ISSN 2279-9362

Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models

Filippo Ascolani and Giacomo Zanella

No. 716

April 2024

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models

Filippo Ascolani* and Giacomo Zanella†

Abstract

Gibbs samplers are popular algorithms to approximate posterior distributions arising from Bayesian hierarchical models. Despite their popularity and good empirical performances, however, there are still relatively few quantitative results on their convergence properties, e.g. much less than for gradient-based sampling methods. In this work we analyse the behaviour of total variation mixing times of Gibbs samplers targeting hierarchical models using tools from Bayesian asymptotics. We obtain dimension-free convergence results under random data-generating assumptions, for a broad class of two-level models with generic likelihood function. Specific examples with Gaussian, binomial and categorical likelihoods are discussed.

1 Introduction

Gibbs samplers [12] are a family of Markov Chain Monte Carlo (MCMC) algorithms [10] commonly used in various scientific fields. In the context of Bayesian Statistics, they are routinely employed to draw samples from posterior distributions of unknown parameters conditional to the observed data [28, 37]. Like most MCMC methods, they are guaranteed to converge to the correct posterior distribution as the number of iterations tends to infinity under mild assumptions [54]. However, understanding how quickly this convergence occurs, for example by quantifying the so-called mixing time of the Markov chain generated by the algorithm, is in general a hard task. In this paper we address this question for Gibbs samplers targeting certain classes of high-dimensional Bayesian hierarchical models. Analysing convergence properties, such as mixing times, is the key technical step needed to rigorously quantify the computational cost of MCMC algorithms.

1.1 Hierarchical models

Our motivating example is given by classical Bayesian hierarchical models of the form

$$\begin{aligned} Y_j \mid \theta_j &\sim f(\cdot \mid \theta_j) \quad j = 1, \dots, J, \\ \theta_j \mid \psi &\overset{\text{iid}}{\sim} p(\cdot \mid \psi) \quad j = 1, \dots, J, \\ \psi &\sim p_0(\cdot)\,. \end{aligned} \tag{1}$$

Here the observed dataset $Y_{1:J} = (Y_j)_{j=1,\dots,J}$ is divided into J groups, with data for each group typically containing multiple observations, e.g. $Y_j = (Y_{j1}, \dots, Y_{jm})$. Each group features some local (i.e. group-specific) parameters $\theta_j \in \mathbb{R}^\ell$, while $\psi \in \mathbb{R}^D$ are global (hyper)-parameters. Above $f(\cdot \mid \theta)$, $p(\cdot \mid \psi)$ and $p_0(\cdot)$ denote some likelihood function, local prior and global prior, respectively. See Section 4 for the assumptions we require on each

*Department of Statistical Science, Duke University, filippo.ascolani@duke.edu
†Department of Decision Sciences and BIDSA, Bocconi University, giacomo.zanella@unibocconi.it

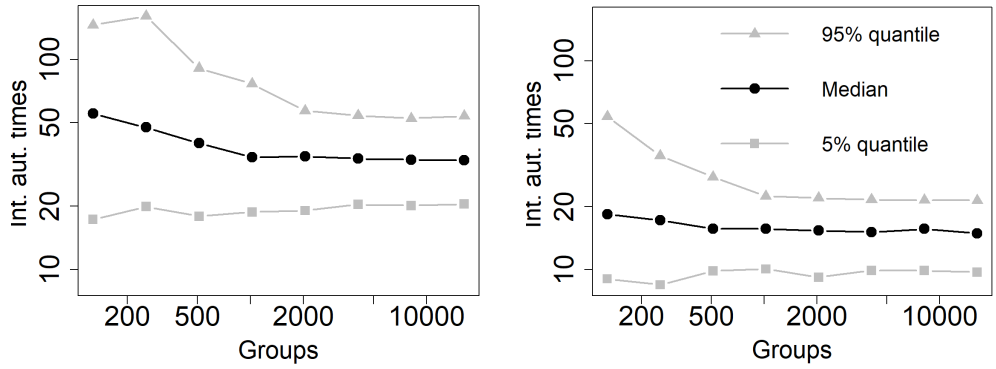

Figure 1: Integrated autocorrelation times (on log-scale) of Gibbs samplers targeting the posterior distribution of model (1) with specification (2). Quantiles refer to repetitions over datasets randomly generated according to the model with true parameters $\mu^* = \tau^* = 1$. Left: $m = 3$. Right: $m = 5$. See Section 5.2 for more details.

of those. Given model (1), posterior inferences are based on the conditional distribution of ψ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)$ given $Y_{1:J}$, which we denote as $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$. Hierarchical models such as (1) are the workhorse of Bayesian Statistics and are commonly employed in many applied contexts (see e.g. [25, 24] and references therein). In this paper, we are mostly interested in the high-dimensional regime where $J \to \infty$, so that both the number of datapoints and parameters, i.e. $n = Jm$ and $p = J\ell + D$ respectively, diverge.

One iteration of a Gibbs sampler targeting $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$ sequentially samples each parameter from its full-conditional distribution, i.e. it performs the updates $\theta_j \sim \mathcal{L}(\mathrm{d}\theta_j | Y_{1:J}, \psi)$ for $j = 1, \dots, J$ and $\psi \sim \mathcal{L}(\mathrm{d}\psi | Y_{1:J}, \boldsymbol{\theta})$. Algorithms based on conditional updates are well-suited to model (1), since they naturally exploit the underlying sparse dependence structure. In particular, the conditional independence of $\theta_1, \dots, \theta_J$ given $Y_{1:J}$ and ψ implies that the sequence of updates from the low-dimensional distributions $\mathcal{L}(\mathrm{d}\theta_j | Y_{1:J}, \psi)$ for $j = 1, \dots, J$ is equivalent to an exact joint update from the high-dimensional distribution $\mathcal{L}(\mathrm{d}\boldsymbol{\theta} | Y_{1:J}, \psi)$. Also, since local parameters interact only with local data conditional on ψ, i.e. $\mathcal{L}(\mathrm{d}\theta_j | Y_{1:J}, \psi) = \mathcal{L}(\mathrm{d}\theta_j | Y_j, \psi)$, one iteration of the Gibbs sampler can typically be implemented with a computational cost that scales linearly with J. For the sake of comparisons, a similar cost is required by a single likelihood evaluation or a single posterior gradient evaluation for model (1). See also Remark 4.2 in Section 4.2 for related discussion.

The key question to properly assess the effectiveness of Gibbs samplers targeting model (1) is how fast the resulting Markov chain converges to its stationary distribution $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$. Interestingly, such chain often enjoys dimension-free convergence speed, meaning that the number of iterations required to converge does not grow (or grows only logarithmically) with J. Figure 1 illustrates numerically this behaviour on a hierarchical logistic model, where the likelihood and prior in (1) are specified as

$$f(y \mid \theta) = \binom{m}{y} \frac{e^{y\theta}}{(1+e^{\theta})^m}, \quad p(\theta \mid \psi) = N(\theta \mid \mu, \tau^{-1}), \quad \psi = (\mu, \tau), \tag{2}$$

with $y \in \{0, \dots, m\}$ and m being a positive integer. The prior for $\psi = (\mu, \tau)$ is set to $\mu \mid \tau \sim N\left(0, 10^3/\tau\right)$ and $\tau \sim \text{Gamma}(0.1, 0.1)$. Full details on the simulation set-up of Figure 1 are described in Section 5.2. The results suggest that the number of iterations required by the Gibbs sampler to draw each sample from $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$ remains bounded as J grows and asymptotes to a finite value as $J \to \infty$. Combined with cost per iteration, this implies a computational complexity that grows linearly with J. Note that this complexity is smaller than the one of popular gradient-based MCMC methods when applied to these models (see Section 1.2 for more details), supporting the idea that Gibbs samplers can achieve state-of-the-art performances for hierarchical models with sparse dependence structures.

In Section 4 we provide rigorous support to the above empirical evidences. In particular, we study the asymptotic behavior of mixing times of Gibbs samplers targeting model (1). There we prove that mixing times remain bounded as $J \to \infty$ under mild assumptions on the likelihood f and the global prior p_0. We instead require stronger assumptions on the local priors $p(\cdot \mid \psi)$, which we assume to be in the exponential family. Our results (see e.g. Theorem 4.2) are average-case ones and hold with high probability with respect to the law of the data-generating process. To do so we assume the observed data $Y_{1:J}$ to be randomly generated. This allows to use tools of Bayesian asymptotics, such as Bernstein-von Mises type statements (see e.g. Chapter 10 of [64]), to characterize the asymptotic posterior behaviour as $J \to \infty$ and then extract information about the limiting behaviour of the associated sequence of MCMC algorithms.

1.2 Related literature

The literature on performances of MCMC methods is very broad. The most well-studied classes of algorithm are probably gradient-based ones, such as Langevin [57] and Hamiltonian [38] Monte Carlo, see e.g. [14, 18, 19] and related literature. Available results suggest that the number of iterations (or target gradient evaluations) required by those algorithm to converge to stationarity increases with dimensionality, e.g. growing as $\mathcal{O}(J^{\alpha})$ with the dimensionality J, for some $\alpha > 0$ that depends on the setup and type of algorithm [51, 7, 66]. In the context of hierarchical models, given that each target gradient evaluation has a linear cost in J, this leads to a computational cost to sample from $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$ that scales super-linearly with J, e.g. as $\mathcal{O}(J^{1+\alpha})$ with $\alpha > 0$. Comparing these results to the one we develop here for Gibbs samplers suggests that, while being state-of-the-art black-box schemes to sample from generic high-dimensional distributions with appropriate regularity conditions (e.g. log-concavity), default gradient-based MCMC schemes can be suboptimal for high-dimensional hierarchical models. See also [46] for related numerical evidences.

Compared to gradient-based MCMC, results for Gibbs-type schemes are less abundant and more model-dependent. Notable recent examples include [67, 30, 49], which provide convergence bounds for hierarchical models, similar to (1), with Gaussian and Poisson likelihoods. Another recent result is given by [48], which provides dimension-free convergence bounds for Gibbs samplers for high-dimensional probit regression models under appropriate regimes. Providing sharp non-asymptotic analyses like the ones above requires proof techniques, such as drift-and-minorization techniques [58] and random mappings [48], that are usually likelihood-specific and potentially hard to construct. For example, they may require to devise and study a suitable Lyapunov function that depends on the specific choices of both likelihood and priors in (1) (see e.g. formulae (6) and (33) in [30] and [67], respectively). On the other hand, these approaches provide non-asymptotic bounds that apply to fixed sample size and dimensionality, thus being complimentary to the high-dimensional asymptotic analysis we develop here.

Interestingly, there are relatively few papers combining the tools of Bayesian asymptotics and MCMC theory in rigorous ways. The work in [6] uses Bernstein-von Mises Theorem to provide polynomial bounds on the convergence of random walk Metropolis-Hastings schemes. After that, very recent papers use similar techniques to provide complexity analysis of MCMC schemes, see e.g. [41, 39, 62] dealing with gradient-based methods, the first in the context of inverse problems. A brief discussion about the use of asymptotic posterior characterisations to study the convergence properties of Gibbs samplers is given in [56]. A more in-depth use of Bayesian asymptotics to study data augmentation procedures is given in [32], which also considers hierarchical models. See Remark 4.2 in Section 4 for more details on the results in [32]. Finally, an interesting exception is given by Bayesian variable selection models, where multiple works have exploited the asymptotic behaviour of the posterior distribution to characterize the computational performances of Bayesian methods [68, 3, 69].

1.3 Sketch of the main arguments and structure of the paper

The argument we employ to study Gibbs samplers targeting $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi \mid Y_{1:J})$ can be decomposed in three main parts. First, if $p(\cdot \mid \psi)$ belongs to the exponential family, there exists a set of sufficient statistics $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{\theta})$, whose dimensionality does not depend on J, such that $\mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}) = \mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J})$. Lemma 4.1 in Section 4.1 shows that, as a result, the Gibbs sampler on $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi \mid Y_{1:J})$ has the same mixing times as the one on $\mathcal{L}(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J})$. This allows to focus on the latter distribution which, unlike the former, is intractable but fixed dimensional. Note that this dimensionality reduction does not require the likelihood f to admit sufficient statistics (see Remark 4.1) and is a peculiar property of Gibbs samplers, since it exploits the presence of exact updates. The second step consists in studying the asymptotic behaviour of $\mathcal{L}(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J})$ as J increases. In particular, Proposition 4.5 shows that a suitable rescaling of $(\boldsymbol{T}, \psi)$ converges to a multivariate Gaussian distribution in total variation distance. The proof combines a classical Bernstein-von Mises Theorem for ψ (Lemma 4.3) with a less standard Central Limit Theorem for $\boldsymbol{T}$ conditional on ψ (Lemma 4.4). More details can be found in Section 4.3. The final and key point is then to connect the convergence of the target distributions, in this case $\{\mathcal{L}(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J})\}_{J \geq 1}$, to the convergence of the associated Gibbs sampler operators. Theorem 2.4 proves that the limiting behaviour of a sequence of Gibbs samplers is equivalent to the behaviour of the Gibbs sampler on the limiting distribution: this is shown in total variation distance and under warm start assumption. The fundamental link is given by Proposition 2.2, which provides an upper bound on the distance between Gibbs sampler operators in terms of the one between the target distributions. Since those results are of independent interest and are not specific to hierarchical models, we start by developing those in a general setup in Section 2. Then, Section 3 recalls the Bernstein-von Mises Theorem and illustrates the results of Section 2 to the fixed-dimensional setting. Section 4 develops the main results of the paper dealing with general hierarchical models (see e.g. Theorem 4.2) and Section 5 verifies the general conditions for some specific likelihood families, e.g. Gaussian, binomial and categorical, together with providing numerical simulations and extension to different graphical model structures. Since a warm start initialization for the sampler is assumed throughout, the availability of feasible starts is discussed in Section 6. Finally, Section 7 discusses extensions and future work.

2 Gibbs sampler and asymptotics

In this section, after recalling basic definitions about Gibbs kernels and mixing times, we connect the convergence of a sequence of target distributions to the convergence of the associated Gibbs kernels. This leads to Theorem 2.4, which characterizes the limiting behaviour of the Gibbs samplers mixing times. Throughout this section, the target distributions are assumed to have fixed dimensionality.

2.1 Setup and notation

Let $(\pi_n)_{n \geq 1} = \left(\pi_n(\cdot \mid Y^{(n)})\right)_{n \geq 1}$ be a sequence of probability distributions on a common product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$, where each π_n is allowed to depend on some observed data $Y^{(n)} \in \mathcal{Y}^{(n)}$. In our applications, $\pi_n(\cdot \mid Y^{(n)})$ represents the posterior distribution of some unknown parameter $\mathbf{x} \in \mathcal{X}$ conditioned on the data $Y^{(n)}$. For the sake of brevity, we will often omit the explicit dependence on $Y^{(n)}$.

Let P_n be the Markov transition kernel of the deterministic-scan Gibbs sampler targeting π_n, defined as the product of K kernels

$$P_n = P_{n,1} \cdots P_{n,K} \,. \tag{3}$$

For each $i \in \{1, \ldots, K\}$, $P_{n,i}$ is the transition kernel on $\mathcal{X}$ that updates the i-th coordinate drawing it from its conditional distribution $\pi_n(\mathrm{d}x_i | \mathbf{x}^{(-i)})$, where $\mathbf{x}^{(-i)} = (x_j)_{j \neq i}$, while

leaving the other components unchanged. Equivalently

$$P_{n,i}\left(\mathbf{x}, S_{\mathbf{x},i,A}\right) = \int_A \pi_n\left(\mathrm{d}y_i \mid \mathbf{x}^{(-i)}\right), \quad A \subset \mathcal{X}_i, \quad i = 1, \ldots, n,$$

with $S_{\mathbf{x},i,A} = \{\mathbf{y} \in \mathcal{X} : y_j = x_j \, \forall \, j \neq i \text{ and } y_i \in A\}$. It is easy to show that $P_{n,i}$ is reversible with respect to π_n for every i, so that π_n is the invariant distribution of P_n [53, 29, 13].

Given $\epsilon \in (0,1)$, define the ϵ-total variation mixing time of P_n with starting distribution $\mu_n \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability distribution on $\mathcal{X}$, as

$$t_{mix}^{(n)}(\epsilon, \mu_n) = \inf\left\{t \geq 0 : \left\|\mu_n P_n^t - \pi_n\right\|_{TV} < \epsilon\right\}, \tag{4}$$

where P^t denotes the t-th power of P, $\mu_n P_n^t(A) = \int_{\mathcal{X}} P_n^t(\mathbf{x}, A)\mu_n(\mathrm{d}\mathbf{x})$ for any $A \subseteq \mathcal{X}$ and $\|\cdot\|_{TV}$ denotes the total variation norm. By definition, mixing times quantify the number of Markov chain's iterations required to obtain a sample from the target distribution π_n up to error ϵ. We will focus on worst-case mixing times with respect to M-warm starts. The set of M-warm starts relative to a distribution π is defined as

$$\mathcal{N}(\pi, M) = \{\mu \in \mathcal{P}(\mathcal{X}) : \mu(A) \leq M\pi(A) \text{ for all } A \subseteq \mathcal{X}\}, \qquad M \geq 1, \, \pi \in \mathcal{P}(\mathcal{X}), \tag{5}$$

and the associated worst-case mixing times for P_n targeting π_n are

$$t_{mix}^{(n)}(\epsilon, M) = \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} t_{mix}^{(n)}(\epsilon, \mu_n). \tag{6}$$

Remark. While being common in the literature, see e.g. [14, 19, 62] for gradient-based methods, the warm start assumption can be quite stringent and potentially unrealistic. In particular, assuming that the algorithm can be initialised by sampling the starting configuration from a warm start with relatively small M (e.g. one that does not grow exponentially fast with dimensionality) may be unrealistic. In Section 6 we show that in the specific case of hierarchical models as in (1) a feasible start, i.e. a starting distribution which can be implemented in practice and allows to control the value of M, is available under some assumptions.

2.2 Assumptions on the sequence of target distributions

We consider settings where a rescaled version of the sequence $(\pi_n)_{n\geq 1}$ converges to a well defined limiting distribution as $n \to \infty$. This is often the case in a Bayesian context where some version of the Bernstein von-Mises theorem holds (see e.g. Theorem 3.1 below). The convergence of $(\pi_n)_{n\geq 1}$ occurs with high probability assuming the data $Y^{(n)}$ is randomly generated from some distribution. In particular, we assume for the rest of this section that $Y^{(n)}$ is random with distribution $Q^{(n)} \in \mathcal{P}\left(\mathcal{Y}^{(n)}\right)$. The following assumption specifies the convergence we require for $(\pi_n)_{n\geq 1}$:

(A1) There exists $\tilde{\pi} \in \mathcal{P}(\mathcal{X})$ and a sequence of transformations $\phi_n : \mathcal{X} \to \mathcal{X}$ that act *coordinate-wise*, i.e. where

$$\phi_n(\mathbf{x}) = \left(\phi_{n,1}(x_1), \ldots, \phi_{n,K}(x_K)\right), \qquad \mathbf{x} \in \mathcal{X} \tag{7}$$

with $\phi_{n,j} : \mathcal{X}_j \to \mathcal{X}_j$ injective and measurable, such that

$$\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0 \qquad \text{as } n \to \infty, \tag{8}$$

in $Q^{(n)}$-probability, i.e. such that $\lim_{n\to\infty} Q^{(n)}(\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} > \epsilon) = 0$ for every $\epsilon \in (0,1)$, where $\tilde{\pi}_n = \pi_n \circ \phi_n^{-1}$ is the law of $\tilde{\mathbf{x}} = \phi_n(\mathbf{x})$ under $\mathbf{x} \sim \pi_n$.

Remark. The necessity of rescaling $\mathbf{x}$ by some transformation ϕ_n in (7) comes from the typical behaviour of posterior distributions in Bayesian models. Indeed, without rescaling, π_n often converges to a random variable which is degenerate to a Dirac delta at a fixed value (e.g. the underlying data-generating parameter). Thus, in order to have a non-trivial limit and total variation convergence, which is essential for our purposes, a suitable rescaling is needed. In our context the specific form of this transformation is dictated by the theory of Bayesian asymptotics, see e.g. Theorem 3.1 below. Moreover, we assume ϕ_n to act coordinate-wise because this class of transformations leaves Gibbs samplers invariant (see e.g. Lemma 2.1 below), while general one-to-one transformations can alter the Gibbs sampler dynamics and change its convergence speed [45].

Remark. The results we develop below could be extended to more general versions of assumption (A1), including ones where the co-domain of ϕ_n is not equal to the domain, i.e. $\phi_n : \mathcal{X} \to \mathcal{Z}$ for some $\mathcal{Z}$, and where the limiting distribution $\tilde{\pi}$ is random, i.e. allowed to depend on the sequence $(Y^{(n)})_n$. Since $(A1)$ is enough for our purposes and motivating applications, we do not consider such extensions here to keep notation simple.

Let $\tilde{P}$ and $\tilde{P}_n$ be the kernels of the Gibbs samplers targeting $\tilde{\pi}$ and $\tilde{\pi}_n$, respectively. The following lemma shows that studying total variation convergence from M-warm starts for the sequence of kernels $(P_n)_{n\geq 1}$ is equivalent to doing it for the sequence $(\tilde{P}_n)_{n\geq 1}$. The proof, which can be found in Appendix C, relies on the coordinate-wise and bijective requirements of (A1).

Lemma 2.1. *Under Assumption (A1) we have*

$$\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^t - \pi_n \right\|_{TV} = \sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} .$$

2.3 Convergence of Gibbs samplers operators

Since by $(A1)$ the stationary distribution of $\tilde{P}_n$, the Gibbs samplers targeting $\tilde{\pi}_n$, converges to the one of $\tilde{P}$, one may be tempted to translate such convergence at the level of the kernels, e.g. $\|\tilde{P}_n(\mathbf{x}, \cdot) - \tilde{P}(\mathbf{x}, \cdot)\|_{TV} \to 0$ for ($\tilde{\pi}$-almost) every $\mathbf{x} \in \mathcal{X}$. However this is not only false for generic Markov operators, but even in the special class of Gibbs sampler operators: one can have $\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0$ as $n \to \infty$, while $\|\tilde{P}_n(\mathbf{x}, \cdot) - \tilde{P}(\mathbf{x}, \cdot)\|_{TV} \nrightarrow 0$ for any $\mathbf{x} \in \mathcal{X}$, see e.g. Example $A.1$ in Appendix A. The reason is that convergence of the joint distribution $\tilde{\pi}_n$ in total variation distance does not imply convergence of the associated conditional distributions, that are the building blocks of the Gibbs sampler operator. However, it turns out that a control on the total variation distance between two target distributions is in general sufficient to control the distance between the corresponding Gibbs sampler operators applied to warm starts. The following Proposition makes the connection precise. Interestingly, no assumptions on the target distribution and Gibbs samplers are required.

Proposition 2.2. *Let P_1 and P_2 be the transition kernels of Gibbs samplers targeting $\pi_1 \in \mathcal{P}(\mathcal{X})$ and $\pi_2 \in \mathcal{P}(\mathcal{X})$, respectively. Then we have*

$$\|\mu P_1 - \mu P_2\|_{TV} \leq 2MK \|\pi_1 - \pi_2\|_{TV} , \tag{9}$$

for every $\mu \in \mathcal{N}(\pi_1, M) \cup \mathcal{N}(\pi_2, M)$ and $M \geq 1$.

Proposition 2.2 translates convergence of the stationary distributions, given by $(A1)$, into convergence of the Gibbs samplers operators when a warm start is considered. It is worth noting that a bound of this form cannot hold for generic Markov transition kernels. Indeed, consider transition kernels P_1 and P_2 with the same stationary distribution π: by basic properties of the total variation distance it holds $\|\mu P_1 - \mu P_2\|_{TV} \leq 2 \|\mu - \pi\|_{TV}$. The latter bound cannot be improved in general, meaning that it is possible to find ergodic kernels P_1 and P_2 that get arbitrarily close to the above upper bound, see Example A.2 in Appendix A.

Proposition 2.2 is used in the proof of Theorem 2.4, which shows that the limiting behaviour of P_n, in terms of distance to stationarity from M-warm starts, is completely characterized by the behaviour of the limiting operator $\tilde{P}$. The proof of Theorem 2.4 also relies on the fact that the total variation distance between π_1 and π_2 provides a control on the distance between the two sets $\mathcal{N}(\pi_1, M)$ and $\mathcal{N}(\pi_2, M)$, as shown in the following Lemma.

Lemma 2.3. *Let $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$. Then, for every $\mu_1 \in \mathcal{N}(\pi_1, M)$, there exists $\mu_2 \in \mathcal{N}(\pi_2, M)$ such that $\|\mu_1 - \mu_2\|_{TV} \leq M \|\pi_1 - \pi_2\|_{TV}$.*

Lemma 2.3 implies that, under assumption $(A1)$, for every $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ there exists a sequence $\{\tilde{\mu}_n\}_n$ such that $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ and $\|\tilde{\mu}_n - \tilde{\mu}\|_{TV} \to 0$ as $n \to \infty$ in $Q^{(n)}$-probability. We can now state Theorem 2.4.

Theorem 2.4. *Let assumption $(A1)$ holds. Then for every $t \in \mathbb{N}$ and $M \geq 1$ it holds*

$$\lim_{n\to\infty} \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\|\mu_n P_n^t - \pi_n\right\|_{TV} = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV},$$

in $Q^{(n)}$-probability.

Remark. An alternative approach to derive convergence statements on the sequence of Gibbs kernels would be to consider stronger forms of convergence for the sequence $(\tilde{\pi}_n)_{n\geq 1}$ than the one in total variation distance in (8). However, we prefer to derive results under weaker convergence requirements for $(\tilde{\pi}_n)_{n\geq 1}$ to allow for a more direct use of standard asymptotic results in the Bayesian literature (e.g. common formulations of the Bernstein-von Mises theorem), which are usually derived in terms of weaker metrics such as total variation one.

2.4 Implications for mixing times

Denote the mixing times of $\tilde{P}$ as

$$\tilde{t}_{mix}(\epsilon, M) = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \inf \left\{ t \geq 1 : \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} < \epsilon \right\}.$$

The following corollary of Theorem 2.4 shows how to use $\tilde{t}_{mix}(\epsilon, M)$ to deduce statements on the behaviour of the sequence of mixing times of interest, $(t_{mix}^{(n)}(\epsilon, M))_{n\geq 1}$.

Corollary 2.5. *Let assumption $(A1)$ holds. If $(M, \epsilon) \in [1, \infty) \times (0, 1)$ is such that $\tilde{t}_{mix}(\epsilon, M) < \infty$, then*

$$Q^{(n)} \left(t_{mix}^{(n)}(\epsilon, M) \leq \tilde{t}_{mix}(\epsilon, M) \right) \to 1 \tag{10}$$

as $n \to \infty$. Otherwise, if $(M, \epsilon) \in [1, \infty) \times (0, 1)$ is such that $\tilde{t}_{mix}(\epsilon, M) = \infty$, then it holds

$$Q^{(n)} \left(t_{mix}^{(n)}(\underline{\epsilon}, M) < T \right) \to 0$$

as $n \to \infty$, for every $\underline{\epsilon} < \epsilon$ and $T > 0$.

Remark (Mixing times bounded in probability). When $\tilde{t}_{mix}(\epsilon, M) < \infty$, the statement in (10) implies that $t_{mix}^{(n)}(\epsilon, M) = \mathcal{O}_P(1)$ as $n \to \infty$, i.e. that the sequence of random variables $(t_{mix}^{(n)}(\epsilon, M))_{n\geq 1}$ is bounded in probability. The latter means that for every $\delta > 0$ there exist an integer N_δ and a real constant $B_\delta < \infty$ such that $Q^{(n)}(t_{mix}^{(n)}(\epsilon, M) \leq B_\delta) \geq 1 - \delta$ for every $n \geq N_\delta$, which holds by (10) taking $B_\delta = \tilde{t}_{mix}(\epsilon, M)$.

By Corollary 2.5, establishing whether $\tilde{P}$ is ergodic (in the sense of yielding finite mixing times) or not is enough to discriminate between sequences of kernels $(P_n)_{n\geq 1}$ whose mixing times diverge as $n \to \infty$ as opposed to ones that do not (see e.g. Figure 4 in Section 5 for an illustration). Since ergodicity of Gibbs samplers can be established under very mild assumptions [54], in practice one can expect $\tilde{P}$ to be ergodic and thus $(t^{(n)}_{mix}(\epsilon, M))_{n\geq 1}$ to be bounded in probability whenever $(A1)$ holds for a well-behaved, non-singular limiting distribution $\tilde{\pi}$. Sections 4 and 5 combine Corollary 2.5 with dimensionality reduction techniques to provide results on Gibbs samplers targeting high-dimensional hierarchical models.

Remark (Alternative metrics). It is natural to wonder whether the result of Corollary 2.5 may hold for weaker metrics, like the one induced by the Wasserstein distance. However, it is possible to find examples where the convergence of the stationary distributions (in Wasserstein distance) does not imply convergence of the associated mixing times (neither the ones defined based on the TV distance nor the ones defined based on the Wasserstein one). The intuition is that the limiting distribution in weaker metrics (e.g. Wasserstein, weak convergence, etc) may ignore features of the joint distribution, such as full conditionals behaviours, that have a relevant impact on Gibbs sampler dynamics. For example, a sequence of increasingly correlated random variables (whose Gibbs samplers converge slower and slower) may converge to a single point mass, for which independence and immediate convergence automatically holds. See Example $A.3$ in Appendix A.

2.5 Explicit limiting bounds

Corollary 2.5 can also be used to derive quantitative bounds on the limiting behaviour of the mixing times $(t^{(n)}_{mix}(\epsilon, M))_{n\geq 1}$. In particular, if one is able to establish explicit bounds on $\tilde{t}_{mix}(\epsilon, M)$, then (10) implies a corresponding bound in high probability on $t^{(n)}_{mix}(\epsilon, M)$ for large n. While deriving quantitative bounds on Gibbs samplers mixing times is in general hard, the limiting distribution $\tilde{\pi}$ is often more tractable than the original sequence $(\pi_n)_{n\geq 1}$, a common case being the one where $\tilde{\pi}$ is multivariate Gaussian while $(\pi_n)_{n\geq 1}$ is not. In those scenarios explicit bounds on $\tilde{t}_{mix}(\epsilon, M)$ can be derived using available results on the convergence properties of Gibbs samplers targeting multivariate Gaussian distributions, see e.g. [1, 33, 55]. For example, Theorem 2 in [1] provides an explicit bound for deterministic scan Gibbs samplers on Gaussian targets in L^2-distance (and therefore total variation [2]).

In Sections 4 and 5 we will apply this strategy mostly to cases where $K = 2$, meaning that $\tilde{P}$ is a two-block Gibbs sampler. In this situation, one can use spectral gaps to bound Gibbs samplers mixing times, as shown in the Corollary 2.6. Given a π-invariant kernel P with $\pi \in \mathcal{P}(\mathcal{X})$ we define its spectral gap as

$$\text{Gap}(P) = \inf_{f\,:\,\pi(f^2)<\infty,\,\text{Var}_\pi(f)>0} \left\{ \frac{\int_{\mathcal{X}^2} \left[f(\mathbf{y}) - f(\mathbf{x})\right]^2 \pi(\mathrm{d}\mathbf{x}) P(\mathbf{x}, \mathrm{d}\mathbf{y})}{2\text{Var}_\pi(f)} \right\},$$

where $f : \mathcal{X} \to \mathbb{R}$ are measurable functions, $\pi(f) = \int_{\mathcal{X}} f(\mathbf{x})\pi(\mathrm{d}\mathbf{x})$ and $\text{Var}_\pi(f) = \int_{\mathcal{X}} \left[f(\mathbf{x}) - \pi(f)\right]^2 \pi(\mathrm{d}\mathbf{x})$. We refer to [59] and the proof of Corollary 2.6 for discussion on why spectral gaps, which are commonly used for π-reversible chains, can be used to analyse two-block Gibbs samplers, which are technically not reversible. We also note that Corollary 2.6 is only one possible approach to bound $\tilde{t}_{mix}(\epsilon, M)$ and that any quantitative bound on the latter can be combined with Corollary 2.5 to deduce limiting statements on $(t^{(n)}_{mix}(\epsilon, M))_{n\geq 1}$.

Corollary 2.6. *Let $K = 2$, assumption $(A1)$ be satisfied and Gap$(\tilde{P}) > 0$. Then, for every*

$(M, \epsilon) \in [1, \infty) \times (0, 1)$ *it holds*

$$Q^{(n)}\left(t_{mix}^{(n)}(\epsilon, M) \leq 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log(1 - Gap(\tilde{P}))}\right) \to 1 \quad \text{as } n \to \infty\,.$$

Given the result of Corollary 2.6, it is natural to ask whether the convergence proved in Theorem 2.4 could be rephrased in terms of spectral gaps, i.e. $\mathrm{Gap}(P_n) \to \mathrm{Gap}(\tilde{P})$. However, once again, convergence in total variation is too weak for this purpose: indeed it is not difficult to find examples where $(A1)$ holds and the associated Gibbs sampler spectral gaps do not converge, even under the stronger condition requiring $\|\tilde{P}_n(\mathbf{x}, \cdot) - \tilde{P}(\mathbf{x}, \cdot)\|_{TV} \to 0$ for any $\mathbf{x} \in \mathcal{X}$, see Example $A.4$ in Appendix A. Controlling directly the spectral gaps would require extremely stringent conditions on the convergence of $\tilde{\pi}_n$ to $\tilde{\pi}$ that are rarely satisfied (e.g. uniform convergence of the associated densities on the log-scale, i.e. $\sup_{\mathbf{x}\in\mathcal{X}} |\log \tilde{\pi}_n(\mathbf{x}) - \log \tilde{\pi}(\mathbf{x})| \to 0$). An alternative approach to the direct warm-start mixing time analysis that we perform here, would be to consider asymptotic behaviours of *approximate* spectral measures, such as approximate spectral gaps, see e.g. [3, 62].

3 Illustrative example: fixed-dimensional parametric models

We first consider the fixed-dimensional case. While this is not our main interest or motivating application, it allows to show the type of results we will derive and also introduce notation about classical Bayesian asymptotic results that we will use. In this setting $\pi_n(\mathrm{d}\psi) = p(\mathrm{d}\psi \mid Y^{(n)})$ is the posterior distribution of the Bayesian model defined as

$$Y_i \mid \psi \overset{iid}{\sim} f(Y \mid \psi), \quad \psi \sim p_0(\psi), \tag{11}$$

where $\psi = (\psi_1, \ldots, \psi_K)$, with $\mathcal{X} \in \mathbb{R}^K$, and $Y^{(n)} = (Y_1, \ldots, Y_n)$, with $Y_i \in \mathcal{Y}$, $i = 1, \ldots, n$, so that $\mathcal{Y}^{(n)} = \mathcal{Y}^n$. Moreover, if $Y_i \overset{\text{iid}}{\sim} Q$ for some $Q \in \mathcal{P}(\mathcal{Y})$, we denote with $Q^{(n)}$ and $Q^{(\infty)}$ the associated product measures. We study the mixing times of the Gibbs sampler that updates one coordinate of ψ at the time as n grows. In order to apply the results of Theorem 2.4 we need a suitable transformation of ψ, that is given by the celebrated Bernstein-von Mises Theorem, which we now recall. The version we provide here, which makes stronger than needed assumptions, can be obtained combining Theorem 10.1 in [64], with other remarks in Chapter 10 therein, incuding Lemmas 10.4 and 10.6.

Theorem 3.1 (Bernstein-von Mises)**.** *Consider model* (11) *and let the map* $\psi \to f(\cdot \mid \psi)$ *be one-to-one. Let the map* $\psi \to \sqrt{f(y \mid \psi)}$ *be continously differentiable for every* $y \in \mathcal{Y}$*, with non-singular and continuous Fisher Information* $\mathcal{I}(\psi)$*. Let the prior measure be absolutely continuous in a neighborhood of* $\psi^* \in \mathcal{X}$ *with a continuous positive density at* ψ^**. Finally, let* Ψ *be a compact neighborhood of* ψ^* *for which there exists a sequence of tests* u_n *such that*

$$\begin{aligned} &\int_{\mathcal{Y}^{(n)}} u_n(y_1, \ldots, y_n) \prod_{i=1}^n f(dy_i \mid \psi^*) \to 0, \\ &\sup_{\psi \notin \Psi} \int_{\mathcal{Y}^{(n)}} [1 - u_n(y_1, \ldots, y_n)] \prod_{i=1}^n f(dy_i \mid \psi) \to 0, \quad \text{as } n \to \infty\,. \end{aligned} \tag{12}$$

Then, if $Y_i \overset{iid}{\sim} Q_{\psi^*}$ *for* $i = 1, 2, \ldots$ *with* Q_{ψ^*} *admitting density* $f(y \mid \psi^*)$*, it holds*

$$\left\| \mathcal{L}\left(d\tilde{\psi} \mid Y^{(n)}\right) - N\left(\mathcal{I}^{-1}(\psi^*)\Delta_{n,\psi^*}, \mathcal{I}^{-1}(\psi^*)\right) \right\|_{TV} \to 0, \qquad \text{as } n \to \infty$$

in $Q_{\psi^*}^{(\infty)}$*-probability, where* $\tilde{\psi} = \sqrt{n}(\psi - \psi^*)$ *and* $\Delta_{n,\psi^*} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla \log f(Y_i \mid \psi)\big|_{\psi=\psi^*}$.

Remark. Differentiability of $\sqrt{f(y \mid \psi)}$ and continuity of $\mathcal{I}(\psi)$ imply that the model is *differentiable in quadratic mean*, which allows to prove local asymptotic normality of the log-likelihood function. See Theorem 7.2 and Lemma 7.6 in [64].

