Journal of Machine Learning Research 26 (2025) 1-75 Submitted 12/24; Revised 6/25; Published 7/25

Frontiers to the learning of nonparametric hidden Markov models

Kweku Abraham LKWA2@CAM.AC.UK
University of Cambridge, Statistical Laboratory
Wilberforce Road
Cambridge CB3 0WB, UK

Elisabeth Gassiat ELISABETH.GASSIAT@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS
Laboratoire de mathématiques d'Orsay
91405, Orsay, France

Zacharie Naulet ZNAULET@INRAE.FR
Université Paris-Saclay, INRAE
MaIAGE
78350, Jouy-en-Josas, France

Editor: Bharath Sriperumbudur

Abstract

Hidden Markov models (HMMs) are flexible tools for clustering dependent data coming from unknown populations, allowing nonparametric modelling of the population densities. Identifiability fails when the data is in fact independent and identically distributed (i.i.d.), and we study the frontier between learnable and unlearnable two-state nonparametric HMMs. Learning the parameters of the HMM requires solving a nonlinear inverse problem whose difficulty depends not only on the smoothnesses of the populations but also on the distance to the i.i.d. boundary of the parameter set. The latter difficulty is mostly ignored in the literature in favour of assumptions precluding nearly independent data. This is the first work conducting a precise nonasymptotic, nonparametric analysis of the minimax risk taking into account all aspects of the hardness of the problem, in the case of two populations. Our analysis reveals an unexpected interplay between the distance to the i.i.d. boundary and the relative smoothnesses of the two populations: a surprising and intriguing transition occurs in the rate when the two densities have differing smoothnesses. We obtain upper and lower bounds revealing that, close to the i.i.d. boundary, it is possible to "borrow strength" from the estimator of the smoother density to improve the risk of the other.

Keywords: Hidden Markov Models, Mixture Models, Inverse Problems, Nonparametric Estimation, Minimax

1. Introduction

1.1 Context and aim

Hidden Markov Models (HMMs) are a class of probabilistic models that play an important role in computer science and machine learning, particularly in the analysis of data sequences. They are widely used in various applications, including speech recognition and

natural language processing, due to their ability to model hidden states that evolve over time. This makes them ideal for capturing the evolution of sequences from different populations, effectively functioning as time-varying mixture models. Mixture models used for i.i.d. data require modelling assumptions on the population distributions (also called emission distributions), for example that they come from a parametric distribution; an advantage of HMMs is that identification can be obtained without such prior modelling (Gassiat, 2019). Thus, HMMs can be viewed as nonparametric mixture models that allow for greater flexibility in the emission distributions, making them particularly valuable in machine learning for their adaptability and robustness (Couvreur and Couvreur, 2000; Lefèvre, 2003; Lambert et al., 2003; Shang and Chan, 2009; Yau et al., 2011). Such flexibility has been discovered and studied in the recent years, see Section 1.3 for references and discussion. However, all theoretical results in this literature are asymptotic in nature, that is with the length n of the data sequence tending to infinity while model parameters are fixed. When the sequence of data is not far from being a sequence of i.i.d. observations, algorithms become unstable, making the output of the algorithms questionable (Rau et al., 2020). This is due to the fact that nonparametric mixtures are highly nonidentifiable and that identification algorithms for nonparametric HMMs proposed in previous literature involve tuning parameters for which no clues are given to address this issue. Indeed, in HMMs, the set of hidden Markov chain parameters and emission distributions can be divided into two subsets, the one for which the observations are not independent random variables (where identification is possible) and the one for which they form an i.i.d. sequence (where identification becomes impossible), and these two subsets share a boundary. Approaching the boundary makes learning more difficult.

The aim of our paper is to understand, in the possible learning properties of nonparametric HMMs, the interplay between the closeness to this boundary and the number of observations. The method we adopt for this purpose is to obtain nonasymptotic minimax rates in which the dependence to the i.i.d. frontier appears clearly together with the usual parameters such as the number of observations and the smoothness of probability emission densities. To obtain the upper bound, we propose a new estimation method which is straightforward to implement.

1.2 Contributions

We consider a two-state HMM with real-valued emissions, in which we observe the first n entries of a sequence $\boldsymbol{Y} = (Y_1, Y_2, \dots) \in [0,1]^{\mathbb{N}}$ which, under a parameter $\theta = (p, q, f_0, f_1)$, satisfies

$$\begin{aligned} \mathbb{P}_\theta(Y_n \in A \mid \boldsymbol{X}) &= \int_A f_{X_n}(y)\mathrm{d}y, \\ \boldsymbol{X} = (X_n)_{n\in\mathbb{N}} &\sim \mathrm{Markov}(\pi, Q_\theta), \end{aligned} \tag{1}$$

with the $Y_n,\ n \in \mathbb{N}$ conditionally independent given $\boldsymbol{X}$. The vector $\boldsymbol{X}$ of 'hidden states', which we assume is started from its invariant distribution $X_1 \sim \pi$, takes values in $\{0,1\}^{\mathbb{N}}$. The transition matrix of the chain is given by

$$Q = Q_\theta := \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}, \tag{2}$$

with the convention that for $j \geq 1$, $\mathbb{P}_\theta(X_{j+1} = 0 \mid X_j = 0) = 1 - p < 1$ and $\mathbb{P}_\theta(X_{j+1} = 0 \mid X_j = 1) = q > 0$. The functions $f_0, f_1 \in L^2([0,1])$ are density functions. Thus all Y_k, $k \geq 1$ follow the mixture distribution $\pi_0 f_0 + \pi_1 f_1$.

The goal is to estimate the parameter θ. This is a nonlinear inverse problem known to be solvable, up to a label-switching issue, even without any modelling assumptions on f_0 and f_1 (Gassiat et al., 2016; Alexandrovich et al., 2016): specifically, given that the highly non-identifiable i.i.d. nonparametric mixture is a degenerate submodel of a HMM, under conditions which rule out independence. There are three ways in which the data $(Y_n)_{n\in\mathbb{N}}$ can fail to exhibit dependence: when the hidden states themselves are in reality independently distributed; when the emission distributions are identical; or when only one population is observed. We adopt the minimax paradigm and we analyse the smallest maximum risk attainable over the following class of parameters. We define for some $\delta, \epsilon \in (0,1)$ and some $\zeta, s_0, s_1, R > 0$

$$\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \coloneqq \{\theta \ : p, q \geq \delta,\ |1-p-q| \geq \epsilon,\ \|f_0 - f_1\|_{L^2} \geq \zeta,\ \|f_i\|_{B^{s_i}_{2,\infty}} \leq R\}. \tag{3}$$

Here $\|\cdot\|_{B^s_{2,\infty}}$ denotes a Besov norm whose precise definition as used in this paper is delayed to equation (15) below. The space $B^s_{2,\infty}$ can be thought to be similar to the subspace of s-times differentiable functions with continuous s-derivative that are square-integrable, but it allows for slightly more general functions with comparable smoothness. We refer to Triebel (1983) for a thorough introduction to Besov spaces and their history. The quantities δ, ϵ and ζ lower bound the "distance" to the i.i.d. submodel. Indeed if $\delta = 0$, we may be unable to estimate both f_0 and f_1 since we may see data from one of these alone; if $\zeta = 0$ we may be unable to estimate p and q; and if $\epsilon = 0$ then we may be unable to identify the contributions of f_0 and f_1 to the mixture $\pi_0 f_0 + \pi_1 f_1$. We use concentration inequalities for Markov chains (Paulin, 2015) to build our estimators. This requires us to slightly shrink the set $\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$ and restrict our attention to parameters that are also in

$$\Sigma_{\gamma^*}(L) \coloneqq \{\theta \ : \ 1 - |1-p-q| \geq \gamma^*,\ \max_{j=0,1} \|f_j\|_\infty \leq L\}, \tag{4}$$

i.e. parameters with uniformly bounded emission densities (here $\|\cdot\|_\infty$ denotes the usual supremum norm) and having an absolute spectral gap. The assumption that the Markov chain starts from its stationary distribution could be relaxed as explained in (Paulin, 2015, Section 3.3), at the price of increasing the constants in the upper bounds, and longer proofs. We throughout use $\mathbb{P}_\theta$ to denote the law of $(\boldsymbol{X}, \boldsymbol{Y})$, and all induced marginal and conditional laws.

We are mainly interested in the regimes where δ, ϵ, ζ can be eventually small, and how the minimax risks for Q and f_0, f_1 over $\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$ are affected in these regimes.

The main message of our theorems may now be stated informally as follows (up to label switching and technical details relative to smoothnesses). The symbol $\asymp$ in the theorem means that expressions on the left and right side of $\asymp$ are proportional with a proportionality constant eventually depending on R, L and the absolute spectral gap of the chain $\boldsymbol{X}$, but nothing else.

Theorem 1 (Informal) *The minimax rate for estimating the transition matrix Q satisfies, for any norm $\|\cdot\|$,*

$$\inf_{\hat{Q}} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\big(\|\hat{Q} - Q\|^2\big) \asymp \frac{\max(\delta, \epsilon\zeta)^2}{\delta^2 \epsilon^4 \zeta^6} \frac{1}{n}.$$

The minimax rates for estimating f_0 and f_1 when $s_0 = s_1 = s$ satisfy

$$\inf_{\hat{f}_j} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\big(\|\hat{f}_j - f_j\|_{L^2}^2\big) \asymp \Big(\frac{1}{\delta^2 \epsilon^2 \zeta^2 n}\Big)^{2s/(2s+1)} + \frac{1}{\delta^2 \epsilon^4 \zeta^4 n},$$

while if $s_0 > s_1$ they satisfy

$$\inf_{\hat{f}_0} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\big(\|\hat{f}_0 - f_0\|_{L^2}^2\big) \asymp \Big(\frac{1}{\delta^2 \epsilon^2 \zeta^2 n}\Big)^{2s_0/(2s_0+1)} + \frac{1}{\delta^2 \epsilon^4 \zeta^4 n}$$

$$\inf_{\hat{f}_1} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\big(\|\hat{f}_1 - f_1\|_{L^2}^2\big) \asymp \Big(\frac{1}{\delta^2 n}\Big)^{2s_1/(2s_1+1)} + \frac{1}{\delta^2 \epsilon^4 \zeta^4 n},$$

and correspondingly if $s_0 < s_1$.

For a formal and rigorous statement of the minimax lower and upper bounds, we refer to Theorems 2, 3 (lower bounds), and to Theorems 5, 6, 8, and their Corollaries 7, 9 (upper bounds). The precise theorems are stated in a nonasymptotic manner. The asymptotic leading terms given in the above main results are in the case where the "distance" to frontier is large compared to n^{-a} for some (precisely defined) a. In this regime, the transition between the situation where emission densities have similar or different smoothnesses can be described as "$s_0 = s_1$" or "$s_0 > s_1$", but the transition appears in a more intricate manner when taking a nonasymptotic point of view. However, the main message is that some transition in the minimax rate occurs depending on the relative smoothnesses of the emission densities.

The transition in the rates arises due to a simple but unexpected phenomenon we call "sharing estimation strength", that can be described informally as follows. It is possible to estimate the combination $\psi_1 = \pi_0 f_0 + \pi_1 f_1$ at a good rate because it is simply the invariant density of Y_n. Hence a reasonable density estimator can estimate ψ_1 at rate $n^{-s/(1+2s)}$ where s is the smoothness of ψ_1, with no dependence on ϵ, δ, ζ. In the case where f_0 is much smoother than f_1, it may be more efficient to estimate f_0 and ψ_1, and estimate f_1 by plug in, rather than directly estimating f_1. This is reflected both in the upper bounds (see Theorem 6 and Theorem 8) and the lower bounds (see Theorem 3). The precise analysis of how one can "borrow" strength from the estimator of the smoother emission density to improve on the estimation rate for the rougher emission density is more involved, but this is the inspiration behind it.

1.3 Related work

It has been proved in (Gassiat et al., 2016; Alexandrovich et al., 2016) that once i.i.d. submodels are excluded, consistent estimation is possible for nonparametric HMMs without

prior modelling assumptions of the emission distributions. Moreover, no cost is incurred relative to the case where the underlying labels are observed. For s-smooth probability densities, the minimax rate $n^{-s/(1+2s)}$ is achieved using tensor methods in (De Castro et al., 2017) and using penalized least-squares estimation in (De Castro et al., 2016). This rate can be achieved adaptively in a "state-by-state" manner: up to a label-switching issue, one can achieve the rate $n^{-s_j/(1+2s_j)}$ if f_j has smoothness s_j, without knowledge of $(s_j,\ j = 0, 1)$, see (Lehéricy, 2018). See also (Lecestre, 2023) for robust estimation of the law of the observations in finite state space HMMs.

Earlier works do not consider the tradeoff between the required sample size and the required "distance" from independence, and it is this tradeoff that forms the focus of the current work, continuing from the previous article (Abraham et al., 2022b) in which we considered the model (1) but with f_0, f_1 densities with respect to counting measure on $\{1, \dots, K\}$ with known K. Discrete modeling is restrictive and extending the study to continuous densities with nonparametric modeling is important for applications. Some of the results in the continuous case mirror their discrete counterparts. For instance the minimax rate for estimating Q remains unchanged, though this is less trivial than it appears. While this might look obvious because for any function $h : [0, 1] \to \{1, \dots, K\}$, the pairs $((X_n, h(Y_n))_{n\geq 0}$ form a hidden Markov model with the same transition matrix Q. Finding a h for which Q is still identifiable from $(h(Y_n))_{n\geq 0}$ is however not straightforward, and it turns out that estimating Q requires first to solve a nonparametric problem (see Section 3.3). Moreover the nonparametric setting exhibits striking *qualitative*, as well as *quantitative*, differences relative to the discrete case. The rates for f_0 and f_1 in the nonparametric setting arise from delicate interplay between the smoothnesses s_0, s_1 and the parameters δ, ϵ, ζ. Also, the dependence of these rates in δ, ϵ, ζ differ between the discrete and the continuous case. A detailed comparison between this work and Abraham et al. (2022b) can be found in Section 3.8.

One additional novelty relative to other HMM papers in the nonparametric setting is that we use a wavelet block thresholding estimator. This allows us to adapt to the smoothnesses s_0 and s_1 without needing to use Lepski's method, and is thus, at least in principle, more computationally feasible.

1.4 Organisation of the paper

In Section 2 we give the lower bounds on the minimax risk for estimating Q and the densities f_0 and f_1. In Section 3 we derive the matching upper bounds. It is worth noting that the upper bounds are obtained via construction of estimators that are explicit and can be computed efficiently. Section 4 is devoted to the discussion of questions left open in our work. Proofs are relegated to the appendices.

2. Lower bounds

We give a lower bound for each component p and q of Q separately, which implies a bound for estimating (a permutation of) Q in any norm (since Q is a 2×2 matrix). The proof of Theorem 2 can be found in Section B.1. In the theorem $\epsilon_0 > 0$ is a constant whose precise value can in principle be computed.

Theorem 2 *Assume $n\delta^2\epsilon^4\zeta^6 \geq 1$, $\zeta \leq 1/(4\sqrt{3})$, $\epsilon \leq \epsilon_0$, $\delta \leq 1/6$, $R \geq 5/4 + 1/(8\sqrt{3})$ and $L \geq 5/8$. Then there exists a constant $c > 0$ such that*

$$\inf_{\hat{p}} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\big(|\hat{p} - p|^2\big) \geq \frac{c \max(\delta^2, \epsilon^2\zeta^2)}{\delta^2\epsilon^4\zeta^6} \frac{1}{n}$$

where the infimum is over all estimators $\hat{p}$ based on $Y_1, \dots, Y_n$. The same lower bound holds for the estimation of q.

We now consider the lower bounds for the estimation risk of the emission densities. Note that the lower bounds do not follow from standard density estimation (as in (Tsybakov, 2009)) because density estimation is not a submodel of HMM when one excludes the i.i.d. boundary of the parameter set. Surprisingly this fact appears to have been overlooked until the recent work of (Abraham et al., 2022a) where the first rigorous minimax lower bounds for estimating the densities have been established (see Section C of therein). The arguments therein rely on reducing to the simpler model where $\boldsymbol{X}$ is observed (so that the problem reduces to standard density estimation with two independent samples); this reduction is too severe to characterise the precise dependence of the minimax risk on δ, ϵ and ζ. To bypass the reduction to density estimation requires understanding the Kullback–Leibler divergence between arbitrary HMM distributions, which is challenging because of dependency. We establish the rates with the correct constants in the next theorem, whose proof can be found in Section B.2.

Theorem 3 *Assume $n\delta^2\epsilon^2\zeta^4 \geq 1$, $\zeta \leq 1/(4\sqrt{3})$, $\epsilon \leq \epsilon_0$ for a suitable $\epsilon_0 > 0$, $\delta \leq 1/6$, $R \geq 5/4 + 1/(8\sqrt{3})$ and $L \geq 5/8$. Then there exists a constant $c > 0$ such that*

$$\inf_{\check{f}_0} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\Big(\|\check{f}_0 - f_0\|_{L^2}^2\Big) \geq c \left\{ \frac{1}{\delta^2\epsilon^4\zeta^4 n} + \left(\frac{1}{\delta^2 n}\right)^{2s_0/(2s_0+1)} \right\}. \tag{5}$$

If moreover it holds $(n\delta^2\epsilon^2\zeta^4)^{-s_0/(1+2s_0)} \leq c_0\zeta$ and $\delta^{2s_1+1}(n\epsilon^2\zeta^2)^{(s_1-s_0)} \leq c_1$ for suitable constants c_0 and c_1, then there exists a constant $c > 0$ such that

$$\inf_{\check{f}_0} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\Big(\|\check{f}_0 - f_0\|_{L^2}^2\Big) \geq c \left\{ \frac{1}{\delta^2\epsilon^4\zeta^4 n} + \left(\frac{1}{\delta^2\epsilon^2\zeta^2 n}\right)^{2s_0/(2s_0+1)} \right\}. \tag{6}$$

The infima are over all estimators $\check{f}_0$ based on $Y_1, \dots, Y_n$. The same lower bounds hold for the estimation of f_1 by exchanging the role of s_0 and s_1 in the conditions and in the bounds.

Note that if $\boldsymbol{X}$ was observed, then we would on average see $n\pi_0 \gtrsim n\delta$ i.i.d. samples from f_0, hence we would be able to estimate f_0 with maximum risk $\lesssim (n\delta)^{-2s_0/(2s_0+1)}$ which is faster than the rates derived in Theorem 3 by at least a factor of $\delta^{-2s_0/(2s_0+1)}$. This shows that the inverse problem is fundamentally harder than standard density estimation.

This theorem calls for a number of comments. The first part of the theorem states that for the estimation of the emission densities, the minimax risk is lower bounded by a parametric term, and a nonparametric term with the usual rate $n^{-2s_0/(2s_0+1)}$ corrected with δ^2, that is with an effective sample size $\delta^2 n$ replacing n. The second part of the

theorem is more involved. It states that, if one of the emission density is smooth enough compared to the other one and relative to "frontier" parameters, the lower bound can be made larger, reducing the effective sample size to $\delta^2\epsilon^2\zeta^2 n$. If $s_0 > s_1$, this will eventually occur under the asymptotic regime where δ, ϵ, ζ do not decay too quickly to zero. Thus, the smoother emission density has a smaller effective sample size when getting close to the frontier (though still has a faster estimation rate overall).

3. Upper bounds

In this section we construct estimators whose maximum risk over $\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)$ match those established in the lower bounds of Theorems 2 and 3 in most cases.

3.1 The estimation procedure

Here we describe the heuristic we use to build a near minimax optimal estimator of $\theta = (p, q, f_0, f_1)$. As noted previously (Gassiat et al., 2016), understanding the law of three consecutive observations is key to recovering the model parameters. A reparametrisation simplifies the expression for said law, and allows the dependence on the parameters δ, ϵ and ζ to appear more naturally. Set

$$\phi(\theta) = \big(\tfrac{q-p}{p+q}\ ,\quad 1-p-q\ ,\quad \|f_0 - f_1\|_{L^2}\big), \quad \psi(\theta) = \big(\tfrac{qf_0+pf_1}{p+q}\ ,\quad \tfrac{f_0-f_1}{\|f_0-f_1\|_{L^2}}\big). \tag{7}$$

For $m \geq 1$, let $P^{(m)}_{\phi,\psi}$ denote the law of $(Y_1, \dots, Y_m)$ under parameter (ϕ, ψ), and let $p^{(m)}_{\phi,\psi}$ denote the corresponding density with respect to Lebesgue measure on $[0,1]^m$. In the parametrisation (7), defining for $\phi = (\phi_1, \phi_2, \phi_3)$

$$r(\phi) = \tfrac{1}{4}(1-\phi_1^2)\phi_2\phi_3^2, \tag{8}$$

one computes, with $f \otimes g$ defined by $(f \otimes g)(x, y) = f(x)g(y)$,

$$\begin{aligned} p^{(3)}_{\phi,\psi} = \psi_1 \otimes \psi_1 \otimes \psi_1 + r(\phi)\big(\psi_2 \otimes \psi_2 \otimes \psi_1 + \psi_1 \otimes \psi_2 \otimes \psi_2\big) \\ + \phi_2 r(\phi)\psi_2 \otimes \psi_1 \otimes \psi_2 - \phi_1\phi_2\phi_3 r(\phi)\psi_2 \otimes \psi_2 \otimes \psi_2. \end{aligned} \tag{9}$$

The parametrisation $\theta \mapsto (\phi, \psi)$ is invertible and has a simple inversion formula,

$$p = \tfrac{1}{2}(1-\phi_2)(1-\phi_1),\ q = \tfrac{1}{2}(1-\phi_2)(1+\phi_1), \tag{10}$$

$$f_0 = \psi_1 - \tfrac{1}{2}\phi_1\phi_3\psi_2 + \tfrac{1}{2}\phi_3\psi_2,\ f_1 = \psi_1 - \tfrac{1}{2}\phi_1\phi_3\psi_2 - \tfrac{1}{2}\phi_3\psi_2. \tag{11}$$

It is also possible to invert the map $(\phi, \psi) \mapsto p^{(3)}_{\phi,\psi}$ up to label switching issues. We now illustrate how this can be done to recover ϕ, ψ from $p^{(3)}_{\phi,\psi}$; we only describe ϕ_2 since it is the simplest to invert, but the same idea is applied to recover ϕ_1 (and consequently p, q) and the wavelet coefficients of f_0 and f_1 in Sections 3.4, 3.5 and 3.6. From formula (9), noting that $\langle \psi_2, 1\rangle = 0$ and $\langle \psi_1, 1\rangle = 1$, it is seen that for any bounded function h on $[0,1]$

$$\begin{aligned} r(\phi)\langle \psi_2, h\rangle^2 &= \mathbb{E}_\theta(h \otimes h) - \mathbb{E}_\theta(h)^2, \\ r(\phi)\phi_2\langle \psi_2, h\rangle^2 &= \mathbb{E}_\theta(h \otimes 1 \otimes h) - \mathbb{E}_\theta(h)^2. \end{aligned}$$

Provided $\langle \psi_2, h \rangle \neq 0$ the previous formula can be inverted to express ϕ_2 as a function of the "moments" $\mathbb{E}_\theta(h)$, $\mathbb{E}_\theta(h \otimes h)$ and $\mathbb{E}_\theta(h \otimes 1 \otimes h)$:

$$\phi_2 = \frac{\mathbb{E}_\theta(h \otimes 1 \otimes h) - \mathbb{E}_\theta(h)^2}{\mathbb{E}_\theta(h \otimes h) - \mathbb{E}_\theta(h)^2}. \tag{12}$$

Analogous formulas show that $(\phi, \psi) \mapsto p^{(3)}_{\phi,\psi}$ can be inverted (up to label-switching) upon computating of suitable moments of $p^{(3)}_{\phi,\psi}$, see Lemmas 13 and 14 for the other parameters. Then (p, q, f_0, f_1) is retrieved by using (10) and (11).

We propose to estimate (p, q) and the wavelet coefficients of f_0 and f_1 using the method of moments. In the inversion procedure described above we replace the moments by their empirical versions computed using

$$\mathbb{P}_n^{(s)}(H) \coloneqq \frac{1}{n-s+1} \sum_{i=1}^{n-s+1} H(Y_i, \dots, Y_{i+s-1}), \quad H : [0,1]^s \to \mathbb{R}, \ s \geq 1. \tag{13}$$

As suggested by equation (12), the formula for computing (ϕ, ψ) given the moments is unstable if the function h is chosen poorly, so that the estimates may be far from the true values if $\langle \psi_2, h \rangle$ is too small even if empirical moments are close to their means. No fixed choice of h works uniformly over the parameter space: given h, there exists a parameter (ϕ, ψ) such that $\langle \psi_2, h \rangle$ is arbitrarily small, resulting in an arbitrarily large maximum risk over $\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$. To avoid a deteriorated maximum risk, it is therefore necessary to estimate h from the data. The oracle choice for h would maximize $h \mapsto \frac{|\langle \psi_2, h \rangle|}{\|h\|}$ and hence be given by $h = \psi_2$. Thus, a crucial step in our estimation procedure is to provide an initial (crude) estimator $\tilde{\psi}_2$ of ψ_2 such that $\|\tilde{\psi}_2\|_{L^2} = 1$ and such that

$$\tilde{\mathcal{I}} \coloneqq \langle \psi_2, \tilde{\psi}_2 \rangle \tag{14}$$

is sufficiently bounded away from zero with high probability under each parameter $(\phi, \psi) \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)$. For this reason we describe $\tilde{\psi}_2$ as a separating function: since $\psi_2 = (f_0 - f_1)/\|f_0 - f_1\|_{L^2}$, finding $\tilde{\psi}_2$ is tantamount to finding an hyperplane in $L^2[0,1]$ which separates f_0 and f_1 sufficiently well. The estimator $\tilde{\psi}_2$ is built in Section 3.3.

Algorithm 1 summarizes the complete estimation procedure. A full version of the estimation algorithm with discussion of its computational complexity is deferred to Section 3.7. Computing our estimator involves only elementary operations, namely: (i) determining the leading eigenvector of a relatively small matrix, (ii) calculating empirical averages, and (iii) performing straightforward algebraic manipulations. This makes our estimator both practical to implement and computationally efficient. Notably, unlike certain alternative estimators – such as the least squares estimator (De Castro et al., 2016) – our approach does not require solving a nonconvex optimization problem, ensuring that the estimator can always be reliably computed. Also, our procedure exploits the appealing adaption properties of wavelet estimators, avoiding to use Lepski's method to achieve rate adaptation (Lehéricy, 2018).

Before entering the details of the estimation procedure, we recall some classical results about wavelets and Besov spaces in Section 3.2.

Algorithm 1 Estimation procedure

Require: An observed chain $(Y_1, \dots, Y_n)$.
Ensure: Estimators $\hat{p}$, $\hat{q}$, $\hat{f}_0$ and $\hat{f}_1$.
1: Estimation of a good separating function $\tilde{\psi}_2$ (see Section 3.3)
2: Estimation of (ϕ_1, ϕ_2) and then (p, q) (see Section 3.4)
3: Estimation of (f_0, f_1) using block thresholding with estimators of the wavelet coefficients (see Section 3.5 for the case $s_0 = s_1$, or Section 3.6 otherwise).

3.2 Preliminaries on wavelets and the Besov norm we use

Throughout the paper we use the S-regular boundary-corrected wavelet basis of (Cohen et al., 1993), see also e.g. (Giné and Nickl, 2016, Section 4.3.5), denoted $\{\{\Phi_{Jk} : k = 0, \dots, 2^{J-1}\}, \{\Psi_{jk} : j \geq J, k = 0, \dots, 2^j - 1\}\}$, with initial resolution level J chosen as in the latter reference. As is common, we will refer to the (Φ_{Jk}) as father wavelets and to the (Ψ_{jk}) as mother wavelets. Any $f \in L^2[0,1]$ has the series expansion

$$f = \sum_{k=0}^{2^J-1} \langle \Phi_{Jk}, f \rangle \Phi_{Jk} + \sum_{j=J}^{\infty} \sum_{k=0}^{2^j-1} \langle \Psi_{jk}, f \rangle \Psi_{jk}$$

with convergence of the series in $L^2[0,1]$. In fact, as our densities will be assumed regular enough, wavelet series expansions for f_0 and f_1 will also converge uniformly (e.g. Giné and Nickl, 2016, eq. (4.71))). Furthermore, it is well-known that the Besov space $B^s_{2,\infty}$ can be characterised via the wavelet coefficients. Indeed the norm for $B^s_{2,\infty}$ that we will use (see e.g. (Giné and Nickl, 2016, Equation (4.166))) is given by

$$\|f\|^2_{B^s_{2,\infty}} := \sum_{k=0}^{2^J-1} \langle \Phi_{Jk}, f \rangle^2 + \sup_{j \geq J} 2^{2js} \sum_{k=0}^{2^j-1} \langle \Psi_{jk}, f \rangle^2. \tag{15}$$

3.3 Estimation of a separating hyperplane

As explained in Section 3.1, our estimation procedure is based on computing empirical averages of the type $\mathbb{P}_n^{(2)}(\tilde{\psi}_2 \otimes f) = \frac{1}{n-1} \sum_{i=1}^{n-1} \tilde{\psi}_2(Y_i) f(Y_{i+1})$ where $\tilde{\psi}_2$ is a crude estimator of ψ_2. If $\tilde{\psi}_2$ is also estimated from $(Y_1, \dots, Y_n)$, it is not clear at all that these empirical average approach $\mathbb{E}_\theta(\psi_2(Y_1) f(Y_2))$, as they are sum of somewhat complex dependent random variables, each term of which depends on the whole sample $(Y_1, \dots, Y_n)$. A classical trick is to estimate $\tilde{\psi}_2$ using a sample $(\tilde{Y}_1, \dots, \tilde{Y}_n)$ that is independent from the sample $(Y_1, \dots, Y_n)$ used to compute the average $\frac{1}{n-1} \sum_{i=1}^{n-1} \tilde{\psi}_2(Y_i) f(Y_{i+1})$. In the context of HMM, however, the sample cannot be split into two independent parts. Fortunately this is not too worrisome. As explained in Section A, it is possible to split the sample $(Y_1, \dots, Y_n)$ into three parts, and then use the first third to estimate $\tilde{\psi}_2$ and the last third for empirical averages. This way, deviation inequalities for the empirical averages of functions involving $\tilde{\psi}_2$ can be achieved as if $\tilde{\psi}_2$ were independent of the observations used in the empirical averages, up to a term $Ce^{-c\gamma^* n}$ for C and c universal constants. Thus, to facilitate reading, we will throughout assume that $\tilde{\psi}_2$ is estimated using a sample $(\tilde{Y}_1, \dots, \tilde{Y}_n) \sim P^{(n)}_{\phi,\psi}$ independent of $(Y_1, \dots, Y_n)$.

For notational convenience, we define the set of wavelet indices

$$\Lambda(M) \coloneqq \{0, 1, \dots, 2^{J-1}\} \cup \{(j,k) \,:\, j = J, \dots, M,\ k = 0, \dots, 2^j - 1\}$$

including all father indices and mother indices of levels $J \le j \le M$, and for all $\lambda \in \Lambda(M)$ we set $e_\lambda = \Phi_{Jk}$ if $\lambda = k$ and $e_\lambda = \Psi_{jk}$ if $\lambda = (j,k)$.

For M large enough (see Theorem 4 below) compute the $2^M \times 2^M$ matrix $\tilde{\mathcal{G}}$ with entries

$$\tilde{\mathcal{G}}_{\lambda,\lambda'} = \frac{1}{2}\tilde{\mathbb{P}}_n^{(2)}(e_\lambda \otimes e_{\lambda'} + e_{\lambda'} \otimes e_\lambda) - \tilde{\mathbb{P}}_n^{(1)}(e_\lambda)\tilde{\mathbb{P}}_n^{(1)}(e_{\lambda'}).$$

The matrix $\tilde{\mathcal{G}}$ is an estimator of the matrix $\mathcal{G}$ with entries

$$\mathcal{G}_{\lambda,\lambda'} = \frac{1}{2}\mathbb{E}_\theta(e_\lambda \otimes e_{\lambda'} + e_{\lambda'} \otimes e_\lambda) - \mathbb{E}_\theta(e_\lambda)\mathbb{E}_\theta(e_\lambda) = r(\phi)\langle\psi_2, e_\lambda\rangle\langle\psi_2, e_{\lambda'}\rangle$$

where the second equality follows from equation (9). Hence, $\mathcal{G}$ is proportional to the Gram matrix of the vector $V_\theta \propto (\langle\psi_2, e_\lambda\rangle \,:\, \lambda \in \Lambda(M))$. The matrices $\tilde{\mathcal{G}}$ and $\mathcal{G}$ are real symmetric, and thus by the spectral theorem are always diagonalizable. By concentration arguments, we expect that $\tilde{\mathcal{G}}$ will have an eigenvalue approximately equal to $r(\phi)$ (which can be positive or negative) and the rest of eigenvalues will be smaller in absolute value. The eigenvector $\tilde{V}$ (chosen such that $\|\tilde{V}\| = 1$) corresponding to the leading eigenvalue is then an estimator of $\pm V_\theta/\|V_\theta\|$. We suggest to set

$$\tilde{\psi}_2(x) \coloneqq \frac{\max\big(-\tau,\, \min\big(\tau, \sum_{\lambda\in\Lambda(M)} \tilde{V}_\lambda e_\lambda(x)\big)\big)}{\Big(\int_0^1 \max\big(-\tau,\, \min\big(\tau, \sum_{\lambda\in\Lambda(M)} \tilde{V}_\lambda e_\lambda(y)\big)\big)^2 dy\Big)^{1/2}}$$

where the truncation $\tau \ge 1$ is intended to prevent technicalities within the proofs. The next theorem shows that $\tilde{\psi}_2$ is well aligned with ψ_2 with high probability under $\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)$. The proof of Theorem 4 can be found in Section C.7.

Theorem 4 *Suppose for some $L \ge 1$, $\zeta > 0$, $R > 0$, $s_* > 0$, $M \ge J$ we have*

$$\tau \ge \frac{L}{\zeta}, \qquad 2^{-Ms_*} \le \frac{\zeta\sqrt{2^{2s_*} - 1}}{4R}.$$

There exists a constant $C > 0$ such that for all $S \ge s_0, s_1 \ge s_*$ and all $\gamma^* > 0$

$$\sup_{\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)} \tilde{\mathbb{P}}_\theta\Big(|\langle\tilde{\psi}_2, \psi_2\rangle| \le \frac{7}{8}\Big) \le 2 \cdot 24^{2^M} \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^2\zeta^4}{L^3 + 2^M\sqrt{L}\delta\epsilon\zeta^2}\right).$$

3.4 Parametric part

Define $m(\phi) = (m(\phi)_1, m(\phi)_2, m(\phi)_3)$ by

$$\begin{aligned}
m(\phi)_1 &\coloneqq \mathbb{E}_\theta[\tilde{\psi}_2(Y_1)\tilde{\psi}_2(Y_2) \mid \tilde{\psi}_2] - \mathbb{E}_\theta[\tilde{\psi}_2(Y_1)^2 \mid \tilde{\psi}_2],\\
m(\phi)_2 &\coloneqq \mathbb{E}_\theta[\tilde{\psi}_2(Y_1)\tilde{\psi}_2(Y_3) \mid \tilde{\psi}_2] - \mathbb{E}_\theta[\tilde{\psi}_2(Y_1)^2 \mid \tilde{\psi}_2],\\
m(\phi)_3 &\coloneqq -\mathbb{E}_\theta[\tilde{\psi}_2(Y_1)\tilde{\psi}_2(Y_2)\tilde{\psi}_2(Y_3) \mid \tilde{\psi}_2] + \mathbb{E}_\theta[\tilde{\psi}_2(Y_1)^3 \mid \tilde{\psi}_2] + \big(2m(\phi)_1 + m(\phi)_2\big)\mathbb{E}_\theta[\tilde{\psi}_2(Y_1) \mid \tilde{\psi}_2].
\end{aligned}$$

This can be estimated by the following empirical quantities:

$$\hat{m}_1 \coloneqq \mathbb{P}_n^{(2)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{P}_n^{(1)}(\tilde{\psi}_2)^2,$$
$$\hat{m}_2 \coloneqq \mathbb{P}_n^{(3)}(\tilde{\psi}_2 \otimes 1 \otimes \tilde{\psi}_2) - \mathbb{P}_n^{(1)}(\tilde{\psi}_2)^2,$$
$$\hat{m}_3 \coloneqq -\mathbb{P}_n^{(3)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2) + \mathbb{P}_n^{(1)}(\tilde{\psi}_2)^3 + \big(2\hat{m}_1 + \hat{m}_2\big)\mathbb{P}_n^{(1)}(\tilde{\psi}_2).$$

Easy computations lead to (recall $\tilde{\mathcal{I}} \coloneqq \langle \psi_2, \tilde{\psi}_2 \rangle$ in (14) and $r(\phi) = (1/4)(1-\phi_1^2)\phi_2\phi_3^2$ in (8))

$$m(\phi) \equiv \big(r(\phi)\tilde{\mathcal{I}}^2,\, r(\phi)\phi_2\tilde{\mathcal{I}}^2,\, r(\phi)\phi_1\phi_2\phi_3\tilde{\mathcal{I}}^3\big), \tag{16}$$

see Lemma 13 in Appendix C. The moments in the previous display can be inverted modulo label-switching. Namely, it is possible to express $\phi_1 \mathrm{sgn}(\tilde{\mathcal{I}})$, ϕ_2, and $\phi_3|\tilde{\mathcal{I}}|$ as functions of $m(\phi)$. The inversion formulas for m are given in Lemma 14. By replacing $m(\phi)$ with the empirical estimates in the inversion formula we define

$$\hat{\phi}_1 \coloneqq \frac{\hat{m}_3}{[4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2]^{1/2}}, \qquad \hat{\phi}_2 \coloneqq \max\Big(-1,\, \min\Big(\frac{\hat{m}_2}{\hat{m}_1},\, 1\Big)\Big).$$

Notice that since $m(\phi)_2 \geq 0$, we replaced $\hat{m}_2$ by $(\hat{m}_2)_+ = \max(\hat{m}_2, 0)$. We then build an estimator of p and q justified by (10) by letting

$$\hat{p} = \tfrac{1}{2}(1-\hat{\phi}_1)(1-\hat{\phi}_2),$$
$$\hat{q} = \tfrac{1}{2}(1+\hat{\phi}_1)(1-\hat{\phi}_2).$$

To account for label switching, write Q^σ for the matrix with entries $(Q^\sigma)_{ij} = Q_{\sigma(i),\sigma(j)}$ for a permutation σ. We consider the loss relative to the Frobenius norm $\|\cdot\|_{\mathrm{F}} \coloneqq \sum_{i,j}(\cdot)_{i,j}^2$. The proof of Theorem 5 can be found in Section C.4

Theorem 5 *Assume that $\zeta \leq 1$, that τ and M are chosen as prescribed in Theorem 4, and that $n\gamma^* \geq \tau^6/L^3$. Then there are universal constants $B, C > 0$ such that*

$$\sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} \inf_{\sigma} \mathbb{E}_\theta\Big(\|Q_{\hat{\theta}}^\sigma - Q_\theta\|_{\mathrm{F}}^2\Big) \leq 2 \cdot 24^{2^M} \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^2\zeta^4}{L^3 + 2^M\sqrt{L}\delta\epsilon\zeta^2}\right)$$
$$+ B\exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^4\zeta^6}{L^3 + \max(\tau, \sqrt{L})^3\delta\epsilon^2\zeta^3}\right) + \frac{BL^3\max(\delta^2, \epsilon^2\zeta^2)}{\delta^2\epsilon^4\zeta^6}\frac{1}{n\gamma^*}.$$

In an asymptotic regime, the first terms in the bound in Theorem 5 can be neglected and our estimator achieves the rate of convergence $\frac{L^3\max(\delta^2,\epsilon^2\zeta^2)}{\delta^2\epsilon^4\zeta^6}\frac{1}{n\gamma^*}$, which is, up to constants, the minimax rate established in Theorem 2. We note that the parametric part $Q_{\hat{\theta}}$ achieves the same rate in the nonparametric setting as in the multinomial setting (Abraham et al., 2022b); at first glance this seems unsurprising in view of the fact that the pairs $((X_n, h(Y_n))_{n\geq 0}$ form a hidden Markov model with transition matrix Q_θ for any function h, so that for a suitable h we can reduce to a parametric setting. However, reducing to a parametric setting in which Q_θ is still identifiable is in fact a nonparametric problem (as alluded to in Section 1.3, or see Section 3.1 for more details), so that getting the same

minimax parametric rate is not a priori guaranteed. Indeed, to construct an estimator for the parametric part (p, q), we must first solve the nonparametric problem of estimating ψ_2. This step does not harm the risk of our estimator and we are able to match the semiparametric rate given in Theorem 2. This is because the estimator $\tilde{\psi}_2$ does not need to be a *good* estimator of ψ_2 (it is not required even to be consistent), but must only guarantee that $\tilde{\mathcal{I}} = \langle \psi_2, \tilde{\psi}_2 \rangle$ does not get too small.

