr{M}}\right) &+ n^{-1}\sum_{i=1}^{n} h^2\left(\overline{P},P_i\right) \\ & n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ &+ (s+1)L^2K^Ld(d+3)\frac{\log n}{n}. \end{aligned} \tag{54}$$

  In particular under Assumption 2 there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \leq h^2\left(P^*,\mathscr{M}\right) + (s+1)L^2K^Ld(d+3)\frac{\log n}{n},$$

  where P^* is given by (27).

- *Let $\hat{P}_s = \hat{P}_{s,\delta}$ be the estimator given by (37) with $\overline{\mathscr{M}} = \mathscr{H}\left(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma)\right)$ and δ given by (38). There exists a positive constant C such that for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \leq h^2\left(\overline{P},\overline{\mathscr{M}}\right) &+ n^{-1}\sum_{i=1}^{n} h^2\left(\overline{P},P_i\right) \\ &+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ &+ (s+1)L^2K^Ld\frac{\log n}{n}, \end{aligned} \tag{55}$$

for any Σ in $Cov_{+*}(d)$. In particular under Assumption 2 there exist positive constants $C(Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have

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

where P^* is given by (27).

Inequalities (54) and (55) are consequences of Corollary 4 and do not require any assumption on the data. We deduce bounds on the convergence rate of our estimator over $\mathscr{H}^*\left(K,\mathscr{G}_d,\ldots,\mathscr{G}_d\right)$ and $\mathscr{H}^*\left(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma)\right)$. For the optimal choice of s we obtain the rate $n^{-1}\log^2 n$ with respect to the squared Hellinger distance both for $P^* \in \mathscr{H}^*\left(K,\mathscr{G}_d,\ldots,\mathscr{G}_d\right)$ and $P^* \in \mathscr{H}^*\left(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma)\right)$. This rate is optimal up to a logarithmic factor. We can see that the dependence on the dimension d is linear for the model $\mathscr{H}\left(K,\mathscr{G}_{loc}(\Sigma),\ldots,\mathscr{G}_{loc}(\Sigma)\right)$ while its quadratic for $\mathscr{H}^*\left(K,\mathscr{G}_d,\ldots,\mathscr{G}_d\right)$.

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 ρ-estimators of π^*, Q^* and F^* if $P_{\hat{w},\hat{Q},\hat{F}} = \hat{P}_{s,\delta}$ is an estimator of P^* given by (37). If we consider models of densities that are uniformly bounded, we can use (3) and Theorem 9 of Lehéricy [15] to deduce risk bounds for the parameter estimators. It is also possible to use the results of Ibragimov and Has'minskiĭ [10] for regular parametric models.

We consider that Assumption 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_{θ_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_θ given by (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 ϕ 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

$$\overline{\mathscr{M}} := \left\{P_\phi = p(\cdot;\phi)\cdot\mu; \phi\in\overline{\Phi}\right\} \tag{56}$$

and

$$p(\mathbf{x};\phi) = \sum_{1\leq k_1,\ldots,k_L\leq K} w_{k_1}Q(k_2|k_1)\ldots Q(k_L|k_{L-1})\prod_{l=1}^{L} f_{\theta_{k_l}}(x_l).$$

We need the following assumption to make sure we can deduce ϕ from P_ϕ.

Assumption 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 η_k and A_k are of class $\mathcal{C}^1$ on $\overline{\Theta}_k$;*

- *for all θ_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 D.5 and shows that under some conditions we can deduce the parameters from the distribution P_ϕ.

Proposition 2. *Under Assumption 5 the information matrix I function given by*

$$I_{ij} : \phi \mapsto I\left(\phi\right)_{ij} = \int_{\mathscr{X}^L} \partial_{\phi_i} p(\mathbf{x};\phi) \partial_{\phi_j} p(\mathbf{x};\phi) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}$$

is well-defined and continuous on $\overline{\Phi}$. We define the subset $\Phi^* \subset \overline{\Phi}$ by

$$\Phi^* := \left\{ \overline{\phi} \in \overline{\Phi}; \begin{array}{c} I\left(\overline{\phi}\right) \text{ is definite positive and} \\ \inf_{\substack{||\overline{\phi} - \phi|| \geq a \\ \phi \in \overline{\Phi}}} h^2\left(P_{\overline{\phi}}, P_\phi\right) > 0, \forall a > 0, \end{array} \right\} \tag{57}$$

For all $\phi^* \in \Phi^*$, there exists a positive constant $C(\phi^*)$ such that

$$C(\phi^*) \left[||w^* - w||_2^2 + ||Q^* - Q||_2^2 + \sum_{k=1}^K ||\theta_k^* - \theta_k||_2^2 \wedge 1 \right] \leq h^2\left(P_{\phi^*}, P_\phi\right), \tag{58}$$

for all ϕ 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 ϕ^* induced by the Hellinger distance on $\mathscr{M}$. The next result is a consequence of Proposition 2 and Corollary 4.

Theorem 8. *Let $N \geq K + L$ and $Y_1, \dots, Y_N$ be arbitrary random variables. Let $P_{\hat{\phi}} = \hat{P}_{s,\delta}$ be the estimator given by (37) with δ given by (38). Under Assumption 5, for all $\overline{\phi} \in \Phi^*$ there exists a positive constant $C(\overline{\phi})$ such that*

$$\begin{aligned} C\left(\overline{\phi}\right) \mathbb{E} &\left[||\overline{w} - \hat{w}||_2^2 + \left|\left|\overline{Q} - \hat{Q}\right|\right|_2^2 + \sum_{k=1}^K \left|\left|\overline{\theta}_k - \hat{\theta}_k\right|\right|_2^2 \wedge 1 \right] \\ &\leq n^{-1} \sum_{i=1}^n h^2\left(P_{\overline{\phi}}, P_i\right) + n^{-1} \sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind}\right) \\ &\quad + (s+1) L K^{L-1} \left(K + L(d_1 + \dots + d_K)\right) \frac{\log n}{n}. \end{aligned} \tag{59}$$

In particular under Assumption 2, there exist positive constants $C(\overline{\phi}, Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*) \log n$ we have

$$\begin{aligned} C\left(\overline{\phi}, Q^*\right) \mathbb{E} &\left[||\overline{w} - \hat{w}||_2^2 + \left|\left|\overline{Q} - \hat{Q}\right|\right|_2^2 + \sum_{k=1}^K \left|\left|\overline{\theta}_k - \hat{\theta}_k\right|\right|_2^2 \wedge 1 \right] \\ &\leq h^2\left(P^*, P_{\overline{\phi}}\right) + L K^{L-1} \left(K + L(d_1 + \dots + d_K)\right) \frac{s \log n}{n}, \end{aligned} \tag{60}$$

where P^* is given by (27).

Inequality (59) is a consequence of Proposition 2 and Corollary 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 Φ^* 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 (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

$$\overline{\mathscr{E}} := \left\{ f_\theta \cdot \nu; f_\theta \in \mathcal{E}\left(\overline{\Theta}, \mathrm{id}_\Theta, -\mathrm{id}_{\mathscr{X}}, 1, 0\right)\right\} \tag{61}$$

where $\overline{\Theta} = (0,\infty)$, $\mathscr{X} = [0,\infty)$, ν 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 5, the last condition being a direct consequence of the dominated convergence theorem. We define $\overline{\Phi}$ by

$$\overline{\Phi} = O_K^{K+1} \times \left\{\theta \in \Theta^K; \theta_1 > \theta_2 > \cdots > \theta_K\right\}, \tag{62}$$

and $\overline{\mathscr{M}}$ as in (56). The condition on the parameters θ ensures identifiability over $\overline{\Phi}$ and $\overline{\Phi}^* = \overline{\Phi}$. The choice $L = 3$ is enough to obtain the result of Proposition 2. The next theorem is proven in Section D.6.

Theorem 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 (37) with δ given by (38). For any $\overline{\phi}$ in $\overline{\Phi}$ there exists a positive constant $C(\overline{\phi})$ such that we have*

$$\begin{aligned} &C\left(\overline{\phi}\right)\mathbb{E}\left[\|\overline{w} - \hat{w}\|_2^2 + \left\|\overline{Q} - \hat{Q}\right\|_2^2 + \sum_{k=1}^K \left(\overline{\theta}_k - \hat{\theta}_k\right)^2 \wedge 1\right] \\ &\leq n^{-1}\sum_{i=1}^n h^2\left(P_{\overline{\phi}}, P_i\right) + n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}\|\mathbf{P}^{ind}_{s,b}\right) + (s+1)K^3\frac{\log n}{n}. \end{aligned}$$

In particular under Assumption 2, there exist positive constants $C(\overline{\phi}, Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have

$$\begin{aligned} &C\left(\overline{\phi}, Q^*\right)\mathbb{E}\left[\|\overline{w} - \hat{w}\|^2 + \left\|\overline{Q} - \hat{Q}\right\|^2 + \sum_{k=1}^K \left(\overline{\theta}_k - \hat{\theta}_k\right)^2 \wedge 1\right] \\ &\leq h^2\left(P^*, P_{\overline{\phi}}\right) + sK^3\frac{\log n}{n}, \end{aligned} \tag{63}$$

where P^* is given by (27).

Our different parameter estimators all reach the usual parametric rate up to a logarithmic factor. One can notice that the ordering of the θ_k in (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 Φ^* given by (57).

4.3.3 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 ν be the Lebesgue measure on $\mathbb{R}$ and α be in (0,1). We denote by f_α the probability density function with respect to ν 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

$$\overline{\mathscr{M}} = \left\{P_{w,q,z}; w,q_{12},q_{21} \in [0,1], z \in \mathbb{R}\right\},$$

where

$$\begin{aligned} P_{w,q,z} &= wF_{\alpha,0} \otimes \left[(1-q_{12})F_{\alpha,0} + q_{12}F_{\alpha,z}\right] \\ &\quad + (1-w)F_{\alpha,z} \otimes \left[q_{21}F_{\alpha,0} + (1-q_{21})F_{\alpha,z}\right]. \end{aligned}$$

The distributions $P_{w,q,z}$ correspond to translation hidden Markov models with one known location parameter. The following result is proven in Section D.8 and shows that we can deduce the parameters from the distribution $P_{w,q,z}$.

Proposition 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 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 D.7.

Theorem 10. *Let $N \geq K+2$ and $P_{\hat{w},\hat{q},\hat{z}} = \hat{P}_{s,\delta}$ be the estimator given by (37) with δ given by (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*

$$\begin{aligned} &C\left(\alpha,\overline{z},\overline{w},\overline{q}\right) \mathbb{E}\left[(\overline{w}-\hat{w})^2 + (\overline{q}_{12}-\hat{q}_{12})^2 + (\overline{q}_{12}-\hat{q}_{12})^2 + (|\overline{z}-\hat{z}| \wedge 1)^2\right] \\ &\leq \frac{1}{n}\sum_{i=1}^{n} h^2\left(P_{\overline{w},\overline{q},\overline{z}}, P_i\right) + \frac{1}{n}\sum_{b=1}^{s+1} \mathbf{K}\left(\mathbf{P}^*_{s,b} || \mathbf{P}^{ind}_{s,b}\right) + (s+1)\frac{\log n}{n}. \end{aligned} \tag{64}$$

In particular under Assumption 2, there exist positive constants $C(\overline{\phi},Q^*)$ and $c(Q^*)$ such that for $s \geq c(Q^*)\log n$ we have

$$\begin{aligned} &C\left(\overline{\phi}, Q^*\right)\mathbb{E}\left[(\overline{w}-\hat{w})^2 + (\overline{q}_{12}-\hat{q}_{12})^2 + (\overline{q}_{21}-\hat{q}_{21})^2 + (|\overline{z}-\hat{z}| \wedge 1)^{1-\alpha}\right] \\ &\qquad \leq h^2\left(P^*, P_{\overline{w},\overline{q},\overline{z}}\right) + \frac{s\log n}{n}, \end{aligned} \tag{65}$$

where P^* is given by (27).

Inequality (64) does not require any assumption on the data. It is a consequence of Proposition 3 and Theorem 4. We can deduce convergence rates for our parameter estimators from (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 [10] (Chapter VI), noticing that f_α 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_α 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_α is not square integrable for $\alpha \in [1/2,1)$.

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.

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

$$n_1(s,b) = \left\lfloor \frac{n_1+s+1-b}{1+s} \right\rfloor. \tag{66}$$

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 ρ-dimension function (see Section 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 (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 ρ-estimator $\hat{P} = \hat{P}\left(n_2,\mathbf{X}^{(2)},\widehat{\mathscr{M}_S}\right)$ given by (7). We denote by $\hat{s}$ the value of s such that $\hat{P} = \hat{P}_{\hat{s}}$ and we write

$$\hat{P} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right). \tag{67}$$

We make the following assumption.

Assumption 6. *The random variables*

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

are independent.

The following result is proven in Section E.1.

Theorem 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 (67). Under Assumption 6, there exists a positive constant $C > 0$ such that for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] &\leq n_1^{-1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right) + n_2^{-1}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) \\ &+ \inf_{t\in[n_2]}\left\{\frac{t}{n_2}\left(1+\log(|S|)\right) + \lceil n_2/t\rceil \beta_t\left(\mathbf{X}^{(2)}\right)\right\} \\ &+ \inf_{s\in S}\left\{h^2\left(\overline{P},\mathscr{M}_s\right) + \frac{(s+1)D_{n_1(s,1)}(\mathscr{M}_s)}{n_1} + n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right\}, \end{aligned} \tag{68}$$

where the mixing coefficient $\beta_t\left(\mathbf{X}^{(2)}\right)$ is given by (1.2.5) in Dedecker et al. [9].

One can check that we do not need any assumption other than Assumption 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 $n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)$ 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 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 $\overline{P} \in \mathscr{M}$.

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 $i \in [N_j]$,

$$X_i^{(j)} = E_i\overline{X}_i^{(j)} + (1-E_i^{(j)})Z_i^{(j)}. \tag{69}$$

For $s \in \{0,1,\ldots,s_{\max}\}$ and $b \in [s+1]$, we define the distributions

$$\overline{\mathbf{P}}^*_{s,b} = \mathcal{L}\left(\overline{X}_b^{(1)},\ldots,\overline{X}_{b+n_1(s,b)(s+1)}^{(1)}\right) \text{ and } \overline{\mathbf{P}}^{ind}_{s,b} = \bigotimes_{i=1}^{n_1(s,b)} \mathcal{L}\left(\overline{X}_{b+(i-1)(s+1)}^{(1)}\right).$$

The next result is a complement of Lemma 2 and is proven in Section E.2.