Remark. A *test* is a measurable function $u : \mathcal{Y}^{(n)} \to [0, 1]$. The integrals in (12) represent probabilities of errors of first and second kind, respectively, when the null hypothesis $H_0 : \psi = \psi^*$ is rejected with probability $u(y_1, \dots, y_n)$.

Loosely speaking, Theorem 3.1 implies that, if the model is well-specified and ψ is suitably rescaled, the posterior distribution converges to a multivariate normal. The result holds under some identifiability requirements: first of all, the true parameter ψ^* must belong to the support of the prior; moreover, we must be able to separate ψ^* from the complements of its neighborhood, given infinitely many data. Such assumption is mild in most interesting cases and it is implied by the existence of uniformly consistent estimators for ψ (that is guaranteed if the support of p_0 is compact). See Chapter 10 in [64] for more details. Finally, the Fisher Information matrix must be non singular.

Remark. Notice that Theorem 3.1 requires the model to be (perfectly) well-specified, which rarely happens in practice. However there exist extended versions for the case of misspecified likelihoods [34], where the limiting distribution is still Gaussian with a different covariance matrix. Indeed, we expect the results of this and the following sections to hold in a similar way under misspecification: of course the different limiting distribution will have an impact on the final result, especially in the application of Corollary 2.6.

We can now use Theorem 2.4 and Corollary 2.5 to bound the mixing times of the Gibbs sampler associated to model (11) as n diverges.

Proposition 3.2. *Let model* (11) *satisfy the hypotheses of Theorem 3.1 and let P_n be the Gibbs sampler kernel targeting $\pi_n(d\psi) = p(d\psi \mid Y^{(n)})$ by updating one coordinate of $\psi = (\psi_1, \dots, \psi_K)$ at a time. Then, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$ there exists $T(\psi^*, \epsilon, M) < \infty$ such that*

$$\lim_{n\to\infty} Q^{(n)}_{\psi^*}\left(t^{(n)}_{mix}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) = 1\,.$$

Proposition 3.2 shows that, under the conditions of Theorem 3.1 and starting from an M-warm distribution, the number of iterations required to get ϵ-close to the posterior distribution does not grow as $n \to \infty$. An application to the normal model with unknown mean and precision is given by Corollary $C.7$ in Section $C.10$ of Appendix C.

The main take-away of this Section is that, under relatively mild conditions, the Gibbs sampler behaves well with models of fixed dimensionality and growing number of observations. In the remaining of the paper we consider the more challenging setting of hierarchical models, where the number of parameters grows with the number of observations: in particular we will explore situations in which the number of required iterations remains fixed even with a growing dimensionality of the problem.

4 Hierarchical models with exponential family priors and generic likelihood

We consider a general class of hierarchical models, with data divided in J groups, each having a set of group-specific parameters θ_j. The latter share a common prior with hyperparameters ψ. Recalling (1), the model under consideration is

$$Y_j \mid \theta_j \sim f(\cdot \mid \theta_j)\,, \quad \theta_j \mid \psi \overset{\text{iid}}{\sim} p(\cdot \mid \psi)\,, \quad \psi \sim p_0(\cdot). \tag{13}$$

We assume that the prior for $\theta_j \in \mathbb{R}^\ell$ belongs to the exponential family, that is

$$p(\theta \mid \psi) = h(\theta)\exp\left\{\sum_{s=1}^{S} \eta_s(\psi)T_s(\theta) - A(\psi)\right\}, \tag{14}$$

where $\psi \in \mathbb{R}^D$, $h : \mathbb{R}^\ell \to \mathbb{R}_+$ is a non-negative function and $\eta_s(\psi)$, $T_s(\theta)$ and $A(\psi)$ are known real-valued functions with domains $\mathbb{R}^D$, $\mathbb{R}^\ell$ and $\mathbb{R}^D$ respectively. We will always assume the family to be minimal, that is both $(\eta_1(\psi), \dots, \eta_S(\psi))$ and $(T_1(\theta), \dots, T_S(\theta))$ are linearly independent. On the other hand, we let $f(y \mid \theta)$ be an arbitrary likelihood function with data $y \in \mathbb{R}^m$ and parameters $\theta \in \mathbb{R}^\ell$, dominated by a suitable σ-finite measure (usually Lebesgue or counting one).

Denoting $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)$, $Y_{1:J} = (Y_1, \dots, Y_J)$ and $\pi_J(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi) = \mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi \mid Y_{1:J})$, we are interested in studying the two-block Gibbs sampler targeting $\pi_J(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi)$, i.e. the kernel defined as

$$P_J\left(\left(\boldsymbol{\theta}^{(t-1)}, \psi^{(t-1)}\right), \left(\mathrm{d}\boldsymbol{\theta}^{(t)}, \mathrm{d}\psi^{(t)}\right)\right) = \pi_J\left(\mathrm{d}\boldsymbol{\theta}^{(t)} \mid \psi^{(t-1)}\right) \pi_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{\theta}^{(t)}\right). \tag{15}$$

Throughout Section 4 we denote by $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ the Markov chain with operator P_J, and by $t^{(J)}_{mix}$ the associated mixing times, i.e.

$$t^{(J)}_{mix}(\epsilon, \mu) = \inf\left\{t \geq 0 : \left\|\mu P_J^t - \pi_J\right\|_{TV} < \epsilon\right\}, \quad t^{(J)}_{mix}(\epsilon, M) = \sup_{\mu \in \mathcal{N}(\pi_J, M)} t^{(J)}_{mix}(\epsilon, \mu).$$

4.1 Dimensionality reduction

In order to apply Corollary 2.5 to characterize $t^{(J)}_{mix}$, we would need to study the asymptotic distribution of π_J as $J \to \infty$. The latter is a distribution over $\ell J + D$ parameters, therefore its dimensionality grows with the size of the data. However, the next lemma shows that the convergence properties of P_J can be described through a Gibbs sampler on an intractable, but fixed-dimensional target, namely $\hat{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi) = \mathcal{L}(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J})$ where $\boldsymbol{T} = \left(\sum_{j=1}^J T_1(\theta_j), \dots, \sum_{j=1}^J T_S(\theta_j)\right)$, with T_s as in (14). Let $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1} = \left(\boldsymbol{T}(\boldsymbol{\theta}^{(t)}), \psi^{(t)}\right)_{t\geq 1}$ be the stochastic process obtained as a time-wise mapping of $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ under $(\boldsymbol{\theta}, \psi) \mapsto (\boldsymbol{T}(\boldsymbol{\theta}), \psi)$. The latter process contains all the information characterising the convergence of $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$, in the sense made precise in the following lemma. Below we denote by $\hat{P}_J$ the kernel of the two-block Gibbs sampler targeting $\hat{\pi}_J$.

Lemma 4.1. *For each $J \geq 1$, the process $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ is a Markov chain, its transition kernel coincides with $\hat{P}_J$, and its mixing times $\hat{t}^{(J)}_{mix}$ satisfy*

$$\sup_{\mu \in \mathcal{N}(\pi_J, M)} t^{(J)}_{mix}(\epsilon, \mu) = \sup_{\nu \in \mathcal{N}(\hat{\pi}_J, M)} \hat{t}^{(J)}_{mix}(\epsilon, \nu) \qquad (M, \epsilon) \in [1, \infty) \times (0, 1).$$

Remark (Prior and likelihood assumptions). In order to reduce the dimensionality of the Markov chain under consideration, Lemma 4.1 requires the existence of sufficient statistics only for the prior density of the group-specific parameters. It does not require any condition on the likelihood function in model (13). In particular, we have $\mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}) = \mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J})$, while $\mathcal{L}(\mathrm{d}Y_{1:J} \mid \boldsymbol{\theta}, \psi) \neq \mathcal{L}(\mathrm{d}Y_{1:J} \mid \boldsymbol{T}(\boldsymbol{\theta}), \psi)$ in general.

Lemma 4.1 allows to focus the analysis on the convergence speed of $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$, which is a chain whose dimensionality does not grow with the size of the data. Note that its target distribution $\hat{\pi}_J$ is usually not available in closed form, and the corresponding two-block Gibbs sampler $\hat{P}_J$ cannot be implemented directly (unless by implementing the original algorithm P_J and keeping track of $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$). In this sense the latter chain is useful for convergence analysis purposes but less so as an algorithmic shortcut.

The result of Lemma 4.1 is a peculiar property of the Gibbs sampler, which naturally ignores ancillary information about ψ in $\boldsymbol{\theta}$. Indeed, the proof of Lemma 4.1 crucially relies on the fact that the algorithm is performing exact conditional updates and analogous reductions do not occur for most other MCMC schemes (e.g. Metropolis-Hastings based schemes, including gradient-based ones).

This dimensionality reduction trick can be applied beyond hierarchical models and has already been employed in similar settings, mainly with the idea of obtaining suitable drift functions [58]: for example, in [48] it is used to derive the convergence complexity of a data augmentation algorithm for the Bayesian probit regression model, while in [50] a similar tecnique allows to study the geometric convergence rate of a Gibbs sampler for high dimensional Bayesian linear regression.

4.2 Regularity assumptions and main result

In order to apply the techniques of Theorem 2.4, we need to provide an asymptotic characterization of $\hat{\pi}_J$. To do so we require the technical assumptions listed in this section. The assumptions will be verified in specific examples in Section 5.1 and 5.2.

The approach we use to analyse $\hat{\pi}_J$, which is discussed after Theorem 4.2, is based on the decomposition $\hat{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi) = \hat{\pi}_J(\mathrm{d}\psi)\hat{\pi}_J(\mathrm{d}\boldsymbol{T} \mid \psi)$. The first set of assumptions contains standard regularity and identifiability conditions to study the marginal distribution $\hat{\pi}_J(\mathrm{d}\psi)$. In particular, assumptions $(B1)-(B3)$ allow the application of Theorem 3.1 to the posterior distribution of ψ. Their applicability has been discussed in Section 3. We denote the marginal likelihood of the model, obtained by integrating out the group specific parameter θ, as

$$g(y \mid \psi) = \int_{\mathbb{R}^\ell} f(y \mid \theta) p(\theta \mid \psi) \, \mathrm{d}\theta \,, \tag{16}$$

and its Fisher Information matrix as

$$[\mathcal{I}(\psi)]_{d,d'} = E\Big[\{\partial_{\psi_d} \log g(Y \mid \psi)\} \, \{\partial_{\psi_{d'}} \log g(Y \mid \psi)\}\Big], \quad d, d' = 1, \ldots, D.$$

We will assume the following:

- $(B1)$ There exists $\psi^* \in \mathbb{R}^D$ such that $Y_j \overset{\text{iid}}{\sim} Q_{\psi^*}$ for $j = 1, 2, \ldots$, where Q_{ψ^*} admits density $g(y \mid \psi^*)$. Moreover the map $\psi \to g(\cdot \mid \psi)$ is one-to-one and the map $\psi \to \sqrt{g(x \mid \psi)}$ is continuously differentiable for every x. Finally, the prior density p_0 is continuous and strictly positive in a neighborhood of ψ^*.
- $(B2)$ There exist a compact neighborhood Ψ of ψ^* and a sequence of tests $u_j : \mathbb{R}^{mJ} \to [0,1]$ such that $\int_{\mathbb{R}^{mJ}} u_j(y_1, \ldots, y_J) \prod_{j=1}^J g(y_j \mid \psi^*) \, \mathrm{d}y_{1:J} \to 0$ and $\sup_{\psi \notin \Psi} \int_{\mathbb{R}^{mJ}} [1 - u_j(y_1, \ldots, y_J)] \prod_{j=1}^J g(y_j \mid \psi) \, \mathrm{d}y_{1:J} \to 0$, as $J \to \infty$.
- $(B3)$ The Fisher Information matrix $\mathcal{I}(\psi)$ is non-singular and continuous w.r.t. ψ.

The second set of regularity assumptions (B4)-(B6) are described and discussed in Appendix B. They deal with smoothness and regularity of the conditional distribution $\hat{\pi}_J(\boldsymbol{T}|\psi)$ and they allow to derive a suitable conditional Central Limit Theorem in total variation for $\hat{\pi}_J(\boldsymbol{T}|\psi)$ as $J \to \infty$.

We can now state the main result of this section. Below we denote the product measures associated to Q_{ψ^*} by $Q_{\psi^*}^{(J)}$ and $Q_{\psi^*}^{(\infty)}$.

Theorem 4.2. *Consider model* (13) *and the Gibbs sampler defined as in* (15)*, with mixing times* $t_{mix}^{(J)}(\epsilon, M)$*. Then, under assumptions* $(B1)$*-*$(B6)$*, for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *there exists* $T(\psi^*, \epsilon, M) < \infty$ *such that*

$$Q_{\psi^*}^{(J)} \left(t_{mix}^{(J)}(\epsilon, M) \leq T(\psi^*, \epsilon, M) \right) \to 1,$$

as $J \to \infty$*. It follows that* $t_{mix}^{(J)}(\epsilon, M) = \mathcal{O}_P(1)$ *as* $J \to \infty$*.*

Remark. Theorem 4.2 provides a formal proof of the linear in J cost for Gibbs samplers on hierarchical models. Indeed, it proves that a bounded (in J) number of iterations suffices to get a good mixing: assuming that the cost of a single iteration scales linearly with J, which

is typically the case, this implies an overall computational cost of order $\mathcal{O}_P(J)$. Note that a single evaluation of the likelihood of $(\boldsymbol{\theta}, \psi)$, or the associated gradients, which is required at every iteration of usual gradient-based methods, yields a cost of the same order.

Remark. The conclusions of Theorem 4.2 are similar in spirit to those of [32, Thm.1]. Also there the convergence of Gibbs Samplers targeting two-level hierarchical models is studied using tools from Bayesian asymptotics. The results therein, which deal with convergence of ergodic averages when the algorithm is started in stationarity, are quite different from ours, which deal with mixing times. Nonetheless they also support the idea that Gibbs samplers targeting two-level hierarchical models can exhibit $\mathcal{O}_P(1)$ convergence as $J \to \infty$.

4.3 Posterior convergence lemmas for Theorem 4.2

The proof of Theorem 4.2 can be found in Appendix C. It relies on Lemma 4.1, which allows to focus on the two-blocks Gibbs sampler targeting $\hat{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi)$, and on Lemmas 4.3 and 4.4 below. These two lemmas imply that $\hat{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi)$ satisfies assumption $(A1)$ as $J \to \infty$ and that the associated limiting kernel is ergodic, thus allowing to apply Corollary 2.5.

In order to prove $(A1)$ for $\hat{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi) = \mathcal{L}\left(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J}\right)$, we need to identify a suitable transformation of $(\boldsymbol{T}, \psi)$, denoted by $\left(\tilde{\boldsymbol{T}}, \tilde{\psi}\right)$. We define a one-to-one transformation of ψ as

$$\tilde{\psi} = \sqrt{J}\left(\psi - \psi^*\right) - \Delta_J, \quad \Delta_J = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \mathcal{I}^{-1}(\psi^*) \nabla \log g(Y_j \mid \psi^*). \tag{17}$$

The asymptotic distribution of $\tilde{\psi}$ follows directly through Theorem 3.1, as summarized in the next lemma.

Lemma 4.3. *Define $\tilde{\psi}$ as in* (17)*. Under assumptions $(B1) - (B3)$ it holds*

$$\left\| \mathcal{L}(d\tilde{\psi} \mid Y_{1:J}) - N\left(\mathbf{0}, \mathcal{I}^{-1}(\psi^*)\right) \right\|_{TV} \to 0,$$

as $J \to \infty$, in $Q_{\psi^*}^{(\infty)}$-probability.

Let $M^{(1)}(\psi \mid y) = \left(M_1^{(1)}(\psi \mid y), \ldots, M_S^{(1)}(\psi \mid y)\right) \in \mathbb{R}^S$ with $M_s^{(1)}(\psi \mid y) = E\left[T_s(\theta_j) \mid Y_j = y, \psi\right]$ and

$$[C(\psi)]_{s,d} = E_{Y_j}\left[\partial_{\psi_d} M_s^{(1)}\left(\psi \mid Y_j\right)\right], \quad [V(\psi)]_{s,s'} = E_{Y_j}\left[\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi\right)\right], \tag{18}$$

with $s, s' = 1, \ldots S$ and $d = 1, \ldots, D$. We use the notation $E_{Y_j}[\cdot]$ for expectations with respect to the law of Y_j as defined in $(B1)$. Then we define a one-to-one transformation of $\boldsymbol{T}$ as

$$\tilde{\boldsymbol{T}} = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \left[T(\theta_j) - M^{(1)}\left(\psi^* \mid Y_j\right)\right] - C(\psi^*)\Delta_J, \tag{19}$$

with $C(\psi^*)$ defined in (38). The next lemma proves the required asymptotic normality of $\tilde{\boldsymbol{T}}$, conditional to $\tilde{\psi}$.

Lemma 4.4. *Let $\tilde{\boldsymbol{T}}$ be as in* (19)*. Under assumptions $(B1)$-$(B6)$ for every $\tilde{\psi}$ it holds*

$$\left\| \mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) - N\left(C(\psi^*)\tilde{\psi}, V(\psi^*)\right) \right\|_{TV} \to 0,$$

as $J \to \infty$, for $Q_{\psi^*}^{(\infty)}$-almost every $(Y_1, Y_2, \ldots)$.

Lemma $C.18$ in Section $C.14$ of Appendix C combines Lemmas 4.3 and 4.4 to prove that $\mathcal{L}(\mathrm{d}\tilde{\boldsymbol{T}}, \tilde{\psi} \mid Y_{1:J})$ converges in total variation to a multivariate Gaussian vector with non singular covariance matrix, which allows to apply Corollary 2.5 as desired.

Remark. The definition of $\tilde{\boldsymbol{T}}$ and Lemma 4.4 are an important part of the proof of Theorem 4.2. Lemma 4.4 relies on the fact that, conditional to $\tilde{\psi}$ and $Y_{1:J}$, $\boldsymbol{T}$ is a sum of independent (but not identically distributed) terms. The proof of convergence in total variation requires more than the usual tools from Lindeberg-Feller Central Limit Theorem, as discussed in Appendix B after assumptions $(B5)$ and $(B6)$.

4.4 Analysis of the limiting chain

As a byproduct of the proof of Theorem 4.2, it is possible to characterize the limiting distribution of the rescaled vector $\left(\tilde{\boldsymbol{T}}, \tilde{\psi}\right)$, as the next proposition shows.

Proposition 4.5. *Consider the same assumptions of Theorem 4.2. Then*

$$\left\| \mathcal{L}(d\tilde{\boldsymbol{T}}, d\tilde{\psi} \mid Y_{1:J}) - N\left(\mathbf{0}, \Sigma\right) \right\|_{TV} \to 0,$$

as $J \to \infty$, in $Q_{\psi^*}^{(\infty)}$-probability, where

$$\Sigma = \begin{bmatrix} V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & C(\psi^*)\mathcal{I}^{-1}(\psi^*) \\ \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & \mathcal{I}^{-1}(\psi^*) \end{bmatrix} \tag{20}$$

with $C(\psi^*)$ and $V(\psi^*)$ defined in (38).

The expression for the limiting covariance in (20) can be used to investigate the convergence properties of the limiting Gibbs sampler, since the spectral gap is explicitly computable from that. We can then apply Corollary 2.6 and obtain the following result.

Corollary 4.6. *Under the assumptions of Theorem 4.2, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$, we have $Q_{\psi^*}^{(J)}\left(t_{mix}^{(J)}(\epsilon, M) \leq T\left(\psi^*, \epsilon, M\right)\right) \to 1$ as $J \to \infty$, with*

$$T\left(\psi^*, \epsilon, M\right) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma(\psi^*)\right)},$$
$$\gamma(\psi^*) = \min\left\{ \frac{1}{1+\lambda_i} \,:\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) \right\}.$$

Thus, once the limiting distribution is obtained, an upper bound on the mixing times can be derived by computing the eigenvalues of a $S \times S$ matrix. As an application, the next corollary provides the value of γ when $S = D = 1$.

Corollary 4.7. *Consider the same setting of Corollary 4.6, with $S = D = 1$. Then we have*

$$\gamma(\psi^*) = \frac{Var_{Y_j}\left(E\left[T(\theta_j) \mid \psi^*, Y_j\right]\right)}{Var\left(T(\theta_j) \mid \psi^*\right)}. \tag{21}$$

By the law of total variance, we have that $\gamma(\psi^*) \to 0$ if and only if

$$\frac{\mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid \psi^*, Y_j\right]\right)}{E\left[\mathrm{Var}_{Y_j}\left(T(\theta_j) \mid \psi^*, Y_j\right)\right]} \to 0,$$

i.e., loosely speaking, when the data Y_j yield little information about $T(\theta_j)$ and therefore about ψ. This phenomenon arises since model (13) is an example of centered parametrization, see e.g. [23, 43, 44]. The formula in (21) resembles the definition of the so-called Bayesian fraction of missing information [35], with the notable difference of not involving an infimum over a set of test functions.

5 Examples

In this section various examples, which differ by the choice of likelihoods and priors, are discussed.

5.1 Hierarchical normal model

Consider the following hierarchical specification:

$$\begin{aligned} Y_{j,i} \mid \theta_j &\sim N\left(\theta_j, \tau_0^{-1}\right), \quad i = 1, \dots, m, \; j = 1, \dots, J \\ \theta_j \mid \mu, \tau_1 &\overset{\text{iid}}{\sim} N(\mu, \tau_1^{-1}), \qquad\qquad j = 1, \dots, J \\ (\mu, \tau_1) &\sim p_0(\cdot)\,. \end{aligned} \tag{22}$$

where (μ, τ_1) are unknown hyperparameters. In this section we assume τ_0 to be fixed and known, see Section 5.3.1 for the case with τ_0 unknown. The prior p_0 can be any distribution satisfying the assumptions stated in Proposition 5.1 below. It can be seen that (22) is a particular case of model (13), with $f(Y_j \mid \theta_j) = \prod_{i=1}^m N(Y_{j,i} \mid \theta_j, \tau_0^{-1})$, $p(\cdot \mid \mu, \tau_1) = N(\mu, \tau_1^{-1})$. The marginal likelihood of Y_j conditional to (μ, τ_1, τ_0) is given by

$$g(y \mid \mu, \tau_1, \tau_0) = N\left(y \mid \mu, \tau_0^{-1} I + \tau_1^{-1} \mathbb{H}\right) \qquad\qquad y \in \mathbb{R}^m, \tag{23}$$

where I is the $m \times m$ identity matrix and $\mathbb{H}$ is the $m \times m$ matrix of ones.

We consider three Gibbs sampler specifications, which vary depending on which parameters are unknown and treated as random and which blocking rules are used. First, when τ_1 is fixed, we define P_1 as the transition kernel of the Gibbs sampler that targets $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \mu, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\mu \mid \boldsymbol{\theta}, Y_{1:J}\right)$. If instead μ and τ_1 are unknown, we define P_2 and P_3 as the transition kernels of the two Gibbs samplers targeting $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu, \mathrm{d}\tau_1 \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid \tau_1, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$ for P_2; and $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \tau_1, Y_{1:J}\right)$, $\mathcal{L}\left(\mathrm{d}\mu \mid \boldsymbol{\theta}, \tau_1, Y_{1:J}\right)$, $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$ for P_3. In the following we will show that the asymptotic behaviour of P_2 and P_3 is essentially the same.

It is possible to prove that P_1 falls directly in the setting of Theorem 4.2, with $T(\theta_j) = \theta_j$ for P_1. Even if P_2 and P_3 are not exactly particular cases of the general theorem, since different update schemes are considered, it turns out that they can be studied with the same tools introduced in the previous section, with $T(\theta_j) = \left(\theta_j, (\theta_j - \mu^*)^2\right)$.

The next proposition shows that the settings introduced above lead to well-behaved asymptotic regimes. Here $t^{(J)}_{mix,l}(\epsilon, M)$ denotes the mixing times of the Gibbs sampler defined by P_l with $l \in \{1, 2, 3\}$.

Proposition 5.1. *Let $Y_j \overset{iid}{\sim} Q_{\psi^*}$, with Q_{ψ^*} admitting density $g(y \mid \psi^*)$ as in (23), where $\psi^* = (\mu^*, \tau_1^*, \tau_0^*)$, and consider model (22) with $\tau_0 = \tau_0^*$. Consider the Gibbs sampler with operator P_l, with $l \in \{1, 2, 3\}$, and let the prior density p_0 be continuous and strictly positive in a neighborhood of μ^* when $l = 1$ and (μ^*, τ_1^*) when $l \in \{2, 3\}$. Finally, when $l = 1$ let $\tau_1 = \tau_1^*$. Then for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$ there exists $T_l(\psi^*, \epsilon, M) < \infty$ such that*

$$Q^{(J)}_{\psi^*}\left(t^{(J)}_{mix,l}(\epsilon, M) \leq T_l(\psi^*, \epsilon, M)\right) \to 1 \qquad\qquad \text{as } J \to \infty, \; l = 1, 2, 3\,. \tag{24}$$

Under model (22), the matrices in Corollary 4.6 can be explicitly computed, leading to the following result.

Corollary 5.2. *Under the same assumptions and notation of Proposition 5.1, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$, (24) holds with*

$$T_l(\psi^*, \epsilon, M) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma_l(\psi^*)\right)}, \qquad\qquad l = 1, 2, 3\,,$$

where

$$\gamma_1(\psi^*) = \left(1 + \frac{\tau_1^*}{m\tau_0^*}\right)^{-1} \quad \textit{and} \quad \gamma_2(\psi^*) = \gamma_3(\psi^*) = \gamma_1(\psi^*)^2\,. \tag{25}$$

The expressions for the asymptotic gaps in (25) are insightful in many ways. First, μ^* does not appear in any of the spectral gaps, meaning that the limiting value of the mean parameter seems not to play a role in the asymptotic behaviour of the Gibbs sampler. Moreover, the gaps are a function of the ratio $(m\tau_0^*)^{-1}\tau_1^*$, that is the ratio of the prior and likelihood precisions, respectively. In particular the gaps converge to 0, i.e. the upper bound on the mixing times diverges, if and only if $(m\tau_0^*)^{-1}\tau_1^* \to \infty$, which happens when the prior is increasingly more informative than the data. As discussed after Corollary 4.7, such phenomenon arises since all the three formulations are an example of centered parametrization [23, 43]. On the contrary, the gaps converge to 1, i.e. asymptotically a single iteration suffices, if and only if $(m\tau_0^*)^{-1}\tau_1^* \to 0$.

When τ_1 is fixed and $p_0(\mu)$ is Gaussian, then $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid Y_{1:J}\right)$ is a multivariate Gaussian and P_1 is amenable to finite-sample analysis. In fact, the expression for $\gamma_1(\psi^*)$ appeared previously in the literature, see e.g. [43]. The result in Corollary 5.2 is, however, different since it is asymptotic and it applies also to general priors.

On the contrary, a finite-sample analysis of P_2 are P_3 is hard even when $p_0(\mu)$ is Gaussian (see e.g. [30, 49, 67]) and $\gamma_2(\psi^*)$ and $\gamma_3(\psi^*)$ did not appear previously in the literature, to the best of our knowledge. It is interesting that, regardless of the value of $(m, \mu^*, \tau_1^*, \tau_0^*)$, including the random precision parameter, when moving from P_1 to either P_2 or P_3, always slows down the sampler (asymptotically), since $\gamma_1(\psi^*) > \gamma_i(\psi^*)$ for $i = 2, 3$, and that the two blocking rules of P_2 and P_3 are asymptotically equivalent in terms of mixing times, since $\gamma_2(\psi^*) = \gamma_3(\psi^*)$.

5.2 Models with binary and categorical data

Let now $f(y \mid \theta)$ be a probability mass function, whose point masses are denoted by $y_0, \ldots, y_m$, with $m < \infty$, such that for every $\theta \in \mathbb{R}^K$ we have

$$\sum_{r=0}^{m} f(y_r \mid \theta) = 1, \quad f(y_r \mid \theta) > 0, \quad r = 0, \ldots, m. \tag{26}$$

The assumption in (26) is mild and holds for most likelihoods usually employed with categorical data, e.g. multinomial logit and probit. We focus on hierarchical models with normal priors, i.e.

$$Y_j \mid \theta_j \sim f(Y_j \mid \theta_j)\,, \quad \theta_1, \ldots, \theta_J \mid \mu, \tau \overset{\text{iid}}{\sim} N(\mu, \tau^{-1})\,, \quad (\mu, \tau) \sim p_0(\cdot)\,. \tag{27}$$

For example the case $f(y \mid \theta) = \binom{m}{y} \frac{e^{y\theta}}{(1+e^{\theta})^m}$, with $y = 0, \ldots, m$, corresponds to the logistic hierarchical model with Gaussian random effects. The prior p_0 can be any distribution satisfying the assumptions stated in Proposition 5.4 below. We define P as the transition kernel of the Gibbs sampler that targets $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu, \mathrm{d}\tau \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \mu, \tau, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\mu, \mathrm{d}\tau \mid \boldsymbol{\theta}, Y_{1:J}\right)$. This is a particular case of the setting of Theorem 4.2, with $\psi = (\mu, \tau)$ and $T(\theta_j) = (\theta_j, \theta_j^2)$. Notice that usually $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \mu, \tau, Y_{1:J}\right)$ is not known in closed form (with the notable exception of the probit case, see [17]), but nonetheless exact sampling is often feasible through adaptive rejection sampling (see e.g. [26]) since each θ_j is one dimensional. The marginal likelihood is given by

$$g(y \mid \psi) = \int_{\mathbb{R}} f(y \mid \theta) N\left(\theta \mid \mu, \tau^{-1}\right) \mathrm{d}\theta. \tag{28}$$

The next lemma shows that assumptions $(B4)$-$(B6)$ follow directly from (27).

Lemma 5.3. *Consider model* (27) *and let* $Y_j \overset{iid}{\sim} Q_{\psi^*}$*, with* Q_{ψ^*} *admitting density* $g(y \mid \psi^*)$ *as in* (28)*, with* $\psi^* = (\mu^*, \tau^*)$*. Then assumptions* $(B4)$-$(B6)$ *are satisfied.*

Thus, in order to apply Theorem 4.2, it suffices to prove assumptions $(B2)$ and $(B3)$, i.e. that the parameters ψ are identifiable with non singular Fisher Information matrix. Therefore, as formalized in the next proposition, standard identifiability conditions (which are also necessary to consistently estimate ψ) are sufficient to prove boundedness of the mixing times.

Proposition 5.4. *Consider model* (27) *and let* $Y_j \overset{iid}{\sim} Q_{\psi^*}$, *with* Q_{ψ^*} *admitting density* $g(y \mid \psi^*)$ *as in* (28), *where* $\psi^* = (\mu^*, \tau^*)$. *Consider the Gibbs sampler with operator* P *and let* p_0 *be continuous and strictly positive in a neighborhood of* ψ^*. *Let the map* $\psi \to g(\cdot \mid \psi)$ *be one-to-one, with non singular and continuous* $\mathcal{I}(\psi)$. *Finally, assume tests as in* $(B2)$ *exist. Then for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *there exists* $T(\psi^*, \epsilon, M) < \infty$ *such that*

$$Q_{\psi^*}^{(J)} \left(t_{mix}^{(J)}(\epsilon, M) \leq T(\psi^*, \epsilon, M) \right) \to 1 \qquad \text{as } J \to \infty .$$

Remark. In most cases $m \geq 2$ is required to avoid the pair (μ, τ) being not identifiable and the associated Fisher Information matrix being singular. For example Lemma $C.35$ in Section $C.23$ of Appendix C shows that with the logit link $\mathcal{I}(\psi)$ is singular if and only if $m = 1$.

As already discussed in the Section 1, the results of Proposition 5.4 are illustrated on simulated data in Figure 1. Since mixing times are very hard to approximate numerically in high-dimensions, we employ the Integrated Autocorrelation Times (IATs) as an empirical measure of convergence time. The IAT associated to a π-invariant Markov chain $X = \{X^{(t)}\}_{t \geq 1}$ and a test function $f \in L^2(\pi)$ is defined as

$$\text{IAT}(f) = 1 + 2 \sum_{t=2}^{\infty} \text{Corr} \left(f(X^{(1)}), f(X^{(t)}) \right) . \tag{29}$$

Loosely speaking, $\text{IAT}(f)$ is the number of MCMC samples that is equivalent to a single independent sample in terms of estimation of $\int f(x)\pi(dx)$, thus the higher IAT the slower the convergence. When dealing with hierarchical models as in (27), we compute the maximum IAT over all the parameters (both global and group specific). We estimate the IAT with the ratio of the number of iterations and the effective sample size, as described in [27], with the effective sample size computed with the R package *mcmcse* [21]. For a review of different methods to estimate the IATs, see [63]. In Figure 1 we plot the quantiles of the IATs as a function of the number of groups for the Gibbs sampler, implemented using adaptive rejection sampling [26] for the exact updates of local parameters with full conditionals $\mathcal{L}(\mathrm{d}\theta_j \mid \mu, \tau, Y_{1:J})$. As expected by Proposition 5.4, the IATs do not diverge as J increases for both values of m under consideration. Note that variability decreases as J increases and the posterior gets closer to its asymptotic limit.

Corollary 5.5. *Consider the same setting of Proposition 5.4. For every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *define*

$$T(\psi^*, \epsilon, M) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma(\psi^*)\right)},$$

for $\gamma(\psi^*) \in (0, 1)$ *as in Corollary 4.6. Then*

$$Q_{\psi^*}^{(J)} \left(t_{mix}^{(J)}(\epsilon, M) \leq T(\psi^*, \epsilon, M) \right) \to 1 \qquad \text{as } J \to \infty .$$

The study of the limiting spectral properties, i.e. of $\gamma(\psi^*)$, can be useful to predict under which scenarios the Gibbs sampler will perform well or not for large J. We illustrate this by considering model (27) with logit link and known τ set to 1. In this setting, where μ is the only global parameter, the value of $\gamma(\psi^*)$ can be computed as in (21) through simple one-dimensional numerical integration. In Figure 2 we compare the resulting mixing time upper bound, $T(\psi^*, \epsilon, M)$, with the numerical estimates of IATs defined in (29), obtained

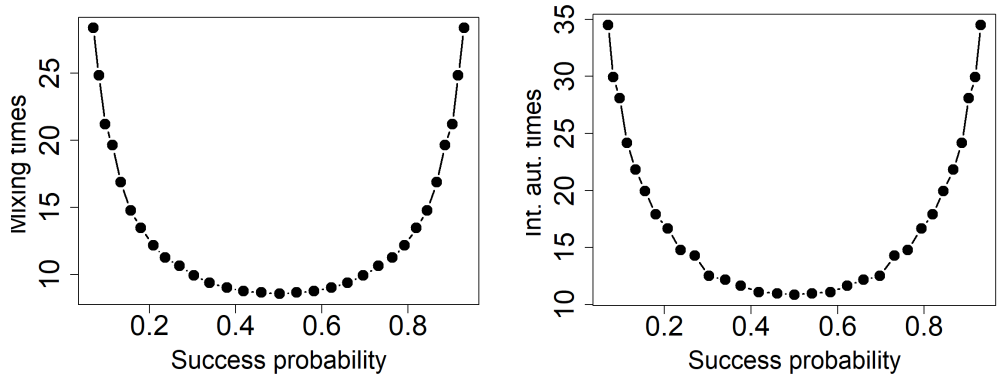

Figure 2: Left: upper bounds on mixing times for model (27) with τ known, where $\tau^* = 1$, $\mu^* \in (-3, 3)$, $m = 1$, $M = 2$ and $\epsilon = 0.2$. A priori $\mu \sim N\left(0, 10^3\right)$. Right: median IATs with $J = 2000$.

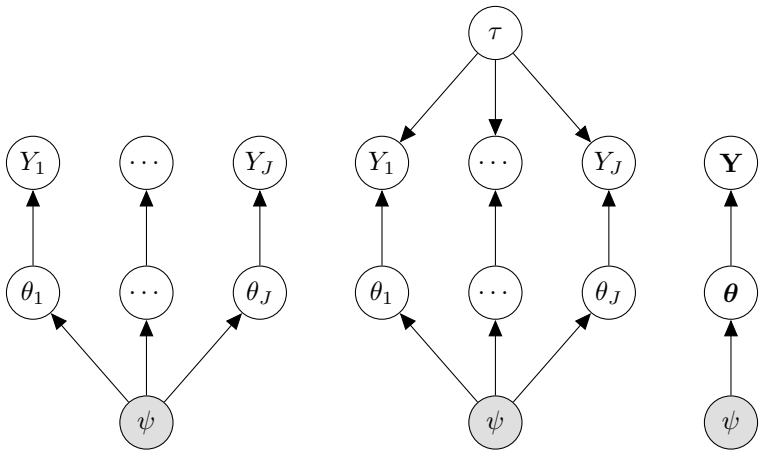

Figure 3: Graphical models of different hierarchical structures. Left: one level nested model as in Theorem 4.2. Center: hyperparameters specifying the likelihood. Right: dependent latent parameters.

by running a long MCMC chain with a moderately large value of J. We compare such quantities for different values of the true success probability induced by μ^*, i.e. $\int_{\mathbb{R}} f(1 \mid \theta) N\left(\theta \mid \mu^*, 1\right) \mathrm{d}\theta$. Both theoretical and empirical measures of convergence highlight that the performances of the Gibbs sampler deteriorate when the problem is not balanced: such conclusion is coherent with the findings in [31], that considers an asymptotic regime with increasing imbalancedness.