3.5 Nonparametric part: case $s_0 = s_1$

Using the ideas like in Section 3.1, the wavelet coefficients of f_0 and f_1 can be extracted from $\{\mathbb{E}_\theta(\tilde{\psi}_2 \otimes \Phi_{Jk})\}$, $\{\mathbb{E}_\theta(\tilde{\psi}_2 \otimes \Psi_{jk})\}$, $\{\mathbb{E}_\theta(\Phi_{Jk})\}$, $\{\mathbb{E}_\theta(\Psi_{jk})\}$ and $\mathbb{E}_\theta(\tilde{\psi}_2)$, and further estimated using their empirical relatives. Given these empirical wavelets coefficients, we construct estimators for f_0 and f_1 based on block-thresholding the coefficients.

For notational convenience, we write $f^{\Phi_{Jk}} := \langle \Phi_{Jk}, f \rangle$ and $f^{\Psi_{jk}} := \langle \Psi_{jk}, f \rangle$. First, using the inversion formulas for m given in Lemma 14 and by replacing $m(\phi)$ with the empirical estimates in the inversion formula we define an estimator of $g := \phi_3 |\tilde{\mathcal{I}}|$ by

$$\hat{g} := \frac{\sqrt{4\hat{m}_1^2 (\hat{m}_2)_+ + \hat{m}_3^2}}{\hat{m}_2} \mathbf{1}_{\{\hat{m}_2 > 0\}}.$$

Now, our goal is to find estimators $\{(\hat{f}_0^{\Phi_{Jk}})_k, (\hat{f}_0^{\Psi_{jk}})_{jk}\}$ of $\{(f_0^{\Phi_{Jk}})_k, (f_0^{\Psi_{jk}})_{jk}\}$ (and similarly for f_1). We use (11) and we set

$$\begin{aligned}
\hat{G}^{\Phi_{Jk}} &:= \mathbb{P}_n^{(2)}(\tilde{\psi}_2 \otimes \Phi_{Jk}) - \mathbb{P}_n^{(1)}(\tilde{\psi}_2)\mathbb{P}_n^{(1)}(\Phi_{Jk}), \\
\hat{f}_0^{\Phi_{Jk}} &:= \mathbb{P}_n^{(1)}(\Phi_{Jk}) + \frac{\hat{g}(1 - \hat{\phi}_1)}{2\hat{m}_1} \mathbf{1}_{\{\hat{m}_1 \neq 0\}} \hat{G}^{\Phi_{Jk}}, \\
\hat{f}_1^{\Phi_{Jk}} &:= \mathbb{P}_n^{(1)}(\Phi_{Jk}) - \frac{\hat{g}(1 + \hat{\phi}_1)}{2\hat{m}_1} \mathbf{1}_{\{\hat{m}_1 \neq 0\}} \hat{G}^{\Phi_{Jk}}.
\end{aligned}$$

The same definition applies *mutatis mutandis* to the estimators of the mother coefficients $\hat{f}_0^{\Psi_{jk}}$, $\hat{f}_1^{\Psi_{jk}}$, and $\hat{G}^{\Psi_{jk}}$. It is customary that not all empirical coefficients be retained in the final estimator, and that small coefficients should be discarded to reduce the risk. It is also well-known (Cai, 2008) that individual coefficient thresholding is sub-optimal with respect to the L^2 loss, as opposed to block-thresholding procedures with carefully chosen blocks (Cai, 1999; Chicken and Cai, 2005). Here, we build the blocks as follows.

Motivated by (Cai, 1999; Chicken and Cai, 2005) we wish to build blocks of consecutive wavelets with size approximately $\log(n)$, which is known to be the best compromise for global versus local adaptation. Since there may be fewer than $\log(n)$ wavelets at small resolution levels j, we will only threshold coefficients with j large enough. We define

$$J_n := \inf\{j \geq J \,:\, 2^j \geq \log(n)\}$$

where the infimum is over the integers. We then let $N := 2^{J_n}$ so that each level with $j \geq J_n$ can be partitioned into an integer number of blocks of N consecutive wavelets. More precisely, for each level $j \geq J_n$, and each $\ell = 0, \ldots, N^{-1}2^j - 1$ we define the blocks of indices

$$\mathfrak{B}_{j\ell} := \{k \in \{0, \ldots, 2^{j-1}\} \,:\, (\ell - 1)N \leq k \leq \ell N - 1\}. \tag{17}$$

For a constant $\tau \geq 1$ we also define $\tilde{\jmath}_n$ as the largest integer such that $2^{\tilde{\jmath}_n} \leq \frac{n}{\log(n)\tau^2}$; we shall assume that $J < J_n < \tilde{\jmath}_n$ which is always satisfied for n large enough. We then let, for $i = 0, 1$,

$$\hat{f}_i \coloneqq \sum_{k=0}^{2^J-1} \hat{f}_i^{\Phi_{Jk}} \Phi_{Jk} + \sum_{j=J}^{J_n-1} \sum_{k=0}^{2^j-1} \hat{f}_i^{\Psi_{jk}} \Psi_{jk} + \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell} \Bigg(\sum_{k \in \mathfrak{B}_{j\ell}} \hat{f}_i^{\Psi_{jk}} \Psi_{jk} \Bigg) \mathbf{1}_{\{\|\hat{f}_i^{\mathfrak{B}_{j\ell}}\| > \Gamma \hat{S}_n\}}$$

where $\|\hat{f}_i^{\mathfrak{B}_{j\ell}}\|^2 \coloneqq \sum_{k \in \mathfrak{B}_{j\ell}} (\hat{f}_i^{\Psi_{jk}})^2$, $\Gamma > 0$ is a tuning parameter, and

$$\hat{S}_n \coloneqq \sqrt{\frac{\log(n)}{n}} \max\Big(1, \frac{\hat{g}}{|\hat{m}_1|}\Big) \mathbf{1}_{\{\hat{m}_1 \neq 0\}}.$$

The above estimators perform well in probability; to ensure good perfomance in expectation we truncate below at 0 and above at some $\check{T}$, defining for $i = 0, 1$

$$\check{f}_i \coloneqq \max\big(0, \min\big(\check{T}, \hat{f}_i\big)\big).$$

Theorem 6 *Assume τ and M are chosen as prescribed in Theorem 4. Suppose $n\gamma^* \geq \max(\tau^3, \frac{\tau^2 \log(n)^2}{L})$, $\tilde{\jmath}_n > J_n$, $L \leq n$, $\check{T} \geq L$, and $\zeta \leq 1$. Then there are universal constants $\beta > 0$, $B > 0$ and $C > 0$ such that for all $\Gamma \geq \beta L^{1/2} \max((L/\gamma^*)^{1/2}, 1/\gamma^*)$ and for $i = 0, 1$, provided $s_* \leq s_i \leq S$ with $S > 0$ the regularity of the wavelet basis,*

$$\begin{aligned}
\sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta \min_{i'=0,1} \Big(\|\check{f}_{i'} - f_i\|_{L^2}^2 \Big) &\leq B\check{T}^2 24^{2^M} \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^2\zeta^4}{L^3 + 2^M\sqrt{L}\delta\epsilon\zeta^2} \right) \\
&+ B\check{T}^2 \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^4\zeta^6}{L^3 + \max(\tau, \sqrt{L})^3\delta\epsilon^2\zeta^3} \right) + \frac{BL^2}{\delta^2\epsilon^2\zeta^2} \frac{\log(n)}{n\gamma^*} + \frac{BL^3}{\delta^2\epsilon^4\zeta^4} \frac{1}{n\gamma^*} \\
&+ \frac{B \max(\tau, \sqrt{L})^6}{\delta^2\epsilon^4\zeta^4} \frac{1}{(n\gamma^*)^2} + \frac{BR^2 \max(1, \frac{L^2}{\Gamma^2\gamma^*})}{\min(1, s_i)} \Big(\frac{\Gamma^2}{R^2\delta^2\epsilon^2\zeta^2 n} \Big)^{2s_i/(2s_i+1)} \\
&+ \frac{BR^2 \max(1, \frac{L^2}{\Gamma^2\gamma^*})}{\min(1, s_i)} \Big(\frac{\tau^2 \log(n)}{n} \Big)^{2s_i}.
\end{aligned}$$

The proof of Theorem 6 is in Section C.5. Of particular interest is the boundary regime, where γ^*, R, L, $\check{T}$ and τ are of constant order while δ, γ and ζ are small. The following corollary is intended to illustrate how the bound simplifies in such settings, provided δ, γ and ζ are not too small. The proof of Corollary 7 is given in Section C.8.

Corollary 7 *Assume that γ^*, R, L, $\check{T}$, and τ remain constant as $n \to \infty$ and $\delta \geq n^{-a}$, $\epsilon \geq n^{-b}$, $1 \geq \zeta \geq n^{-c}$ for constants $a, b, c > 0$ such that $1 - 2a - 4b - 6c > 0$ and such that $2^M = o(n^{(1-a-b-2c)/2})$ (the penultimate requirement corresponds to where the bounds on the right vanish, so that parameters are proved to be learnable). Then the bound in the Theorem 6 simplifies: for large enough n,*

$$\sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta \min_{i'=0,1} \Big(\|\check{f}_{i'} - f_i\|_{L^2}^2 \Big) \leq C \left\{ \frac{1}{\delta^2\epsilon^4\zeta^4 n} + \Big(\frac{1}{\delta^2\epsilon^2\zeta^2 n} \Big)^{2s_i/(1+2s_i)} \right\},$$

for a constant C depending on γ^*, L, R, Γ, B, τ, $\check{T}$, and a, b, c.

3.6 Nonparametric part: case $s_0 < s_1$

In the particular situation where $s_0 = s_1$, the lower bound (6) holds for the estimation of both emission densities, and the estimators $\check{f}_0$ and $\check{f}_1$ are rate minimax adaptive, including to the parameters of interest δ, ϵ, ζ. However, in the situation where $s_0 \neq s_1$, assuming without loss of generality assuming that $s_0 < s_1$, the estimator for the rougher density f_0 is not rate optimal in term of δ, ϵ, ζ. We fill the gap by constructing another estimator for f_0 that attains the optimal rate. The construction of the estimator exploits the "borrowing strength" phenomenon described in the introduction, which we now make more formal. We focus only on estimating f_0 when $s_0 < s_1$; the estimation of f_1 when $s_0 > s_1$ is similar.

The starting point is to remark that

$$f_0 = \frac{2\psi_1}{1+\phi_1} - \left(\frac{1-\phi_1}{1+\phi_1}\psi_1 - \frac{g(1-\phi_1)}{2m_1} G \right) \tag{18}$$

with $G = r(\phi)\tilde{\mathcal{I}}\psi_2$, whose wavelet coefficients can be estimated using $\{\hat{G}^{\Phi_{Jk}}, \hat{G}^{\Psi_{jk}}\}$. Note that $2\psi_1/(1+\phi_1) = \pi_0^{-1}\psi_1$ and the other term involved in (18) equals $(1-\pi_0)\pi_0^{-1} f_1$. We recall the rationale of the borrowing strength phenomenon: ψ_1 is "easy" to estimate (estimating it is a direct problem, not an inverse problem) since it is the stationary distribution of Y_n; also f_1, being smoother than f_0, can be estimated at a better rate. We estimate the father wavelet coefficients of f_0 using the same estimators as before. Regarding the mother coefficients, however, we let $\alpha_0 \coloneqq \pi_0^{-1}\psi_1$ and $\beta_0 \coloneqq f_0 - \alpha_0$ and we estimate separately the coefficients of these two functions using

$$\hat{\alpha}_0^{\Psi_{jk}} \coloneqq \frac{2\hat{\psi}_1^{\Psi_{jk}}}{1+\hat{\phi}_1} \mathbf{1}_{\{\hat{\phi}_1 \neq -1\}}, \quad \hat{\beta}_0^{\Psi_{jk}} \coloneqq -\left(\frac{1-\hat{\phi}_1}{1+\hat{\phi}_1} \mathbf{1}_{\{\hat{\phi}_1 \neq -1\}} \hat{\psi}_1^{\Psi_{jk}} - \frac{\hat{g}(1-\hat{\phi}_1)}{2\hat{m}_1} \mathbf{1}_{\{\hat{m}_1 \neq 0\}} \hat{G}^{\Psi_{jk}} \right).$$

Then, what we shall call the 'rough estimator' (since it only usefully estimates the rougher of the two functions f_0, f_1) is defined as:

$$\begin{aligned} \hat{f}_0^R \coloneqq \sum_{k=0}^{2^{J_n}-1} \hat{f}_0^{\Phi_{Jk}} \Phi_{J_n k} + \sum_{j=J}^{J_n-1} \sum_{k=0}^{2^j-1} \hat{f}_0^{\Psi_{jk}} \Psi_{jk} \\ + \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell=0}^{2^j/N-1} \Big(\sum_{k \in \mathfrak{B}_{j\ell}} \hat{\alpha}_0^{\Psi_{jk}} \Psi_{jk} \Big) \mathbf{1}_{\{\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\| > \Gamma\sqrt{\log(n)/n}\}} \\ + \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell=0}^{2^j/N-1} \Big(\sum_{k \in \mathfrak{B}_{j\ell}} \hat{\beta}_0^{\Psi_{jk}} \Psi_{jk} \Big) \mathbf{1}_{\{\|\hat{\beta}_0^{\mathfrak{B}_{j\ell}}\| > \Gamma \hat{T}_n\}}, \end{aligned} \tag{19}$$

with $\hat{f}_0^{\Phi_{Jk}}$ and $\hat{f}_0^{\Psi_{jk}}$ as previously and

$$\hat{T}_n \coloneqq \sqrt{\frac{\log(n)}{n}} \max\Big(1, \frac{\hat{g}}{|\hat{m}_1|} \mathbf{1}_{\hat{m}_1 \neq 0}, \frac{1}{1-\hat{\phi}_1^2} \mathbf{1}_{\hat{\phi}_1^2 \neq 1}\Big).$$

Note that in (19), thresholding of the estimated coefficients of ψ_1 is done "as usual" for density estimation, whereas thresholding of the $\hat{\beta}_0^{\Psi_{jk}}$'s is done with another carefully chosen threshold.

As previously, we also further require a truncation of the estimator to obtain control in expectation not just in probability, and for some $\check{T} > 0$ we define

$$\check{f}_0^R \coloneqq \max\big(0,\, \min\big(\check{T},\, \hat{f}_0^R\big)\big).$$

The following theorem gives an upper bound on the maximum risk of $\hat{f}_0^R$. The proof of Theorem 8 is detailed in Section C.6.

Theorem 8 *Assume τ and M are chosen as prescribed in Theorem 4. Suppose $n\gamma^* \geq \max(\tau^3, \frac{\tau^2 \log(n)^2}{L})$, $\tilde{\jmath}_n > J_n$, $L \leq n$, $\check{T} \geq L$, $\zeta \leq 1$, and $s_* < s_0 \leq S$, with $S > 0$ the regularity of the wavelet basis. Then there are universal constants $\beta > 0$, $B > 0$ and $C > 0$ such that for all $\Gamma \geq \beta \max(\frac{L}{\sqrt{\gamma^*}}, \frac{\sqrt{L}}{\tau\gamma^*})$*

$$\begin{aligned}
\sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\Big(\|\check{f}_0^R - f_0\|_{L^2}^2\Big) &\leq B\check{T}^2 24^{2^M} \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^2\zeta^4}{L^3 + 2^M\sqrt{L}\delta\epsilon\zeta^2}\right) \\
&+ B\check{T}^2 \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^4\zeta^6}{L^3 + \max(\tau,\sqrt{L})^3\delta\epsilon^2\zeta^3}\right) + \frac{BL^2}{\delta^2\epsilon^2\zeta^2}\frac{\log(n)}{n\gamma^*} + \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} \\
&+ \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2} + \frac{R^2}{\min(1,s_0)}\Big(\frac{\Gamma^2}{nR^2\delta^2}\Big)^{2s_0/(2s_0+1)} \\
&+ \frac{R^2}{\min(1,s_1)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_1/(2s_1+1)} + \frac{BR^2}{\min(1,s_0)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_0}.
\end{aligned}$$

As with Theorem 6 and its Corollary 7, of particular interest is the boundary regime, where γ^*, R, L, $\check{T}$ and τ are of constant order while δ, γ and ζ are small, but not too small. The following corollary is intended to illustrate how the bound simplifies in such setting. The proof of Corollary 9 is given in Section C.9.

Corollary 9 *Assume that γ^*, R, L, $\check{T}$, and τ remain constant as $n \to \infty$ and $\delta \geq n^{-a}$, $\epsilon \geq n^{-b}$, $1 \geq \zeta \geq n^{-c}$ for constants $a, b, c > 0$ with $a, b, c = o(1)$ and $2^M = o(n^{(1-a-b-2c)/2})$ as $n \to \infty$. Then if $s_1 < s_0$ the bound in the Theorem 8 simplifies: for large enough n,*

$$\sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\Big(\|\check{f}_0^R - f_0\|_{L^2}^2\Big) \leq C\left\{\frac{1}{\delta^2\epsilon^4\zeta^4 n} + \left(\frac{1}{\delta^2 n}\right)^{2s_0/(2s_0+1)}\right\},$$

for a constant C depending on γ^*, L, R, Γ, B, τ, $\check{T}$.

In the regime of Corollary 9, *ie.* when δ, ϵ, ζ are small but not too small, the estimator $\check{f}_0$ achieves the lower bound established in Theorem 3. In settings where δ, ϵ, ζ are allowed to be smaller than a polynomial in n, a transition in the rate still occurs according to how s_0 and s_1 compare, but then it may be required to have s_1 much larger than s_0 (depending on δ, ϵ, ζ) to get matching upper and lower bounds.

We conclude this section by mentioning that the "borrowing strength phenomenon" is not specific to the case where the f_j's belong to different Besov bodies $\{f : \|f\|_{B^{s_j}_{2,\infty}} \leq R\}$

with $s_0 \neq s_1$. Indeed, the same phenomenon should occur as long as the f_j's belong to classes $\mathcal{S}_j$ of different "complexities" (which can for instance be measured by the number of balls of finite radius needed to cover $\mathcal{S}_j$); or in other words, as soon as nonparametric estimation over $\mathcal{S}_1$ is easier than over $\mathcal{S}_0$ (or conversely). Thus, the phenomenon would take place if the Besov bodies are replaced by other types of smoothness classes (for instance Hölder balls of finite radius).

3.7 Summary of the algorithm

In this section we present our algorithm in full, self-contained manner, and discuss its computational complexity. To simplify the exposition of the algorithm, let us recall or introduce some notations.

We use the S-regular boundary corrected wavelet basis $\{\{\Phi_{Jk} : k = 0, \dots, 2^{J-1}\}, \{\Psi_{jk} : j \geq J,\ k = 0, \dots, 2^{j-1}\}\}$ constructed in Cohen et al. (1993). We use the notation $\Lambda(m) = \{0, \dots, 2^{J-1}\} \cup \{(j,k) : j = J, \dots, m,\ k = 0, \dots, 2^{j-1}\}$ for $m \geq J$. We also write $e_\lambda = \Phi_{J\lambda}$ if $\lambda \in \{0, \dots, 2^{J-1}\}$ or $e_\lambda = \Psi_{jk}$ if $\lambda = (j,k)$. We also define for real-valued function f and reals $a < b$ the clipping operation $\mathrm{clip}(f, [a,b])$ defined such that $\mathrm{clip}(f,[a,b])(x) = \max(a, \min(f(x), b))$.

Our complete estimation procedure is given in the Algorithm 2. The Algorithm 2 computes the estimator of Q defined in Section 3.4 and the estimators of f_0 and f_1 defined in Section 3.5, where they are proven to be minimax optimal in the case where $s_0 = s_1$. In the case where $s_0 \neq s_1$ and information is available to identify the smoothest emission density, the previous algorithm can be complemented by an additional step to improve the estimator of the roughest density, corresponding to the estimator derived in Section 3.6. We summarize this additional step in the Algorithm 3, assuming without loss of generality that $s_0 < s_1$.

We now discuss the computational complexity of our algorithm. As for the minimax rates, our interest is about the complexity of the algorithm as function of n, δ, ϵ, and ζ. We do assume that pointwise evaluation of wavelets can be done in time $O(1)$. The complexity of step 2 of Algorithm 2 is dominated by the computation of the leading eigenvector of a $2^M \times 2^M$ matrix, which can be done in $O(2^{3M})$ time. The Theorem 4, however, prescribes that 2^M must be at least $\left(\frac{4R}{\zeta\sqrt{2^{2s_*}-1}}\right)^{1/s_*}$, so step 2 of Algorithm 2 is feasible in time $O(\zeta^{-3/s_*})$. The most demanding computation in step 3 of Algorithm 1 is to evaluate $\tilde{\psi}_2(Y_i)$ for all $i = 1, \dots, n$. Since the wavelets are compactly supported, evaluating $\tilde{\psi}_2(Y_i)$ requires only summing $O(M)$ terms, and hence the step 3 can be achieved in time $O(n \cdot M) = O(n \log(1/\zeta))$. In the step 4 of Algorithm 2, we do not need to reevaluate $(\tilde{\psi}_2(Y_i))_{i=1}^n$ since we can keep it in memory from the previous step. Exploiting the compactness of the support of the wavelets, we can compute $(\tilde{\psi}_1^\lambda, \tilde{G}^\lambda)_{\lambda \in \Lambda(\tilde{\jmath}_n}$ in time $O(n \cdot \tilde{\jmath}_n) = O(n \log(n/\tau^2)) = O(n \log(n/\zeta^2))$, again by Theorem 4. The thresholding of the coefficients can be trivially performed in time $O(2^{\tilde{\jmath}_n}) = O(\frac{n}{\log(n)\tau^2}) = O(\frac{n}{\log(n)\zeta^2})$ since there are $2^{\tilde{\jmath}_n}$ coefficients. Gathering all these estimates, it is seen that Algorithm 2 runs in time $O(\max(\zeta^{-3/s_*}, \frac{n}{\log(n)\zeta^2}, n \log(n/\zeta^2))$, which is typically dominated by $n \log(n)$. Furthermore, it is easily seen that running Algorithm 3 does not increase the computational complexity of the overall algorithm.

Algorithm 2 Full algorithm

Require: Data $(Y_1, \dots, Y_{3n})$ and hyperparameters $M \in \{J, J+1, \dots\}$, $\tau > 0$, $\Gamma > 0$, $\check{T} > 0$.
Ensure: Estimators $\hat{Q}$, $\check{f}_0$, and $\check{f}_1$.

Step 1: Sample splitting

Let $(\tilde{Y}_1, \dots, \tilde{Y}_n) = (Y_{2n+1}, \dots, Y_{3n})$.

Step 2: Estimation of the separating hyperplane

Compute the $2^M \times 2^M$ matrix $\tilde{\mathcal{G}}$ with entries $\tilde{\mathcal{G}}_{\lambda,\lambda'} = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (e_\lambda(\tilde{Y}_i) e_{\lambda'}(\tilde{Y}_{i+1}) + e_{\lambda'}(\tilde{Y}_i) e_\lambda(\tilde{Y}_{i+1})) - \frac{1}{n} \sum_{i=1}^n e_\lambda(\tilde{Y}_i) \cdot \frac{1}{n} \sum_{i=1}^n e_{\lambda'}(\tilde{Y}_i)$ for every $\lambda, \lambda' \in \Lambda(M)$.

Compute leading eigenvector v of $\tilde{\mathcal{G}}$.

Let $\tilde{\psi}_2 \propto \mathrm{clip}(\sum_{\lambda \in \Lambda(M)} v_\lambda e_\lambda, [-\tau, \tau])$ with $\|\tilde{\psi}_2\| = 1$.

Step 3: Estimation of the transition matrix Q

Compute

$$\hat{m}_1 = \frac{1}{n-1} \sum_{i=1}^{n-1} \tilde{\psi}_2(Y_i)\tilde{\psi}_2(Y_{i+1}) - \left(\frac{1}{n} \sum_{i=1}^n \tilde{\psi}_2(Y_i)\right)^2,$$

$$\hat{m}_2 = \frac{1}{n-2} \sum_{i=1}^{n-1} \tilde{\psi}_2(Y_i)\tilde{\psi}_2(Y_{i+2}) - \left(\frac{1}{n} \sum_{i=1}^n \tilde{\psi}_2(Y_i)\right)^2,$$

$$\hat{m}_3 = \frac{1}{n-2} \sum_{i=1}^{n-2} \tilde{\psi}_2(Y_i)\tilde{\psi}_2(Y_{i+1})\tilde{\psi}_2(Y_{i+2}) + \left(\frac{1}{n} \sum_{i=1}^n \tilde{\psi}_2(Y_i)\right)^3 + (2\hat{m}_1 + \hat{m}_2)\frac{1}{n} \sum_{i=1}^n \tilde{\psi}_2(Y_i).$$

Compute $\hat{\phi}_1 = \frac{\hat{m}_3}{[4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2]^{1/2}}$ and $\hat{\phi}_2 = \max\left(-1, \min\left(\frac{\hat{m}_2}{\hat{m}_1}, 1\right)\right)$.

Let $\hat{p} = \frac{1}{2}(1 - \hat{\phi}_1)(1 - \hat{\phi}_2)$, $\hat{q} = \frac{1}{2}(1 + \hat{\phi}_1)(1 - \hat{\phi}_2)$, and $\hat{Q} = \begin{pmatrix} 1-\hat{p} & \hat{p} \\ \hat{q} & 1-\hat{q} \end{pmatrix}$.

Step 4: Estimation of the emission densities

Compute $\hat{g} = \frac{\sqrt{4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2}}{\hat{m}_2} \mathbf{1}_{\{\hat{m}_2 > 0\}}$.

Let $\tilde{\jmath}_n = \left\lfloor \log_2\left(\frac{n}{\log(n)\tau^2}\right) \right\rfloor$, $J_n = \lceil \log_2(n) \rceil$, $\hat{S}_n = \sqrt{\frac{\log(n)}{n}} \max\left(1, \frac{\hat{g}}{|\hat{m}_1|}\right) \mathbf{1}_{\{\hat{m}_1 \neq 0\}}$.

For all $\lambda \in \Lambda(\tilde{\jmath}_n)$, compute the empirical wavelet coefficients $\hat{\psi}_1^\lambda = \frac{1}{n} \sum_{i=1}^n e_\lambda(Y_i)$ and $\hat{G}^\lambda = \frac{1}{n-1} \sum_{i=1}^{n-1} \tilde{\psi}_2(Y_i) e_\lambda(Y_{i+1}) - \frac{1}{n} \sum_{i=1}^n \tilde{\psi}_2(Y_i) \cdot \frac{1}{n} \sum_{i=1}^n e_\lambda(Y_i)$.

for m=0,1 **do**

  Compute $\hat{f}_m^\lambda = \hat{\psi}_1^\lambda + (-1)^m \frac{\hat{g}(1+(-1)^{m+1}\hat{\phi}_1)}{2\hat{m}_1} \mathbf{1}_{\{\hat{m}_1 \neq 0\}} \hat{G}^\lambda$ for all $\lambda \in \Lambda(\tilde{\jmath}_n)$.

  Set $\hat{f}_m^{(j,k)} = 0$ for all coefficients (j,k) belonging to blocks $\mathfrak{B}_{j\ell} = \{k \in \{0, \dots, 2^{j-1}\} : (\ell-1)2^{J_n} \leq k \leq \ell 2^{J_n} - 1\}$ such that $\sum_{k \in \mathfrak{B}_{j\ell}} [\hat{f}_m^{(j,k)}]^2 \leq \Gamma^2 \hat{S}_n^2$ and $j \geq J_n$.

  Let $\check{f}_m = \mathrm{clip}(\sum_{\lambda \in \Lambda(\tilde{\jmath}_n)} \hat{f}_m^\lambda e_\lambda, [0, \check{T}])$.

end for

Our algorithm is thus simple and computationally efficient, avoiding any non-convex optimization step. It thus provides a promising alternative to existing methods. Further practical implementation may require additional work on tuning the hyperparameters, which is beyond the scope of this paper and a consideration for future research.

Algorithm 3 Improved estimator of f_0 when $s_0 < s_1$

Require: $\hat{g}$, $\tilde{j}_n$, J_n, $(\hat{\psi}_1^\lambda)_{\lambda\in\Lambda(\tilde{j}_n)}$, $(\hat{G}^\lambda)_{\lambda\in\Lambda(\tilde{j}_n)}$, $(\hat{f}_0^\lambda)_{\lambda\in\Lambda(J_n)}$ as obtained in Step 4 of Algorithm 2, $\check{T} > 0$.

Ensure: Estimator $\hat{f}_0^R$

1: Let $\hat{T}_n = \sqrt{\frac{\log(n)}{n}} \max\left(1,\ \frac{\hat{g}}{|\hat{m}_1|}\mathbf{1}_{\{\hat{m}_1\neq 0\}},\ \frac{1}{1-\hat{\phi}_1^2}\mathbf{1}_{\{\hat{\phi}_1^2\neq 1\}}\right)$

2: Compute $\hat{\alpha}_0^\lambda = \frac{2\hat{\psi}_1^\lambda}{1+\hat{\phi}_1}\mathbf{1}_{\{\hat{\phi}_1\neq -1\}}$ and $\hat{\beta}_0^\lambda = -\left(\frac{1-\hat{\phi}_1}{1+\hat{\phi}_1}\mathbf{1}_{\{\hat{\phi}_1\neq -1\}}\hat{\psi}_1^\lambda - \frac{\hat{g}(1-\hat{\phi}_1)}{2\hat{m}_1}\mathbf{1}_{\{\hat{m}_1\neq 0\}}\hat{G}^\lambda\right)$ for all $\lambda \in \Lambda(\tilde{j}_n)\backslash\Lambda(J_n)$.

3: Set $\hat{\alpha}_0^{(j,k)} = 0$ for all coefficients (j,k) belonging to blocks $\mathfrak{B}_{j\ell} = \{k \in \{0,\dots,2^{j-1}\} : (\ell-1)2^{J_n} \le k \le \ell 2^{J_n} - 1\}$ such that $\sum_{k\in\mathfrak{B}_{j\ell}}[\hat{f}_m^{(j,k)}]^2 \le \Gamma^2 \log(n)/n$ and $j \ge J_n$.

4: Set $\hat{\beta}_0^{(j,k)} = 0$ for all coefficients (j,k) belonging to blocks $\mathfrak{B}_{j\ell} = \{k \in \{0,\dots,2^{j-1}\} : (\ell-1)2^{J_n} \le k \le \ell 2^{J_n} - 1\}$ such that $\sum_{k\in\mathfrak{B}_{j\ell}}[\hat{f}_m^{(j,k)}]^2 \le \Gamma^2 \hat{T}_n^2$ and $j \ge J_n$.

5: Let $\hat{f}_0^R = \mathrm{clip}(\sum_{\lambda\in\Lambda(J_n)} \hat{f}_0^\lambda e_\lambda + \sum_{\lambda\in\Lambda(\tilde{j}_n)\backslash\Lambda(J_n)}(\hat{\alpha}_0^\lambda + \hat{\beta}_0^\lambda)e_\lambda, [0,\check{T}])$.

3.8 Comparison with the case of discrete emissions

To the best of our knowledge, the paper Abraham et al. (2022b) is the only work that has considered the explicit dependence of the distance to the i.i.d frontier in the minimax rates of estimating HMM. In Abraham et al. (2022b) we considered only the case of emissions on $\{1,\dots,K\}$ for known $K \ge 2$. The present work considers the more interesting (for applications) case of continuous emission densities. Although the results of both papers share some similarities, there are some aspects that are crucially different. The major difference between the discrete case and the present paper resides in the necessity of estimating the separating hyperplane described in Section 3.1. This step of the estimation procedure isn't needed for the discrete case, and was overlooked in the previous literature on nonparametric HMMs.

We note that the parametric part $\hat{Q}$ achieves the same rate in the nonparametric setting as in the multinomial setting (first inequality in Theorem 1); at first glance this seems unsurprising in view of the fact that the pairs $((X_n, h(Y_n))_{n\ge 0}$ form a hidden Markov model with transition matrix Q for any function h, so that for a suitable h we can reduce to a parametric setting. This is the *no bias* phenomenon already used in (Gassiat et al., 2018) for multidimensional mixture models and in (Moss and Rousseau, 2024) for finite state space HMMs. Choosing $A_1,\dots,A_K$ partitioning $[0,1]$ and defining h by $h(y) = k$ for $y \in A_k$, we may apply the results from the discrete setting to deduce that Q can be estimated at the parametric rate given in (Abraham et al., 2022b). However in said rate ζ must lower bound the euclidean distance between vectors $(\langle f_0, \mathbb{1}_{A_k}\rangle : k \le K)$ and $(\langle f_1, \mathbb{1}_{A_k}\rangle : k \le K)$. If the A_k are not chosen carefully, this distance may be much smaller than $\|f_0 - f_1\|_{L^2}$, potentially even equal to 0. A suitable choice of $(A_k)_{k=1}^K$ depends on the direction $(f_0 - f_1)/\|f_0 - f_1\|_{L^2} = \psi_2$, which is unknown and *nonparametric*. This is tantamount to estimating the separating hyperplane.

Similarly, the no bias phenomenon could be exploited to build histogram estimators of f_0 and f_1 and thereby reducing the continuous case to the discrete case. Doing so, it is tempting to think that the minimax rates for the continuous case can be deduced from

the results in Abraham et al. (2022b). Unfortunately, in Abraham et al. (2022b) we did not explicit the dependence of the rates in the number of bins K, which do not enable for immediate obtentation of the rates for f_0 and f_1 since in the continuous case the number of bins must be a function of number of observations to ensure the adequate bias-variance tradeoff. Furthermore, the approach considered in this paper offers several advantages compared to the histogram approach: (i) histograms permit optimal estimation only in a very limited range of smoothness, *ie.* $s_0, s_1 \in (0, 2]$, compared to $(0, S]$ in this paper (where S can be made large by choosing the suitable wavelet basis); (ii) making histogram estimators that are adaptive to smoothness requires some form of model selection to choose the optimal number of bins, which is avoided in this paper using thresholding; and (iii) the estimator in Abraham et al. (2022b) is a minimum distance estimator that requires solving a tricky non-convex optimization problem, while in the moment based estimator in the current paper is computable in almost linear time (see Section 3.7).

Finally, the continuous cases offers some curiosities in comparison with the discrete case. First, the minimax rate for estimating f_0 and f_1 in Abraham et al. (2022b) was found to be of order $(\delta^2\epsilon^4\zeta^4 n)^{-1}$. In the continuous case, although the minimax rate is also bounded by a term of order $(\delta^2\epsilon^4\zeta^4 n)^{-1}$, in most regimes of interest[1] the dominating term in the rate is of order $(\delta^2\epsilon^2\zeta^2 n)^{-2s_i/(2s_i+1)}$ for the smoothest density (see Corollary 7) or $(\delta^2 n)^{-2s_i/(2s_i+1)}$ (see Corollary 9). Thus, the constants δ, ϵ, ζ appear with different powers in the dominating term, which is a curiosity for which we do not have a clear intuition. Second, the "borrowing estimation strength" phenomenon described in Section 3.6 came as a big surprise to us when writing this paper. We uncovered this phenomenon when trying to match the minimax upper and lower bounds, realizing that given one of the two densities, the other can be estimated in two ways, leading to different rates. We could't have guessed this phenomenon from our previous work (Abraham et al., 2022b) since its appear only in situations where f_0 and f_1 have different "complexities" – here measured by smoothness s_0, s_1, in Abraham et al. (2022b) measured by K – which we didn't considered earlier.