Lemma 7. *If* $E_1^{(1)},Z_1^{(1)},\dots,E_{n_1}^{(1)},Z_{n_1}^{(1)},E_1^{(2)},Z_1^{(2)},\dots,E_{n_2}^{(2)},Z_{n_2}^{(2)},\overline{\mathbf{X}}^{(1)}$ *and* $\overline{\mathbf{X}}^{(2)}$ *are mutually independent, we have*

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \leq \mathbf{K}\left(\overline{\mathbf{P}}^*_{s,b}||\overline{\mathbf{P}}^{ind}_{s,b}\right), \forall s \in \{0,1,\dots,s_{\max}\}, \forall b \in [s+1], \tag{70}$$

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 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 (67). There exists a positive constant* C *such that in the situation of Lemma 7 and for all* $\overline{P} \in \mathscr{P}_X$*,*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] &\leq n_1^{-1}\sum_{i=1}^{n_1}(1-p_i^{(1)}) + n_2^{-1}\sum_{i=1}^{n_2}(1-p_i^{(2)}) \\ &\quad + \inf_{t\in[n_2]}\left\{\frac{t}{n_2}\left(1+\log(|S|)\right) + \lceil n_2/t\rceil \beta_t\left(\overline{\mathbf{X}}^{(2)}\right)\right\} \\ &\quad + \inf_{s\in S}\left\{h^2\left(\overline{P},\mathscr{M}_s\right) + \frac{(s+1)D_{n_1(s,1)}(\mathscr{M}_s)}{n_1} + n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\overline{\mathbf{P}}^*_{s,b}||\overline{\mathbf{P}}^{ind}_{s,b}\right)\right\}. \end{aligned}$$

This result is a direct consequence of Theorem 11 and Lemma 7. We illustrate the performance of our estimator with hidden Markov models.

5.3 Application to hidden Markov models

Let $Y_1^{(1)},\dots,Y_{N_1}^{(1)},Y_1^{(2)},\dots,Y_{N_2}^{(2)}$ be random variables taking values in the measurable space $(\mathscr{Y},\mathcal{Y})$. Let L be in $\{2,3,\dots,\lfloor (N_1 \wedge N_2)/2\rfloor\}$ and $n_j = N_j + 1 - L$ for $j \in [2]$. We define the new random variables

$$X_i^{(j)} = \left(Y_i^{(j)},Y_{i+1}^{(j)},\dots,Y_{i+L-1}^{(j)}\right), i \in [n_j], j \in [2],$$

taking values in the measurable space $(\mathscr{X},\mathcal{X}) = \left(\mathscr{Y}^L,\mathcal{Y}^{\otimes L}\right)$. We adapt Assumption 2 to this context.

Assumption 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 π^* and we define the distribution P^* by (27). Let $\tau \geq e$ and $J = \lfloor \log_\tau\left(\lfloor (n_1-2)/2\rfloor\right)\rfloor$. Let S be the set given by

$$S = \{0\} \cup \left\{\left\lceil \tau^j \right\rceil; j \in \{0,1,\dots,J\}\right\}. \tag{71}$$

Let $\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K$ be subsets of $\mathscr{P}_Y$ such that Assumption 3 is satisfied. Let $\overline{\mathscr{M}}$ be a non-empty subset of the model $\mathscr{H}\left(K,\overline{\mathscr{F}}_1,\dots,\overline{\mathscr{F}}_K\right)$ defined by (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 (36) and $n_1(s,1)$ given by (66). The following result is proven in Section E.3.

Theorem 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 (67). Under Assumption 6, there is a numeric constant $C>0$ such that for all $\overline{P} \in \mathscr{P}_X$*

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] \leq\ & h^2\left(\overline{P},\overline{\mathscr{M}}\right) + n_1^{-1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right) + n_2^{-1}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) \\ & + L\epsilon^2 + \inf_{t\in[n_2]}\left\{\frac{t\log\log n_1}{n_2} + \lceil n_2/t\rceil \beta_t\left(\mathbf{X}^{(2)}\right)\right\} \\ & + \inf_{s\in S}\left\{\frac{(s+1)L\overline{V}\log n_1}{n_1} + n_1^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right\}. \end{aligned} \tag{72}$$

In particular under Assumption 7, there exists a positive constant $C(Q^*)$ such that

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_{\hat{s}}\right)\right] \leq h^2\left(P^*,\overline{\mathscr{M}}\right) + L\epsilon^2 + \tau L\overline{V}\frac{\log^2 n_1}{n_1} + \frac{\log n_2 \log\log n_1}{n_2}, \tag{73}$$

where P^* is given by (27).

Inequality (72) is a consequence of Theorem 11 and only requires Assumption 6. Under Assumption 7 we can control the different terms and obtain (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 $\mathscr{H}^*\left(K,\mathscr{F}_1,\ldots,\mathscr{F}_K\right)$. When $\epsilon>0$ the quantity $\overline{V}$ depends on ϵ and we need to balance the second and third term in (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 ϵ^{-2} up to logarithmic term and the bound on the convergence rate is of order ϵ^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 6. In both cases, it shows that we recover a value of s that allows 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 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 E.4.

Corollary 6. *Let $\hat{P}_{\hat{s}} = \hat{P}_{\hat{s}}\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)$ be the estimator given by (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 7, there exists a positive constant $C(Q^*)$ such that*

$$\begin{aligned} C(Q^*)\mathbb{E}\left[h^2\left(P^*,\hat{P}_{\hat{s}}\right)\right] \leq\ & \frac{L}{N_1}\sum_{i=1}^{N_1}\left(1-p_i^{(1)}\right) + \frac{L}{N_2}\sum_{i=1}^{N_2}\left(1-p_i^{(2)}\right) \\ & + L\epsilon^2 + \tau L\overline{V}\frac{\log^2 n_1}{n_1} + \frac{\log n_2\log\log n_1}{n_2}, \end{aligned}$$

where P^* is given by (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 ϵ^2 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

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \le L\left[\alpha_{cont} + \frac{\tau\overline{V}\log^2 n}{n}\right].$$

Our bound on the convergence rate is not deteriorated as long as the contamination rate α_{cont} is small compared to $\epsilon^2 + \frac{\tau\overline{V}\log^2 n}{n}$. 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

$$C(Q^*)\mathbb{E}\left[h^2\left(P^*, \hat{P}_s\right)\right] \le L\left[\frac{|I_1|+|I_2|}{N} + \frac{\tau\overline{V}\log^2 n}{n}\right].$$

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 terms.

References

[1] Pierre Alquier and Mathieu Gerber. *Universal Robust Regression via Maximum Mean Discrepancy.* 2020. DOI: 10.48550/ARXIV.2006.00840.

[2] 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. ISSN: 1432-1297. DOI: 10.1007/s00222-016-0673-5.

[3] Yannick Baraud. "Tests and estimation strategies associated to some loss functions". In: *Probability Theory and Related Fields* (Aug. 2021), pp. 799–846. ISSN: 1432-2064. DOI: 10.1007/s00440-021-01065-1.

[4] 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.

[5] Yannick Baraud and Juntong Chen. *Robust estimation of a regression function in exponential families.* 2020. DOI: 10.48550/ARXIV.2011.01657.

[6] 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.

[7] 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. ISSN: 0246-0203. DOI: https://doi.org/10.1016/j.anihpb.2005.04.004.

[8] 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.

[9] 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.

[10] I. A. Ibragimov and Has'minskii R. Z. *Statistical Estimation. Asymptotic Theory.* Springer New York, 1981. DOI: 10.1007/978-1-4899-0027-2.

[11] *Inference in Hidden Markov Models*. New York, NY: Springer New York, 2005. ISBN: 978-0-387-28982-3. DOI: 10.1007/0-387-28982-8.

[12] 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.

[13] 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].

[14] Lecestre, Alexandre. "Robust Estimation in Finite Mixture Models". In: *ESAIM: PS* 27 (2023), pp. 402–460. DOI: 10.1051/ps/2023004.

[15] Luc Lehéricy. "Consistent order estimation for nonparametric hidden Markov models". In: *Bernoulli* 25.1 (2019), pp. 464–498. DOI: 10.3150/17-BEJ993.

[16] 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.

[17] 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.

[18] 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. ISSN: 1886–1784. DOI: 10.1007/s11831-020-09422-4.

[19] Gilles Royer. *Une initiation aux inégalités de Sobolev logarithmiques*. Collection SMF. Cours spécialisés ; Paris: Société mathématique de France, 1999, p. 114.

[20] 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.

[21] 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.

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 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 [1], T-estimators [7] or l-estimators [3]. We can get rid of the independence assumption with the following result.

Proposition 4. *Under Assumption 8, for all random variables $X_1, \ldots, X_n$ with distributions $P_1, \ldots, P_n \in \mathscr{A}$ we have*

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} d\left(P_i, \hat{P}(n, \mathbf{X})\right)\right] \leq C_0 &\inf_{Q\in\mathscr{Q}} \sum_{i=1}^{n} d(P_i, Q) + f(n) \\ &+ g(n)\left[2 + \frac{3}{2}\mathbf{K}\left(\mathbf{P}^* || \mathbf{P}^{ind}\right)\right]^{\beta},\end{aligned}$$

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 1 that we prove hereafter, with $\mathbf{P} = \mathbf{P}^{ind}$ and $\mathbf{Q} = \mathbf{P}^*$.

A.1 Proof of Lemma 1

We use Lemma 48 in [2]. 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 λ demands

$$\log(1-\lambda) - \mathbf{K}(\mathbf{Q}||\mathbf{P}) + \frac{\lambda}{1-\lambda} = 0.$$

Let λ^* be such a number. In that case

$$(\lambda^*)^{-1}\left[\log\left(\frac{1}{1-\lambda^*}\right) + \mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right)\right] = \frac{1}{1-\lambda^*}.$$

We set $a(x) = x - \log(1+x)$ for x in $(0, +\infty)$. Following the proof of Proposition 5 [2], a is increasing and

$$\forall x > 0, a^{-1}(x) \leq x + \sqrt{2x}.$$

Since $\frac{\lambda^*}{1-\lambda^*} = a^{-1}\left(\mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right)\right)$, we get

$$\begin{aligned}\frac{1}{1-\lambda^*} = 1 + \frac{\lambda^*}{1-\lambda^*} &\leq 1 + \mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right) + \sqrt{2\mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right)} \\ &\leq 2 + \frac{3}{2}\mathbf{K}\left(\mathbf{Q}||\mathbf{P}\right).\end{aligned}$$

Finally, with Jensen's inequality we get

$$\mathbb{E}_{\mathbf{Q}}\left[l\left(\hat{\theta}(\mathbf{X}),\theta\right)\right] \leq A + \frac{B + \left(2 + \frac{3}{2}\mathbf{K}\left(P||Q\right)\right)^{\beta}}{n}.$$

B Main results

This section gathers the proofs of Theorem 1, Corollary 1 and Lemmas 2, 3, 4. We first give a formal definition of the ρ-dimension function that is originally introduced in Baraud & Birgé [4]. We slightly modify some notation to adapt it to our context. The function ψ defined by (4) satisfies Assumption 2 [4] with $a_0 = 4, a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see Proposition 3 [4]). 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 σ-finite measure ν such that $\mathscr{M} = \{Q = q\cdot\nu; q \in \mathcal{M}\}$, we write

$$w\left(\nu,\mathcal{M},\mathscr{M},\mathbf{P}^{ind},P,y\right) = \mathbb{E}_{\mathbf{X}\sim\mathbf{P}^{ind}}\left[\sup_{Q\in\mathscr{B}^{\mathscr{M}}\left(\mathbf{P}^{ind},P,y\right)} \left|\mathbf{Z}_n\left(\mathbf{X},p,q\right)\right|\right],$$

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 (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 σ-finite measure ν. We define the ρ-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 2 we consider cases for which we have a uniform bound over the ρ-dimension function. More precisely we assume there is a non-increasing function $m \mapsto D_m(\mathscr{M})$ such that

$$D^{\mathscr{M}}\left(\mathbf{P}^{ind},P^{\otimes m}\right) \leq D_m(\mathscr{M}), \forall \mathbf{P}^{ind} \in \mathscr{P}_X^{\otimes m}, \forall P \in \mathscr{M}.$$

B.1 Proof of Theorem 1

From Theorem 1 of Baraud & Birgé [4], 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 γ and κ are given in [4] and satisfy $\gamma \leq 150$ and $\frac{4\kappa}{a_1} \leq 5014$ (see proof of Theorem 1 [5], page 32). We can take the infimum for Q over $\mathscr{M}$ and it shows we satisfy Assumption 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 4, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)}, \hat{P}_s\right)\right] &\leq 150 \inf_{Q\in\mathscr{Q}} \sum_{i=1}^{n(s,b)} h^2\left(P_{b+(i-1)(s+1)},Q\right) \\ &\quad + 5014\left(\frac{D_{n(s,b)}(\mathscr{M})}{4.7} + 3.49 + \frac{3}{2}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\right),\end{aligned}$$

for all $b \in [s+1]$. From (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

$$\begin{aligned}\mathbb{E}_{\mathbf{P}^*}\left[\sum_{i=1}^{n} h^2\left(P_i, \hat{P}_s\right)\right] &\leq 602\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2\left(P_i, Q\right) + \frac{20056}{4.7}(s+1)\left[D_{n(s,1)}(\mathscr{M}) + 17\right] \\ &\quad + 30084\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right).\end{aligned}$$

B.2 Proof of Lemma 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 8. *For random variables A,B,C such that $\mathcal{L}(A) \ll \mathcal{L}(B)$, we have*

$$\mathbf{K}\left(\mathcal{L}(A)||\mathcal{L}(B)\right) \leq \mathbf{K}\left(\mathcal{L}\left(A,C\right)||\mathcal{L}(B) \otimes \mathcal{L}(C)\right). \tag{74}$$

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 to conclude.

B.2.1 Proof of Lemma 8

Let μ_1 and μ_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.

B.3 Proof of Corollary 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 1 and Lemma 2.

B.4 Proof of Lemma 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.

B.5 Proof of Lemma 4

If $(\mathbf{X},\mathbf{H})$ a hidden Markov chain, with Lemma 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

$$\mathbf{K}\left(\mathcal{L}\left(\mathbf{X}\right)||\mathcal{L}\left(X_1\right)\otimes\cdots\otimes\mathcal{L}\left(X_n\right)\right)\leq\sum_{i=2}^{n}\mathbf{K}\left(\mathcal{L}\left(H_{i-1},H_i\right)||\mathcal{L}\left(H_{i-1}\right)\otimes\mathcal{L}\left(H_i\right)\right).$$

C Kolmogorov processes

This section gathers the proofs of Theorems 2, 3 and Lemmas 5, 6.