5.3 Different graphical models structure

In the previous subsections we have studied applications of Theorem 4.2 for some specification of the hierarchical model in (13). These correspond to the graphical models in the leftmost panel of Figure 3. While this structure is very common in Bayesian modeling and it constitutes our main motivating application, the techniques we developed - and in particular the dimensionality reduction and posterior asymptotic approach - can be applied to different classes of models, including other widely used ones. Here we provide two examples, the first is a relatively direct extension of the model in (13) with the addition of parameters in the likelihood, the second is a more different setting of Gaussian Process regression where the latent parameters are not independent. See respectively the center and rightmost panels in Figure 3 for the resulting graphical models. More generally, we expect our methodology to be potentially useful to analyse samplers for models that fea-

ture a fixed set of hyperparameters ψ, conditional to which a growing set of parameters or latent variables is tractable enough for posterior sampling.

5.3.1 Likelihood parameters

Consider again the hierarchical normal model

$$Y_{j,i} \mid \theta_j, \tau_0 \sim N\left(\theta_j, \tau_0^{-1}\right), \quad \theta_j \mid \mu, \tau_1 \overset{\text{iid}}{\sim} N(\mu, \tau_1^{-1}), \quad (\mu, \tau_1, \tau_0) \sim p_0(\cdot), \tag{30}$$

with $i = 1, \dots, m$ and $j = 1, \dots, J$. The unknown parameters are now given by the triplet $\psi = (\mu, \tau_1, \tau_0)$. We denote with P the transition kernel of the Gibbs sampler targeting $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu, \mathrm{d}\tau_1, \mathrm{d}\tau_0 \mid Y_{1:J})$ by alternating updates from $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid \tau_1, \tau_0, Y_{1:J})$ and $\mathcal{L}(\mathrm{d}\tau_1, \mathrm{d}\tau_0 \mid \boldsymbol{\theta}, \mu, Y_{1:J})$. This cannot be seen as a specific case of Theorem 4.2 with $\psi = (\mu, \tau_1, \tau_0)$, since τ_0 is a parameter of the likelihood f and therefore there is no conditional independence between Y_j and ψ, given θ_j. However, an approach similar to the one of the previous section can be employed. In particular, a result analogous to Lemma 4.1 can be derived, with $T(\theta_j) = \left(\left(\theta_j - \bar{Y}_j\right)^2, (\theta_j - \mu)^2\right)$ playing the role of the sufficient statistics and $\bar{Y}_j = \frac{1}{m}\sum_{i=1}^m Y_{j,i}$. It is interesting to notice that T in this case depends also on the data $Y_{1:J}$, exactly because the group specific parameters $\boldsymbol{\theta}$ do not contain all the information regarding ψ. The next proposition shows that also this specification leads to a well-behaved asymptotic regime.

Proposition 5.6. *Consider model* (30) *with* $m \geq 2$ *and let* $Y_j \overset{iid}{\sim} Q_{\psi^*}$*, with* Q_{ψ^*} *admitting density* $g(y \mid \psi^*)$ *as in* (23)*, where* $\psi^* = (\mu^*, \tau_1^*, \tau_0^*)$*. Consider the Gibbs sampler with operator* P *and let the prior density* p_0 *be a continuous and strictly positive in a neighborhood of* ψ^**. Then for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *there exists* $T(\psi^*, \epsilon, M) < \infty$ *such that*

$$Q_{\psi^*}^{(J)}\left(t_{mix}^{(J)}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) \to 1 \qquad \text{as } J \to \infty. \tag{31}$$

An explicit value for $T(\psi^*, \epsilon, M)$ can be found through Corollary 2.6, as shown in the next corollary.

Corollary 5.7. *Consider the same setting of Proposition 5.6. Then, for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$*,* (31) *holds with*

$$T(\psi^*, \epsilon, M) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma(\psi^*)\right)},$$

where

$$\gamma(\psi^*) = \left(1 + \frac{1}{m-1}\left(1 - \frac{\tau_1^*}{m\tau_0^*}\right)^2 + \left(\frac{\tau_1^*}{m\tau_0^*}\right)^2\right)^{-1}.$$

Remark. The assumption $m \geq 2$ cannot be relaxed: indeed, if a single observation per group is available, the pair (τ_1, τ_0) is not identifiable and the Fisher Information matrix is singular. For an empirical illustration of the issues arising in this context, see the top left panel in Figure 4 or Section 6.2 of [50].

Unlike the case of Corollary 5.2, in this setting the limiting gap does not depend on m only through the ratio of prior and likelihood precisions, but also directly on its value. Loosely speaking, a higher value of m allows to better recover the relation between τ_0 and τ_1.

The results of Proposition 5.6 and Corollary 5.7 are illustrated on simulated data in Figure 4, which depicts the Integrated Autocorrelations Times (IATs) as defined in (29). When the model is not identifiable, i.e. $m = 1$ (top left panel), the IATs diverge with the number of groups, while with $m = 3$ and $m = 5$ they stabilize as J increases. Differently from the binomial setting of Figure 4, the IATs grow for small values of J before the asymptotic regime kicks in.

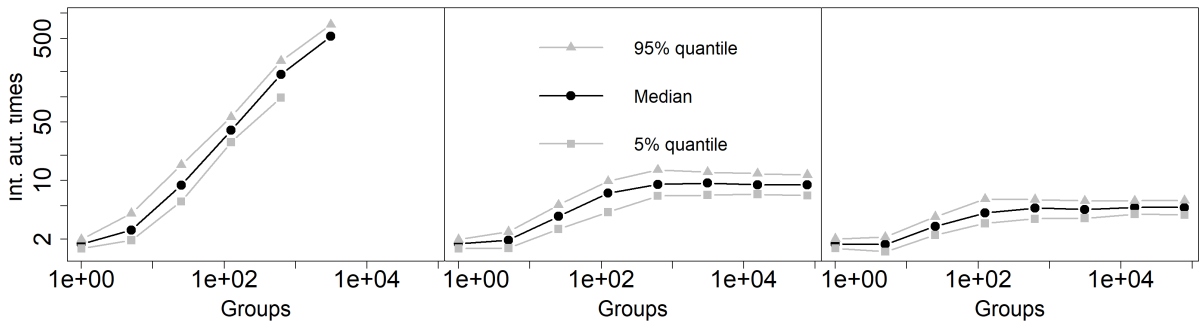

Figure 4: Quantiles of the integrated autocorrelations times (on log-scale) for model (30) with $\mu^* = 4$, $\tau_0^* = 1$ and $\tau_1^* = 3$. A priori $(\tau_0, \tau_1) \overset{\text{i.i.d.}}{\sim} \text{Gamma}(1,1)$ and $p_0(\mu) \propto 1$. Top left: $m = 1$ (last points not plotted due to numerical instability). Center: $m = 3$. Top right: $m = 5$.

5.3.2 Gaussian processes

We now consider the popular setting where the groups are identified by a continuous covariate (e.g. location) and group specific parameters are modeled through a Gaussian process. It turns out that the main arguments of the paper, namely dimensionality reduction and impact of posterior asymptotic characterization, can be applied also in this context. This section, compared to the previous ones, aims to provide a proof of concept rather than a detailed analysis, e.g. we directly assume limiting statements on the posterior distributions of interest. Nonetheless we find it useful to show how widely our methodology could be applied and illustrate interesting directions of ongoing work.

Assume to observe n data points $Y(s_i)$ with $i = 1, \dots, n$, at a set of locations $(s_1, \dots, s_n)$, together with input variables or covariates $x(s_i) \in \mathbb{R}$. We consider Gaussian Process regression models of the form

$$\begin{aligned} Y(s_i) \mid \boldsymbol{\beta} &\sim f(\cdot \mid \beta(s_i), x(s_i)), \quad i = 1, \dots, n \\ \boldsymbol{\beta}^{(n)} \mid \psi &\sim N(\theta \mathbf{1}, \tau_\beta^{-1} R^{(n)}) \\ \psi &\sim p_0(\cdot). \end{aligned} \tag{32}$$

where $\boldsymbol{\beta} = (\beta(s_1), \dots, \beta(s_n))^\top$ is a Gaussian Process (GP) observed at $(s_1, \dots, s_n)$ and f is a density function with respect to a suitable dominating measure. Here $\mathbf{1}_n = (1, \dots, 1)^\top$ is an n-dimensional vector and $R^{(n)} = (R_{ij})_{i,j=1,\dots,n}$ is a $n \times n$ correlation matrix, with $R_{ij} = \text{Corr}\,(\beta(s_i), \beta(s_j))$, defined through a suitable kernel function, that we assume to be fixed and known. Typically, strength of correlation among coefficients at different locations depends on their distance, with R_{ij} defined e.g. through a kernel of the Matérn family (see e.g. Section 4.2.1 in [65]). In this Section we focus on a single real covariate for notational convenience, but everything could be restated on a general p-dimensional space with little effort: direct analogues of the next lemma and corollaries similarly follow. We first consider cases where the likelihood function has no specific hyper-parameters, such as in the common binary case where $Y(s_j) \mid \boldsymbol{\beta} \sim \text{Bernoulli}(\sigma(\beta(s_j)x(s_j)))$, with σ logistic link function and $Y(s_j) \in \{0, 1\}$.

Let P_n be the kernel of the Gibbs sampler which targets $\pi_n(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta) = \mathcal{L}\left(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta \mid Y^{(n)}\right)$, by sequentially performing updates from the full conditionals of $\boldsymbol{\beta}$, θ and τ_β. Despite the different graphical model structure, the analysis of mixing times of P_n as $n \to \infty$ can be approached with the techniques we developed above, regardless of the specific likelihood used in (32). The first step is to perform a dimensionality reduction analogous to the one in Section 4.1. Define $\psi = (\theta, \tau_\beta)$ and $\boldsymbol{T}(\boldsymbol{\beta}) = \left(T_\theta, T_{\tau_\beta}\right)$, where $T_\theta = \mathbf{1}^\top R^{-1} \boldsymbol{\beta}$, $T_{\tau_\beta} = \boldsymbol{\beta}^\top R^{-1} \boldsymbol{\beta}$, which play the same role of global parameters and sufficient statistics in

Lemma 4.1. Indeed it holds $\mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\beta}, Y^{(n)}\right) = \mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\beta}), Y^{(n)})$ and we can provide an analogue of Lemma 4.1 for model (32).

Lemma 5.8. *Let π_n and P_n be defined as above for model (32). Let $\hat{P}_n$ be the transition kernel of Gibbs sampler targeting $\hat{\pi}_n(d\boldsymbol{T}, d\theta, d\tau_\beta) = \mathcal{L}\left(d\boldsymbol{T}, d\theta, d\tau_\beta \mid Y^{(n)}\right)$ which sequentially performs updates from the full conditionals of $\boldsymbol{T}$, θ and τ_β. Let $(\boldsymbol{T}^{(t)}, d\theta^{(t)}, d\tau_\beta^{(t)})_{t\geq 1}$ be the stochastic process obtained as a time-wise transformation of $(\boldsymbol{\beta}^{(t)}, d\theta^{(t)}, d\tau_\beta^{(t)})_{t\geq 1}$. Then $(\boldsymbol{T}^{(t)}, d\theta^{(t)}, d\tau_\beta^{(t)})_{t\geq 1}$ is a Markov chain, its transition kernel coincides with $\hat{P}_n$, and its mixing times $\hat{t}_{mix}^{(n)}$ satisfy*

$$\sup_{\mu \in \mathcal{N}(\pi_n, M)} t_{mix}^{(n)}(\epsilon, \mu) = \sup_{\nu \in \mathcal{N}(\hat{\pi}_n, M)} \hat{t}_{mix}^{(n)}(\epsilon, \nu) \qquad\qquad M \geq 1\,.$$

Also, provided a rescaled version of $(\boldsymbol{T}, \theta, \tau_\beta)$ converges to a suitable limit conditional on the data, the mixing times are bounded with respect to the number of observations.

Corollary 5.9. *Under model (32), let $\hat{\pi}_n$ satisfy assumption (A1) for a given data generating process $Y^{(n)} \sim Q^{(n)}$, with limiting distribution $\tilde{\pi}$. If $(M, \epsilon) \in [1, \infty) \times (0, 1)$ is such that $\tilde{t}_{mix}(\epsilon, M) < \infty$, then it holds*

$$Q^{(n)}\left(t_{mix}^{(n)}(\epsilon, M) \leq \tilde{t}_{mix}(\epsilon, M)\right) \to 1 \qquad \text{as } n \to \infty\,. \tag{33}$$

In some cases the likelihood contains some unknown parameters that are also included in the Bayesian model. A common example is the likelihood precision τ_ϵ in normal linear models with spatially varying regression coefficients (see e.g. [22] or Section 2 in [65]), where

$$Y(s_i) \mid \boldsymbol{\beta} \sim N(\beta(s_i)x(s_i), \tau_\epsilon^{-1}), \qquad\qquad i = 1, \ldots, n. \tag{34}$$

Let P_n be the Gibbs sampler kernel targeting $\pi_n(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta, \mathrm{d}\tau_\epsilon) = \mathcal{L}\left(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta, \mathrm{d}\tau_\epsilon \mid Y^{(n)}\right)$, by sequentially performing updates from the full conditionals of $\boldsymbol{\beta}$, θ, τ_β and τ_ϵ. Analogously to Section 5.3.1, the results of Lemma 5.8 and Corollary 5.9 extend to this context with $\psi = (\theta, \tau_\beta, \tau_\epsilon)$ and $\boldsymbol{T}$ defined as $\boldsymbol{T} = \left(T_\theta, T_{\tau_\beta}, T_{\tau_\epsilon}\right)$, where $T_{\tau_\epsilon} = \left(Y^{(n)} - D\boldsymbol{\beta}\right)^\top \left(Y^{(n)} - D\boldsymbol{\beta}\right)$ and D is the $n \times n$ diagonal matrix with values $(x(s_1), \ldots, x(s_n))$. This is summarized in the next corollary.

Corollary 5.10. *Under model (32) with likelihood as in (34), assume the conditions of Corollary 5.9 are satisfied with $\psi = (\theta, \tau_\beta, \tau_\epsilon)$ and $\boldsymbol{T} = \left(T_\theta, T_{\tau_\beta}, T_{\tau_\epsilon}\right)$. Then (33) holds.*

Similarly to the hierarchical normal case, studied in Section 5.1, if the precisions $(\tau_\beta, \tau_\epsilon)$ are fixed in specification (34), then the spectral gap of P_n can be explicitly studied to deduce limiting bounds on mixing times (see e.g. [5]); while if the precisions are unknown, as it is mostly the case in applications, the performances of P_n have only been empirically studied through simulations. The methodology we introduce here can be used to formally analyze the behaviour of these samplers as $n \to \infty$.

To conclude this section, it is important to note that in this context the kernel P_n may or may not be directly implementable, depending on the specific model formulation. In the commonly used linear case, the full conditional distribution $\pi_n(\mathrm{d}\boldsymbol{\beta} \mid \psi)$ is normal, so that sampling becomes accessible and P_n is directly the algorithm used to sample from π_n. See e.g. Appendix 2 of [5] for details on the implementation, including expressions for the full conditionals. In other cases, e.g. for log-concave likelihoods such as the binary regression ones, adaptive rejection sampling techniques (e.g. [26]) can be used in low dimensions. In the more general case the exact update from $\pi_n(\mathrm{d}\boldsymbol{\beta} \mid \psi)$ is commonly replaced with a Metropolis update from $\pi_n(\mathrm{d}\boldsymbol{\beta} \mid \psi)$ (using e.g. a gradient-based kernel such as MALA or HMC). In the latter case, the Gibbs kernel P_n we analyse here is an idealized version of the practically used Metropolis-within-Gibbs kernel. Under suitable (mild) assumptions, we expect the convergence properties of this idealized scheme to provide a lower bound to

the Metropolis-within-Gibbs schemes used in practice. Also, we expect the convergence of the two kernels to be of the same order when the kernel used for the Metropolis updates on the full conditional mixes fast. Providing quantitative results in this direction is an interesting area for future work, which we are currently pursuing. This would extend the applicability of the proof techniques developed in this work to broad classes of non conditionally-conjugate models, such as Gaussian Processes with non-Gaussian likelihood discussed above. See Section 7 for more details.

6 Feasible start

All the previous results are stated in terms of mixing times from worst case M-warm start, as defined in (5). Since starting from $\mu \in \mathcal{N}(\pi_J, M)$ with small M (e.g. not increasing with J) may be in principle infeasible, it is of interest to provide an explicit example of a starting distribution that can be implemented in practice, a so-called feasible start, where the associated value of M can be controlled. In the setting of Theorem 4.2, the properties of the Gibbs samplers combined with the probabilistic structure of hierarchical models allow to translate the problem of feasible starts into the one of having a good initialisation for the hyper-parameters ψ, as we now show. Indeed, assume that the maximum marginal likelihood estimator $\hat{\psi}_J = \arg\max \prod_{j=1}^{J} g(Y_j \mid \psi)$, with g as in (16), is well-defined. Let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ be given by

$$\mu_J(B) = \int_B \mathrm{Unif}\left(\hat{\psi}_J, c/\sqrt{J}\right)(\mathrm{d}\psi) \prod_{j=1}^{J} p(\theta_j \mid Y_j, \psi)\, \mathrm{d}\boldsymbol{\theta} \qquad B \subset \mathbb{R}^{lJ+D} \tag{35}$$

where $c > 0$ is a fixed constant and $\mathrm{Unif}(\psi, r)$ denotes the uniform distribution over the closed ball of center ψ and radius $r > 0$. Therefore, the initial point is obtained by sampling from the uniform distribution around the maximum likelihood estimator for ψ and, conditional on this value, from the posterior distribution of the groups specific parameters. The next theorem shows that this choice leads to a good asymptotic behaviour of the mixing times.

Theorem 6.1. *Consider the same setting of Theorem 4.2 and let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ as in (35). Then, for every $\epsilon \in (0,1)$ there exists $T(\psi^*, \epsilon, c) < \infty$ such that*

$$\liminf_{J\to\infty} Q_{\psi^*}^{(J)}\left(t_{mix}^{(J)}(\epsilon, \mu_J) \leq T(\psi^*, \epsilon, c)\right) \to 1 \qquad \text{as } J \to \infty\,.$$

The difference with Theorem 4.2 is in the specification of the starting distribution, that is now made explicit. Note that whether or not μ_J is a feasible start in practice depends on whether the maximum likelihood estimate $\hat{\psi}_J$ can be computed, using e.g. an Expectation-Maximization algorithm, up to a $\mathcal{O}(1/\sqrt{J})$ error.

Remark. By its definition in (3), the Gibbs sampler does not depend on the starting point of the first block. Therefore Theorem 6.1 extends to any $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ such that

$$\mu_J\left(\mathbb{R}^{lJ} \times A\right) = \mathrm{Unif}\left(\hat{\psi}_J, c/\sqrt{J}\right)(A) \qquad A \subset \mathbb{R}^D\,.$$

7 Future works

A first natural extension in this context would be the case where no fixed dimensional sufficient statistic is available, i.e. $p(\cdot \mid \psi)$ in (1) does not belong to the exponential family. Since the above dimensionality reduction does not apply there, a possibility is to study the marginal chain induced on ψ; indeed the latter has the same properties of the Gibbs

sampler on $(\boldsymbol{\theta}, \psi)$, see e.g. [52]. Also, in this work we have focused on the case with well-specified likelihoods but, as discussed after Theorem 3.1, we expect the misspecified setting to behave in qualitatively similar ways.

Secondly, when dealing with Gibbs samplers, it is often the case that some of the conditional updates cannot be performed exactly. A natural solution is to employ more general coordinate-wise schemes, where exact sampling is replaced by Markov updates with stationary measure given by the conditional distribution. For example in hierarchical models for categorical data (see Section 5.2), while in principle exact conditional sampling is feasible, the parameters θ_j are often sampled in a Metropolis-within-Gibbs fashion, for reasons of computational efficiency and easiness of implementation. While algorithmically convenient, the modification makes theoretical analysis significantly more involved: in particular Proposition 2.2 ceases to hold and the dimensionality reduction given by Lemma 4.1 is not available without exact sampling. In ongoing work we are considering a different strategy, by providing lower bounds on the approximate conductance [36]: our preliminary results suggest that, provided the conditional Markov updates have good spectral properties, general coordinate-wise schemes can enjoy the same dimension-free convergence of the Gibbs sampler. Another interesting direction would be to derive results analogous to the ones in Section 2 for other MCMC kernels (e.g. gradient-based ones) under appropriate regularity assumptions on the sequence of target distribution, potentially exploiting tools from the recent work in [11].

Finally, we expect (at least parts of) our methodology to be applicable much beyond hierarchical models as in (1). For example, when fitting (finite or infinite) Bayesian mixture models, it is customary to use a Gibbs sampler over a properly augmented space by introducing latent allocation variables (see e.g. [16]): this leads to a problem of increasing dimensionality, since the number of latent variables grows linearly with n. An asymptotic analysis, as performed in this paper, seems accessible: indeed, posterior concentration results are available [40] and a dimensionality reduction similar to Lemma 4.1 can be exploited. However there are still significant challenges to perform a rigorous analysis in this setting: for example posterior contraction is often proved using Wasserstein distance, that is in general too weak for our purposes. We leave the discussion of such issues to a future work.

Funding. GZ acknowledges support from the European Research Council (ERC), through StG "PrSc-HDBayLe" grant ID 101076564.

References

[1] Amit, Y. (1991). On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions. *J. Multivar. Anal. 38*, 82–99.

[2] Andrieu, C., A. Lee, S. Power, and A. Q. Wang (2022). Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. *arXiv preprint arXiv:2211.08959*.

[3] Atchadé, Y. F. (2021). Approximate Spectral Gaps for Markov Chain Mixing Times in High Dimensions. *SIAM. J. MATH. DATA SCI. 3*, 854–872.

[4] Bally, V. and L. Caramellino (2015). Asymptotic development for the CLT in total variation distance. *Bernoulli 22*, 2442–2485.

[5] Bass, M. R. and S. K. Sahu (2016). A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC. *Stat. Comput. 27*, 1491–1512.

[6] Belloni, A. and V. Chernozhukov (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist. 37*, 2011–2055.

[7] Beskos, A., N. Pillai, G. Roberts, J. Sanz-Serna, and A. Stuart (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli 19*, 1501–1534.

[8] Bhattacharya, R. N. and R. R. Rao (2010). *Normal Approximations and Asymptotic Expansions*. Society for Industrial and Applied Mathematics.

[9] Bobkov, S. G., G. P. Chistyakov, and F. Götze (2014). Berry-Essen bounds in the entropic central limit theorem. *Probab. Theory Relat. Fields 159*, 435–478.

[10] Brooks, S., A. Gelman, G. L. Jones, and X. Meng (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall.

[11] Caprio, R. and A. Johansen (2023). A calculus for Markov chain Monte Carlo: studying approximations in algorithms. *arXiv preprint arXiv:2310.03853*.

[12] Casella, G. and E. I. george (1992). Explaining the Gibbs Sampler. *Am. Stat. 46*, 167–174.

[13] Chlebicka, I., K. Latuszynski, and B. Miasojedow (2023). Solidarity of Gibbs Samplers: the spectral gap. *arXiv preprint arXiv:2304.02109*.

[14] Dalalyan, A. S. (2017). Theoretical Guarantees for Approximate Sampling from Smooth and Log-Concave Densities. *J. R. Stat. Soc. Ser. B. 79*, 651–676.

[15] Diaconis, P., K. Khare, and L. Saloff-Coste (2008). Gibbs Sampling, Exponential Families and Orthogonal Polynomials. *Stat. Sci. 23*, 151–178.

[16] Diebolt, J. and C. P. Robert (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *J. R. Stat. Soc. Ser. B. 56*, 363–375.

[17] Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika 106*, 765–779.

[18] Durmus, A. and E. Moulines (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab. 27*, 1551–1587.

[19] Dwivedi, R., Y. Chen, M. J. Wainwright, and B. Yu (2019). Log–concave sampling: Metropolis–Hastings algorithms are fast! *J. Mach. Learn. Res. 20*, 1–42.

[20] Feller, W. (1970). *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons.

[21] Flegal, J. M., J. Hughes, D. Vats, K. Gupta, and U. Maji (2021). mcmcse: Monte Carlo Standard Errors for MCMC. *R package*.

[22] Gelfand, A. E., H. J. Kim, C. Sirmans, and S. Banerjee (2003). Spatial Modelling With Spatially Varying Coefficient Processes. *J. Am. Stat. Assoc. 98*, 387–396.

[23] Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient Parametrisations for Normal Linear Mixed Models. *Biometrika 82*, 479–488.

[24] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. CRC press.

[25] Gelman, A. and J. L. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[26] Gilks, W. R. and P. Wild (1992). Adaptive Rejection Sampling for Gibbs Sampling. *J. R. Stat. Soc. Ser. C 41*, 337–348.

[27] Gong, L. and J. M. Flegal (2015). A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo. *J. Comput. Graph. Stat. 25*, 684–700.

[28] Green, P. J., K. Latuszynski, M. Pereyra, and C. P. Robert (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Stat. Comput. 25*, 835–862.

[29] Hobert, J. P. (2011). The data augmentation algorithm: Theory and methodology. *Handbook of Markov chain Monte Carlo*, 253–293.

[30] Jin, Z. and J. P. Hobert (2022). Dimension free convergence rates for Gibbs samplers for Bayesian linear mixed models. *Stoch. Process. Their Appl. 148*, 25–67.

[31] Johndrow, J. E., A. Smith, N. Pillai, and D. B. Dunson (2019). MCMC for Imbalanced Categorical Data. *J. Am. Stat. Assoc. 114*, 1394–1403.

[32] Kamatani, K. (2014). Local consistency of Markov chain Monte Carlo methods. *Ann. Inst. Stat. Math. 66*, 63–74.

[33] Khare, K. and H. Zhou (2009). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. Appl. Probab. 2*, 737–777.

[34] Kleijn, B. J. K. and A. W. van der Vaart (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat. 6*, 353–381.

[35] Liu, J. S. (1994). Fraction of Missing Information and Convergence Rate for Data Augmentation. In *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface.*

[36] Lovász, L. and M. Simonovits (1993). Random Walks in a Convex Body and an Improved Volume Algorithm. *Random Struct. and Alg. 4*, 359–412.

[37] Martin, G. M., D. T. Frazier, and C. P. Robert (2023). Computing Bayes: From Then 'Til Now. *Stat. Sci. In press.*

[38] Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 113–162.

[39] Negrea, J., J. Yang, H. Feng, D. M. Roy, and J. H. Huggins (2022). Statistical Inference with Stochastic Gradient Algorithms. *arXiv preprint arXiv:2207.12395*.

[40] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist. 41*, 370–400.

[41] Nickl, R. and S. Wang (2022). On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *J. Eur. Math. Soc.*.

[42] Papaspiliopoulos, O., G. Roberts, and G. Zanella (2020). Scalable inference for crossed random effects models. *Biometrika 107*, 25–40.

[43] Papaspiliopoulos, O., G. O. Roberts, and M. Sköld (2003). Non-Centered Parameterizations for Hierarchical Models and Data Augmentation (with discussion). In *Bayesian Statistics (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.)*, pp. 307–326.

[44] Papaspiliopoulos, O., G. O. Roberts, and M. Sköld (2007). A General Framework for the Parametrization of Hierarchical Models. *Stat. Sci.*, 59–73.

[45] Papaspiliopoulos, O., G. O. R. Roberts, and M. Sköld (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, 59–73.

[46] Papaspiliopoulos, O., T. Stumpf-Fétizon, and G. Zanella (2023). Scalable computation for Bayesian hierarchical models. *arXiv preprint arXiv:2103.10875*.

[47] Petrov, V. V. (1956). A local theorem for densities of sums of independent random variables. *Theory Probab. Appl. 84*, 316–322.

[48] Qin, Q. and J. P. Hobert (2019). Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *Ann. Statist. 47*, 2320–2347.

[49] Qin, Q. and J. P. Hobert (2022). Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *Ann, Appl. Prob. 32*, 124–166.

[50] Rajaratnam, B. and D. Sparks (2015). MCMC-Based Inference in the Era of Big Data: A Fundamental Analysis of the Convergence Complexity of High-Dimensional Chains. *arXiv preprint arXiv:1508.00947*.

[51] Roberts, G. O. and J. S. Rosenthal (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B 60*, 255–268.

[52] Roberts, G. O. and J. S. Rosenthal (2001). Markov Chains and De-Initializing Processes. *Scand. J. Stat. 28*, 489–504.

[53] Roberts, G. O. and J. S. Rosenthal (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv. 60*, 255–268.

[54] Roberts, G. O. and S. H. Sahu (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stoch. Process. Their Appl. 49*, 207–216.

[55] Roberts, G. O. and S. H. Sahu (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *J. R. Stat. Soc. Ser. B 59*, 291–317.

[56] Roberts, G. O. and S. H. Sahu (2001). Approximate Predetermined Convergence Properties of the Gibbs Sampler. *J. Comput. Graph. Statist. 10*, 216–229.

[57] Roberts, G. O. and R. L. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli 2*, 341–363.

[58] Rosenthal, J. S. (1995). Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *J. Am. Stat. Assoc 90*, 558–566.

[59] Rosenthal, J. S. and P. Rosenthal (2015). Spectral bounds for certain two-factor non-reversible MCMC algorithms. *Electron. Commun. Probab.. 20*, 1–10.

[60] Ross, N. (2011). Fundamentals of Stein's method. *Probab. Surv. 8*, 210–293.

[61] Smith, W. L. (1953). A frequency-function form of the central limit theorem. *Math. Proc. Camb. Philos. Soc. 49*, 462–472.

[62] Tang, R. and Y. Yang (2022). Computational Complexity of Metropolis-Adjusted Langevin Algorithms for Bayesian Posterior Sampling. *arXiv preprint arXiv:2206.06491*.

[63] Thompson, M. A Comparison of Methods for Computing Autocorrelation Time. *Technical Report No. 1007, Department of Statistics, University of Toronto*.

[64] Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

[65] Williams, C. K. and C. E. Rasmussen (2006). *Gaussian Processes for Machine Learning*. Cambridge MA: MIT press.

[66] Wu, K., S. Schmidler, and Y. Chen (2022). Minimax Mixing Time of the Metropolis-Adjusted Langevin Algorithm for Log-Concave Sampling. *J. Mach. Learn. Res. 23*, 1–63.

[67] Yang, J. and J. S. Rosenthal (2022). Complexity results for MCMC derived from quantitative bounds. *Ann. Appl. Prob. 33*, 1459–1500.

[68] Yang, J., M. J. Wainwright, and M. I. Jordan (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist. 44*, 2497–2532.

[69] Zhou, Q., J. Yang, D. Vats, G. O. Roberts, and J. S. Rosenthal (2022). Dimension-free mixing for high-dimensional Bayesian variable selection. *J. R. Stat. Soc. Ser. B 84*, 1751–1784.

Appendix A Simple counter-examples for Section 2

A.1 Convergence of the stationary distribution does not imply pointwise convergence of Gibbs operators

Let $\mathcal{X} = [0,1]^2$ and define $A_n = \left[\frac{r_n}{l_n}, \frac{r_n+1}{l_n}\right]$, where

$$r_n = n - 2^{k_n}, \quad l_n = 2^{k_n}, \quad k_n = \lfloor \log_2 n \rfloor,$$

with $\lfloor a \rfloor$ denoting the integer part of a and $n \geq 2$. Therefore $\{A_n\}_n$ is a collection of intervals with decreasing length, such that $x \in A_n$ infinitely often, for every $x \in [0,1]$. We define a sequence $\{\pi_n\}_n \subset \mathcal{P}(\mathcal{X})$ as

$$\pi_n(\mathrm{d}x_1 \mid x_2) = \begin{cases} \mathbb{1}_{[0,1]}(x_1)\,\mathrm{d}x_1, & x_2 \notin A_n \\ \delta_0(\mathrm{d}x_1), & x_2 \in A_n \end{cases}, \quad \pi_n(\mathrm{d}x_2) = \mathbb{1}_{[0,1]}(x_2)\,\mathrm{d}x_2,$$

where $\mathbb{1}_A(x)\,\mathrm{d}x$ denotes the uniform measure on A. Define now

$$\pi(\mathrm{d}x_1, \mathrm{d}x_2) = \mathbb{1}_{[0,1]}(x_1)\mathbb{1}_{[0,1]}(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

and denote $C = \{0\} \times A_n$. For every $B \subset \mathcal{X}$ we have

$$\begin{aligned} |\pi_n(B) - \pi(B)| &\leq |\pi_n(B \cap C) - \pi(B \cap C)| + |\pi_n(B \cap C^c) - \pi(B \cap C^c)| \\ &= \pi_n(B \cap C) \leq \pi_n(C). \end{aligned}$$

Therefore we conclude

$$\|\pi_n - \pi\|_{TV} \leq \pi_n(C) \to 0,$$

as $n \to \infty$. However, if P_n and P are the operators of the associated Gibbs samplers, for every $\mathbf{x} \in \mathcal{X}$ it holds

$$\|P_n(\mathbf{x}, \cdot) - P(\mathbf{x}, \cdot)\|_{TV} \geq |P_n(\mathbf{x}, C) - P(\mathbf{x}, C)|,$$

so that, since $x_2 \in A_n$ infinitely often, we get

$$\|P_n(\mathbf{x}, \cdot) - P(\mathbf{x}, \cdot)\|_{TV} = 1$$

infinitely often. Incidentally, it is not difficult to show that $\mathrm{Gap}(P_n) = 0$ for every n, while $\mathrm{Gap}(P) = 1$. Example 1.4 shows that this mismatch may hold under significantly less pathological scenarios.

A.2 Equality of the stationary distributions does not imply closeness of the transition operators

Let $\pi_1 = \pi_2 = \pi$, with π the standard Gaussian distribution. Moreover, let

$$P_1(x, \cdot) = \epsilon\pi(\cdot) + (1-\epsilon)\delta_x(\cdot) \quad \text{and} \quad P_2(x, \cdot) = \epsilon\pi(\cdot) + (1-\epsilon)\delta_{-x}(\cdot),$$

with $\epsilon \in [0,1)$. P_1 and P_2 are uniformly ergodic transition operators with invariant distribution π. Let μ be the truncation of π on the positive real numbers: it is easy to show that $\mu \in \mathcal{N}(\pi, 2)$. However

$$\|\mu P_1 - \mu P_2\|_{TV} \geq (1-\epsilon)\left[\mu((0,\infty)) - \mu((-\infty, 0])\right] = 1 - \epsilon.$$

Moreover, it holds that $\|\mu - \pi\|_{TV} = 1/2$, so that we conclude

$$2\|\mu - \pi\|_{TV} - \epsilon \leq \|\mu P_1 - \mu P_2\|_{TV} \leq 2\|\mu - \pi\|_{TV}.$$

A.3 Convergence of the stationary distribution in Wasserstein distance does not imply convergence of the mixing times for Gibbs sampler operators

Let $\mathcal{X} = \mathbb{R}^2$ and $\bar{\pi}_n(\mathrm{d}\mathbf{x}) = N(x_1 \mid 0, 1/n)N(x_2 \mid 0, 1/n)\mathrm{d}x_1\mathrm{d}x_2$. Define π_n to be the truncation of $\bar{\pi}_n$ on the set

$$A = \{(-\infty, 0] \times (-\infty, 0]\} \bigcup \{[0, +\infty) \times [0, +\infty)\}.$$

Let $f : \mathcal{X} \to \mathbb{R}$ be a Lipschitz function with constant 1. Then it holds

$$\int_{\mathcal{X}} [f(x_1, x_2) - f(0, 0)]\, \pi_n(\mathrm{d}\mathbf{x}) \leq \int_{\mathcal{X}} \sqrt{x_1^2 + x_2^2}\, \pi_n(\mathrm{d}\mathbf{x}) \to 0,$$

as $n \to \infty$, so that $\|\pi_n - \pi\|_W \to 0$, where $\pi(\mathrm{d}\mathbf{x}) = \delta_{(0,0)}(\mathbf{x})$ and $\|\cdot\|_W$ denotes the Wasserstein distance.

If P is the kernel of the Gibbs sampler targeting π, then it is immediate to show that

$$\sup_{\mu \in \mathcal{N}(\pi, M)} \|\mu P - \pi\|_W = 0$$

for every $M \geq 1$, so that the mixing times in Wasserstein distance are equal to 1 for every $\epsilon > 0$.

Instead, denote with μ_n the truncation of π_n on $A_1 = (-\infty, 0] \times (-\infty, 0]$. It is easy to show that $\mu_n \in \mathcal{N}(\pi_n, 2)$, but

$$\mu_n P_n^t(A_1) - \pi_n(A_1) = \frac{1}{2}$$

for every n and t, where P_n is the kernel of the Gibbs sampler targeting π_n. Since the Wasserstein distance is stronger than the weak one, there exists an absolute constant c such that $\|\mu_n P_n^t - \pi_n\|_W \geq c$ for every n and t. Therefore, with ϵ small enough and $M \geq 2$, the mixing times of P_n in Wasserstein distance are equal to infinity for every n.

A.4 Convergence of the stationary distribution does not imply convergence of the spectral gaps for Gibbs operators

Let $\mathcal{X} = \mathbb{R}^2$ and

$$\pi(\mathrm{d}\mathbf{x}) = N(x_1 \mid 0, 1)N(x_2 \mid 0, 1)\mathrm{d}x_1\mathrm{d}x_2,$$

where $N(x \mid \mu, \sigma^2)$ is the density function of a gaussian distribution with mean μ and variance σ^2. Define π_n to be the truncation of π on the set A_n, where

$$A_n = \{(-\infty, n] \times (-\infty, n]\} \bigcup \{[n, +\infty) \times [n, +\infty)\}.$$

If P_n and P are the operators of the associated Gibbs samplers, it is not difficult to show that

$$\|\pi_n - \pi\|_{TV} \to 0 \quad \text{and} \quad \|P_n(\mathbf{x}, \cdot) - P(\mathbf{x}, \cdot)\|_{TV} \to 0$$

as $n \to \infty$, for every $\mathbf{x} \in \mathcal{X}$. However, if $B_n = (-\infty, n] \times (-\infty, n]$ we have

$$\pi_n(B_n) > 0 \quad \text{and} \quad \int_{B_n} P_n\left(\mathbf{x}, B_n^c\right) \pi_n(\mathrm{d}\mathbf{x}) = 0,$$

so that $\mathrm{Gap}(P_n) = 0$ for every n, while $\mathrm{Gap}(P) = 1$.