4. Conclusion and open questions

In this paper, we obtain precise behaviour of the minimax risk of all parameters in a nonparametric hidden Markov models, with exact constants regarding the distance to the i.i.d. frontier where the parameters become non-identifiable (we were not interested in the exact dependence of the constants with respect to L, R and γ^*). In particular, we prove a surprising transition in the minimax rates depending on relative smoothnesses of the emission densities.

Similarly to wavelet density estimation with i.i.d. data, the parameter Γ used in the optimal threshold must be chosen depending on the upper L for the supremum norms of f_0, f_1. In the i.i.d. case a simple workaround to adapt to L is to obtain a consistent estimator of the density in L^∞ norm, see (Giné and Nickl, 2016, Exercise 8.2.1), and plug into the threshold. In the HMM situation, it is not obvious how to obtain an asymptotically valid value for L empirically. Our optimal threshold also depends on γ^*, which requires the preliminary step of the separation hyperplane estimation, itself requiring L. For the estimation of the separating hyperplane, we assume lower bounds on $\min\{s_0, s_1\}$ and on ζ.

1. *ie.* δ, ϵ, ζ small but not too small, as in Corollaries 7 and 9.

If neither L nor γ^* is known, the interconnectedness of the parametric and nonparametric part causes us difficulty in fully adapting.

The main open question concerns full adaptation to get the right constants in the upper bound when a transition occurs due to different smoothnesses. From results herein one deduces the existence of pairs of estimators $(\check{f}_0, \check{f}_1)$, $(\check{f}_0^R, \check{f}_1)$, $(\check{f}_0, \check{f}_1^R)$, $(\check{f}_0^R, \check{f}_1^R)$ of which one pair is minimax optimal. When it is known which pair to use, we indeed get minimax optimal estimators. The question of the possibility or impossibility of choosing the correct pair without oracle guidance is of distinguished interest, yet challenging. It will be the subject of a future work.

Finally, we remark that we only investigated the minimax rates over Besov $B_{2,\infty}^{s_j}$ bodies. But our results can easily be extended to $B_{2,q_j}^{s_j}$ for any $1 \leq q_j \leq \infty$. Indeed, it is trivial that $\|\cdot\|_{B_{2,\infty}^s} \leq \|\cdot\|_{B_{2,q}^s}$ for all $s > 0$ and all $1 \leq q \leq \infty$, from which it is deduced that $B_{2,\infty}^{s_j}$ balls are larger than $B_{2,q_j}^{s_j}$ balls, hence all our upper bounds remain valid if $B_{2,\infty}^{s_j}$ is replaced by $B_{2,q_j}^{s_j}$. On the other direction, we prove the lower bounds using a classical reduction to a multiple hypotheses testing problem, and it can be seen in our proofs (see for instance Section B.2) that the hypotheses we choose all belong to $\{(f_0, f_1) \, : \, \max_{i=0,1} \|f_i\|_{B_{2,1}^{s_i}} \leq R\}$. Hence our minimax lower bounds indeed hold over $B_{2,1}^{s_j}$ bodies, and thus extend trivially to $B_{2,q_j}^{s_j}$ bodies for any $1 \leq q_j \leq \infty$, by the same embedding argument as before. A natural direction for the next would be to investigate the rates over $B_{p,q}^s$ bodies $1 \leq p, q \leq \infty$, $s > 0$, with loss measured in L_r norm for $1 \leq r \leq \infty$, as it is classical in nonparametric estimation (see for instance the seminal paper of Donoho et al. (1996)). In this situation, we expect that the rates will exhibit the same "elbow" uncovered by Donoho et al. (1996), but it would be interesting to figure out the interplay between $(\delta, \epsilon, \zeta)$ and (p, q, s, r), which is beyond the scope of the present paper.

Acknowledgments

Kweku Abraham is supported bythe EPSRC Programme Grant on the Mathematics of Deep Learning, under the project: EP/V026259/1. Élisabeth Gassiat is supported by Institut Universitaire de France. Élisabeth Gassiat and Zacharie Naulet are support by the ANR under projects ANR-21-CE23-0035-02 and ANR-23-CE40-0018-02.

Appendix A. About the assumption of two independent samples

We assumed in the paper that we first get $\tilde{\psi}_2$ based on an independent sample of the HMM. Suppose we are given a single stationary HMM of length $3n$ with distribution $\mathbb{P}_\theta$ such that the hidden Markov chain has absolute spectral gap γ^*. Let $Y' = (Y_1, \dots, Y_n)$, $\tilde{Y}' = (Y_{2n+1}, \dots, Y_{3n})$, and denote $\mathbb{P}_{(Y',\tilde{Y}')}$ the distribution of $(Y', \tilde{Y}')$. Denote also $\mathbb{P}_{Y'}$ the distribution of Y' (which is the same as the distribution of $\tilde{Y}'$ by stationarity). For $j = 1, \dots, 4$ let $\hat{\theta}_j$ denote our estimator of θ_j. Notice that $\hat{\theta}_j$ (resp. θ_j) is non-negative and bounded by 2 (resp. 1) for $j = 1, 2$ and $\check{T}$ (resp. L) for $j = 3, 4$, so that, denoting M (resp. $\tilde{M}$) the upper bound, we have $\|\hat{\theta}_j - \theta_j\| \le M \vee \tilde{M}$, $\|\cdot\|$ being the euclidean norm for $j = 1, 2$ and the $L^2[0,1]$-norm for $j = 3, 4$. Then,

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{(Y',\tilde{Y}')}} \left(\|\hat{\theta}_j - \theta_j\|^2 \right) \\
&= \int_0^{M \vee \tilde{M}} \mathbb{P}_{(Y',\tilde{Y}')} \left(\|\hat{\theta}_j - \theta_j\|^2 \ge t \right) dt \\
&= \mathbb{E}_{\mathbb{P}_{Y'}^{\otimes 2}} \left(\|\hat{\theta}_j - \theta_j\|^2 \right) + \int_0^{M \vee \tilde{M}} \left[\mathbb{P}_{(Y',\tilde{Y}')} \left(\|\hat{\theta}_j - \theta_j\|^2 \ge t \right) - \mathbb{P}_{Y'}^{\otimes 2} \left(\|\hat{\theta}_j - \theta_j\|^2 \ge t \right) \right] dt \\
&\le \mathbb{E}_{\mathbb{P}_{Y'}^{\otimes 2}} \left(\|\hat{\theta}_j - \theta_j\|^2 \right) + \left(M \vee \tilde{M} \right) \|\mathbb{P}_{(Y',\tilde{Y}')} - \mathbb{P}_{Y'}^{\otimes 2}\|_{\mathrm{TV}},
\end{aligned}
$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm. Using Proposition 10 below, we deduce that the first term on the right side of the last display dominates the second, hence the only cost of using one sample for the whole procedure is a multiplicative constant factor.

Proposition 10 *There exist universal constants C and c such that*

$$\|\mathbb{P}_{(Y',\tilde{Y}')} - \mathbb{P}_{Y'}^{\otimes 2}\|_{\mathrm{TV}} \le C e^{-c\gamma^* n}.$$

Proof Denote $Z_i = (X_i, Y_i)$, $i = 1, \dots, 3n$, where $(X_1, \cdots, X_n)$ is the hidden Markov chain. Using similar notations, we have

$$\|\mathbb{P}_{(Y',\tilde{Y}')} - \mathbb{P}_{Y'}^{\otimes 2}\|_{\mathrm{TV}} \le \|\mathbb{P}_{(Z',\tilde{Z}')} - \mathbb{P}_{Z'}^{\otimes 2}\|_{\mathrm{TV}}.$$

Now, for any $(x_1, \dots, x_n, x_{2n+1}, \dots, x_{3n})$, the distribution of $(Y_1, \dots, Y_n, Y_{2n+1}, \dots, Y_{3n})$ conditional on $(X_1, \dots, X_n, X_{2n+1}, \dots, X_{3n}) = (x_1, \dots, x_n, x_{2n+1}, \dots, x_{3n})$ is the same under $\mathbb{P}_{(Y',\tilde{Y}')}$ and $\mathbb{P}_{Y'}^{\otimes 2}$, so that

$$\|\mathbb{P}_{(Z',\tilde{Z}')} - \mathbb{P}_{Z'}^{\otimes 2}\|_{\mathrm{TV}} \le 2\|\mathbb{P}_{(X',\tilde{X}')} - \mathbb{P}_{X'}^{\otimes 2}\|_{\mathrm{TV}}$$

and the result follows from the uniform geometric ergodicity of the binary chain. ■

Appendix B. Proofs for the lower bounds

For proving our lower bounds, we shall follow the usual path, in which we need at some point upper bounds for distances between joint distributions $P_\theta^{(n)}$ for different values of θ.

We shall use the same trick as the one used in (Abraham et al., 2022b), that is an upper bound on the Kullback-Leibler divergence using a pseudo-distance ρ between parameters, see the end of Section III in (Abraham et al., 2022b) for heuristics explaining the importance of ρ interpreted as a fundamental statistical distance in HMM learning.

The following result is Proposition 2 in (Abraham et al., 2022b), for which a close look at the proof shows that it still holds with emission densities on $[0,1]$ instead of probability mass functions.

Proposition 11 *Assume there exists $c > 0$ such that $\min(f_0, f_1, \tilde{f}_0, \tilde{f}_1) \geq c$ uniformly on $[0,1]$. Then*

$$K(P_\theta^{(n)}, P_{\tilde{\theta}}^{(n)}) \leq Cn\rho(\phi(\theta), \psi(\theta); \phi(\tilde{\theta}), \psi(\tilde{\theta}))^2, \tag{20}$$

where, as in (Abraham et al., 2022b), we have defined

$$\begin{aligned}\rho(\phi, \psi; \tilde{\phi}, \tilde{\psi}) = \max\big\{ & |r(\phi) - r(\tilde{\phi})|, |\phi_2 r(\phi) - \tilde{\phi}_2 r(\tilde{\phi})|, \\ & |\phi_1\phi_2\phi_3 r(\phi) - \mathrm{sgn}\big(\langle \psi_2, \tilde{\psi}_2\rangle\big)\tilde{\phi}_1\tilde{\phi}_2\tilde{\phi}_3 r(\tilde{\phi})|, \\ & \|\psi_1 - \tilde{\psi}_1\|_{L^2}, \max(|r(\phi)|, |r(\tilde{\phi})|)\|\psi_2 - \mathrm{sgn}\big(\langle \psi_2, \tilde{\psi}_2\rangle\big)\tilde{\psi}_2\|_{L^2}\big\}. \end{aligned} \tag{21}$$

[Recall $r(\phi) = (1/4)(1-\phi_1^2)\phi_2\phi_3^2$.]

B.1 Proof of Theorem 2

To prove Theorem 2, we shall use a standard two-points argument using Le Cam's method (Le Cam (1986), see also Yu (1997) for a review of lower bound ideas): if θ and $\tilde{\theta}$ in $\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)$ are such that $|p - \tilde{p}|^2 \geq R_n$ and $K(P_\theta^{(n)}, P_{\tilde{\theta}}^{(n)}) \leq \alpha < 1$, then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta\Big(|\hat{p} - p|^2\Big) \geq \frac{R_n}{4}\left(1 - \sqrt{\alpha}\right).$$

We follow the method in the multinomial case (see Abraham et al., 2022b) used to choose the two points in proving Theorems 1 and 3 therein, except that rather than defining ψ according to Lemma 3 therein we choose $\psi_1 = 1$ and $\psi_2(x) = \sqrt{3}(2x-1)$. This choice of $\tilde{\psi} = \psi$ leads to lower bounded f_0 and f_1 (so that we can apply Proposition 11) when $\|f_0 - f_1\|_{L^2} = \zeta \leq 1/(4\sqrt{3})$, $\|f_i\|_\infty \leq 5/8$ and $\|f_i\|_{B_{2,\infty}^{s_i}} \leq 5/4 + 1/(8\sqrt{3})$, $i = 0, 1$, as a consequence of the inversion formulae (Lemma 12). Under the assumption that for a suitable $\epsilon_0 > 0$ we have $\zeta \leq 1/(4\sqrt{3})$, $\gamma^* \leq 1/3$, $\epsilon \leq \epsilon_0$, $\delta \leq 1/6$, the proof of the lower bounds for ϕ in Theorem 3 and the lower bound for p in Theorem 1 in (Abraham et al., 2022b) goes through to get the result. That is:
When $\delta > \epsilon\zeta$, we choose $\phi = (1-3\delta, \epsilon, \zeta(1+S)^{1/2}$ with $S = (2-6\delta-\sqrt{R_n})\sqrt{R_n}/(6\delta-9\delta)$ and $R_n = c/(n\epsilon^4\zeta^6)$, and we choose $\tilde{\phi} = (1-3\delta-\sqrt{R_n}, \epsilon, \zeta)$, so that $r(\phi) = r(\tilde{\phi})$, $\rho(\phi, \psi; \tilde{\phi}, \tilde{\psi}) \leq 6cn^{-1/2}$ and $|p - \tilde{p}|^2 \geq c/(n\epsilon^4\zeta^6)$.
When now $\delta \leq \epsilon\zeta$, we choose $\phi = (1 - 3\delta, \epsilon, \zeta(1 + \sqrt{R_n}/\epsilon)^{1/2}$ with $R_n = c/(n\epsilon^2\delta^2\zeta^4)$ and $\tilde{\phi} = (1 - 3\delta, \epsilon + \sqrt{R_n}, \zeta)$, so that again $r(\phi) = r(\tilde{\phi})$, $\rho(\phi, \psi; \tilde{\phi}, \tilde{\psi}) \leq cCn^{-1/2}$ for some constant C, and $|p - \tilde{p}|^2 \geq c/(n\epsilon^2\delta^2\zeta^4)$. The theorem follows by setting c small enough.

B.2 Proof of Theorem 3

For the parametric term in the lower bound, we are again able to copy the proof of (Abraham et al., 2022b) Theorems 1 and 3 up to the choice of ψ. Under the assumption that for a suitable $\epsilon_0 > 0$ we have $\zeta \leq 1/(4\sqrt{3})$, $\gamma^* \leq 1/3$, $\epsilon \leq \epsilon_0$, $\delta \leq 1/6$, as with proving Theorem 2 we choose $\psi_1 = 1$, $\psi_2(x) = \sqrt{3}(2x-1)$, $\tilde{\psi} = \psi$ and the proof of the lower bound for f_0 in (Abraham et al., 2022b, Theorem 1) goes through. That is we choose $\phi = (1-3\delta, \epsilon, \zeta(1+S)^{1/2}$ with $S = (2-6\delta-\sqrt{R_n})\sqrt{R_n}/(6\delta-9\delta)$ and $R_n = c/(n\epsilon^4\zeta^6)$, and we choose $\tilde{\phi} = (1-3\delta-\sqrt{R_n}, \epsilon, \zeta)$. Again $\rho(\phi, \psi; \tilde{\phi}, \tilde{\psi}) \leq 6cn^{-1/2}$ and now $\|f_0 - \tilde{f}_0\|_{L^2}^2 \geq c/(n\delta^2\epsilon^4\zeta^4)$.

We now prove the lower bound given in the second part of the theorem

$$R_{\text{smooth}} = (n\delta^2\epsilon^2\zeta^2)^{-s_0/(2s_0+1)}$$

We proceed via a usual reduction to multiple testing, see for instance (Tsybakov, 2009). For a suitable c, α, it suffices to construct function $f_{0,m} \in \{f : \|f\|_{B^{s_0}_{2,\infty}} \leq R\}$, $f_{1,m} \in \{f : \|f\|_{B^{s_1}_{2,\infty}} \leq R\}$, $0 \leq m \leq M = \lceil 2^{c2^j} \rceil$, for some j, such that

$$K\Big(P_m^{(n)}, P_0^{(n)}\Big) \leq c\alpha 2^j, \quad \|f_{0,m} - f_{0,m'}\|_{L^2} \geq cR_{\text{smooth}}, \tag{22}$$

where $P_m^{(n)}$ denotes the law of $(Y_1, \ldots, Y_n)$ under parameter $\theta_m = (p_m, q_m, f_{0,m}, f_{1,m})$ (for suitable choices of the parameters p_m, q_m). Indeed, given such functions, we note that

$$\frac{1}{M \log M} \sum_{m=1}^{M} K\big(P_m^{(n)}, P_0^{(n)}\big) \leq \alpha,$$

so that applying (Giné and Nickl, 2016, Theorem 6.3.2) yields the claim (for example $\alpha = 1/16$ suffices). We closely follow the proof of (Giné and Nickl, 2016, Theorem 6.3.9) to construct $f_{0,m}$, and use ideas inspired by (Abraham et al., 2022b) to choose the remaining parameters of θ_m.

Define

$$f_{0,0} = 1, \quad f_{1,0} = f_{0,0} + \zeta\psi_{2,0},$$
$$\psi_{2,0}(x) = \sqrt{3}(2x-1).$$

Note that $f_{0,0}, f_{1,0} \geq 3/4$ pointwise (recall we assumed $\zeta \leq (4\sqrt{3})^{-1}$) and hence any small perturbations of these will remain bounded away from zero.

We choose perturbations $f_{0,m}$ of f_0 to satisfy the second condition of equation (22), and we choose the remaining parameters $f_{1,m}$, p_m, q_m to ensure the Kullback–Leibler condition holds. Proposition 11, which upper bounds the KL divergence by a " distance" ρ will be of help for the latter.

Define the parameters $\theta_m = (p_m, q_m, f_{0,m}, f_{1,m})$ as follows: First, choose $\phi_{1,m} = -1 + c\delta$ and $\phi_{2,m} = \epsilon$ for all $m \geq 0$ and define p_m, q_m according to the inversion formulae in Lemma 12. Next, for $m \geq 1$, for g_m to be chosen define

$$f_{0,m} = f_{0,0} + g_m, \quad f_{1,m} = f_{1,0} - \frac{1+\phi_1}{1-\phi_1} g_m.$$

Writing $\psi_{1,m}, \psi_{2,m}, \phi_{3,m}$ for the corresponding alternative parametrisation as in Section 3.1, the above choice ensures that $\psi_{1,m} = \psi_{1,0}$ regardless of the choice of g_m. We will choose g_m (depending on n) such that $\|\psi_{2,m} - \psi_{2,0}\|_{L^2} \to 0$ (uniformly in m) as $n \to \infty$ so that in particular it is less than 2 eventually, hence

$$\langle \psi_{2,m}, \psi_{2,0} \rangle = 1 - \tfrac{1}{2}\|\psi_{2,m} - \psi_{2,0}\|_{L^2}^2 \geq 0.$$

Under the condition that $\phi_{3,m} \asymp \zeta$, one sees that

$$\rho\big((\phi,\psi)(\theta_m); (\phi,\psi)(\theta_0))\big) = C \max\Big\{\delta\epsilon\zeta|\phi_{3,m} - \phi_{3,0}|, \delta\epsilon\zeta^2\|\psi_{2,m} - \psi_{2,0}\|_{L^2}\Big\}.$$

We calculate $f_{0,m} - f_{1,m} = f_{0,0} - f_{1,0} + \frac{2}{2-c\delta} g_m$ and hence, using that $\|f_{0,0} - f_{1,0}\|_{L^2} = \phi_{3,0} = \zeta$,

$$|\phi_{3,m} - \phi_{3,0}| = \|f_{0,m} - f_{1,m}\|_{L^2} - \|f_{0,0} - f_{1,0}\|_{L^2} \leq \tfrac{2}{2-c\delta}\|g_m\|_{L^2},$$

and

$$\begin{aligned}\|\psi_{2,m} - \psi_{2,0}\|_{L^2} &= \Big\|\frac{f_{0,m} - f_{1,m}}{\phi_{3,m}} - \frac{f_{0,0} - f_{1,0}}{\phi_{3,0}}\Big\|_{L^2} \\ &\leq \frac{|\phi_{3,0} - \phi_{3,m}|}{\phi_{3,m}} + \frac{2\|g_m\|_{L^2}}{2 - c\delta\phi_{3,m}} \lesssim \zeta^{-1}\|g_m\|_{L^2},\end{aligned}$$

yielding

$$\rho\big((\phi,\psi)(\theta_m); (\phi,\psi)(\theta_0))\big) \leq C'\delta\epsilon\zeta\|g_m\|_{L^2}. \tag{23}$$

[provided $c\delta \leq 1$, say, and the condition $\phi_{3,m} \asymp \zeta$ reduces to $\|g_m\|_{L^2} \leq \zeta/3$, say.].

Now we verify that there are M valid choices of g_m such that $f_{0,m}$ and $f_{0,m'}$ are suitably separated in L^2 distance but suitably close in Kullback–Leibler divergence as in (22), and $f_{0,m}$ and $f_{1,m}$ are in the appropriate Sobolev balls. Fix $S \geq s_0$, and let φ_{jk}, $k \leq 2^j$ be a collection of wavelet functions supported in the interior of $[0,1]$ given as scaled translates $\varphi_{jk} = 2^{j/2}\varphi(2^j(\cdot) - k)$ of an S-regular Daubechies wavelet function φ supported in $[1, 2N]$ for some $N = N(S)$. We may choose a collection of $c_0 2^j$ of these functions whose supports are pairwise disjoint for some $c_0 = c_0(S) > 0$; we denote these $\{\varphi_{jp} : 1 \leq p \leq c_0 2^j\}$ in a slight abuse of notation. By the Varsharmov–Gilbert bound (Giné and Nickl, 2016, Example 3.1.4) there exist $c_1, c_2 > 0$ such that we may choose a set $\mathcal{M} = \{\beta_{m,\cdot} \in \{-1,1\}^{c_0 2^j} : m \leq 2^{c_1 2^j}\}$ for which

$$\sum_p |\beta_{mp} - \beta_{m'p}|^2 \geq c_2 2^j, \quad \forall p' \neq p.$$

Set $g_m = \alpha_1 \sum_p \beta_{m,p}\varphi_{jp}$ for α_1 to be chosen and observe that

$$\begin{aligned}\|f_{0,m}\|_{B^{s_0}_{2,\infty}} \leq 1 + \|g_m\|_{B^{s_0}_{2,\infty}} &= 1 + \alpha_1 2^{js_0}\big(\sum_p \beta_{m,p}^2\big)^{1/2} = 1 + c_0\alpha_1 2^{j(s_0+1/2)}, \\ \|g_m\|_{L^2}^2 &= \alpha_1^2 \sum_p \beta_{m,p}^2 \|\varphi_{jp}\|_{L^2}^2 = c_0\alpha_1^2 2^j, \\ \|f_{0,m} - f_{0,m'}\|^2 = \|g_m - g_{m'}\|_{L^2}^2 &= \alpha_1^2 \sum_p |\beta_{m,p} - \beta_{m,p'}|^2 \geq c_2\alpha_1^2 2^j.\end{aligned}$$

The first line ensures that $\|f_{0,m}\|_{B^{s_0}_{2,\infty}} \leq R$ if $\alpha_1^2 \asymp 2^{-j(2s_0+1)}$; note also that consequently $\|f_{1,m}\|_{B^{s_1}_{2,\infty}} \leq 1 + \delta\|g_m\|_{B^{s_1}_{2,\infty}} \lesssim 1 + \delta 2^{j[s_1-s_0]}$. For this choice of α_1, the second line, in conjunction with (23) and Proposition 11 yields that $K(P_m^{(n)}, P_0^{(n)}) \lesssim n\delta^2\epsilon^2\zeta^2 2^{-2js_0}$, so that choosing j such that $2^{j(2s_0+1)} \asymp n\delta^2\epsilon^2\zeta^2$ gives the required bound on Kullback–Leibler divergences in (22). Note also that $\|g\|_\infty \asymp \alpha_1 2^{j/2}$ so that for this choice of j we have $f_{0,m} \geq 1/2, f_{1,m} \geq 1/2$ on $[0,1]$ for n large, hence Proposition 11 indeed applies, and as soon as $(n\delta^2\epsilon^2\zeta^2)^{-s_0/(1+2s_0)} \lesssim \zeta$ we get as needed $\phi_{3,m} \asymp \zeta$. Also, $f_{1,m}$ is in the appropriate Sobolev ball if $\delta^{2s_1+1}(n\epsilon^2\zeta^2)^{s_1-s_0} \lesssim 1$. Finally, for these choices of α_1 and j, the third line yields $\|f_{0,m} - f_{0,m'}\|_{L^2} \gtrsim (n\delta^2\epsilon^2\zeta^2)^{-s_0/(2s_0+1)}$.

We finally prove the general lower bound

$$R_{\text{rough}} = (n\delta^2)^{-s_0/(2s_0+1)},$$

again using a reduction to multiple testing. As before choose $\phi_{1,m} = -1 + c\delta, \phi_{2,m} = \epsilon$, and choose $f_{0,0}, f_{1,0}$ as in proving R_{smooth}. Now set

$$f_{0,m} = f_{0,0} + g_m, \quad f_{1,m} = f_{1,0}.$$

We now have $f_{0,m} - f_{1,m} = f_{0,0} - f_{1,0} + g_m$ which is of the same form as before up to the coefficient $2/(2 - c\delta) \in [1,2]$ which no longer appears. The calculations for ρ then go through fundamentally unchanged except that we no longer have $\psi_{1,m} = \psi_{1,0}$, hence

$$\rho\big((\phi,\psi)(\theta_m); (\phi,\psi)(\theta_0)\big) \leq C' \max\big(\delta\epsilon\zeta\|g_m\|_{L^2}, \|\psi_{1,m} - \psi_{1,0}\|_{L^2}\big).$$

We calculate

$$\psi_{1,m} - \psi_{1,0} = \tfrac{1}{2}(1+\phi_{1,m})f_{0,m} + \tfrac{1}{2}(1-\phi_{1,m})f_{1,m} = \tfrac{1}{2}c\delta g_m,$$

hence calculating the upper bound $C''\delta\|g_m\|_{L^2}$ for ρ.

Choosing $M = \lfloor 2^{c2^j} \rfloor$ functions g_m as before, we again choose the factor α_1 proportional $2^{-j(2s_0+1)}$ to ensure $\|f_{0,m}\|_{B^{s_0}_{2,\infty}} \leq R$; note now that $\|f_{1,m}\|_{B^{s_1}_{2,\infty}} = \|f_{1,0}\|_{B^{s_1}_{2,\infty}}$ for all m so that these are suitably bounded.

Where before we chose $2^{j(2s_0+1)} \asymp n\delta^2\epsilon^2\zeta^2$ to obtain the required bound on the KL divergences in equation (22), we now must choose $2^{j(2s_0+1)} \asymp n\delta^2$. This leads to $\|f_{0,m} - f_{0,m'}\|_{L^2} \gtrsim (n\delta^2)^{-s_0/(2s_0+1)}$ so that equation (22) holds with $R_{\text{rough}} = (n\delta^2)^{-s_0/(2s_0+1)}$ in place of R_{smooth}. This yields the claim.

Appendix C. Proofs for the upper bounds

C.1 Overview of the proofs

The proofs of the upper bounds proceed according to the following strategy.

In Section C.2, we first state a series of lemmas whose purpose is to simplify further proofs. These lemmas are elementary – yet crucial – results about the reparameterization $\theta \mapsto \phi$ and its inversion, the simplification of the expression of $m(\phi)$ given in (16), and

the (quasi) inversion formula to recover ϕ from $m(\phi)$. The section also contains auxiliary results on $p_\theta^{(k)}$ and useful concentration inequalities for Markov chains.

The Section C.3 establishes deviation inequalities for $|\hat{m}_j - m(\phi)_j|$ and $|\hat{m}_j/m(\phi)_j - 1|$. These deviation inequalities are used many times after when using the method of moments to estimate θ. It is worth noting that $m(\phi)$, together with the coefficients $\{\psi_1^{\Phi_{Jk}}\}$, $\{\psi_1^{\Psi_{jk}}\}$, $\{G^{\Phi_{Jk}}\}$ and $\{G^{\Psi_{jk}}\}$ are all easy functionals of $p_\theta^{(s)}$ for some $s \geq 1$, and can all be estimated at a universal marginal rate $c\sqrt{n}$, with c eventually depending on γ_* and L but nothing else. Thus one can think of estimating those quantities as solving the *direct* problem. The main challenge is to translate the inequalities for the direct problem onto inequalities for the *inverse* problem, *i.e.* for (Q_θ, f_0, f_1), which is the purpose of the subsequent subsections.

The Section C.4 proves the Theorem 5, *ie.* the minimax upper bound for estimating Q_θ. This is done in many steps, that can be on a high level summarized by upper bounding $|\hat{p} - p|$ (similarly $|\hat{q} - q|$) by a parameter dependent term times $\max_{j=1,2,3} |\hat{m}_j - m(\phi)_j|$, and then using the concentration inequalities for $|\hat{m}_j - m(\phi)_j|$ to conclude. Here, we emphasize that the obtention of a tight upper bound for $|\hat{p} - p|$ in terms of the deviation of the moments is crucial in obtaining the exact minimax rate and requires substantial work.

The Section C.5 proves the Theorem 6, *ie.* the minimax upper bounds to estimate f_0 and f_1 when $s_0 = s_1$. The proof relies on a somewhat classical decomposition of the risk when studying block-threshold wavelet density estimators, with additional cares to be taken due to the optimal threshold depending on the parameters and being estimated. Modulo these additional cares, the proofs follows the classical steps and is based on deviation inequalities for $\|\hat{f}_m^{\mathfrak{B}_{j\ell}} - f_m^{\mathfrak{B}_{j\ell}}\|$ and similar quantities, to establish that the chosen threshold balances the bias and variance optimally. In contrast with classical density estimation, estimation of the empirical wavelets coefficients requires here to solve an inverse problem. This is done by upper bounding $\|\hat{f}_m^{\mathfrak{B}_{j\ell}} - f_m^{\mathfrak{B}_{j\ell}}\|$ in term of $\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\|$, $\|\hat{G}_1^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\|$ and $\max_{j=1,2,3} |\hat{m}_1 - m(\phi)_j|$, and then using deviation inequalities for the the direct problem.

The Section C.6 proves the Theorem 8, *ie.* the minimax upper bounds to estimate f_0 and f_1 when $s_0 < s_1$. The ideas of the proof are very similar to Theorem 6. The main difference resides in the definition of the empirical wavelet coefficients.

The Section C.7 proves the Theorem 4 about the estimation of the separating hyperplane. Recall that the estimator of the hyperplane is obtained by estimating the leading eigenvector of a certain gram matrix $\mathcal{G}$ from the leading eigenvector of its empirical version $\tilde{\mathcal{G}}$. The proof of the theorem is based on the celebrated Davis-Kahan theorem and the obtention of a deviation bound for $\|\tilde{\mathcal{G}} - \mathcal{G}\|_{\text{op}}$, which is based on a ε-net argument together with concentration inequalities for Markov chains.

Finally, the Sections C.8 and C.9 proves the Corollaries 7 and 9, respectively. Those follow immediately from the Theorems 6 and 8 and straightforward computations.

C.2 Useful lemmas

Lemma 12 *The parametrisation $\theta \mapsto (\phi, \psi)$ from (7) is invertible:*

$$\begin{aligned} p &= \tfrac{1}{2}(1-\phi_2)(1-\phi_1), \\ q &= \tfrac{1}{2}(1-\phi_2)(1+\phi_1), \\ f_0 &= \psi_1 - \tfrac{1}{2}\phi_1\phi_3\psi_2 + \tfrac{1}{2}\phi_3\psi_2, \\ f_1 &= \psi_1 - \tfrac{1}{2}\phi_1\phi_3\psi_2 - \tfrac{1}{2}\phi_3\psi_2. \end{aligned}$$

Defining $p_\pm = \frac{1}{2}(1 \mp \tilde{s}\phi_1)(1-\phi_2)$, where $\tilde{s} \coloneqq \operatorname{sgn}(\langle \psi_2, \tilde{\psi}_2 \rangle)$ we have

$$(p_+, p_-) \coloneqq \begin{cases} (p, q) & \text{if } \tilde{s} > 0, \\ (q, p) & \text{if } \tilde{s} < 0. \end{cases}$$

Recalling the definition (16) of m, define

$$g \coloneqq \phi_3 |\tilde{\mathcal{I}}| = \frac{\sqrt{4m_1^2 m_2 + m_3^2}}{m_2},$$

and define

$$f_\pm \coloneqq \psi_1 \pm \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1} G, \qquad G \coloneqq \frac{m_1 \psi_2}{\tilde{\mathcal{I}}}.$$

Then

$$(f_+, f_-) \coloneqq \begin{cases} (f_0, f_1) & \text{if } \tilde{s} > 0, \\ (f_1, f_0) & \text{if } \tilde{s} < 0. \end{cases}$$

The proof is elementary. Note that $\mathbb{P}_n^{(1)}(\Phi_{Jk})$ is the empirical estimator of $\mathbb{E}_\theta[\Phi_{Jk}] = \langle \Phi_{Jk}, \psi_1 \rangle$, hence the above lemma justifies the use of $\hat{f}_0^{\Phi_{Jk}}$, $\hat{f}_1^{\Phi_{Jk}}$ from Section 3.5.

Lemma 13 *Given $p_{\phi,\psi}^{(3)}$ as defined in (9) and any function $\tilde{\psi}_2$, one can compute*

$$\begin{aligned} r(\phi)\tilde{\mathcal{I}}^2 &= \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)^2 \\ r(\phi)\phi_2\tilde{\mathcal{I}}^2 &= \mathbb{E}_\theta(\tilde{\psi}_2 \otimes 1 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)^2 \\ r(\phi)\phi_1\phi_2\phi_3\tilde{\mathcal{I}}^3 &= -\mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2) + \mathbb{E}_\theta(\tilde{\psi}_2)^3 + \Big(2r(\phi)\tilde{\mathcal{I}}^2 + r(\phi)\phi_2\tilde{\mathcal{I}}^2\Big)\mathbb{E}_\theta(\tilde{\psi}_2). \end{aligned}$$

Also if $G = m_1\psi_2/\tilde{\mathcal{I}}$, then $\langle \Phi_{Jk}, G \rangle = \mathbb{E}[\tilde{\psi}_2 \otimes \Phi_{Jk}] - \mathbb{E}_\theta[\tilde{\psi}_2]\mathbb{E}_\theta[\Phi_{Jk}]$.

Proof We compute, from the expression for $p_{\phi,\psi}^{(3)}$, applied for example to $\tilde{\psi}_2 \otimes 1 \otimes 1$ and using that $\langle \psi_1, 1 \rangle = \int \psi_1 = 1$, $\langle \psi_2, 1 \rangle = 0$,

$$\begin{aligned} \mathbb{E}_\theta(\tilde{\psi}_2) &= \langle \psi_1, \tilde{\psi}_2 \rangle \\ \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2) &= \langle \psi_1, \tilde{\psi}_2 \rangle^2 + r(\phi)\langle \psi_2, \tilde{\psi}_2 \rangle^2 \\ \mathbb{E}_\theta(\tilde{\psi}_2 \otimes 1 \otimes \tilde{\psi}_2) &= \langle \psi_1, \tilde{\psi}_2 \rangle^2 + r(\phi)\phi_2\langle \psi_2, \tilde{\psi}_2 \rangle^2 \\ \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2) &= \langle \psi_1, \tilde{\psi}_2 \rangle^3 + (2r(\phi) + r(\phi)\phi_2)\langle \psi_2, \tilde{\psi}_2 \rangle^2 \langle \psi_1, \tilde{\psi}_2 \rangle - r(\phi)\phi_1\phi_2\phi_3\langle \psi_2, \tilde{\psi}_2 \rangle^3 \end{aligned}$$

Then $m := (r(\phi)\tilde{\mathcal{I}}^2, r(\phi)\phi_2\tilde{\mathcal{I}}^2, r(\phi)\phi_1\phi_2\phi_3\tilde{\mathcal{I}}^3)$, $\tilde{\mathcal{I}} := \langle \psi_2, \tilde{\psi}_2 \rangle$ is easily extracted.

Similarly, $\mathbb{E}_\theta[\tilde{\psi}_2 \otimes \Phi_{Jk}] = \langle \psi_1, \tilde{\psi}_2 \rangle \langle \psi_1, \Phi_{Jk} \rangle + r(\phi)\tilde{\mathcal{I}} \langle \psi_2, \Phi_{Jk} \rangle$, and the expression for the coefficient of G can be extracted. ■

Lemma 14 (Inversion formulas for m) *Let $m(\phi) = (r(\phi)\tilde{\mathcal{I}}^2, r(\phi)\phi_2\tilde{\mathcal{I}}^2, r(\phi)\phi_1\phi_2\phi_3\tilde{\mathcal{I}}^3)$ with $\tilde{\mathcal{I}} \neq 0$. Then,*

$$\begin{aligned} \operatorname{sgn}(\tilde{\mathcal{I}})\phi_1 &= \frac{m_3(\phi)}{\sqrt{4m_1(\phi)^2 m_2(\phi) + m_3(\phi)^2}}, \\ \phi_2 &= \frac{m_2(\phi)}{m_1(\phi)}, \\ \phi_3|\tilde{\mathcal{I}}| &= \frac{\sqrt{4m_1(\phi)^2 m_2(\phi) + m_3(\phi)^2}}{m_2(\phi)}. \end{aligned}$$

Proof This can be checked via direct computations. ■

The following bounds are immediate from the definition of the parameter space (3) and the reparametrisation (7) (recall also the definition (8) of r).