C.1 Proof of Theorems 2 and 3

From Proposition 6 [4], we can take $D_n(\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]) = 9\log(2|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]|)$. From Theorem 1 there exists a positive constant C such that

$$\begin{aligned} C\mathbb{E}_{\mathbf{P}^*}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \le{}& h^2\left(\overline{P},\mathscr{F}_{\lambda_-,\lambda_+,M}\right) + \epsilon^2 + n^{-1}\sum_{i=1} h^2\left(P_i,\overline{P}\right) \\ &+ n^{-1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \\ &+ \frac{s+1}{n}\left[1+\log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|)\right]. \end{aligned}$$

Given the bounds on $\log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|)$ given by Lemma 6, we obtain the following inequalities.

- For $d=1$ we have $\epsilon^2 = n^{-4/5}\log^{4/5} n$ and

$$\begin{aligned} \log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|) &\le \log(9/\eta_1) + \frac{7}{2}\log M + \overline{K}_1\epsilon^{-1/2} \\ &= \log(9/\eta_1) + \frac{9}{2}\overline{K}_1 n^{1/5}\log^{-1/5} n. \end{aligned}$$

- For $d=2$ we have $\epsilon^2 = n^{-2/3}\log^{5/3} n$ and

$$\begin{aligned} \log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|) &\le \log\left(\frac{3^8\pi}{\eta_2^3}\right) + 9\log M + \overline{K}_2\epsilon^{-1}\log_{++}^{3/2}(1/\epsilon) \\ &\le \log\left(\frac{3^8\pi}{\eta_2^3}\right) + \frac{28}{3}\overline{K}_2 n^{1/3}\log^{2/3} n. \end{aligned}$$

- For $d=3$ we have $\epsilon^2 = n^{-1/4}\log^{1/4} n$ and

$$\begin{aligned} \log(2|\mathcal{F}_{\lambda_-,\lambda_+,M}[\delta]|) &\le \log\left(\frac{2^7 3^{27/2}\pi^3}{\eta_3^6}\right) + \frac{33}{2}\log M + \overline{K}_3\epsilon^{-2} \\ &= \log\left(\frac{2^7 3^{27/2}\pi^3}{\eta_3^6}\right) + \frac{33}{2}\overline{K}_3 n^{1/2}\log^{-1/2} n. \end{aligned}$$

This proves the bound (22). Lemma 5 allows to conclude the proof of Theorem 2.

For $d \ge 4$ we have $\epsilon^2 = n^{-\frac{2}{d+1}}\log^{d+2+\frac{2}{d+1}} n$ and

$$\begin{aligned} \log(|\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]|) &\le \log C_d + \left(\overline{K}_d + 2 + \frac{1}{d} + \frac{1}{d^2}\right)\epsilon^{-(d-1)}\log^{(d+1)(d+2)/2}(\epsilon^{-1}) \\ &\le \log C_d + \frac{1}{d+1}\left(\overline{K}_d + 2 + \frac{1}{d} + \frac{1}{d^2}\right) n^{\frac{d-1}{d+1}}\log^{\frac{2}{d+1}+d+1} n. \end{aligned}$$

Lemma 5 allows to conclude the proof of Theorem 3.

C.2 Proof of Lemma 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 $(Y_t)_{t\ge 0}$ 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 (17) satisfying $Y_0^x = x$. We follow the proof of Theorem 3.2.7 [19] with their notation. From (44) therein we have

$$A_x(s) \leq \mathbb{E}\left[\left(\log(Z) + U(x) + U(W_s) - 2v(W_s) - \frac{1}{2}\int_0^s [|\nabla U|^2 - \Delta U](W_t)dt\right) F\right], \tag{75}$$

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.
$$\exp(-2v(y)) = (2\pi t)^{-d/2}\exp\left(-\frac{(x-y)^2}{2s}\right), \forall y \in \mathbb{R}^d. \tag{76}$$

Let us check that the right-hand side of (75) is finite. From (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 (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

$$\begin{aligned}
A_x(s) &\leq \log(Z) + U(x) - \frac{d}{2}\log(2\pi s) - \frac{Cs}{2} + e^{U(x)-\frac{Cs}{2}}||g||_\infty \\
&\leq \log(Z) - \frac{d}{2}\log(2\pi s) - \frac{Cs}{2} + e^{U(x)}||g||_\infty\left(1 + e^{-\frac{Cs}{2}}\right).
\end{aligned} \tag{77}$$

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 [19], for all $s_0 > 0$, we have

$$A_x(s) \leq A_x(s_0)\exp\left(-2m(s - s_0)\right), \forall s > s_0. \tag{78}$$

Therefore with (77) and (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 α.

C.3 Proof of Lemma 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

$$
N(B_2(M), ||\cdot||_2, \epsilon) \leq \left(\frac{3M}{\epsilon}\right)^d, \tag{79}
$$

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 ϵ-net of $\tilde{F}_d^{1,\eta_d}$ with respect to the Hellinger distance. We define

$$
\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon] := \left\{ (\det\Sigma)^{-1/2} g\left(\Sigma^{-1/2}\left(\cdot - \overline{x}\right)\right); \begin{array}{l} \overline{x} \in B_2(M)\left[\sqrt{\lambda_-}\right], \\ \Sigma \in \mathrm{Sym}(\lambda_-,\lambda_+)[\eta_d\lambda_-], \\ g \in \tilde{F}_d^{1,\eta_d}[\epsilon] \end{array} \right\}
$$

and we show it is an ϵ-net of $\mathcal{F}_{\lambda_-,\lambda_+,M}$ with respect to the Hellinger distance. For $f \in \mathcal{F}_{\lambda_-,\lambda_+,M}$, there is Σ 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 ϵ-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 ϵ, 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 9. *We have*

$$N\left(\epsilon, \mathit{Sym}(\lambda_-,\lambda_+), ||\cdot||_{op}\right) \leq \begin{cases} \frac{3(\lambda_+-\lambda_-)}{\epsilon} \textit{ for } d=1, \\ \left(\frac{9}{\epsilon}\right)^3 (\lambda_+-\lambda_-)^2\lambda_+\pi \textit{ for } d=2, \\ 2\left(\frac{2\cdot 3^{5/4}\sqrt{\lambda_+(\lambda_+-\lambda_-)\pi}}{\epsilon}\right)^6 \textit{ for } d=3. \end{cases} \tag{80}$$

In higher dimensions, we have

$$\begin{aligned} N\left(\epsilon, \mathit{Sym}(\lambda_-,\lambda_+), ||\cdot||_{op}\right) \leq C\left(\frac{3}{4}\right)^d & \frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}}(2\lambda_+)^{d(d-1)/2}(\lambda_+-\lambda_-)^d \\ & \times (d+1)^{d(d+1)/2} d^{(d-1)(d+2)/2}(d-1)^{(d-1)/2}\epsilon^{-d(d+1)/2}, \end{aligned}$$

with $C = \frac{e^{1/2}}{3^{1/2}2^3}$.

Theorem 4 [12] gives a bound on $|\tilde{F}_d^{1,\eta_d}[\epsilon]|$ which allows to conclude the proof of Theorem 2.

Case $d \geq 4$. We use Theorem 3 of Kur *et al.* [13]. We follow some of their notation. Let $d \geq 4$. There exist positive constants ξ_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 ϵ-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

$$\mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon] := \left\{ (\det\Sigma)^{-1/2} g\left(\Sigma^{-1/2}(\cdot - \overline{x})\right);\ \begin{array}{c} \overline{x} \in B_2(M)\left[\xi_d\sqrt{\lambda_-}\right], \\ \Sigma \in \mathrm{Sym}(\lambda_-,\lambda_+)[\lambda_-/3], \\ g \in \mathcal{F}_{d,\tilde{I}}[\epsilon] \end{array} \right\}$$

and we show that $\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon] = \{f(x)dx; f \in \mathcal{F}_{\lambda_-,\lambda_+,M}[\epsilon]\}$ is an ϵ-net of $\mathscr{F}_{\lambda_-,\lambda_+,M}$ with respect to the Hellinger distance. For $f \in \mathcal{F}_{\lambda_-,\lambda_+,M}$, there is Σ 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 ϵ-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 9 we get

$$\begin{aligned}
|\mathscr{F}_{\lambda_-,\lambda_+,M}[\epsilon]| &\leq C\left(\frac{3M}{\xi_d\sqrt{\lambda_-}}\right)^d \left(\frac{3}{4}\right)^d \frac{\pi^{d(d-1)/2}}{e^{(d-1)(d-2)/4}} (2\lambda_+)^{d(d-1)/2} (\lambda_+ - \lambda_-)^d \\
&\quad \times (d+1)^{d(d+1)/2} d^{(d-1)(d+2)/2} (d-1)^{(d-1)/2} \left(\frac{\lambda_-}{3}\right)^{-d(d+1)/2} \\
&\quad \times \exp\left(\overline{K}_d \epsilon^{-(d-1)} \log(\epsilon^{-1})^{(d+1)(d+2)/2}\right) \\
&\leq C_d \frac{\lambda_+^{d(d-1)/2} M^d (\lambda_+ - \lambda_-)^d}{\lambda_-^{d(d+1)/2}} \exp\left(\overline{K}_d \epsilon^{-(d-1)} \log(\epsilon^{-1})^{(d+1)(d+2)/2}\right).
\end{aligned}$$

C.3.1 Proof of Lemma 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 Σ can be written as $\Sigma = UDU^T$ where D is the diagonal matrix containing the real eigenvalues of Σ 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 ϵ 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

  $$N\left(\mathrm{ON}(3), ||\cdot||, \epsilon\right) \leq \left(N\left([0,2\pi], |\cdot|, \epsilon/\sqrt{3}\right)\right)^3 \leq 2\left(\frac{3\sqrt{3}\pi}{\epsilon}\right)^3, \tag{81}$$

  where the factor 2 comes from the presence of ϵ 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 10. *For $d \geq 3$, we can build an ϵ-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 $\epsilon_2 = \frac{\epsilon}{2\lambda_+}\frac{d-1}{d+1}$.

C.3.2 Proof of Lemma 10

We prove this by induction. From (81) we have the desired inequality for $d = 3$ with $C_3 = \frac{e^{1/2}}{3^{1/2}2^3}$. Let ϵ be in (0,1] and $d \geq 3$. Let us now assume that for $\lambda_1 > 0$ we have a λ_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 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 $ON[\epsilon] = O[\lambda_1,\lambda_2]$.

C.3.3 Proof of Lemma 11

Let $C = (C_1 \dots C_{d+1})$ be in $ON(d+1)$. There is θ in $[0,2\pi]^d$ such that $C_1 = A_1(\theta)$. Let B be the matrix in $ON(d)$ given by

$$A(\theta)^T C = \begin{pmatrix} 1 & 0 \\ 0 & B \end{pmatrix}.$$

For $\theta \in [0,2\pi]^d$ 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 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

$$||A(\theta) - A(\psi_{i_1,\dots,i_d})||_{op} \leq d^{1/2}\pi\epsilon.$$

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

$$\begin{aligned}
||C - C'||_{op} &\leq \left|\left|A(\theta)\begin{pmatrix} 0 & 0 \\ 0 & B-B' \end{pmatrix}\right|\right|_{op} + \left|\left|(A(\theta) - A(\psi_{i_1,\dots,i_d}))\begin{pmatrix} 1 & 0 \\ 0 & B' \end{pmatrix}\right|\right|_{op} \\
&\leq ||B-B'||_{op} + ||A(\theta) - A(\psi_{i_1,\dots,i_d})||_{op} \\
&\leq \lambda_1 + d^{1/2}\pi\lambda_2.
\end{aligned}$$

C.3.4 Proof of Lemma 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

$$\begin{aligned}
A_{ij}^{(k+1)} - A_{ij}^{(k)} &= \begin{cases}
0 \text{ if } i \leq k \\
\prod\limits_{l \leq k} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right) \\
\times \left[\cos\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + h_{k+1}\right) - \cos\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2}\right)\right] \text{ if } i = k+1 \\
\prod\limits_{\substack{l<i \\ l \neq k+1}} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + \mathbb{1}_{l \leq k}\right) \times \cos\left(\theta_i + \mathbb{1}_{i<j}\frac{\pi}{2}\right) \\
\times \left[\sin\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + h_{k+1}\right) - \sin\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2}\right)\right] \text{ if } i > k+1,
\end{cases} \\
&= 2\sin\left(\frac{h_{k+1}}{2}\right) \prod_{l \leq k} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2} + h_l\right) \\
&\times \begin{cases}
0 \text{ if } i \leq k \\
-\sin\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + \frac{h_{k+1}}{2}\right) \text{ if } i = k+1 \\
\cos\left(\theta_{k+1} + \mathbb{1}_{k+1<j}\frac{\pi}{2} + \frac{h_{k+1}}{2}\right) \prod\limits_{k+1<l<i} \sin\left(\theta_l + \mathbb{1}_{l<j}\frac{\pi}{2}\right) \\
\times \cos\left(\theta_i + \mathbb{1}_{i<j}\frac{\pi}{2}\right) \text{ if } i > k+1.
\end{cases}
\end{aligned}$$

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

$$\begin{aligned} ||A^{(d)} - A^{(0)}||_{op} &\leq \sum_{k=0}^{d-1} ||(A^{(k+1)} - A^{(k)})^T||_{op} \\ &\leq \sum_{k=0}^{d-1} (d-k) h_{k+1}^2. \end{aligned}$$

D Hidden Markov models

This section gathers the proof of Theorems 4, 5, 6, 9, 10, Corollary 3 and Proposition 1, 2, 3.

D.1 Proof of Theorem 4

The next result is proven in Section D.1.1 and gives a bound on the ρ-dimension function.

Proposition 5. *Under Assumption 3 and with $\delta(s)$ given by (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 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 D.1.2 and tells us how well $\mathscr{M}_\delta$ approximates $\mathscr{M}$.

Proposition 6. *For $K \geq 2$, w,v in $\mathcal{W}_K$, Q,R in $\mathcal{T}_K$ and probability distributions $F_1, \dots, F_K, G_1, \dots, G_K$ 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 6 and inequality (B.5) in Lecestre [14] we have

$$h^2\left(P, \mathscr{M}_\delta\right) \leq (K-1)L\delta + L\epsilon^2, \forall P \in \mathscr{M}. \tag{82}$$

With the choice of δ given in (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 4. The next result is proven later in Section D.1.3.

Lemma 13. *Under Assumption 2, there are positive constants $C(Q^*)$ and $r(Q^*)$ that only depend on Q^* such that*

$$n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \le C(Q^*)e^{-r(Q^*)s}, \forall s \ge L-1, \forall b \in [s+1],$$

and $h^2\left(P^*,P_i\right) \le C(Q^*)e^{-r(Q^*)i}$ for all $i \in [n]$.

In this situation, for $\overline{P} = P^*$ and $s \ge L-1$ we have

$$C\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_s\right)\right] \le h^2\left(\overline{P},\mathscr{M}\right) + \frac{C(Q^*)}{n(e^{r(Q^*)}-1)} + C(Q^*)e^{-r(Q^*)s} \\ + L\epsilon^2 + (s+1)L\overline{V}\frac{\log n}{n},$$

for some positive constant. The condition on s leads to the desired inequality.