Appendix B Regularity assumptions (B4)-(B6) for Theorem 4.2

Let

$$M_s^{(p)}(\psi \mid y) = E\left[T_s^p(\theta_j) \mid Y_j = y, \psi\right], \tag{36}$$

$$M_{s,s'}^{(p)}(\psi \mid y) = E\left[T_s^p(\theta_j) T_{s'}^p(\theta_j) \mid Y_j = y, \psi\right], \tag{37}$$

be the posterior moments of $\boldsymbol{T}$ given ψ, denote $M^{(p)}(\psi \mid y) = \left(M_1^{(p)}(\psi \mid y), \ldots, M_S^{(p)}(\psi \mid y)\right) \in \mathbb{R}^S$ and

$$[C(\psi)]_{s,d} = E_{Y_j}\left[\partial_{\psi_d} M_s^{(1)}(\psi \mid Y_j)\right], \quad [V(\psi)]_{s,s'} = E_{Y_j}\left[\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi\right)\right], \tag{38}$$

with $s, s' = 1, \ldots S$ and $d = 1, \ldots, D$. Moreover we write B_δ for the ball of center ψ^* and radius δ, and denote expectations with respect to the law of Y_j as defined in $(B1)$ by $E_{Y_j}[\cdot]$.

$(B4)$ The expectation $M_s^{(p)}(\psi \mid y)$ is well defined for every y and $p = 1, \ldots, 6$. Moreover, there exist $\delta_4 > 0$ and C finite constant such that for every $\psi \in B_{\delta_4}$ it holds $E_{Y_j}\left[\left|\partial_{\psi_d} M_s^{(6)}(\psi \mid Y_j)\right|\right] < C$, $E_{Y_j}\left[\left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}(\psi \mid Y_j)\right|\right] < C$, $E_{Y_j}\left[\left|\partial_{\psi_d} M_{s,s'}^{(1)}(\psi \mid Y_j)\right|\right] < C$ and $E_{Y_j}\left[\left|\partial_{\psi_d}\left\{M_s^{(1)}(\psi \mid Y_j) M_{s'}^{(1)}(\psi \mid Y_j)\right\}\right|\right] < C$ for $s, s' = 1, \ldots, S$ and $d, d' = 1, \ldots, D$. Finally, the matrix $V(\psi^*)$ defined in (38) is non singular.

Assumption $(B4)$ can be understood as a smoothness condition. The posterior distribution of $\boldsymbol{T}$ should not change considerably, if we move from ψ^* to a sufficiently close ψ: this is measured in terms of the derivative of the posterior moments, that must be finite in average. Thanks to $(B4)$ we can prove a suitable conditional Central Limit Theorem to show convergence of a rescaled version of $\boldsymbol{T}$, conditional to ψ and $Y_{1:J}$.

We define the posterior characteristic function of $T(\theta_j) = (T_1(\theta_j), \ldots, T_S(\theta_j))$ and $\sum_{j=1}^k T(\theta_j)$, given ψ, as $\varphi(t \mid Y_j, \psi) = E\left[e^{it^\top T(\theta_j)} \mid Y_j, \psi\right]$ for $t \in \mathbb{R}^S$. and $\varphi^{(k)}(t \mid Y_{1:k}, \psi) = \prod_{j=1}^k \varphi(t \mid Y_j, \psi)$, respectively. We will assume:

$(B5)$ There exist $k \geq 1$ and $\delta_5 > 0$ such that

$$\sup_{\psi \in B_{\delta_5}} \int_{\mathbb{R}^S} \left|\varphi^{(k)}(t \mid Y_{1:k}, \psi)\right|^2 \mathrm{d}t < \infty,$$

for almost every $Y_1, \ldots, Y_k \overset{\text{iid}}{\sim} Q_{\psi^*}$.

$(B6)$ There exist $k' \geq 1$ and $\delta_6 > 0$ such that

$$\sup_{\psi \in B_{\delta_6}} \sup_{|t| > \epsilon} \left|\varphi^{(k')}(t \mid Y_{1:k'}, \psi)\right| < \phi(\epsilon),$$

for almost every $Y_1, \ldots, Y_k \overset{\text{iid}}{\sim} Q_{\psi^*}$, with $\phi(\epsilon) < 1$ for every $\epsilon > 0$.

Assumptions $(B5)$ and $(B6)$ allow the convergence of $\boldsymbol{T}$ to hold for the total variation distance, that is stronger than the weak one, proved through $(B4)$. Loosely speaking, integrability of the characteristic function and its strictly positive distance from 1 guarantee that the distribution is far from being discrete: the latter is exactly the case where weak convergence does not translate to stronger metrics. The problem of proving Central Limit theorems in total variation distance has received considerable attention over the decades: it can be tackled with Fourier-based techniques [47, 61], as we do here, but also with Stein's method (see [60] for a survey), Malliavin calculus (e.g. [4]) or through bounds based on entropy (e.g. [9]). Conditions $(B5)$ and $(B6)$ are somewhat reminiscent of the ones in Theorem 19.3 in [8].

Appendix C Proofs

C.1 Statement and proof of Lemma C.1

Lemma C.1. *Let $\mathcal{N} \subset \mathcal{P}(\mathcal{X})$ and $\pi \in \mathcal{P}(\mathcal{X})$. Then*

$$\sup_{\mu\in\mathcal{N}} \inf\left\{t \geq 1 \,:\, \left\|\mu P^t - \pi\right\|_{TV} < \epsilon\right\} = \inf\left\{t \geq 1 \,:\, \sup_{\mu\in\mathcal{N}} \left\|\mu P^t - \pi\right\|_{TV} < \epsilon\right\},$$

for every Markov transition kernel P.

Proof. Let

$$t^{(1)} = \sup_{\mu\in\mathcal{N}} \inf\left\{t \geq 1 \,:\, \left\|\mu P^t - \pi\right\|_{TV} < \epsilon\right\}, \quad t^{(2)} = \inf\left\{t \geq 1 \,:\, \sup_{\mu\in\mathcal{N}} \left\|\mu P^t - \pi\right\|_{TV} < \epsilon\right\}.$$

Assume $t^{(1)} < \infty$. Then $\left\|\mu P^{t^{(1)}} - \pi\right\|_{TV} < \epsilon$ for every $\mu \in \mathcal{N}$. This implies

$$\sup_{\mu\in\mathcal{N}} \left\|\mu P^{t^{(1)}} - \pi\right\|_{TV} < \epsilon,$$

i.e. $t^{(2)} \leq t^{(1)}$. With a similar reasoning, if $t^{(2)} < \infty$ we have $t^{(1)} \leq t^{(2)}$. Therefore $t^{(1)} = t^{(1)}$ if either $t^{(1)} < \infty$ or $t^{(2)} < \infty$.

Assume now $t^{(1)} = \infty$ and fix $t^* > 0$. By definition of $t^{(1)}$ there exists $\mu \in \mathcal{N}$ such that

$$\left\|\mu P^{t^*} - \pi\right\|_{TV} \geq \epsilon,$$

that implies

$$\sup_{\mu\in\mathcal{N}} \left\|\mu P^{t^*} - \pi\right\|_{TV} \geq \epsilon,$$

i.e. $t^{(2)} > t^*$. Since t^* is arbitrary, we have $t^{(2)} = \infty$. With a similar reasoning, if $t^{(2)} = \infty$ it holds $t^{(1)} = \infty$. □

C.2 Statement and proof of Lemma C.2

Lemma C.2. *Let $M \geq 1$, $\pi \in \mathcal{P}(\mathcal{X})$, $\mu \in \mathcal{N}(\pi, M)$ and P be a π-invariant Markov transition kernel. Then $\mu P^t \in \mathcal{N}(\pi, M)$, for every $t \in \mathbb{N}$.*

Proof. Let $A \subseteq \mathcal{X}$. Since $\mu \in \mathcal{N}(\pi, M)$ and P is π-invariant, we have $(\mu P)(A) \leq M(\pi P)(A) = M\pi(A)$. Thus $\mu P \in \mathcal{N}(\pi, M)$ and the result follows by induction on t. □

C.3 Proof of Lemma 2.1

Proof. Let $\hat{P}_n = P_n \circ \phi_n^{-1}$ be the push-forward operator of P_n under ϕ_n, defined as

$$\hat{P}_n(\mathbf{x}, B) = P_n\left(\phi_n^{-1}(\mathbf{x}), \phi_n^{-1}(B)\right) \tag{39}$$

for every $\mathbf{x} \in \phi_n(\mathcal{X})$ and $B \subseteq \mathcal{X}$. Since ϕ_n is an injective transformation, $\hat{P}_n$ is a well-defined Markov transition kernel (see e.g. Lemma 1 in [42]). Moreover, since ϕ_n is coordinate-wise as in (7) we have $\hat{P}_n = \hat{P}_{n,1} \dots \hat{P}_{n,K}$, where

$$\begin{aligned}\hat{P}_{n,i}\left(\mathbf{x}, S_{\mathbf{x},i,A}\right) = P_{n,i}\left(\phi_n^{-1}(\mathbf{x}), S_{\phi_n^{-1}(\mathbf{x}),i,\phi_{n,i}^{-1}(A)}\right) &= \int_{\phi_{n,i}^{-1}(A)} \pi_n\left(\mathrm{d}y_i \mid \phi_n^{-1}(\mathbf{x})^{(-i)}\right) \\ &= \int_A \tilde{\pi}_n\left(\mathrm{d}y_i \mid \mathbf{x}^{(-i)}\right), \quad A \subset \mathcal{X}_i,\end{aligned}$$

so that $\hat{P}_n$ is exactly the operator of the Gibbs sampler targeting $\tilde{\pi}_n$, i.e. $\tilde{P}_n = \hat{P}_n$.

Therefore, since ϕ_n is an injective transformation, by Corollary 2 in [52] we have

$$\left\|\mu_n P_n^t - \pi_n\right\|_{TV} = \left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV},$$

with $\tilde{\mu}_n = \mu_n \circ \phi_n^{-1}$. To conclude the proof, we show that $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ if and only if $\mu_n \in \mathcal{N}(\pi_n, M)$. Indeed, to prove the implication from right to left, by definition of push-forward measure we have

$$\tilde{\mu}_n(A) = \mu_n\left(\phi_n^{-1}(A)\right) = \int_{\phi_n^{-1}(A)} \frac{\mathrm{d}\mu_n}{\mathrm{d}\pi_n}(\mathbf{x})\, \pi_n(\mathrm{d}\mathbf{x}) \leq M\pi_n\left(\phi_n^{-1}(A)\right) = M\tilde{\pi}_n(A),$$

for every set $A \subset \mathcal{X}$. Equivalently we obtain the other implication. □

C.4 Proof of Proposition 2.2

For any $\pi \in \mathcal{P}(\mathcal{X})$ and Q Markov transition kernel with state space $\mathcal{X}$, we define $(\pi \otimes Q) \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ as

$$(\pi \otimes Q)(B) = \int_B Q(\mathbf{x}, \mathrm{d}\mathbf{y})\pi(\mathrm{d}\mathbf{x})$$

for every $B \subseteq \mathcal{X} \times \mathcal{X}$.

Lemma C.3. *Let $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$ and Q be a Markov transition kernel with state space $\mathcal{X}$. Then*

$$\|\pi_1 \otimes Q - \pi_2 \otimes Q\|_{TV} = \|\pi_1 - \pi_2\|_{TV}.$$

Proof. By definition of total variation distance we have

$$\begin{aligned}
&\|\pi_1 \otimes Q - \pi_2 \otimes Q\|_{TV} \\
&= \sup_{f\,:\,\mathcal{X}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y}) Q(\mathbf{x}, \mathrm{d}\mathbf{y})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y}) Q(\mathbf{x}, \mathrm{d}\mathbf{y})\pi_2(\mathrm{d}\mathbf{x}) \right| \\
&= \sup_{f\,:\,\mathcal{X}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(\mathbf{x},\mathbf{y}) Q(\mathbf{x}, \mathrm{d}\mathbf{y}) \right) \pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(\mathbf{x},\mathbf{y}) Q(\mathbf{x}, \mathrm{d}\mathbf{y}) \right) \pi_2(\mathrm{d}\mathbf{x}) \right| \\
&\leq \sup_{g\,:\,\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} g(\mathbf{x})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} g(\mathbf{x})\pi_2(\mathrm{d}\mathbf{x}) \right| = \|\pi_1 - \pi_2\|_{TV}.
\end{aligned}$$

Also, taking $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x})$ for every $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}$ we have

$$\begin{aligned}
\|\pi_1 - \pi_2\|_{TV} &= \sup_{g\,:\,\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} g(\mathbf{x})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} g(\mathbf{x})\pi_2(\mathrm{d}\mathbf{x}) \right| \\
&\leq \sup_{f\,:\,\mathcal{X}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y}) Q(\mathbf{x}, \mathrm{d}\mathbf{y})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y}) Q(\mathbf{x}, \mathrm{d}\mathbf{y})\pi_2(\mathrm{d}\mathbf{x}) \right| \\
&= \|\pi_1 \otimes Q - \pi_2 \otimes Q\|_{TV}.
\end{aligned}$$

□

For $j = 1, 2$, denote the kernel of the Gibbs sampler targeting π_j as $P_j = P_{j,1} \ldots P_{j,K}$, where

$$P_{j,i}(\mathbf{x}, S_{\mathbf{x},i,A}) = \int_A \pi_j\left(\mathrm{d}y_i \mid \mathbf{x}^{(-i)}\right), \quad A \subset \mathcal{X}_i,$$

with $S_{\mathbf{x},i,A} = \{\mathbf{y} \in \mathcal{X} : y_j = x_j \,\forall\, j \neq i \text{ and } y_i \in A\}$ as in the main. By definition, $P_i(\mathbf{x}, \mathrm{d}\mathbf{y})$ depends only on $\mathbf{x}^{(-i)}$. Thus we can define $\left(\pi^{(-i)} \otimes Q\right) \in \mathcal{P}\left(\mathcal{X}^{(-i)} \times \mathcal{X}\right)$ as

$$\left(\pi^{(-i)} \otimes P_i\right)(B) = \int_B P_i\left(\mathbf{x}^{(-i)}, \mathrm{d}\mathbf{y}\right) \pi\left(\mathrm{d}\mathbf{x}^{(-i)}\right),$$

for every $B \subset \mathcal{X}^{(-i)} \times \mathcal{X}$ and similarly for

$$\left(\pi^{(-1)} \otimes P\right) \in \mathcal{P}\left(\mathcal{X}^{(-1)} \times \mathcal{X}\right) \quad \text{and} \quad \left(\pi^{(-i)} \otimes \prod_{j \geq i} P_j\right) \in \mathcal{P}\left(\mathcal{X}^{(-i)} \times \mathcal{X}\right),$$

with $i = 1 \ldots, K$. Given this notation we have the following Lemmas.

Lemma C.4. *We have*

$$\|\mu P_1 - \mu P_2\|_{TV} \leq M \left\|\pi_2^{(-1)} \otimes P_1 - \pi_2^{(-1)} \otimes P_2\right\|_{TV}$$

for every $\mu \in \mathcal{N}(\pi_2, M)$ and $M \geq 1$.

Proof. By definition of total variation distance

$$\|\mu P_1 - \mu P_2\|_{TV} = \sup_{f\,:\,\mathcal{X} \to [0,1]} \left| \int_{\mathcal{X}} f(\mathbf{y}) \mu P_1(\mathrm{d}\mathbf{y}) - \int_{\mathcal{X}} f(\mathbf{y}) \mu P_2(\mathrm{d}\mathbf{y}) \right|.$$

Then, by definition of $\mathcal{N}(\pi_2, M)$, it holds

$$\begin{aligned}
&\|\mu P_1 - \mu P_2\|_{TV} \\
&= M \sup_{f\,:\,\mathcal{X} \to [0,1]} \left| \int_{\mathcal{X}^K} \frac{f(\mathbf{y})}{M} \int_{\mathcal{X}^{(-1)}} \frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi_2^{(-1)}}(\mathbf{x}^{(-1)}) P_1(\mathbf{x}^{(-1)}, \mathrm{d}\mathbf{y}) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right. \\
&\qquad \left. - \int_{\mathcal{X}} \frac{f(\mathbf{y})}{M} \int_{\mathcal{X}^{(-1)}} \frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi_2^{(-1)}}(\mathbf{x}^{(-1)}) P_2(\mathbf{x}^{(-1)}, \mathrm{d}\mathbf{y}) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right| \\
&\leq M \sup_{g\,:\,\mathcal{X}^{(-1)} \times \mathcal{X} \to [0,1]} \left| \int_{\mathcal{X}^{(-1)} \times \mathcal{X}} g(\mathbf{x}^{(-1)}, \mathbf{y}) P_1\left(\mathbf{x}^{(-1)}, \mathrm{d}\mathbf{y}\right) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right. \\
&\qquad \left. - \int_{\mathcal{X}^{(-1)} \times \mathcal{X}} g(\mathbf{x}^{(-1)}, \mathbf{y}) P_2\left(\mathbf{x}^{-1}, \mathrm{d}\mathbf{y}\right) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right| \\
&= M \left\|\pi_2^{(-1)} \times P_1 - \pi_2^{(-1)} \otimes P_2\right\|_{TV}.
\end{aligned}$$

□

Lemma C.5. *We have*

$$\begin{aligned}
\left\|\pi_1^{(-i)} \otimes \prod_{j \geq i} P_{1,j} - \pi_2^{(-i)} \otimes \prod_{j \geq i} P_{2,j}\right\|_{TV} &\leq 2 \|\pi_1 - \pi_2\|_{TV} \\
&+ \left\|\pi_1^{(-(i+1))} \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j \geq i+1} P_{2,j}\right\|_{TV}
\end{aligned} \tag{40}$$

for every $i = 1, \ldots, K-1$ and

$$\left\|\pi_1^{(-K)} \otimes P_{1,K} - \pi_2^{(-K)} \otimes P_{2,K}\right\|_{TV} = \|\pi_1 - \pi_2\|_{TV}.$$

Proof. We start by proving (40). Notice that, by definition of $P_{1,i}$ and $P_{2,i}$, we have

$$\begin{aligned}
&\int_{\mathcal{X}^{(-i)} \times \mathcal{X}} g\left(\mathbf{x}^{(-i)}, \mathbf{y}\right) \prod_{j \geq i} P_{1,j}\left(\mathbf{x}^{(-i)}, \mathrm{d}\mathbf{y}\right) \pi_1^{(-i)}\left(\mathrm{d}\mathbf{x}^{(-i)}\right) \\
&\qquad = \int_{\mathcal{X} \times \mathcal{X}^{(-i)}} h\left(\mathbf{x}, \mathbf{y}^{(-i)}\right) \prod_{j \geq i+1} P_{1,j}\left(\mathbf{x}^{(-i-1)}, \mathrm{d}\mathbf{y}\right) \pi_1(\mathrm{d}\mathbf{x})
\end{aligned}$$

and

$$\int_{\mathcal{X}^{(-i)}\times\mathcal{X}} g\left(\mathbf{x}^{(-i)},\mathbf{y}\right)\prod_{j\geq i}P_{2,j}\left(\mathbf{x}^{(-i)},\mathrm{d}\mathbf{y}\right)\pi_2^{(-i)}\left(\mathrm{d}\mathbf{x}^{(-i)}\right)$$
$$=\int_{\mathcal{X}\times\mathcal{X}^{(-i)}} h\left(\mathbf{x},\mathbf{y}^{(-i)}\right)\prod_{j\geq i+1}P_{2,j}\left(\mathbf{x}^{(-i-1)},\mathrm{d}\mathbf{y}\right)\pi_2\left(\mathrm{d}\mathbf{x}\right),$$

where $g:\mathcal{X}^{(-i)}\times\mathcal{X}\to\mathbb{R}$ is any measurable function and h is the composition of g and the function $c:\mathcal{X}^{(-i)}\times\mathcal{X}\to\mathcal{X}\times\mathcal{X}^{(-i)}$ that relocates the $(K-1+i)$-th element of a vector after the $(i-1)$-th element. Since there is a one-to-one relationship between functions g and h, we have

$$\left\|\pi_1^{(-i)}\otimes\prod_{j\geq i}P_{1,j}-\pi_2^{(-i)}\otimes\prod_{j\geq i}P_{2,j}\right\|_{TV}=\left\|\pi_1\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}. \tag{41}$$

Then by triangular inequality and Lemma C.3 we have

$$\begin{aligned}\left\|\pi_1\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}&\leq\left\|\pi_1\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{1,j}\right\|_{TV}\\&\quad+\left\|\pi_2\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}\\&\leq\|\pi_1-\pi_2\|_{TV}+\left\|\pi_2\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}.\end{aligned} \tag{42}$$

Notice that $\prod_{j\geq i+1}P_{1,j}$ and $\prod_{j\geq i+1}P_{2,j}$ do not depend on x_{i+1} by construction, that implies

$$\left\|\pi_2\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}$$

$$=\sup_{h:\,\mathcal{X}\times\mathcal{X}\to[0,1]}\left|\int_{\mathcal{X}\times\mathcal{X}}h\left(\mathbf{x},\mathbf{y}\right)\prod_{j\geq i+1}P_{1,j}\left(\mathbf{x}^{(-(i+1))},\mathrm{d}\mathbf{y}\right)\pi_2\left(\mathrm{d}\mathbf{x}\right)\right.$$
$$\left.-\int_{\mathcal{X}\times\mathcal{X}}h\left(\mathbf{x},\mathbf{y}\right)\prod_{j\geq i+1}P_{2,j}\left(\mathbf{x}^{(-(i+1))},\mathrm{d}\mathbf{y}\right)\pi_2\left(\mathrm{d}\mathbf{x}\right)\right|,$$

so that we have

$$\left\|\pi_2\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}$$

$$=\sup_{h:\,\mathcal{X}\times\mathcal{X}\to[0,1]}\left|\int_{\mathcal{X}^{(-(i+1))}\times\mathcal{X}}\int_{\mathcal{X}_{i+1}}h\left(\mathbf{x},\mathbf{y}\right)\pi_2\left(\mathrm{d}x_{i+1}\mid x^{(-(i+1))}\right)\prod_{j\geq i+1}P_{1,j}\left(\mathbf{x}^{(-(i+1))},\mathrm{d}\mathbf{y}\right)\pi_2\left(\mathrm{d}\mathbf{x}^{(-(i+1))}\right)\right.$$
$$\left.-\int_{\mathcal{X}^{(-(i+1))}\times\mathcal{X}}\int_{\mathcal{X}_{i+1}}h\left(\mathbf{x},\mathbf{y}\right)\pi_2\left(\mathrm{d}x_{i+1}\mid x^{(-(i+1))}\right)\prod_{j\geq i+1}P_{2,j}\left(\mathbf{x}^{(-(i+1))},\mathrm{d}\mathbf{y}\right)\pi_2\left(\mathrm{d}\mathbf{x}^{(-(i+1))}\right)\right|$$

$$\leq\left\|\pi_2^{(-(i+1))}\otimes\prod_{j\geq i+1}P_{1,j}-\pi_2^{(-(i+1))}\otimes\prod_{j\geq i+1}P_{2,j}\right\|_{TV}.$$

Moreover, it is clear that

$$\left\| \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV} \leq \left\| \pi_2 \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV},$$

thus combining the two above inequalities we get

$$\left\| \pi_2 \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV} = \left\| \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV}. \tag{43}$$

Combining (41), (42) and (43) with the fact that

$$\begin{aligned}\left\| \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV} &\leq \|\pi_1 - \pi_2\|_{TV} \\ &+ \left\| \pi_1^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV}\end{aligned}$$

we finally obtain (40). When $i = K$ the result follows by noticing that

$$\pi_1^{(-K)} \otimes P_{1,K} = \pi_1 \quad \text{and} \quad \pi_2^{(-K)} \otimes P_{2,K} = \pi_2$$

by definition. □

Proof of Proposition 2.2. Without loss of generality, let $\mu \in \mathcal{N}(\pi_2, M)$. By Lemma C.4 and the triangle inequality we have

$$\begin{aligned}\|\mu P_1 - \mu P_2\|_{TV} &\leq M \left\| \pi_2^{(-1)} \otimes P_1 - \pi_2^{(-1)} \otimes P_2 \right\|_{TV} \\ &\leq M \|\pi_1 - \pi_2\|_{TV} + M \left\| \pi_1^{(-1)} \otimes P_1 - \pi_2^{(-1)} \otimes P_2 \right\|_{TV}\end{aligned}$$

and the result follows by applying K times Lemma C.5. □

C.5 Proof of Lemma 2.3

Proof. With an abuse of notation, let $\pi_1(x)$, $\pi_2(x)$ and $\mu_1(x)$ be densities of π_1, π_2 and μ_1 with respect to a common dominating measure, such as $\tau = \pi_1 + \pi_2$. Let $\bar{\mu}$ be the measure on $\mathcal{X}$ with density $\bar{\mu}(x) = \min\{\mu_1(x), M\pi_2(x)\}$ for $x \in \mathcal{X}$. By construction $\bar{\mu}$ is a sub-probability since

$$\bar{\mu}(\mathcal{X}) = \int_{\mathcal{X}} \bar{\mu}(x)\tau(\mathrm{d}x) \leq \int_{\mathcal{X}} \mu_1(x)\tau(\mathrm{d}x) = 1.$$

Therefore, we can define a probability distribution $\mu_2 \in \mathcal{P}(\mathcal{X})$ with density

$$\mu_2(x) = \bar{\mu}(x) + \alpha \max\{M\pi_2(x) - \mu_1(x), 0\}, \qquad x \in \mathcal{X}$$

where

$$\alpha = \frac{1 - \int \bar{\mu}(x)\tau(\mathrm{d}x)}{\int_{\mathcal{X}} \max\{M\pi_2(x) - \mu_1(x), 0\}\,\tau(\mathrm{d}x)} \in (0, 1).$$

Notice that $\mu_2(x) \leq M\pi_2(x)$ for every $x \in \mathcal{X}$ since

$$\mu_2(x) = \begin{cases} M\pi_2(x), & \text{if } \mu_1(x) > M\pi_2(x), \\ (1-\alpha)\mu_1(x) + \alpha M\pi_2(x), & \text{if } \mu_1(x) \leq M\pi_2(x). \end{cases}$$

Thus $\mu_2 \in \mathcal{N}(\pi_2, M)$. By definition of total variation distance and of $\tilde{\mu}$, we have

$$\begin{aligned}\|\mu_1 - \mu_2\|_{TV} = &\int_{\mathcal{X}} \max\{\mu_1(x) - \mu_2(x), 0\}\, \tau(\mathrm{d}x) = \int_{\mathcal{X}} \max\{\mu_1(x) - M\pi_2(x), 0\}\, \tau(\mathrm{d}x) \\ \leq & M \int_{\mathcal{X}} \max\{\pi_1(x) - \pi_2(x), 0\}\, \tau(\mathrm{d}x) = M \|\pi_1 - \pi_2\|_{TV}\,.\end{aligned}$$

□

C.6 Proof of Theorem 2.4

Proof. By Lemma 2.1 the statement is equivalent to

$$\lim_{n\to\infty} \sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} \tag{44}$$

in $Q^{(n)}$-probability, where $\tilde{P}_n$ is the kernel of the Gibbs sampler targeting $\tilde{\pi}$.

Consider $\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\|_{TV}$ with $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$. By Lemma 2.3, there exists $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ such that

$$\|\tilde{\mu}_n - \tilde{\mu}\|_{TV} \leq M \|\tilde{\pi}_n - \tilde{\pi}\|_{TV}\,. \tag{45}$$

By the triangular inequality we can decompose $\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\|_{TV}$ as follows

$$\left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} \leq \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\mu} \tilde{P}_n^t \right\|_{TV} + \left\| \tilde{\mu} \tilde{P}_n^t - \tilde{\mu} \tilde{P}^t \right\|_{TV} + \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} + \|\tilde{\pi}_n - \tilde{\pi}\|_{TV}\,. \tag{46}$$

Combining (45) with the monotonicity of the total variation distance with respect to the application of transition kernels, we obtain

$$\left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\mu} \tilde{P}_n^t \right\|_{TV} \leq \|\tilde{\mu}_n - \tilde{\mu}\|_{TV} \leq M \|\tilde{\pi}_n - \tilde{\pi}\|_{TV}\,. \tag{47}$$

For the second term in (46), we want to prove that if $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ we have

$$\left\| \tilde{\mu} \tilde{P}_n^t - \tilde{\mu} \tilde{P}^t \right\|_{TV} \leq 2MKt \|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \tag{48}$$

for every $t \geq 1$. Indeed, the case $t = 1$ holds by Proposition 2.2. Assume now (48) holds for $t-1$, with $t \geq 2$. Then by the triangular inequality we have

$$\begin{aligned}\left\| \tilde{\mu} \tilde{P}_n^t - \tilde{\mu} \tilde{P}^t \right\|_{TV} &\leq \left\| \tilde{\mu} \tilde{P}_n^{\,t} - \mu \tilde{P}^{t-1} \tilde{P}_n \right\|_{TV} + \left\| \mu \tilde{P}^t - \mu \tilde{P}^{t-1} \tilde{P}_n \right\|_{TV} \\ &\leq \left\| \tilde{\mu} \tilde{P}_n^{t-1} - \tilde{\mu} \tilde{P}^{t-1} \right\|_{TV} + \left\| \mu \tilde{P}^{t-1} \tilde{P} - \mu \tilde{P}^{t-1} \tilde{P}_n \right\|_{TV}.\end{aligned}$$

By induction hypothesis we have

$$\left\| \tilde{\mu} \tilde{P}_n^{t-1} - \tilde{\mu} \tilde{P}^{t-1} \right\|_{TV} \leq 2MK(t-1) \|\tilde{\pi}_n - \tilde{\pi}\|_{TV}\,. \tag{49}$$

Moreover, by Lemma C.2 we have that $\tilde{\mu} \tilde{P}^{t-1} \in \mathcal{N}(\tilde{\pi}, M)$, so that from the case $t = 1$ we obtain

$$\left\| \mu \tilde{P}^{t-1} \tilde{P} - \mu \tilde{P}^{t-1} \tilde{P}_n \right\|_{TV} \leq 2MK \|\tilde{\pi}_n - \tilde{\pi}\|_{TV}\,. \tag{50}$$

Then (48) follows by (49) and (50). Combining (46), (47) and (48), for every $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ there exists $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ such that

$$\left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} \leq (2MKt + M + 1) \|\tilde{\pi}_n - \tilde{\pi}\|_{TV} + \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi}_n \right\|_{TV}.$$

Thus

$$\sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} \leq (2MKt + M + 1) \left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV} + \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV}.$$

It follows that, for any $\epsilon > 0$, we have

$$\begin{aligned} & Q^{(n)} \left(\sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} - \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} \geq \epsilon \right) \\ & \leq Q^{(n)} \left(\left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV} \geq (2MKt + M + 1)^{-1} \epsilon \right) \to 0, \end{aligned} \tag{51}$$

as $n \to \infty$ by $(A1)$ and $(2MKt + M + 1)^{-1}\epsilon > 0$.

We now prove the reverse inequality of (51) to establish (44). Given $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$, by Lemma 2.3, there exists $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ such that $\left\| \tilde{\mu} - \tilde{\mu}_n \right\|_{TV} \leq M \left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV}$. Then we proceed analogously to above, first decomposing $\left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV}$ as

$$\left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} \leq \left\| \tilde{\mu} \tilde{P}^t - \tilde{\mu} \tilde{P}_n^t \right\|_{TV} + \left\| \tilde{\mu} \tilde{P}_n^t - \tilde{\mu}_n \tilde{P}_n^t \right\|_{TV} + \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} + \left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV} \tag{52}$$

and then applying Proposition 2.2 using an argument analogous to above to get

$$\left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} \leq \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} + (2MKt + M + 1) \left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV}.$$

It follows

$$\sup_{\mu_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} \geq \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} - (2MKt + M + 1) \left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV}.$$

Fixing $\epsilon > 0$ arbitrary constant we have

$$\begin{aligned} & Q^{(n)} \left(\sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\| \tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n \right\|_{TV} - \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} \leq -\epsilon \right) \\ & \leq Q^{(n)} \left(\left\| \tilde{\pi}_n - \tilde{\pi} \right\|_{TV} \geq \frac{\epsilon}{2MKt + M + 1} \right) \to 0, \end{aligned} \tag{53}$$

as $n \to \infty$ by $(A1)$ and $(2MKt + M + 1)^{-1}\epsilon > 0$. The result follows by combining (51) and (53). □

C.7 Proof of Corollary 2.5

Proof. Thanks to Lemma C.1 we can write

$$t_{mix}^{(n)}(\epsilon, M) = \inf \left\{ t \geq 1 \, : \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^t - \pi_n \right\|_{TV} < \epsilon \right\}$$

and

$$\tilde{t}_{mix}(\epsilon, M) = \inf \left\{ t \geq 1 \, : \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} < \epsilon \right\}.$$

Assume $(A1)$ and denote $t^* = \tilde{t}_{mix}(\epsilon, M) < \infty$ for brevity. By definition of t^* we have $\delta = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^{t^*} - \tilde{\pi} \right\|_{TV} < \epsilon$. Thus

$$\begin{aligned} Q^{(n)} \left(t_{mix}^{(n)}(\epsilon, M) \leq t^* \right) &= Q^{(n)} \left(\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^{t^*} - \pi_n \right\|_{TV} < \epsilon \right) \\ &= Q^{(n)} \left(\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^{t^*} - \pi_n \right\|_{TV} - \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^{t^*} - \tilde{\pi} \right\|_{TV} < \epsilon - \delta \right) \\ &\to 1, \end{aligned}$$

as $n \to \infty$ by Theorem 2.4.

As regards the second part of the statement, let $(A1)$ hold and fix $T > 0$. Denote $\delta = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^T - \tilde{\pi} \right\|_{TV}$ and notice that by assumption $\delta \geq \epsilon > \underline{\epsilon}$. Thus

$$\begin{aligned}
\liminf_{n \to \infty} Q^{(n)} \left(t_{mix}^{(n)}(\underline{\epsilon}, M) < T \right) &= \liminf_{n \to \infty} Q^{(n)} \left(\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^T - \pi_n \right\|_{TV} < \underline{\epsilon} \right) \\
&= \liminf_{n \to \infty} Q^{(n)} \left(\delta - \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^T - \pi_n \right\|_{TV} \geq \delta - \underline{\epsilon} \right) \\
&\to 0,
\end{aligned}$$

as $n \to \infty$ by Theorem 2.4. □

C.8 Proof of Corollary 2.6

We need a preliminary well known lemma, whose proof we include for self-containedness.

Lemma C.6. *Let P be a Gibbs sampler kernel with $K = 2$ and target $\pi \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$. Then*

$$\left\| \mu P^t - \pi \right\|_{TV} \leq \frac{M}{2} \left(1 - Gap(P) \right)^t ,$$

for every $\mu \in \mathcal{N}(\pi, M)$ and $t \geq 1$.