Lemma 15 *For ϕ corresponding to $\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$ we have the bounds*

$$-\frac{1-\delta}{1+\delta} \leq \phi_1 \leq \frac{1-\delta}{1+\delta}, \quad \epsilon \leq |\phi_2| \leq 1-2\delta, \quad \phi_3 \geq \zeta, \quad \delta\epsilon\zeta^2/4 \leq |r(\phi)| \leq \phi_3^2/4.$$

Lemma 16 *Let m_1, m_2, m_3 be defined as in (16) and let $v := 4m_1^2 m_2 + m_3^2$. Then $0 \leq m_2 \leq |m_1|$ and $\sqrt{v} = \tilde{\mathcal{I}}^3 r(\phi)\phi_2\phi_3 = \tilde{\mathcal{I}} m_2 \phi_3$. Furthermore, for every $\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$ and $0 < \delta \leq 1$, $0 < \epsilon \leq 1$, and $0 < \zeta \leq 1$:*

$$\Big|\frac{g}{m_1}\Big| \leq \frac{4}{\delta\epsilon\zeta|\tilde{\mathcal{I}}|}, \qquad \frac{\max(1,g)}{m_2} \leq \frac{4}{\delta\epsilon^2\zeta^2|\tilde{\mathcal{I}}|^2}, \qquad \frac{\max(1,g)}{g m_2} \leq \frac{4}{\delta\epsilon^2\zeta^3|\tilde{\mathcal{I}}|^3}.$$

Proof Observe that $m_2 = m_1\phi_2$ and $|\phi_2| \leq 1$. Also, $m_2 = r(\phi)\phi_2\tilde{\mathcal{I}}^2 = \frac{1}{4}(1-\phi_1^2)\phi_2^2\phi_3^2\tilde{\mathcal{I}}^2 \geq 0$. Similarly,

$$v = 4r(\phi)^2\tilde{\mathcal{I}}^4 \cdot r(\phi)\phi_2\tilde{\mathcal{I}}^2 + r(\phi)^2\phi_1^2\phi_2^2\phi_3^2\tilde{\mathcal{I}}^6 = r(\phi)^2\tilde{\mathcal{I}}^6\Big(4r(\phi)\phi_2 + \phi_1^2\phi_2^2\phi_3^2\Big) = r(\phi)^2\phi_2^2\phi_3^2\tilde{\mathcal{I}}^6.$$

Next, observe that $\frac{g}{m_1} = \frac{\phi_3|\tilde{\mathcal{I}}|}{\frac{1}{4}(1-\phi_1^2)\phi_2\phi_3^2|\tilde{\mathcal{I}}|^2} = \frac{4}{(1-\phi_1^2)\phi_2\phi_3|\tilde{\mathcal{I}}|}$. But $0 \geq 1-\phi_1^2 \geq \frac{4\delta}{(1+\delta)^2} \geq \delta$, $|\phi_2| \geq \epsilon$, and $\phi_3 \geq \zeta$ by Lemma 15. Similarly, since $g = \phi_3|\tilde{\mathcal{I}}| \leq \zeta \leq 1$, $0 \leq \frac{\max(1,g)}{m_2} = \frac{1}{m_2} = \frac{4}{(1-\phi_1^2)\phi_2^2\phi_3^2|\tilde{\mathcal{I}}|^2} \leq \frac{4}{\delta\epsilon^2\zeta^2|\tilde{\mathcal{I}}|^2}$. ■

Lemma 17 *For any $k \geq 1$,*

$$\|p_\theta^{(k)}\|_\infty \leq \max\big(\|f_0\|_\infty,\ \|f_1\|_\infty\big)^k.$$

Consequently, for any $\theta \in \Sigma_{\gamma^*}(L)$ and any measurable function $h : \mathbb{R}^k \to \mathbb{R}$, we have

$$\mathbb{E}_\theta[h(Y_1, \ldots, Y_k)^2] \leq L^k \|h\|_{L^2}^2.$$

Proof Observe that $p_\theta^{(k)}(y_1,\dots,y_k) = \sum_{x_1,\dots,x_k} \mathbb{P}_\theta(X_1 = x_1,\dots,X_k = x_k)\prod_{i=1}^k f_{x_i}(y_i)$. The first conclusion is immediate, and the second follows from

$$\mathbb{E}_\theta h(Y_1,\dots,Y_k)^2 = \int p_\theta^{(k)}(y_1,\dots,y_k)h(y_1,\dots,y_k)\mathrm{d}y_1\cdots\mathrm{d}y_k \le \|p^{(k)}\|_\infty \|h\|_{L^2}^2.$$

■

Remark 18 *The proof adapts to yield $E_\theta[h(Y_1,Y_3)^2] \le L^2\|h\|_{L^2}^2$ rather than the weaker bound $L^3\|h\|_{L^2}^2$ directly obtainable using the lemma. Indeed, we have*

$$\sup_{y_1,y_3}\Big|\int p_\theta^{(3)}(y_1,y_2,y_3)\mathrm{d}y_2\Big| = \sum_{x_1,x_2,x_3} \mathbb{P}_\theta(X_1 = x_1, X_2 = x_2, X_3 = x_3) f_{x_1}(y_1) f_{x_3}(y_3) \le L^2,$$

and the rest of the proof is the same.

Lemma 19 *For all $\theta \in \Sigma_{\gamma^*}(L)$, $\phi_3 \le \sqrt{2L}$.*

Proof We compute $\phi_3^2 = \int_0^1 (f_0 - f_1)^2 \le \|f_0 - f_1\|_\infty \int_0^1 (|f_0| + |f_1|) = 2\|f_0 - f_1\|_\infty$. Since we have the pointwise bounds $0 \le f_0, f_1 \le L$ for every $\theta \in \Sigma_{\gamma^*}(L)$, it follows that $\phi_3^2 \le 2L$. We remark that this upper bound is tight since it is attained for instance when f_0 is the uniform density on $[0, 1/L]$ and f_1 the uniform density on $[1 - 1/L, 1]$. ■

We now recall the following result, which is adapted from (Paulin, 2015) and will be key to getting deviation inequalities of empirical ingredients in our procedures.

Lemma 20 *Let $1 \le k \le 3$ and let $h : \mathbb{R}^k \to \mathbb{R}$ be measurable. There is a universal constant $C > 0$ such that for all θ, all $n \ge 4$ such that $n\gamma^* \ge 1/99$, and all $t \ge 0$*

$$\mathbb{P}_\theta\Big(|\mathbb{P}_n^{(k)}(h) - \mathbb{E}_\theta(h)| \ge t\Big) \le \exp\Big(-\frac{Cnt^2\gamma^*}{\mathbb{E}_\theta(h^2) + \|h\|_\infty t}\Big).$$

This in particular implies that there is a is a universal constant $C > 0$ such that for all θ, all $n \ge 4$ such that $n\gamma^* \ge 1/99$, and all $x \ge 0$

$$\mathbb{P}_\theta\left(|\mathbb{P}_n^{(k)}(h) - \mathbb{E}_\theta(h)| \ge C\sqrt{\frac{\mathbb{E}_\theta[h^2]x}{n\gamma^*}} + \frac{C\|h\|_\infty x}{n\gamma^*}\right) \le e^{-x}.$$

Proof Since $1 \le k \le 3$, we can view any function $h : \mathbb{R}^k \to \mathbb{R}$ as $\tilde{h} : \mathbb{R}^6 \to \mathbb{R}$ with $h(Y_i,\dots,Y_{i+k}) = \tilde{h}(X_i, X_{i+1}, X_{i+2}, Y_i, Y_{i+1}, Y_{i+2})$. The process $\big((X_i, X_{i+1}, X_{i+2}, Y_i, Y_{i+1}, Y_{i+2})\big)_{i\ge1}$ is a stationary Markov Chain with pseudo spectral gap (defined as in Paulin (2015)) $\gamma_{\mathrm{ps}} \ge \gamma^*/8$, by our assumptions. Indeed, calculations in (Abraham et al., 2022b, Lemma 1) based on the relationship between the pseudo spectral gap and the mixing time show that $\gamma_{\mathrm{ps}} \ge 0.5((\log 4/\gamma^*) + 2)^{-1}$, and the bound $\max(\gamma^*, \log 2) \le 1$ yields the claimed bound.

By Theorem 3.4 in (Paulin, 2015) (though note there is an updated version of the paper on arXiv), for $S_n \coloneqq \sum_{i=1}^{n-k+1} \tilde{h}(X_i, X_{i+1}, X_{i+2}, Y_i, Y_{i+1}, Y_{i+2})$ we do have for any $t \geq 0$

$$\mathbb{P}_\theta\Big(|S_n - \mathbb{E}_\theta(S_n)| \geq t\Big) \leq \exp\Big(-\frac{t^2\gamma_{\mathrm{ps}}}{8(n-k+1+1/\gamma_{\mathrm{ps}})\mathbb{E}_\theta(h^2) + 20\|h\|_\infty t}\Big).$$

Dividing S_n by $n-k+1$ and replacing $n-k+1$ and γ_{ps} by the respective lower bounds $n/2$ and $\gamma^*/8$, we find that

$$\begin{aligned}\mathbb{P}_\theta\Big(|\mathbb{P}_n^{(k)}(h) - \mathbb{E}_\theta(h)| \geq t\Big) &\leq \exp\Big(-\frac{nt^2\gamma^*/16}{8(1+\frac{16}{n\gamma^*})\mathbb{E}_\theta(h^2) + 20\|h\|_\infty t}\Big) \\ &\leq \exp\Big(-\frac{nt^2\gamma^*}{16\times 8\times(1+16\times 99)\times\mathbb{E}_\theta(h^2) + 320\|h\|_\infty t}\Big)\end{aligned}$$

under the assumption that $n\gamma^* \geq 1/99$. The result follows by taking $t = C\sqrt{\mathbb{E}_\theta[h^2]x/(n\gamma^*)} + C\|h\|_\infty x/(n\gamma^*)$ for C a sufficiently large constant that the argument of the exponential is smaller than $-x$ (by splitting into cases based on which of the two terms in the denominator is larger it can be seen that it suffices to take $C = \max(\sqrt{2\times 16\times 8\times(1+16\times 99)}, 640) = 640$), yielding the claim. ■

The following consequence of deviation inequalities to get bounds in expectation will also be used.

Lemma 21 *Suppose X is a non-negative random variable and there exist $a, b, c > 0$ such that $\mathbb{P}(X > b\sqrt{x/n} + ax/n) \leq ce^{-x}$ for all $x > 0$. There for all $d \geq 0$*

$$\mathbb{E}(X^2\mathbf{1}_{\{X>d\}}) \leq c\Big(d^2 + \frac{5b^2}{4n} + \frac{7a^2}{2n^2}\Big)\exp\Big(-\frac{nd^2}{2b^2+8ad}\Big).$$

Proof Applying the standard identity $\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y > y)\mathrm{d}y$ for any non-negative random variable Y to $Y = X^2\mathbf{1}_{\{X>d\}}$ and making the substitution $y = u^2$ we obtain

$$\begin{aligned}\mathbb{E}(X^2\mathbf{1}_{\{X>d\}}) &= \int_0^\infty \mathbb{P}\big(X^2\mathbf{1}_{\{X>d\}} > y\big)\mathrm{d}y \\ &= \int_0^\infty \mathbb{P}\big(X > \max(d, \sqrt{y})\big)\mathrm{d}y \\ &= \int_0^{d^2} \mathbb{P}(X > d)\mathrm{d}y + \int_{d^2}^\infty \mathbb{P}\big(X > \sqrt{y}\big)\mathrm{d}y \\ &= d^2\mathbb{P}(X > d) + \int_d^\infty 2u\mathbb{P}(X > u)\mathrm{d}u.\end{aligned}$$

Define $\varphi(x) \coloneqq \frac{b}{2a}\big(\sqrt{1+4ax/b^2} - 1\big)$. For the change of variables $u = b\sqrt{x/n} + ax/n$ one calculates that $x = n\varphi(u)^2$ and hence computes, using Cauchy–Schwarz for the penultimate

line,

$$\begin{aligned}\int_d^\infty u\mathbb{P}(X>u)\mathrm{d}u &= \int_{n\varphi(d)^2}^\infty \Big(b\sqrt{\frac{x}{n}}+a\frac{x}{n}\Big)\Big(\frac{b}{2\sqrt{nx}}+\frac{a}{n}\Big)\mathbb{P}\Big(X>b\sqrt{\frac{x}{n}}+a\frac{x}{n}\Big)\mathrm{d}x\\ &\le c\int_{n\varphi(d)^2}^\infty \Big(\frac{b^2}{2n}+\frac{3}{2}\frac{b}{\sqrt{n}}\frac{a\sqrt{x}}{n}+\frac{a^2x}{n^2}\Big)e^{-x}\mathrm{d}x\\ &\le c\int_{n\varphi(d)^2}^\infty \Big(\frac{5b^2}{4n}+\frac{7a^2x}{4n^2}\Big)e^{-x}\mathrm{d}x\\ &= \frac{c}{4}\Big(\frac{5b^2}{n}+\frac{7a^2}{n^2}(n\varphi(d)^2+1)\Big)e^{-n\varphi(d)^2}.\end{aligned}$$

Similarly one has

$$\mathbb{P}(X>d)=\mathbb{P}\Big(X>b\sqrt{\frac{n\varphi(d)^2}{n}}+a\frac{n\varphi(d)^2}{n}\Big)\le ce^{-n\varphi(d)^2}.$$

To obtain the final expression, we remark that $xe^{-x}\le \frac{2}{e}e^{-x/2}$, that $2/e+1\le 2$ and that for all $x>0$

$$\varphi(x)\ge \frac{b}{2a}\frac{4ax/b^2}{2\sqrt{1+4ax/b^2}}=\frac{x/b}{\sqrt{1+4ax/b^2}}.$$

■

C.3 Inequalities for the m functional

Recall the definitions

$$\begin{aligned}\hat{m}_1 &\coloneqq \mathbb{P}_n^{(2)}(\tilde{\psi}_2\otimes\tilde{\psi}_2)-\mathbb{P}_n^{(1)}(\tilde{\psi}_2)^2,\\ \hat{m}_2 &\coloneqq \mathbb{P}_n^{(3)}(\tilde{\psi}_2\otimes 1\otimes\tilde{\psi}_2)-\mathbb{P}_n^{(1)}(\tilde{\psi}_2)^2\\ \hat{m}_3 &= -\mathbb{P}_n^{(3)}(\tilde{\psi}_2\otimes\tilde{\psi}_2\otimes\tilde{\psi}_2)+\mathbb{P}_n^{(1)}(\tilde{\psi}_2)^3+\big(2\hat{m}_1+\hat{m}_2\big)\mathbb{P}_n^{(1)}(\tilde{\psi}_2),\end{aligned}$$

estimators of the functional m defined in (16) as $m=(r(\phi)\tilde{\mathcal{I}}^2, r(\phi)\phi_2\tilde{\mathcal{I}}^2, r(\phi)\phi_1\phi_2\phi_3\tilde{\mathcal{I}}^3)$ with $\tilde{\mathcal{I}}=\langle\psi_2,\tilde{\psi}_2\rangle$, and deduced from Lemma 13 to be equal to what is obtained in the expressions for $\hat{m}$ on replacing every instance of an empirical estimator by the expectation operator. [This does not mean that $\mathbb{E}_\theta\hat{m}=m$, since there are powers and products in the expressions.] In this section, we prove deviation inequalities for the estimators of m, from which we deduce bounds in expectation. The results of this section will be used to prove Theorem 5 and Theorem 6.

We remark that the results are mostly uniform over the whole class $\Sigma_{\gamma^*}(L)$, not our final parameter set $\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)$. The need to intersect with $\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$ arises for ensuring the parameters θ are identifiable from m.

Proposition 22 *Let $n\gamma^*\ge 1/99$. Then there exists a universal constant $C>0$ such that for all $x\ge 0$*

$$\sup_{\theta\in\Sigma_{\gamma^*}(L)}\mathbb{P}_\theta\left(\max_{j=1,2}|\hat{m}_j-m_j|\ge CL\sqrt{\frac{x}{n\gamma^*}}+C\max(\tau,\sqrt{L})^2\frac{x}{n\gamma^*}\right)\le 3e^{-x}.$$

Proposition 23 *Let $n\gamma^* \geq 1/99$. Then there exists a universal constant $C > 0$ such that for all $x \geq 0$*

$$\sup_{\theta \in \Sigma_{\gamma^*}(L)} \mathbb{P}_\theta \left(\max_{j=1,2,3} |\hat{m}_j - m_j| \geq CL^{3/2} \sqrt{\frac{x}{n\gamma^*}} + C \max(\tau, \sqrt{L})^3 \frac{x}{n\gamma^*} \right) \leq 4e^{-x}.$$

Proposition 24 *There exists a constant $K > 0$ such that whenever $n\gamma^* \geq 1/99$,*

$$\sup_{\theta \in \Sigma_{\gamma^*}(L)} \mathbb{E}_\theta \Big(\max_{j=1,2,3} |\hat{m}_j - m_j|^2 \Big) \leq K \Big(\frac{L^3}{n\gamma^*} + \frac{\max(\tau, \sqrt{L})^6}{(n\gamma^*)^2} \Big).$$

Proposition 25 *Assume $n\gamma^* \geq 1/99$, $|\tilde{\mathcal{I}}| \geq 7/8$ and $\zeta \leq 1$, and define the event*

$$\Omega_n := \left\{ \max_{j=1,2} \left| \frac{\hat{m}_j}{m_j} - 1 \right| \leq \frac{1}{2}, \ \max_{j=1,2,3} |\hat{m}_j - m_j| \leq \frac{g m_2}{44 \max(1, g)} \right\}. \tag{24}$$

Then there exists a universal constant $C > 0$ such that

$$\sup_{\theta \in \Sigma_{\gamma^*}(L)} \mathbb{P}_\theta(\Omega_n^c) \leq 7 \exp \left(- \frac{C n \gamma^* g^2 m_2^2 / \max(1, g)^2}{L^3 + \max(\tau, \sqrt{L})^3 g m_2 / \max(1, g)} \right),$$

$$\sup_{\theta \in \Theta^{s_0, s_1}_{\delta, \epsilon, \zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{P}_\theta(\Omega_n^c) \leq 7 \exp \left(- \frac{C n \gamma^* \delta^2 \epsilon^4 \zeta^6}{L^3 + \max(\tau, \sqrt{L})^3 \delta \epsilon^2 \zeta^3} \right).$$

The proof of Proposition 23 is the most involved of these, and we outline how to prove the other results before addressing it.

Proof [Proof of Proposition 22] The proof is similar to the proof of Proposition 23, where $\max_{j=1,2,3} |\hat{m}_j - m_j|$ is controlled. Here, since only $\hat{m}_1$ and $\hat{m}_2$ are involved, the proxy variance is no more than L since only $\mathbb{P}_n^{(2)}$ is involved (versus $L^{3/2}$ when $\mathbb{P}_n^{(3)}$ is involved). ■

Proof [Proof of Proposition 24] In view of Proposition 23 we may apply Lemma 21 with $a = C \max(\tau, \sqrt{L})^3 / \gamma^*$, $b = CL^{3/2} / \sqrt{\gamma^*}$, $c = 8$ and $d = 0$ to obtain the claimed bound. ■

Proof [Proof of Proposition 25] The first inequality essentially follows from Propositions 22 and 23 and a change of variables: see Lemmas 27 and 28 (and the sentence after the former) below where this change of variables is explicitly made. The second inequality follows from the fact that $\frac{\max(1,g)}{g m_2} \leq \frac{16}{\delta \epsilon^2 \zeta^3 \tilde{\mathcal{I}}^2}$ on $\Theta^{s_0, s_1}_{\delta, \epsilon, \zeta}(R)$ by Lemma 16. ■

Proof [Proof of Proposition 23] We have that $\max_{j=1,2,3} |\hat{m}_j - m_j| \leq 16 \|\tilde{\psi}_2\|_\infty^3 \leq 16\tau^3$ by construction. Hence whenever $x > n\gamma^*$ we have with probability $1 \geq 1 - e^{-x}$ under $\mathbb{P}_\theta$ that

$$\max_{j=1,2,3} |\hat{m}_j - m_j| \leq 16\tau^3 \leq CL^{3/2} \sqrt{\frac{x}{n\gamma^*}} + C \max(\tau, \sqrt{L})^3 \frac{x}{n\gamma^*}$$

Next we address the case $x \leq n\gamma^*$. It is seen that

$$\hat{m}_1 - m_1 = \mathbb{P}_n^{(2)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2) - \Big(\mathbb{P}_n^{(1)}(\tilde{\psi}_2)^2 - \mathbb{E}_\theta(\tilde{\psi}_2)^2\Big)$$

ie.

$$\hat{m}_1 - m_1 = \Big(\mathbb{P}_n^{(2)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2)\Big) - 2\mathbb{E}_\theta(\tilde{\psi}_2)\Big(\mathbb{P}_n^{(1)}(\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)\Big) - \Big(\mathbb{P}_n^{(1)}(\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)\Big)^2.$$

Noting that $\mathbb{E}_\theta(|\tilde{\psi}_2|) \leq \mathbb{E}_\theta(\tilde{\psi}_2^2)^{1/2} \leq \sqrt{L}\|\tilde{\psi}_2\|_{L^2} = \sqrt{L}$ whenever $\theta \in \Sigma_{\gamma^*}(L)$ by Lemma 17, we deduce

$$|\hat{m}_1 - m_1| \leq |Z_2| + 2\sqrt{L}|Z_1| + |Z_1|^2,$$

where $Z_1 = \mathbb{P}_n^{(1)}(\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)$ and $Z_2 = \mathbb{P}_n^{(2)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2)$. The same reasoning yields, with , $Z_3 = \mathbb{P}_n^{(3)}(\tilde{\psi}_2 \otimes 1 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2 \otimes 1 \otimes \tilde{\psi}_2)$,

$$|\hat{m}_2 - m_2| \leq |Z_3| + 2\sqrt{L}|Z_1| + |Z_1|^2.$$

The decomposition for $\hat{m}_3 - m_3$ is similar but slightly more involved. Since $m_3 = -\mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2) + \mathbb{E}_\theta(\tilde{\psi}_2)^3 + \big(2m_1 + m_2\big)\mathbb{E}_\theta(\tilde{\psi}_2)$, we deduce

$$\begin{aligned}
\hat{m}_3 - m_3 = &-\Big(\mathbb{P}_n^{(3)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2)\Big) \\
&+ \mathbb{P}_n^{(1)}(\tilde{\psi}_2)^3 - \mathbb{E}_\theta(\tilde{\psi}_2)^3 \\
&+ \big[(2\hat{m}_1 + \hat{m}_2) - (2m_1 + m_2)\big]\mathbb{E}_\theta(\tilde{\psi}_2) \\
&+ (2m_1 + m_2)\big(\mathbb{P}_n^{(1)}(\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)\big) \\
&+ \big[(2\hat{m}_1 + \hat{m}_2) - (2m_1 + m_2)\big]\big(\mathbb{P}_n^{(1)}(\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)\big).
\end{aligned}$$

But $\mathbb{P}_n^{(1)}(\tilde{\psi}_2)^3 - \mathbb{E}_\theta(\tilde{\psi}_2)^3 = 3\mathbb{E}_\theta(\tilde{\psi}_2)^2 Z_1 + 3\mathbb{E}_\theta(\tilde{\psi}_2)Z_1^2 + Z_1^3$, and thus recalling $\mathbb{E}_\theta(|\tilde{\psi}_2|) \leq \sqrt{L}$ and $m_2 \leq |m_1| \leq \frac{1}{4}\phi_3^2 \leq \frac{1}{2}L$ by Lemmas 19 and 15, writing $Z_4 = \mathbb{P}_n^{(3)}(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2 \otimes \tilde{\psi}_2 \otimes \tilde{\psi}_2)$ we have

$$\begin{aligned}
|\hat{m}_3 - m_3| \leq |Z_4| + 3L|Z_1| + 3\sqrt{L}|Z_1|^2 + |Z_1|^3 + 2\sqrt{L}|\hat{m}_1 - m_1| + \sqrt{L}|\hat{m}_2 - m_2| \\
+ \frac{3L}{2}|Z_1| + 2|\hat{m}_1 - m_1||Z_1| + |\hat{m}_2 - m_2||Z_1|.
\end{aligned}$$

It follows (recall $L \geq 1$ necessarily)

$$\begin{aligned}
\max_{j=1,2,3} |\hat{m}_j - m_j| \leq |Z_4| + \sqrt{L}|Z_3| + 2\sqrt{L}|Z_2| + 10.5L|Z_1| \\
+ 9\sqrt{L}Z_1^2 + 4|Z_1|^3 + 2|Z_1 Z_2| + |Z_1 Z_3|.
\end{aligned}$$

Feeding in bounds on the Z_i from Lemma 26 below, we deduce with probability at least $1-4e^{-x}$ under $\mathbb{P}_\theta$ that

$$\begin{aligned}\max_{j=1,2,3}|\hat{m}_j - m_j| \leq\ & C\left(L^{3/2}\sqrt{\frac{x}{n\gamma^*}} + \tau^3\frac{x}{n\gamma^*}\right) + 3C\left(L^{3/2}\sqrt{\frac{x}{n\gamma^*}} + L^{1/2}\tau^2\frac{x}{n\gamma^*}\right)\\ &+ 10.5C\left(L^{3/2}\sqrt{\frac{x}{n\gamma^*}} + L\tau\frac{x}{n\gamma^*}\right) + 9C^2\sqrt{L}\left(L^{1/2}\sqrt{\frac{x}{n\gamma^*}} + \tau\frac{x}{n\gamma^*}\right)^2\\ &+ 4C^3\left(L^{1/2}\sqrt{\frac{x}{n\gamma^*}} + \tau\frac{x}{n\gamma^*}\right)^3\\ &+ 3C^2\left(L^{1/2}\sqrt{\frac{x}{n\gamma^*}} + \tau\frac{x}{n\gamma^*}\right)\left(L\sqrt{\frac{x}{n\gamma^*}} + \tau^2\frac{x}{n\gamma^*}\right).\end{aligned}$$

Grouping together the terms with same powers, still with probability at least $1-8e^{-x}$ under $\mathbb{P}_\theta$

$$\begin{aligned}\max_{j=1,2,3}|\hat{m}_j - m_j| \leq\ & 14.5CL^{3/2}\Big(\frac{x}{n\gamma^*}\Big)^{1/2} + C\Big(\tau^3 + 3L^{1/2}\tau^2 + 10.5L\tau + 12CL^{3/2}\Big)\frac{x}{n\gamma^*}\\ &+ C^2\Big(18\tau L + 4CL^{3/2} + 3\tau^2\sqrt{L} + 3\tau L\Big)\Big(\frac{x}{n\gamma^*}\Big)^{3/2}\\ &+ C^2\Big(9\sqrt{L}\tau^2 + 12C\tau L + 3\tau^3\Big)\Big(\frac{x}{n\gamma^*}\Big)^2 + 12C^3\tau^2\sqrt{L}\Big(\frac{x}{n\gamma^*}\Big)^{5/2}\\ &+ 4C^3\tau^3\Big(\frac{x}{n\gamma^*}\Big)^3.\end{aligned}$$

The conclusion follows since we are in the case where $x \leq n\gamma^*$, and because $L \geq 1$ and $\tau \geq 1$. ■

Lemma 26 *Assume $\theta \in \Sigma_{\gamma^*}(L)$ and $n\gamma^* \geq 1/99$. Write $Z_1 = \mathbb{P}_n^{(1)}(\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2)$, $Z_2 = \mathbb{P}_n^{(2)}(\tilde{\psi}_2\otimes\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2\otimes\tilde{\psi}_2)$, $Z_3 = \mathbb{P}_n^{(3)}(\tilde{\psi}_2\otimes 1\otimes\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2\otimes 1\otimes\tilde{\psi}_2)$, and $Z_4 = \mathbb{P}_n^{(3)}(\tilde{\psi}_2\otimes\tilde{\psi}_2\otimes\tilde{\psi}_2) - \mathbb{E}_\theta(\tilde{\psi}_2\otimes\tilde{\psi}_2\otimes\tilde{\psi}_2)$. Then*

$$\mathbb{P}_\theta\Big(|Z_1| \geq C\sqrt{\frac{Lx}{n\gamma^*}} + C\tau\frac{x}{n\gamma^*}\Big) \leq e^{-x},$$

$$\mathbb{P}_\theta\Big(|Z_j| \geq CL\sqrt{\frac{x}{n\gamma^*}} + C\tau^2\frac{x}{n\gamma^*}\Big) \leq e^{-x}, \quad j = 2,3,$$

$$\mathbb{P}_\theta\Big(|Z_4| \geq CL^{3/2}\sqrt{\frac{x}{n\gamma^*}} + C\tau^3\frac{x}{n\gamma^*}\Big) \leq e^{-x}.$$

Proof For Z_4, use Lemma 20 together with the facts that $\|\tilde{\psi}_2\otimes\tilde{\psi}_2\otimes\tilde{\psi}_2\|_\infty = \|\tilde{\psi}_2\|_\infty^3 \leq \tau^3$ and that $\mathbb{E}_\theta[(\tilde{\psi}_2\otimes\tilde{\psi}_2\otimes\tilde{\psi}_2)^2] \leq L^3\|\tilde{\psi}_2\|_{L^2}^6 = L^3$ by Lemma 17. The arguments are similar for $j = 1,2,3$, though note for $j = 3$ we use Remark 18 rather than Lemma 17 itself. ■

Lemma 27 *Let $n\gamma^* \geq 1/99$. Then, there exists a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$*

$$\mathbb{P}_\theta\Big(\max_{j=1,2}\Big|\frac{\hat{m}_j}{m_j} - 1\Big| \geq \frac{1}{2}\Big) \leq 3\exp\bigg(-\frac{Cn\gamma^* m_2^2}{L^2 + \max(\tau, \sqrt{L})^2 m_2}\bigg).$$

Note that $\frac{gm_2}{\max(1,g)} \leq m_2$ and that $L \geq 1$ necessarily, hence the the absolute value of the exponent in Lemma 27 is larger than that in Lemma 28.

Proof We apply Proposition 22 with $x \geq 0$ such that

$$CL\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau, \sqrt{L})^2\frac{x}{n\gamma^*} = \frac{m_2}{2},$$

i.e.,

$$\begin{aligned}\sqrt{\frac{x}{n\gamma^*}} &= \frac{L}{2\max(\tau, \sqrt{L})^2}\Big(\sqrt{1 + \frac{2\max(\tau, \sqrt{L})^2 m_2}{CL^2}} - 1\Big)\\ &\geq \frac{L}{2}\frac{m_2/(CL^2)}{\sqrt{1 + \frac{2\max(\tau,\sqrt{L})^2 m_2}{CL^2}}}.\end{aligned}$$

Then, using that $0 \leq m_2 \leq |m_1|$, (Lemma 16), we have

$$\begin{aligned}\mathbb{P}_\theta\Big(\max_{j=1,2}\Big|\frac{\hat{m}_j}{m_j} - 1\Big| \geq \frac{1}{2}\Big) &\leq \mathbb{P}_\theta\Big(\max_{j=1,2}\Big|\hat{m}_j - m_j\Big| \geq \frac{m_2}{2}\Big)\\ &\leq 6\exp\bigg(-\frac{n\gamma^* m_2^2}{2C^2L^2 + 2C\max(\tau, \sqrt{L})^2|m_2|}\bigg)\end{aligned}$$

concluding the proof. ∎

Lemma 28 *Let $n\gamma^* \geq 1/99$. Then, there exists a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$*

$$\mathbb{P}_\theta\Big(\max_{j=1,2,3}|\hat{m}_j - m_j| \geq \frac{gm_2}{44\max(1,g)}\Big) \leq 4\exp\bigg(-\frac{Cn\gamma^* g^2 m_2^2/\max(1,g)^2}{L^3 + \max(\tau, \sqrt{L})^3 gm_2/\max(1,g)}\bigg).$$

Proof By Proposition 23, applied with $x \geq 0$ such that

$$CL^{3/2}\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau, \sqrt{L})^3\frac{x}{n\gamma^*} = \frac{gm_2}{44\max(1,g)}$$

ie,

$$\begin{aligned}\sqrt{\frac{x}{n\gamma^*}} &= \frac{L^{3/2}}{2\max(\tau, \sqrt{L})^3}\left(\sqrt{1 + \frac{4\max(\tau, \sqrt{L})^3 gm_2}{44CL^3\max(1,g)}} - 1\right)\\ &\geq \frac{1}{44CL^{3/2}}\frac{gm_2/\max(1,g)}{\sqrt{1 + \frac{4\max(\tau,\sqrt{L})^3 gm_2}{44CL^3\max(1,g)}}},\end{aligned}$$

we obtain the result. ■

C.4 Proof of Theorem 5

In the whole proof, since $\tilde{\psi}_2$ is computed independently of the rest, we assume for convenience and without loss of generality that $\tilde{\psi}_2$ is non random and we work implicitly conditional on $\tilde{\psi}_2$. It is assumed that $\tilde{\psi}_2$ satisfies the properties stated in the Theorem 4. Since the loss function is almost-surely bounded by 1, the contribution of estimating $\tilde{\psi}_2$ to the risk is easily deduced from the Theorem 4.

Due to label switching, $\hat{\phi}_1$ may be either an estimator of ϕ_1 or $-\phi_1$, depending on the value of $\tilde{s} \coloneqq \mathrm{sgn}(\langle \psi_2, \tilde{\psi}_2 \rangle)$. In the proofs, rather than allow an arbitrary permutation, we define $p_\pm$ as an (unobserved) permutation of (p, q) and we define $\hat{p}_+, \hat{p}_-$ such that $\hat{p}_\pm$ estimates $p_\pm$. To this end, define $p_\pm = \frac{1}{2}(1 \mp \tilde{s}\phi_1)(1 - \phi_2)$ (as in Lemma 12 already) and define $\hat{p}_\pm$ accordingly:

$$\hat{p}_\pm = \tfrac{1}{2}(1 \mp \hat{\phi}_1)(1 - \hat{\phi}_2). \tag{25}$$

It is noted in Lemma 12 that we may equivalently define

$$(p_+, p_-) \coloneqq \begin{cases} (p, q) & \text{if } \tilde{s} > 0, \\ (q, p) & \text{if } \tilde{s} < 0. \end{cases}$$

Recall the definitions $g \coloneqq \phi_3 |\tilde{\mathcal{I}}| = m_2^{-1}\sqrt{4m_1^2 m_2 + m_3^2}$, $m_1 \coloneqq r(\phi)\tilde{\mathcal{I}}^2$, $m_2 \coloneqq r(\phi)\phi_2\tilde{\mathcal{I}}^2$, and $m_3 \coloneqq r(\phi)\phi_1\phi_2\phi_3\tilde{\mathcal{I}}^3$. Also recall the event Ω_n defined in Proposition 25, and proved therein to satisfy $\sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)} \mathbb{P}_\theta(\Omega_n^c) \leq 14 \exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^4\zeta^6}{L^3 + \max(\tau, \sqrt{L})^3 \delta\epsilon^2\zeta^3}\right)$ for a constant $C > 0$:

$$\Omega_n \coloneqq \left\{ \max_{j=1,2} \left| \frac{\hat{m}_j}{m_j} - 1 \right| \leq \frac{1}{2}, \ \max_{j=1,2,3} |\hat{m}_j - m_j| \leq \frac{g m_2}{44 \max(1, g)} \right\}.$$

Its definition is according to the needs of the proof of Theorem 6 which are more stringent than those of the current result. In particular, note that on Ω_n we have $\max_{j=1,2,3} |\hat{m}_j - m_j| \leq \frac{|\tilde{\mathcal{I}}|^3 r(\phi)\phi_2\phi_3}{20 \max(|\phi_1|, (1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|)}$, as a consequence of the fact that $|\phi_1| \leq 1$; this latter bound is what we will use for the current theorem.

We decompose

$$\begin{aligned} \mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2\Big) &= \mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2 \mathbf{1}_{\Omega_n^c}\Big) + \mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2 \mathbf{1}_{\Omega_n}\Big) \\ &\leq \mathbb{P}_\theta(\Omega_n^c) + \mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2 \mathbf{1}_{\Omega_n}\Big), \end{aligned}$$

We have

$$\hat{p}_\pm - p_\pm = -\frac{1}{2}(\hat{\phi}_2 - \phi_2) \mp \frac{1}{2}(\hat{\phi}_1 - \tilde{s}\phi_1) \pm \frac{\tilde{s}\phi_1}{2}(\hat{\phi}_2 - \phi_2) \mp \frac{\hat{\phi}_2}{2}(\hat{\phi}_1 - \tilde{s}\phi_1),$$

hence, using that $|\hat{\phi}_2| \leq 1$ and $|\phi_1| \leq 1$,

$$|\hat{p}_\pm - p_\pm| \leq |\hat{\phi}_1 - \phi_1| + |\hat{\phi}_2 - \phi_2|.$$

Using Lemmas 29 and 31 below and Proposition 24, we get for a constant K

$$\begin{aligned}\mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2\mathbf{1}_{\Omega_n}\Big) &\leq 2\mathbb{E}_\theta\Big(|\hat{\phi}_1 - \tilde{s}\phi_1|^2\mathbf{1}_{\Omega_n}\Big) + 2\mathbb{E}_\theta\Big(|\hat{\phi}_2 - \phi_2|^2\mathbf{1}_{\Omega_n}\Big)\\ &\leq 2\Big(\frac{53^2\max(1,g^2)}{\phi_2^4\phi_3^6|\tilde{\mathcal{I}}|^6} + \frac{16}{m_1^2}\Big)\mathbb{E}_\theta\Big(\max_{j=1,2,3}|\hat{m}_j - m_j|^2\Big)\\ &\leq 2K\Big(\frac{53^2\max(1,g^2)}{\phi_2^4\phi_3^6|\tilde{\mathcal{I}}|^6} + \frac{16}{m_1^2}\Big)\Big(\frac{L^3}{n\gamma^*} + \frac{\max(\tau,\sqrt{L})^6}{(n\gamma^*)^2}\Big).\end{aligned}$$

Therefore, there is a universal constant $B \geq 1$ such that

$$\begin{aligned}&\sup_{\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)}\mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2\mathbf{1}_{\Omega_n}\Big)\\ &\qquad\leq \frac{BL^3\max(\delta^2,\epsilon^2\zeta^2)}{\delta^2\epsilon^4\zeta^6}\frac{1}{n\gamma^*} + \frac{B\max(\tau,L)^6\max(\delta^2,\epsilon^2\zeta^2)}{\delta^2\epsilon^4\zeta^6}\frac{1}{(n\gamma^*)^2}\\ &\qquad\leq \frac{2BL^3\max(\delta^2,\epsilon^2\zeta^2)}{\delta^2\epsilon^4\zeta^6}\frac{1}{n\gamma^*},\end{aligned}$$

since $L \geq 1$ and $\sup_{\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)}\mathbb{E}_\theta\Big(|\hat{p}_\pm - p_\pm|^2\mathbf{1}_{\Omega_n}\Big) \leq 1$. Lemmas 29 and 31 therefore conclude the proof.