D.1.1 Proof of Proposition 5

From Proposition A.1. [14], we have

$$D^{\mathscr{H}_{\delta(s)}}\left(\bigotimes_{i=1}^{n(s,b)}P_i, Q^{\otimes n(s,b)}\right) \le 545.3\overline{V}\left[5.82 + \log\left(\frac{(K^L+1)^2}{\delta(s)^L}\right) + \log_+\left(\frac{n(s,b)}{\overline{V}}\right)\right].$$

- If $\overline{V} \le n(s,1)(K-1)/K$, we have

$$\begin{aligned}\log\left(\frac{(K^L+1)^2}{\delta(s,b)^L}\right) + \log_+\left(\frac{n(s,b)}{\overline{V}}\right) &\le \log\left(\frac{(K^L+1)^2 n(s,1)^L(K-1)^L}{\overline{V}^L}\frac{n(s,1)}{\overline{V}}\right) \\ &= \log\left(\frac{(K^L+1)^2(K-1)^L}{K^{L+1}}\right) + \log\left(\frac{K^{L+1}n(s,1)^{L+1}}{\overline{V}^{L+1}}\right) \\ &= \log\left(\frac{(K^L+1)^2(K^2-1)^L}{K^{L+1}(K+1)^L}\right) + (L+1)\log\left(\frac{Kn(s,1)}{\overline{V}}\right).\end{aligned}$$

  One can check that for $L \ge 2$, we have $\frac{(K^L+1)^2(K^2-1)^L}{K^{L+1}(K+1)^L} \le K^{2L-1}$ for all $K \ge 1$. Therefore,

$$\begin{aligned}\log\left(\frac{(K^L+1)^2}{\delta(s)^L}\right) + \log_+\left(\frac{n(s,b)}{\overline{V}}\right) &\le (2L-1)\log K + (L+1)\log\left(\frac{Kn(s,1)}{\overline{V}}\right) \\ &\le 3L\log\left(\frac{KN}{\overline{V}}\right) = 3L\log\left(\frac{KN}{\overline{V}\wedge N}\right).\end{aligned}$$

- 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

$$\begin{aligned}\log\left(\frac{(K^L+1)^2}{\delta(s)^L}\right) + \log_+\left(\frac{n(s,b)}{\overline{V}}\right) &\le \log\left(\frac{(K^L+1)^2K^Ln(s,1)}{\overline{V}}\right) \\ &= \log\left(\frac{Kn(s,1)}{\overline{V}}\right) + (L-1)\log K + 2\log\left(1+K^L\right) \\ &\le 3L\log\left(\frac{Kn(s,1)}{\overline{V}\wedge n(s,1)}\right) + 2\log(1+K^{-L}) \\ &\le 2\log 2 + 3L\log\left(\frac{Kn(s,1)}{\overline{V}\wedge n(s,1)}\right).\end{aligned}$$

D.1.2 Proof of Proposition 6

With Lemma B.3 [14], 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

$$wQ^{\bigcirc L}(k_1,\ldots,k_L) = w_{k_1}Q_{k_1,k_2}\ldots Q_{k_{L-1},k_L}, \forall k_1,\ldots,k_L \in [K]. \tag{83}$$

Let ρ 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 to conclude the proof.

D.1.3 Proof of Lemma 13

Let s not be 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 4, we have

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \leq \sum_{i=1}^{n(s,b)-1} \mathbf{K}\left(\mathcal{L}\left(H_i^{(L,s,b)}, H_{i+1}^{(L,s,b)}\right)||\mathcal{L}\left(H_i^{(L,s,b)}\right)\otimes\mathcal{L}\left(H_{i+1}^{(L,s,b)}\right)\right).$$

We can use the following result to bound the terms in the sum on the right-hand side of the inequality.

Lemma 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 $k_1,\dots,k_{2L}\in[K^*]$, we have

$$\begin{aligned}&\mathbb{P}\left(H_{i+1}^{(L,s,b)}=(k_{L+1},\dots,k_{2L})|H_i^{(L,s,b)}=(k_1,\dots,k_L)\right)\\&=Q^*_{k_{2L-1},k_{2L}}\dots Q^*_{k_{L+1},k_{L+2}}(Q^*)^{s+2-L}_{k_L,k_{L+1}}\end{aligned}$$

Therefore, we have

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right)\leq 2\sum_{i=1}^{n(s,b)-1}\sum_{k\in[K^*]}d_{TV}\left((Q^*)^{s+2-L}_{k,\cdot},\nu_iQ^{s+2-L}\right),$$

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 π^* 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

$$h^2\left(P^*,P_i\right)\leq d_{TV}\left(P^*,P_i\right)=d_{TV}\left(\pi^*,w^*(Q^*)^{i-1}\right)\leq C(Q^*)e^{-r(Q^*)(i-1)}.$$

D.1.4 Proof of Lemma 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

$$\mathbf{K}\left(\mathcal{L}(A,B)||\mathcal{L}(A)\otimes\mathcal{L}(B)\right)\leq\sum_{a\in\mathscr{A}}2d_{TV}\left(\mathcal{L}\left(B|A=a\right),\mathcal{L}\left(B\right)\right).$$

D.2 Proof of Corollary 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

$$\begin{aligned} h^2\left(P_i, P^*\right) &\leq p_i p_{i+1} \dots p_{i+L-1} h^2\left(P_i', P^*\right) + \left(1 - p_i p_{i+1} \dots p_{i+L-1}\right) \\ &\leq h^2\left(P_i', P^*\right) + (1 - p_i) + \cdots + (1 - p_{i+L-1}), \end{aligned}$$

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 4, Lemma 2 and Lemma 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 $s \geq L - 1$.

D.3 Proof of Theorems 5 and 6

With (45) and Theorem 4, we have

$$\begin{aligned} C\mathbb{E}\left[h^2\left(\overline{P}, \hat{P}_s\right)\right] \leq{}& h^2\left(\overline{P}, \mathscr{M}\right) + n^{-1}\sum_{i=1}^{n} h^2\left(P_i, \overline{P}\right) \\ &+ n^{-1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}_{s,b}^* || \mathbf{P}_{s,b}^{ind}\right) \\ &+ L\epsilon^2 + (s+1)L^2K^L\log(2|\mathscr{F}_{\lambda_-,\lambda_+,M}|[\epsilon])\frac{\log n}{n}. \end{aligned}$$

We can simply follow the proof of Theorems 2 and 3 to conclude.

D.4 Proof of Proposition 1

The proof relies on the following lemma.

Lemma 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 is met with

$$\overline{V} = 3K^L + K^{L-1}L\sum_{k=1}^{K} d_k \leq K^L\left(3 + L\max_{1\leq k\leq K} d_k\right).$$

D.4.1 Proof of Lemma 15

We have

$$\begin{aligned}\mathcal{A} &= \left\{(x_1,\ldots,x_L) \mapsto f_{\theta_1}(x_1)\ldots f_{\theta_L}(x_L); \theta_i \in \overline{\Theta}_i, \forall i \in \{1,\ldots,L\}\right\}\\ &= \exp\circ\left\{(x_1,\ldots,x_L) \mapsto \sum_{i=1}^{L}\langle\eta_i(\theta_i), T_i(x_i)\rangle + A_i(\theta_i) + B_i(x_i), \forall i \in \{1,\ldots,L\}\right\}\\ &\subset \exp\circ(V+B)\end{aligned}$$

with $B : (x_1,\ldots,x_L) \mapsto B_i(x_1) + \cdots + B_i(x_L)$ and

$$V = \left\{(x_1,\ldots,x_L) \mapsto A + \sum_{i=1}^{K}\langle\eta_i, T_i(x_i)\rangle; \eta_i \in \mathbb{R}^d, \forall i \in \{1,\ldots,L\}, A \in \mathbb{R}\right\}.$$

The set V is a vector space of dimension $1 + d_1 + \cdots + d_L$ and exp is monotone, therefore, from Proposition 42-(i,ii) [2] and Lemma 2.6.15 [21] and Lemma 2.6.18-(v) [21], the class of functions $\mathcal{A}$ is VC-subgraph with VC-index $V(\mathcal{A}) \leq 3 + d_1 + \cdots + d_L$.

D.5 Proof of Proposition 2

We first need the following lemma to apply results of regular parametric models.

Lemma 16. *Under Assumption 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*

$$I : \phi \mapsto I(\phi) = \int_{\mathscr{X}^L} \partial_\phi p(\mathbf{x};\phi)\,(\partial_\phi p(\mathbf{x};\phi))^T\,\frac{\mu(\mathbf{x})}{p(\mathbf{x};\phi)}$$

is well-defined and continuous.

We can now apply results of Ibragimov and Has'minskiĭ [10], in particular (7.20) which is a consequence of Theorem 7.6. Let κ be a compact subset of $\overline{\Phi}$ such that $\overline{\Phi}$ belongs to the interior of κ. 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

$$\begin{aligned}\forall\phi \in \overline{\Phi}, h^2\left(P_\phi, P_{\overline{\phi}}\right) &\geq \mathbb{1}_{\phi\in\kappa}\frac{a(\kappa)}{1+b(\kappa)}||\phi-\overline{\phi}||^2 + \mathbb{1}_{\phi\in\overline{\Phi}\setminus\kappa}c(\kappa)\\ &\geq C(\overline{\phi})\left[||\overline{w}-w||^2 + \left\|\overline{Q}-Q\right\|^2 + \sum_{k=1}^{K}\left\|\overline{\theta}-\theta\right\|^2 \wedge 1\right].\end{aligned}$$

D.5.1 Proof of Lemma 16

For $k_1,\dots,k_L \in [K]$ we have

$$p(\mathbf{x};\phi) \geq w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l). \tag{84}$$

- Since η_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 η_k are differentiable for all $k \in [K]$. For all $\overline{k} \in [K]$ and $j \in [e_k]$,

$$\begin{aligned}\partial_{\theta_{\overline{k},j}} p(\mathbf{x};\phi) &= \sum_{k_1,\dots,k_L} w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \sum_{l=1}^{L} \mathbb{1}_{k_l=\overline{k}} \left(\prod_{i\neq l} f_{\theta_{k_j}}(x_j) \right) \partial_{\theta_{\overline{k},j}} f_{\theta_{\overline{k}}}(x_l) \\ &= \sum_{k_1,\dots,k_L} w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \prod_{i=1}^{L} f_{\theta_{k_i}}(x_i) \\ &\times \sum_{l=1}^{L} \mathbb{1}_{k_l=\overline{k}} \left[\langle \partial_{\theta_{\overline{k},j}} \eta_{\overline{k}}(\theta_{\overline{k}}), T_{\overline{k}}(x_l) \rangle + \partial_{\theta_{\overline{k},j}} A_{\overline{k}}(\theta_{\overline{k}}) \right].\end{aligned} \tag{85}$$

For $\overline{k} \in [K-1]$ and $k' \in [K]$ we have

$$\begin{aligned}\partial_{w_{\overline{k}}} p(\mathbf{x};\phi) = &\sum_{k_2,\dots,k_L} Q_{\overline{k},k_2} \dots Q_{k_{L-1},k_L} f_{\theta_{\overline{k}}}(x_1) \prod_{l=2}^{L} f_{\theta_{k_l}}(x_l) \\ &- \sum_{k_2,\dots,k_L} Q_{K,k_2} \dots Q_{k_{L-1},k_L} f_{\theta_K}(x_1) \prod_{l=2}^{L} f_{\theta_{k_l}}(x_l)\end{aligned} \tag{86}$$

and

$$\begin{aligned}\partial_{Q_{k',\overline{k}}} p(\mathbf{x};\phi) &= \sum_{k_1,k_2,\dots,k_L} w_{k_1} \partial_{Q_{k',\overline{k}}} \left[Q_{k_1,k_2} \dots Q_{k_{L-1},k_L} \right] \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) \\ &= \sum_{k_1,k_2,\dots,k_L} w_{k_1} \prod_{i=1}^{L} f_{k_i,\theta_{k_i}}(x_i) \sum_{l=2}^{L} \left[\mathbb{1}_{(k',\overline{k})=(k_{l-1},k_l)} - \mathbb{1}_{(k',K)=(k_{l-1},k_l)} \right] \prod_{\substack{2\leq j\leq L,\\ j\neq l}} Q_{k_{j-1},k_j}.\end{aligned} \tag{87}$$

Since A_k and η_k are $\mathcal{C}^1$, we just need to check that the functions

$$\phi \mapsto \int_{\mathscr{Y}^L} T_{\overline{k},j}(x_i) T_{\overline{k}',j'}(x_{i'}) \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}, \tag{88}$$

$$\phi \mapsto \int_{\mathscr{Y}^L} T_{\overline{k},j}(x_i) \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}, \tag{89}$$

$$\phi \mapsto \int_{\mathscr{Y}^L} \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)}, \tag{90}$$

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 (84) repeatedly.

- We have

$$
\begin{aligned}
0 &\leq \int_{\mathscr{Y}^L} \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)} \\
&\leq \Big(w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L}\Big)^{-1} \int_{\mathscr{Y}^L} \prod_{l=1}^{L} f_{\theta_{k'_j}}(x_j) \mu(d\mathbf{x}) \\
&= \Big(w_{k_1} Q_{k_1,k_2} \dots Q_{k_{L-1},k_L}\Big)^{-1} < \infty,
\end{aligned}
$$

and (90) is well defined. Similarly

$$
\begin{aligned}
0 &\leq \int_{\mathscr{Y}^L} \left| T_{\overline{k},j}(x_i) \right| \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)} \\
&\leq \Big(w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}\Big)^{-1} \int_{\mathscr{Y}} \left| T_{\overline{k},j}(x_i) \right| f_{\theta_{k_i}}(x_i) \nu(dx_i) \\
&\leq \Big(w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}\Big)^{-1} \sqrt{\int_{\mathscr{Y}} \left| T_{\overline{k},j}(x_i) \right|^2 f_{\theta_{k_i}}(x_i) \nu(dx_i)} < \infty,
\end{aligned}
$$

and (89) is well defined. Finally

$$
\begin{aligned}
0 &\leq \int_{\mathscr{Y}^L} \left| T_{\overline{k},j}(x_i) T_{\overline{k}',j'}(x_{i'}) \right| \prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \frac{\mu(d\mathbf{x})}{p(\mathbf{x};\phi)} \\
&\leq \Big(w_{k_1} w_{k'_1} Q_{k_1,k_2} Q_{k'_1,k'_2} \dots Q_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}\Big)^{-1/2} \\
&\times \int_{\mathscr{Y}^L} \left| T_{\overline{k},j}(x_i) T_{\overline{k}',j'}(x_{i'}) \right| \sqrt{\prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l) \mu(d\mathbf{x})} \\
&\leq \Big(w_{k_1} w_{k'_1} Q_{k_1,k_2} Q_{k'_1,k'_2} \dots Q_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}\Big)^{-1/2} \\
&\times \sqrt{\int_{\mathscr{Y}} \left| T_{\overline{k},j}(x_i) \right|^2 f_{\theta_{k_i}}(x_i) \nu(dx_i)} \sqrt{\int_{\mathscr{Y}} \left| T_{\overline{k}',j'}(x_{i'}) \right|^2 f_{\theta_{k'_{i'}}}(x_{i'}) \nu(dx_{i'})} < \infty,
\end{aligned}
$$

and (88) is well defined. The Fisher information matrix $I(\phi)$ is well-defined for all ϕ. We now turn to continuity.