Proof. Let $\mu \in \mathcal{N}(\pi, M)$ and $t \geq 1$. By Corollary 1 in [52] we have

$$\left\| \mu P^t - \pi \right\|_{TV} = \left\| \mu^{(-1)} \hat{P}^t - \pi^{(-1)} \right\|_{TV} , \tag{54}$$

where $\hat{P}$ is the Markov transition kernel on $\mathcal{X}_2$ defined as

$$\hat{P}(x_2, \mathrm{d}y_2) = \int_{\mathcal{X}_1} \pi(\mathrm{d}y_2 \mid y_1) \pi(\mathrm{d}y_1 \mid x_2) \qquad\qquad x_2 \in \mathcal{X}_2 .$$

Note that $\hat{P}$ is $\pi^{(-1)}$-reversible. Also, for every $f \in L^2(\pi^{(-1)})$, i.e. $f : \mathcal{X}_2 \to \mathbb{R}$ such that $\|f\|_2^2 = \pi^{(-1)}(f^2)$ is finite, we have

$$\begin{aligned}
&\int_{\mathcal{X}_2^2} f(x_2) f(y_2) \hat{P}(x_2, \mathrm{d}y_2) \pi(\mathrm{d}x_2) \\
&\quad = \int_{\mathcal{X}_2^2} f(x_2) f(y_2) \int_{\mathcal{X}_1} \pi(\mathrm{d}y_2 \mid y_1) \pi(\mathrm{d}y_1 \mid x_2) \pi(\mathrm{d}x_2) \\
&\quad = \int_{\mathcal{X}_1} \left[\int_{\mathcal{X}_2} f(y_2) \pi(\mathrm{d}y_2 \mid y_1) \right] \left[\int_{\mathcal{X}_2} f(x_2) \pi(\mathrm{d}x_2 \mid y_1) \right] \pi(\mathrm{d}y_1) \\
&\quad = \int_{\mathcal{X}_1} \left[\int_{\mathcal{X}_2} f(y_2) \pi(\mathrm{d}y_2 \mid y_1) \right]^2 \pi(\mathrm{d}y_1) \geq 0,
\end{aligned}$$

so that $\hat{P}$ is also positive semi-definite. Since $\hat{P}$ is reversible and positive semi-definite, we have (see e.g. equation (5) in [2]) that

$$\left\| \hat{P}^t(f) \right\|_2 \leq ||f||_2 \left(1 - \mathrm{Gap}(\hat{P}) \right)^t , \tag{55}$$

for every f such that $\pi(f) = 0$. Choosing $f = \frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi^{(-1)}} - 1$ and using the reversibility of $\hat{P}$ (see e.g. Section 2.1 in [33]) we also have

$$\left\| \mu^{(-1)} \hat{P}^t - \pi^{(-1)} \right\|_{TV} \leq \frac{1}{2} \left\| \mu^{(-1)} \hat{P}^t(f) \right\|_2 , \tag{56}$$

where $\mu^{(-1)}\hat{P}^t(f) = \int f(x_2)\mu^{(-1)}\hat{P}^t(\mathrm{d}x_2)$. With the same choice of f, we have

$$||f||_2^2 = \int \left(\frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi^{(-1)}}(x_2) - 1\right)^2 \pi^{-1}(\mathrm{d}x_2) \leq M^2$$

since $\mu^{(-1)} \in \mathcal{N}(\pi^{(-1)}, M)$. Thus, combining (55) with (56) we obtain

$$\left\|\mu P^t - \pi\right\|_{TV} \leq \frac{M}{2}\left(1 - \mathrm{Gap}(\hat{P})\right)^t.$$

Finally, for every $f : \mathcal{X}_2 \to \mathbb{R}$ with $||f||_2 < \infty$ it holds

$$\frac{\int_{\mathcal{X}_2^2}\left[f(y_2) - f(x_2)\right]^2 \pi(\mathrm{d}x_2)\hat{P}(x_2, \mathrm{d}y_2)}{2\mathrm{Var}_\pi^{(-1)}(f)} = \frac{\int_{\mathcal{X}^2}\left[g(\mathbf{y}) - g(\mathbf{x})\right]^2 \pi(\mathrm{d}\mathbf{x})P(\mathbf{x}, \mathrm{d}\mathbf{y})}{2\mathrm{Var}_\pi(f)},$$

where $g(\mathbf{x}) = f(x_2)$. Therefore $\mathrm{Gap}(\hat{P}) \geq \mathrm{Gap}(P)$ and we get

$$\left\|\mu P^t - \pi\right\|_{TV} \leq \frac{M}{2}\left(1 - \mathrm{Gap}(P)\right)^t,$$

as desired. □

Proof of Corollary 2.6. By Lemma C.6 we obtain

$$\tilde{t}_{mix}(\epsilon, M) \leq 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \mathrm{Gap}(\tilde{P})\right)},$$

and the result follows by the first part of Corollary 2.5. □

C.9 Proof of Proposition 3.2

Proof. By Theorem 3.1, assumption $(A1)$ is satisfied with

$$\phi_n(\psi) = \sqrt{n}(\psi - \psi^*) - \mathcal{I}^{-1}(\psi^*)\Delta_{n,\psi^*},$$

and $\tilde{\pi} = N\left(\mathbf{0}, \mathcal{I}^{-1}(\psi^*)\right)$. Since $\tilde{\pi}$ is the distribution of a multivariate normal with non singular covariance matrix, then it is easy to show $\tilde{t}_{mix}(\epsilon, M) < \infty$ for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$, see e.g. Theorem 2 in [1]. □

C.10 Statement and proof of Corollary C.7

We illustrate the result of Proposition 3.2 on a simple example of model (11) with normal likelihood and unknown mean and precision, that is

$$f(y \mid \mu, \tau) = N\left(y \mid \mu, \tau^{-1}\right), \tag{57}$$

where $K = 2$ and $\psi = (\mu, \tau)$. Notice that, even if a conjugate prior exists, it is common to place independent priors on μ and τ, for which the Gibbs sampler defined in (3) becomes a reasonable option.

Corollary C.7. *Consider model* (11) *with likelihood as in* (57). *Let* $Y_i \overset{iid}{\sim} Q_{\psi^*}$, *with* Q_{ψ^*} *admitting density* $f(y \mid \psi^*)$ *and* $\psi^* = (\mu^*, \tau^*) \in \mathbb{R} \times \mathbb{R}_+$. *Moreover let* p_0 *be absolutely continuous in a neighborhood of* ψ^* *with a continuous positive density at* ψ^*. *Consider the Gibbs sampler defined in* (3). *Then, for every* $M \geq 1$ *and* $\epsilon > 0$ *we have*

$$Q_{\psi^*}^{(n)}\left(t_{mix}^{(n)}(\epsilon, M) \leq 1\right) \to 1,$$

as $n \to \infty$.

For the proof we need a preliminary Lemma, whose proof we include for self-containedness and because it will be useful to refer to later on.

Lemma C.8. *Consider the same setting of Corollary C.7. Then conditions* (12) *are satisfied.*

Proof of Lemma C.8. Define

$$\Psi = \Psi_1 \times \Psi_2 = [\mu^* - 1, \mu^* + 1] \times \left[\frac{\tau^*}{2}, 2\tau^*\right]$$

compact neighborhood of ψ^* and

$$u_n(Y_1, \dots, Y_n) = 1 - \mathbb{1}_{g_1(Y_{1:n}) \leq c_1} \, \mathbb{1}_{g_2(Y_{1:n}) \leq c_2},$$

where $c_1 = 1/2$, $c_2 = (2\tau^*)^{-1}$ and

$$g_1(Y_{1:n}) = \left|\bar{Y} - \mu^*\right|, \quad \text{and} \quad g_2(Y_{1:n}) = \left|\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 - \frac{1}{\tau^*}\right|,$$

with $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Since $Y_i \overset{\text{iid}}{\sim} N(\mu, \tau^{-1})$, then $g_1(Y_{1:n})$ and $g_2(Y_{1:n})$ are equal in distribution, respectively, to

$$h_1(Z_{1:n}, \mu, \tau) = \left|\frac{1}{\sqrt{\tau}}\bar{Z} + \mu - \mu^*\right|, \quad h_2(Z_{1:n}, \mu, \tau) = \left|\frac{1}{\tau}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 - \frac{1}{\tau^*}\right|,$$

where $Z_i \overset{\text{iid}}{\sim} N(0,1)$. By the Law of Large numbers we have

$$\bar{Z} \to 0, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 \to 1$$

almost surely as $n \to \infty$. This implies

$$\begin{aligned}\int u_n(y_1, \dots, y_n) \prod_{i=1}^{n} f(\mathrm{d}y_i \mid \psi^*) \leq & P\left(h_1(Z_{1:n}, \mu^*, \tau^*) > c_1\right) \\ & + P\left(h_2(Z_{1:n}, \mu^*, \tau^*) > c_2\right) \to 0,\end{aligned}$$

as $n \to \infty$. Also, we have

$$\begin{aligned}\sup_{\psi \notin \Psi} \int \left[1 - u_n(y_1, \dots, y_n)\right] \prod_{i=1}^{n} f(\mathrm{d}y_i \mid \psi) \leq & \sup_{\tau \notin \Psi_2} P\left(h_2(Z_{1:n}, \mu, \tau) \leq c_2\right) \\ & + \sup_{\mu \notin \Psi_1,\, \tau \in \Psi_2} P\left(h_1(Z_{1:n}, \mu, \tau) \leq c_1\right).\end{aligned}$$

Now notice that by the reverse triangle inequality we have

$$\begin{aligned}\sup_{\tau \notin \Psi_2} P\left(h_2(Z_{1:n}, \mu, \tau) \leq c_2\right) &= \sup_{\tau \notin \Psi_2} P\left(\left|\frac{1}{\tau}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 - \frac{1}{\tau^*}\right| \leq c_2\right) \\ &\leq \sup_{\tau \notin \Psi_2} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 - 1\right| \geq \left|1 - \frac{\tau}{\tau^*}\right| - c_2\tau\right) \to 0,\end{aligned}$$

by definition of Ψ_2, as $n \to \infty$. Finally, again by reverse triangle inequality, we have

$$\sup_{\mu \notin \Psi_1,\, \tau \in \Psi_2} P\left(h_1(Z_{1:n}, \mu, \tau) \leq c_1\right) \leq \sup_{\mu \notin \Psi_1,\, \tau \in \Psi_2} P\left(|\bar{Z}| \geq \sqrt{\tau}\left(|\mu - \mu^*| - c_1\right)\right) \to 0,$$

as $n \to \infty$. □

Proof of Corollary C.7. In this case $\psi = (\mu, \tau)$ and

$$f(y \mid \psi) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(y-\mu)^2}.$$

By Lemma C.8 conditions (12) are satisfied. Also, the map $\psi \to f(y \mid \psi)$ is one-to-one, the map $\psi \to \sqrt{f(y \mid \psi)}$ is continuously differentiable, and the Fisher information matrix is

$$\mathcal{I}(\psi) = \begin{bmatrix} \frac{\tau}{2} & 0 \\ 0 & \frac{1}{2\tau} \end{bmatrix},$$

which is non singular and continuous as a function of ψ. Thus the conditions of Theorem 3.1 and Proposition 3.2 are satisfied. Finally, since we are considering a two-blocks Gibbs sampler, by Corollary 2.6 we have

$$T(\psi^*, \epsilon, M) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \mathrm{Gap}(\tilde{P})\right)},$$

where $\tilde{P}$ is the Gibbs sampler targeting a bivariate normal distribution with covariance matrix given by $\mathcal{I}^{-1}(\psi^*)$. Since the latter is diagonal, the Gibbs sampler coincides with independent sampling, so that $\mathrm{Gap}(\tilde{P}) = 1$. □

C.11 Proof of Lemma 4.1

Proof. Denote by $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t \geq 1}$ the Markov chain with kernel P_J defined in (15). The Markovianity of the induced sequence $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t \geq 1}$ follows by the one of $\left(\psi^{(t)}\right)_{t \geq 1}$, which is well known [15, 52]. We now show that $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t \geq 1}$ admits $\hat{P}_J$ as kernel. The conditional distribution of $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)$ given $\left(\boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right)$ is given by

$$\begin{aligned}\mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right) &= \mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right) \mathcal{L}\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}, \boldsymbol{T}^{(t-1)}\right) \\ &= \hat{\pi}_J\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \psi^{(t-1)}\right) \mathcal{L}\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right),\end{aligned}$$

where the last equality follows by (15) and the definition of $\hat{\pi}_J$. By the exponential family assumption in (14), $\boldsymbol{T}$ is a set of sufficient statistics for ψ, so that

$$\pi_J(\mathrm{d}\psi \mid \boldsymbol{\theta}) = \mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}) = \mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}) = \hat{\pi}_J(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta})). \tag{58}$$

Combining (15) and (58) we have

$$\begin{aligned}\mathcal{L}\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right) &= \int \pi_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{\theta}\right) \pi_J\left(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right) \\ &= \int \hat{\pi}_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}(\boldsymbol{\theta})\right) \pi_J\left(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right) = \hat{\pi}_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}\right)\end{aligned} \tag{59}$$

since $\boldsymbol{T}(\boldsymbol{\theta}) = \boldsymbol{T}^{(t)}$ almost surely under $\pi_J\left(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right)$. Thus we can conclude

$$\begin{aligned}\mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right) &= \hat{\pi}_J\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \psi^{(t-1)}\right) \hat{\pi}_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}\right) \\ &= \hat{P}_J\left(\left(\boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right), \left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\psi^{(t)}\right)\right),\end{aligned}$$

as desired. From the above one can easily deduce that $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t \geq 1}$ and $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t \geq 1}$ are *co-deinitializing* as in [52] and thus, by Corollary 2 therein, for every $\mu \in \mathcal{P}\left(\mathbb{R}^{\ell J} \times \mathbb{R}^D\right)$ we have

$$\left\|\mu P_J^t - \pi_J\right\|_{TV} = \left\|\nu \hat{P}_J^t - \hat{\pi}_J\right\|_{TV}, \tag{60}$$

where $\nu \in \mathcal{P}\left(\mathbb{R}^S \times \mathbb{R}^D\right)$ is the push forward of μ under $(\boldsymbol{\theta}, \psi) \mapsto (\boldsymbol{T}(\boldsymbol{\theta}), \psi)$. Moreover, by (5) we have that $\nu \in \mathcal{N}(\hat{\pi}_J, M)$ whenever $\mu \in \mathcal{N}(\pi_J, M)$. It follows that $\sup_{\mu \in \mathcal{N}(\pi_J, M)} t^{(J)}_{mix}(\epsilon, \mu) \leq \sup_{\nu \in \mathcal{N}(\hat{\pi}_J, M)} \hat{t}^{(J)}_{mix}(\epsilon, \nu)$. For the reverse inequality, fix $\nu \in \mathcal{N}(\hat{\pi}_J, M)$ and take $\mu(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi) = \int \pi_J(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}, \psi)\, \nu(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi)$. By (5) we have $\mu \in \mathcal{N}(\pi_J, M)$ and thus (60). It follows $\sup_{\nu \in \mathcal{N}(\hat{\pi}_J, M)} \hat{t}^{(J)}_{mix}(\epsilon, \nu) \leq \sup_{\mu \in \mathcal{N}(\pi_J, M)} t^{(J)}_{mix}(\epsilon, \mu)$ as desired. □

C.12 Proof of Lemma 4.3

Proof. The result follows immediately from Theorem 3.1, whose assumptions are given exactly by assumption $(B1) - (B3)$, with likelihood $g(y \mid \psi)$. □

C.13 Proof of Lemma 4.4

The proof is divided in two main steps: in Section C.13.1 the result is proved under the weak metric (Lemma C.11) and it is extended to the total variation distance in Section C.13.2.

First of all we need two technical lemmas, that we prove for completeness.

Lemma C.9. *Let S and p be two positive integers. Then there exists a constant $C = C(S, p)$ such that*

$$|\boldsymbol{x}|^p \leq 1 + C \sum_{s=1}^{S} x_s^{2p}$$

for every $\boldsymbol{x} \in \mathbb{R}^S$.

Proof. Since $(1 - |\mathbf{x}|^p)^2 \geq 0$, we have $|\mathbf{x}|^p \leq 1 + |\mathbf{x}|^{2p}$. Moreover, by the Multinomial Theorem, we get

$$|\mathbf{x}|^{2p} = \left(\sum_{s=1}^{S} x_s^2\right)^p = \sum_{\boldsymbol{k} \in \mathbb{P}} \binom{p}{k_1 \ \dots \ k_S} \prod_{s=1}^{S} x_s^{2k_s},$$

where $\mathbb{P} = \left\{\boldsymbol{k} = (k_1, \dots, k_S) \,:\, k_s \text{ positive integer}, \ \sum_{s=1}^{S} k_s = p\right\}$. Since

$$\prod_{s=1}^{S} x_s^{2k_s} \leq \left(\max_s |x_s|\right)^{2p} \leq \sum_{s=1}^{S} x_s^{2p},$$

the result follows by choosing $C = \sum_{\boldsymbol{k} \in \mathbb{P}} \binom{p}{k_1 \dots k_S}$. □

Lemma C.10. *Under assumption $(B3)$, the random variables $\Delta_J = (\Delta_{J,1}, \dots, \Delta_{J,D})$ defined in (17) are such that for every $\beta > 0$ we have*

$$\frac{1}{J^\beta} \Delta_{J,d} \quad \to \quad 0,$$

$Q^{(\infty)}_{\psi^*}$-almost surely as $J \to \infty$ for every $d = 1, \dots, D$.

Proof. Recall that

$$\Delta_{J,d} = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \left[\mathcal{I}^{-1}(\psi^*) \nabla \log g(Y_j \mid \psi^*)\right]_d =: \frac{1}{\sqrt{J}} \sum_{j=1}^{J} X_{j,d}$$

and $\mathcal{I}^{-1}(\psi^*) \partial_{\psi_d} \log g(Y_j \mid \psi^*)$ has zero mean and finite variance, by $(B3)$. Therefore, by Chebychev inequality

$$P\left(\left|\frac{1}{J^\beta} \Delta_{J,d}\right| > \epsilon\right) \leq \frac{\mathrm{Var}\,(X_{1,d})}{\epsilon^2 J^{1+2\beta}},$$

for every $\epsilon > 0$. This implies

$$\sum_{J=1}^{\infty} P\left(\left|\frac{1}{J^{\beta}}\Delta_{J,d}\right| > \epsilon\right) \leq \sum_{J=1}^{\infty} \frac{\mathrm{Var}\left(X_{1,d}\right)}{\epsilon^2 J^{1+2\beta}} < \infty,$$

and the result follows by Borel-Cantelli Lemma. □

C.13.1 Weak convergence

In order to ease the following exposition, denote

$$\psi^{(J)} := \psi^* + \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}, \quad J \geq 1\,. \tag{61}$$

The next lemma proves convergence of $\tilde{\boldsymbol{T}}$ using the weak metric, denoted by $\|\cdot\|_W$.

Lemma C.11. *Define $\tilde{\psi}$ and $\tilde{\boldsymbol{T}}$ as in* (17) *and* (19)*, respectively. Under assumptions $(B1)-(B4)$, for every $\tilde{\psi} \in \mathbb{R}^D$ it holds*

$$\left\|\mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) - N\left(C(\psi^*)\tilde{\psi}, V(\psi^*)\right)\right\|_W \to 0, \tag{62}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$.

Proof. For ease of notation, denote

$$\mu = C(\psi^*)\tilde{\psi} \quad \text{and} \quad \Xi := V(\psi^*).$$

By definition of $M_s^{(p)}$, we have

$$E\left[T_s^p(\theta_j) \mid Y_j, \psi^{(J)}\right] = M_s^{(p)}\left(\psi^{(J)} \mid Y_j\right).$$

Conditional on $\tilde{\psi}$, the group specific statistics $T_s(\theta_j)$ are independent across $j = 1, \ldots, J$. Thus, by Lyapunov version of Central Limit Theorem, in order to obtain (62) it suffices to show

$$\frac{1}{\sqrt{J}}\sum_{j=1}^{J}\left[M^{(1)}\left(\psi^{(J)} \mid Y_j\right) - M^{(1)}\left(\psi^* \mid Y_j\right)\right] - C(\psi^*)\Delta_J \quad \to \quad \mu \tag{63}$$

$$\frac{1}{J}\sum_{j=1}^{J}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^{(J)}\right) \quad \to \quad \Xi_{s,s'} \tag{64}$$

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J} E_{Y_j}\left[\left|T(\theta_j) - M^{(1)}\left(\psi^* \mid Y_j\right)\right|^3 \mid Y_j, \psi^{(J)}\right] \quad \to \quad 0, \tag{65}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$, with $s, s' = 1, \ldots, S$. We prove the three above results sequentially below, which concludes the proof of (62). □

Proof of (63). For any $s = 1, \ldots, S$, by (61) and the multivariate Taylor formula it holds

$$M_s^{(1)}\left(\psi^{(J)} \mid Y_j\right) - M_s^{(1)}\left(\psi^* \mid Y_j\right) = \sum_{d=1}^{D}\frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}}\partial_{\psi_d} M_s^{(1)}\left(\psi^* \mid Y_j\right) + R_2(Y_j),$$

where

$$R_2(Y_j) = \sum_{d,d'=1}^{D}\frac{(\tilde{\psi}_d + \Delta_{J,d})(\tilde{\psi}_{d'} + \Delta_{J,d'})}{J}\int_0^1 (1-t)\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \mid Y_j\right)\mathrm{d}t.$$

Therefore

$$\begin{aligned}\frac{1}{\sqrt{J}}\sum_{j=1}^{J}\left[M_s^{(1)}\left(\psi^{(J)}\mid Y_j\right)-M_s^{(1)}\left(\psi^*\mid Y_j\right)\right]=\\ =\sum_{d=1}^{D}(\tilde{\psi}_d+\Delta_{J,d})\frac{1}{J}\sum_{j=1}^{J}\partial_{\psi_d}M_s^{(1)}\left(\psi^*\mid Y_j\right)+\frac{1}{\sqrt{J}}\sum_{j=1}^{J}R_2(Y_j)\,,\end{aligned}\tag{66}$$

where

$$\begin{aligned}&\frac{1}{\sqrt{J}}\sum_{j=1}^{J}R_2(Y_j)=\\ &\sum_{d,d'=1}^{D}\frac{(\tilde{\psi}_d+\Delta_{J,d})(\tilde{\psi}_{d'}+\Delta_{J,d'})}{J^{1/4}}\frac{1}{J^{5/4}}\sum_{j=1}^{J}\int_0^1(1-t)\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\mid Y_j\right)\mathrm{d}t.\end{aligned}\tag{67}$$

As regards (67), for every $d,d'=1,\dots,D$ by Lemma C.10 it holds

$$\frac{(\tilde{\psi}_d+\Delta_{J,d})(\tilde{\psi}_{d'}+\Delta_{J,d'})}{J^{1/4}}=\frac{\tilde{\psi}_d\tilde{\psi}_{d'}}{J^{1/4}}+\tilde{\psi}_d\frac{\Delta_{J,d'}}{J^{1/4}}+\tilde{\psi}_{d'}\frac{\Delta_{J,d}}{J^{1/4}}+\frac{\Delta_{J,d}}{J^{1/8}}\frac{\Delta_{J,d'}}{J^{1/8}}\quad\to\quad 0,\tag{68}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$. Moreover, with the change of variables $x=t/J^{1/4}$ we have

$$\begin{aligned}&\left|\frac{1}{J^{5/4}}\sum_{j=1}^{J}\int_0^1(1-t)\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\mid Y_j\right)\mathrm{d}t\right|\\ &\le\int_0^{J^{1/4}}\frac{1}{J}\sum_{j=1}^{J}\left|\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+x\frac{\tilde{\psi}+\Delta_J}{J^{1/4}}\mid Y_j\right)\right|\mathrm{d}x\\ &\le\int_{-J^{1/4}}^{J^{1/4}}\frac{1}{J}\sum_{j=1}^{J}\left|\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+x\mid Y_j\right)\right|\mathrm{d}x,\end{aligned}$$

where the last inequality follows from $\left|\frac{\tilde{\psi}+\Delta_J}{J^{1/4}}\right|\le 1$ for J high enough, thanks to Lemma C.10. Moreover, $\frac{1}{J^{1/4}}<\delta_4$ for J high enough, so that

$$\begin{aligned}&\left|\frac{1}{J^{5/4}}\sum_{j=1}^{J}\int_0^1(1-t)\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\mid Y_j\right)\mathrm{d}t\right|\\ &\le\int_{\delta_4}^{\delta_4}\frac{1}{J}\sum_{j=1}^{J}\left|\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+x\mid Y_j\right)\right|\mathrm{d}x\\ &=\frac{1}{J}\sum_{j=1}^{J}\int_{\delta_4}^{\delta_4}\left|\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+x\mid Y_j\right)\right|\mathrm{d}x.\end{aligned}$$

By the Law of Large Numbers and $(B4)$ it holds

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}\int_{\delta_4}^{\delta_4}\left|\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+x\mid Y_j\right)\right|\mathrm{d}x\\ &\qquad\to\int_{-\delta_4}^{\delta_4}E\left[\left|\partial_{\psi_d}\partial_{\psi_{d'}}M_s^{(1)}\left(\psi^*+x\mid Y_j\right)\right|\right]\mathrm{d}x<2C\delta_4.\end{aligned}\tag{69}$$

By combining (68) and (69), we can conclude

$$\left| \frac{1}{\sqrt{J}} \sum_{j=1}^{J} R_2(Y_j) \right| \quad \to \quad 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. As regards (66), by the Law of Large Numbers we have

$$\frac{1}{J} \sum_{j=1}^{J} \partial_{\psi_d} M_s^{(1)} \left(\psi^* \mid Y_j \right) \quad \to \quad E \left[\partial_{\psi_d} M_s^{(1)} \left(\psi^* \mid Y_j \right) \right] = C_{s,d}(\psi^*),$$

that is finite thanks to $(B4)$. Therefore, we can conclude that for any $s = 1, \ldots, S$ we have

$$M_s^{(1)} \left(\psi^{(J)} \mid Y_j \right) - M_s^{(1)} \left(\psi^* \mid Y_j \right) - \sum_{d=1}^{D} C_{s,d}(\psi^*) \Delta_{J,d} \quad \to \quad \sum_{d=1}^{D} C_{s,d}(\psi^*) \tilde{\psi}_d,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$ and thus (63) holds. □

Proof of (64). For every $s, s' = 1, \ldots, S$ by multivariate Taylor formula it holds

$$\mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^{(J)} \right) = \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* \right) + R_{1,cov}(Y_j),$$

where

$$R_{1,cov}(Y_j) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}} \int_0^1 (1-t) \partial_{\psi_d} \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + t \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \right) \mathrm{d}t.$$

Notice that

$$\frac{1}{J} \sum_{j=1}^{J} R_{1,cov}(Y_j) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{J^{1/4}} \int_0^1 (1-t) \frac{1}{J^{5/4}} \sum_{j=1}^{J} \partial_{\psi_d} \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + t \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \right) \mathrm{d}t.$$

With the same arguments of before we have $\frac{\tilde{\psi}_d + \Delta_{J,d}}{J^{1/4}} \to 0$ and

$$\begin{aligned}
&\left| \int_0^1 (1-t) \frac{1}{J^{5/4}} \sum_{j=1}^{J} \partial_{\psi_d} \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + t \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \right) \mathrm{d}t \right| \\
&\qquad \leq \frac{1}{J} \sum_{j=1}^{J} \int_{-\delta_4}^{\delta_4} \left| \partial_{\psi_d} \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + x \right) \right| \mathrm{d}x \\
&\qquad \to \quad \int_{-\delta_4}^{\delta_4} E \left[\left| \partial_{\psi_d} \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + x \right) \right| \right] \mathrm{d}x
\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Notice that by $(B4)$ we have

$$\begin{aligned}
&E \left[\left| \partial_{\psi_d} \mathrm{Cov} \left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + x \right) \right| \right] \\
&\quad \leq E \left[\left| \partial_{\psi_d} M_{s,s'}^{(1)} \left(\psi^* + x \mid Y_j \right) \right| \right] + E \left[\left| \partial_{\psi_d} \left\{ M_s^{(1)} \left(\psi^* + x \mid Y_j \right) M_{s'}^{(1)} \left(\psi^* + x \mid Y_j \right) \right\} \right| \right] \\
&\quad \leq 2C,
\end{aligned}$$

for every $x \in (-\delta_4, \delta_4)$. Therefore, we can conclude

$$\left| \frac{1}{J} \sum_{j=1}^{J} R_{1,cov}(Y_j) \right| \to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Thus, by the Law of Large Numbers we have

$$\frac{1}{J}\sum_{j=1}^{J} \mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^*\right) \quad \to \quad E\left[\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^*\right)\right],$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. □

Proof of (65). By Lemma C.9 we have

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J} E_{Y_j}\left[\left|T(\theta_j) - M^{(1)}\left(\psi^* \mid Y_j\right)\right|^3 \mid Y_j, \psi^{(J)}\right]$$
$$\leq \frac{1}{\sqrt{J}} + C\frac{1}{J^{3/2}}\sum_{s=1}^{S}\sum_{j=1}^{J} M^{(6)}\left(\psi^{(J)} \mid Y_j\right) + C\frac{1}{J^{3/2}}\sum_{s=1}^{S}\sum_{j=1}^{J}\left[M^{(1)}\left(\psi^* \mid Y_j\right)\right]^6.$$

By Jensen inequality $\left[M^{(1)}\left(\psi^* \mid Y_j\right)\right]^6 \leq M^{(6)}\left(\psi^* \mid Y_j\right)$ and by the Law of Large Numbers

$$\frac{1}{J}\sum_{s=1}^{S}\sum_{j=1}^{J} M^{(6)}\left(\psi^* \mid Y_j\right) \quad \to \quad \sum_{s=1}^{S} E\left[T_s^6(\theta_j) \mid \psi^*\right] < \infty$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Thus to prove (65) it suffices to show

$$\frac{1}{J^{3/2}}\sum_{s=1}^{S}\sum_{j=1}^{J} M^{(6)}\left(\psi^{(J)} \mid Y_j\right) \quad \to \quad 0$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. For every $s = 1, \ldots, S$ by multivariate Taylor formula it holds

$$M_s^{(6)}\left(\psi^{(J)} \mid Y_j\right) = M_s^{(6)}\left(\psi^* \mid Y_j\right) + R_{1,6}(Y_j),$$

where

$$R_{1,6}(Y_j) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}} \int_0^1 (1-t)\partial_{\psi_d} M_s^{(6)}\left(\psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \mid Y_j\right) \mathrm{d}t.$$

Notice that

$$\frac{1}{J}\sum_{j=1}^{J} R_{1,6}(Y_j) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{J^{1/4}} \int_0^1 (1-t)\frac{1}{J^{5/4}}\sum_{j=1}^{J} \partial_{\psi_d} M_s^{(6)}\left(\psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \mid Y_j\right) \mathrm{d}t,$$

and with the same arguments of before we have $\frac{\tilde{\psi}_d + \Delta_{J,d}}{J^{1/4}} \to 0$ $Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$ and

$$\left|\int_0^1 (1-t)\frac{1}{J^{5/4}}\sum_{j=1}^{J} \partial_{\psi_d} M_s^{(6)}\left(\psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \mid Y_j\right) \mathrm{d}t\right|$$
$$\leq \frac{1}{J}\sum_{j=1}^{J}\int_{-\delta_4}^{\delta_4}\left|\partial_{\psi_d} M_s^{(6)}(\psi^* + x \mid Y_j)\right| \mathrm{d}x$$
$$\to \int_{-\delta_4}^{\delta_4} E\left[\left|\partial_{\psi_d} M_s^{(6)}(\psi^* + x \mid Y_j)\right|\right] \mathrm{d}x < 2\delta_4 C,$$

by $(B4)$. Therefore, we can conclude

$$\left|\frac{1}{J}\sum_{j=1}^{J} R_{1,6}(Y_j)\right| \to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Moreover, by the Law of Large Numbers we have

$$\frac{1}{J}\sum_{j=1}^{J} M_s^{(6)}\left(\psi^* \mid Y_j\right) \quad \to \quad E\left[M_s^{(6)}\left(\psi^* \mid Y_j\right)\right] = E\left[T_s^6(\theta_j) \mid \psi^*\right],$$

by $(B1)$ and the definition of conditional expectation. Therefore

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J} M_s^{(6)}\left(\psi^* + \frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}} \mid Y_j\right) \to 0,$$

from which (65) follows. □

C.13.2 Total variation convergence

We extend the weak convergence to total variation using characteristic functions, in particular exploiting the conditions in Lemma C.15. Here we first state some other technical lemmas that will be required later on.

Lemma C.12. *Let X be a $\mathbb{R}^S$-valued random vector with zero mean and characteristic function $\varphi_X(u)$. Then for every $u \in \mathbb{R}^S$*

$$\varphi_X(u) = 1 - \frac{1}{2}E\left[(u^\top X)^2\right] + \frac{\theta}{6}E\left[|u^\top X|^3\right],$$

for some $\theta = \theta(u) \in \mathbb{C}$ such that $|\theta| \leq 1$.

Proof. Taylor formula for the complex exponential reads

$$e^{ix} = 1 + ix - \frac{x^2}{2} + \frac{x^3}{6}e^{iz},$$

where $z \in \mathbb{C}$ is such that $0 \leq |z| \leq |x|$. By $x = u^\top X$, we have

$$\varphi_X(u) = 1 + iE\left[u^\top X\right] - \frac{1}{2}E\left[\left(u^\top X\right)^2\right] + \frac{\theta}{6}E\left[\left|u^\top X\right|^3\right],$$

with $\theta = e^{iz}$, recalling that $|e^{iz}| \leq 1$ for any z. The result follows from $E\left[u^\top X\right] = 0$. □

Lemma C.13. *Let $X \in \mathbb{R}^S$ and $Y \in \mathbb{R}^S$ be independent random vectors with the same distribution. Then*

$$\varphi_{X-Y}(u) = \left|\varphi_X(u)\right|^2.$$

Proof. By independence we can write

$$\varphi_{X-Y}(u) = E\left[e^{iu^\top X}\right] E\left[e^{-iu^\top X}\right],$$

where

$$E\left[e^{iu^\top X}\right] = E\left[\cos u^\top X\right] + iE\left[\sin u^\top X\right] = a + ib,$$

for suitable a and b. Since $\cos x$ is even and $\sin x$ is odd, we can write

$$|\varphi_{X-Y}(u)| = |(a+ib)(a-ib)| = a^2 + b^2 = |\varphi_X(u)|^2$$

Since $X - Y$ has a symmetric density by construction $|\varphi_{X-Y}(u)| = \varphi_{X-Y}(u)$ and the result follows. □

Corollary C.14. *Let X be a $\mathbb{R}^S$-valued random vector with characteristic function $\varphi_X(u)$. Then*

$$|\varphi_X(u)|^2 \leq e^{-u^\top Var(X)u + \frac{2|u|^3}{3}\left[1 + C\sum_{s=1}^S E\left[X_i^6\right]\right]},$$

for $u \in \mathbb{R}^S$, where C is a finite constant independent of u.

Proof. Let Y be an independent copy of X. By Lemma C.13, it holds

$$|\varphi_X(u)|^2 = \varphi_{X-Y}(u),$$

where $\varphi_{X-Y}(u)$ is a real function, since it is the characteristic function of a random variable with symmetric density. Therefore, by Lemma C.12 it holds

$$\varphi_{X-Y}(u) = 1 - \frac{1}{2}E\left[(u^\top Z)^2\right] + \frac{\theta}{6}E\left[|u^\top Z|^3\right],$$

where $Z = X - Y$ and $\theta = \theta(u) \in \mathbb{R}$. Recalling that $e^x \geq 1 + x$ for every x, we have

$$\varphi_{X-Y}(u) \leq e^{-\frac{1}{2}E\left[(u^\top Z)^2\right] + \frac{\theta}{6}E\left[|u^\top Z|^3\right]}.$$

By Lemma 8.8 in [8] it holds

$$E\left[(u^\top Z)^2\right] = 2E\left[(u^\top X)^2\right] = 2u^\top \mathrm{Var}(X)u$$

and

$$E\left[(u^\top Z)^3\right] \leq 4E\left[(u^\top X)^3\right] \leq 4|u|^3 E\left[|X|^3\right].$$

Moreover by Lemma C.9 we have

$$E\left[|X|^3\right] \leq 1 + C\sum_{s=1}^S E\left[X_i^6\right].$$

Therefore

$$\varphi_{X-Y}(u) \leq e^{-u^\top \mathrm{Var}(X)u + \frac{2|u|^3\theta}{3}\left[1 + C\sum_{s=1}^S E\left[X_i^6\right]\right]}$$

and the result follows from $|\theta| \leq 1$. □

The following lemma is a minor variation of commonly used techniques to prove total variation Central Limit Theorems.

Lemma C.15. *Let $(X_J)_{J\geq 1}$ and X be $\mathbb{R}^S$-valued random variables with characteristic functions $(\varphi_J)_{J\geq 1}$ and φ, respectively. Denote by $L^1(\mathbb{R}^S)$ the space of complex-valued integrable functions with domain $\mathbb{R}^S$. If*

(a) X_J converges weakly to X as $J \to \infty$

(b) φ belongs to $L^1(\mathbb{R}^S)$, i.e. $\int_{\mathbb{R}^S} |\varphi(t)|\, dt < \infty$

(c) $\lim_{A\to\infty} \limsup_{J\to\infty} \int_{|t|\geq A} |\varphi_J(t)|\, dt = 0$.

then X_J converges to X in total variation as $J \to \infty$.

Proof. First we prove that $\lim_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} = 0$. By the triangle inequality, for every $A > 0$ we have

$$\|\varphi_J - \varphi\|_{L^1} \leq \int_{|t|<A} |\varphi_J(t) - \varphi(t)|\, \mathrm{d}t + \int_{|t|\geq A} |\varphi_J(t)|\, \mathrm{d}t + \int_{|t|\geq A} |\varphi(t)|\, \mathrm{d}t. \tag{70}$$

Since weak convergence implies pointwise convergence of characteristic functions, assumption (a) implies that $\varphi_J(t) \to \varphi(t)$ as $J \to \infty$ for every $t \in \mathbb{R}^S$. Thus by the Dominated

Convergence Theorem and $|\varphi_J(t) - \varphi(t)| \le |\varphi_J(t)| + |\varphi(t)| = 2$, we have $\int_{|t|<A} |\varphi_J(t) - \varphi(t)| \, \mathrm{d}t \to 0$ as $J \to \infty$ for every $A > 0$. It follows by (70) that

$$0 \le \limsup_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} \le \int_{|t|\ge A} |\varphi(t)| \, \mathrm{d}t + \limsup_{J\to\infty} \int_{|t|\ge A} |\varphi_J(t)| \, \mathrm{d}t \,, \tag{71}$$

for every $A > 0$. By assumption (b) $\lim_{A\to\infty} \int_{|t|\ge A} |\varphi(t)| \, \mathrm{d}t = 0$. Combining with assumption (c), taking the limit $A \to \infty$ we obtain $\limsup_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} \le 0$ and thus $\lim_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} = 0$.