Lemma 29 *Suppose*

$$\max_{j=1,2}\Big|\frac{\hat{m}_j}{m_j} - 1\Big| \leq \frac{1}{2}, \quad \text{and,} \quad \max_{j=1,2,3}|\hat{m}_j - m_j| \leq \frac{|\tilde{\mathcal{I}}|^3 r(\phi)\phi_2\phi_3}{20\max(|\phi_1|,(1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|)}.$$

Then,

$$|\hat{\phi}_1 - \tilde{s}\phi_1| \leq \frac{53\max(1,\phi_3|\tilde{\mathcal{I}}|)}{\phi_2^2\phi_3^3|\tilde{\mathcal{I}}|^3}\max_{j=1,2,3}|\hat{m}_j - m_j|.$$

Proof We use the notations $\Delta_1 = \hat{m}_1 - m_1$, $\Delta_2 = (\hat{m}_2)_+ - m_2$, and $\Delta_3 = \hat{m}_3 - m_3$. Then, we define

$$\begin{aligned}\hat{v} &\coloneqq 4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2,\\ v &\coloneqq 4m_1^2m_2 + m_3^2,\\ h &\coloneqq \hat{v} - v,\\ \xi &\coloneqq 8m_1m_2\Delta_1 + 4m_1^2\Delta_2 + 8m_1\Delta_1\Delta_2 + 4m_2\Delta_1^2 + 4\Delta_1^2\Delta_2,\\ \eta &\coloneqq 2m_3\Delta_3 + \Delta_3^2.\end{aligned}$$

Lemma 30 below tells us that $|h| \leq 10\max(|\phi_1|,(1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|)|r(\phi)\phi_2\phi_3\tilde{\mathcal{I}}^3|\max_{j=1,2,3}|\Delta_j|$. Furthermore, it is seen that $\sqrt{v} = |\tilde{\mathcal{I}}|^3 r(\phi)\phi_2\phi_3 = |\tilde{\mathcal{I}}|m_2\phi_3$ (see Lemma 16) and then under the conditions of this lemma, we have $|h| \leq v/2$ and $|\Delta_3| \leq (1/2)|m_3| = (1/2)\phi_1\phi_3\tilde{\mathcal{I}}|m_2| \leq \sqrt{v}/2$. Consequently, $1 - \frac{\Delta_3^2}{(\sqrt{v+h}+\sqrt{v})^2} \geq 3/4$ and $(v+h)^{1/2}[(v+h)^{1/2}+v^{1/2}] \geq (1+\sqrt{2})v/2 \geq v$

and hence using Lemma 14

$$\begin{aligned}
|\hat{\phi}_1 - \tilde{s}\phi_1| &\leq \frac{|\phi_1\xi|}{v} + \frac{4}{3v}\Big[2|\Delta_3|(1-\phi_1^2)v^{1/2} + |\phi_1|\Delta_3^2|\xi|v^{-1} + |\Delta_3\xi|v^{-1/2}\Big] \\
&\leq \frac{28}{v}m_1^2 \max_{j=1,2}|\Delta_j| + \frac{8}{3}(1-\phi_1^2)v^{-1/2}|\Delta_3| + \frac{4}{3}|\xi|[1/2 + |\phi_1|/4] \\
&\leq 28\frac{m_1^2}{v}\max_{j=1,2}|\Delta_j| + \frac{8}{3}(1-\phi_1^2)v^{-1/2}|\Delta_3| + 56\frac{m_1^2}{v}\max_{j=1,2}|\Delta_j| \\
&\leq 42(\phi_2^2\phi_3^2\tilde{\mathcal{I}}^2)^{-1}\max_{j=1,2}|\Delta_j| + \frac{32}{3}(\phi_2^2\phi_3^3\tilde{\mathcal{I}}^3)^{-1}|\Delta_3| \\
&\leq 53(\phi_2^2\phi_3^3\tilde{\mathcal{I}}^3)^{-1}\max(\phi_3\tilde{\mathcal{I}}, 1)\max_{j=1,2,3}|\Delta_j|.
\end{aligned}$$

The conclusion follows since $x \mapsto (x)_+$ is 1-Lipschitz and thus $|\Delta_2| = |(\hat{m}_2)_+ - m_2| = |(\hat{m}_2)_+ - (m_2)_+| \leq |\hat{m}_2 - m_2|$, so that $\max_{j=1,2,3}|\Delta_j| \leq \max_{j=1,2,3}|\hat{m}_j - m_j|$. ■

Lemma 30 *Define* $v = 4m_1^2m_2 + m_3^2$, $\hat{v} = 4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2$. *Then*

$$|\hat{v} - v| \leq 10\max(|\phi_1|, (1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|)|r(\phi)\phi_2\phi_3\tilde{\mathcal{I}}^3|\max_{j=1,2,3}|\Delta_j|,$$

where $\Delta_j = \hat{m}_j - m_j$, $j = 1, 3$ *and* $\Delta_2 = (\hat{m}_2)_+ - m_2$.

Proof Define

$$\begin{aligned}
h &\coloneqq \hat{v} - v, \\
\xi &\coloneqq 8m_1m_2\Delta_1 + 4m_1^2\Delta_2 + 8m_1\Delta_1\Delta_2 + 4m_2\Delta_1^2 + 4\Delta_1^2\Delta_2, \\
\eta &\coloneqq 2m_3\Delta_3 + \Delta_3^2.
\end{aligned}$$

Note that $h = \xi + \eta$. By Lemma 14 and mimicking the proof of (Abraham et al., 2022b, Proposition 3), it is found that

$$\hat{\phi}_1 - \tilde{s}\phi_1 = \frac{\phi_1\xi + \frac{-2\Delta_3(1-\phi_1^2)v^{1/2} + \frac{\phi_1\Delta_3^2\xi}{((v+h)^{1/2}+v^{1/2})^2} - \frac{\Delta_3\xi}{(v+h)^{1/2}+v^{1/2}}}{1-\Delta_3^2/((v+h)^{1/2}+v^{1/2})^2}}{(v+h)^{1/2}[(v+h)^{1/2} + v^{1/2}]}$$

We note that the assumptions of the lemma imply that $|\Delta_j| \leq |m_j|$ for $j = 1, 2, 3$; recall also that $0 \leq m_2 = m_1 \leq |m_1|$. Thus,

$$\begin{aligned}
|\xi| &= \Big|8m_1m_2\Delta_1 + 4m_1^2\Delta_2 + 8m_1\Delta_1\Delta_2 + 4m_2\Delta_1^2 + 4\Delta_1^2\Delta_2\Big| \\
&\leq 28m_1^2\max_{j=1,2}|\Delta_j|.
\end{aligned}$$

Since $|\eta| \leq 2|m_3\Delta_3| + \Delta_3^2 \leq 3|m_3\Delta_3|$, it also follows that (recall $m_1 = r(\phi)\tilde{\mathcal{I}}^2$, $m_3 = \phi_1\phi_2\phi_3 r(\phi)\tilde{\mathcal{I}}^3$, $r(\phi) = (1/4)(1-\phi_1^2)\phi_2\phi_3^2$)

$$\begin{aligned}
|h| &\leq \big(28m_1^2 + 3|m_3|\big) \max_{j=1,2,3} |\Delta_j| \\
&= |r(\phi)\phi_2\phi_3\tilde{\mathcal{I}}^3|\Big(3|\phi_1| + \frac{28|r(\phi)\tilde{\mathcal{I}}|}{|\phi_2\phi_3|}\Big) \max_{j=1,2,3} |\Delta_j| \\
&= |r(\phi)\phi_2\phi_3\tilde{\mathcal{I}}^3|\Big(3|\phi_1| + 7(1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|\Big) \max_{j=1,2,3} |\Delta_j| \\
&\leq 10\max\big(|\phi_1|, (1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|\big)|r(\phi)\phi_2\phi_3\tilde{\mathcal{I}}^3| \max_{j=1,2,3} |\Delta_j|.
\end{aligned}$$

This concludes the proof. ■

Lemma 31 *The following bounds holds true.*

$$|\hat{\phi}_2 - \phi_2| \leq 2\min\left(1, \frac{2\max_{j=1,2}|\hat{m}_j - m_j|}{|m_1|}\right).$$

Proof We let $\Delta_1 \coloneqq \hat{m}_1 - m_1$ and $\Delta_2 \coloneqq \hat{m}_2 - m_2$. We also let $f(x) \coloneqq \max(-1, \min(x, 1))$. It is easily seen that $|f(x) - f(y)| \leq \min(2, |x-y|)$. Suppose first that $|\Delta_1| > |m_1|/2$. Then, $|\hat{\phi}_2 - \phi_2| \leq 2 \leq \min(2, \frac{4|\Delta_1|}{|m_1|})$. On the other hand, if $|\Delta_1| \leq |m_1|/2$, then, recalling that $m_2 \leq |m_1|$ we have from Lemma 14

$$\begin{aligned}
|\hat{\phi}_2 - \phi_2| &= |f(\hat{m}_2/\hat{m}_1) - f(m_2/m_1)| \\
&\leq \min\Big(2, \Big|\frac{m_2 + \Delta_2}{m_1 + \Delta_1} - \frac{m_2}{m_1}\Big|\Big) \\
&= \min\Big(2, \Big|\frac{m_1\Delta_2 - m_2\Delta_1}{m_1(m_1 + \Delta_1)}\Big|\Big) \\
&\leq \min\Big(2, \frac{2|\Delta_1| + 2|\Delta_2|}{|m_1|}\Big).
\end{aligned}$$

The conclusion follows since $x \mapsto (x)_+$ is 1-Lipschitz and thus $|\Delta_2| = |(\hat{m}_2)_+ - m_2| = |(\hat{m}_2)_+ - (m_2)_+| \leq |\hat{m}_2 - m_2|$. ■

C.5 Proof of Theorem 6

In the whole proof, since $\tilde{\psi}_2$ is computed independently of the rest, we assume for convenience and without loss of generality that $\tilde{\psi}_2$ is non random and we work implicitly conditional on $\tilde{\psi}_2$. It is assumed that $\tilde{\psi}_2$ satisfies the properties stated in the Theorem 4. The loss function is almost-surely bounded by $\check{T}^2$ so the contribution of estimating $\tilde{\psi}_2$ to the risk is easily deduced from the Theorem 4.

As in Appendix C.4, rather than allow an arbitrary permutation to account for the label-switching, we give a specific (unobserved) permutation. We recall the definitions of

the estimators of f_0 and f_1 from Section 3.5, here writing as $\check{f}_\pm$ to align with notation used in Lemma 12. We define (see also Lemma 14)

$$g \coloneqq \phi_3|\tilde{\mathcal{I}}| = \frac{\sqrt{4m_1^2m_2+m_3^2}}{m_2}, \qquad G \coloneqq \frac{m_1\psi_2}{\tilde{\mathcal{I}}}, \qquad f_\pm \coloneqq \psi_1 \pm \frac{g(1\mp\tilde{s}\phi_1)}{2m_1}G,$$

and

$$\hat{g} \coloneqq \frac{\sqrt{4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2}}{\hat{m}_2}\mathbf{1}_{\{\hat{m}_2>0\}}, \qquad \hat{G}^{\Phi_{Jk}} \coloneqq \mathbb{P}_n^{(2)}(\tilde{\psi}_2\otimes\Phi_{Jk}) - \mathbb{P}_n^{(1)}(\tilde{\psi}_2)\mathbb{P}_n^{(1)}(\Phi_{Jk}),$$
$$\hat{f}_\pm^{\Phi_{Jk}} \coloneqq \mathbb{P}_n^{(1)}(\Phi_{Jk}) \pm \frac{\hat{g}(1\mp\hat{\phi}_1)}{2\hat{m}_1}\mathbf{1}_{\{\hat{m}_1\neq 0\}}\hat{G}^{\Phi_{Jk}}.$$

Then, defining $\hat{f}_\pm^{\Psi_{jk}}$ and $\hat{G}^{\Psi_{jk}}$ correspondingly we set

$$\hat{f}_\pm \coloneqq \sum_{k=0}^{2^J-1}\hat{f}_\pm^{\Phi_{Jk}}\Phi_{Jk} + \sum_{j=J}^{J_n-1}\sum_{k=0}^{2^j-1}\hat{f}_\pm^{\Psi_{jk}}\Psi_{jk} + \sum_{j=J_n}^{\tilde{\jmath}_n}\sum_{\ell}\left(\sum_{k\in\mathfrak{B}_{j\ell}}\hat{f}_\pm^{\Psi_{jk}}\Psi_{jk}\right)\mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{S}_n\}},$$
$$\check{f}_\pm \coloneqq \max\big(0,\,\min\big(\check{T},\,\hat{f}_\pm\big)\big),$$

where $J_n \coloneqq \inf\{j\geq J \,:\, 2^j \geq \log(n)\}$, $N = 2^{J_n}$, and $\mathfrak{B}_{j\ell} \coloneqq \{k \,:\, (\ell-1)N \leq k \leq \ell N-1\}$ and $\tilde{\jmath}_n$ is the largest integer such that $2^{\tilde{\jmath}_n} \leq \frac{n}{\log(n)\tau^2}$ (recall we assume that $\tilde{\jmath}_n$ is larger than J_n) and where $\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|^2 \coloneqq \sum_{k\in\mathfrak{B}_{j\ell}}(\hat{f}_\pm^{\Psi_{jk}})^2$, $\Gamma>0$ is a tuning parameter, and

$$\hat{S}_n \coloneqq \sqrt{\frac{\log(n)}{n}}\max\Big(1,\,\frac{\hat{g}}{|\hat{m}_1|}\Big)\mathbf{1}_{\{\hat{m}_1\neq 0\}}.$$

Recall the event $\Omega_n = \Big\{\max_{j=1,2}\Big|\frac{\hat{m}_j}{m_j}-1\Big| \leq \frac{1}{2},\ \max_{j=1,2,3}|\hat{m}_j - m_j| \leq \frac{gm_2}{44\max(1,g)}\Big\}$ defined in Proposition 25 which by the proposition satisfies for a universal constant $C>0$

$$\sup_{\theta\in\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)\cap\Sigma_{\gamma^*}(L)}\mathbb{P}_\theta(\Omega_n^c) \leq 7\exp\left(-\frac{Cn\gamma^*\delta^2\epsilon^4\zeta^6}{L^3+\max(\tau,\sqrt{L})^3\delta\epsilon^2\zeta^3}\right).$$

Decompose

$$\begin{aligned}\mathbb{E}_\theta\Big(\|\check{f}_\pm - f_\pm\|_{L^2}^2\Big) &= \mathbb{E}_\theta\Big(\|\check{f}_\pm - f_\pm\|_{L^2}^2\mathbf{1}_{\Omega_n^c}\Big) + \mathbb{E}_\theta\Big(\|\check{f}_\pm - f_\pm\|_{L^2}^2\mathbf{1}_{\Omega_n}\Big)\\ &\leq \check{T}^2\mathbb{P}_\theta(\Omega_n^c) + \mathbb{E}_\theta\Big(\|\hat{f}_\pm - f_\pm\|_{L^2}^2\mathbf{1}_{\Omega_n}\Big)\end{aligned}$$

where the last line follows because $0\leq f_\pm, \check{f}_\pm \leq \check{T}$ since $\check{T}\geq L$ by assumption, and because $|\check{f}_\pm - f_\pm| \leq |\hat{f}_\pm - f_\pm|$ pointwise. The first term is included in the theorem and it remains to bound the second term. We decompose as follows (recall that $\tilde{\jmath}_n > J_n$ by assumption

and the sum over ℓ is the sum over blocks from $\ell = 0$ to $\ell = 2^j/N - 1$)

$$\begin{aligned}
\mathbb{E}_\theta\Big(\|\hat{f}_\pm - f_\pm\|_{L^2}^2 \mathbf{1}_{\Omega_n}\Big) &= \mathbb{E}_\theta\Big(\|\hat{f}_\pm^{J_n} - f_\pm^{J_n}\|_{L^2}^2 \mathbf{1}_{\Omega_n}\Big) \\
&\quad + \mathbb{E}_\theta\left(\sum_{j=J_n}^{\tilde{\jmath}_n}\sum_{\ell}\|f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|\le\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|\le 2\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\right) \\
&\quad + \mathbb{E}_\theta\left(\sum_{j=J_n}^{\tilde{\jmath}_n}\sum_{\ell}\|f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|\le\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|> 2\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\right) \\
&\quad + \mathbb{E}_\theta\left(\sum_{j=J_n}^{\tilde{\jmath}_n}\sum_{\ell}\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|\le\frac{1}{2}\Gamma\hat{S}_n\}\}}\mathbf{1}_{\Omega_n}\right) \\
&\quad + \mathbb{E}_\theta\left(\sum_{j=J_n}^{\tilde{\jmath}_n}\sum_{\ell}\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac{1}{2}\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\right) \\
&\quad + \mathbb{P}_\theta(\Omega_n)\sum_{j>\tilde{\jmath}_n}\sum_{k=0}^{2^j-1}|f_\pm^{\Psi_{jk}}|^2
\end{aligned}$$

where we have used the convention that for any function f the notation f^{J_n} stands for the projection $f_\pm^{J_n} \coloneqq \sum_{k=0}^{2^J-1} f_\pm^{\Phi_{Jk}}\Phi_{Jk} + \sum_{j=J}^{J_n-1}\sum_{k=0}^{2^j-1} f_\pm^{\Psi_{jk}}\Psi_{jk}$. Recall that $f^{\mathfrak{B}_{j\ell}}$ denotes the vector of coefficients $(\langle f, \Psi_{jk}\rangle \,:\, (j,k)\in\mathfrak{B}_{j\ell})$ and $\|\cdot\|$ the euclidean norm. We call the terms in the previous decomposition $R_1(\theta),\dots,R_6(\theta)$, respectively. To ease the notations in the proof, we also introduce the quantities

$$\hat{\omega}_\pm \coloneqq \pm\frac{\hat{g}(1\mp\hat{\phi}_1)}{\hat{m}_1}\mathbf{1}_{\{\hat{m}_1\neq 0\}}, \qquad \omega_\pm \coloneqq \pm\frac{g(1\mp\tilde{s}\phi_1)}{m_1} \tag{26}$$

and

$$S_n \coloneqq \sqrt{\frac{\log(n)}{n}}\max\Big(1,\,\frac{g}{|m_1|}\Big). \tag{27}$$

In the next subsections we prove the following bounds, uniformly over $\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R) \cap \Sigma_{\gamma^*}(L)$:

$$
\begin{aligned}
R_1(\theta) &\leq \frac{BL^2}{\delta^2\epsilon^2\zeta^2}\frac{\log(n)}{n\gamma^*} + \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2}\\
R_2(\theta) &\leq \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{R^2\delta^2\epsilon^2\zeta^2 n}\Big)^{2s_\pm/(2s_\pm+1)} + \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_\pm}\\
R_3(\theta) &\leq \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2},\\
R_4(\theta) &\leq \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2},\\
R_5(\theta) &\leq \frac{BL^2}{\Gamma^2\gamma^*}\left(\frac{R^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{R^2\delta^2\epsilon^2\zeta^2 n}\Big)^{2s_\pm/(2s_\pm+1)} + \frac{R^2}{\min(1,s_\pm)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_\pm}\right)\\
&\quad + \frac{BL^3}{\delta^2\epsilon^4\zeta^6}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2},\\
R_6(\theta) &\leq \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_\pm}.
\end{aligned}
$$

Combining will yield the theorem.

C.5.1 Control of R_1

Using Lemma 36 to control $\|\hat{f}_\pm^{J_n} - f_\pm^{J_n}\|_{L^2}$ and Proposition 37 in Section C.5.7 to control $|\hat{\omega}_\pm - \omega_\pm|$, the bounds $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and $\|G^{J_n}\|_{L^2} = |m_1|\|\psi_2^{J_n}\|_{L^2}/|\tilde{\mathcal{I}}| \leq (8/7)|m_1|$ allow us to deduce

$$
\begin{aligned}
R_1(\theta) &\coloneqq \mathbb{E}_\theta\Big(\|\hat{f}_\pm^{J_n} - f_\pm^{J_n}\|_{L^2}^2 \mathbf{1}_{\Omega_n}\Big)\\
&\leq 3\mathbb{E}_\theta\Big(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2}^2\Big) + \frac{12g^2}{m_1^2}\mathbb{E}_\theta\Big(\|\hat{G}^{J_n} - G^{J_n}\|_{L^2}^2\Big) + \frac{3\|G^{J_n}\|_{L^2}^2}{4}\mathbb{E}_\theta\Big(|\hat{\omega}_\pm - \omega_\pm|^2\mathbf{1}_{\Omega_n}\Big).
\end{aligned}
$$

ie.

$$
\begin{aligned}
R_1(\theta) \leq 3\mathbb{E}_\theta\Big(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2}^2\Big) + \frac{12g^2}{m_1^2}\mathbb{E}_\theta\Big(\|\hat{G}^{J_n} - G^{J_n}\|_{L^2}^2\Big)\\
+ \frac{3\cdot 8^2\cdot 83^2\max(1,\phi_3^2\tilde{\mathcal{I}}^2)}{4\cdot 7^2 m_2^2}\mathbb{E}_\theta\Big(\max_{j=1,2,3}|\hat{m}_j - m_j|^2\Big). \quad (28)
\end{aligned}
$$

Proposition 32 tells us that

$$
\mathbb{P}_\theta\left(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2} \geq C\sqrt{\frac{Lx}{n\gamma^*}} + C2^{J_n/2}\frac{x}{n\gamma^*}\right) \leq 24\,2^{2^{J_n}}e^{-x},
$$

hence, using that $2^{J_n} \leq 2\log(n)$ for $n \geq 2$, for a sufficient large constant $\alpha > 0$ we may apply Lemma 21 with $a = C\sqrt{2\log(n)}/\gamma^*$, $b = C\sqrt{L/\gamma^*}$, $c = 24^{2^{2\log(n)}}$ and $d^2 =$

$\alpha C^2 L \log(n)/(n\gamma^*)$

$$\begin{aligned}
&\mathbb{E}_\theta\Big(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2}^2\Big) \\
&\quad \le \alpha C^2 L \frac{\log(n)}{n\gamma^*} + \mathbb{E}_\theta\Big(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2}^2 \mathbf{1}_{\{\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2}^2 > \alpha C^2 L \log(n)/(n\gamma^*)\}}\Big) \\
&\quad \le \alpha C^2 L \frac{\log(n)}{n\gamma^*} + c\Big(d^2 + \frac{5b^2}{2n} + \frac{7a^2}{2n^2}\Big) e^{-nd^2/(2b^2+8ad)} \\
&\quad \le \alpha C^2 L \frac{\log(n)}{n\gamma^*} + C^2 24^{2\log(n)} \Big(\frac{\alpha L \log(n)}{n\gamma^*} + \frac{5L}{2n\gamma^*} + \frac{14\log(n)}{2(n\gamma^*)^2}\Big) e^{-nd^2/(2b^2+8ad)} \\
&\quad \le \alpha C^2 L \frac{\log(n)}{n\gamma^*} + C^2 24^{2\log(n)} \Big(\alpha L + \frac{5L}{2} + 7\Big) \log(n) e^{-nd^2/(2b^2+8ad)}
\end{aligned}$$

where the last line follows because $n\gamma^* \ge \tau^3 \ge 1$. Let us now study the argument of the exponential in the last display. If $2b^2 \ge 8ad$, then

$$\frac{nd^2}{2b^2 + 8ad} \ge \frac{nd^2}{4b^2} = \frac{\alpha}{4}\log(n),$$

while if $2b^2 < 8ad$, then

$$\frac{nd^2}{2b^2 + 8ad} \ge \frac{nd^2}{16ad} = \frac{n\gamma^* \sqrt{\alpha C^2 L \log(n)/(n\gamma^*)}}{16C\sqrt{2\log n}} \ge \frac{\sqrt{\alpha L}}{16\sqrt{2}} \sqrt{n\gamma^*} \ge \frac{\sqrt{\alpha}}{16\sqrt{2}} \log(n)$$

because by assumption $n\gamma^* \ge \frac{\log(n)^2}{L}$. Hence, since $L \le n$ and $\gamma^* \le 1$ it is possible to choose $\alpha > 0$ universally such that

$$\mathbb{E}_\theta\Big(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2}^2\Big) \le 2\alpha C^2 L \frac{\log(n)}{n\gamma^*}.$$

Similarly, Proposition 33 tells us that

$$\mathbb{P}_\theta\left(\|\hat{G}^{J_n} - G^{J_n}\|_{L^2} \ge CL\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau 2^{J_n/2}, \sqrt{L}2^{J_n/2}, \tau\sqrt{L})\frac{x}{n\gamma^*}\right) \le 4 \cdot 24^{2^{J_n}} e^{-x},$$

hence, for any $\alpha > 0$, using that $2^{J_n} \le 2\log(n)$ for $n \ge 2$, Lemma 21 with $a = C\tau\sqrt{2L\log(n)}/\gamma^*$, $b = CL/\sqrt{\gamma^*}$, $c = 4 \times 24^{2\log n}$, and $d^2 = \alpha C^2 L^2 \log(n)/(n\gamma^*)$ [and by remarking that $\max(\tau 2^{J_n/2}, \sqrt{L}2^{J_n/2}, \tau\sqrt{L}) \le \tau\sqrt{L}2^{J_n/2}$] yields

$$\begin{aligned}
&\mathbb{E}_\theta\Big(\|\hat{G}^{J_n} - G^{J_n}\|_{L^2}^2\Big) \\
&\quad \le \alpha C^2 L^2 \frac{\log(n)}{n\gamma^*} + c\Big(d^2 + \frac{5b^2}{2n} + \frac{7a^2}{2n^2}\Big) e^{-nd^2/(2b^2+8ad)} \\
&\quad \le \alpha C^2 L^2 \frac{\log(n)}{n\gamma^*} + 4C^2 24^{2\log(n)} \Big(\frac{\alpha L^2 \log(n)}{n\gamma^*} + \frac{5L^2}{2n\gamma^*} + \frac{14\tau^2 L\log(n)}{2(n\gamma^*)^2}\Big) e^{-nd^2/(2b^2+8ad)}.
\end{aligned}$$

Let us study the argument of the exponential in the last display. If $2b^2 \ge 8ad$, then

$$\frac{nd^2}{2b^2 + 8ad} \ge \frac{nd^2}{4b^2} = \frac{\alpha}{4}\log(n)$$

while if $2b^2 < 8ad$, then

$$\frac{nd^2}{2b^2+8ad} \geq \frac{nd^2}{16ad} = \frac{n\gamma^*\sqrt{\alpha C^2 L^2 \log(n)/(n\gamma^*)}}{16C\tau\sqrt{L}2^{J_n/2}} \geq \frac{\sqrt{\alpha L}}{32\tau}\sqrt{n\gamma^*} \geq \frac{\sqrt{\alpha}}{32}\log(n)$$

because by assumption $n\gamma^* \geq \frac{\tau^2\log(n)^2}{L}$. Since by assumption $L \leq n$ and $n\gamma^* \geq \tau^3 \geq 1$, it is possible to choose $\alpha > 0$ universally such that

$$\mathbb{E}_\theta\Big(\|\hat{G}^{J_n} - G^{J_n}\|_{L^2}^2\Big) \leq 2\alpha C^2 L^2 \frac{\log(n)}{n\gamma^*}.$$

Returning to (28) and feeding the bound for $E_\theta \max_j |\hat{m}_j - m_j|^2$ from Proposition 24, we deduce that

$$R_1(\theta) \leq 6\alpha C^2 L\Big(1+\frac{g^2 L}{m_1^2}\Big)\frac{\log(n)}{n\gamma^*} + \frac{3\cdot 83^2\cdot 40 C^2 L^3 \max(1,g^2)}{n\gamma^* m_2^2} + \frac{3\cdot 83^2\cdot 64 C^2 \max(\tau,\sqrt{L})^6}{(n\gamma^*)^2 m_2^2}.$$

Finally, we remark $\frac{g^2}{m_1^2} \leq \frac{16}{\delta^2\epsilon^2\zeta^2\tilde{\mathcal{I}}^2}$ and $\frac{\max(1,g^2)}{m_2^2} \leq \frac{16}{\delta^2\epsilon^4\zeta^4\tilde{\mathcal{I}}^4}$ by Lemma 16 and by the assumption that $\zeta \leq 1$. Thus, there exists a universal constant $B > 0$ such that

$$\sup_{\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)} R_1(\theta) \leq \frac{BL^2}{\delta^2\epsilon^2\zeta^2}\frac{\log(n)}{n\gamma^*} + \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2}.$$

C.5.2 Control of R_2

From equation (15) whenever $\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$ it is the case that $\sup_{j\geq J} 2^{2js_\pm}\sum_k |f_\pm^{\Psi_{jk}}|^2 \leq R^2$. This in particular implies that $\sum_\ell \|f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \leq R^2 2^{-2js_\pm}$. Moreover $\hat{S}_n \leq 4S_n$ on Ω_n by Proposition 38 in Section C.5.7. Then, since $J_n \leq \tilde{\jmath}_n$,

$$\begin{aligned}
R_2(\theta) &\coloneqq \mathbb{E}_\theta\Bigg(\sum_{j=J_n}^{\tilde{\jmath}_n}\sum_\ell \|f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|\leq\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|\leq 2\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\Bigg)\\
&\leq \sum_{j=J_n}^{\tilde{\jmath}_n}\sum_\ell \min\Big(\|f_\pm^{\mathfrak{B}_{j\ell}}\|^2, 8\Gamma S_n\Big)^2\\
&\leq \sum_{j=J_n}^{\tilde{\jmath}_n}\min\Bigg(\sum_\ell \|f_\pm^{\mathfrak{B}_{j\ell}}\|^2,\ \frac{2^j}{N}\cdot 64\Gamma^2 S_n^2\Bigg)\\
&\leq \sum_{j=J_n}^{\tilde{\jmath}_n}\min\Big(R^2 2^{-2js_\pm},\ \frac{2^j}{N}\cdot 64\Gamma^2 S_n^2\Big).
\end{aligned}$$

Define $A = \sup\{0 \leq j \leq \tilde{\jmath}_n : 2^{-j(s_\pm+1/2)} > 8\Gamma S_n/(R\sqrt{N})\}$, so that the first term in the minimum is the smaller exactly when $j > A$. Then we observe that $2^A < (R^2N/(64\Gamma^2 S_n^2))^{1/(2s_\pm+1)}$ and $2^{A+1} \geq \min\{(R^2N/(64\Gamma^2S_n^2))^{1/(2s_\pm+1)}, n/(\tau^2\log n)\}$ (for the latter recall that $\tilde{\jmath}$ is the

largest integer such that $2^{\tilde{\jmath}} \le n/(\tau^2 \log n)$), and we calculate

$$\begin{aligned} R_2(\theta) &\le \frac{64\Gamma^2 S_n^2}{N} \sum_{j=0}^{A} 2^j + R^2 \sum_{j=A+1}^{\infty} 2^{-2js_\pm} \\ &\le \frac{128\Gamma^2 S_n^2}{N} \left(\frac{c^2 R^2 N}{64\Gamma^2 S_n^2} \right)^{1/(2s_\pm+1)} + \frac{R^2}{1-2^{-2s_\pm}} \max\left(\frac{\tau^2 \log(n)}{n}, \left(\frac{64\Gamma^2 S_n^2}{R^2 N} \right)^{1/(2s_\pm+1)} \right)^{2s_\pm} \\ &= 2R^2 \left(\frac{64\Gamma^2 S_n^2}{R^2 N} \right)^{2s_\pm/(2s_\pm+1)} + \frac{R^2}{1-2^{-2s_\pm}} \max\left(\frac{\tau^2 \log(n)}{n}, \left(\frac{64\Gamma^2 S_n^2}{R^2 N} \right)^{1/(2s_\pm+1)} \right)^{2s_\pm}. \end{aligned}$$

Recalling that $S_n = \sqrt{(\log n)/n} \max(1, g/|m_1|)$ and $N > \log n$, we deduce that

$$\begin{aligned} R_2(\theta) \le 2R^2 &\left(\frac{64\Gamma^2 \max(1, g^2/m_1^2)}{R^2 n} \right)^{2s_\pm/(2s_\pm+1)} \\ &+ \frac{R^2}{1-2^{-2s_\pm}} \max\left(\frac{\tau^2 \log(n)}{n}, \left(\frac{64\Gamma^2 \max(1, g^2/m_1^2)}{R^2 n} \right)^{1/(2s_\pm+1)} \right)^{2s_\pm}. \end{aligned}$$

Hence, recalling that $|\tilde{\mathcal{I}}| \ge 7/8$ and the result of Lemma 16, there exists a universal constant $B > 0$ such that

$$\sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} R_2(\theta) \le \frac{BR^2}{\min(1, s_\pm)} \Big(\frac{\Gamma^2}{R^2 \delta^2 \epsilon^2 \zeta^2 n} \Big)^{2s_\pm/(2s_\pm+1)} + \frac{BR^2}{\min(1, s_\pm)} \Big(\frac{\tau^2 \log(n)}{n} \Big)^{2s_\pm}$$

C.5.3 Control of R_3

We remark that on the event $\{\|\hat{f}^{\mathfrak{B}_{j\ell}}\| \le \Gamma \hat{S}_n\} \cap \{\|f^{\mathfrak{B}_{j\ell}}\| > 2\Gamma \hat{S}_n\}$ it must that

$$\|f_\pm^{\mathfrak{B}_{j\ell}}\| \le \|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| + \|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\| \le \|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| + \frac{1}{2}\|f_\pm^{\mathfrak{B}_{j\ell}}\|$$

and thus $\|f_\pm^{\mathfrak{B}_{j\ell}}\| \le 2\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|$. Then, since $\frac{1}{4} S_n \le \hat{S}_n \le 4 S_n$ on the event Ω_n by Proposition 38 in Section C.5.7,

$$\begin{aligned} R_3(\theta) &\coloneqq \mathbb{E}_\theta \Bigg(\sum_{j=J_n}^{\tilde{\jmath}_n} \sum_\ell \|f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\| \le \Gamma \hat{S}_n\}} \mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\| > 2\Gamma \hat{S}_n\}} \mathbf{1}_{\Omega_n} \Bigg) \\ &\le 4 \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_\ell \mathbb{E}_\theta \Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\| \le \Gamma \hat{S}_n\}} \mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\| > 2\Gamma \hat{S}_n\}} \mathbf{1}_{\Omega_n} \Big) \\ &\le 4 \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_\ell \mathbb{E}_\theta \Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| > \Gamma S_n/4\}} \mathbf{1}_{\Omega_n} \Big). \end{aligned}$$

Recalling that $\hat{f}_\pm = \hat{\psi}_1 + \frac{1}{2}\hat{\omega}_\pm \hat{G}$, we define $W_1^{\mathfrak{B}_{j\ell}} \coloneqq \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\|$, $W_2^{\mathfrak{B}_{j\ell}} \coloneqq \frac{4g}{|m_1|}\|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\|$, and $W_3^{\mathfrak{B}_{j\ell}} \coloneqq \frac{1}{2}|\hat{\omega}_\pm - \omega_\pm| \|G^{\mathfrak{B}_{j\ell}}\|$, so that a direct calculation (see Lemma 36)

yields $\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|_{L^2} \leq W_1^{\mathfrak{B}_{j\ell}} + W_2^{\mathfrak{B}_{j\ell}} + W_3^{\mathfrak{B}_{j\ell}}$. We then observe, writing $\bar{W}^{\mathfrak{B}_{j\ell}} = \max(W_1^{\mathfrak{B}_{j\ell}}, W_2^{\mathfrak{B}_{j\ell}}, W_3^{\mathfrak{B}_{j\ell}})$, that

$$
\begin{aligned}
R_3(\theta) \leq\ & 4 \sum_{j=J_n}^{\tilde{j}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| > \Gamma S_n/4\}} \mathbf{1}_{\{\bar{W}^{\mathfrak{B}_{j\ell}} = W_1^{\mathfrak{B}_{j\ell}}\}} \mathbf{1}_{\Omega_n} \Big) \\
& + 4 \sum_{j=J_n}^{\tilde{j}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| > \Gamma S_n/4\}} \mathbf{1}_{\{\bar{W}^{\mathfrak{B}_{j\ell}} = W_2^{\mathfrak{B}_{j\ell}}\}} \mathbf{1}_{\Omega_n} \Big) \\
& + 4 \sum_{j=J_n}^{\tilde{j}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| > \Gamma S_n/4\}} \mathbf{1}_{\{\bar{W}^{\mathfrak{B}_{j\ell}} = W_3^{\mathfrak{B}_{j\ell}}\}} \mathbf{1}_{\Omega_n} \Big)
\end{aligned}
$$

We call these terms $R_{3,1}$, $R_{3,2}$, and $R_{3,3}$, respectively. Let us start with $R_{3,1}$. Observe that on the event $\Omega_n \cap \{\bar{W}^{\mathfrak{B}_{j\ell}} = W_1^{\mathfrak{B}_{j\ell}}\}$ we have $\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| \leq 3W_1^{\mathfrak{B}_{j\ell}}$. Therefore,

$$
R_{3,1} \leq 36 \sum_{j=J_n}^{\tilde{j}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\big(W_1^{\mathfrak{B}_{j\ell}}\big)^2 \mathbf{1}_{\{W_1^{\mathfrak{B}_{j\ell}} > \Gamma S_n/12\}} \Big)
$$

Proposition 34 in Section C.5.7 tells us that, for $n\gamma^* \geq 1/99$, there is a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$ and all $x \geq 0$

$$
\mathbb{P}_\theta\left(\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| \geq C\sqrt{\frac{Lx}{n\gamma^*}} + C2^{j/2}\frac{x}{n\gamma^*} \right) \leq 24^N e^{-x}.
$$