- We have

$$\begin{aligned}
&\left|\frac{\prod_{l=1}^L f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)} - \frac{\prod_{l=1}^L f_{k_l,\theta'_{k_l}}(x_l) f_{k'_l,\theta'_{k'_l}}(x_l)}{p(\mathbf{x};\phi')}\right| \\
&\le \frac{\prod_{l=1}^L f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)} \left|\prod_{l=1}^L f_{\theta_{k_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\right| \\
&+ \prod_{l=1}^L f_{\theta'_{k_l}}(x_l) \prod_{l=1}^L f_{\theta_{k'_l}}(x_l) \left|\frac{1}{p(\mathbf{x};\phi)} - \frac{1}{p(\mathbf{x};\phi')}\right| \\
&+ \frac{\prod_{l=1}^L f_{\theta'_{k_l}}(x_l)}{p(\mathbf{x};\phi')} \left|\prod_{l=1}^L f_{\theta_{k'_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k'_l}}(x_l)\right| \\
&\le \frac{\left|\prod_{l=1}^L f_{\theta_{k_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\right|}{w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}} \\
&+ \frac{|p(\mathbf{x};\phi) - p(\mathbf{x};\phi')|}{w'_{k_1} w_{k'_1} Q'_{k_1,k_2} Q_{k'_1,k'_2} \dots Q'_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}} \\
&+ \frac{\left|\prod_{l=1}^L f_{\theta_{k'_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k'_l}}(x_l)\right|}{w'_{k_1} Q'_{k_1,k_2} \dots Q'_{k_{L-1},k_L}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\int_{\mathscr{Y}^L} \frac{\prod_{l=1}^L f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)} \mu(d\mathbf{x}) - \int_{\mathscr{Y}^L} \frac{\prod_{l=1}^L f_{\theta'_{k_l}}(x_l) f_{\theta'_{k'_l}}(x_l)}{p(\mathbf{x};\phi')} \mu(d\mathbf{x})\right| \\
&\le \frac{2 d_{TV}\left(\bigotimes_{l=1}^L F_{\theta_{k_l}}, \bigotimes_{l=1}^L F_{\theta'_{k_l}}\right)}{w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}} \\
&+ \frac{2 d_{TV}\left(P_\phi, P_{\phi'}\right)}{w'_{k_1} w_{k'_1} Q'_{k_1,k_2} Q_{k'_1,k'_2} \dots Q'_{k_{L-1},k_L} Q_{k'_{L-1},k'_L}} \\
&+ \frac{2 d_{TV}\left(\bigotimes_{l=1}^L F_{\theta_{k'_l}}, \bigotimes_{l=1}^L F_{\theta'_{k'_l}}\right)}{w'_{k_1} Q'_{k_1,k_2} \dots Q'_{k_{L-1},k_L}}.
\end{aligned}$$

Since convergence with respect to the total variation distance and to the Hellinger distance are equivalent, we get continuity of (90) with Proposition 6. Similarly, we have

$$\begin{aligned}
&\left|\int_{\mathscr{Y}^L} \frac{T_{\overline{k},j}(x_i) \prod_{l=1}^L f_{\theta_{k_l}}(x_l) f_{\theta_{k'_l}}(x_l)}{p(\mathbf{x};\phi)} \mu(d\mathbf{x}) - \int_{\mathscr{Y}^L} \frac{T_{\overline{k},j}(x_i) \prod_{l=1}^L f_{\theta'_{k_l}}(x_l) f_{\theta'_{k'_l}}(x_l)}{p(\mathbf{x};\phi')} \mu(d\mathbf{x})\right| \\
&\le \frac{\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \left|\prod_{l=1}^L f_{\theta_{k_l}}(x_l) - \prod_{l=1}^L f_{\theta'_{k_l}}(x_l)\right| \mu(d\mathbf{x})}{w_{k'_1} Q_{k'_1,k'_2} \dots Q_{k'_{L-1},k'_L}} \\
&+ \frac{\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \, |p(\mathbf{x};\phi) - p(\mathbf{x};\phi')| \, \mu(d\mathbf{x})}{w'_{k_1} w_{k'_1} Q'_{k_1,k_2} Q_{k'_1,k'_2} \dots Q'_{k_{L-1};k_L} Q_{k'_{L-1},k'_L}} \\
&+ \frac{\int |T_{k_l}(x_l)| \left|\prod_{i=l}^L f_{\theta_{k'_l}}(x_l) - \prod_{i=1}^L f_{k'_l,\theta'_{k'_l}}(x_l)\right| \mu(d\mathbf{x})}{w'_{k_1} Q'_{k_1,k_2} \dots Q'_{k_{L-1},k_L}}.
\end{aligned}$$

We have

$$\begin{aligned}
&\int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)|\,|p(\mathbf{x};\phi) - p(\mathbf{x};\phi')|\,\mu(d\mathbf{x}) \\
&\leq \sum_{1\leq k_1,\ldots,k_L\leq K} \int_{\mathscr{Y}^L} |T_{\overline{k},j}(x_i)| \left|\prod_{l=1}^{L} f_{\theta_{k_l}}(x_l) - \prod_{l=1}^{L} f_{\theta'_{k_l}}(x_l)\right| \mu(d\mathbf{x})
\end{aligned}$$

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

$$\begin{aligned}
&\int_{\mathscr{Y}} |T_{\overline{k},j}(x)| \left|f_{\theta_k}(x) - f_{\theta'_k}(x)\right| \nu(dx) \\
&\leq \sqrt{\int_{\mathscr{Y}} |T_{\overline{k},j}(x)|^2 \left|f_{\theta_k}(x) - f_{\theta'_k}(x)\right| \nu(dx)} \times \sqrt{2d_{TV}\left(F_{\theta_k}, F_{\theta'_k}\right)} \xrightarrow[\theta'_k\to\theta_k]{} 0.
\end{aligned}$$

for all $k \in [K]$ and $\theta_k \in \Theta_k$, we get continuity of (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 (88).

D.6 Proof of Theorem 9

We start the proof with two lemmas that ensure we fit into the framework of Proposition 2.

Lemma 17. *The information matrix $I(\phi)$ is definite positive for all ϕ in $\overline{\Phi}$.*

Lemma 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 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 2. From Proposition 1, we get $\overline{V} \leq (3+L)K^L = 5K^3$.

D.6.1 Proof of Lemma 17

For $\mathbf{k} = (k_1,\ldots,k_L) \in [K]^L$, the notation $wQ^{\bigcirc L}(\mathbf{k})$ is defined by (83). Following Theorem 1 of Meijer & Ypma [17], 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 (85), (86) and (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

$$\begin{aligned}0 = &-\sum_{\mathbf{k}\in[K]^L} wQ^{\bigcirc L}(\mathbf{k})\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L} \left(\sum_{l=1}^{L} \lambda_{\theta_{k_l}} x_l\right) &(91)\\
&+\sum_{k_1=1}^{K-1} \lambda_{w_{k_1}} \sum_{k_2,\dots,k_L} \frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}}\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L}\\
&-\sum_{k_1=1}^{K-1} \lambda_{w_{k_1}} \sum_{k_2,\dots,k_L} \frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}}\theta_K\theta_{k_2}\dots\theta_{k_L} e^{-\theta_K x_1-\dots-\theta_{k_L}x_L}\\
&+\sum_{l=2}^{L}\sum_{k_l=1}^{K-1}\sum_{k_i;i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}}\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_l}x_l}\\
&-\sum_{l=2}^{L}\sum_{k_l=1}^{K-1}\sum_{k_i;i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}}\theta_{k_1}\dots\theta_{k_{l-1}}\theta_K\theta_{k_{l+1}}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_K x_l-\dots-\theta_{k_l}x_l}.\end{aligned}$$

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

$$\begin{aligned}&0 = -wQ^{\bigcirc L}(\mathbf{k})\theta_{k_1}\dots\theta_{k_L}\left(\sum_{l=1}^{L} \lambda_{\theta_{k_l}} x_l\right) + \lambda_{w_{k_1}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{w_{k_1}}\theta_{k_1}\dots\theta_{k_L}\\
&+\sum_{l=2}^{L} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}}\theta_{k_1}\dots\theta_{k_L} \text{ for almost all } \mathbf{x}\\
&\Rightarrow 0 = \lambda_{\theta_{k_1}} = \dots = \lambda_{\theta_{k_L}} = \frac{\lambda_{w_{k_1}}}{w_{k_1}} + \sum_{l=2}^{L} \frac{\lambda_{Q_{k_{l-1},k_l}}}{Q_{k_{l-1},k_l}}.\end{aligned}$$

This implies $\lambda_{\theta_k} = 0$ for all $k \in [K-1]$ and there are quantities λ_w^* and λ_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, (91) becomes

$$\begin{aligned}0 = &\lambda_w^* \sum_{k_1=1}^{K-1}\sum_{k_2,\dots,k_L} wQ^{\bigcirc L}(\mathbf{k})\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_L}x_L} &(92)\\
&-\lambda_w^* \sum_{k_2,\dots,k_L}\left(\sum_{k_1=1}^{K-1} wQ^{\bigcirc L}(\mathbf{k})\right)\theta_K\theta_{k_2}\dots\theta_{k_L} e^{-\theta_K x_1-\dots-\theta_{k_L}x_L}\\
&+\sum_{l=2}^{L}\sum_{k_l=1}^{K-1}\sum_{k_i:i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}}\theta_{k_1}\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_{k_l}x_l}\\
&-\sum_{l=2}^{L}\sum_{k_l=1}^{K-1}\sum_{k_i:i\neq l} \lambda_{Q_{k_{l-1},k_l}} \frac{wQ^{\bigcirc L}(\mathbf{k})}{Q_{k_{l-1},k_l}}\theta_{k_1}\dots\theta_K\dots\theta_{k_L} e^{-\theta_{k_1}x_1-\dots-\theta_K x_l-\dots-\theta_{k_l}x_l}.\end{aligned}$$

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 $\mathbf{x} \mapsto e^{-\theta_K x_1 - \theta_{k_2} x_2 - \dots - \theta_{k_L} x_L}$ 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 $k \in [K-1]$,

$$\frac{\lambda_{Q_{K,k}}}{Q_{K,k}} = \lambda_w^* \beta_k \text{ with } \beta_k = \frac{\sum_{k'=1}^{K-1} w_{k'} Q_{k',k}}{w_K Q_{K,k}}. \tag{93}$$

Finally (92) becomes

$$\begin{aligned}
0 = & \lambda_w^* \sum_{k_1=1}^{K-1} \sum_{k_2,\dots,k_L} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \dots - \theta_{k_L} x_L} \\
& - \lambda_w^* \sum_{k_2,\dots,k_L} \left(\sum_{k_1=1}^{K-1} wQ^{\bigcirc L}(\mathbf{k}) \right) \theta_K \theta_{k_2} \dots \theta_{k_L} e^{-\theta_K x_1 - \dots - \theta_{k_L} x_L} \\
& + \lambda_Q^* \sum_{l=2}^{L} \sum_{k_{l-1},k_l \in [K-1]} \sum_{\substack{k_i \in [K]; \\ i \notin \{l-1,l\}}} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \dots - \theta_{k_L} x_L} \\
& + \lambda_w^* \sum_{l=2}^{L} \sum_{k_l \in [K-1]} \sum_{\substack{k_i \in [K]; \\ i \neq l}} \beta_{k_l} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_{l-2}} \theta_K \theta_{k_l} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \dots - \theta_K x_{l-1} - \theta_{k_l} x_l - \dots - \theta_{k_L} x_L} \\
& - \lambda_Q^* \sum_{l=2}^{L} \sum_{k_{l-1},k_l \in [K-1]} \sum_{\substack{k_i \in [K]; \\ i \notin \{l-1,l\}}} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_{l-1}} \theta_K \theta_{k_{l+1}} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \dots - \theta_{k_{l-1}} x_{l-1} - \theta_K x_l - \dots - \theta_{k_L} x_L} \\
& - \lambda_w^* \sum_{l=2}^{L} \sum_{k_l=1}^{K-1} \sum_{\substack{k_i \in [K]; \\ i \neq l}} \beta_{k_l} wQ^{\bigcirc L}(\mathbf{k}) \theta_{k_1} \dots \theta_{k_{l-2}} \theta_K \theta_K \theta_{k_{l+1}} \dots \theta_{k_L} e^{-\theta_{k_1} x_1 - \dots - \theta_K x_{l-1} - \theta_K x_l - \dots - \theta_{k_L} x_L}.
\end{aligned}$$

Identification with respect to $\mathbf{x} \mapsto e^{-\theta_K x_1 \dots - \theta_K x_K}$ 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 β_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 $\lambda_Q^* = \lambda K,1 = \dots = \lambda_{K,K-1} = 0$ which means $\lambda = 0$ and therefore the information matrix is definite positive.

D.6.2 Proof of Lemma 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 ψ 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 $\lim\limits_{n\to\infty} h\left(P_{\phi_n}, P_{\overline{\phi}}\right) = 0$, we must have $w_k^* (Q_{k,k}^*)^{L-1} = 0$.

This proves that P_{ϕ_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^*,θ^*) and $(\overline{w},\overline{Q},\overline{\theta})$ are equal up to a permutation σ on $[K]$. The ordering of the $\overline{\theta}_k$ and the θ_k^* ensures that this equality is true, not even up to a permutation.

D.7 Proof of Theorem 10

We just need to check that we satisfy Assumption 3. Then we can combine Proposition 3 and ??. We use Definition 41 [2] that allows to consider functions taking values in $(-\infty, +\infty]$. From Lemma 2.6.15 [21], we have that

$$\{\mathbf{x} \mapsto (x_1 - z_1)(x_2 - z_2); z_1,z_2 \in \mathbb{R}\} \subset \{\mathbf{x} \mapsto ax_1 + bx_2 + x_1x_2 + c; a,b,c \in \mathbb{R}\}$$

is VC-subgraph with VC-dimension smaller than or equal to 4. With Proposition 42-(v) [2], 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 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) [2] 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 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*) [2], we get that $\mathscr{S}_{\alpha,2}$ is VC-subgraph with VC-dimension at most 196. This proves that we satisfy Assumption 3 with

$$\overline{V} = 4 \times 196 = 784.$$

D.7.1 Proof of Lemma 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 $(u_i)_{i in[V+1]}$ which contradicts the fact that its VC-dimension is at most V.