Then, note that $\varphi \in L^1(\mathbb{R}^S)$ and $\|\varphi_J - \varphi\|_{L^1} \to 0$ as $J \to \infty$ imply $\varphi_J \in L^1(\mathbb{R}^S)$ eventually as $J \to \infty$, since by the triangle inequality

$$\|\varphi_J\|_{L^1} \le \|\varphi_J - \varphi\|_{L^1} + \|\varphi\|_{L^1} < \infty$$

for J large enough. Thus, by the Inversion formula, for J large enough X_J and X admit density functions w.r.t. the Lebesgue measure, which can be written as $f_{X_J}(\boldsymbol{t}) = \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi_J(t) \, \mathrm{d}t$ and $f_X(\boldsymbol{t}) = \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi(t) \, \mathrm{d}t$. Thus

$$\begin{aligned} |f_{X_J}(\boldsymbol{t}) - f_X(\boldsymbol{t})| &= \left| \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi_J(t) \, \mathrm{d}t - \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi(t) \, \mathrm{d}t \right| \\ &\le \int_{\mathbb{R}^S} \left| e^{-i\boldsymbol{t}^\top t} (\varphi_J(t) - \varphi(t)) \right| \mathrm{d}t \le \|\varphi_J - \varphi\|_{L^1} \to 0 \end{aligned}$$

as $J \to \infty$ for every $\boldsymbol{t} \in \mathbb{R}^S$. By Scheffé Theorem, total variation convergence is implied by pointwise convergence of the densities. □

Proof of Lemma 4.4. Fix $\tilde{\psi} \in \mathbb{R}^D$ and denote $\mu = C(\psi^*)\tilde{\psi}$ and $\Xi = V(\psi^*)$. We will prove conditions (a), (b) and (c) of Lemma C.15 to show that $\mathcal{L}(\mathrm{d}\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) \overset{TV}{\to} N(\mu, \Xi)$ for $Q_{\psi^*}^{(\infty)}$-almost every Y as $J \to \infty$.

Condition (a) is shown in Proposition C.11. Regarding condition (b), the characteristic function of the limiting distribution $N(\mu, \Xi)$ is $\varphi(t) = e^{i\mu^\top t - \frac{1}{2} t^\top \Xi t}$, which is integrable since Ξ is positive definite by (B4).

We now turn to condition (c). Let

$$\tilde{\varphi}(t \mid Y_{1:J}, \psi) = \mathbb{E}\left[e^{it^\top \tilde{\boldsymbol{T}}} \mid Y_{1:J}, \psi \right] \qquad\qquad t \in \mathbb{R}^S$$

be the characteristic function of $\mathcal{L}\left(\mathrm{d}\tilde{\boldsymbol{T}} \mid Y_{1:J}, \psi\right)$. Using the definition of $\tilde{\boldsymbol{T}}$ in (19), and the fact that $T_s(\theta_j)$ are conditionally independent given $\tilde{\psi}$, we can write $\tilde{\varphi}$ as

$$\tilde{\varphi}(t \mid Y_{1:J}, \tilde{\psi}) = e^{-it^\top \alpha_J} \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)} \right),$$

where $\alpha_J = C(\psi^*)\Delta_J + \frac{1}{\sqrt{J}} \sum_{j=1}^{J} M^{(1)}(\psi^* \mid Y_j)$, $\varphi(t \mid Y_j, \psi) = E\left[e^{it^\top T(\theta_j)} \mid Y_j, \psi \right]$ as in the definition of (B5) and $\psi^{(J)}$ as in (61). Since $\alpha_J \in \mathbb{R}^S$ we have $|e^{-it^\top \alpha_J}| = 1$ and thus

$$|\tilde{\varphi}(t \mid Y_{1:J}, \psi)| = \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi \right) \right|. \tag{72}$$

For every $\epsilon > 0$, by (72) and the subadditivity of lim sup we have

$$\lim_{A\to\infty} \limsup_{J\to\infty} \int_{|t|>A} \left| \tilde{\varphi}(t \mid Y_{1:J}, \tilde{\psi}) \right| \mathrm{d}t \le$$

$$\lim_{A\to\infty} \limsup_{J\to\infty} \int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)} \right) \right| \mathrm{d}t + \limsup_{J\to\infty} \int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)} \right) \right| \mathrm{d}t.$$

Lemma C.16 shows that the second lim sup in the last line is equal to 0 for every $\epsilon > 0$, while Lemma C.17 shows that the $\lim_{A\to\infty} \limsup_{J\to\infty}$ term goes to 0 when ϵ is chosen as in (73). Thus condition (c) follows by taking ϵ as in (73) in the above inequality. □

Lemma C.16. *Under the same setting and notation as in the proof of Lemma 4.4, for every $\epsilon > 0$ we have*

$$\limsup_{J\to\infty} \int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| dt = 0$$

$Q_{\psi^*}^{(\infty)}$*-almost surely.*

Proof. Consider the change of variables $x = t/\sqrt{J}$. Then

$$\int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t = J^{S/2} \int_{|x|>\epsilon} \left| \prod_{j=1}^{J} \varphi\left(x \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}x.$$

Let k and B_{δ_5} be as in $(B5)$ and k' and B_{δ_6} be as in $(B6)$. Take J high enough so that $J \geq 2k$ as well as $\psi^{(J)} \in B := B_{\delta_5} \cap B_{\delta_6}$, so that

$$\int_{|x|>\epsilon} \left| \prod_{j=1}^{J} \varphi\left(x \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}x \leq \sup_{\psi\in B} \int_{|x|>\epsilon} \left| \prod_{j=1}^{2k} \varphi\left(x \mid Y_j, \psi\right) \right| \left| \prod_{j=2k+1}^{J} \varphi\left(x \mid Y_j, \psi\right) \right| \mathrm{d}x\,.$$

For every $a \in \mathbb{R}_+$ denote its integer part as $\lfloor a \rfloor$. By (B6), for every $\psi \in B$ we have

$$\left| \prod_{j=2k+1}^{J} \varphi\left(x \mid Y_j, \psi\right) \right| \leq \prod_{s=1}^{\lfloor \frac{J-2k}{k'} \rfloor} A_s \leq \phi(\epsilon)^{\lfloor \frac{J-2k}{k'} \rfloor}, \qquad \text{with } A_s = \left| \prod_{j=2k+1+(s-1)k'}^{2k+1+sk'} \varphi\left(x \mid Y_j, \psi\right) \right|$$

almost surely, where we exploited the fact that each A_s is distributed as $\varphi^{(k')}\left(t \mid Y_{1:k'}, \psi\right)$ in (B6). Therefore

$$\int_{|x|>\epsilon} \left| \prod_{j=1}^{J} \varphi\left(x \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}x \leq \phi(\epsilon)^{\lfloor \frac{J-2k}{k'} \rfloor} \sup_{\psi\in B} \int_{|x|>\epsilon} \left| \prod_{j=1}^{2k} \varphi\left(x \mid Y_j, \psi\right) \right| \mathrm{d}x.$$

almost surely. By Hölder Inequality and $(B5)$, we have

$$c = \sup_{\psi\in B} \int_{|x|>\epsilon} \left| \prod_{j=1}^{2k} \varphi\left(x \mid Y_j, \psi\right) \right| \mathrm{d}x \leq \sup_{\psi\in B} \int_{\mathbb{R}^S} \left| \prod_{j=1}^{2k} \varphi\left(x \mid Y_j, \psi\right) \right| \mathrm{d}x \leq$$
$$\left\{ \sqrt{\sup_{\psi\in B} \int_{\mathbb{R}^S} \left| \prod_{j=1}^{k} \varphi\left(x \mid Y_j, \psi\right) \right|^2 \mathrm{d}x} \right\} \left\{ \sqrt{\sup_{\psi\in B} \int_{\mathbb{R}^S} \left| \prod_{j=k+1}^{2k} \varphi\left(x \mid Y_j, \psi\right) \right|^2 \mathrm{d}x} \right\} < \infty,$$

almost surely. Therefore it holds

$$\int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t \leq J^{S/2} \phi(\epsilon)^{\lfloor \frac{J-2k}{k'} \rfloor} c,$$

that goes to 0 as $J \to \infty$, since $\phi(\epsilon) < 1$ by $(B6)$. □

Lemma C.17. *Under the same setting and notation as in the proof of Lemma 4.4, let $\lambda > 0$ be such that the matrix $V(\psi^*) - \lambda I$ is positive definite. Such λ can be found, since $V(\psi^*)$ is positive definite by (B4). Then, given*

$$\epsilon = \frac{\lambda}{1 + C \sum_{s=1}^{S} E\left[T_s(\theta_1)^6 \mid \psi^*\right]} \tag{73}$$

we have

$$\lim_{A\to\infty} \limsup_{J\to\infty} \int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| dt = 0$$

$Q_{\psi^*}^{(\infty)}$-almost surely.

Proof. By Corollary C.14, we have

$$|\varphi(u \mid Y_j, \psi)|^2 \le e^{-u^\top \mathrm{Var}(T(\theta_j)|Y_j,\psi)u + \frac{2|u|^3}{3}\left[1 + C\sum_{s=1}^{S} E\left[T_s(\theta_j)^6 | Y_j,\psi\right]\right]},$$

for every $u \in \mathbb{R}^S$ and $\psi \in \mathbb{R}^D$. Therefore

$$\left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi\right)\right|^2 \le e^{-t^\top \frac{1}{J}\sum_{j=1}^{J} \mathrm{Var}(T(\theta_j)|Y_j,\psi)t + \frac{2|t|^3}{3\sqrt{J}}\left[1 + C\frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{S} E\left[T_s(\theta_j)^6 | Y_j,\psi\right]\right]}. \tag{74}$$

Notice that in the proof of (65) we have shown through (B4) that

$$\frac{1}{J}\sum_{j=1}^{J} E\left[T_s(\theta_j)^6 \mid Y_j, \psi^{(J)}\right] \to E\left[T_s(\theta_1)^6 \mid \psi^*\right] \tag{75}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$, for every $s = 1, \dots, S$. Thus, combining (73) and (75), for every $|t| \le \epsilon\sqrt{J}$ we have

$$\left| e^{\frac{2|t|^3}{3\sqrt{J}}\left[1 + C\frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{S} E\left[T_s(\theta_j)^6 | Y_j,\psi\right]\right]}\right|^2 \le e^{\lambda t^\top t}, \tag{76}$$

almost surely for J high enough. Finally by (74) and (76)

$$\int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| \mathrm{d}t \le \int_{|t|>A} e^{-t^\top \Xi^{(J)} t}\, \mathrm{d}t\,, \tag{77}$$

with

$$\Xi^{(J)} = \frac{1}{J}\sum_{j=1}^{J} \mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^{(J)}\right) - \lambda I\,.$$

Since $\Xi^{(J)} \to V(\psi^*) - \lambda I$ by (64), and $V(\psi^*) - \lambda I$ is positive definite by definition of λ, by Dominated Convergence Theorem

$$\limsup_{J} \int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| \mathrm{d}t \le \int_{|t|>A} e^{-t^\top (V(\psi^*) - \lambda I) t}\, \mathrm{d}t\,, \tag{78}$$

Since the right hand side of (78) is integrable the conclusion follows by taking $A \to \infty$. □

C.14 Proof of Theorem 4.2

We first need a technical lemma.

Lemma C.18. *Let $\left\{Y^{(n)}\right\}_n$ be a sequence of random elements with state space $\mathcal{Y}^{(n)}$, such that $Y^{(n)} \sim Q^{(n)}$ with $Q^{(n)} \in \mathcal{P}\left(\mathcal{Y}^{(n)}\right)$. Let $\{\pi_n\}_n$ be a sequence of Markov kernels from $\mathcal{Y}^{(n)}$ to $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and let $\pi \in \mathcal{P}(\mathcal{X})$. If*

$$\|\pi_{n,1}(\cdot) - \pi_1(\cdot)\|_{TV} \to 0 \quad \text{and} \quad \|\pi_n(\cdot \mid x) - \pi(\cdot \mid x)\|_{TV} \to 0, \text{ for } \pi_1\text{-almost every } x \in \mathcal{X}_1,$$

as $n \to \infty$ in $Q^{(n)}$-probability, where $\pi_{n,1}$ and π_1 are the marginal distributions on $\mathcal{X}_1$ of π_n and π respectively, then

$$\|\pi_n(\cdot) - \pi(\cdot)\|_{TV} \to 0,$$

as $n \to \infty$ in $Q^{(n)}$-probability

Proof. Let $f : \mathcal{X} \to [0,1]$ be a measurable function. By the triangular inequality we have

$$\begin{aligned}\left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_1,\mathrm{d}x_2) - \int_{\mathcal{X}} f(x_1,x_2)\pi(\mathrm{d}x_1,\mathrm{d}x_2)\right| \leq \\ \left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_{n,1}(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| + \\ \left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right|.\end{aligned}$$

Notice that

$$\begin{aligned}\sup_f \left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_{n,1}(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| \\ \leq \|\pi_{n,1}(\cdot) - \pi_1(\cdot)\|_{TV} \to 0,\end{aligned}$$

as $n \to \infty$ in $Q^{(n)}$-probability, by assumption. Moreover we have

$$\begin{aligned}\sup_f \left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| \leq \\ \int_{\mathcal{X}_1} \sup_f \left|\int_{\mathcal{X}_2} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1) - \int_{\mathcal{X}_2} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\right| \pi_1(\mathrm{d}x_1).\end{aligned}$$

The integrand on the right hand side goes to 0 as $n \to \infty$ in $Q^{(n)}$-probability, by assumption. Therefore, by Dominated Convergence Theorem, we have

$$\sup_f \left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| \to 0,$$

as $n \to \infty$ in $Q^{(n)}$-probability, as desired. □

Proof of Theorem 4.2. Lemma 4.3 shows that $\tilde{\psi}$ converges to a Normal distribution with zero mean and non-singular covariance matrix $\mathcal{I}^{-1}(\psi^*)$. Similarly, Lemma 4.4 shows that, conditional to every $\tilde{\psi}$, $\tilde{\boldsymbol{T}}$ converges to a Normal distribution with mean and variance (denoted by $E_\infty[\cdot]$ and $\mathrm{Var}_\infty(\cdot)$) given by

$$E_\infty[\tilde{\boldsymbol{T}} \mid \tilde{\psi}] = C(\psi^*)\tilde{\psi}, \quad \mathrm{Var}_\infty\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right) = V(\psi^*).$$

Therefore, by Lemma C.18, we conclude that $\left(\tilde{\boldsymbol{T}}, \tilde{\psi}\right)$ converges in total variation to a $(S+D)$-dimensional Gaussian distribution $\tilde{\pi}$ with zero mean and covariance matrix Σ given by

$$\Sigma = \begin{bmatrix} \Sigma_{\tilde{\boldsymbol{T}}} & \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}}^\top \\ \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}} & \Sigma_{\tilde{\psi}} \end{bmatrix},$$

where $\Sigma_{\tilde{\psi}} = \mathcal{I}^{-1}(\psi^*) \in \mathbb{R}^{D\times D}$ and $\Sigma_{\tilde{\boldsymbol{T}}} \in \mathbb{R}^{S\times S}$ are the limiting variances of $\tilde{\psi}$ and $\tilde{\boldsymbol{T}}$, while $\Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}} \in \mathbb{R}^{D\times S}$ is the limiting covariance. Thus, thanks to standard properties of the multivariate Gaussian distribution, the determinant of Σ can be computed as

$$\begin{aligned}\det(\Sigma) &= \det(\Sigma_{\tilde{\psi}})\det\left(\Sigma_{\tilde{\boldsymbol{T}}} - \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}}^{\top}\Sigma_{\tilde{\psi}}^{-1}\Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}}\right) = \det(\Sigma_{\tilde{\psi}})\det\left(\mathrm{Var}_{\infty}\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right)\right)\\ &= \det\left(\mathcal{I}^{-1}(\psi^*)\right)\det\left(V(\psi^*)\right),\end{aligned}$$

which implies that Σ is non singular. Indeed, $\det\left(\mathcal{I}^{-1}(\psi^*)\right) > 0$ by $(B3)$, while $\det\left(V(\psi^*)\right) > 0$ by (B4). Therefore, by Theorem 1 in [55], the Gibbs sampler on the limit Gaussian target has a strictly positive spectral gap. Moreover, since the Gibbs sampler in (15) has two blocks, by Lemma C.6 we have $\tilde{t}_{mix}(\epsilon, M) < \infty$ for every M and ϵ: thus the result follows by Corollary 2.5. □

C.15 Proof of Proposition 4.5

Proof. Using the notation $E_{\infty}[\cdot]$,$\mathrm{Var}_{\infty}(\cdot)$ and $\mathrm{Cov}_{\infty}(\cdot,\cdot)$ for the limiting mean, variance and covariance, by Propositions 4.3 and 4.4 we have

$$E_{\infty}[\tilde{\psi}] = \mathbf{0}_D, \quad \mathrm{Var}_{\infty}(\tilde{\psi}) = \mathcal{I}^{-1}(\psi^*)$$

and

$$E_{\infty}[\tilde{\boldsymbol{T}} \mid \tilde{\psi}] = C(\psi^*)\tilde{\psi}, \quad \mathrm{Var}_{\infty}\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right) = V(\psi^*).$$

By standard properties of the multivariate Gaussian distribution we have

$$E_{\infty}[\tilde{\boldsymbol{T}}] = \mathbf{0}_S, \quad \mathrm{Cov}_{\infty}\left(\boldsymbol{T}, \tilde{\psi}\right) = C(\psi^*)\mathrm{Var}_{\infty}(\tilde{\psi}) = C(\psi^*)\mathcal{I}^{-1}(\psi^*)$$

and

$$\begin{aligned}\mathrm{Var}_{\infty}(\boldsymbol{T}) &= \mathrm{Var}_{\infty}\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right) + \mathrm{Cov}_{\infty}\left(\boldsymbol{T}, \tilde{\psi}\right)\mathrm{Var}_{\infty}^{-1}(\tilde{\psi})\mathrm{Cov}_{\infty}^{\top}\left(\boldsymbol{T}, \tilde{\psi}\right)\\ &= V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^{\top}(\psi^*),\end{aligned}$$

as desired. □

C.16 Proof of Corollary 4.6

We need three preliminary lemmas. The first one is a special version of well-known results (e.g. [55]).

Lemma C.19. *The Gibbs sampler targeting the distribution in Proposition 4.5 can be written as*

$$\begin{bmatrix}\tilde{\boldsymbol{T}}^{(t)}\\ \tilde{\psi}^{(t)}\end{bmatrix} = B\begin{bmatrix}\tilde{\boldsymbol{T}}^{(t-1)}\\ \tilde{\psi}^{(t-1)}\end{bmatrix} + \begin{bmatrix}U_1\\ U_2\end{bmatrix},$$

where

$$B = \begin{bmatrix}\boldsymbol{O}_{S\times S} & C(\psi^*)\\ \boldsymbol{O}_{D\times S} & \mathcal{I}^{-1}(\psi^*)C^{\top}(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^{\top}(\psi^*)\right\}^{-1}C(\psi^*)\end{bmatrix}$$

and

$$\begin{bmatrix}U_1\\ U_2\end{bmatrix} \sim N\left(\boldsymbol{0}_{S+D}, \Sigma - B\Sigma B^{\top}\right)$$

Proof. By Proposition 4.4 we have

$$E\left[\tilde{\boldsymbol{T}}^{(t)} \mid \tilde{\boldsymbol{T}}^{(t-1)}, \tilde{\psi}^{(t-1)}\right] = C(\psi^*)\tilde{\psi}^{(t-1)}.$$

Moreover, by Proposition 4.5 and standard properties of the multivariate Gaussian distribution, we have

$$\begin{aligned}
&E\big[\tilde{\psi}^t \mid \tilde{\boldsymbol{T}}^{(t-1)}, \tilde{\psi}^{(t-1)}\big] \\
&\quad = E\left[\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1}\tilde{\boldsymbol{T}}^{(t)} \mid \tilde{\boldsymbol{T}}^{(t-1)}, \tilde{\psi}^{(t-1)}\right] \\
&\quad = \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\tilde{\psi}^{(t-1)},
\end{aligned}$$

as desired. □

Lemma C.20. *Let*

$$M = \begin{bmatrix} \boldsymbol{O}_{S\times S} & A \\ \boldsymbol{O}_{D\times S} & W \end{bmatrix},$$

with $A \in \mathbb{R}^{S\times D}$ and $W \in \mathbb{R}^{D\times D}$. Then M and W have the same non null eigenvalues.

Proof. Let $\mu \neq 0$ be an eigenvalue of M, with eigenvector $x = [x_S^\top, x_D^\top]^\top$. We have

$$Mx = \mu x \quad \Leftrightarrow \quad \begin{bmatrix} Ax_D \\ Wx_D \end{bmatrix} = \begin{bmatrix} \mu x_S \\ \mu x_D \end{bmatrix},$$

so that μ is an eigenvalue of W with eigenvector x_D. Indeed, x_D is different from the null vector, since $\mu \neq 0$.

Let $\lambda \neq 0$ be an eigenvalue of W with eigenvector x_D. Then

$$M\begin{bmatrix} \frac{Ax_D}{\lambda} \\ x_D \end{bmatrix} = \begin{bmatrix} Ax_D \\ Wx_D \end{bmatrix} = \lambda \begin{bmatrix} \frac{Ax_D}{\lambda} \\ x_D \end{bmatrix},$$

so that λ is an eigenvalue of M, with eigenvector

$$\begin{bmatrix} \frac{Ax_D}{\lambda} \\ x_D \end{bmatrix},$$

as desired. □

Lemma C.21. *Let $A \in \mathbb{R}^{D\times S}$ and $B \in \mathbb{R}^{S\times D}$. Then the matrices AB and BA have the same non-null eigenvalues.*

Proof. Let $\lambda \neq 0$ be an eigenvalue of AB, with eigenvector $v \in \mathbb{R}^D$. Then

$$\lambda Bv = B(AB)v = (BA)Bv.$$

Since $Bv \neq \mathbf{0}$ we conclude that λ is an eigenvalue of BA with eigenvector Bv. □

Proof of Corollary 4.6. With B as in Lemma C.19, by Theorem 1 in [55] the spectral gap of the Gibbs sampler with operator $\tilde{P}$ is given by

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |\lambda_i| \,:\, \lambda_i \text{ eigenvalue of } B\right\}$$

Thus, by Lemma C.20, with $M := B$ and

$$W = \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*),$$

we have

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |\lambda_i| \,:\, \lambda_i \text{ eigenvalue of } \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\right\}.$$

By Lemma C.21 with

$$A = \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*), \quad B = \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)$$

we deduce

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |\lambda_i| \,:\, \lambda_i \text{ eigenvalue of } \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}.$$

Notice that

$$\begin{aligned}&\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\\ &= I - \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} V(\psi^*).\end{aligned}$$

Since λ is an eigenvalue of A if and only if $1-\lambda$ is an eigenvalue of $I - A$, it follows that

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |1 - \lambda_i| \,;\, \lambda_i \text{ eigenvalue of } \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} V(\psi^*)\right\}.$$

Moreover the eigenvalues of the inverse are the inverse of the eigenvalues, so that the rate of convergence is equal to

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - \left|1 - \frac{1}{\lambda_i}\right| \,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}\right\}.$$

Since

$$V^{-1}(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\} = I + V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*),$$

we have

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - \left|1 - \frac{1}{1+\lambda_i}\right| \,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}.$$

Moreover both $V^{-1}(\psi^*)$ and $C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)$ are positive semi-definite, so that also their product is positive semi-definite and has positive eigenvalues. Therefore we conclude

$$\mathrm{Gap}(\tilde{P}) = \min\left\{\frac{1}{1+\lambda_i} \,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}$$

and the result follows by Corollary 2.6. □

C.17 Proof of Corollary 4.7

We need a preliminary lemma, that we prove for self-containedness.

Lemma C.22. *Let $p(\theta \mid \psi)$ be as in (14). Then it holds*

$$E[T(\theta) \mid \psi] = \frac{\partial_\psi A(\psi)}{\partial_\psi \eta(\psi)}, \quad Var(T(\theta) \mid \psi) = \left\{\partial^2_\psi A(\psi) - \frac{\partial^2_\psi \eta(\psi)\partial_\psi A(\psi)}{\partial_\psi \eta(\psi)}\right\}[\partial_\psi \eta(\psi)]^{-2}.$$

Proof. Differentiating the following equality

$$1 = \int p(\theta \mid \psi)\,\mathrm{d}\theta, \tag{79}$$

by the regularity properties of the exponential family we get

$$0 = \int \partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta = \partial_\psi \eta(\psi) E[T(\theta) \mid \psi] + \partial_\psi A(\psi),$$

and the formula for the expected value follows. As regards the variance, differentiating (79) twice, we obtain

$$0 = \partial_\psi^2\eta(\psi)E[T(\theta) \mid \psi] - \partial_\psi^2 A(\psi) + \left[\partial_\psi\eta(\psi)\right]^2 E[T^2(\theta) \mid \psi] - 2\left[\partial_\psi\eta(\psi)\right]^2 E^2[T(\theta) \mid \psi] + \left[\partial_\psi A(\psi)\right]^2.$$

Noticing that

$$\left[\partial_\psi\eta(\psi)\right]^2 E^2[T(\theta) \mid \psi] = \left[\partial_\psi A(\psi)\right]^2$$

and rearranging, we get

$$\partial_\psi^2 A(\psi) - \partial_\psi^2\eta(\psi)E[T(\theta) \mid \psi] = \left[\partial_\psi\eta(\psi)\right]^2 \mathrm{Var}(T(\theta) \mid \psi),$$

from which the result follows. □

Proof of Corollary 4.7. By Corollary 4.6, we have

$$\gamma(\psi^*) = \frac{1}{1+\lambda} \quad \text{with } \lambda = \frac{C^2(\psi^*)}{V(\psi^*)\mathcal{I}(\psi^*)},$$

where

$$\begin{aligned} C(\psi) &= E_{Y_j}\left[\partial_\psi E[T(\theta_j) \mid Y_j, \psi]\right], \\ V(\psi) &= E_{Y_j}\left[\mathrm{Var}(T(\theta_j) \mid Y_j, \psi)\right], \\ \mathcal{I}(\psi) &= -E_{Y_j}\left[\partial_\psi^2 \log g(Y_j \mid \psi)\right], \end{aligned}$$

with $g(y \mid \psi)$ as in (16). As regards $C(\psi)$, notice that

$$\begin{aligned} \partial_\psi E[T(\theta) \mid Y, \psi] =& \frac{\int T(\theta) f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} - \\ & \frac{\left[\int T(\theta) f(Y \mid \theta) p(\theta \mid \psi)\,\mathrm{d}\theta\right]\left[\int f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta\right]}{g^2(Y \mid \psi)} \\ =& \partial_\psi\eta(\psi)E\left[T^2(\theta) \mid Y, \psi\right] - \partial_\psi\eta(\psi)E^2\left[T(\theta) \mid Y, \psi\right] \\ =& \partial_\psi\eta(\psi)\mathrm{Var}\left(T(\theta) \mid Y, \psi\right). \end{aligned}$$

Therefore

$$C^2(\psi^*) = \left[\partial_\psi\eta(\psi^*)\right]^2 E_{Y_j}^2\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right]. \tag{80}$$

As regards $\mathcal{I}(\psi)$, notice that

$$\partial_\psi \log g(Y_j \mid \psi) = \frac{\int f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} = \partial_\psi\eta(\psi)\frac{\int T(\theta) f(Y \mid \theta) p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} - \partial_\psi A(\psi)$$

and

$$\begin{aligned} \partial_\psi^2 \log g(Y_j \mid \psi) &= \partial_\psi^2\eta(\psi)E\left[T(\theta) \mid Y, \psi\right] - \partial_\psi^2 A(\psi) + \partial_\psi\eta(\psi)\frac{\int T(\theta) f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} \\ &\quad - \partial_\psi\eta(\psi)\frac{\left[\int T(\theta) f(Y \mid \theta) p(\theta \mid \psi)\,\mathrm{d}\theta\right]\left[\int f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta\right]}{g^2(Y \mid \psi)} \\ &= \partial_\psi^2\eta(\psi)E\left[T(\theta) \mid Y, \psi\right] - \partial_\psi^2 A(\psi) + \left[\partial_\psi\eta(\psi)\right]^2 \mathrm{Var}\left(T(\theta) \mid Y, \psi\right). \end{aligned}$$

Noticing that, by Lemma C.22, we have

$$\begin{aligned} \partial_\psi^2\eta(\psi)E\left[T(\theta) \mid Y, \psi\right] - \partial_\psi^2 A(\psi) &= \left\{\partial_\psi^2 A(\psi) - \frac{\partial_\psi^2\eta(\psi)\partial_\psi A(\psi)}{\partial_\psi\eta(\psi)}\right\} \\ &= \left[\partial_\psi\eta(\psi)\right]^2 \mathrm{Var}\left(T(\theta) \mid \psi\right), \end{aligned}$$

we get

$$\begin{aligned}\mathcal{I}(\psi^*) &= \left[\partial_\psi \eta(\psi^*)\right]^2 \mathrm{Var}\left(T(\theta_j) \mid \psi^*\right) - \left[\partial_\psi \eta(\psi^*)\right]^2 E_{Y_j}\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right] \\ &= \left[\partial_\psi \eta(\psi^*)\right]^2 \mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid Y_j, \psi^*\right]\right),\end{aligned} \tag{81}$$

by the Law of Total Variance. Combining (80) and (81), it holds

$$\lambda = \frac{E^2_{Y_j}\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right]}{V(\psi^*)\mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid Y_j, \psi^*\right]\right)} = \frac{E_{Y_j}\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right]}{\mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid Y_j, \psi^*\right]\right)}.$$

The expression for $\gamma(\psi^*)$ follows by rearranging and applying the Law of Total Variance. □

C.18 Proof of Proposition 5.1

First of all notice that, by Bayes' Theorem, we have

$$\theta_j \mid Y_j, \mu, \tau_1 \overset{\text{ind.}}{\sim} N\left(m_j, (m\tau_0 + \tau_1)^{-1}\right), \tag{82}$$

where

$$m_j = \frac{m\tau_0}{m\tau_0 + \tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0 + \tau_1}\mu.$$

Recall that by (B1) we have

$$Y_j \overset{\text{iid}}{\sim} g(\cdot \mid \psi^*) = N\left(\mu^*, (\tau_0^*)^{-1} I + (\tau_1^*)^{-1}\mathbb{H}\right),$$

so that

$$\bar{Y}_j = \frac{1}{m}\sum_{i=1}^{m} Y_{j,i} \overset{\text{iid}}{\sim} N\left(\mu^*, \frac{1}{\tau_1^*} + \frac{1}{m\tau_0^*}\right). \tag{83}$$

Moreover we need some preliminary lemmas.

Lemma C.23. *Let $X \sim N(\nu, \sigma^2)$. Then*

$$E[X^p] = \sum_{i=0}^{p} \binom{p}{i} \nu^i \sigma^{p-i} E[Z^{p-i}],$$

where $Z \sim N(0,1)$ and

$$E[Z^s] = \begin{cases} 0 & \text{if } s \text{ is odd} \\ 2^{-s/2}\frac{s!}{(s/2)!} & \text{if } s \text{ is even} \end{cases}$$

Proof. The result follows by noticing $X = \nu + \sigma Z$ and applying Netwon's Binomial Theorem. □

Lemma C.24. *Let A be $m \times m$ matrix such that $A = aI + b\mathbb{H}$, with $a \neq b$ and $a \neq (1-m)b$. Then $det(A) = [a + mb]a^{m-1}$ and $A^{-1} = \frac{1}{a}\mathbb{I} - \frac{b}{a(a+mb)}\mathbb{H}$.*

Proof. We start by the determinant

$$\det\begin{pmatrix} c & d & \cdots & d \\ d & c & \cdots & d \\ \vdots & \vdots & \ddots & \vdots \\ d & d & \cdots & c \end{pmatrix} = [c + (m-1)d]\det\begin{pmatrix} 1 & 1 & \cdots & 1 \\ d & c & \cdots & d \\ \vdots & \vdots & \ddots & \vdots \\ d & d & \cdots & c \end{pmatrix}$$

$$= [c + (m-1)d]\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & c-d & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c-d \end{pmatrix} = [c + (m-1)d](c-d)^{m-1},$$

where the first equality comes by adding to the first row all the others, while the second comes by subtracting the first row (scaled by d) from all the others. In our case $c = a + b$ and $d = b$, that is $\det(A) = [a + mb]a^{m-1}$, as desired. With our assumptions we get that the determinant is different from zero.

As regards the inverse we prove $A^{-1} = xI + y\mathbb{H}$ for suitable x and y. Indeed

$$(aI + b\mathbb{H})(xI + y\mathbb{H}) = axI + ay\mathbb{H} + bx\mathbb{H} + by\mathbb{H}^2 = axI + (ay + bx + mby)\mathbb{H}.$$

Setting the above equal to I, we obtain $x = 1/a$ and

$$ay + bx + mby = 0 \quad \Rightarrow \quad y(a + mb) = -\frac{b}{a} \quad \Rightarrow \quad y = -\frac{b}{a(a + mb)}$$

as desired. □

Lemma C.25. *Consider the marginal likelihood as in* (23)*, with* $\psi^* = (\mu^*, \tau_1^*, \tau_0^*)$*. Then we have*

$$\mathcal{I}(\psi^*) = \begin{pmatrix} \frac{m\tau_0^*\tau_1^*}{\tau_1^* + m\tau_0^*} & 0 & 0 \\ 0 & \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2(\tau_1^* + m\tau_0^*)^2} & \frac{m}{2(\tau_1^* + m\tau_0^*)^2} \\ 0 & \frac{m}{2(\tau_1^* + m\tau_0^*)^2} & \frac{m-1}{2(\tau_0^*)^2} + \frac{(\tau_1^*)^2}{2(\tau_0^*)^2(\tau_1^* + m\tau_0^*)^2} \end{pmatrix} \tag{84}$$

Proof. The log–likelihood $l(\psi) = \log g(y \mid \psi)$ is given by

$$l(\mu, \tau_0, \tau_1) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}(Y_1 - \mu I)^t \Sigma^{-1}(Y_1 - \mu I),$$

with $\Sigma = \tau_0^{-1} I + \tau_1^{-1}\mathbb{H}$. By Lemma C.24 with $a = \tau_0^{-1}$ and $b = \tau_1^{-1}$ we have

$$\det(\Sigma) = [\tau_0^{-1} + m\tau_1^{-1}](\tau_0^{-1})^{m-1}, \quad \Sigma^{-1} = \tau_0 I - \frac{\tau_0^2}{\tau_1 + m\tau_0}\mathbb{H}.$$

Thus, the log–likelihood becomes

$$\begin{aligned} l(\mu, \tau_0, \tau_1) = &-\frac{1}{2}\log 2\pi + \frac{m-1}{2}\log\tau_0 - \frac{1}{2}\log(\tau_0^{-1} + m\tau_1^{-1}) - \frac{\tau_0}{2}\sum_{i=1}^m (Y_{1,i} - \mu)^2 \\ &+ \frac{\tau_0^2}{2(\tau_1 + m\tau_0)}(Y_1 - \mu I)^t \mathbb{H}(Y_1 - \mu I). \end{aligned}$$

Rewriting the last expression we get

$$\begin{aligned} l(\mu, \tau_0, \tau_1) = &-\frac{1}{2}\log 2\pi + \frac{m-1}{2}\log\tau_0 - \frac{1}{2}\log(\tau_0^{-1} + m\tau_1^{-1}) - \frac{\tau_0}{2}\sum_{i=1}^m (Y_{1,i} - \mu)^2 \\ &+ \frac{\tau_0^2}{2(\tau_1 + m\tau_0)}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2. \end{aligned}$$

The required derivatives are given by

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{m\tau_0\tau_1}{\tau_1 + m\tau_0}, \quad \frac{\partial^2 l}{\partial \tau_1^2} = -\frac{m\tau_0(2\tau_1 + m\tau_0)}{2\tau_1^2(\tau_1 + m\tau_0)^2} + \frac{\tau_0^2}{(\tau_1 + m\tau_0)^3}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2,$$

$$\frac{\partial^2 l}{\partial \tau_0^2} = -\frac{m-1}{2\tau_0^2} - \frac{\tau_1(\tau_1 + 2m\tau_0)}{2\tau_0^2(\tau_1 + m\tau_0)^2} + \frac{(\tau_1 + m\tau_0)^2 - 2m\tau_0\tau_1 - m^2\tau_0^2}{(\tau_1 + m\tau_0)^3}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2,$$

$$\frac{\partial^2 l}{\partial \mu \partial \tau_0} = \sum_{i=1}^m (Y_{1,i} - \mu) - \frac{2m\tau_0\tau_1 + m^2\tau_0^2}{(\tau_1 + m\tau_0)^2}\sum_{i=1}^m (Y_{1,i} - \mu),$$

$$\frac{\partial^2 l}{\partial \mu \partial \tau_1} = \frac{\tau_0^2}{(\tau_1 + m\tau_0)^2}\sum_{i=1}^m (Y_{1,i} - \mu), \quad \frac{\partial^2 l}{\partial \tau_0 \partial \tau_1} = \frac{m}{2(\tau_1 + m\tau_0)^2} - \frac{\tau_0\tau_1}{(\tau_1 + m\tau_0)^3}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2.$$

The entries of the Fisher Information matrix reported in (84) can then be computed from the above expressions by taking expectations with respect to Y_1 and exploiting that

$$\begin{aligned}
&\mathbb{E}[Y_{1,i}-\mu]=0, \quad \mathbb{E}\left[(Y_{1,i}-\mu)^2\right]=Var(Y_{1,i}-\mu)=\frac{\tau_0+\tau_1}{\tau_0\tau_1},\\
&\mathbb{E}\left[\left(\sum_{i=1}^{m}(Y_{1,i}-\mu)\right)^2\right]=Var\left(\sum_{i=1}^{m}(Y_{1,i}-\mu)\right)=[1,\dots,1]Var(Y_1)[1,\dots,1]^t\\
&\qquad\qquad=[1,\dots,1]\left(\tau_0^{-1}I+\tau_1^{-1}\mathbb{H}\right)[1,\dots,1]^t\\
&\qquad\qquad=m\left(\frac{m\tau_0+\tau_1}{\tau_0\tau_1}\right).
\end{aligned}$$

Thus we can compute the entries of the Fisher Information matrix as

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 l}{\partial\tau_0^2}\right]&=-\frac{m-1}{2\tau_0^2}-\frac{\tau_1(\tau_1+2m\tau_0)}{2\tau_0^2(\tau_1+m\tau_0)^2}+\frac{m(\tau_1+m\tau_0)^2-2m^2\tau_0\tau_1-m^3\tau_0^2}{\tau_0\tau_1(\tau_1+m\tau_0)^2}\\
&=-\frac{m-1}{2\tau_0^2}-\frac{\tau_1^2}{2\tau_0^2(\tau_1+m\tau_0)^2},\\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\tau_1^2}\right]&=-\frac{m\tau_0(2\tau_1+m\tau_0)}{2\tau_1^2(\tau_1+m\tau_0)^2}+\frac{m\tau_0}{\tau_1(\tau_1+m\tau_0)^2}=-\frac{m^2\tau_0^2}{2\tau_1^2(\tau_1+m\tau_0)^2},\\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\mu\partial\tau_0}\right]&=0, \quad \mathbb{E}\left[\frac{\partial^2 l}{\partial\mu\partial\tau_1}\right]=0,\\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\tau_0\partial\tau_1}\right]&=\frac{m}{2(\tau_1+m\tau_0)^2}-\frac{m}{(\tau_1+m\tau_0)^2}=-\frac{m}{2(\tau_1+m\tau_0)^2},
\end{aligned}$$

as desired. □

Lemma C.26. *Let $X\sim N(\nu,\sigma^2)$. Then*

$$\left|E\left[e^{i(aX^2+bX)}\right]\right|\leq\frac{e^{-\frac{\sigma^2}{2}\frac{(2\nu a+b)^2}{1+4a^2\sigma^4}}}{(1+4a^2\sigma^4)^{1/4}},$$

for every $(a,b)\in\mathbb{R}_2$.