Then by Lemma 21 with $a = C2^{j/2}/\gamma^*$, $b = C\sqrt{L/\gamma^*}$, $c = 24^N \leq 24^{2\log(n)}$ [$n \geq 2$ so $N \leq 2\log(n)$], we find that

$$
\begin{aligned}
R_{3,1} &\leq 36 \cdot 24^N \sum_{j=J_n}^{\tilde{j}_n} \sum_{\ell} \Big(\frac{\Gamma^2 S_n^2}{144} + \frac{5C^2 L}{2n\gamma^*} + \frac{7C^2 2^j}{2(n\gamma^*)^2} \Big) \exp\Big(- \frac{n\gamma^*\Gamma^2 S_n^2/144}{2C^2 L + 8C2^{j/2}\Gamma S_n/12} \Big). \\
&\leq 36 \cdot 24^N \Big(\frac{\Gamma^2 \max(1, g^2/m_1^2)}{144} + 5C^2 Ln + \frac{14C^2 n^2}{2} \Big) \exp\Big(- \frac{n\gamma^*\Gamma^2 S_n^2/144}{2C^2 L + 8C2^{j/2}\Gamma S_n/12} \Big)
\end{aligned}
$$

where the last line follows since there are $2^j/N \leq 2^j$ blocks at each level j, and because $2^{\tilde{j}_n} \leq n$ by construction whenever $n \geq 3$, and because $n\gamma^* \geq \tau^3 \geq 1$. Let us analyse the argument of the exponential in the last display. Firstly if $8C2^{j/2}\Gamma S_n/12 \leq 2C^2 L$, it is the case that

$$
\frac{n\gamma^*\Gamma^2 S_n^2/144}{2C^2 L + 8C2^{j/2}\Gamma S_n/12} \geq \frac{n\gamma^*\Gamma^2 S_n^2}{576C^2 L} \geq \frac{\gamma^*\Gamma^2}{576C^2 L}\log(n)
$$

since $S_n = \sqrt{\log(n)/n}\max(1, g/|m_1|)$. Secondly, if $8C2^{j/2}\Gamma S_n/12 > 2C^2 L$, it is the case that for any $j \leq \tilde{j}_n$

$$
\frac{n\gamma^*\Gamma^2 S_n^2/144}{2C^2 L + 8C2^{j/2}\Gamma S_n/12} \geq \frac{n\gamma^*\Gamma S_n}{192C2^{j/2}} \geq \frac{\gamma^*\Gamma}{192C} 2^{-\tilde{j}_n/2}\sqrt{n\log(n)} \geq \frac{\gamma^*\Gamma}{192C}\log(n)
$$

since by construction $2^{\tilde{\jmath}_n} \leq \frac{n}{\tau^2 \log(n)} \leq \frac{n}{\log(n)}$. Therefore since $L \leq n$ by assumption, for any $A > 0$ there exists $c_0 > 0$ such that whenever $\Gamma \geq c_0 \max(L^{1/2}(\gamma^*)^{-1/2}, (\gamma^*)^{-1})$:

$$R_{3,1} \leq \max\Big(1, \frac{g^2}{m_1^2}\Big) n^{-A}.$$

We now control $R_{3,2}$. With the same argument as before,

$$R_{3,2} \leq 36 \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\big(W_2^{\mathfrak{B}_{j\ell}}\big)^2 \mathbf{1}_{\{W_2^{\mathfrak{B}_{j\ell}} > \Gamma S_n/12\}} \Big).$$

Proposition 35 tells us that

$$\mathbb{P}_\theta\Big(\|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| \geq CL\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau 2^{j/2}, \sqrt{L}2^{j/2}, \tau\sqrt{L})\frac{x}{n\gamma^*} \Big) \leq 4 \cdot 24^N e^{-x}.$$

Thus, applying Lemma 21 with $a = \frac{4Cg}{|m_1|\gamma^*}\tau\sqrt{L}2^{j/2}$, $b = \frac{4CLg}{|m_1|\sqrt{\gamma^*}}$, $c = 24^N$, and $d = \Gamma S_n/12$ [note that $\max(\tau 2^{j/2}, \sqrt{L}2^{j/2}, \tau\sqrt{L}) \leq \tau\sqrt{L}2^{j/2}$], we find that

$$R_{3,2} \leq 36 \cdot 24^N \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell} \left(\frac{\Gamma^2 S_n^2}{144} + \frac{10C^2L^2g^2}{n\gamma^* m_1^2} + \frac{7 \cdot 4^2 C^2 \tau^2 L 2^j g^2}{2(n\gamma^*)^2 m_1^2} \right) \times \exp\left(- \frac{n\gamma^*\Gamma^2 S_n^2/144}{\frac{8C^2L^2g^2}{m_1^2} + \frac{16C\tau\sqrt{L}2^{j/2}g}{12|m_1|}\Gamma S_n} \right)$$

ie.

$$R_{3,2} \leq 36 \cdot 24^N \max\Big(1, \frac{g^2}{m_1^2}\Big)\Big(\frac{\Gamma^2}{144} + 20C^2L^2 n + \frac{14 \cdot 4^2\tau^2 L n^2}{2}\Big) \times \exp\left(- \frac{n\gamma^*\Gamma^2 S_n^2/144}{\frac{8C^2L^2g^2}{m_1^2} + \frac{16C\tau\sqrt{L}2^{j/2}g}{12|m_1|}\Gamma S_n} \right)$$

Let us analyse the argument of the exponential in the previous display. Firstly, in the case where $\frac{16C\tau\sqrt{L}2^{j/2}g}{12|m_1|}\Gamma S_n \leq \frac{8C^2L^3g^2}{m_1^2}$,

$$\frac{n\gamma^*\Gamma^2 S_n^2/144}{\frac{8C^2L^2g^2}{m_1^2} + \frac{16C\tau\sqrt{L}2^{j/2}g}{12|m_1|}\Gamma S_n} \geq \frac{n\gamma^*\Gamma^2 S_n^2}{\frac{2304C^2L^2g^2}{m_1^2}} \geq \frac{\gamma^*\Gamma^2}{2304C^2L^2}\log(n)$$

since $S_n = \sqrt{\log(n)/n}\max(1, g/|m_1|)$. Secondly, in the case where $\frac{16C\tau\sqrt{L}2^{j/2}g}{12|m_1|}\Gamma S_n \leq \frac{8C^2L^2g^2}{m_1^2}$, for any $j \leq \tilde{\jmath}_n$

$$\frac{n\gamma^*\Gamma^2 S_n^2/144}{\frac{8C^2L^2g^2}{m_1^2} + \frac{16C\tau\sqrt{L}2^{j/2}g}{12|m_1|}\Gamma S_n} \geq \frac{n\gamma^*\Gamma S_n}{\frac{384C\tau\sqrt{L}2^{j/2}g}{|m_1|}} \geq \frac{\gamma^*\Gamma}{384C\tau\sqrt{L}} 2^{-\tilde{\jmath}_n/2}\sqrt{n\log(n)} \geq \frac{\gamma^*\Gamma}{384C\sqrt{L}}\log(n)$$

since by construction $2^{\tilde{J}_n} \leq \frac{n}{\tau^2}\log(n)$. Therefore, for any $A > 0$ there exits a constant $c_0 > 0$ such that whenever $\Gamma \geq c_0 L^{1/2}\max(L^{1/2}(\gamma^*)^{-1/2}, (\gamma^*)^{-1})$

$$R_{3,2} \leq \max\Big(1, \frac{g^2}{m_1^2}\Big) n^{-A}.$$

We now control $R_{3,3}$. With the same argument as before,

$$\begin{aligned} R_{3,3} &\leq 36 \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\big(W_3^{\mathfrak{B}_{j\ell}}\big)^2 \mathbf{1}_{\{W_3^{\mathfrak{B}_{j\ell}} > \Gamma S_n/12\}} \mathbf{1}_{\Omega_n}\Big) \\ &\leq 36 \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \mathbb{E}_\theta\Big(\big(W_3^{\mathfrak{B}_{j\ell}}\big)^2 \mathbf{1}_{\Omega_n}\Big). \end{aligned}$$

Proposition 37 in Section C.5.7 tells us that $|\hat{\omega}_\pm - \omega_\pm| \leq \frac{83\max(1,\phi_3|\tilde{\mathcal{I}}|)}{|m_1 m_2|}\max_{j=1,2,3}|\hat{m}_j - m_j|$ on the event Ω_n, hence

$$\begin{aligned} R_{3,3} &\leq \frac{9\cdot 83^2 \max(1,\phi_3^2\tilde{\mathcal{I}}^2)}{m_1^2 m_2^2} \mathbb{E}_\theta\Big(\max_{j=1,2,3} |\hat{m}_j - m_j|^2\Big) \sum_{j=J_n}^{\tilde{J}_n}\sum_{\ell} \|G^{\mathfrak{B}_{j\ell}}\|^2 \\ &\leq \frac{9\cdot 83^2 \max(1,\phi_3^2\tilde{\mathcal{I}}^2)}{m_2^2} \mathbb{E}_\theta\Big(\max_{j=1,2,3} |\hat{m}_j - m_j|^2\Big) \end{aligned}$$

because $\|G\|_{L^2} = |m_1|\|\psi_2\|_{L^2} = |m_1|$. Furthermore, by Proposition 24, we deduce

$$R_{3,3} \leq \frac{9\cdot 83^2 \cdot 40 C^2 L^3 \max(1,g^2)}{n\gamma^* m_2^2} + \frac{9\cdot 83^2\cdot 64 C^2 \max(\tau,\sqrt{L})^6 \max(1,g^2)}{(n\gamma^*)^2 m_2^2}.$$

In the end for every $A > 0$ there exists $c_0 > 0$ such that whenever the threshold constant satisfies $\Gamma \geq c_0 L^{1/2}\max(L^{1/2}(\gamma^*)^{-1/2}, (\gamma^*)^{-1})$

$$\begin{aligned} R_3(\theta) \leq 2\max\Big(1, \frac{g^2}{m_1^2}\Big) n^{-A} + \frac{9\cdot 83^2\cdot 40 C^2 L^3 \max(1,g^2)}{n\gamma^* m_2^2} \\ + \frac{9\cdot 83^2\cdot 64 C^2 \max(\tau,\sqrt{L})^6 \max(1,g^2)}{(n\gamma^*)^2 m_2^2}. \end{aligned}$$

By choosing $\beta > 0$ carefully enough, there is a universal constant $B > 0$ such that

$$\sup_{\theta\in\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)\cap\Sigma_{\gamma^*}(L)} R_3(\theta) \leq \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2}.$$

C.5.4 Control of R_4

Observe that

$$\begin{aligned}
R_4(\theta) &\coloneqq \mathbb{E}_\theta\Bigg(\sum_{j=J_n}^{\tilde{J}_n}\sum_{\ell}\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|\leq\frac{1}{2}\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\Bigg)\\
&\leq \mathbb{E}_\theta\Bigg(\sum_{j=J_n}^{\tilde{J}_n}\sum_{\ell}\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac{1}{2}\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\Bigg)\\
&\leq \mathbb{E}_\theta\Bigg(\sum_{j=J_n}^{\tilde{J}_n}\sum_{\ell}\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac{1}{8}\Gamma S_n\}}\mathbf{1}_{\Omega_n}\Bigg)
\end{aligned}$$

since $\hat{S}_n \geq S_n/4$ on the event Ω_n by Proposition 38 in Section C.5.7. From here, we see that the bounds derived for R_3 adapts mutatis mutandis by letting $\Gamma \mapsto \Gamma/2$. In the end it is found that for $\beta > 0$ chosen carefully enough

$$\sup_{\theta\in\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)\cap\Sigma_{\gamma^*}(L)} R_4(\theta) \leq \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2}.$$

C.5.5 Control of R_5

First see that, since $\hat{S}_n \geq S_n/4$ on the event Ω_n by Proposition 38,

$$\begin{aligned}
R_5(\theta) &\coloneqq \mathbb{E}_\theta\Bigg(\sum_{j=J_n}^{\tilde{J}_n}\sum_{\ell}\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\{\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{S}_n\}}\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac{1}{2}\Gamma\hat{S}_n\}}\mathbf{1}_{\Omega_n}\Bigg)\\
&\leq \sum_{j=J_n}^{\tilde{J}_n}\sum_{\ell}\mathbb{E}_\theta\Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\Omega_n}\Big)\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac{1}{8}\Gamma S_n\}}.
\end{aligned}$$

Let $W_j^{\mathfrak{B}_{j\ell}}$ be defined as in Section C.5.3. Then, by Lemma 36 in Section C.5.7,

$$\mathbb{E}_\theta\Big(\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\Omega_n}\Big) \leq 3\mathbb{E}_\theta\Big(\big(W_1^{\mathfrak{B}_{j\ell}}\big)^2\Big) + 3\mathbb{E}_\theta\Big(\big(W_2^{\mathfrak{B}_{j\ell}}\big)^2\Big) + 3\mathbb{E}_\theta\Big(\big(W_3^{\mathfrak{B}_{j\ell}}\big)^2\mathbf{1}_{\Omega_n}\Big)$$

By computations made in Section C.5.3, for any $A > 0$ we can choose $\alpha > 0$ such that

$$\begin{aligned}
\mathbb{E}_\theta\Big(\big(W_1^{\mathfrak{B}_{j\ell}}\big)^2\Big) &\leq \alpha^2C^2L\frac{\log(n)}{n\gamma^*} + \mathbb{E}_\theta\Big(\big(W_1^{\mathfrak{B}_{j\ell}}\big)^2\mathbf{1}_{\{W_1^{\mathfrak{B}_{j\ell}}>\alpha C\sqrt{L\log(n)/(n\gamma^*)}\}}\Big)\\
&\leq \alpha^2C^2L\frac{\log(n)}{n\gamma^*} + \max\Big(1,\frac{g^2}{m_1^2}\Big)2^{-\tilde{J}_n}n^{-A}\\
&\leq \frac{\alpha^2C^2LS_n^2}{\gamma^*} + \max\Big(1,\frac{g^2}{m_1^2}\Big)2^{-\tilde{J}_n}n^{-A}.
\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}_\theta\Big(\big(W_2^{\mathfrak{B}_{j\ell}}\big)^2\Big) &\le \alpha^2C^2L^2\frac{g^2\log(n)}{n\gamma^* m_1^2} + \mathbb{E}_\theta\Big(\big(W_2^{\mathfrak{B}_{j\ell}}\big)^2\mathbf{1}_{\{W_2^{\mathfrak{B}_{j\ell}}>\frac{\alpha CLg}{|m_1|}\sqrt{\log(n)/(n\gamma^*)}\}}\Big)\\ &\le \alpha^2C^2L^2\frac{g^2\log(n)}{n\gamma^* m_1^2} + \max\Big(1,\frac{g^2}{m_1^2}\Big)2^{-\tilde\jmath_n}n^{-A}\\ &\le \frac{\alpha^2C^2L^2S_n^2}{\gamma^*} + \max\Big(1,\frac{g^2}{m_1^2}\Big)2^{-\tilde\jmath_n}n^{-A}.\end{aligned}$$

Also, by computations made in Section C.5.3, we know that

$$\begin{aligned}\sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\Big(\big(W_3^{\mathfrak{B}_{j\ell}}\big)^2\mathbf{1}_{\Omega_n}\Big) \le{}& \frac{9\cdot 83^2\cdot 40C^2L^3\max(1,g^2)}{36n\gamma^* m_2^2}\\ &+\frac{9\cdot 83^2\cdot 64C^2\max(\tau,\sqrt{L})^6\max(1,g^2)}{36(n\gamma^*)^2m_2^2}.\end{aligned}$$

Consequently,

$$\begin{aligned}R_5(\theta) \le{}& \frac{6\alpha^2C^2L^2S_n^2}{\gamma^*}\sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac18\Gamma S_n\}}\\ &+\frac{27\cdot 83^2\cdot 40C^2L^3\max(1,g^2)}{36n\gamma^*m_2^2}+\frac{27\cdot 83^2\cdot 64C^2\max(\tau,\sqrt{L})^6\max(1,g^2)}{36(n\gamma^*)^2m_2^2}\\ &+2\max\Big(1,\frac{g^2}{m_1^2}\Big)n^{-A}.\end{aligned}$$

Whenever $\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)$, it is the case (recall (33)) that $\sup_{j\ge J_n}2^{2js_\pm}\sum_k|f_\pm^{\Psi_{jk}}|^2\le R^2$. This in particular implies that for all $j\ge J_n$

$$\begin{aligned}R^2 2^{-2js_\pm} &\ge \sum_\ell \|f_\pm^{\mathfrak{B}_{j\ell}}\|^2\\ &\ge \sum_\ell \|f_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac18\Gamma S_n\}}\\ &\ge \frac{\Gamma^2S_n^2}{64}\sum_\ell \mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac18\Gamma S_n\}}.\end{aligned}$$

Since there are $2^j/N$ blocks at level j, deduce that

$$\sum_\ell \mathbf{1}_{\{\|f_\pm^{\mathfrak{B}_{j\ell}}\|>\frac18\Gamma S_n\}} \le \min\Big(\frac{2^j}{N},\frac{64R^2}{\Gamma^2S_n^2}2^{-2js_\pm}\Big) = \frac{1}{\Gamma^2S_n^2}\min\Big(\frac{2^j}{N}\Gamma^2S_n^2, 64R^2 2^{-2js_\pm}\Big)$$

Therefore,

$$R_5(\theta) \leq \frac{6\alpha^2 C^2 L^2}{\Gamma^2 \gamma^*} \sum_{j=J_n}^{\tilde{j}_n} \min\Big(\frac{2^j}{N}\Gamma^2 S_n^2,\, 64R^2 2^{-2js_\pm}\Big) \\ + \frac{27\cdot 83^2 \cdot 40 C^2 L^3 \max(1,g^2)}{36 n\gamma^* m_2^2} + \frac{27\cdot 83^2 \cdot 64 C^2 \max(\tau,\sqrt{L})^6 \max(1,g^2)}{36(n\gamma^*)^2 m_2^2} \\ + 2\max\Big(1, \frac{g^2}{m_1^2}\Big) n^{-A}.$$

Then by inspecting the proof of the bound of $R_2(\theta)$ and by choosing α sufficiently large it follows immediately that there exists a universal constant $B > 0$ such that

$$\sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} R_5(\theta) \leq \frac{BL^3}{\delta^2\epsilon^4\zeta^6}\frac{1}{n\gamma^*} + \frac{B\max(\tau,\sqrt{L})^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2} \\ + \frac{BL^2}{\Gamma^2\gamma^*}\Bigg(\frac{R^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{R^2\delta^2\epsilon^2\zeta^2 n}\Big)^{2s_\pm/(2s_\pm+1)} + \frac{R^2}{\min(1,s_\pm)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_\pm}\Bigg).$$

C.5.6 Control of R_6

Whenever $\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)$, it is the case (recall equation (33)) that $\sup_{j\geq J_n} 2^{2js_\pm} \sum_k |f_\pm^{\Psi_{jk}}|^2 \leq R^2$. Therefore,

$$R_6(\theta) \coloneqq \mathbb{P}_\theta(\Omega_n) \sum_{j>\tilde{j}_n} \sum_{k=0}^{2^j-1} |f_\pm^{\Psi_{jk}}|^2 \leq R^2 \sum_{j>\tilde{j}_n} 2^{-2js_\pm} = \frac{L^2}{2^{2s_\pm}-1} 2^{-2\tilde{j}_n s_\pm} \\ \leq \frac{R^2}{2^{2s_\pm}-1}\Big(\frac{2\tau^2\log(n)}{n}\Big)^{2s_\pm}$$

because by construction $2^{\tilde{j}_n+1} > \frac{n}{\tau^2\log(n)}$. Hence, there is a universal constant $B > 0$ such that

$$\sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} R_6(\theta) \leq \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_\pm}.$$

C.5.7 Auxiliary results

Proposition 32 *Let $n\gamma^* \geq 1/99$. Then, there is a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$ and all $x \geq 0$*

$$\mathbb{P}_\theta\Bigg(\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2} \geq C\sqrt{\frac{Lx}{n\gamma^*}} + C2^{J_n/2}\frac{x}{n\gamma^*}\Bigg) \leq 24\, 2^{2^{J_n}} e^{-x}.$$

Proof The strategy is classical and consists on remarking that $\|\hat{\psi}_1^{J_n} - \psi_1^{J_n}\|_{L^2} = \sup_{u\in U}\langle \hat{\psi}_1^{J_n} - \hat{\psi}_1^{J_n}, u\rangle$ where U is the unit ball of the appropriate vector space (which has dimension

$2^J + \sum_{j=J}^{J_n-1} 2^j = 2^{J_n}$). Then, letting $\mathcal{N}$ be a (1/2)-net over U and $\pi : U \to \mathcal{N}$ mapping any point $u \in U$ to its closest element in $\mathcal{N}$, we see that

$$\begin{aligned}\|\hat\psi_1^{J_n} - \psi_1^{J_n}\|_{L^2} &= \sup_{u\in U} \langle \hat\psi_1^{J_n} - \hat\psi_1^{J_n}, u\rangle \\ &= \sup_{u\in U} \Big(\langle \hat\psi_1^{J_n} - \hat\psi_1^{J_n}, \pi(u)\rangle + \langle \hat\psi_1^{J_n} - \hat\psi_1^{J_n}, u - \pi(u)\rangle \Big) \\ &\le \max_{u\in\mathcal{N}} \langle \hat\psi_1^{J_n} - \hat\psi_1^{J_n}, u\rangle + \frac{1}{2}\|\hat\psi_1^{J_n} - \psi_1^{J_n}\|_{L^2}\end{aligned}$$

and hence

$$\|\hat\psi_1^{J_n} - \psi_1^{J_n}\|_{L^2} \le 2 \max_{u\in\mathcal{N}} \langle \hat\psi_1^{J_n} - \hat\psi_1^{J_n}, u\rangle.$$

It follows that

$$\mathbb{P}_\theta\Big(\|\hat\psi_1^{J_n} - \psi_1^{J_n}\|_{L^2} \ge 2x \Big) \le |\mathcal{N}| \max_{u\in\mathcal{N}} \mathbb{P}_\theta\Big(\langle \hat\psi_1^{J_n} - \hat\psi_1^{J_n}, u\rangle \ge x \Big)$$

The conclusion follows by Lemma 20 applied to the function $h(y) = \sum_{k=0}^{2^J-1} u_{Jk}\Phi_{Jk}(y) + \sum_{j=J}^{J_n}\sum_{k=0}^{2^j-1} u_{jk}\psi_{jk}(y)$, because $\mathbb{E}_\theta(h^2) \le L\|h\|_{L^2}^2 = L$ for every $\theta \in \Sigma_{\gamma^*}(L)$ by Lemma 17, because $\|h\|_\infty \le c2^{J_n/2}$ for a universal $c > 0$, by standard localization properties of wavelets (Giné and Nickl, 2016, Theorem 4.2.10 or Definition 4.2.14) and because $\mathcal{N}$ can be chosen so that $|\mathcal{N}| \le 24^{2^{J_n}}$ because $\mathcal{N}$ can always be chosen to have cardinality no more than $24^{2^{J_n}}$ (e.g. Giné and Nickl, 2016, Theorem 4.3.34). ■

Proposition 33 *Let $n\gamma^* \ge 1/99$ and $\|\tilde\psi_2\|_\infty \le \tau$. Then, there is a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$ and all $x \ge 0$*

$$\mathbb{P}_\theta\left(\|\hat G^{J_n} - G^{J_n}\|_{L^2} \ge CL\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau 2^{J_n/2}, \sqrt{L}2^{J_n/2}, \tau\sqrt{L})\frac{x}{n\gamma^*} \right) \le 4\cdot 24^{2^{J_n}} e^{-x}.$$

Proof We remark that $\hat G^{\Phi_{Jk}} = \mathbb{P}_n^{(2)}(\tilde\psi_2 \otimes \Phi_{Jk}) - \mathbb{P}_n^{(1)}(\tilde\psi_2)\mathbb{P}_n^{(1)}(\Phi_{Jk})$; similarly for $\hat G^{\Psi_{jk}}$. Recall that $\|\tilde\psi_2\|_\infty \le \tau$ by assumption. Hence, $\|\hat G^{J_n}\|_{L^2} \le c\tau 2^{J_n/2}$ for a universal constant $c > 0$. Similarly $\|G^{J_n}\|_{L^2} \le c\tau 2^{J_n/2}$. Hence with probability $1 \ge 1 - e^{-x}$, whenever $x > n\gamma^*$

$$\|\hat G^{J_n} - G^{J_n}\|_{L^2} \le 2c\tau 2^{J_n/2} \le CL^{3/2}\sqrt{\frac{x}{n\gamma^*}}$$

provided $C > 2c$. We now consider the case where $0 \le x \le n\gamma^*$. We decompose

$$\begin{aligned}\hat G^{J_n} - G^{J_n} = &\sum_{k=0}^{2^J-1} \Big(\mathbb{P}_n^{(2)}\big(\tilde\psi_2 \otimes \Phi_{Jk}\big) - \mathbb{E}_\theta\big(\tilde\psi_2 \otimes \Phi_{Jk}\big) \Big)\Phi_{Jk} \\ &+ \sum_{j=J}^{J_n}\sum_{k=0}^{2^j-1} \Big(\mathbb{P}_n^{(2)}\big(\tilde\psi_2 \otimes \Psi_{jk}\big) - \mathbb{E}_\theta\big(\tilde\psi_2 \otimes \Psi_{jk}\big) \Big)\Psi_{jk} \\ &- \mathbb{E}_\theta(\tilde\psi_2)\Big(\hat\psi_1^{J_n} - \psi_1^{J_n} \Big) - \psi_1^{J_n}\Big(\mathbb{P}_n^{(1)}(\tilde\psi_2) - \mathbb{E}_\theta(\tilde\psi_2) \Big) \\ &- \Big(\mathbb{P}_n^{(1)}(\tilde\psi_2) - \mathbb{E}_\theta(\tilde\psi_2) \Big)\Big(\hat\psi_1^{J_n} - \psi_1^{J_n} \Big).\end{aligned}$$

But $\|\psi_1^{J_n}\|_{L^2} \le \|\psi_1\|_{L^2} \le \max(\|f_0\|_{L^2}, \|f_1\|_{L^2})$ and $\|f_j\|_{L^2}^2 = \int_0^1 f_j^2 \le \|f_j\|_\infty \int_0^1 f_j \le L$ whenever $\theta \in \Sigma_{\gamma^*}(L)$. Thus $\|\psi_1^{J_n}\|_{L^2} \le \sqrt{L}$. Similarly by Cauchy-Schwarz' $|\mathbb{E}_\theta(\tilde\psi_2)| \le \mathbb{E}_\theta(\tilde\psi_2^2)^{1/2} \le \|\psi_1\|_\infty^{1/2}\|\tilde\psi_2\|_{L^2} \le \sqrt{L}$. Therefore, letting $v^{J_n} := \sum_{k=0}^{2^J-1} \mathbb{E}_\theta(\tilde\psi_2 \otimes \Phi_{Jk})\Phi_{Jk} + \sum_{j=J}^{J_n}\sum_{k=0}^{2^{j-1}} \mathbb{E}_\theta(\tilde\psi_2 \otimes \Psi_{jk})\Psi_{jk}$ and its empirical counterpart $\hat v^{J_n}$ defined similarly:

$$\|\hat G^{J_n} - G^{J_n}\|_{L^2} \le \|\hat v^{J_n} - v^{J_n}\|_{L^2} + \sqrt{L}\|\hat\psi_1^{J_n} - \psi_1^{J_n}\|_{L^2} + \sqrt{L}\Big|\mathbb{P}_n^{(1)}(\tilde\psi_2) - \mathbb{E}_\theta(\tilde\psi_2)\Big| + \Big|\mathbb{P}_n^{(1)}(\tilde\psi_2) - \mathbb{E}_\theta(\tilde\psi_2)\Big|\|\hat\psi_1^{J_n} - \psi_1^{J_n}\|_{L^2}.$$

Using the same ε-net argument as in the proof of Proposition 32, we find that

$$\mathbb{P}_\theta\Big(\|\hat v^{J_n} - v^{J_n}\|_{L^2} \ge CL\sqrt{\frac{x}{n\gamma^*}} + C\tau 2^{J/2}\frac{x}{n\gamma^*}\Big) \le 24^{2^{J_n}} \sup_{u\in U} \mathbb{P}_\theta\Big(\langle \hat v^{J_n} - v^{J_n}, u\rangle \ge CL\sqrt{\frac{x}{n\gamma^*}} + C\tau 2^{J/2}\frac{x}{n\gamma^*}\Big) \le 24^{2^{J_n}} e^{-x}$$

where the last inequality follows from Lemma 20 applied to the function $h(y_1, y_2) = \sum_{k=0}^{2^J-1} u_{Jk}\tilde\psi_2(y_1)\Phi_{Jk}(y_2) + \sum_{j=J}^{J_n}\sum_{k=0}^{2^j-1} u_{jk}\tilde\psi_2(y_1)\Psi_{jk}(y_1)$ which satisfies $\mathbb{E}_\theta(h^2) \le L^2\|h\|_{L^2}^2 = L^2$ for every $\theta \in \Sigma_{\gamma^*}(L)$ by Lemma 17, and $\|h\|_\infty \le c\|\tilde\psi_2\|_\infty 2^{J/2} \le c\tau 2^{J/2}$ by standard localization properties of wavelets (Giné and Nickl, 2016, Theorem 4.2.10 or Definition 4.2.14). Also by Lemma 20 applies to $h = \tilde\psi_2$,

$$\mathbb{P}_\theta\left(\Big|\mathbb{P}_n^{(1)}(\tilde\psi_2) - \mathbb{E}_\theta(\tilde\psi_2)\Big| \ge C\sqrt{\frac{Lx}{n\gamma^*}} + C\tau\frac{x}{n\gamma^*}\right) \le e^{-x}$$

and using Proposition 32

$$\mathbb{P}_\theta\left(\|\hat\psi_1^{J_n} - \psi_1^{J_n}\| \ge C\sqrt{\frac{Lx}{n\gamma^*}} + C2^{J/2}\frac{x}{n\gamma^*}\right) \le 24^{2^{J_n}} e^{-x}.$$

Therefore with probability at least $1 - (2\cdot 24^{2^{J_n}} + 1)e^{-x}$ under $\mathbb{P}_\theta$

$$\begin{aligned}\|\hat G^{J_n} - G^{J_n}\|_{L^2} &\le C\left(\sqrt{\frac{L^2x}{n\gamma^*}} + \tau 2^{J_n/2}\frac{x}{n\gamma^*}\right) + C\sqrt{L}\left(\sqrt{\frac{Lx}{n\gamma^*}} + 2^{J_n/2}\frac{x}{n\gamma^*}\right)\\ &\quad + C\sqrt{L}\left(\sqrt{\frac{Lx}{n\gamma^*}} + \tau\frac{x}{n\gamma^*}\right) + C^2\left(\sqrt{\frac{Lx}{n\gamma^*}} + 2^{J_n/2}\frac{x}{n\gamma^*}\right)\left(\sqrt{\frac{Lx}{n\gamma^*}} + \tau\frac{x}{n\gamma^*}\right)\\ &\le 3CL\sqrt{\frac{x}{n\gamma^*}} + C\Big(\tau 2^{J_n/2} + \sqrt{L}2^{J_n/2} + \tau\sqrt{L} + CL\Big)\frac{x}{n\gamma^*}\\ &\quad + C^2\Big(\tau\sqrt{L} + 2^{J_n/2}\sqrt{L}\Big)\frac{x^{3/2}}{(n\gamma^*)^{3/2}} + C^2\tau 2^{J_n/2}\frac{x^2}{(n\gamma^*)^2}.\end{aligned}$$

The conclusion follows since $x \le n\gamma^*$ which implies that the last two terms are bounded by the second term, and the $Lx/(n\gamma^*)$ part of second term is bounded by the first term. ■

Proposition 34 *Let $n\gamma^* \geq 1/99$. Then, there is a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$, all $j \geq J_n$, all ℓ, and all $x \geq 0$,*

$$\mathbb{P}_\theta\left(\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| \geq C\sqrt{\frac{Lx}{n\gamma^*}} + C2^{j/2}\frac{x}{n\gamma^*} \right) \leq 24^N e^{-x}.$$

Proof The proof is identical to Proposition 32. (Note the vector $\psi_1^{\mathfrak{B}_{j\ell}}$ is in $\mathbb{R}^N$, where ψ_1^Φ was in $\mathbb{R}^{2^{J_n}}$.) ■

Proposition 35 *Let $n\gamma^* \geq 1/99$. Then, there is a universal constant $C > 0$ such that for all $\theta \in \Sigma_{\gamma^*}(L)$, all $j \geq J_n$, all ℓ, and all $x \geq 0$*

$$\mathbb{P}_\theta\left(\|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| \geq CL\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau 2^{j/2}, \sqrt{L}2^{j/2}, \tau\sqrt{L})\frac{x}{n\gamma^*} \right) \leq 4 \cdot 24^N e^{-x}.$$

Proof The proof is identical to Proposition 33. ■

Lemma 36 *On the event Ω_n, for all $j \geq J_n$ and all ℓ:*

$$\|\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}}\| \leq \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| + \frac{4g\|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\|}{|m_1|} + \frac{|\hat{\omega}_\pm - \omega_\pm|\|G^{\mathfrak{B}_{j\ell}}\|}{2},$$

and similarly for $\|\hat{f}_\pm^{J_n} - f_\pm^{J_n}\|_{L^2}$.

Proof Trivially,

$$\hat{f}_\pm^{\mathfrak{B}_{j\ell}} - f_\pm^{\mathfrak{B}_{j\ell}} = \hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}} + \frac{\hat{\omega}_\pm}{2}\left(\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\right) + \frac{\hat{\omega}_\pm - \omega_\pm}{2}G^{\mathfrak{B}_{j\ell}}.$$

The conclusion follows since on Ω_n, Proposition 38 implies that $\hat{g} \leq 2g$ and $|\hat{m}_1| \geq |m_1|/2 > 0$. (Recall also that $|\phi_1| \leq 1$.) ■

Proposition 37 *On the event Ω_n*

$$|\hat{\omega}_\pm - \omega_\pm| \leq \frac{83\max(1, \phi_3|\tilde{\mathcal{I}}|)}{|m_1 m_2|}\max_{j=1,2,3}|\hat{m}_j - m_j|.$$

Proof On Ω_n we have $\hat{g} \le 2g$ by Proposition 38 to follow, and note that $|\hat{m}_1| \ge |m_1|/2 > 0$. Consequently, by straightforward computations, using Lemmas 29 and 39,

$$\begin{aligned}
|\hat{\omega}_\pm - \omega_\pm| &= \Big|\frac{1}{m_1}(\hat{g}-g)(1 \mp \hat{\phi}_1) + \frac{g}{m_1}(1 \mp \hat{\phi}_1 - (1 \mp \tilde{s}\phi_1)) + \hat{g}(1 \mp \hat{\phi}_1)(\frac{1}{\hat{m}_1} - \frac{1}{m_1})\Big| \\
&\le \frac{2|\hat{g}-g|}{|m_1|} + \frac{g|\hat{\phi}_1 - \tilde{s}\phi_1|}{|m_1|} + \frac{8g|\hat{m}_1 - m_1|}{m_1^2} \\
&\le \left(\frac{22\max(1,\phi_3|\tilde{\mathcal{I}}|)}{|m_1 m_2|} + \frac{53\max(1,\phi_3|\tilde{\mathcal{I}}|)g}{|m_1|\phi_2^2\phi_3^3|\tilde{\mathcal{I}}|^3} + \frac{8g}{m_1^2}\right)\max_{j=1,2,3}|\hat{m}_j - m_j| \\
&\le \frac{83\max(1,\phi_3|\tilde{\mathcal{I}}|)}{|m_1 m_2|}\max_{j=1,2,3}|\hat{m}_j - m_j|
\end{aligned}$$

because $m_2 = \frac{1}{4}(1-\phi_1^2)\phi_2^2\phi_3^2\tilde{\mathcal{I}}^2$, because $g = \phi_3|\tilde{\mathcal{I}}|$, and because $m_2 = m_1\phi_2 \le |m_1|$. ■

Proposition 38 *On the event Ω_n, we have $|\frac{\hat{g}}{g} - 1| \le \frac{1}{2}$. Consequently, $\frac{1}{4}S_n \le \hat{S}_n \le 4S_n$ and $|\hat{\omega}_\pm| \le 8|g/m_1|$ on Ω_n.*

Proof It suffices to remark that

$$\frac{gm_2}{44\max(1,g)} \le \frac{gm_2}{20\max(|\phi_1|,(1-\phi_1^2)\phi_3|\tilde{\mathcal{I}}|)},$$

since $-1 \le \phi_1 \le 1$, so that Lemma 39 applies. Replacing $\max_j|\hat{m}_j - m_j|$ by its bound $gm_2/44\max(1,g)$ on the event Ω_n yields the result for $\hat{g}$. For S_n, recalling the definitions $S_n = \sqrt{(\log n)/n}\max(1, g/|m_1|)$, $\hat{S}_n = \sqrt{(\log n)/n}\max(1,\hat{g}/|\hat{m}_1|)\mathbb{1}\{\hat{m}_1 \ne 0\}$ and inserting the bounds $g/2 \le \hat{g} \le 2g$, $|m_1|/2 \le \hat{m}_1 \le 2|m_1|$ yields the bounds for $\hat{S}_n$. ■

Lemma 39 *Suppose*

$$\max_{j=1,2}\Big|\frac{\hat{m}_j}{m_j} - 1\Big| \le \frac{1}{2}, \quad \text{and,} \quad \max_{j=1,2,3}|\hat{m}_j - m_j| \le \frac{m_2 g}{20\max(|\phi_1|,(1-\phi_1^2)g)}$$

Then,

$$|\hat{g} - g| \le \frac{22\max(1,g)}{m_2}\max_{j=1,2,3}|\hat{m}_j - m_j|.$$

Recall that $g = \phi_3|\tilde{\mathcal{I}}|$ and $m_2 = \phi_2 r(\phi)\tilde{\mathcal{I}}^2$, so that the conditions of Lemma 39 match those of Lemma 29.