D.7.2 Proof of Lemma 20

We adapt the proof of Lemma 2.6.18 [21]. 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 $n \leq V$.

D.8 Proof of Proposition 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 π 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 21 which proves the desired inequality on $\mathscr{Y}$.

Lemma 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 (π,z) in $\mathscr{Y}$.

We also get that g is lower bounded out of $\mathscr{Y}$ with the following lemma.

Lemma 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 π 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 π,z. We now relate the distance to π^* to the distance to (w^*,q^*) with the following result.

Lemma 23. *For $w,q_{12},q_{21} \in [0,1]$ we have*

$$\begin{aligned} &(\pi_{11} - \pi^*_{11})^2 + (\pi_{12} - \pi^*_{12})^2 + (\pi_{21} - \pi^*_{21})^2 \\ &\geq \max\left(\frac{1}{2}(w - w^*)^2, \frac{(1 - w^*)^2}{3} (q^*_{21} - q_{21})^2, (w^*)^2 (q_{12} - q^*_{12})^2 \right). \end{aligned}$$

This last result allows to conclude the proof of Proposition 3.

D.8.1 Proof of Lemma 21

We will repeatedly use the following inequality

$$\forall x,y > 0, \left| x^{1-\gamma} - y^{1-\gamma} \right| \geq \frac{(1-\gamma)|x - y|}{(x \vee y)^\gamma}. \tag{94}$$

Let (π,z) be in $\mathscr{Y}$. Our goal is to lower bound a on subsets of $\mathscr{Y}$ in a way that makes appearing the difference between some parameters. Inequalities (95), (96), (98) and (99) will be proved later.

- For $I_{11} = [-1,b)^2$ with $b = (z^* \wedge z \wedge 1) - 1$, we have

$$\int_{I_{11}} a(x_1,x_2)^2 dx_1 dx_2 \geq \frac{(1-\alpha)\left(1 \wedge |z^*|/2\right)^2}{16} (\pi^*_{11} - \pi_{11})^2. \tag{95}$$

- 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

$$\int_{I_{22}} a^2(x_1,x_2) dx \geq \frac{\alpha^2}{4^3} \left(\frac{3-\alpha}{2-\alpha} \right)^2 \left(\frac{1-\alpha}{5-3\alpha} \right)^{1-\alpha} (\pi^*_{22})^{1/\alpha} \left(1 \wedge |z^*|/2\right)^{1-\alpha} |z - z^*|^{1-\alpha}. \tag{96}$$

- Let $\beta \in (0,1]$. For

$$\begin{aligned} I_{12} &:= (-1, -(1 - z \wedge z^*)_+) \times (z \vee z^* + b_-, z \vee z^* + b_+), \\ I_{21} &:= (z \vee z^* + b_-, z \vee z^* + b_+) \times (-1, -(1 - z \wedge z^*)_+), \end{aligned}$$

with

$$\begin{aligned} b_+ &= \mathbb{1}_{z \vee z^* \geq \beta}(1 - |z - z^*|) + \mathbb{1}_{z \vee z^* < \beta} \frac{z \vee z^*(1-\beta)}{\beta} \\ &\geq \mathbb{1}_{z^* \geq \beta}(1 - \beta) + \mathbb{1}_{z^* < \beta} \frac{z^*(1-\beta)}{\beta} = (1 \wedge |z^*|/\beta)(1-\beta) \qquad (97) \end{aligned}$$

and $b_- = b_+ \delta$, $\delta \in (0,1)$. We have

$$\int_{I_{12}} a^2(x_1,x_2)dx \geq (\pi_{12}^* - \pi_{12})^2 \frac{(1-\alpha)^2 (1 \wedge |z^*|/2)}{8^2} (b_+)^{1-\alpha} (1-\delta) \mathbb{1}_{\Omega_{12}}, \qquad (98)$$

$$\int_{I_{21}} a^2(x_1,x_2)dx \geq (\pi_{21}^* - \pi_{21})^2 \frac{(1-\alpha)^2 (1 \wedge |z^*|/2)}{8^2} (b_+)^{1-\alpha} (1-\delta) \mathbb{1}_{\Omega_{21}}, \qquad (99)$$

with

$$\begin{aligned} I_{12} &:= \left\{ |\pi_{12}^* - \pi_{12}| \geq 2 \left[\frac{\alpha|z - z^*|}{\delta b_+} + |\pi_{11} - \pi_{11}^*|(1-\beta)^\alpha \right] \right\}, \\ I_{21} &:= \left\{ |\pi_{21}^* - \pi_{21}| \geq 2 \left[\frac{\alpha|z - z^*|}{\delta b_+} + |\pi_{11} - \pi_{11}^*|(1-\beta)^\alpha \right] \right\}. \end{aligned}$$

Combining (95), (96), (98) and (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 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 Ω_2 and Ω_3 are subsets of Θ given by

$$\Omega_i := \left\{ \theta \in \Theta; A_i(\theta) \geq C_A A_1(\theta) + C_B B(\theta) \right\}.$$

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 θ in Θ.

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 β and δ, which gives β^* depending only z^*, α and π^*. This concludes the proof of Lemma 21. We now prove the different inequalities.
Proof of (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 (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 (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

$$\int_{I_{22}} a^2(x_1,x_2)dx \geq \frac{\alpha^2\left(\pi^*_{22}\right)^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}}{4^2}$$
$$\times\left(\frac{2+(1-\alpha)(\pi^*_{22})^{1/\alpha}}{2+(1-\alpha)\left(1+(\pi^*_{22})^{1/\alpha}\right)}\right)^2\left(\frac{1-\alpha}{(1-\alpha)(2\pi+1)+2}\right)^{1-\alpha}$$
$$\geq \frac{\alpha^2\left(\pi^*_{22}\right)^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}}{4^2}$$
$$\times\left(\frac{3-\alpha}{2+2(1-\alpha)}\right)^2\left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}$$
$$= \frac{\alpha^2(3-\alpha)^2}{4^3\left(2-\alpha\right)^2}\left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}\left(\pi^*_{22}\right)^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}.$$

Finally, we always have have

$$\int_{I_{22}} a^2(x_1,x_2)dx \geq \frac{\alpha^2}{4^3}\left(\frac{3-\alpha}{2-\alpha}\right)^2\left(\frac{1-\alpha}{5-3\alpha}\right)^{1-\alpha}\left(\pi^*_{22}\right)^{1/\alpha}\left(1 \wedge |z^*|/2\right)^{1-\alpha}|z-z^*|^{1-\alpha}.$$

Proof of (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

$$\mathbb{1}_{x_2\leq 1}\frac{|x_2-z|}{|x_2|}\leq U(z,c_-,c_+):=\begin{cases}\frac{c_+}{z+c_+} & \text{if } z+c_+\leq 1,\\ 1-z & \text{if } z+c_-<1<z+c_+,\\ 0 & \text{if } 1\leq z+c_-.\end{cases}$$

For $c_+=\mathbb{1}_{z\geq\beta^*}(1-|z-z^*|)+\mathbb{1}_{z<\beta^*}\frac{z(1-\beta^*)}{\beta^*}$ we have $U(z,c_-,c_+)\leq 1-\beta^*$. We also have

$$\begin{aligned}1-\frac{|x_2-z|^\alpha}{|x_2-z^*|^\alpha}&\leq 1-\left(\frac{c_-}{c_-+|z-z^*|}\right)^\alpha\\ &\leq \alpha\frac{\frac{|z-z^*|}{c_-+|z-z^*|}}{\left(\frac{c_-}{c_-+|z-z^*|}\right)^{1-\alpha}}=\alpha\frac{|z-z^*|}{c_-}\left(\frac{c_-}{c_-+|z-z^*|}\right)^\alpha\\ &\leq \frac{\alpha|z-z^*|}{c_-}.\end{aligned}$$

Therefore, with $c_-=c_+\delta$, $\delta\in(0,1)$, on I_{12} we have

$$\frac{2|x_1|^{\alpha/2}|x_2-z|^{\alpha/2}}{1-\alpha}a(x_1,x_2)\geq\frac{\left[|\pi_{12}^*-\pi_{12}|-\frac{\alpha|z^*-z|}{b_-}-(\pi_{11}^*-\pi_{11})(1-\beta^*)^\alpha\right]_+}{2}.$$

If $|\pi_{12}^*-\pi_{12}|\geq 2\left[|\pi_{11}^*-\pi_{11}|(1-\beta^*)^\alpha+\alpha|z-z^*|/c_-\right]$ then

$$\frac{2|x_1|^{\alpha/2}|x_2-z|^{\alpha/2}}{1-\alpha}a(x_1,x_2)\geq\frac{|\pi_{12}^*-\pi_{12}|}{4}$$

and

$$\begin{aligned}\int_{I_{12}}a^2(x_1,x_2)dx&\geq\frac{(\pi_{12}^*-\pi_{12})^2}{8^2}\left[1-(1-z\wedge z^*)_+^{1-\alpha}\right](c_+)^{1-\alpha}\left[1-\delta^{1-\alpha}\right]\\ &\geq(\pi_{12}^*-\pi_{12})^2\frac{(1-\alpha)^2(1\wedge|z|\wedge|z^*|)(c_+)^{1-\alpha}(1-\delta)}{8^2}.\end{aligned}$$

Otherwise we have $|\pi_{12}^*-\pi_{12}|<2\left[|\pi_{11}^*-\pi_{11}|(1-\beta^*)^\alpha+\alpha|z-z^*|/b_-\right]$.

D.8.2 Proof of Lemma 22

We need to go through numerous cases and subcases. Let β^* be given in Lemma 21. Without loss of generality we are going to assume that $z^*>0$.
Case 1: $z\geq 0$ and $|z-z^*|\geq\beta^*$. Let c be a positive constant.

- *Subcase 1.1*: $z^*>z$ or ($z^*<z$ and $\pi_{22}\geq c^2\pi_{22}^*$). For $x\in I=(z\vee z^*+\beta^*,z\vee z^*+1)^2$, we have

$$a(x_1,x_2)=\frac{(1-\alpha)\left(\mathbb{1}_{z>z^*}\pi_{22}+\mathbb{1}_{z^*>z}\pi_{22}^*\right)}{2|x_1-z\vee z^*|^{\alpha/2}|x_2-z\vee z^*|^{\alpha/2}},$$

and therefore

$$\begin{aligned}\int_I a^2(x_1,x_2)dx&=\frac{\mathbb{1}_{z>z^*}\pi_{22}+\mathbb{1}_{z^*>z}\pi_{22}^*}{4}\left(1-(\beta^*)^{1-\alpha}\right)^2\\ &\geq\frac{c^2\pi_{22}^*(1-\alpha)^2}{4}(1-\beta^*)^2.\end{aligned}$$

- *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 $b = b' \left(\pi_{22}^*\right)^{1/2\alpha}$ we get

$$\begin{aligned} \int_I a^2(x_1,x_2)dx &\geq \frac{(z^*)^{2(1-\alpha)} \left(\pi_{22}^*\right)^{(1-\alpha)/\alpha} (b')^{2(1-\alpha)}}{4} \left(\sqrt{\pi_{22}^*} - \sqrt{\pi_{22}^*} \left(\frac{b'}{1 - (\pi_{22}^*)^{1/2\alpha} b'} \right)^{\alpha} \right)_+^2 \\ &\geq \frac{(z^*)^{2(1-\alpha)} \left(\pi_{22}^*\right)^{1/\alpha} (b')^{2(1-\alpha)}}{4} \alpha^2 \left(1 - \frac{b'}{1 - (\pi_{22}^*)^{1/2\alpha} b'} \right)_+^2. \end{aligned}$$

For $b' = \frac{1}{1 + 2\left(\pi_{22}^*\right)^{1/2\alpha} + \frac{1}{1-\alpha}}$, we have

$$\int_I a^2(x_1,x_2)dx \geq \frac{(z^*)^{2(1-\alpha)} \left(\pi_{22}^*\right)^{1/\alpha} \alpha^2}{4 \left(1 + 2 \left(\pi_{22}^*\right)^{1/2\alpha} + \frac{1}{1-\alpha} \right)^{2(1-\alpha)}} \left(1 - \frac{1}{1 + (\pi_{22}^*)^{1/2\alpha} + \frac{1}{1-\alpha}} \right)^2.$$

- *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

$$\begin{aligned}\int_{(z^*,z^*+b|z-z^*|)^2} a^2(x_1,x_2)dx &\geq \frac{|z-z^*|^{2(1-\alpha)}(1-\alpha)^{2(1-\alpha)/\alpha}\left(\pi^*_{22}\right)^{(1-\alpha)/\alpha}}{4}\pi^*_{22}\alpha^2 \\ &\geq \frac{\alpha^2\left(|z^*|/2\right)^{2(1-\alpha)}(1-\alpha)^{2(1-\alpha)/\alpha}\left(\pi^*_{22}\right)^{1)/\alpha}}{4}.\end{aligned}$$

D.8.3 Proof of Lemma 24.