Proof. By definition of expectation we have

$$E\left[e^{i(aX^2+bX)}\right]=\int_{\mathbb{R}}e^{i(az^2+bz)}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(z-\nu)^2}{2\sigma^2}}\,\mathrm{d}z=\frac{e^{-\frac{\nu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}e^{-\frac{1}{2}\left[z^2\left(\frac{1}{\sigma^2}-2ia\right)-2z\left(\frac{\nu}{\sigma^2}+ib\right)\right]}\,\mathrm{d}z$$

Notice that

$$\begin{aligned}
z^2\left(\frac{1}{\sigma^2}-2ia\right)-2z\left(\frac{\nu}{\sigma^2}+ib\right)&=\left(\frac{1-2ia\sigma^2}{\sigma^2}\right)\left[z^2-2z\frac{\nu+ib\sigma^2}{1-2ia\sigma^2}+\left(\frac{\nu+ib\sigma^2}{1-2ia\sigma^2}\right)-\left(\frac{\nu+ib\sigma^2}{1-2ia\sigma^2}\right)^2\right]\\
&=\left(\frac{1-2ia\sigma^2}{\sigma^2}\right)\left(z-\frac{\nu+i\sigma^2 b}{1-2ia\sigma^2}\right)^2-\frac{(\nu+ib\sigma^2)^2}{\sigma^2(1-2ia\sigma^2)},
\end{aligned}$$

so that

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}e^{-\frac{1}{2}\left[z^2\left(\frac{1}{\sigma^2}-2ia\right)-2z\left(\frac{\nu}{\sigma^2}+ib\right)\right]}\,\mathrm{d}z=\frac{e^{\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)}}}{\sqrt{1-2ia\sigma^2}}.$$

Finally, we get

$$E\left[e^{i(aX^2+bX)}\right]=e^{-\frac{\nu^2}{2\sigma^2}}\frac{e^{\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)}}}{\sqrt{1-2ia\sigma^2}}. \tag{85}$$

With simple computations we obtain

$$\begin{aligned}\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)} &= \frac{(\nu^2+2i\nu b\sigma^2-b^2\sigma^4)(1+2ia\sigma^2)}{2\sigma^2(1+4a^2\sigma^4)}\\ &= \frac{\nu^2+2i\nu b\sigma^2-b^2\sigma^4+2i\nu^2a\sigma^2-4\nu ab\sigma^2-2i\sigma^6ab^2}{2\sigma^2(1+4a^2\sigma^4)}\\ &= \frac{\nu^2+2i(\nu b\sigma^2+\nu^2a\sigma^2-\sigma^6ab^2)-4\nu ab\sigma^4-\sigma^4b^2}{2\sigma^2(1+4a^2\sigma^4)}.\end{aligned}$$

Thus, by (85) we can write

$$E\left[e^{i(aX^2+bX)}\right] = e^{-\frac{\nu^2}{2\sigma^2}}\frac{e^{\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)}}}{\sqrt{1-2ia\sigma^2}},$$

that implies

$$\left|E\left[e^{i(aX^2+bX)}\right]\right| \le \frac{e^{-\frac{4\nu^2a^2\sigma^4+4\nu ab\sigma^4+b^2\sigma^4}{2\sigma^2(1+4a^2\sigma^4)}}}{|\sqrt{1-2ia\sigma^2}|} = \frac{e^{-\frac{\sigma^2}{2}\frac{(2\nu a+b)^2}{1+4a^2\sigma^4}}}{(1+4a^2\sigma^4)^{1/4}},$$

as desired. □

Define

$$\psi=(\mu,\tau_1) \quad \text{and} \quad \boldsymbol{T}=\boldsymbol{T}(\boldsymbol{\theta})=\left(\sum_{j=1}^J\theta_j,\sum_{j=1}^J(\theta_j-\mu^*)^2\right). \tag{86}$$

Next three lemmas show that assumptions $(B1)-(B6)$ are satisfied for $(\boldsymbol{T},\psi)$ as defined above.

Lemma C.27. *Consider the setting of Proposition 5.1. Then assumptions* $(B1)-(B3)$ *are satisfied for* $(\boldsymbol{T},\psi)$ *as in* (86).

Proof. It is easy to show that assumption $(B1)$ is satisfied, with $g(\cdot)$ as in (23). As regards $(B2)$, suitable tests can be defined analogously to Lemma C.8.

Finally, by Lemma C.25, the Fisher Information is given by

$$\frac{m\tau_0^*\tau_1^*}{m\tau_0^*+\tau_1^*}$$

for $l=1$ and by

$$\begin{bmatrix}\frac{m\tau_0^*\tau_1^*}{m\tau_0^*+\tau_1^*} & 0\\ 0 & \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2(\tau_1^*+m\tau_0^*)^2}\end{bmatrix},$$

for $l=2,3$. Therefore $(B3)$ is satisfied for any ψ^*. □

Lemma C.28. *Consider the setting of Proposition 5.1. Then assumption* $(B4)$ *is satisfied for* $(\boldsymbol{T},\psi)$ *as in* (86).

Proof. Since $T(\theta_j)=(\theta_j,(\theta_j-\mu^*)^2)$ it holds

$$M_s^{(p)}(\mu,\tau_1\mid Y_j)=E\left[\theta_j^{sp}\mid\mu,\tau_1\right], \quad M_{1,2}^{(1)}(\mu,\tau_1\mid Y_j)=E\left[\theta_j(\theta_j^*-\mu^*)^2\mid\mu,\tau_1\right].$$

By Lemma C.23 and (82), we obtain

$$E\left[\theta_j^k\mid\mu,\tau_1\right]=\sum_{i=0}^k\binom{k}{i}\left(\frac{m\tau_0}{m\tau_0+\tau_1}\bar{Y}_j+\frac{\tau_1}{m\tau_0+\tau_1}\mu\right)^i\left(\frac{1}{m\tau_0+\tau_1}\right)^{(k-i)/2}E[Z^{k-i}].$$

It is a finite sum of infinitely times differentiable terms (with respect to μ and τ_1). Moreover, for every $k \geq 1$, thanks to Lemma C.23 and (83), $E_{Y_j}\left[|\bar{Y}_j|^k \mid \mu, \tau_1\right]$ is uniformly bounded over (μ, τ_1) belonging to a bounded set.

Therefore, choosing $\delta_4 < \tau_1^*$, it is easy to find $C < \infty$ that satisfies assumption $(B4)$. $\square$

Lemma C.29. *Consider the setting of Proposition 5.1. Then assumptions* $(B5)$ *and* $(B6)$ *are satisfied for* $(\boldsymbol{T}, \psi)$ *as in* (86).

Proof. Assume $\mu^* = 0$, the general case follows by similar calculations. Recall that the posterior distribution of θ_j is given by $N(m_j, \sigma^2)$, with m_j as in (82) and

$$\sigma^2 = \frac{1}{m\tau_0 + \tau_1}.$$

By Lemma C.26 we have

$$\left| E\left[e^{i(t_1\theta_j + t_2\theta_j^2)} \mid Y_j, \mu, \tau_1 \right] \right|^2 \leq \frac{e^{-\sigma^2 \frac{(2m_j t_2 + t_1)^2}{1+4t_2^2\sigma^4}}}{\left(1 + 4t_2^2\sigma^4\right)^{1/2}}. \tag{87}$$

Moreover, notice that

$$\int_{\mathbb{R}} e^{-c\sigma^2 \frac{(2m_j t_2 + t_1)^2}{1+4t_2^2\sigma^4}} \, \mathrm{d}t_1 = \sqrt{\frac{\pi}{c\sigma^2}} \sqrt{1 + 4t_2^2\sigma^4},$$

for any $c > 0$. Since θ_j are independent, given μ and τ_1, by Hölder inequality we write

$$\begin{aligned}
&\int_{\mathbb{R}_2} \left| E\left[e^{i(t_1 \sum_{j=1}^3 \theta_j + t_2 \sum_{j=1}^3 \theta_j^2)} \mid Y, \mu, \tau_1 \right] \right|^2 \mathrm{d}t_1 \mathrm{d}t_2 = \int_{\mathbb{R}_2} \prod_{j=1}^3 \left| E\left[e^{i(t_1\theta_j + t_2\theta_j^2)} \mid Y_j, \mu, \tau_1 \right] \right|^2 \mathrm{d}t_1 \mathrm{d}t_2 \\
&\leq \int_{\mathbb{R}_2} \prod_{j=1}^3 \frac{e^{-\sigma^2 \frac{(2m_j t_2 + t_1)^2}{1+4t_2^2\sigma^4}}}{\left(1 + 4t_2^2\sigma^4\right)^{1/2}} \mathrm{d}t_1 \mathrm{d}t_2 = \int_{\mathbb{R}} \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{3/2}} \left(\int_{\mathbb{R}} \prod_{j=1}^3 e^{-\sigma^2 \frac{(2\nu_j t_2 + t_1)^2}{1+4t_2^2\sigma^4}} \, \mathrm{d}t_1 \right) \mathrm{d}t_2 \\
&\leq \int_{\mathbb{R}} \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{3/2}} \prod_{j=1}^3 \left(\int_{\mathbb{R}} e^{-3\sigma^2 \frac{(2\nu_j t_2 + t_1)^2}{1+4t_2^2\sigma^4}} \, \mathrm{d}t_1 \right)^{1/3} \mathrm{d}t_2 \\
&= \sqrt{\frac{\pi}{3\sigma^2}} \int_{\mathbb{R}} \frac{1}{1 + 4t_2^2\sigma^4} \, \mathrm{d}t_2.
\end{aligned}$$

Therefore

$$\int_{\mathbb{R}^2} \left| \varphi^{(3)}\left(t \mid Y, \psi\right) \right|^2 \mathrm{d}t \leq \sqrt{\frac{\pi}{3\sigma^2}} \int_{\mathbb{R}} \frac{1}{1 + 4t_2^2\sigma^4} \, \mathrm{d}t_2 < \infty,$$

where the right hand side does not depend on the data and it is a continuous function of μ and τ_1. This implies $(B5)$ is satisfied with $k = 3$.

As regards $(B6)$, by Lemma C.26 if $t_2 \neq 0$ we have

$$|\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{1/4}},$$

while if $t_2 = 0$ then

$$|\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq e^{-\frac{\sigma^2}{2}t_1^2}.$$

Therefore

$$|\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq \max\left\{ \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{1/4}}, e^{-\frac{\sigma^2}{2}t_1^2} \right\},$$

so that

$$\sup_{|t|>\epsilon} |\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq \max\left\{ \frac{1}{(1+\epsilon^2\sigma^4)^{1/4}}, e^{-\frac{\sigma^2}{8}\epsilon^2} \right\},$$

since at least one between t_1 and t_2 must be larger than $\epsilon/2$. Notice that the right hand side does not depend on Y_j and is strictly smaller than 1 for every triplet (μ, τ_1, τ_0). Since σ^2 is a continuous function of μ and τ_1, assumption $(B6)$ is satisfied by choosing $\delta_6 < \tau_1^*$ and $k' = 1$.

□

Proof of Proposition 5.1. The result for P_1 follows directly by Theorem 4.2, whose assumptions are satisfied by Lemmas C.27, C.28 and C.29. As regards P_2 and P_3, they are not particular cases of Theorem 4.2, since the two operators are different by the one in (15). However, the result follows by very similar arguments, that we briefly summarize. Since by construction

$$\mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}\right)$$

a direct analogue of Lemma 4.1 holds. Moreover, following the proof of Theorem 4.2, Lemmas 4.3, 4.4 and C.18 hold for $\boldsymbol{T}$ in (86). Finally, Corollary 5.2 proves that the limiting spectral gaps associated to P_2 and P_3 are strictly positive: by Lemma C.6 this implies $\tilde{t}_{mix}(\epsilon, M) < \infty$ for P_2, being a two-block Gibbs sampler. The same holds for P_3, since in the limit it can be reduced to a two-block Gibbs sampler, as it will be clear by the proof of Corollary 5.2. □

C.19 Proof of Corollary 5.2

We split the proof in two different cases.

C.19.1 Proof of Corollary 5.2 for $\gamma_1(\psi^*)$

Proof. By Corollary 4.7, the spectral gap is equal to

$$\gamma_1(\psi) = \frac{\mathrm{Var}_{Y_j}\left(E\left[\theta_j \mid \psi, Y_j\right]\right)}{\mathrm{Var}\left(\theta_j \mid \psi\right)}.$$

By (82) and (83) we have

$$\mathrm{Var}_{Y_j}\left(E\left[\theta_j \mid \psi, Y_j\right]\right) = \left(\frac{m\tau_0}{m\tau_0+\tau_1}\right)^2, \quad \mathrm{Var}\left(\bar{Y}_j\right) = \frac{m\tau_0}{\tau_1(m\tau_0+\tau_1)},$$

and $\mathrm{Var}\left(\theta_j \mid \psi\right) = \tau_1^{-1}$, that leads to

$$\gamma_1(\psi^*) = \frac{m\tau_0^*}{m\tau_0^*+\tau_1^*},$$

as desired. □

C.19.2 Proof of Corollary 5.2 for $\gamma_2(\psi^*)$ and $\gamma_3(\psi^*)$

We need a technical Lemma.

Lemma C.30. *Consider the setting of Proposition 5.1. Then*

$$C(\psi^*) = \begin{bmatrix} \frac{\tau_1^*}{m\tau_0^*+\tau_1^*} & 0 \\ 0 & -\frac{\tau_1^*+2m\tau_0^*}{\tau_1^*(m\tau_0^*+\tau_1^*)^2} \end{bmatrix}, \quad V(\psi^*) = \begin{bmatrix} \frac{1}{m\tau_0^*+\tau_1^*} & 0 \\ 0 & \frac{2\tau_1^*+4m\tau_0^*}{\tau_1^*(m\tau_0^*+\tau_1^*)^2} \end{bmatrix},$$

with $C(\psi^*)$ and $V(\psi^*)$ as in (38).

Proof. Recall that, in the context of Proposition 5.1, we define $T_1(\theta_j) = \theta_j$ and $T_2(\theta_j) = (\theta_j - \mu^*)^2$. By (82) we have

$$E[T_1(\theta_j) \mid Y_j, \psi] = \frac{m\tau_0}{m\tau_0 + \tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0 + \tau_1}\mu,$$
$$E[T_2(\theta_j) \mid Y_j, \psi] = \frac{1}{m\tau_0 + \tau_1} + \left(\frac{m\tau_0}{m\tau_0 + \tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0 + \tau_1}\mu - \mu^*\right)^2.$$

Therefore we can compute $C(\psi^*)$ as

$$E_{Y_j}\left[\partial_\mu M_1(\psi^* \mid Y_j)\right] = \frac{\tau_1^*}{m\tau_0^* + \tau_1^*},$$
$$E_{Y_j}\left[\partial_\mu M_2(\psi^* \mid Y_j)\right] = E_{Y_j}\left[\frac{2\tau_1^*}{m\tau_0^* + \tau_1^*}\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\bar{Y}_j - \frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\mu^*\right)\right] = 0,$$
$$E_{Y_j}\left[\partial_{\tau_1} M_1(\psi^* \mid Y_j)\right] = E_{Y_j}\left[-\frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\bar{Y}_j + \frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\mu^*\right] = 0,$$
$$\begin{aligned} E_{Y_j}\left[\partial_{\tau_1} M_2(\psi^* \mid Y_j)\right] =& -\frac{1}{(m\tau_0^* + \tau_1^*)^2} + \\ & E_{Y_j}\left[2\left(-\frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\bar{Y}_j + \frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\mu^*\right)\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\bar{Y}_j - \frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\mu^*\right)\right] \\ =& -\frac{1}{(m\tau_0^* + \tau_1^*)^2} - 2\frac{(m\tau_0^*)^2}{(m\tau_0^* + \tau_1^*)^3}E_{Y_j}\left[(\bar{Y}_j - \mu^*)^2\right] \\ =& -\frac{1}{(m\tau_0^* + \tau_1^*)^2} - 2\frac{m\tau_0^*}{\tau_1^*(m\tau_0^* + \tau_1^*)^2}, \end{aligned}$$

by (83).

We now consider $V(\psi^*)$. Given $X \sim N(\mu, \sigma^2)$, we have

$$\mathrm{Cov}(X, X^2) = 2\mu\sigma^2, \quad \mathrm{Var}(X^2) = 2\sigma^4 + 4\mu^2\sigma^2,$$

which can be easily derived by computing the first four moments of X using Lemma C.23, which are $E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$, $E[X^3] = 3\mu\sigma^2 + \mu^3$ and $E[X^4] = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$. By (82) we have

$$\mathrm{Var}(\theta_j \mid Y_j, \psi^*) = \frac{1}{m\tau_0^* + \tau_1^*},$$
$$\mathrm{Cov}(\theta_j, (\theta_j - \mu^*)^2 \mid Y_j, \psi^*) = \mathrm{Cov}(\theta_j - \mu^*, (\theta_j - \mu^*)^2 \mid Y_j, \psi^*) = 2\frac{m_j - \mu^*}{m\tau_0^* + \tau_1^*},$$
$$\begin{aligned} \mathrm{Var}((\theta_j - \mu^*)^2 \mid Y_j, \psi^*) &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4}{m\tau_0^* + \tau_1^*}(m_j - \mu^*)^2 \\ &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4}{m\tau_0^* + \tau_1^*}\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}(\bar{Y}_j - \mu^*) + \mu^*\right)^2. \end{aligned}$$

Therefore, we conclude

$$E_{Y_j}\left[\mathrm{Cov}(\theta_j, \theta_j^2 \mid Y_j, \psi^*)\right] = 0$$

and

$$\begin{aligned} E_{Y_j}\left[\mathrm{Var}(\theta_j^2 \mid Y_j, \psi^*)\right] &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4}{m\tau_0^* + \tau_1^*}E_{Y_j}\left[\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\right)^2(\bar{Y}_j - \mu^*)^2\right] \\ &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4m^2(\tau_0^*)^2}{(m\tau_0^* + \tau_1^*)^3}E_{Y_j}\left[(\bar{Y}_j - \mu^*)^2\right] \\ &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4m\tau_0^*}{\tau_1^*(m\tau_0^* + \tau_1^*)^2}, \end{aligned}$$

as desired. □

Lemma C.31. *Consider the same assumptions of Proposition 5.1. Then*

$$\left\| \mathcal{L}(d\tilde{\boldsymbol{T}}, d\tilde{\psi} \mid Y_{1:J}) - N(\mathbf{0}, \Sigma) \right\|_{TV} \to 0,$$

as $J \to \infty$*, in* $Q_{\psi^*}^{(\infty)}$*-probability, where* $(\tilde{\boldsymbol{T}}, \tilde{\psi})$ *are derived by* (86) *with transformations* (17) *and* (19) *and where*

$$\Sigma = \begin{bmatrix} 2\frac{\tau_1^* + 2m\tau_0^*}{m^2(\tau_0^*)^2\tau_1^*} & 0 & -2\frac{\tau_1^*(\tau_1^* + 2m\tau_0^*)}{m^2(\tau_0^*)^2} & 0 \\ 0 & \frac{1}{m\tau_0^*} & 0 & \frac{1}{m\tau_0^*} \\ -2\frac{\tau_1^*(\tau_1^* + 2m\tau_0^*)}{m^2(\tau_0^*)^2} & 0 & 2\frac{(\tau_1^*)^2(\tau_1^* + m\tau_0^*)^2}{m^2(\tau_0^*)^2} & 0 \\ 0 & \frac{1}{m\tau_0^*} & 0 & \frac{m\tau_0^* + \tau_1^*}{m\tau_0^*\tau_1^*} \end{bmatrix} \tag{88}$$

Proof. The result follows by an argument similar to the proof of Proposition 4.5, where

$$\Sigma = \begin{bmatrix} V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & C(\psi^*)\mathcal{I}^{-1}(\psi^*) \\ \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & \mathcal{I}^{-1}(\psi^*) \end{bmatrix}$$

The entries of Σ can be computed through Lemmas C.25 and C.30. □

Proof of Corollary 5.2 for $\gamma_2(\psi^*)$ *and* $\gamma_3(\psi^*)$*.* Recall that P_2 is the transition kernel of the Gibbs sampler that alternates updates from $\mathcal{L}(\mathrm{d}\mu, \mathrm{d}\boldsymbol{\theta} \mid \tau_1, Y_{1:J})$ and $\mathcal{L}(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J})$. Through the same reasoning of Lemma 4.1, the mixing times of P_2 are the same of the Gibbs sampler targeting $\mathcal{L}(\mathrm{d}\mu, \mathrm{d}\tau_1, \mathrm{d}\boldsymbol{T} \mid Y_{1:J})$ by alternating updates from $\mathcal{L}(\mathrm{d}\mu, \mathrm{d}\boldsymbol{T} \mid \tau_1, Y_{1:J})$ and $\mathcal{L}(\mathrm{d}\tau_1 \mid \mu, \boldsymbol{T}, Y_{1:J})$. Indeed

$$\mathcal{L}(\mathrm{d}\tau_1 \mid \mu, \boldsymbol{\theta}, Y_{1:J}) = \mathcal{L}(\mathrm{d}\tau_1 \mid \mu, \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}).$$

Therefore, by Corollary 2.6 $\gamma_2(\psi^*)$ is the spectral gap of the Gibbs sampler alternating updates from $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\mu}, \mathrm{d}\tilde{\boldsymbol{T}}_1, \mathrm{d}\tilde{\boldsymbol{T}}_2 \mid \tilde{\tau}_1\right)$ and $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\tau}_1 \mid \tilde{\mu}, \tilde{\boldsymbol{T}}_1, \tilde{\boldsymbol{T}}_2\right)$, where $\tilde{\mathcal{L}}(\cdot)$ is the law identified in Lemma C.31. By inspection of the matrix (88), $\left(\tilde{\mu}, \tilde{\boldsymbol{T}}_1\right)$ is independent from $\tilde{\tau}_1$ and $\tilde{\boldsymbol{T}}_2$ according to $\tilde{\mathcal{L}}$, so that $\left(\tilde{\mu}, \tilde{\boldsymbol{T}}_1\right)$ is sampled independently from everything else at each iteration. Therefore by the same arguments of the proof of Corollary 4.6 we have

$$\gamma_2(\psi^*) = 1 - \frac{\Sigma_{24}^2}{\Sigma_{22}\Sigma_{44}} = \left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\right)^2.$$

Instead, recall that P_3 is the transition kernel of the Gibbs sampler that alternates updates from $\mathcal{L}(\mathrm{d}\boldsymbol{\theta} \mid \tau_1, Y_{1:J})$, $\mathcal{L}(\mathrm{d}\mu \mid \boldsymbol{\theta}, \tau_1, Y_{1:J})$ and $\mathcal{L}(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J})$. Reasoning as before, by Corollary 2.6 $\gamma_3(\psi^*)$ is the spectral gap of the Gibbs sampler alternating updates from $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\boldsymbol{T}} \mid \tilde{\mu}, \tilde{\tau}_1\right)$, $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\mu} \mid \tilde{\tau}_1, \tilde{\boldsymbol{T}}\right)$ and $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\tau}_1 \mid \tilde{\mu}, \tilde{\boldsymbol{T}}\right)$, where $\tilde{\mathcal{L}}(\cdot)$ is the law identified in Lemma C.31. By inspection of the matrix (88), the pair $(\tilde{\mu}, \tilde{\boldsymbol{T}}_1)$ is independent from $(\tilde{\tau}_1, \tilde{\boldsymbol{T}}_2)$, according to $\tilde{\mathcal{L}}$. By standard properties of the Gibbs samplers (e.g. Lemma 2 in [42]), the spectral gap is given by the minimum of the spectral gaps of the Gibbs samplers associated to the two pairs, i.e.

$$\gamma_3(\psi^*) = \min\left\{1 - \frac{\Sigma_{24}^2}{\Sigma_{22}\Sigma_{44}}, 1 - \frac{\Sigma_{13}^2}{\Sigma_{11}\Sigma_{33}}\right\} = \left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\right)^2.$$

Notice that the result of Lemma C.6 holds even if P_3 has three blocks: indeed, by inspection of the matrix (88), $\tilde{\mu}$ and $\tilde{\tau}_1$ are independent according to $\tilde{\mathcal{L}}$, so that the updates $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\mu} \mid \tilde{\tau}_1, \tilde{\boldsymbol{T}}\right)$ and $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\tau}_1 \mid \tilde{\mu}, \tilde{\boldsymbol{T}}\right)$ can be equivalently seen as a single one. □

C.20 Proof of Lemma 5.3

Since it will be useful in the following, we denote

$$c(\mu,\tau) = \min_{r\in\{0,\dots,m\}} g(y_r \mid \mu,\tau)\,,$$

with $g(y_r \mid \mu,\tau)$ defined in (28). Notice that by construction, see e.g. (26), we have $0 < c(\mu,\tau) \leq 1$. Also, $g(y_r \mid \mu,\tau)$ is continuous w.r.t. (μ,τ) since it is defined in (28) as the integral of a bounded function, $\theta \mapsto f(y \mid \theta)$, with respect to the normal kernel which is continuous w.r.t. (μ,τ). It follows that also $c(\mu,\tau)$ is continuous, since it is the minimum of a finite number of continuous functions. Define

$$c := \inf_{(\mu,\tau)\in B} c(\mu,\tau) > 0 \tag{89}$$

where B is the largest of the three balls – namely B_{δ_4}, B_{δ_5} and B_{δ_6} – centered at $\psi^* = (\mu^*,\tau^*)$ defined in (B4), (B5) and (B6), respectively. The positivity of c follows from the continuity of $c(\mu,\tau)$ and the compactness of B.

Recall that $T(\theta_j) = \left(\theta_j, \theta_j^2\right)$. Thus we need three lemmas.

Lemma C.32. *Consider the setting of Lemma 5.3. Then assumption (B4) is satisfied.*

Proof. First of all, consider $V(\psi^*)$, as defined in (38). For every $y = 0,\dots,m$, we have that the posterior distribution of θ_j admits a density with respect to the Lebesgue measure of the form

$$p(\theta_j \mid y,\mu,\tau) \propto f(y_r \mid \theta_j) N(\theta_j \mid \mu,\tau),$$

which implies that

$$\mathrm{Var}(\theta_j \mid y,\psi^*) > 0, \quad \mathrm{Var}(\theta_j^2 \mid y,\psi^*) > 0, \quad |\mathrm{Corr}(\theta_j,\theta_j^2 \mid y,\psi^*)| < 1.$$

Consequently $V(\psi^*)$ is a sum of positive definite matrices and is therefore non singular.

Secondly, let $s,p = 1,2$. Then by Bayes' Theorem it follows

$$M_s^{(p)}(y_r \mid \mu,\tau) = \frac{\int_{\mathbb{R}} \theta^{sp} f(y_r \mid \theta) N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta}{\int_{\mathbb{R}} f(y_r \mid \theta) N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta}, \quad r = 0,\dots,m.$$

Therefore

$$\begin{aligned}|\partial_\mu M_1^{(p)}(y_r \mid \mu,\tau)| \leq & \left|\frac{\int_{\mathbb{R}} \theta^p f(y_r \mid \theta)\partial_\mu N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta}{\int_{\mathbb{R}} f(y_r \mid \theta) N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta}\right| + \\ & \left|\frac{\left(\int_{\mathbb{R}} \theta^p f(y_r \mid \theta) N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta\right)\left(\int_{\mathbb{R}} f(y_r \mid \theta)\partial_\mu N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta\right)}{\left(\int_{\mathbb{R}} f(y_r \mid \theta) N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta\right)^2}\right|.\end{aligned}$$

By definition of c we have

$$\begin{aligned}|\partial_\mu M_1^{(p)}(y_r \mid \mu,\tau)| \leq & \frac{1}{c}\int_{\mathbb{R}} |\theta|^p \left|\partial_\mu N(\theta \mid \mu,\tau^{-1})\right|\,\mathrm{d}\theta + \\ & \frac{1}{c^2}\left(\int_{\mathbb{R}} |\theta|^p N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta\right)\left(\int_{\mathbb{R}} |\theta|^p \left|\partial_\mu N(\theta \mid \mu,\tau^{-1})\right|\,\mathrm{d}\theta\right) \\ = & \frac{\tau}{c}\int_{\mathbb{R}} |(\theta-\mu)\theta^p| N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta + \\ & \frac{\tau}{c^2}\left(\int_{\mathbb{R}} |(\theta-\mu)\theta|^p N(\theta \mid \mu,\tau^{-1})\,\mathrm{d}\theta\right)\left(\int_{\mathbb{R}} |\theta|^p f\left|N(\theta \mid \mu,\tau^{-1})\right|\,\mathrm{d}\theta\right).\end{aligned}$$

The right hand side does not depend on the data, so that

$$E_{Y_j}\left[|\partial_\mu M_1^{(p)}(y_r \mid \mu,\tau)|\right] \leq m\frac{\tau}{c}E[|(\theta_j-\mu)\theta_j^p| \mid \mu,\tau] + m\frac{\tau}{c^2}E[|(\theta_j-\mu)\theta_j^p| \mid \mu,\tau]E[|\theta_j|^p \mid \mu,\tau].$$

By the specification of model (27), the prior absolute moments are all finite and continuous function of μ and τ: therefore the right hand side is uniformly bounded for every bounded neighborhood of (μ^*, τ^*). Using a similar argument for all the other quantities involved, it is easy to see that assumption $(B4)$ holds for every $\delta_4 < \tau^*$. □

Lemma C.33. *Consider the setting of Lemma 5.3. Then assumption $(B5)$ is satisfied with $k = 5$.*

Proof. Consider the random vector $X = (X_1, X_2) = (\sum_{j=1}^5 \theta_j, \sum_{j=1}^5 \theta_j^2)$. First of all we prove that X admits a density function with respect to the Lebesgue measure on $\mathbb{R}^2$, conditional to (μ, τ). By Lemma C.26 and conditional independence of θ_j we have

$$\left| E\left[e^{i(t_1 X_1 + t_2 X_2)} \mid \mu, \tau_1 \right] \right| \leq \frac{e^{-5\frac{\sigma^2}{2}\frac{(2\mu t_2 + t_1)^2}{1+4t_2^2\sigma^4}}}{(1+4t_2^2\sigma^4)^{5/4}},$$

where we denote $\sigma^2 = \tau^{-1}$, so that we can write

$$\begin{aligned} \int_{\mathbb{R}^2} |\varphi_X(t \mid \mu, \tau)| \, \mathrm{d}t &= \int_{\mathbb{R}^2} \left| E\left[e^{i(t_1 X_1 + t_2 \sum_{j=1}^3 X_2)} \mid Y, \mu, \tau_1 \right] \right| \, \mathrm{d}t_1 \mathrm{d}t_2 \\ &\leq \int_{\mathbb{R}} \frac{1}{(1+4t_2^2\sigma^4)^{5/4}} \left(\int_{\mathbb{R}} e^{-5\frac{\sigma^2}{2}\frac{(2\mu t_2 + t_1)^2}{1+4t_2^2\sigma^4}} \, \mathrm{d}t_1 \right) \mathrm{d}t_2 \\ &= \sqrt{\frac{2\pi}{5\sigma^2}} \int_{\mathbb{R}} \frac{1}{(1+4t_2^2\sigma^4)^{3/4}} \, \mathrm{d}t_2 < \infty. \end{aligned} \tag{90}$$

Therefore, by the Inversion Formula we have that X admits a density $p(x \mid \mu, \tau)$ with respect to the Lebesgue measure on $\mathbb{R}^2$. Thus, by Bayes' Theorem we can write

$$p(x \mid Y_{1:5}, \mu, \tau) = \frac{f(Y_{1:5} \mid x, \mu, \tau) p(x \mid \mu, \tau)}{\int_{\mathbb{R}^2} f(Y_{1:5} \mid x, \mu, \tau) p(x \mid \mu, \tau) \, \mathrm{d}x},$$

where $f(Y_{1:5} \mid x, \mu, \tau) = \int \prod_{j=1}^5 f(Y_j \mid \theta_j) \mathcal{L}(\mathrm{d}\theta_{1:5} \mid x, \mu, \tau)$. It is easy to see that $f(Y_{1:5} \mid x, \mu, \tau) \leq 1$ and

$$\int_{\mathbb{R}^2} f(Y_{1:5} \mid x, \mu, \tau) p(x \mid \mu, \tau) \, \mathrm{d}x = \prod_{j=1}^5 g(Y_j \mid \mu, \tau) \geq c^5,$$

for every $(\mu, \tau) \in B_{\delta_5}$, with δ_5 to be fixed. We can therefore conclude that

$$p(x \mid Y_{1:5}, \mu, \tau) \leq \frac{p(x \mid \mu, \tau)}{c^5}.$$

We can now apply the Plancherel identity to get

$$\int_{\mathbb{R}^2} \left| \varphi^{(5)}(t \mid Y, \mu, \tau) \right|^2 \mathrm{d}t = \int_{\mathbb{R}^2} p^2(x_1, x_2 \mid Y, \mu, \tau) \, \mathrm{d}x \leq \frac{1}{c^{10}} \int_{\mathbb{R}^2} p^2(x_1, x_2 \mid \mu, \tau) \, \mathrm{d}x.$$

Applying again the Plancherel identity we obtain

$$\int_{\mathbb{R}^2} \left| \varphi^{(5)}(t \mid Y, \mu, \tau) \right|^2 \mathrm{d}t \leq \frac{1}{c^{10}} \int_{\mathbb{R}^2} |\varphi_X(t \mid \mu, \tau)|^2 \, \mathrm{d}t \leq \frac{1}{c^{10}} \int_{\mathbb{R}^2} |\varphi_X(t \mid \mu, \tau)| \, \mathrm{d}t < \infty,$$

by (90) for every $\tau > 0$. Therefore assumption $(B5)$ follows with $\delta_5 < \tau^*$. □

Lemma C.34. *Consider the setting of Lemma 5.3. Then assumption $(B6)$ is satisfied with $k' = 5$.*

Proof. As shown in the proof of Lemma C.33, the vector $(\sum_{j=1}^5 \theta_j, \sum_{j=1}^5 \theta_j^2)$ admits a density with respect to the Lebesgue measure on $\mathbb{R}^2$, conditional to Y and (μ^*, τ^*). Therefore, by Lemma 4 in Chapter 15 of [20], $|\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)| < 1$ for every $t = (t_1, t_2)$. Moreover, by Riemann-Lebesgue Lemma we have

$$|\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)| \quad \to \quad 0,$$

as $|t| \to \infty$. We conclude

$$\sup_{|t| \geq \epsilon} \left| \varphi^{(5)}(t \mid Y, \mu^*, \tau^*) \right| < 1.$$

Let $\delta_6 > 0$ to be chosen later and $(\mu, \tau) \in B_{\delta_6}$. Then by Taylor formula we get

$$|\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 = |\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)|^2 + (\mu^* - \mu)\partial_\mu |\varphi^{(5)}(t \mid Y, \bar{\mu}, \bar{\tau})|^2 + (\tau^* - \tau)\partial_\tau |\varphi^{(5)}(t \mid Y, \bar{\mu}, \bar{\tau})|^2, \tag{91}$$

where $(\bar{\mu}, \bar{\tau}) \in B_{\delta_6}$. Notice that

$$\begin{aligned}
|\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 = &\left(\int_{\mathbb{R}^3} \cos\left(t_1 \sum_{j=1}^5 \theta_j + t_2 \sum_{j=1}^5 \theta_j^2 \right) \left\{ \prod_{j=1}^5 \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right)^2 \\
&+ \left(\int_{\mathbb{R}^5} \sin\left(t_1 \sum_{j=1}^5 \theta_j + t_2 \sum_{j=1}^5 \theta_j^2 \right) \left\{ \prod_{j=1}^5 \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right)^2,
\end{aligned}$$

which implies

$$\begin{aligned}
\left| \partial_\mu |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \right| \leq &2 \left| \int_{\mathbb{R}^5} \cos\left(t_1 \sum_{j=1}^5 \theta_j + t_2 \sum_{j=1}^5 \theta_j^2 \right) \partial_\mu \left\{ \prod_{j=1}^5 \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right| \\
&+ 2 \left| \int_{\mathbb{R}^5} \sin\left(t_1 \sum_{j=1}^5 \theta_j + t_2 \sum_{j=1}^5 \theta_j^2 \right) \partial_\mu \left\{ \prod_{j=1}^5 \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right|
\end{aligned}$$

and therefore

$$\begin{aligned}
\left| \partial_\mu |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \right| &\leq 4 \int_{\mathbb{R}^5} \left| \partial_\mu \left\{ \prod_{j=1}^5 \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \right| \mathrm{d}\theta_{1:5} \\
&= 4 \sum_{j=1}^5 \int_{\mathbb{R}} \left| \partial_\mu \left\{ \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \right| \mathrm{d}\theta_j.
\end{aligned} \tag{92}$$

Moreover, for every $r = 0, \ldots, m$, we have

$$\begin{aligned}
&\left| \partial_\mu \left\{ \frac{f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(y_r \mid \psi) N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi} \right\} \right| \leq \left| \left\{ \frac{f(y_r \mid \theta) \partial_\mu N(\theta \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(y_r \mid \psi) N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi} \right\} \right| \\
&+ \left| \left\{ \frac{f(y_r \mid \theta) \partial_\mu N(\theta \mid \mu, \tau^{-1}) \left(\int_{\mathbb{R}} f(y_r \mid \psi) \partial_\mu N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi \right)}{\left(\int_{\mathbb{R}} f(y_r \mid \psi) N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi \right)^2} \right\} \right| \\
&\leq \frac{|\partial_\mu N(\theta \mid \mu, \tau^{-1})|}{c} + \frac{1}{c^2} |\partial_\mu N(\theta \mid \mu, \tau^{-1})| \left(\int_{\mathbb{R}} |\partial_\mu N(\psi \mid \mu, \tau^{-1})| \mathrm{d}\psi \right) \\
&= 2\tau \frac{|\theta - \mu| N(\theta \mid \mu, \tau)}{c} + \frac{4\tau^2}{c^2} |\theta - \mu| N(\theta \mid \mu, \tau^{-1}) \left(\int_{\mathbb{R}} |\psi - \mu| N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi \right).
\end{aligned}$$

Therefore, by (92) there exists $C(\delta_6) < \infty$ which does not depend on μ and τ such that

$$\begin{aligned}
\left| \partial_\mu |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \right| &\leq 40\tau \frac{\int_{\mathbb{R}} |\theta - \mu| N(\theta \mid \mu, \tau^{-1}) \, \mathrm{d}\theta}{c} + 80\tau^2 \left(\frac{\int_{\mathbb{R}} |\theta - \mu| N(\theta \mid \mu, \tau^{-1}) \, \mathrm{d}\theta}{c} \right)^2 \\
&\leq C(\delta_6),
\end{aligned}$$

for every $(\mu,\tau) \in B_{\delta_6}$ Notice that $C(\delta_6)$ becomes smaller as δ_6 decreases. Similarly holds for $\partial_\tau |\varphi^{(3)}(t \mid Y,\mu,\tau)|^2$, so that by (91) we have

$$\begin{aligned}|\varphi^{(5)}(t \mid Y,\mu,\tau)|^2 &\leq |\varphi^{(5)}(t \mid Y,\mu^*,\tau^*)|^2 + |\mu^* - \mu| C(\delta_6) + |\tau^* - \tau| C(\delta_6) \\ &\leq |\varphi^{(5)}(t \mid Y\mu^*,\tau^*)|^2 + 2\delta_6 C(\delta_6).\end{aligned}$$

Since $\sup_{|t|\geq\epsilon} |\varphi^{(5)}(t \mid Y,\mu^*,\tau^*)|^2 < 1$, by choosing δ_6 small enough we have

$$\sup_{(\mu,\tau)\in B_{\delta_6}} \sup_{|t|\geq\epsilon} |\varphi^{(5)}(t \mid Y,\mu,\tau)|^2 \leq \sup_{|t|\geq\epsilon} |\varphi^{(5)}(t \mid Y,\mu^*,\tau^*)|^2 + 2\delta_6 C(\delta_6) < 1,$$

and $(B6)$ is satisfied. □

Proof of Lemma 5.3. Assumption (B4) is satisfied by Lemma C.32, assumption (B5) by Lemma C.33 and assumption (B6) by Lemma C.34. □

C.21 Proof of Proposition 5.4

Proof. Requirements $(B1)-(B3)$ of Theorem 4.2 are satisfied by assumption, while $(B4)-(B6)$ hold by Lemma 5.3. □

C.22 Proof of Corollary 5.5

Proof. The result is a direct consequence of Corollary 4.6. □

C.23 Statement and proof of Lemma C.35

Let

$$f(y \mid \theta) = \binom{m}{y} \frac{e^{y\theta}}{(1+e^\theta)^m}, \tag{93}$$

where $y = 0,\dots,m$. It means that for each group, conditional to θ, m independent Bernoulli trials are performed, with probability of success given by $e^\theta/(1+e^\theta)$. The following Section is devoted to the proof of the following lemma.