Proof We let $\Delta_1 = \hat{m}_1 - m_1$, $\Delta_2 = (\hat{m}_2)_+ - m_2$, $\Delta_3 = \hat{m}_3 - m_3$, $\hat{v} \coloneqq 4\hat{m}_1^2(\hat{m}_2)_+ + \hat{m}_3^2$, $v \coloneqq 4m_1^2 m_2 + m_3^2$, and $h \coloneqq \hat{v} - v$. Then, since $\hat{m}_2 \ge m_2/2 > 0$ under the assumption of

the lemma

$$\begin{aligned}\hat{g} - g &= \frac{\sqrt{v+h}}{m_2+\Delta_2} - \frac{\sqrt{v}}{m_2} \\ &= \frac{\sqrt{v+h}-\sqrt{v}}{m_2+\Delta_2} - \frac{\Delta_2\sqrt{v}}{m_2(m_2+\Delta_2)} \\ &= \frac{h}{(\sqrt{v+h}+\sqrt{v})(m_2+\Delta_2)} - \frac{\Delta_2\sqrt{v}}{m_2(m_2+\Delta_2)}.\end{aligned}$$

Hence it must be that

$$|\hat{g} - g| \leq \frac{2|h|}{m_2\sqrt{v}} + \frac{2\sqrt{v}}{m_2^2}|\Delta_2|.$$

Lemma 30, together with the fact that $|\phi_1| \leq 1$, tells us that

$$|h| \leq 10\max(1, \phi_3|\tilde{\mathcal{I}}|)|r(\phi)\phi_2\phi_3\tilde{\mathcal{I}}^3| \max_{j=1,2,3}|\Delta_j|$$

and $\sqrt{v} = |\tilde{\mathcal{I}}|^3 r(\phi)\phi_2\phi_3 = |\tilde{\mathcal{I}}|m_2\phi_3 \leq m_2\max(1, \phi_3|\tilde{\mathcal{I}}|)$, thus

$$|\hat{g} - g| \leq \frac{20\max(1, \phi_3|\tilde{\mathcal{I}}|)}{m_2} \max_{j=1,2,3}|\Delta_j| + \frac{2\max(1, \phi_3|\tilde{\mathcal{I}}|)}{m_2}|\Delta_2|$$

concluding the proof. ■

C.6 Proof of Theorem 8

In the whole proof, since $\tilde{\psi}_2$ is computed independently of the rest, we assume for convenience and without loss of generality that $\tilde{\psi}_2$ is non random and we work implicitly conditional on $\tilde{\psi}_2$. It is assumed that $\tilde{\psi}_2$ satisfies the properties stated in the Theorem 4. The loss function is almost-surely bounded by $\check{T}^2$ so the contribution of estimating $\tilde{\psi}_2$ to the risk is easily deduced from the Theorem 4.

C.6.1 Definitions and rationale

To avoid issues with the non-identifiability, we once again define $p_\pm$ and $f_\pm$ as in Lemma 12. The starting point of the proof is to remark that $f_\pm$ can be rewritten as

$$f_\pm = \left[\frac{2\psi_1}{1 \pm \tilde{s}\phi_1}\right] + \left[-\left(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}\psi_1 \mp \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1}G\right)\right].$$

Then each of the two functions in brackets in the previous display is estimated separately using block-thresholded wavelets estimators. The population mother coefficients are defined as

$$\alpha_\pm^{\Psi_{jk}} := \frac{2\psi_1^{\Psi_{jk}}}{1 \pm \tilde{s}\phi_1}, \qquad \beta_\pm^{\Psi_{jk}} := -\left(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}\psi_1^{\Psi_{jk}} \mp \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1}G^{\Psi_{jk}}\right)$$

and the corresponding empirical versions are

$$\hat{\alpha}_{\pm}^{\Psi_{jk}} \coloneqq \frac{2\hat{\psi}_1^{\Psi_{jk}}}{1 \pm \hat{\phi}_1} \mathbf{1}_{\{\hat{\phi}_1 \neq \mp 1\}}, \qquad \hat{\beta}_{\pm}^{\Psi_{jk}} \coloneqq -\left(\frac{1 \mp \hat{\phi}_1}{1 \pm \hat{\phi}_1} \mathbf{1}_{\{\hat{\phi}_1 \neq \mp 1\}} \hat{\psi}_1^{\Psi_{jk}} \mp \frac{\hat{g}(1 \mp \hat{\phi}_1)}{2\hat{m}_1} \mathbf{1}_{\{\hat{m}_1 \neq 0\}} \hat{G}^{\Psi_{jk}} \right).$$

Then, the untruncated estimators can be rewritten as (here $\hat{f}_{\pm}^{\Phi_{Jk}}$ are the father coefficients that were defined in the beginning of Section C.5)

$$\begin{aligned}
\hat{f}_{\pm}^{R} \coloneqq & \sum_{k=0}^{2^{J_n}-1} \hat{f}_{\pm}^{\Phi_{Jk}} \Phi_{J_n k} + \sum_{j=J}^{J_n-1} \sum_{k=0}^{2^j-1} \hat{f}_{\pm}^{\Psi_{jk}} \Psi_{jk} \\
& + \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell=0}^{2^j/N-1} \left(\sum_{k \in \mathfrak{B}_{j\ell}} \hat{\alpha}_{\pm}^{\Psi_{jk}} \Psi_{jk} \right) \mathbf{1}_{\{\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\| > \Gamma\sqrt{\log(n)/n}\}} \\
& + \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell=0}^{2^j/N-1} \left(\sum_{k \in \mathfrak{B}_{j\ell}} \hat{\beta}_{\pm}^{\Psi_{jk}} \Psi_{jk} \right) \mathbf{1}_{\{\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\| > \Gamma\hat{T}_n\}}
\end{aligned}$$

while the truncated versions are

$$\check{f}_{\pm}^{R} \coloneqq \max\big(0,\, \min\big(\check{T},\, \hat{f}_{\pm}^{R}\big)\big).$$

C.6.2 Decomposition of the error

We define auxiliary events

$$\Xi_n^{(1)} \coloneqq \Big\{ \forall j = J_n, \ldots, \tilde{\jmath}_n,\ \forall \ell,\ \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| \leq c_0 \Gamma \sqrt{\log(n)/n} \Big\},$$

and

$$\Xi_n^{(2)} \coloneqq \Big\{ \forall j = J_n, \ldots, \tilde{\jmath}_n,\ \forall \ell,\ \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| \leq c_1 \Gamma \sqrt{\log(n)/n} \Big\}$$

. We let Ξ_n denote the intersection of both of these events. Then by the same argument used in Section C.5

$$\mathbb{E}_\theta\big(\|\check{f}_{\pm}^{R} - f_{\pm}\|_{L^2}^2\big) \leq \check{T}^2\big(\mathbb{P}_\theta(\Omega_n^c) + \mathbb{P}_\theta(\Xi_n^c)\big) + \mathbb{E}_\theta\big(\|\hat{f}_{\pm}^{R} - f_{\pm}\|_{L^2}^2 \mathbf{1}_{\Omega_n \cap \Xi_n}\big).$$

The probability of the event Ω_n^c is bounded in Proposition 25, while the probability of Ξ_n^c is bounded in Lemma 40 (to follow). We bound the remaining term by decomposing it into several terms. For this matter, we introduce the events

$$E_{j\ell} \coloneqq \left\{ \max_{j=1,2,3} |\hat{m}_j - m_j| \|G^{\mathfrak{B}_{j\ell}}\| \leq c_2 |m_1 m_2| \Gamma T_n / \max(1, g) \right\}$$

and we decompose

$$
\begin{aligned}
\mathbb{E}_\theta\big(\|\hat f_\pm^R - f_\pm\|_{L^2}^2 \mathbf{1}_{\Omega_n\cap\Xi_n}\big) &= \mathbb{E}_\theta\Big(\|\hat f_\pm^{J_n} - f_\pm^{J_n}\|_{L^2}^2 \mathbf{1}_{\Omega_n\cap\Xi_n}\Big) \\
&+ \sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\Big(\big\|\hat\alpha_\pm^{\mathfrak{B}_{j\ell}} + \hat\beta_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\big\|^2 \\
&\qquad\qquad \times \mathbf{1}_{\Omega_n\cap\Xi_n}\mathbf{1}_{n\|\hat\psi_1^{\mathfrak{B}_{j\ell}}\|^2>\Gamma^2\log(n)}\mathbf{1}_{\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat T_n}\mathbf{1}_{E_{j\ell}^c}\Big) \\
&+ \sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\Big(\|\hat\alpha_\pm^{\mathfrak{B}_{j\ell}} + \hat\beta_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\|^2 \\
&\qquad\qquad \times \mathbf{1}_{\Omega_n\cap\Xi_n}\mathbf{1}_{n\|\hat\psi_1^{\mathfrak{B}_{j\ell}}\|^2>\Gamma^2\log(n)}\mathbf{1}_{\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat T_n}\mathbf{1}_{E_{j\ell}}\Big) \\
&+ \sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\Big(\|\hat\alpha_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\|^2 \\
&\qquad\qquad \times \mathbf{1}_{\Omega_n\cap\Xi_n}\mathbf{1}_{n\|\hat\psi_1^{\mathfrak{B}_{j\ell}}\|^2>\Gamma^2\log(n)}\mathbf{1}_{\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}}\|\le\Gamma\hat T_n}\mathbf{1}_{E_{j\ell}}\mathbf{1}_{\|\psi_1^{\mathfrak{B}_{j\ell}}\|>\frac{g(1\pm\tilde s\phi_1)}{|m_1|}\|G^{\mathfrak{B}_{j\ell}}\|}\Big) \\
&+ \sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\Big(\|\hat\alpha_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\|^2 \\
&\qquad\qquad \times \mathbf{1}_{\Omega_n\cap\Xi_n}\mathbf{1}_{n\|\hat\psi_1^{\mathfrak{B}_{j\ell}}\|^2>\Gamma^2\log(n)}\mathbf{1}_{\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}}\|\le\Gamma\hat T_n}\mathbf{1}_{E_{j\ell}}\mathbf{1}_{\|\psi_1^{\mathfrak{B}_{j\ell}}\|\le\frac{g(1\pm\tilde s\phi_1)}{|m_1|}\|G^{\mathfrak{B}_{j\ell}}\|}\Big) \\
&+ \sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\big(\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\Omega_n\cap\Xi_n}\mathbf{1}_{n\|\hat\psi_1^{\mathfrak{B}_{j\ell}}\|^2\le\Gamma^2\log(n)}\mathbf{1}_{\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat T_n}\mathbf{1}_{E_{j\ell}}\big) \\
&+ \sum_{j=J_n}^{\tilde\jmath_n}\sum_\ell \mathbb{E}_\theta\big(\|\alpha_\pm^{\mathfrak{B}_{j\ell}} + \beta_\pm^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\Omega_n\cap\Xi_n}\mathbf{1}_{n\|\hat\psi_1^{\mathfrak{B}_{j\ell}}\|^2\le\Gamma^2\log(n)}\mathbf{1}_{\|\hat\beta_\pm^{\mathfrak{B}_{j\ell}}\|\le\Gamma\hat T_n}\mathbf{1}_{E_{j\ell}}\big) \\
&+ \sum_{j>\tilde\jmath_n}\sum_k |f_\pm^{\Psi_{jk}}|^2\mathbb{P}_\theta(\Omega_n\cap\Xi_n)
\end{aligned}
$$

where we have used the same convention as in Section C.5 to define $\hat f_\pm^{J_n}$ and $f_\pm^{J_n}$. We call $R_1(\theta),\dots,R_8(\theta)$, respectively, each of the terms of the previous right hand side. In the next subsections, after stating a couple of preliminary results, we prove the following bounds,

uniformly over $\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)$ and for a universal constant $B > 0$:

$$
\begin{aligned}
R_1(\theta) &\le \frac{BL^2}{\delta^2\epsilon^2\zeta^2}\frac{\log(n)}{n\gamma^*} + \frac{BL^3}{\delta^2\epsilon^4\zeta^4}\frac{1}{n\gamma^*} + \frac{B\max(\tau,L)^6}{\delta^2\epsilon^4\zeta^4}\frac{1}{(n\gamma^*)^2}.\\
R_2(\theta) &\le \frac{B}{\delta^2\epsilon^4\zeta^4}\Big(\frac{L^3}{n\gamma^*} + \frac{\max(\tau,L)^6}{(n\gamma^*)^2}\Big).\\
R_3(\theta) &\le \frac{BR^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}.\\
R_4(\theta) &\le \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{nR^2\delta^2}\Big)^{2s_\pm/(2s_\pm+1)} + \frac{BR^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}.\\
R_5(\theta) &\le \frac{B}{\delta^2\epsilon^4\zeta^4}\Big(\frac{L^3}{n\gamma^*} + \frac{\max(\tau,L)^6}{(n\gamma^*)^2}\Big) + \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{nR^2\delta^2}\Big)^{2s_\pm/(2s_\pm+1)}\\
&\quad + \frac{BR^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}\\
R_6(\theta) &\le \frac{BR^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}.\\
R_7(\theta) &\le \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{nR^2\delta^2}\Big)^{2s_\pm/(2s_\pm+1)} + \frac{R^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{BR^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}.\\
R_8(\theta) &\le \frac{BR^2}{\min(1,s_\pm)}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_\pm}.
\end{aligned}
$$

C.6.3 Preliminary computations

Lemma 40 *For all $A > 0$ and for all choice of $c_0, c_1 > 0$ there exists a constant $\beta_0 > 0$ such that if $\Gamma \ge \beta\max(\frac{L}{\sqrt{\gamma^*}}, \frac{\sqrt{L}}{\tau\gamma^*})$ with $\beta \ge \beta_0$ then*

$$\mathbb{P}_\theta(\Xi_n^c) \le n^{-A}.$$

Proof By a union bound,

$$
\begin{aligned}
\mathbb{P}_\theta\big((\Xi_n^{(1)})^c\big) &\le \sum_{j=J_n}^{\tilde\jmath_n}\sum_{\ell}\mathbb{P}_\theta\Big(\|\hat\psi_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| > c_0\Gamma\sqrt{\log(n)/n}\Big)\\
&\le \frac{2^{\tilde\jmath_n+1}}{N}\max_{j\le\tilde\jmath_n}\max_{\ell}\mathbb{P}_\theta\Big(\|\hat\psi_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| > c_0\Gamma\sqrt{\log(n)/n}\Big)\\
&\le n\max_{j\le\tilde\jmath_n}\max_{\ell}\mathbb{P}_\theta\Big(\|\hat\psi_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| > c_0\Gamma\sqrt{\log(n)/n}\Big).
\end{aligned}
$$

Then choose $x = B\log(n)$ for some $B > 0$ to be chosen accordingly. Observe that for all $j \le \tilde\jmath_n$ (recall $L \ge 1$)

$$
\begin{aligned}
C\sqrt{\frac{Lx}{n\gamma^*}} + C2^{j/2}\frac{x}{n\gamma^*} &\le \frac{C\sqrt{BL}}{\sqrt{\gamma^*}}\cdot\sqrt{\frac{\log(n)}{n}} + C\sqrt{\frac{n}{\log(n)\tau^2}}\frac{B\log(n)}{n\gamma^*}\\
&\le \frac{C\sqrt{B} + CB}{\beta}\Gamma\sqrt{\log(n)/n}.
\end{aligned}
$$

Hence by choosing $c_0 = (C\sqrt{B} + CB)/\beta$ we deduce from the Proposition 34 that

$$\mathbb{P}_\theta\big((\Xi_n^{(1)})^c\big) \leq 24^N n^{-B+1}.$$

The probability of $\Xi_n^{(2)}$ is bounded similarly, remarking that for $x = B\log(n)/n$ we have for all $j \leq \tilde{j}_n$

$$\begin{aligned}
&CL\sqrt{\frac{x}{n\gamma^*}} + C\max(\tau 2^{j/2}, \sqrt{L}2^{j/2}, \tau\sqrt{L})\frac{x}{n\gamma^*} \\
&\qquad \leq \frac{CL\sqrt{B}}{\sqrt{\gamma^*}}\sqrt{\frac{\log(n)}{n}} + \frac{CB}{\gamma^*}\max\left(\sqrt{\frac{n}{\log(n)}}, \frac{\sqrt{L}}{\tau}\sqrt{\frac{\log(n)}{n}}, \tau\sqrt{L}\right)\frac{\log(n)}{n} \\
&\qquad \leq \frac{CL\sqrt{B}}{\sqrt{\gamma^*}}\sqrt{\frac{\log(n)}{n}} + \frac{CB\sqrt{L}}{\gamma^*\tau}\sqrt{\frac{\log(n)}{n}} \\
&\qquad \leq \frac{C\sqrt{B} + CB}{\beta}\Gamma\sqrt{\log(n)/n},
\end{aligned}$$

where the third line is true because by assumption $1 \leq 2^{J_n} \leq 2^{\tilde{j}_n} \leq \frac{n}{\log(n)\tau^2}$ and hence $\tau \leq \sqrt{n/\log(n)}$ necessarily. We then deduce from Proposition 35 that

$$\mathbb{P}_\theta\big((\Xi_n^{(1)})^c\big) \leq 4 \cdot 24^N n^{-B+1}$$

which concludes the proof by taking B sufficiently large. ■

Lemma 41 *On the event* Ω_n

$$\frac{1}{2} \leq \frac{1 \pm \tilde{s}\phi_1}{1 \pm \hat{\phi}_1} \leq 2, \qquad \text{and,} \qquad \frac{1}{2} \leq \frac{1 \mp \tilde{s}\phi_1}{1 \mp \hat{\phi}_1} \leq 2.$$

Proof Observe that

$$\frac{1 \pm \tilde{s}\phi_1}{1 \pm \hat{\phi}_1} = \frac{1}{1 + \frac{\hat{\phi}_1 - \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}}$$

But on the event Ω_n, by Lemma 29

$$\begin{aligned}
|\hat{\phi}_1 - \tilde{s}\phi_1| &\leq \frac{53\max(1,g)}{gm_2} \cdot \frac{1-\phi_1^2}{4} \cdot \max_{j=1,2,3}|\hat{m}_j - m_j| \\
&\leq \frac{53}{4 \cdot 44}(1 \pm \tilde{s}\phi_1)(1 \mp \tilde{s}\phi_1) \\
&\leq \frac{1 \pm \tilde{s}\phi_1}{2}
\end{aligned}$$

which proves the first claim. The second claim is proven similarly. ■

Lemma 42 *On the event* Ω_n *we have* $\hat{m}_1 \neq 0$ *and* $\hat{\phi}_1^2 \neq 1$.

Proof The fact that $\hat{m}_1 \neq 0$ follows immediately from the definition of Ω_n. The fact that $\hat{\phi}_1^2 \neq 1$ follows from Lemma 41 (either one of the two inequalities would not hold if $\hat{\phi}_1^2 = 1$). ■

The next Proposition controls the empirical threshold $\hat{T}_n$ in term of its population version defined as

$$T_n \coloneqq \sqrt{\frac{\log(n)}{n}} \max\Big(1, \frac{g}{|m_1|}, \frac{1}{1-\phi_1^2}\Big).$$

Lemma 43 *On the event Ω_n, $\frac{1}{4}T_n \leq \hat{T}_n \leq 4T_n$.*

Proof Notice that $T_n = \max\big(S_n, \frac{\sqrt{\log(n)/n}}{1-\phi_1^2}\big)$. Thus, in view of Proposition 38 it is enough to show that $\frac{1-\phi_1^2}{4} \leq 1 - \hat{\phi}_1^2 \leq 4(1-\phi_1^2)$. But,

$$1-\hat{\phi}_1^2 = (1 \pm \hat{\phi}_1)(1 \mp \hat{\phi}_1) = \frac{1 \pm \hat{\phi}_1}{1 \pm s\tilde{\phi}_1}\frac{1 \mp \hat{\phi}_1}{1 \mp \tilde{s}\phi_1}(1 \mp \tilde{s}\phi_1)(1 \pm \tilde{s}\phi_1) = \frac{1 \pm \hat{\phi}_1}{1 \pm s\tilde{\phi}_1}\frac{1 \mp \hat{\phi}_1}{1 \mp \tilde{s}\phi_1}(1-\phi_1^2).$$

Thus the conclusion follows from Lemma 41. ■

Lemma 44 *It is possible to choose c_0, c_1, c_2 such that on the event $E_{j\ell} \cap \Xi_n \cap \Omega_n$:*

1. $\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\| > \Gamma\hat{T}_n \implies \|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| > \frac{1}{32}\Gamma T_n$;
2. $\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\| \leq \Gamma\hat{T}_n \implies \|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \leq 32\Gamma T_n$;
3. $\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\| > \Gamma\sqrt{\log(n)/n} \implies \|\psi_1^{\mathfrak{B}_{j\ell}}\| > \frac{1}{2}\Gamma\sqrt{\log(n)/n}$;
4. $\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\| \leq \Gamma\sqrt{\log(n)/n} \implies \|\psi_1^{\mathfrak{B}_{j\ell}}\| \leq \frac{3}{2}\Gamma\sqrt{\log(n)/n}$.

Proof Before proving the items, we first remark that we never have $\hat{\phi}_1^2 = 1$ nor $\hat{m}_1 = 0$ on the event Ω_n thanks to Lemma 42.

We establish Item 1. Notice that

$$\begin{aligned}
\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\| > \Gamma\hat{T}_n &\iff \Big\|\frac{1 \mp \hat{\phi}_1}{1 \pm \hat{\phi}_1}\hat{\psi}_1^{\mathfrak{B}_{j\ell}} \mp \frac{\hat{g}(1 \mp \hat{\phi}_1)}{2\hat{m}_1}\hat{G}^{\mathfrak{B}_{j\ell}}\Big\| > \Gamma\hat{T}_n \\
&\iff \Big\|\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}\hat{\psi}_1^{\mathfrak{B}_{j\ell}} \mp \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1}\frac{\hat{g}(1 \mp \tilde{s}\phi_1)}{2\hat{m}_1}\hat{G}^{\mathfrak{B}_{j\ell}}\Big\| > \frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}\frac{1 \pm \hat{\phi}_1}{1 \mp \hat{\phi}_1}\Gamma\hat{T}_n \\
&\implies \Big\|\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}\hat{\psi}_1^{\mathfrak{B}_{j\ell}} \mp \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1}\frac{\hat{g}(1 \mp \tilde{s}\phi_1)}{2\hat{m}_1}\hat{G}^{\mathfrak{B}_{j\ell}}\Big\| > \frac{1}{16}\Gamma T_n
\end{aligned}$$

ie.

$$\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\| > \Gamma\hat{T}_n \implies$$
$$\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| > \frac{1}{16}\Gamma T_n - \frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1}\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| - \Big\|\frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1}\frac{\hat{g}}{\hat{m}_1}\hat{G}^{\mathfrak{B}_{j\ell}} - \frac{g}{m_1}G^{\mathfrak{B}_{j\ell}}\Big\|$$

where we have used Lemmas 41 and 43. But on the event $E_{j\ell} \cap \Xi_n \cap \Omega_n$

$$\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| \leq \frac{(1 \mp s\tilde{\phi}_1)^2}{1 - \phi_1^2} \cdot c_0 \Gamma \sqrt{\log(n)/n} \leq c_0 \Gamma T_n$$

and

$$\begin{aligned}
&\Big\| \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{\hat{g}}{\hat{m}_1} \hat{G}^{\mathfrak{B}_{j\ell}} - \frac{g}{m_1} G^{\mathfrak{B}_{j\ell}} \Big\| \\
&\leq \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{\hat{g}}{|\hat{m}_1|} \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| + \Big| \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{\hat{g}}{|\hat{m}_1|} - \frac{g}{m_1} \Big| \|G^{\mathfrak{B}_{j\ell}}\| \\
&\leq \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{\hat{g}}{|\hat{m}_1|} \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| \\
&\qquad + \left(\frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{|\hat{g} - g|}{|\hat{m}_1|} + \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{g|\hat{m}_1 - m_1|}{|\hat{m}_1 m_1|} + \frac{g}{|m_1|} \Big| \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} - 1 \Big| \right) \|G^{\mathfrak{B}_{j\ell}}\| \\
&\leq \frac{8g}{|m_1|} \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| + \left(\frac{4|\hat{g} - g|}{|m_1|} + \frac{4g|\hat{m}_1 - m_1|}{m_1^2} + \frac{g|\hat{\phi}_1 - \tilde{s}\phi_1|}{(1 - \phi_1^2)|m_1|} \right) \|G^{\mathfrak{B}_{j\ell}}\|
\end{aligned}$$

where the last line holds true on Ω_n by Lemmas 38 and 41. Therefore by Lemmas 29 and 39, there is a universal constant $C > 0$ such that

$$\begin{aligned}
&\Big\| \frac{1 \pm \hat{\phi}_1}{1 \pm \tilde{s}\phi_1} \frac{\hat{g}}{\hat{m}_1} \hat{G}^{\mathfrak{B}_{j\ell}} - \frac{g}{m_1} G^{\mathfrak{B}_{j\ell}} \Big\| \\
&\qquad \leq \frac{8g}{|m_1|} \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| + \frac{C \max(1, g)}{|m_1 m_2|} \max_{j=1,2,3} |\hat{m}_j - m_j| \|G^{\mathfrak{B}_{j\ell}}\| \leq (8c_1 + Cc_2) \Gamma T_n
\end{aligned}$$

on the event $E_{j\ell} \cap \Xi_n \cap \Omega_n$ by definitions of these events. Therefore by choosing c_0, c_1, c_2 small enough, the Item 1 claim follows. The proof of the Item 2 is nearly identical. Items 3 and 4 are immediate from the definition of Ξ_n provided $c_0 \leq 1/2$. ■

In the next we make use of the symbol $\lesssim$ to denote inequalities that are valid up to a universal multiplicative constant. Furthermore, since $\hat{m}_1 \neq 0$ and $\hat{\phi}_1^2 \neq 1$ on the event Ω_n thanks to Lemma 42, and since all the terms we wish to control are conditional on Ω_n, we will assume throughout the rest of the proof that $\hat{m}_1 \neq 0$ and $\hat{\phi}_1^2 \neq 1$ without justification.

C.6.4 Control of R_1

This has already been done in Section C.5.1. We recall the result:

$$\sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} R_1(\theta) \leq \frac{BL^2}{\delta^2 \epsilon^2 \zeta^2} \frac{\log(n)}{n\gamma^*} + \frac{BL^3}{\delta^2 \epsilon^4 \zeta^4} \frac{1}{n\gamma^*} + \frac{B \max(\tau, L)^6}{\delta^2 \epsilon^4 \zeta^4} \frac{1}{(n\gamma^*)^2}.$$

C.6.5 Control of R_2

$$\begin{aligned}
&\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}\mathbf{1}_{\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\|>\Gamma\sqrt{\log(n)/n}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\mathbf{1}_{\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{T}_n}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \\
&\quad=\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}-\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}\mathbf{1}_{\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\|\leq\Gamma\sqrt{\log(n)/n}}-\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\mathbf{1}_{\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\|\leq\Gamma\hat{T}_n}\| \\
&\quad\leq\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|+\frac{2\Gamma\sqrt{\log(n)/n}}{1\pm\hat{\phi}_1}+\Gamma\hat{T}_n \\
&\quad\leq\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|+8\Gamma T_n
\end{aligned}$$

on the event Ω_n by Lemmas 41 and 43. Furthermore, letting $\hat{f}_{\pm}^{\mathfrak{B}_{j\ell}}$ and $f_{\pm}^{\mathfrak{B}_{j\ell}}$ as defined in Section C.5, it is easily seen that

$$\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}=\hat{f}_{\pm}^{\mathfrak{B}_{j\ell}}-f_{\pm}^{\mathfrak{B}_{j\ell}}.$$

Hence by Lemma 36, on the event $\Xi_n\cap\Omega_n$,

$$\begin{aligned}
\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|&\leq c_0\Gamma\sqrt{\log(n)/n}+\frac{4g}{|m_1|}c_1\Gamma\sqrt{\log(n)/n}+\frac{1}{2}|\hat{\omega}_{\pm}-\omega_{\pm}|\|G^{\mathfrak{B}_{j\ell}}\| \\
&\leq(c_0+4c_1)\Gamma T_n+\frac{1}{2}|\hat{\omega}_{\pm}-\omega_{\pm}|\|G^{\mathfrak{B}_{j\ell}}\| \\
&\leq(c_0+4c_1)\Gamma T_n+\frac{41.5\max(1,g)}{|m_1m_2|}\max_{j=1,2,3}|\hat{m}_j-m_j|\|G^{\mathfrak{B}_{j\ell}}\|
\end{aligned}\tag{29}$$

where the last line follows by Proposition 37. Deduce from the definition of $E_{j\ell}$ that on the event $E_{j\ell}^c\cap\Xi_n\cap\Omega_n$ we must have

$$\begin{aligned}
&\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}}\mathbf{1}_{\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\|>\Gamma\sqrt{\log(n)/n}}+\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\mathbf{1}_{\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}}\|>\Gamma\hat{T}_n}-\alpha_{\pm}^{\mathfrak{B}_{j\ell}}-\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \\
&\qquad\leq\Big(\frac{8+c_0+4c_1}{c_2}+41.5\Big)\frac{\max(1,g)}{|m_1m_2|}\max_{j=1,2,3}|\hat{m}_j-m_j|\|G^{\mathfrak{B}_{j\ell}}\|.
\end{aligned}$$

From this we obtain the estimate

$$\begin{aligned}
R_2(\theta)&\lesssim\frac{\max(1,g)^2}{m_1^2m_2^2}\mathbb{E}_\theta\Big(\max_{j=1,2,3}|\hat{m}_j-m_j|^2\Big)\sum_{j\geq J_n}\sum_{\ell}\|G^{\mathfrak{B}_{j\ell}}\|^2 \\
&\lesssim\frac{\max(1,g)^2}{m_2^2}\Big(\frac{C^2L^3}{n\gamma^*}+\frac{C^2\max(\tau,L)^6}{(n\gamma^*)^2}\Big)
\end{aligned}$$

where the last line follows from Proposition 24. Therefore we deduce from Lemma 16 that

$$\sup_{\theta\in\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)\cap\Sigma_{\gamma^*}(L)}R_2(\theta)\lesssim\frac{1}{\delta^2\epsilon^4\zeta^4}\Big(\frac{L^3}{n\gamma^*}+\frac{\max(\tau,L)^6}{(n\gamma^*)^2}\Big).$$

C.6.6 Control of R_3

By equation (29) and the definition of $E_{j\ell}$, it is found that on the event $E_{j\ell} \cap \Xi_n \cap \Omega_n$,

$$\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}} + \hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}} - \alpha_{\pm}^{\mathfrak{B}_{j\ell}} - \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \leq (c_0 + 2c_1 + 41.5c_2)\Gamma T_n.$$

Then we deduce from Lemma 44 that

$$R_3(\theta) \lesssim \Gamma^2 T_n^2 \sum_{j=J_n}^{\tilde{\jmath}_n} \sum_{\ell} \mathbf{1}_{\{\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| > \frac{1}{32}\Gamma T_n\}}.$$

Noting $\beta_{\pm} = -\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1} f_{\mp}$ and mimicking the proof in Section C.5.5, it is found that

$$\sum_{\ell} \mathbf{1}_{\{\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| > \frac{1}{32}\Gamma T_n\}} \leq \min\left(\frac{2^j}{N}, \Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2 \frac{R^2 2^{-2js_{\mp}}}{\Gamma^2 T_n^2}\right) \tag{30}$$

Letting $A = \sup\{0 \leq j \leq \tilde{\jmath}_n : 2^{-j(s_{\mp}+1/2)} > \frac{\Gamma T_n}{R\sqrt{N}} \frac{1\pm s\tilde{\phi}_1}{1\mp\tilde{s}\phi_1}\}$ it is found that

$$\begin{aligned}
R_3(\theta) &\lesssim \Gamma^2 T_n^2 \sum_{j=0}^{A} \frac{2^j}{N} + \Gamma^2 T_n^2 \sum_{j>A} \Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2 \frac{R^2 2^{-2js_{\mp}}}{\Gamma^2 T_n^2} \\
&\lesssim \frac{\Gamma^2 T_n^2}{N} 2^A + \Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2 R^2 \frac{2^{-2As_{\mp}}}{2^{2s_{\mp}} - 1} \\
&\lesssim \frac{\Gamma^2 T_n^2}{N} \left(\Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2 \frac{R^2 N}{\Gamma^2 T_n^2}\right)^{1/(2s_{\mp}+1)} \\
&\quad + \Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2 R^2 \frac{1}{2^{2s_{\mp}} - 1} \left(\Big(\frac{1\pm\tilde{s}\phi_1}{1\mp\tilde{s}\phi_1}\Big)^2 \frac{\Gamma^2 T_n^2}{R^2 N}\right)^{2s_{\mp}/(2s_{\mp}+1)} \\
&\lesssim \frac{R^2}{\min(1, s_{\mp})} \Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^{2/(2s_{\mp}+1)} \Big(\frac{\Gamma^2 T_n^2}{R^2 N}\Big)^{2s_{\mp}/(2s_{\mp}+1)}.
\end{aligned}$$

It follows using the definition of T_n and $\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)$ together with Lemma 16 (recall that $\zeta \leq 1$ by assumption) that

$$\sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) \cap \Sigma_{\gamma^*}(L)} R_3(\theta) \lesssim \frac{R^2}{\min(1, s_{\mp})} \frac{1}{\delta^2} \Big(\frac{\Gamma^2}{R^2 n \epsilon^2 \zeta^2}\Big)^{2s_{\mp}/(2s_{\mp}+1)}.$$

C.6.7 Control of R_4

When $\|\psi_1^{\mathfrak{B}_{j\ell}}\| > \frac{g(1\pm\tilde{s}\phi_1)}{|m_1|}\|G^{\mathfrak{B}_{j\ell}}\|$

$$\begin{aligned}
\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| &= \Big\|\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\psi_1^{\mathfrak{B}_{j\ell}} \mp \frac{g(1\mp\tilde{s}\phi_1)}{2m_1} G^{\mathfrak{B}_{j\ell}}\Big\| \\
&\geq \frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\|\psi_1^{\mathfrak{B}_{j\ell}}\| - \frac{g(1\mp\tilde{s}\phi_1)}{2|m_1|}\|G^{\mathfrak{B}_{j\ell}}\| \\
&\geq \frac{1}{2}\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\|\psi_1^{\mathfrak{B}_{j\ell}}\|.
\end{aligned}$$

Consequently,

$$\begin{aligned}\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}} - \alpha_{\pm}^{\mathfrak{B}_{j\ell}} - \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| &= \Big\| \frac{2}{1 \pm \hat{\phi}_1} \hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \frac{2}{1 \pm \tilde{s}\phi_1} \psi_1^{\mathfrak{B}_{j\ell}} + \Big(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \psi_1^{\mathfrak{B}_{j\ell}} \mp \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1} G^{\mathfrak{B}_{j\ell}} \Big) \Big\| \\ &= \Big\| \frac{2(\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}})}{1 \pm \hat{\phi}_1} + \Big(\frac{1 \mp \hat{\phi}_1}{1 \pm \hat{\phi}_1} \psi_1^{\mathfrak{B}_{j\ell}} \mp \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1} G^{\mathfrak{B}_{j\ell}} \Big) \Big\| \\ &\leq \frac{2\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\|}{1 \pm \hat{\phi}_1} + \frac{1 \mp \hat{\phi}_1}{1 \pm \hat{\phi}_1} \|\psi_1^{\mathfrak{B}_{j\ell}}\| + \frac{g(1 \mp \tilde{s}\phi_1)}{2|m_1|} \|G^{\mathfrak{B}_{j\ell}}\|\end{aligned}$$

Then on the event $\Xi_n \cap \Omega_n$, by Lemma 41

$$\begin{aligned}\|\hat{\alpha}_{\pm}^{\mathfrak{B}_{j\ell}} - \alpha_{\pm}^{\mathfrak{B}_{j\ell}} - \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| &\leq \frac{4c_0\Gamma\sqrt{\log(n)/n}}{1 \pm \tilde{s}\phi_1} + \frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big(4\|\psi_1^{\mathfrak{B}_{j\ell}}\| + \frac{g(1 \pm \tilde{s}\phi_1)}{2|m_1|} \|G^{\mathfrak{B}_{j\ell}}\| \Big) \\ &\leq \frac{4c_0\Gamma\sqrt{\log(n)/n}}{1 \pm \tilde{s}\phi_1} + 5\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \|\psi_1^{\mathfrak{B}_{j\ell}}\| \\ &\leq \frac{4c_0\Gamma\sqrt{\log(n)/n}}{1 \pm \tilde{s}\phi_1} + 10\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|.\end{aligned}$$

Deduce from Lemma 44 that

$$R_4(\theta) \lesssim \frac{\Gamma^2 \log(n)/n}{(1 \pm \tilde{s}\phi_1)^2} \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \mathbf{1}_{\{\|\psi_1^{\mathfrak{B}_{j\ell}}\| > \frac{1}{2}\Gamma\sqrt{\log(n)/n}\}} + \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \leq 32\Gamma T_n\}}.$$

Observe that $2\psi_1 = (1 + \tilde{s}\phi_1)f_+ + (1 - \tilde{s}\phi_1)f_-$. Therefore, for all $j \geq J_n$

$$\begin{aligned}\sum_k |\psi_1^{\Psi_{jk}}|^2 &\leq \frac{(1 + \tilde{s}\phi_1)^2}{2} \sum_k |f_+^{\Psi_{jk}}|^2 + \frac{(1 - \tilde{s}\phi_1)^2}{2} \sum_k |f_-^{\Psi_{jk}}|^2 \\ &\leq R^2 \frac{(1 + \tilde{s}\phi_1)^2 2^{-2js_+} + (1 - \tilde{s}\phi_1)^2 2^{-2js_-}}{2}, \end{aligned} \tag{31}$$

whenever $\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)$ (recall equation (15)). Deduce that (see also Section C.5.5)

$$\begin{aligned}\sum_{\ell} \mathbf{1}_{\{\|\psi_1^{\mathfrak{B}_{j\ell}}\| > \frac{1}{2}\Gamma\sqrt{\log(n)/n}\}} &\leq \min\left(\frac{2^j}{N}, \frac{2nR^2\big((1 + \tilde{s}\phi_1)^2 2^{-2js_+} + (1 - \tilde{s}\phi_1)^2 2^{-2js_-}\big)}{\Gamma^2 \log(n)} \right) \\ &\leq \frac{1}{2} \min\Big(\frac{2^j}{N}, \frac{4nR^2(1 + \tilde{s}\phi_1)^2 2^{-2js_+}}{\Gamma^2 \log(n)} \Big) \\ &\quad + \frac{1}{2} \min\Big(\frac{2^j}{N}, \frac{4nR^2(1 - \tilde{s}\phi_1)^2 2^{-2js_-}}{\Gamma^2 \log(n)} \Big)\end{aligned}$$

by convexity of $x \mapsto \min(2^j/N, x)$. Deduce that,

$$\begin{aligned}
&\frac{\Gamma^2 \log(n)/n}{(1 \pm \tilde{s}\phi_1)^2} \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \mathbf{1}_{\{\|\psi_1^{\mathfrak{B}_{j\ell}}\| > \frac{1}{2}\Gamma\sqrt{\log(n)/n}\}} \\
&\quad \lesssim \frac{\Gamma^2}{n(1 \pm \tilde{s}\phi_1)^2} \Big(\frac{nR^2(1+\tilde{s}\phi_1)^2}{\Gamma^2} \Big)^{1/(2s_+ +1)} \\
&\qquad + \frac{1}{2^{2s_+}-1} \frac{R^2(1+\tilde{s}\phi_1)^2}{(1 \pm \tilde{s}\phi_1)^2} \Big(\frac{\Gamma^2}{nR^2(1+\tilde{s}\phi_1)^2} \Big)^{2s_+/(2s_+ +1)} \\
&\qquad + \frac{\Gamma^2}{n(1 \pm \tilde{s}\phi_1)^2} \Big(\frac{nR^2(1-\tilde{s}\phi_1)^2}{\Gamma^2} \Big)^{1/(2s_- +1)} \\
&\qquad + \frac{1}{2^{2s_-}-1} \frac{R^2(1-\tilde{s}\phi_1)^2}{(1 \pm \tilde{s}\phi_1)^2} \Big(\frac{\Gamma^2}{nR^2(1-\tilde{s}\phi_1)^2} \Big)^{2s_-/(2s_- +1)}.
\end{aligned}$$