- For θ in $\Omega_2 \cap \Omega_3$, we have

$$\begin{aligned}g(\theta) &\geq D_1 A_1^2 + D_{2,3}\left(A_2^2 + A_3^2\right) + D_B B^{1-\alpha} \\ &\geq \min\left(D_1, D_{2,3}, D_B\right)\left[A_1^2 + A_2^2 + A_3^2 + B^{1-\alpha}\right].\end{aligned}$$

- For θ in $\Omega_2 \cap \Omega_3^C$, we have

$$g(\theta) \geq D_1 A_1^2 + D_{2,3}A_2^2 + D_B B^{1-\alpha}$$

  and

$$A_3^2 < \left(C_A A_1 + C_B B\right)^2 \leq 2C_A^2 A_1^2 + 2C_B^2 B^{1-\alpha}.$$

  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 θ in $\Omega_2^C \cap \Omega_3^C$, we have

$$g(\theta) \geq D_1 A_1^2 + D_B B^{1-\alpha}$$

  and

$$A_2^2 + A_3^2 < 2\left(C_A A_1 + C_B B\right)^2 \leq 4C_A^2 A_1^2 + 4C_B^2 B^{1-\alpha}.$$

  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

$$g(\theta) \geq \min\left(\frac{D_B}{1+4C_B^2}, \frac{D_1}{1+4C_A^2}, D_{2,3}\right)\left[A_1^2 + A_2^2 + A_3^2 + B^{1-\alpha}\right].$$

D.8.4 Proof of Lemma 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

$$\begin{aligned}(\pi_{11}-\pi_{11}^*)^2 + (\pi_{12}-\pi_{12}^*)^2 &= w^2\left[2\left(q_{12}-\frac{1}{2}\right)^2+\frac{1}{2}\right] \\ &\quad - 2ww^*\left[2\left(q_{12}^*-\frac{1}{2}\right)\left(q_{12}-\frac{1}{2}\right)+\frac{1}{2}\right] \\ &\quad + (w^*)^2\left[2\left(q_{12}^*-\frac{1}{2}\right)^2+\frac{1}{2}\right] \\ &= \frac{1}{2}(w-w^*)^2 + 2\left(w\left(q_{12}-\frac{1}{2}\right) - w^*\left(q_{12}^*-\frac{1}{2}\right)\right)^2 \\ &\geq \frac{1}{2}(w-w^*)^2. \qquad (100)\end{aligned}$$

Therefore, we also have

$$\begin{aligned}&(\pi_{11}-\pi_{11}^*)^2 + (\pi_{12}-\pi_{12}^*)^2 + (\pi_{21}-\pi_{21}^*)^2 \\ &\geq \frac{1}{2}(w-w^*)^2 + ((1-w)q_{21}-(1-w^*)q_{21}^*)^2 \\ &= (1-w)^2\left[\frac{1}{2}+q_{21}^2\right] + (1-w^*)^2\left[\frac{1}{2}+(q_{21}^*)^2\right] \\ &- (1-w)(1-w^*)\left[1+2q_{21}q_{21}^*\right] \\ &= \left[\frac{1}{2}+q_{21}^2\right]\left((1-w)-(1-w^*)\frac{1+2q_{21}q_{21}^*}{1+2q_{21}^2}\right)^2 \\ &+ (1-w^*)^2\left[\frac{1}{2}+(q_{21}^*)^2\right] - \left[\frac{1}{2}+q_{21}^2\right](1-w^*)^2\left(\frac{1+2q_{21}q_{21}^*}{1+2q_{21}^2}\right)^2 \\ &\geq \frac{(1-w^*)^2}{2\left(1+2q_{21}^2\right)}\left[(1+2(q_{21}^*)^2)(1+2q_{21}^2) - (1+2q_{21}q_{21}^*)^2\right] \\ &= \frac{(1-w^*)^2}{1+2q_{21}^2}\left(q_{21}^*-q_{21}\right)^2 \\ &\geq \frac{(1-w^*)^2}{3}\left(q_{21}^*-q_{21}\right)^2. \qquad (101)\end{aligned}$$

- Similarly, we have

$$\begin{aligned}
(\pi_{11}-\pi_{11}^*)^2+(\pi_{12}-\pi_{12}^*)^2 &= w^2\left[q_{12}^2+(1-q_{12})^2\right]+(w^*)^2\left[(q_{12}^*)^2+(1-q_{12}^*)^2\right]\\
&-2ww^*\left[q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*)\right]\\
&=\left[q_{12}^2+(1-q_{12})^2\right]\left(w-w^*\frac{q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*)}{q_{12}^2+(1-q_{12})^2}\right)^2\\
&+(w^*)^2\left[(q_{12}^*)^2+(1-q_{12}^*)^2-\frac{(q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*))^2}{q_{12}^2+(1-q_{12})^2}\right]\\
&\geq\frac{(w^*)^2}{q_{12}^2+(1-q_{12})^2}\left[\left((q_{12}^*)^2+(1-q_{12}^*)^2\right)\left((q_{12})^2+(1-q_{12})^2\right)\right.\\
&\left.-(q_{12}q_{12}^*+(1-q_{12})(1-q_{12}^*))^2\right]\\
&=(w^*)^2\frac{(q_{12}-q_{12}^*)^2}{q_{12}^2+(1-q_{12})^2}\\
&\geq(w^*)^2(q_{12}-q_{12}^*)^2. \qquad (102)
\end{aligned}$$

Finally, with (100),(101) and (102), we get

$$\begin{aligned}
&(\pi_{11}-\pi_{11}^*)^2+(\pi_{12}-\pi_{12}^*)^2+(\pi_{21}-\pi_{21}^*)^2\\
&\geq\max\left(\frac{1}{2}(w-w^*)^2,\frac{(1-w^*)^2}{3}(q_{21}^*-q_{21})^2,(w^*)^2(q_{12}-q_{12}^*)^2\right).
\end{aligned}$$

E Selection of the spacing parameter

This section gathers the proofs of Theorem 11, 12, Lemma 7 and Corollary 6.

E.1 Proof of Theorem 11

We first need the following result.

Lemma 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 σ-finite measure μ. Let $\hat{P}=\hat{P}(n,\mathbf{X},\mathcal{M})$ be the ρ-estimator given by (7). For $t\in[n]$, there is an event Ω^* 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*

$$\begin{aligned}
\mathbb{1}_{\Omega^*}\sum_{i=1}^n h^2\left(P_i,\hat{P}\right)\leq\left(\frac{4a_0}{a_1}+1\right)&\inf_{Q\in\mathscr{M}}\sum_{i=1}^n h^2(P_i,Q)\\
&+\frac{8}{3a_1}(\xi+1.47)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right]+\frac{16.48}{a_1},
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n h^2\left(P_i,\hat{P}\right)\right]&\leq n\mathbb{P}\left((\Omega^*)^C\right)+\int_0^\infty\mathbb{P}\left(\mathbb{1}_{\Omega^*}\sum_{i=1}^n h^2\left(P_i,\hat{P}\right)\geq u\right)du\\
&\leq n\lceil n/t\rceil\beta_t(\mathbf{X})+\left(\frac{4a_0}{a_1}+1\right)\inf_{Q\in\mathscr{M}}\sum_{i=1}^n h^2(P_i,\mathscr{M})+\frac{16.48}{a_1}\\
&+\frac{8}{3a_1}(2.47+\log(2|\mathcal{M}|))\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right].
\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_{\hat{s}}\right)\Big|\mathbf{X}^{(1)}\right] \leq\ & c_0' \inf_{s\in S}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)}, \hat{P}_s\left(\mathbf{X}^{(1)}\right)\right)\\ &+ c_1'\left(2.47+\log(2|S|)\right)\left[1+96\sqrt{2.35}t\right]\\ &+ c_2' + n_2\lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right),\end{aligned}$$

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 no t in the upper bound. Let $\overline{P}$ be in $\mathscr{P}_X$. We get

$$\begin{aligned}\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] \leq\ & \frac{2}{n_2}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_{\hat{s}}\right)\right] + \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right)\\ \leq\ & \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{2c_0'}{n_2}\inf_{s\in S}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_s\right)\right]\\ &+ \inf_{t\in[n_2]}\left\{\frac{c_1'}{n_2}\left(2.47+\log(2|S|)\right)\left[1+96\sqrt{2.35}t\right] + 2\lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right)\right\}\\ &+ \frac{2c_2'}{n_2}.\end{aligned}$$

From (14), for s in S, we have

$$\begin{aligned}\frac{1}{n_2}\mathbb{E}\left[\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\hat{P}_s\right)\right] \leq\ & \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{4}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)\\ &+ \frac{4}{n_1}\mathbb{E}\left[\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\hat{P}_s\right)\right]\\ \leq\ & \frac{2}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{4}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)\\ &+ \frac{4c_0}{n_1}\inf_{Q\in\mathscr{M}_s}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},Q\right) + 4c_1\frac{(s+1)}{n_1}\left[17+D_{n(s,1)}(\mathscr{M}_s)\right]\\ &+ \frac{4c_2}{n_1}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}_{s,b}^*||\mathbf{P}_{s,b}^{ind}\right).\end{aligned}$$

We get

$$\begin{aligned}\mathbb{E}\left[h^2\left(\overline{P},\hat{P}_{\hat{s}}\right)\right] \leq\ & \frac{2+4c_0'}{n_2}\sum_{i=1}^{n_2} h^2\left(P_i^{(2)},\overline{P}\right) + \frac{8c_0'}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right)\\ &+ \inf_{t\in[n_2]}\left\{\frac{c_1'}{n_2}\left(2.47+\log(2|S|)\right)\left[1+96\sqrt{2.35}t\right] + 2\lceil n_2/t\rceil\beta_t\left(\mathbf{X}^{(2)}\right)\right\}\\ &+ \frac{2c_2'}{n_2} + \frac{8c_0'}{n_1}\inf_{s\in S}\left\{c_0\inf_{Q\in\mathscr{M}_s}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},Q\right)\right.\\ &\left.+ c_1(s+1)\left[D_{n(s,1)}(\mathscr{M})+17\right] + c_2\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}_{s,b}^*||\mathbf{P}_{s,b}^{ind}\right)\right\}.\end{aligned}$$

We also have

$$\frac{1}{n_1}\inf_{Q\in\mathscr{M}_s}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},Q\right) \leq 2h^2(\overline{P},\mathscr{M}_s) + \frac{2}{n_1}\sum_{i=1}^{n_1} h^2\left(P_i^{(1)},\overline{P}\right).$$

E.1.1 Proof of Lemma 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 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 Ω^* 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

$$\mathbf{P}^*\left(\sup_{q\in\mathcal{M}}\left\{\left|\mathbf{Z}_n(\mathbf{X},p,q)\right| \mathbb{1}_{\Omega^*} - \frac{a_1}{2}H^2_{P,Q}\right\} > \frac{2(\upsilon+\xi)}{3}\left[1+\sqrt{1+18ta_2^2\alpha}\right]\right) \leq 2|\mathcal{M}|e^{-\xi},$$

with $\mathbf{P}^* = \mathcal{L}(\mathbf{X})$ and $\alpha \geq \alpha_0(t) = \frac{32ta_2^2\delta}{a_1^2} + \frac{8}{3a_1}$.

We take $\delta = 1.175$ and $\upsilon = 1.47$ as in [4] Section A.1. Let $\xi > 0$ and $p \in \mathcal{M}$. On the event Ω^* defined by Lemma 26 and with Proposition 3 [4], we have for all $q \in \mathcal{M}$,

$$\begin{aligned}
\mathbf{T}_n\left(\mathbf{X},p,q\right) &\leq \mathbb{E}\mathbf{T}_n\left(\mathbf{X},p,q\right) + \left|\mathbf{Z}\left(\mathbf{X},p,q\right)\right| \\
&\leq \sum_{i=1}^{n}\left[a_0h^2\left(P_i,P\right) - a_1h^2\left(P_i,Q\right)\right] \\
&+ \frac{a_1}{2}H^2_{P,Q} + \frac{2(\xi+\upsilon)}{3}\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right] \\
&= \sum_{i=1}^{n}\left[\left(a_0+\frac{a_1}{2}\right)h^2\left(P_i,P\right) - \frac{a_1}{2}h^2\left(P_i,Q\right)\right] \\
&+ \frac{2}{3}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right].
\end{aligned}$$

Then,

$$\begin{aligned}
\boldsymbol{\Upsilon}_n\left(\mathbf{X},p\right) &= \sup_{q\in\mathcal{M}}\mathbf{T}_n\left(\mathbf{X},p,q\right) \\
&\leq \left(a_0+\frac{a_1}{2}\right)\sum_{i=1}^{n}h^2\left(\mathbf{P}_i^{ind},P\right) \\
&- \frac{a_1}{2}\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n}h^2\left(P_i,Q\right) \\
&+ \frac{2}{3}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right],
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Upsilon}_n\left(\mathbf{X},q\right) &= \sup_{q'\in\mathcal{M}}\mathbf{T}_n\left(\mathbf{X},q,p\right) \\
&\geq \mathbf{T}_n\left(\mathbf{X},q,p\right) = -\mathbf{T}_n\left(\mathbf{X},p,q\right) \\
&\geq -\left(a_0+\frac{a_1}{2}\right)\sum_{i=1}^{n}h^2\left(P_i,P\right) + \frac{a_1}{2}\sum_{i=1}^{n}h^2\left(P_i,Q\right) \\
&- \frac{2}{3}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right].
\end{aligned}$$

Since $\boldsymbol{\Upsilon}_n(\mathbf{X},\hat{p}) < \boldsymbol{\Upsilon}_n(\mathbf{X},p) + 8.24$, we have

$$\begin{aligned}\frac{a_1}{2}\sum_{i=1}^{n} h^2\left(P_i,\hat{P}\right) \leq 2\left(a_0+\frac{a_1}{2}\right)\sum_{i=1}^{n} h^2\left(P_i,P\right) - \frac{a_1}{2}\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2\left(P_i,\mathscr{M}\right) \\ + \frac{4}{3}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right] + 8.24.\end{aligned}$$

Given that $\mathscr{M}$ is finite we can take P such that

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

Hence we have

$$\begin{aligned}\sum_{i=1}^{n} h^2\left(P_i,\hat{P}\right) \leq \left(\frac{4a_0}{a_1}+1\right)\inf_{Q\in\mathscr{M}}\sum_{i=1}^{n} h^2\left(P_i,Q\right) \\ + \frac{8}{3a_1}(\xi+\upsilon)\left[1+\sqrt{1+18ta_2^2\alpha_0(t)}\right] + \frac{16.48}{a_1}.\end{aligned}$$

E.1.2 Proof of Lemma 26

Lemma 27. *For t in $[n]$, there is an event Ω^* such that $\mathbb{P}(\Omega^*) \geq 1 - \lceil n/t\rceil\beta_t(\mathbf{X})$ and*

$$\forall q,q' \in \mathcal{M}, \forall x > 0, \mathbb{P}\left(\left|\mathbf{Z}_n\left(\mathbf{X},q,q'\right)\right|\mathbb{1}_{\Omega^*} > \frac{2x}{3}\left[1+\sqrt{1+\frac{18ta_2^2H_{Q,Q'}^2}{x}}\right]\right) \leq 2e^{-x}. \tag{103}$$

Let $\xi > 0$ and $\alpha > 0$. We define $x_0 = \upsilon + \xi$ and for $j \geq 0$,

$$y_{j+1}^2 = \delta y_j^2 = \delta\alpha x_j. \tag{104}$$

Let q,q' be in $\mathcal{M}$. We apply Lemma 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

$$\alpha \geq \alpha_0(t) := \frac{32\delta ta_2^2}{a_1} + \frac{8}{3a_1}. \tag{105}$$

- 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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{q\in\mathcal{M}}\left\{\left|\mathbf{Z}_n(\mathbf{X},\overline{p},q)\right|\mathbb{1}_{\Omega^*}-\frac{a_1}{2}H^2_{\overline{p},q}\right\}>\frac{2x_0}{3}\left[1+\sqrt{1+18ta_2^2\alpha}\right]\right)\\
&\leq\sum_{\substack{q\in\mathcal{M}:\\ H^2_{\overline{P},Q}<y_0^2}}\mathbb{P}\left(\left|\mathbf{Z}_n(\mathbf{X},\overline{p},q)\right|\mathbb{1}_{\Omega^*}-\frac{a_1}{2}H^2_{\overline{p},q}>\frac{2x_0}{3}\left[1+\sqrt{1+18ta_2^2\alpha}\right]\right)\\
&+\sum_{j\geq 0}\sum_{\substack{q\in\mathcal{M}:\\ y_j^2\leq H^2_{\overline{p},q}<y^2_{j+1}}}\mathbb{P}\left(\left|\mathbf{Z}_n(\mathbf{X},\overline{p},q)\right|\mathbb{1}_{\Omega^*}-\frac{a_1}{2}H^2_{\overline{p},q}>0\right)\\
&\leq\sum_{\substack{q\in\mathcal{M}:\\ H^2_{\overline{p},q}<y_0^2}}2e^{-x_0}+\sum_{j\geq 0}\sum_{\substack{q\in\mathcal{M}:\\ y_j^2\leq H^2_{\overline{p},q}<y^2_{j+1}}}2e^{-x_j}\\
&\leq 2|\mathcal{M}|\left(e^{-x_0}+\sum_{j\geq 1}e^{-x_j}\right)=2|\mathcal{M}|\left(e^{-(\upsilon+\xi)}+\sum_{j\geq 1}e^{-\delta^j(\upsilon+\xi)}\right)\\
&\leq 2|\mathcal{M}|e^{-\xi}\left(e^{-\upsilon}+\sum_{j\geq 1}e^{-\delta^j\upsilon}\right)\leq 2|\mathcal{M}|e^{-\xi}.
\end{aligned}$$

E.1.3 Proof of Lemma 27

We follow the proof of Sart [20] (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 $B_1^*,\dots,B_{2lt}^*$ such that

- For $i=1,\dots,l$, the random vectors

$$B_{i,1}=\left(X_{2(i-1)t+1},\dots,X_{(2i-1)t}\right)\text{ and }B^*_{i,1}=\left(X^*_{2(i-1)t+1},\dots,X^*_{(2i-1)t}\right)\tag{106}$$

  have the same distribution, and so have the random vectors

$$B_{i,2}=\left(X_{(2i-1)t+1},\dots,X_{2it}\right)\text{ and }B^*_{i,2}=\left(X^*_{(2i-1)t+1},\dots,X^*_{2it}\right).\tag{107}$$

- The random vectors $B^*_{1,1},\dots,B*_{l,1}$ are independent. The random vectors $B^*_{1,2},\dots,B*_{l,2}$ 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

$$\begin{aligned}
Z^*_{q,q',1}:=&\sum_{i=1}^{l}\sum_{j=1}^{t}\left\{\psi\left(\sqrt{\frac{q'}{q}\left(X^*_{2(i-1)t+j}\right)}\right)-\mathbb{E}\left[\psi\left(\sqrt{\frac{q'}{q}\left(X^*_{2(i-1)t+j}\right)}\right)\right]\right\}\mathbb{1}_{2(i-1)t+j\leq n}\\
=&\sum_{i=1}^{l}\sum_{j=1}^{t}z^{q,q'}_{2(i-1)t+j}\mathbb{1}_{2(i-1)t+j\leq n}
\end{aligned}$$

and

$$\begin{aligned}
Z^*_{q,q',2} &:= \sum_{i=1}^{l}\sum_{j=1}^{t}\left\{\psi\left(\sqrt{\frac{q'}{q}}\left(X^*_{(2i-1)t+j}\right)\right) - \mathbb{E}\left[\psi\left(\sqrt{\frac{q'}{q}}\left(X^*_{(2i-1)t+j}\right)\right)\right]\right\}\mathbb{1}_{(2i-1)t+j\leq n} \\
&= \sum_{i=1}^{l}\sum_{j=1}^{m} z^{q,q'}_{(2i-1)t+j}\mathbb{1}_{(2i-1)t+j\leq n}.
\end{aligned}$$

Let ξ be a positive real number. Since

$$|Z_{q,q'}|\mathbb{1}_{\Omega^*} > \xi \Rightarrow |Z^*_{q,q',1}|\mathbb{1}_{\Omega^*} > \xi/2 \text{ or } |Z^*_{q,q',2}|\mathbb{1}_{\Omega^*} > \xi/2, \tag{108}$$

we have

$$\begin{aligned}
\mathbb{P}\left(|Z_{q,q'}|\,\mathbb{1}_{\Omega^*} > \xi\right) &\leq \mathbb{P}\left(|Z^*_{q,q',1}|\mathbb{1}_{\Omega^*} > \xi/2\right) + \mathbb{P}\left(|Z^*_{q,q',2}|\mathbb{1}_{\Omega^*} > \xi/2\right) \\
&\leq \mathbb{P}\left(|Z^*_{q,q',1}| > \xi/2\right) + \mathbb{P}\left(|Z^*_{q,q',2}| > \xi/2\right).
\end{aligned}$$

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

$$\begin{aligned}
V_{q,q',1} &= \sum_{i=1}^{l}\mathbb{E}\left[\left(\sum_{j=1}^{t} z_{q,q'}^{2(i-1)t+j}\mathbb{1}_{2(i-1)t+j}\right)^2\right] \\
&\leq \sum_{i=1}^{l}\sum_{j=1}^{t} t\mathbb{E}\left[\left(z_{q,q'}^{2(i-1)t+j}\right)^2\mathbb{1}_{2(i-1)t+j}\right] \\
&\leq t\sum_{i=1}^{n}\operatorname{Var}\left(\psi\left(\sqrt{\frac{q'}{q}}\left(X_i^*\right)\right)\right) \\
&\leq t\sum_{i=1}^{n} a_2^2\left[h^2(P_i,Q) + h^2(P_i,Q')\right] = ta_2^2H^2_{Q,Q'}.
\end{aligned}$$

The last inequality comes from Proposition 3 in Baraud & Birgé [4] 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 [16]) 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

$$|Z_{q,q'}|\mathbb{1}_{\Omega^*} > \frac{2x}{3}\left[1 + \sqrt{1 + \frac{18ta_2^2H^2_{Q,Q'}}{x}}\right]. \tag{109}$$

E.2 Proof of Lemma 7

We have

$$\begin{aligned}
\beta_t(\mathbf{Y}) &= \sup_i \beta\left(\sigma(Y_1,\ldots,Y_i);\sigma(Y_{i+t},\ldots,Y_n)\right) \\
&= \sup_i d_{TV}\left(\mathcal{L}\left(Y_1,\ldots,Y_i\right)\otimes\mathcal{L}\left(Y_{i+t},\ldots,Y_n\right),\mathcal{L}\left(Y_1,\ldots,Y_i,Y_{i+t},\ldots,Y_n\right)\right).
\end{aligned}$$

We use the notation $X_a^b = (X_a, \dots, X_b)$ and similarly for $\mathbf{E}$, $\mathbf{Y}$ and $\mathbf{Z}$. The triangle inequality implies

$$\begin{aligned}
&d_{TV}\left(\mathcal{L}\left(Y_1^i\right)\otimes\mathcal{L}\left(Y_{i+t}^n\right), \mathcal{L}\left(Y_1^n\right)\right)\\
&\le \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right) d_{TV}\left(\mathcal{L}(Y_1^i|E_1^i=e_1^i)\otimes\mathcal{L}(Y_{i+t}^n|E_{i+t}^N=e_{i+t}^n), \mathcal{L}(Y_1^i,Y_{i+t}^n|E_1^i=e_1^i,E_{i+t}^n=e_{i+t}^n)\right)\\
&= \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right)\beta\left(\sigma((X_j)_{\substack{j\le i,\\ e_j=1}}), \sigma((X_j)_{\substack{j\ge i+k,\\ e_j=1}})\right).
\end{aligned}$$

We now need the following result to conclude.

Lemma 28. *For any random variables A_1,A_2,B_1,B_2, we have*

$$\beta\left(\sigma(A_1),\sigma(A_2)\right) \le \beta\left(\sigma(A_1,B_1),\sigma(A_2,B_2)\right).$$

Combining the different inequalities above, we get

$$\begin{aligned}
\beta_t\left(\mathbf{Y}\right) &\le \sup_i \beta\left(\sigma(Y_1^i);\sigma(Y_{i+t}^n)\right)\\
&= \sup_i \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right)\beta\left(\sigma((X_j)_{\substack{j\le i,\\ e_j=1}}), \sigma((X_j)_{\substack{j\ge i+t,\\ e_j=1}})\right)\\
&\le \sup_i \sum_{\mathbf{e}\in\{0,1\}^n} \mathbb{P}\left(\mathbf{E}=\mathbf{e}\right)\beta\left(\sigma((X_j)_{j\le i}), \sigma((X_j)_{j\ge i+t})\right) = \beta_t\left(\mathbf{X}\right).
\end{aligned}$$

E.2.1 Proof of Lemma 28

Let μ_1, μ_2, ν_1 and ν_2 be measures dominating respectively $\mathcal{L}(A_1)$, $\mathcal{L}(A_2)$, $\mathcal{L}(B_1)$ and $\mathcal{L}(B_2)$. We have

$$\begin{aligned}
&\beta\left(\sigma(A_1),\sigma(A_2)\right)\\
&= \frac{1}{2}\int |p_A(a_1,a_2) - p_{A_1}(a_1)p_{A_2}(a_2)|\mu_1(da_1)\mu_2(da_2)\\
&= \frac{1}{2}\int |\int \left(p_{A,B}(a_1,b_1,a_2,b_2) - p_1(a_1,b_1)p_2(a_2,b_2)\right)\nu_1(db_1)\nu_2(db_2)|\mu_1(da_1)\mu_2(da_2)\\
&\le \frac{1}{2}\int |p_{A,B}(a_1,b_1,a_2,b_2) - p_1(a_1,b_1)p_2(a_2,b_2)|\nu_1(db_1)\nu_2(db_2)\mu_1(da_1)\mu_2(da_2\\
&= \beta\left(\sigma(A_1,B_1);\sigma(A_2,B_2)\right),
\end{aligned}$$

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 $p_2 = \frac{d\mathcal{L}(A_2,B_2)}{d\mu_2\otimes\mu_2}$.

E.3 Proof of Theorem 12

From (82) we have

$$\begin{aligned}
h^2\left(\overline{P},\mathscr{M}_s\right) &\le 2L\epsilon^2 + 2L(K-1)\delta(s) + 2h^2\left(\overline{P},\overline{\mathscr{M}}\right)\\
&\le 2L\epsilon^2 + 2h^2\left(\overline{P},\mathscr{M}\right) + 2(s+1)L\frac{\overline{V}}{n_1}.
\end{aligned}$$

From Proposition 5 we have $D_{n_1(s,1)}\left(\mathscr{M}_s\right) \le CL\overline{V}\log n_1$, for a constant C. For S defined by (71), we have

$$|S| = 2 + \lfloor \log_\tau(\lfloor (n_1-2)/2\rfloor)\rfloor \le 2 + \frac{\log n_1}{\log\tau} \le C\log n_1,$$

for some positive constant C. Theorem 11 allows to obtain (72).

The following result is proven in Section E.3.1.

Lemma 29. *Under Assumption 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*

$$h^2\left(P_i^{(j)},P^*\right) \leq C(Q^*)e^{-r(Q^*)i}, \tag{110}$$

- *for all* $t \in [n_2]$, *we have*

$$\beta_t\left(\mathbf{X}^{(2)}\right) \leq C(Q^*)e^{-r(Q^*)t/2}, \tag{111}$$

- *for all* $s \geq L-1$, *all* b *in* $[s+1]$,

$$\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) \leq n(s,b)C(Q^*)e^{-r(Q^*)s}. \tag{112}$$

From (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 (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 (112) gives

$$\begin{aligned}\sum_{b=1}^{s+1}\mathbf{K}\left(\mathbf{P}^*_{s,b}||\mathbf{P}^{ind}_{s,b}\right) &\leq C(Q^*)n_1e^{-r(Q^*)s} \\ &\leq C(Q^*)n_1\left(2\frac{2+\tau r(Q^*)^{-1}\log n_1}{n_1} \vee \frac{1}{n_1}\right) = 2C(Q^*)(2+\tau r(Q^*)^{-1}\log n_1).\end{aligned}$$

These last inequalities give (73).

E.3.1 Proof of Lemma 29

We just have to follow the proof of Lemma 13. We already have (110) and (112). The inequality (111) can be deduced from the inequality

$$d_{TV}\left(Q^t_{k,\cdot},\pi\right) \le Ce^{-rt},$$

and from the definition of β_t.

E.4 Proof of Corollary 6

We have

$$\mathbb{P}\left(X_i^{(j)} = \left(\overline{Y}_i^{(j)},\dots,\overline{Y}_{i+L-1}^{(j)}\right)\right) \ge \mathbb{P}\left(E_i^{(j)} = \dots = E_{i+L-1}^{(j)} = 1\right) = p_i^{(j)}p_{i+1}^{(j)}\dots p_{i+L-1}^{(j)},$$

and with the convexity of the squared Hellinger distance

$$\begin{aligned} h^2\left(P_i^{(j)},P^*\right) &\le p_i^{(j)}p_{i+1}^{(j)}\dots p_{i+L-1}^{(j)}h^2\left(\overline{P}_i^{(j)},P^*\right) + \left(1-p_i^{(j)}p_{i+1}^{(j)}\dots p_{i+L-1}^{(j)}\right) \\ &\le h^2\left(\overline{P}_i^{(j)},P^*\right) + \left(1-p_i^{(j)}\right) + \dots + \left(1-p_{i+L-1}^{(j)}\right), \end{aligned}$$

where $\overline{P}_i^{(j)} = \mathcal{L}\left(\overline{Y}_i^{(j)},\dots,\overline{Y}_{i+L-1}^{(j)}\right)$. One can check that $n \ge 1 + N/2$ with our conditions on L. With Theorem 12, Lemma 7 and Lemma 29 we have

$$\begin{aligned} C\mathbb{E}\left[h^2\left(P^*,\hat{P}_s\right)\right] \le{}& h^2\left(P^*,\mathscr{M}\right) + \frac{C(Q^*)}{n_1(e^{r(Q^*)}-1)} + \frac{C(Q^*)}{n_2(e^{r(Q^*)}-1)} \\ &+ L\epsilon^2 + \frac{L}{N_1}\sum_{i=1}^{N_1}\left(1-p_i^{(1)}\right) + \frac{L}{N_2}\sum_{i=1}^{N_2}\left(1-p_i^{(2)}\right) \\ &+ \inf_{t\in[n_2]}\left\{\frac{t\log\log n_1}{n_2} + \lceil n_2/t\rceil C(Q^*)e^{-r(Q^*)t/2}\right\} \\ &+ \inf_{s\in S}\left\{(s+1)L\overline{V}\frac{\log n_1}{n_1} + e^{-r(Q^*)s}\right\}, \end{aligned}$$

for some positive constant C and $s \ge L-1$. We can control the last terms with reasonable choices of t and s following the proof of Theorem 12.