That is,

$$\begin{aligned}
&\frac{\Gamma^2 \log(n)/n}{(1 \pm \tilde{s}\phi_1)^2} \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \mathbf{1}_{\{\|\psi_1^{\mathfrak{B}_{j\ell}}\| > \frac{1}{2}\Gamma\sqrt{\log(n)/n}\}} \\
&\quad \lesssim \frac{R^2}{\min(1, s_+)} \Big(\frac{1+\tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^2 \Big(\frac{\Gamma^2}{nR^2(1+\tilde{s}\phi_1)^2} \Big)^{2s_+/(2s_+ +1)} \\
&\qquad + \frac{R^2}{\min(1, s_-)} \Big(\frac{1-\tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^2 \Big(\frac{\Gamma^2}{nR^2(1-\tilde{s}\phi_1)^2} \Big)^{2s_-/(2s_- +1)}.
\end{aligned}$$

Regarding the remaining term, recall that $\beta_\pm = -\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} f_\mp$ and observe that

$$\begin{aligned}
\sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\{\|\beta_\pm^{\mathfrak{B}_{j\ell}}\| \le 32\Gamma T_n\}} &\lesssim \sum_{j=J_n}^{\tilde{J}_n} \min\Big(\sum_{\ell} \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|^2, \frac{2^j \Gamma^2 T_n^2}{N} \Big) \\
&\lesssim \sum_{j=J_n}^{\tilde{J}_n} \min\Big(\sum_{\ell} \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|^2, \frac{2^j \Gamma^2 T_n^2}{N} \Big) \\
&\lesssim \sum_{j=J_n}^{\tilde{J}_n} \min\Big(R^2 \Big(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^2 2^{-2js_\mp}, \frac{2^j \Gamma^2 T_n^2}{N} \Big) \\
&\lesssim \frac{R^2}{\min(1, s_\mp)} \Big(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^{2/(2s_\mp +1)} \Big(\frac{\Gamma^2 T_n^2}{R^2 N} \Big)^{2s_\mp/(2s_\mp +1)} \qquad (32)
\end{aligned}$$

where the last line follows from the estimate in (30) and subsequent iterates. In the end,

$$\begin{aligned}
R_4(\theta) &\lesssim \frac{R^2}{\min(1, s_+)} \Big(\frac{1+\tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^2 \Big(\frac{\Gamma^2}{nR^2(1+\tilde{s}\phi_1)^2} \Big)^{2s_+/(2s_+ +1)} \\
&\quad + \frac{R^2}{\min(1, s_-)} \Big(\frac{1-\tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^2 \Big(\frac{\Gamma^2}{nR^2(1-\tilde{s}\phi_1)^2} \Big)^{2s_-/(2s_- +1)} \\
&\quad + \frac{R^2}{\min(1, s_\mp)} \Big(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \Big)^{2/(2s_\mp +1)} \Big(\frac{\Gamma^2 T_n^2}{R^2 N} \Big)^{2s_\mp/(2s_\mp +1)}.
\end{aligned}$$

Taking the suprema of each terms, with the help of Lemma 16 it is found that

$$\begin{aligned}\sup_{\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)} R_4(\theta) \lesssim{}& \frac{R^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{nR^2\delta^2}\Big)^{2s_\pm/(2s_\pm+1)}\\ &+\frac{R^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{nR^2}\Big)^{2s_\mp/(2s_\mp+1)}\\ &+\frac{R^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}.\end{aligned}$$

Namely,

$$\begin{aligned}\sup_{\theta\in\Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)} R_4(\theta) \lesssim{}& \frac{R^2}{\min(1,s_\pm)}\Big(\frac{\Gamma^2}{nR^2\delta^2}\Big)^{2s_\pm/(2s_\pm+1)}\\ &+\frac{R^2}{\min(1,s_\mp)}\frac{1}{\delta^2}\Big(\frac{\Gamma^2}{R^2n\epsilon^2\zeta^2}\Big)^{2s_\mp/(2s_\mp+1)}.\end{aligned}$$

C.6.8 Control of R_5

When $\|\psi_1^{\mathfrak{B}_{j\ell}}\| \leq \frac{g(1\pm\tilde{s}\phi_1)}{|m_1|}\|G^{\mathfrak{B}_{j\ell}}\|$,

$$\begin{aligned}\|\hat{\alpha}_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\| &\leq \|\hat{\alpha}_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}}\| + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|\\ &= \Big\|\frac{2}{1\pm\hat{\phi}_1}\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \frac{2}{1\pm\tilde{s}\phi_1}\psi_1^{\mathfrak{B}_{j\ell}}\Big\| + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|\\ &\leq \frac{2}{1\pm\hat{\phi}_1}\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| + 2\|\psi_1^{\mathfrak{B}_{j\ell}}\|\Big|\frac{1}{1\pm\hat{\phi}_1} - \frac{1}{1\pm\tilde{s}\phi_1}\Big| + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|\\ &\leq \frac{2}{1\pm\hat{\phi}_1}\|\hat{\psi}_1^{\mathfrak{B}_{j\ell}} - \psi_1^{\mathfrak{B}_{j\ell}}\| + 2\|\psi_1^{\mathfrak{B}_{j\ell}}\|\frac{|\hat{\phi}_1 - \tilde{s}\phi_1|}{(1\pm\hat{\phi}_1)(1\pm\tilde{s}\phi_1)} + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|\end{aligned}$$

So by Lemmas 29 and 41, it holds on the event $E_{j\ell}\cap\Xi_n\cap\Omega_n$

$$\begin{aligned}&\|\hat{\alpha}_\pm^{\mathfrak{B}_{j\ell}} - \alpha_\pm^{\mathfrak{B}_{j\ell}} - \beta_\pm^{\mathfrak{B}_{j\ell}}\|\\ &\quad\leq \frac{4c_0\Gamma\sqrt{\log(n)/n}}{1\pm\tilde{s}\phi_1 l} + \frac{800\max(1,g)}{\phi_2^2\phi_3^2\tilde{\mathcal{I}}^2 g}\frac{\max_{j=1,2,3}|\hat{m}_j - m_j|}{(1\pm\tilde{s}\phi_1)^2}\|\psi_1^{\mathfrak{B}_{j\ell}}\| + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|\\ &\quad\leq \frac{4c_0\Gamma\sqrt{\log(n)/n}}{1\pm\tilde{s}\phi_1} + \frac{800\max(1,g)}{\phi_2^2\phi_3^2\tilde{\mathcal{I}}^2}\frac{\max_{j=1,2,3}|\hat{m}_j - m_j|}{|m_1|(1\pm\tilde{s}\phi_1)}\|G^{\mathfrak{B}_{j\ell}}\| + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|\\ &\quad\leq \frac{4c_0\Gamma\sqrt{\log(n)/n}}{1\pm\tilde{s}\phi_1} + \frac{800\max(1,g)}{|m_1m_2|}\max_{j=1,2,3}|\hat{m}_j - m_j|\|G^{\mathfrak{B}_{j\ell}}\| + \|\beta_\pm^{\mathfrak{B}_{j\ell}}\|.\end{aligned}$$

From here, it is seen that an upper bound on the supremum of R_5 is obtained by adding the bounds obtained on R_2 together with the bound on R_4, eventually up to a universal multiplicative constant.

C.6.9 Control of R_6

$$\begin{aligned}
&\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}} - \alpha_{\pm}^{\mathfrak{B}_{j\ell}} - \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \\
&\quad \le \|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}} - \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| + \|\alpha_{\pm}^{\mathfrak{B}_{j\ell}}\| \\
&\quad = \Big\| \frac{1 \mp \hat{\phi}_1}{1 \pm \hat{\phi}_1} \hat{\psi}_1^{\mathfrak{B}_{j\ell}} \mp \frac{\hat{g}(1 \mp \hat{\phi}_1)}{2\hat{m}_1} \hat{G}^{\mathfrak{B}_{j\ell}} - \Big(\frac{1 \mp \tilde{s}\phi_1}{1 \pm \tilde{s}\phi_1} \psi_1^{\mathfrak{B}_{j\ell}} \mp \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1} G^{\mathfrak{B}_{j\ell}} \Big) \Big\| \\
&\qquad + \frac{2}{1 \pm \tilde{s}\phi_1} \|\psi_1^{\mathfrak{B}_{j\ell}}\| \\
&\quad \le \frac{3}{1 \pm \tilde{s}\phi_1} \|\psi_1^{\mathfrak{B}_{j\ell}}\| + \frac{1}{1 \pm \hat{\phi}_1} \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\| + \Big\| \frac{\hat{g}(1 \mp \hat{\phi}_1)}{2\hat{m}_1} \hat{G}^{\mathfrak{B}_{j\ell}} - \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1} G^{\mathfrak{B}_{j\ell}} \Big\|
\end{aligned}$$

but by Proposition 37 on the event Ω_n we have

$$\begin{aligned}
\Big\| \frac{\hat{g}(1 \mp \hat{\phi}_1)}{2\hat{m}_1} \hat{G}^{\mathfrak{B}_{j\ell}} - \frac{g(1 \mp \tilde{s}\phi_1)}{2m_1} G^{\mathfrak{B}_{j\ell}} \Big\| &= |\hat{\omega}_{\mp}| \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| + |\hat{\omega}_{\mp} - \omega_{\mp}| \|G^{\mathfrak{B}_{j\ell}}\| \\
&\lesssim \frac{g}{|m_1|} \|\hat{G}^{\mathfrak{B}_{j\ell}} - G^{\mathfrak{B}_{j\ell}}\| \\
&\quad + \frac{\max(1, g)}{|m_1 m_2|} \max_{j=1,2,3} |\hat{m}_j - m_j| \|G^{\mathfrak{B}_{j\ell}}\|.
\end{aligned}$$

Therefore on the event $E_{j\ell} \cap \Xi_n \cap \Omega_n$

$$\begin{aligned}
\|\hat{\beta}_{\pm}^{\mathfrak{B}_{j\ell}} - \alpha_{\pm}^{\mathfrak{B}_{j\ell}} - \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| &\lesssim \frac{\|\psi_1^{\mathfrak{B}_{j\ell}}\| + \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\|}{1 \pm \tilde{s}\phi_1} + \frac{g}{|m_1|} c_1 \Gamma \sqrt{\log(n)/n} + c_2 \Gamma T_n \\
&\le \frac{\|\psi_1^{\mathfrak{B}_{j\ell}}\| + \|\hat{\psi}_1^{\mathfrak{B}_{j\ell}}\|}{1 \pm \tilde{s}\phi_1} + (c_1 + c_2) \Gamma T_n.
\end{aligned}$$

Deduce by Lemma 44 that

$$R_6(\theta) \lesssim \Gamma^2 T_n^2 \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \mathbf{1}_{\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| > \frac{1}{32} \Gamma T_n}.$$

Therefore, $R_6(\theta)$ admits the same upper bound as $R_3(\theta)$, eventually up to a universal multiplicative factor.

C.6.10 Control of R_7

$$\|\alpha_{\pm}^{\mathfrak{B}_{j\ell}} + \beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \le \|\alpha_{\pm}^{\mathfrak{B}_{j\ell}}\| + \|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| = \frac{2}{1 \pm \tilde{s}\phi_1} \|\psi_1^{\mathfrak{B}_{j\ell}}\| + \|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|$$

Therefore, we obtain from Lemma 44 that

$$\begin{aligned}
R_7(\theta) \le \frac{2}{(1 \pm \tilde{s}\phi_1)^2} &\sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \|\psi_1^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\|\psi_1^{\mathfrak{B}_{j\ell}}\| \le \frac{3}{2} \Gamma \sqrt{\log(n)/n}} \\
&+ 2 \sum_{j=J_n}^{\tilde{J}_n} \sum_{\ell} \|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|^2 \mathbf{1}_{\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\| \le 32 \Gamma T_n}
\end{aligned}$$

From equation (31),

$$\begin{aligned}
&\frac{2}{(1\pm\tilde{s}\phi_1)^2}\sum_{j=J_n}^{\tilde{j}_n}\sum_{\ell}\|\psi_1^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\|\psi_1^{\mathfrak{B}_{j\ell}}\|\leq\frac{3}{2}\Gamma\sqrt{\log(n)/n}}\\
&\quad\leq\frac{2}{(1\pm\tilde{s}\phi_1)^2}\sum_{j=J_n}^{\tilde{j}_n}\min\left(\frac{9\Gamma^2\log(n)}{4n}\frac{2^j}{N},\sum_{\ell}\|\psi_1^{\mathfrak{B}_{j\ell}}\|^2\right)\\
&\quad\lesssim\frac{1}{(1\pm\tilde{s}\phi_1)^2}\frac{\Gamma^2\log(n)}{n}\sum_{j=J_n}^{\tilde{j}_n}\min\left(\frac{2^j}{N},nR^2\frac{(1+\tilde{s}\phi_1)^2 2^{-2js_+}+(1-\tilde{s}\phi_1)^2 2^{-2js_-}}{\Gamma^2\log(n)}\right)
\end{aligned}$$

Then deduce from the series of estimates after (31) that

$$\begin{aligned}
&\frac{2}{(1\pm\tilde{s}\phi_1)^2}\sum_{j=J_n}^{\tilde{j}_n}\sum_{\ell}\|\psi_1^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\|\psi_1^{\mathfrak{B}_{j\ell}}\|\leq\frac{3}{2}\Gamma\sqrt{\log(n)/n}}\\
&\qquad\lesssim\frac{R^2}{\min(1,s_+)}\Big(\frac{1+\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2\Big(\frac{\Gamma^2}{nR^2(1+\tilde{s}\phi_1)^2}\Big)^{2s_+/(2s_++1)}\\
&\qquad\qquad+\frac{R^2}{\min(1,s_-)}\Big(\frac{1-\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^2\Big(\frac{\Gamma^2}{nR^2(1-\tilde{s}\phi_1)^2}\Big)^{2s_-/(2s_-+1)}.
\end{aligned}$$

Next, it has been already established in (32) that

$$\sum_{j=J_n}^{\tilde{j}_n}\sum_{\ell}\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|^2\mathbf{1}_{\|\beta_{\pm}^{\mathfrak{B}_{j\ell}}\|\leq 32\Gamma T_n}\lesssim\frac{R^2}{\min(1,s_{\mp})}\Big(\frac{1\mp\tilde{s}\phi_1}{1\pm\tilde{s}\phi_1}\Big)^{2/(2s_{\mp}+1)}\Big(\frac{\Gamma^2T_n^2}{R^2N}\Big)^{2s_{\mp}/(2s_{\mp}+1)}.$$

Consequently, when passing to the supremum, R_7 will obey the same upper bound as R_4, eventually up to a universal multiplicative constant.

C.6.11 Control of R_8

This has already been done in Section C.5.6. We recall the result:

$$\sup_{\theta\in\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)\cap\Sigma_{\gamma^*}(L)}R_8(\theta)\leq\frac{BR^2}{\min(1,s_{\pm})}\Big(\frac{\tau^2\log(n)}{n}\Big)^{2s_{\pm}}.$$

C.7 Proof of Theorem 4

Recall $\tilde{V}$ is the leading eigenvector of the empirical Gram matrix $\tilde{\mathcal{G}}$ and V_θ the leading eigenvector of the Gram matrix $\mathcal{G}$ normalized such that $\|\tilde{V}\|=\|V_\theta\|=1$. We use a Davis-Kahan argument to bound the norm $\|\tilde{V}-\mathrm{sgn}(\langle\tilde{V},V_\theta\rangle)V_\theta\|$. In particular using the version of Davis-Kahan's theorem given in the Corollary 1 of (Yu et al., 2015), we know that

$$\|\tilde{V}-\mathrm{sgn}(\langle\tilde{V},V_\theta\rangle)V_\theta\|\leq\frac{2\sqrt{2}\|\tilde{\mathcal{G}}-\mathcal{G}\|_{\mathrm{op}}}{|\lambda|}$$

where λ is the unique non-zero eigenvalue of $\mathcal{G}$, and $\|\cdot\|_{\mathrm{op}}$ stands for the operator norm. It is rapidly seen that

$$\lambda = r(\phi) \sum_{\lambda \in \Lambda(M)} \langle \psi_2, e_\lambda \rangle^2 = r(\phi) \left(\sum_{k=0}^{2^J-1} \langle \psi_2, \Phi_{Jk} \rangle^2 + \sum_{j=J}^{M} \sum_{k=0}^{2^j-1} \langle \psi_2, \Psi_{jk} \rangle^2 \right).$$

We now bound $\|\tilde{\mathcal{G}} - \mathcal{G}\|_{\mathrm{op}}$. By definition of the operator norm and then by a duality argument [here U denotes the unit ball of $\mathbb{R}^{\Lambda(M)}$]

$$\begin{aligned}
\|\tilde{\mathcal{G}} - \mathcal{G}\|_{\mathrm{op}} &= \sup_{u \in U} \|\tilde{\mathcal{G}} u - \mathcal{G} u\| \\
&= \sup_{u \in U} \sup_{v \in U} v^T (\tilde{\mathcal{G}} - \mathcal{G}) u \\
&= \sup_{u \in U} \sup_{v \in U} \Big[\Big(\frac{u+v}{2} \Big)^T (\tilde{\mathcal{G}} - \mathcal{G}) \frac{u+v}{2} - \Big(\frac{u-v}{2} \Big)^T (\tilde{\mathcal{G}} - \mathcal{G}) \frac{u-v}{2} \Big] \\
&\leq \sup_{u \in U} \sup_{v \in U} \Big[u^T (\tilde{\mathcal{G}} - \mathcal{G}) u - v^T (\tilde{\mathcal{G}} - \mathcal{G}) v \Big] \\
&\leq 2 \sup_{u \in U} u^T (\tilde{\mathcal{G}} - \mathcal{G}) u.
\end{aligned}$$

Then, let $\mathcal{N}$ be a $(1/8)$-net over U in the euclidean norm, and let $\pi : U \to \mathcal{N}$ denote the map that projects elements of U onto their closest element in $\mathcal{N}$. Then,

$$\begin{aligned}
\sup_{u \in U} u^T (\tilde{\mathcal{G}} - \mathcal{G}) u = \sup_{u \in U} \Big[& \pi(u)^T (\tilde{\mathcal{G}} - \mathcal{G}) \pi(u) + 2\pi(u)^T (\tilde{\mathcal{G}} - \mathcal{G})(u - \pi(u)) \\
& + (u - \pi(u))^T (\tilde{\mathcal{G}} - \mathcal{G})(u - \pi(u)) \Big] \\
\leq \max_{u \in \mathcal{N}} u^T (\tilde{\mathcal{G}} - \mathcal{G}) u + \frac{3}{8} & \|\tilde{\mathcal{G}} - \mathcal{G}\|_{\mathrm{op}}
\end{aligned}$$

and thus

$$\|\tilde{\mathcal{G}} - \mathcal{G}\|_{\mathrm{op}} \leq 8 \max_{u \in \mathcal{N}} u^T (\tilde{\mathcal{G}} - \mathcal{G}) u.$$

Next, we decompose $\tilde{\mathcal{G}} - \mathcal{G} = \Delta^{(1)} + \Delta^{(2)} + \Delta^{(3)} + \Delta^{(4)}$ with

$$\begin{aligned}
\Delta^{(1)}_{\lambda\lambda'} &:= \frac{1}{2} \Big(\tilde{\mathbb{P}}^{(1)}_n (e_\lambda \otimes e_{\lambda'} + e_{\lambda'} \otimes e_\lambda) - \mathbb{E}_\theta (e_\lambda \otimes e_{\lambda'} + e_{\lambda'} \otimes e_\lambda) \Big) \\
\Delta^{(2)}_{\lambda\lambda'} &:= -\mathbb{E}_\theta (e_{\lambda'}) \Big(\tilde{\mathbb{P}}^{(1)}_n (e_\lambda) - \mathbb{E}_\theta (e_\lambda) \Big) \\
\Delta^{(3)}_{\lambda\lambda'} &:= -\mathbb{E}_\theta (e_\lambda) \Big(\tilde{\mathbb{P}}^{(1)}_n (e_{\lambda'}) - \mathbb{E}_\theta (e_{\lambda'}) \Big) \\
\Delta^{(4)}_{\lambda\lambda'} &:= -\Big(\tilde{\mathbb{P}}^{(1)}_n (e_\lambda) - \mathbb{E}_\theta (e_\lambda) \Big) \Big(\tilde{\mathbb{P}}^{(1)}_n (e_{\lambda'}) - \mathbb{E}_\theta (e_{\lambda'}) \Big)
\end{aligned}$$

Using Lemma 20 applied to the function $h(y_1, y_2) = \frac{1}{2} \sum_{\lambda, \lambda' \in \Lambda(M)} u_\lambda u_{\lambda'} \big(e_\lambda(y_1) e_{\lambda'}(y_2) + e_{\lambda'}(y_1) e_\lambda(y_2) \big)$ we find that

$$\begin{aligned}
\mathbb{P}_\theta \Big(\max_{u \in \mathcal{N}} |u^T \Delta^{(1)} u| \geq x \Big) &\leq |\mathcal{N}| \max_{u \in |\mathcal{N}|} \mathbb{P}_\theta \Big(|u^T \Delta^{(1)} u| \geq x \Big) \\
&\leq 24^{2^M} \exp \left(- \frac{C n \gamma^* x^2}{L^2 + 2^M x} \right)
\end{aligned}$$

because $\mathcal{N}$ can always be chosen to have cardinality no more than 24^{2^M} (e.g. Giné and Nickl, 2016, Theorem 4.3.34), because $\mathbb{E}_\theta(h^2) \le L^2\|h\|_{L^2}^2 = L^2$ for all $\theta \in \Sigma_{\gamma^*}(L)$ by Lemma 17, and because

$$\begin{aligned}\|h\|_\infty &\le \sup_{y_1,y_2}\Big|\sum_{\lambda\in\Lambda(M)} u_\lambda e_\lambda(y_1)\sum_{\lambda'\in\Lambda(M)} u_{\lambda'}e_{\lambda'}(y_2)\Big| \\ &\le \Big(\sup_y \sum_{\lambda\in\Lambda(M)}\big|e_\lambda(y)\big|\Big)^2 \\ &\le c2^M\end{aligned}$$

for a constant $c > 0$ depending only on the wavelet basis by a standard localization properties of wavelets (Giné and Nickl, 2016, Theorem 4.2.10 or Definition 4.2.14). Next, note that

$$u^T\Delta^{(2)}u = u^T\Delta^{(3)}u = -\mathbb{E}_\theta\bigg(\sum_{\lambda\in\Lambda(M)} u_\lambda e_\lambda\bigg)\bigg(\sum_{\lambda\in\Lambda(M)} u_\lambda\Big(\tilde{\mathbb{P}}_n^{(1)}(e_\lambda) - \mathbb{E}_\theta(e_\lambda)\Big)\bigg)$$

and,

$$u^T\Delta^{(4)}u = -\bigg(\sum_{\lambda\in\Lambda(M)} u_\lambda\Big(\tilde{\mathbb{P}}_n^{(1)}(e_\lambda) - \mathbb{E}_\theta(e_\lambda)\Big)\bigg)^2.$$

Again using Lemma 20, this time applied to the function $h(y) = \sum_{\lambda\in\Lambda(M)} u_\lambda e_\lambda(y)$ which satisfies $\mathbb{E}_\theta(h^2) \le L$ for all $\theta \in \Sigma_{\gamma^*}(L)$ and $\|h\|_\infty \le c2^{M/2}$ for a universal constant $c > 0$, we deduce that

$$\mathbb{P}_\theta\bigg(\max_{u\in\mathcal{N}}\Big|\sum_{\lambda\in\Lambda(M)} u_\lambda\Big(\tilde{\mathbb{P}}_n^{(1)}(e_\lambda) - \mathbb{E}_\theta(e_\lambda)\Big)\Big| \ge x\bigg) \le 24^{2^M}\exp\bigg(-\frac{Cn\gamma^*x^2}{L + 2^{M/2}x}\bigg).$$

Since $|\mathbb{E}_\theta h| \le [\mathbb{E}_\theta h^2]^{1/2} \le \sqrt{L}$, using that $L, 2^{M/2} \ge 1$, we deduce that

$$\mathbb{P}_\theta\Big(\tfrac{1}{8}\|\tilde{\mathcal{G}} - \mathcal{G}\|_{\mathrm{op}} \ge (2\sqrt{L}+1)x + x^2\Big) \le 2\cdot 24^{2^M}\exp\bigg(-\frac{Cn\gamma^*x^2}{L^2 + 2^M x}\bigg)$$

for a constant $C > 0$. This entails that

$$\mathbb{P}_\theta\bigg(\|\tilde{V} - \mathrm{sgn}(\langle\tilde{V}, V_\theta\rangle)V_\theta\| \ge \frac{16\sqrt{2}\big((2\sqrt{L}+1)x + x^2\big)}{|r(\phi)|\sum_{\lambda\in\Lambda(M)}\langle\psi_2, e_\lambda\rangle^2}\bigg) \le 2\cdot 24^{2^M}\exp\bigg(-\frac{Cn\gamma^*x^2}{L^2 + 2^M x}\bigg)$$

Let us remark that the wavelets coefficients of ψ_2 are those of $(f_0 - f_1)/\phi_3$. Hence, whenever $\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)$, from the definition of $\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)$ and of the Besov norm in equation (15) it must be that

$$\sup_{j\ge J} 2^{2js_*}\sum_{k=0}^{2^j-1}|\langle\psi_2, \Psi_{jk}\rangle|^2 \le \frac{4R^2}{\phi_3^2}, \tag{33}$$

Consequently since $\|\psi_2\|_{L^2} = 1$:

$$\begin{aligned}
1 &= \sum_{k=0}^{2^J-1} \langle \psi_2, \Phi_{Jk}\rangle^2 + \sum_{j\geq J}\sum_{k=0}^{2^j-1} \langle \psi_2, \Psi_{jk}\rangle^2 \\
&\leq \sum_{k=0}^{2^J-1} \langle \psi_2, \Phi_{Jk}\rangle^2 + \sum_{j=J}^{M}\sum_{k=0}^{2^j-1} \langle \psi_2, \Psi_{jk}\rangle^2 + \frac{4R^2}{\phi_3^2}\sum_{j>M} 2^{-2js_*} \\
&= \sum_{\lambda\in\Lambda(M)} \langle \psi_2, e_\lambda\rangle^2 + \frac{4R^2}{\phi_3^2}\frac{2^{-2Ms_*}}{2^{2s_*}-1}.
\end{aligned}$$

and hence $\sum_{\lambda\in\Lambda(M)}\langle \psi_2, e_\lambda\rangle^2 \geq 3/4$ under the assumptions of the theorem. Observe that $|r(\phi)| \leq \phi_3^2/4 \leq L/2$ by Lemmas 19 and 15. Then taking $x = \kappa|r(\phi)|/\sqrt{L}$ for a small enough constant κ, we find that for some $C > 0$

$$\mathbb{P}_\theta\left(\|\tilde{V} - \mathrm{sgn}(\langle \tilde{V}, V_\theta\rangle)V_\theta\| \geq \frac{1}{5}\right) \leq 2\cdot 24^{2^M}\exp\left(-\frac{Cn\gamma^* r(\phi)^2}{L^3 + 2^M\sqrt{L}|r(\phi)|}\right).$$

Next, let define $t \coloneqq \sum_{\lambda\in\Lambda(M)}\tilde{V}_\lambda e_\lambda$ and $f(x) \coloneqq \max(-\tau, \min(\tau, x))$. Observe that

$$\|\psi_2\|_\infty = \frac{\|f_0 - f_1\|_\infty}{\phi_3} \leq \frac{L}{\zeta}$$

since $0 \leq f_0, f_1 \leq L$ and $\phi_3 \geq \zeta$ when $\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)\cap\Sigma_{\gamma^*}(L)$. Then by assumption $|\psi_2(x)| \leq \tau$ for all x, and thus $\psi_2(x) = f(\psi_2(x))$. Also f is 1-Lipschitz, and thus

$$\|f\circ t - \tilde{s}\psi_2\|_{L^2} = \|f\circ t - f\circ(\tilde{s}\psi_2)\|_{L^2} \leq \|t - \tilde{s}\psi_2\|_{L^2} = \|\tilde{V} - \mathrm{sgn}\big(\langle \tilde{V}, V_\theta\rangle\big)V_\theta\|.$$

Since $\tilde{\psi}_2 = f\circ t/\|f\circ t\|_{L^2}$, we use that for any norm $\|a/\|a\| - b/\|b\|\| \leq 2\|a-b\|/(1-\|a-b\|)$ if $\|b\| = 1$, $\|a-b\| < 1$ to deduce that

$$\|\tilde{\psi}_2 - \tilde{s}\psi_2\|_{L^2} \leq \frac{2\|\tilde{V} - \mathrm{sgn}\big(\langle \tilde{V}, V_\theta\rangle\big)V_\theta\|}{1 - \|\tilde{V} - \mathrm{sgn}\big(\langle \tilde{V}, V_\theta\rangle\big)V_\theta\|}.$$

The conclusion follows since $\|\tilde{\psi}_2 - \tilde{s}\psi_2\|_{L^2}^2 = 2 - 2|\langle\tilde{\psi}_2, \psi_2\rangle|$, and hence $|\langle\tilde{\psi}_2, \psi_2\rangle| \geq 1 - \frac{\|\tilde{\psi}_2 - \tilde{s}\psi_2\|_{L^2}^2}{2}$.

C.8 Proof of Corollary 7

Suppose $2^M = O(1)$, then $\frac{n\gamma^2\delta^2\epsilon^2\zeta^4}{L^3+2^M\sqrt{L}\delta\epsilon\zeta^2} \gtrsim n^{1-2a-2b-2c}$ so that the first exponential in the bound of Theorem 6 is smaller than $\exp(-Kn^{1-2a-2b-2c})$ for some $K > 0$, which is negligible. If 2^M is not $O(1)$, then in the considered regime $\frac{n\gamma^2\delta^2\epsilon^2\zeta^4}{L^3+2^M\sqrt{L}\delta\epsilon\zeta^2} \gtrsim n2^{-M}\delta\epsilon\zeta^2 \gg 2^M$ so that the first exponential in the bound of Theorem 6 is smaller than $\exp(-Kn^{(1-a-b-2c)/2})$ for some $K > 0$, which is negligible.

Also $n\delta^2\epsilon^4\zeta^6 \geq n^{1-2a-4b-6c}$ while $L^3 + \max(\tau, \sqrt{L})^3\delta\epsilon^2\zeta^3 \leq L^3 + \max(\tau, \sqrt{L})^3$ since $\delta\epsilon^2\zeta^3 \leq 1$. Hence, the second exponential term in the bound of Theorem 6 is smaller than $\exp(-Kn^{1-2a-4b-6c})$ for some $K > 0$ and is negligible.

We claim that the term $\frac{1}{\delta^2\epsilon^2\zeta^2}\frac{\log(n)}{n}$ never dominates. Indeed, for this term to dominate, it is necessary that $\epsilon^2\zeta^2 \gg \frac{1}{\log(n)}$ to dominate the term $\frac{1}{\delta^2\epsilon^4\zeta^4 n}$ and that $\delta^2\epsilon^2\zeta^2 n = O(\log(n)^{2s_i+1})$ to dominate the term $(\delta^2\epsilon^2\zeta^2 n)^{-2s_i/(2s_i+1)}$, *ie.* $\epsilon^2\zeta^2 = O(\frac{\log(n)^{2s_i+1}}{n\delta^2}) = O(\frac{\log(n)^{2s_i+1}}{n^{1-2a}})$. Since $1-2a > 0$, the two requirements cannot be fulfilled simultaneously for n large.

Finally, the term $\frac{1}{\delta^2\epsilon^4\zeta^6 n^2}$ is clearly dominated by the term $\frac{1}{\delta^2\epsilon^4\zeta^6 n}$ and the remaining term is clearly dominated by the term $(\delta^2\epsilon^2\zeta^2 n)^{-2s_i/(2s_i+1)}$.

C.9 Proof of Corollary 9

As for the proof of Corollary 7 the two first exponential terms in the bound of Theorem 8 cannot dominate in the considered regime. It has been shown in Corollary 7 that the term $\frac{\log(n)}{\delta^2\epsilon^2\zeta^2 n}$ cannot simultaneously dominate the terms $\frac{1}{\delta^2\epsilon^4\zeta^4 n}$ and $\delta^{-2}(n\epsilon^2\zeta^2)^{-2s_1/(2s_1+1)}$ [observe that $\delta^{-2}(n\epsilon^2\zeta^2)^{-2s_1/(2s_1+1)} \geq (n\delta^2\epsilon^2\zeta^2)^{-2s_1/(2s_1+1)}$]. Also using the arguments in the proof of Corollary 7 it is trivial that the terms $\frac{1}{\delta^2\epsilon^4\zeta^4 n^2}$ and $(\log(n)/n)^{2s_0}$ cannot dominate.

To finish the proof, it is enough to show that the term $\delta^{-2}(n\epsilon^2\zeta^2)^{-2s_1/(2s_1+1)}$ is dominated by the term $(n\delta^2)^{-2s_0/(2s_0+1)}$. But in the considered regime $\delta^{-2}(n\epsilon^2\zeta^2)^{-2s_1/(2s_1+1)} = n^{-2s_1/(2s_1+1)+o(1)}$ and $(n\delta^2)^{-2s_0/(2s_0+1)} = n^{-2s_0/(2s_0+1)+o(1)}$. The conclusion follows since $s_1 > s_0$ by assumption.

References

Kweku Abraham, Ismaël Castillo, and Elisabeth Gassiat. Multiple testing in nonparametric hidden Markov models: an empirical Bayes approach. *J. Mach. Learn. Res.*, 23:Paper No. [94], 57, 2022a. ISSN 1532-4435.

Kweku Abraham, Elisabeth Gassiat, and Zacharie Naulet. Fundamental limits for learning hidden markov model parameters. *IEEE Transactions on Information Theory*, pages 1–1, 2022b. doi: 10.1109/TIT.2022.3213429.

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016.

T Tony Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *The Annals of statistics*, 27(3):898–924, 1999.

T Tony Cai. On information pooling, adaptability and superefficiency in nonparametric function estimation. *Journal of Multivariate Analysis*, 99(3):421–436, 2008.

Eric Chicken and T Tony Cai. Block thresholding for density estimation: local and global adaptivity. *Journal of Multivariate Analysis*, 95(1):76–106, 2005.

Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993. ISSN 1063-5203. doi: 10.1006/acha.1993.1005. URL `https://doi-org.ezproxy.universite-paris-saclay.fr/10.1006/acha.1993.1005`.

L. Couvreur and C. Couvreur. Wavelet based non-parametric HMMs: theory and methods. In *ICASSP '00 Proceedings*, pages 604–607, 2000.

Y. De Castro, É. Gassiat, and C. Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 17:Paper No. 111, 43, 2016. ISSN 1532-4435.

Y. De Castro, É. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Trans. Inform. Theory*, 63(8):4758–4777, 2017.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of statistics*, pages 508–539, 1996.

É. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26(1-2):61–71, 2016.

Elisabeth Gassiat. Mixtures of nonparametric components and hidden Markov models. In *Handbook of mixture analysis*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 343–360. CRC Press, Boca Raton, FL, 2019.

Elisabeth Gassiat, Judith Rousseau, and Elodie Vernet. Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electron. J. Stat.*, 12 (1):703–740, 2018. doi: 10.1214/17-EJS1387. URL `https://doi-org.ezproxy.universite-paris-saclay.fr/10.1214/17-EJS1387`.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016.

M. F. Lambert, J. P. Whiting, and A. V. Metcalfe. A non-parametric hidden Markov model for climate state identification. *Hydrology and Earth System Sciences*, 7 (5):652–667, 2003.

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, 1986.

Alexandre Lecestre. Robust estimation for ergodic markovian processes, 2023.

F. Lefèvre. Non-parametric probability estimation for HMM-based automatic speech recognition. *Computer Speach and Language*, 17:113–136, 2003.

L. Lehéricy. State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 19:Paper No. 39, 46, 2018. ISSN 1532-4435.

Daniel Moss and Judith Rousseau. Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models. *Electron. J. Stat.*, 18(1):1815–1886, 2024. ISSN 1935-7524. doi: 10.1214/23-ejs2201. URL `https://doi.org/10.1214/23-ejs2201`.

D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20:no. 79, 32, 2015.

N. Rau, J. Lucke, and A. K. Hartmann. Phase transition for parameter learning of Hidden Markov Models. *arXiv:2003.11680*, 2020.

L. Shang and K.P. Chan. Nonparametric discriminant HMM and application to facial expression recognition. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 2090–2096, 2009.

Hans Triebel. *Theory of Function Spaces*. Birkhäuser, 1983.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer-Verlag, New York, 2009.

C. Yau, O. Papaspiliopoulos, G. O. Roberts, and C. Holmes. Bayesian non-parametric hidden Markov models with applications in genomics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(1):37–57, 2011. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2010.00756.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2010.00756.x`.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, 1997.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. ISSN 0006-3444. doi: 10.1093/biomet/asv008. URL `https://doi-org.ezproxy.universite-paris-saclay.fr/10.1093/biomet/asv008`.