Lemma C.35. *Consider the setting of Proposition 5.4 with likelihood* (93)*. The Fisher Information Matrix $I(\mu,\tau)$ is non-singular if and only if $m \geq 2$, for every (μ,τ).*

First of all we need few preliminary results.

Lemma C.36. *Consider the setting of Proposition 5.4 with likelihood* (93) *and fix (μ,τ). Let $h(y \mid \mu,\tau) = \log g(y \mid \mu,\tau)$, with $g(\cdot)$ as in* (28)*. Then it holds*

$$E_Y\left[\frac{\partial}{\partial\mu} h(Y \mid \mu,\tau)\right] = E_Y\left[\frac{\partial}{\partial\tau} h(Y \mid \mu,\tau)\right] = 0$$

and

$$E_Y\left[\left(\frac{\partial}{\partial\mu} h(Y \mid \mu,\tau)\right)^2\right] < \infty, \quad E_Y\left[\left(\frac{\partial}{\partial\tau_1} h(Y \mid \mu,\tau)\right)^2\right] < \infty.$$

Moreover, for every $y = 0,\dots,m$ we have

$$\frac{\partial}{\partial\mu} g(y \mid,\mu,\tau) = \binom{m}{y} \int \frac{e^{y\theta}\left[y + ye^\theta - me^\theta\right]}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, d\theta$$

and

$$\frac{\partial}{\partial\tau} g(y \mid,\mu,\tau) = -\binom{m}{y} \frac{1}{2\tau} \int (\theta-\mu) \frac{e^{y\theta}\left[y + ye^\theta - me^\theta\right]}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, d\theta.$$

Proof. Through Dominated Convergence Theorem it is easy to verify that

$$\frac{\partial}{\partial\mu}g(y\mid\mu,\tau)=\binom{m}{y}\int\frac{e^{y\theta}}{(1+e^{\theta})^{m}}\frac{\partial}{\partial\mu}\left\{\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\right\}\mathrm{d}\theta$$

and

$$\frac{\partial}{\partial\tau}g(y\mid\mu,\tau)=\binom{m}{y}\int\frac{e^{y\theta}}{(1+e^{\theta})^{m}}\frac{\partial}{\partial\tau}\left\{\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\right\}\mathrm{d}\theta,$$

that is integrals and derivatives can be exchanged. Therefore

$$\frac{\partial}{\partial\mu}h(y\mid,\mu,\tau)=E\left[\theta-\mu\mid y,\mu,\tau\right],\quad\frac{\partial}{\partial\mu}h(y\mid,\mu,\tau)=\frac{1}{2\tau}-\frac{1}{2}E\left[(\theta-\mu)^{2}\mid y,\mu,\tau\right]$$

and the statements on $h(y\mid\mu,\tau)$ easily follow. Moreover

$$\begin{aligned}\frac{\partial}{\partial\mu}g(y\mid\mu,\tau)&=\binom{m}{y}\int\frac{e^{y\theta}}{(1+e^{\theta})^{m}}(\theta-\mu)\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta\\&=\binom{m}{y}\int\frac{e^{y\theta}\left[y+ye^{\theta}-me^{\theta}\right]}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta\end{aligned}$$

integrating by parts. Similarly

$$\begin{aligned}\frac{\partial}{\partial\tau}g(y\mid\mu,\tau)&=\binom{m}{y}\frac{1}{2\tau}\int\frac{e^{y\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta\\&\quad-\binom{m}{y}\frac{1}{2}\int\frac{e^{y\theta}}{(1+e^{\theta})^{m}}(\theta-\mu)^{2}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta\\&=-\binom{m}{y}\frac{1}{2\tau}\int(\theta-\mu)\frac{e^{y\theta}\left[y+ye^{\theta}-me^{\theta}\right]}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta.\end{aligned}$$

□

Lemma C.37. *Consider the setting of Proposition 5.4 with likelihood* (93) *and let* $y,y'\in\{0,1,\dots,m\}$ *be such that* $y<y'$ *and* $m\geq1$. *Then*

$$E\left[\theta\mid y,\mu,\tau\right]<E\left[\theta\mid y',\mu,\tau\right]$$

for every (μ,τ_1).

Proof. Fix (μ,τ). Consider the function

$$r(x)=\frac{\int\theta\frac{e^{x\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau_1}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta}{\int\frac{e^{x\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta}.$$

with $x\in(0,m)$. Notice that

$$r(y)=E\left[\theta\mid y,\mu,\tau\right]\quad\text{and}\quad r(y')=E\left[\theta\mid y',\mu,\tau\right].$$

Notice that

$$\frac{\mathrm{d}}{\mathrm{d}x}r(x)=\frac{\int\theta^{2}\frac{e^{x\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta}{\int\frac{e^{x\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta}-\left[\frac{\int\theta\frac{e^{x\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta}{\int\frac{e^{x\theta}}{(1+e^{\theta})^{m}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^{2}}\,\mathrm{d}\theta}\right]^{2}>0$$

for every $x\in(0,m)$ by Jensen inequality. Therefore $r(x)$ is strictly increasing and $r(y)<r(y')$. □

Lemma C.38. *Consider the setting of Proposition 5.4 with likelihood* (93)*. Then the Fisher Information Matrix $I(\mu,\tau)$ is non-singular in (μ,τ) if and only if there exists $\alpha = \alpha(\mu,\tau) \neq 0$ such that*

$$\frac{\partial}{\partial\mu} g(y \mid \mu,\tau) = \alpha \frac{\partial}{\partial\tau} g(y \mid \mu,\tau)$$

for every $y = 0,\dots,m$.

Proof. Fix a pair (μ,τ). By Lemma C.36 the matrix $I(\mu,\tau)$ is well-defined. The determinant is given by

$$E_Y\left[\left(\frac{\partial}{\partial\mu} h(Y \mid \mu,\tau)\right)^2\right] E_Y\left[\left(\frac{\partial}{\partial\tau} h(Y \mid \mu,\tau)\right)^2\right] - E^2\left[\left(\frac{\partial}{\partial\mu} h(Y \mid \mu,\tau)\right)\left(\frac{\partial}{\partial\tau} h(Y \mid \mu,\tau)\right)\right].$$

By Cauchy–Schwartz inequality, the above formula is always non-negative and it is equal to 0 if and only if $\frac{\partial}{\partial\mu} h(Y \mid \mu,\tau)$ and $\frac{\partial}{\partial\tau} h(Y \mid \mu,\tau)$ are linearly dependent, that is

$$\frac{\partial}{\partial\mu} h(y \mid \mu,\tau) = \alpha \frac{\partial}{\partial\tau} h(y \mid \mu,\tau) + \beta \tag{94}$$

for every $y \in \{0,1,\dots,m\}$ and for constants α and β. By Lemma C.36 it is immediate to prove $\beta = 0$. Moreover, by Lemma C.37, we deduce that $\alpha \neq 0$. Multiplying by $g(y \mid \mu,\tau)$ on both sides of (94) we get the final result. □

Proof of Lemma C.35. Fix (μ,τ) and let $m = 1$. Define

$$\alpha := \frac{\frac{\partial}{\partial\mu} g(0 \mid \mu,\tau)}{\frac{\partial}{\partial\tau} g(0 \mid \mu,\tau)}.$$

Notice that α is well defined, since $\frac{\partial}{\partial\tau} g(0 \mid \mu,\tau) \neq 0$ for every (μ,τ). Then by construction

$$\frac{\partial}{\partial\mu} g(0 \mid \mu,\tau) = \alpha \frac{\partial}{\partial\tau} g(0 \mid \mu,\tau)$$

and

$$\frac{\partial}{\partial\mu} g(1 \mid \mu,\tau) = -\frac{\partial}{\partial\mu} g(0 \mid \mu,\tau) = -\alpha \frac{\partial}{\partial\tau} g(0 \mid \mu,\tau) = \alpha \frac{\partial}{\partial\tau} g(1 \mid \mu,\tau),$$

so that the Fisher Information matrix is singular by Lemma C.38.

Let $m \geq 2$ and fix (μ,τ). Assume by contradiction that $I(\mu,\tau)$ is singular. By Lemma C.38 we have that there exists $\alpha \neq 0$ such that

$$\frac{\partial}{\partial\mu} g(y \mid \mu,\tau) = \alpha \frac{\partial}{\partial\tau} g(y \mid \mu,\tau)$$

for every $y \in \{0,1,\dots,m\}$. By the second part of Lemma C.36 for $y = 0$ and $y = m$ it implies

$$-m \int \frac{e^\theta}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau_1}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta = \alpha \frac{m}{2\tau} \int (\theta-\mu) \frac{e^\theta}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta$$

and

$$m \int \frac{e^{m\theta}}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta = -\alpha \frac{m}{2\tau} \int (\theta-\mu) \frac{e^{m\theta}}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta.$$

Since $\alpha \neq 0$, we conclude

$$\frac{\int (\theta-\mu) \frac{e^{m\theta}}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta}{\int \frac{e^{m\theta}}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta} = \frac{\int (\theta-\mu) \frac{e^{\theta}}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta}{\int \frac{e^{\theta}}{(1+e^\theta)^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta},$$

that means

$$E[\theta \mid m, \mu, \tau] = E[\theta \mid 1, \mu, \tau].$$

Since $m > 1$, the above equality directly contradicts Lemma C.37. Therefore the Fisher Information matrix is non singular. □

C.24 Proof of Proposition 5.6

Define a one-to-one transformation of $\psi = (\mu, \tau_1, \tau_0)$ as

$$\tilde{\psi} = \sqrt{J}\,(\psi - \psi^*) - \Delta_J, \quad \Delta_J = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \mathcal{I}^{-1}(\psi^*) \nabla \log g(Y_j \mid \psi^*), \tag{95}$$

with $g(\cdot)$ as in (23) and $\mathcal{I}(\psi^*)$ as in (84).

Lemma C.39. *Consider the assumptions of Proposition 5.6. Then it holds*

$$\left|\left| \mathcal{L}(d\tilde{\psi} \mid Y_{1:J}) - N\left(\mathbf{0}, \mathcal{I}^{-1}(\psi^*)\right) \right|\right|_{TV} \to 0,$$

as $J \to \infty$ in $Q_{\psi^*}^{(\infty)}$-probability, with $\mathcal{I}(\psi^*)$ non singular matrix as in (84).

Proof. The result follows by Theorem 3.1. Indeed, the map $\psi \to g(y \mid \psi)$ clearly satisfies identifiability and smoothness requirements. Moreover, by Lemma C.25 we have

$$\det\left(\mathcal{I}(\psi^*)\right) = \frac{m^3(m-1)\tau_0^*}{4\tau_1^*(\tau_1^* + m\tau_0^*)^3},$$

that is strictly positive for every ψ^*, with $m \geq 2$. As regards the testing conditions, analogously to Lemma C.8 define

$$\Psi = \Psi_1 \times \Psi_2 \times \Psi_3 = [\mu^* - 1, \mu^* + 1] \times \left[\frac{\tau_1^*}{2}, 2\tau_1^*\right] \times \left[\frac{\tau_0^*}{2}, 2\tau_0^*\right]$$

compact neighborhood of ψ^* and

$$u_J(Y_{1:J}) = 1 - \mathbb{1}_{g_1(Y_{1:J}) \leq c_1}\, \mathbb{1}_{g_2(Y_{1:J}) \leq c_2}\, \mathbb{1}_{g_3(Y_{1:J}) \leq c_3},$$

where (c_1, c_2, c_3) are positive constants to be fixed and

$$g_1(Y_{1:J}) = \left|\bar{Y} - \mu^*\right|, \quad g_2(Y_{1:J}) = \left| \frac{1}{J} \sum_{j=1}^{J} \left(\bar{Y}_j - \bar{Y}\right)^2 - \frac{1}{\tau_1^*} - \frac{1}{m\tau_0^*} \right|,$$

$$g_3(Y_{1:J}) = \left| \frac{1}{J} \sum_{j=1}^{J} \left(Y_{j,1} - \hat{Y}_1\right)\left(Y_{j,2} - \hat{Y}_2\right) - \frac{1}{\tau_1^*} \right|,$$

where

$$\bar{Y} = \frac{1}{J} \sum_{j=1}^{J} \bar{Y}_j, \quad \hat{Y}_i = \frac{1}{J} \sum_{j=1}^{J} Y_{j,i}.$$

By definition of $g(\cdot)$ in (23), by the Law of Large numbers we have

$$\int u_J(y_{1:J}) \prod_{j=1}^{J} g(\mathrm{d}y_j \mid \psi^*)$$
$$\leq P\left(g_1(Y_{1:J}) > c_1\right) + P\left(g_2(Y_{1:J}) > c_2\right) + P\left(g_3(Y_{1:J}) > c_3\right) \to 0,$$

as $J \to \infty$ for every strictly positive constants (c_1, c_2, c_3). Moreover, notice that

$$\sup_{\psi \notin \Psi} \int [1 - u_J(y_{1:J})] \prod_{j=1}^{J} g(\mathrm{d}y_j \mid \psi)$$
$$\leq \sup_{\tau_1 \notin \Psi_2} P\left(g_3(Y_{1:J}) \leq c_3\right) + \sup_{\tau_1 \in \Psi_2,\, \tau_0 \notin \Psi_3} P\left(g_2(Y_{1:J}) \leq c_2\right) + \sup_{\mu \notin \Psi_1,\, \tau_0 \in \Psi_3,\, \tau_1 \in \Psi_2} P\left(g_1(Y_{1:J}) > c_1\right).$$

With the same reasoning of the proof of Lemma C.8, we can find (c_1, c_2, c_3) such that the three suprema goes to 0 as $J \to \infty$. □

We need another technical Lemma.

Lemma C.40. *Consider the setting of Proposition 5.6. Then we have*

$$E\left[(\theta_j - \mu)^2 \mid Y, \psi\right] = \frac{1}{m\tau_0 + \tau_1} + \left(\frac{m\tau_0}{m\tau_0 + \tau_1}\right)^2 (\bar{Y}_j - \mu)^2,$$
$$E\left[(\theta_j - \bar{Y}_j)^2 \mid Y, \psi\right] = \frac{1}{m\tau_0 + \tau_1} + \left(\frac{\tau_1}{m\tau_0 + \tau_1}\right)^2 (\bar{Y}_j - \mu)^2$$

and

$$Var\left((\theta_j - \mu)^2 \mid Y, \psi\right) = \frac{2}{(m\tau_0 + \tau_1)^2} + 4\frac{m^2\tau_0^2}{(m\tau_0 + \tau_1)^3}(\bar{Y}_j - \mu)^2,$$
$$Var\left((\theta_j - \bar{Y}_j)^2 \mid Y, \psi\right) = \frac{2}{(m\tau_0 + \tau_1)^2} + 4\frac{\tau_1^2}{(m\tau_0 + \tau_1)^3}(\bar{Y}_j - \mu)^2$$

and

$$Cov\left((\theta_j - \mu)^2, (\theta_j - \bar{Y}_j)^2 \mid Y, \psi\right) = \frac{2}{(m\tau_0 + \tau_1)^2} - 4\frac{m\tau_0\tau_1}{(m\tau_0 + \tau_1)^3}(\bar{Y}_j - \mu)^2.$$

Proof. Notice that by (82) we have

$$(\theta_j - \mu) \mid Y_j, \psi \sim N\left(\frac{m\tau_0}{m\tau_0 + \tau_1}(\bar{Y}_j - \mu), (m\tau_0 + \tau_1)^{-1}\right)$$

and

$$(\theta_j - \bar{Y}_j) \mid Y_j, \psi \sim N\left(\frac{\tau_1}{m\tau_0 + \tau_1}(\mu - \bar{Y}_j), (m\tau_0 + \tau_1)^{-1}\right).$$

Therefore we have

$$E\left[(\theta_j - \mu)^2 \mid Y, \psi\right] = \frac{1}{m\tau_0 + \tau_1} + \left(\frac{m\tau_0}{m\tau_0 + \tau_1}\right)^2 (\bar{Y}_j - \mu)^2,$$

and similarly for the other case. If $X \sim N(\mu, \sigma^2)$, by Lemma C.23 we have $E[X^4] = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$. In our case, considering $\sigma = (m\tau_0 + \tau_1)^{-1/2}$ and $\mu = \frac{m\tau_0}{m\tau_0 + \tau_1}(\bar{Y}_j - \mu)$, we have

$$E\left[(\theta_j - \mu)^4 \mid Y, \psi\right] = \frac{3}{(m\tau_0 + \tau_1)^2} + 6\frac{m^2\tau_0^2}{(m\tau_0 + \tau_1)^3}(\bar{Y}_j - \mu)^2 + \left(\frac{m\tau_0}{m\tau_0 + \tau_1}\right)^4 (\bar{Y}_j - \mu)^4$$

and

$$E^2\left[(\theta_j - \mu)^2 \mid Y, \psi\right] = \frac{1}{(m\tau_0 + \tau_1)^2} + 2\frac{m^2\tau_0^2}{(m\tau_0 + \tau_1)^3}(\bar{Y}_j - \mu)^2 + \left(\frac{m\tau_0}{m\tau_0 + \tau_1}\right)^4 (\bar{Y}_j - \mu)^4.$$

Therefore

$$\mathrm{Var}\left((\theta_j - \mu)^2 \mid Y, \psi\right) = \frac{2}{(m\tau_0 + \tau_1)^2} + 4\frac{m^2\tau_0^2}{(m\tau_0 + \tau_1)^3}(\bar{Y}_j - \mu)^2,$$

and similarly for the other one. Finally, again by Lemma C.23, if $Z \sim N(0,1)$ we have $E[(\sigma Z + \mu_1)^2 (\sigma Z + \mu_2)^2] = 3\sigma^4 + \sigma^2(\mu_1^2 + 4\mu_1\mu_2 + \mu_2^2) + \mu_1^2\mu_2^2$. In our case, considering $\sigma = (m\tau_0 + \tau_1)^{-1/2}$, $\mu_1 = \frac{m\tau_0}{m\tau_0+\tau_1}(\bar{Y}_j - \mu)$ and $\mu_2 = \frac{\tau_1}{m\tau_0+\tau_1}(\mu - \bar{Y}_j)$, we have

$$\begin{aligned} E\left[(\theta_j - \mu)^2(\theta_j - \bar{Y}_j)^2 \mid Y, \psi\right] =& \frac{3}{(m\tau_0+\tau_1)^2} + \frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j - \mu)^2 + \frac{\tau_1^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j - \mu)^2 \\ & - 4\frac{m\tau_0\tau_1}{(m\tau_0+\tau_1)^3}(\bar{Y}_j - \mu)^2 + \frac{m^2\tau_0^2\tau_1^2}{(m\tau_0+\tau_1)^4}(\bar{Y}_j - \mu)^4 \end{aligned}$$

and

$$\begin{aligned} & E\left[(\theta_j - \mu)^2 \mid Y, \psi\right] E\left[(\theta_j - \bar{Y}_j)^2 \mid Y, \psi\right] = \\ & \frac{1}{(m\tau_0+\tau_1)^2} + \frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j - \mu)^2 + \frac{\tau_1^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j - \mu)^2 + \frac{m^2\tau_0^2\tau_1^2}{(m\tau_0+\tau_1)^4}(\bar{Y}_j - \mu)^4. \end{aligned}$$

Therefore

$$\mathrm{Cov}\left((\theta_j - \mu)^2, (\theta_j - \bar{Y}_j)^2 \mid Y, \psi\right) = \frac{2}{(m\tau_0+\tau_1)^2} - 4\frac{m\tau_0\tau_1}{(m\tau_0+\tau_1)^3}(\bar{Y}_j - \mu)^2,$$

as desired. □

Define

$$C(\psi) = \begin{bmatrix} 0 & \frac{1}{(m\tau_0+\tau_1)^2} & \frac{m}{(m\tau_0+\tau_1)^2} \\ 0 & \frac{1}{(m\tau_0+\tau_1)^2} & \frac{m}{(m\tau_0+\tau_1)^2} \end{bmatrix}, \quad V(\psi) = \begin{bmatrix} \frac{2}{(m\tau_0+\tau_1)^2} + 4\frac{m\tau_0(\tau_1)^{-1}}{(m\tau_0+\tau_1)^2} & -\frac{2}{(m\tau_0+\tau_1)^2} \\ -\frac{2}{(m\tau_0+\tau_1)^2} & \frac{2}{(m\tau_0+\tau_1)^2} + 4\frac{\tau_1(m\tau_0)^{-1}}{(m\tau_0+\tau_1)^2} \end{bmatrix}. \tag{96}$$

Now we define a linear rescaling of $\boldsymbol{T} = \left(\sum_{j=1}^J (\theta_j - \bar{Y}_j)^2, \sum_{j=1}^J (\theta_j - \mu)^2\right)$ as

$$\tilde{\boldsymbol{T}} = \frac{1}{\sqrt{J}} \sum_{j=1}^J \begin{bmatrix} (\theta_j - \bar{Y}_j)^2 - \frac{1}{m\tau_0^*+\tau_1^*} - \left(\frac{\tau_1^*}{m\tau_0^*+\tau_1^*}\right)^2 \left(\bar{Y}_j - \mu^*\right)^2 \\ (\theta_j - \mu)^2 - \frac{1}{m\tau_0^*+\tau_1^*} - \left(\frac{m\tau_0^*}{m\tau_0^*+\tau_1^*}\right)^2 \left(\bar{Y}_j - \mu^*\right)^2 \end{bmatrix} - C(\psi^*)\Delta_J, \tag{97}$$

with Δ_J as in (95). The next lemma shows the asymptotic distribution of $\tilde{\boldsymbol{T}}$ using the weak topology.

Lemma C.41. *Define $\tilde{\psi}$ and $\tilde{\boldsymbol{T}}$ as in* (95) *and* (97)*, respectively. For every $\tilde{\psi} \in \mathbb{R}^D$ it holds*

$$\left\| \mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) - N\left(C(\psi^*)\tilde{\psi}, V(\psi^*)\right) \right\|_W \to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$.

Proof. The result follows by arguments similar to the proof of Lemma C.11. First of all notice that $C(\psi)$ defined in (96) is such that

$$C(\psi) = \begin{bmatrix} E_{Y_j}\left[\partial_\mu E[(\theta_j - \bar{Y}_j)^2 \mid Y_j, \psi]\right] & E_{Y_j}\left[\partial_{\tau_1} E[(\theta_j - \bar{Y}_j)^2 \mid Y_j, \psi]\right] & E_{Y_j}\left[\partial_{\tau_0} E[(\theta_j - \bar{Y}_j)^2 \mid Y_j, \psi]\right] \\ E_{Y_j}\left[\partial_\mu E[(\theta_j - \mu)^2 \mid Y_j, \psi]\right] & E_{Y_j}\left[\partial_{\tau_1} E[(\theta_j - \mu)^2 \mid Y_j, \psi]\right] & E_{Y_j}\left[\partial_{\tau_0} E[(\theta_j - \mu)^2 \mid Y_j, \psi]\right] \end{bmatrix},$$

since by Lemma C.40 we have

$$\begin{aligned} & E_{Y_j}\left[\partial_\mu E[(\theta_j - \bar{Y}_j)^2 \mid Y_j, \psi]\right] = E_{Y_j}\left[\partial_\mu E[(\theta_j - \mu)^2 \mid Y_j, \psi]\right] = 0, \\ & E_{Y_j}\left[\partial_{\tau_0} E[(\theta_j - \bar{Y}_j)^2 \mid Y_j, \psi]\right] = E_{Y_j}\left[\partial_{\tau_0} E[(\theta_j - \mu)^2 \mid Y_j, \psi]\right] = \frac{m}{(m\tau_0+\tau_1)^2}, \\ & E_{Y_j}\left[\partial_{\tau_1} E[(\theta_j - \bar{Y}_j)^2 \mid Y_j, \psi]\right] = E_{Y_j}\left[\partial_{\tau_1} E[(\theta_j - \mu^*)^2 \mid Y_j, \psi]\right] = \frac{1}{(m\tau_0+\tau_1)^2}. \end{aligned}$$

By the same reasoning in the proofs of (63) and (64) we get

$$E_{Y_j}\left[\tilde{T} \mid Y_{1:J}, \psi^* + \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right] \quad \to \quad C(\psi^*)\tilde{\psi}$$

and

$$\left|\mathrm{Cov}\left(\tilde{\boldsymbol{T}} \mid Y_{1:J}, \psi^* + \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right) - \mathrm{Cov}\left(\tilde{\boldsymbol{T}} \mid Y_{1:J}, \psi^*\right)\right| \quad \to \quad 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Then by (83), Lemma C.40 and the Law of Large Numbers we have

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^{J}(\theta_j - \bar{Y}_j)^2 \mid Y_{1:J}, \psi^*\right) &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + 4\frac{(\tau_1^*)^2}{(m\tau_0^* + \tau_1^*)^3}\frac{1}{J}\sum_{j=1}^{J}(\bar{Y}_j - \mu^*)^2 \\ &\to \frac{2}{(m\tau_0^* + \tau_1^*)^2} + 4\frac{(m\tau_0^*)^{-1}\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^{J}(\theta_j - \mu^*)^2 \mid Y_{1:J}, \psi^*\right) &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + 4\frac{(m\tau_0^*)^2}{(m\tau_0^* + \tau_1^*)^3}\frac{1}{J}\sum_{j=1}^{J}(\bar{Y}_j - \mu^*)^2 \\ &\to \frac{2}{(m\tau_0^* + \tau_1^*)^2} + 4\frac{m\tau_0^*(\tau_1^*)^{-1}}{(m\tau_0^* + \tau_1^*)^2}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{Cov}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^{J}(\theta_j - \bar{Y}_j)^2, \frac{1}{\sqrt{J}}\sum_{j=1}^{J}(\theta_j - \mu^*)^2 \mid Y_{1:J}, \psi^*\right) = \frac{2}{(m\tau_0 + \tau_1)^2} - 4\frac{m\tau_0\tau_1}{(m\tau_0 + \tau_1)^3}\frac{1}{J}\sum_{j=1}^{J}(\bar{Y}_j - \mu)^2 \\ &\to -\frac{2}{(m\tau_0^* + \tau_1^*)^2},\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Finally, by the Law of Large Numbers and calculations similar to Lemma C.40, we have

$$E\left[(\theta_j - \bar{Y}_j)^{12} \mid Y_J, \psi\right] < \infty, \quad E\left[(\theta_j - \mu)^{12} \mid Y_J, \psi\right] < \infty$$

for every ψ. Therefore, with the same arguments in the proof of (65) we conclude that

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J} E\left[(\theta_j - \bar{Y}_j)^{12} \mid Y_j, \psi^* + \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right] \to 0, \quad \frac{1}{J^{3/2}}\sum_{j=1}^{J} E\left[(\theta_j - \mu^*)^{12} \mid Y_j, \psi^* + \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right] \to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely, as $J \to \infty$. The result then follows by Lyapunov version of Central Limit Theorem. □

We need another technical Lemma.

Lemma C.42. *Consider the assumptions of Proposition 5.6. Then it holds*

$$\left|E\left[e^{it_1(\theta_j - \mu)^2 + it_2(\theta_j - \bar{Y}_j)^2} \mid Y_j, \psi\right]\right| \leq \frac{e^{-\frac{2\sigma^2[\nu_j(t_1+t_2) - (t_1\mu + t_2\bar{Y}_j)]^2}{1+4\sigma^4(t_1+t_2)^2}}}{[1 + 4(t_1+t_2)^2\sigma^4]^{1/4}},$$

with $(t_1, t_2) \in \mathbb{R}_2$ *and*

$$\nu_j = \frac{m\tau_0}{m\tau_0 + \tau_1}\mu + \frac{\tau_1}{m\tau_0 + \tau_1}\bar{Y}_j, \quad \sigma^2 = \frac{1}{m\tau_0 + \tau_1}.$$

Proof. By simple computations we get

$$t_1(\theta_j-\mu)^2+t_2(\theta_j-\bar{Y}_j)^2=(t_1+t_2)\theta_j^2-2\theta_j(t_1\mu+t_2\bar{Y}_j)+t_1\mu^2+t_2\bar{Y}_j^2.$$

Therefore

$$\left|E\left[e^{it_1(\theta_j-\mu)^2+it_2(\theta_j-\bar{Y}_j)^2}\right]\right|\leq\left|E\left[e^{i\left((t_1+t_2)\theta_j^2-2\theta_j(\mu+\bar{Y}_j)\right)}\right]\right|.$$

Then we can apply Lemma C.26, with

$$a=t_1+t_2,\quad b=-2(t_1\mu+t_2\bar{Y}_j),\quad \nu=\frac{m\tau_0}{m\tau_0+\tau_1}\mu+\frac{\tau_1}{m\tau_0+\tau_1}\bar{Y}_j,\quad \sigma^2=\frac{1}{m\tau_0+\tau_1}.$$

□

Consistently with the previous Sections, we denote

$$\varphi(t\mid Y_j,\psi)=E\left[e^{it_1(\theta_j-\bar{Y}_j)^2+it_2(\theta_j-\mu)^2}\mid Y_j,\psi\right],\quad \tilde{\varphi}(t\mid Y_{1:J},\psi)=\mathbb{E}\left[e^{it^\top\tilde{\boldsymbol{T}}}\mid Y_{1:J},\psi\right]$$

for every ψ and $t=(t_1,t_2)\in\mathbb{R}^2$. The next lemma proves the same convergence of Lemma C.41 using the total variation distance.

Lemma C.43. *Define $\tilde{\psi}$ and $\tilde{\boldsymbol{T}}$ as in* (95) *and* (97)*, respectively. For every $\tilde{\psi}\in\mathbb{R}^D$ it holds*

$$\left\|\mathcal{L}(d\tilde{\boldsymbol{T}}\mid Y_{1:J},\tilde{\psi})-N\left(C(\psi^*)\tilde{\psi},V(\psi^*)\right)\right\|_{TV}\to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$.

Proof. Since the result holds under the weak metric by Lemma C.41, with the same reasoning of Lemma C.15 it suffices to prove

$$\lim_{A\to\infty}\lim_{B\to\infty}\limsup_{J\to\infty}\int_{\left((t_1+t_2)^2\leq A,t_1^2\leq B\right)^c}\left|\tilde{\varphi}(t\mid Y_{1:J},\psi^{(J)})\right|\,\mathrm{d}t=0$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$, where

$$\psi^{(J)}=\psi^*+\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}$$

Analogously, denote also

$$\mu^{(J)}=\mu^*+\frac{\tilde{\mu}+\Delta_{J,1}}{\sqrt{J}},\quad \tau_1^{(J)}=\tau_1^*+\frac{\tilde{\tau}_1+\Delta_{J,2}}{\sqrt{J}},\quad \tau_0^{(J)}=\tau_0^*+\frac{\tilde{\tau}_0+\Delta_{J,3}}{\sqrt{J}}.$$

As in (72) we have

$$|\tilde{\varphi}(t\mid Y_{1:J},\psi)|=\left|\prod_{j=1}^J\varphi\left(\frac{t}{\sqrt{J}}\mid Y_j,\psi\right)\right|.$$

Therefore, with the change of variables $u=t_1+t_2$ and $v=t_1$, we have

$$\begin{aligned}&\int_{\left((t_1+t_2)^2\leq A,t_1^2\leq B\right)^c}\left|\tilde{\varphi}(t\mid Y_{1:J},\psi^{(J)})\right|\,\mathrm{d}t\\&\qquad=\int_{(u^2\leq A,v^2\leq B)^c}\prod_{j=1}^J\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\,\mathrm{d}u\mathrm{d}v\end{aligned}$$

Moreover it is easy to see that

$$\left\{(u,v)\mid u^2\leq A\text{ and }v^2\leq B\right\}^c\subset\left\{(u,v)\mid u^2>A\right\}\cup\left\{(u,v)\mid u^2\leq A\text{ and }v^2>B\right\},$$

so that

$$\int_{(u^2\le A,v^2\le B)^c}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\,\mathrm{d}u\mathrm{d}v\le\int_{u^2>A}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\,\mathrm{d}u\mathrm{d}v$$
$$+\int_{(u^2\le A,v^2>B)}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\,\mathrm{d}u\mathrm{d}v. \tag{98}$$

For every ψ, by Lemma C.42 with

$$\nu_j=\frac{m\tau_0}{m\tau_0+\tau_1}\mu+\frac{\tau_1}{m\tau_0+\tau_1}\bar{Y}_j,\quad \sigma^2=\frac{1}{m\tau_0+\tau_1}$$

we have

$$\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi\right)\right|\le\frac{e^{-\frac{2\sigma^2\frac{1}{J}\sum_{j=1}^{J}\left[u(\nu_j-\bar{Y}_j)-v(\mu-\bar{Y}_j)\right]^2}{1+4\sigma^4u^2}}}{\left[1+4u^2\sigma^4\right]^{J/4}}.$$

Notice that

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}\left[u(\nu_j-\bar{Y}_j)-v(\mu-\bar{Y}_j)\right]^2=\\
&\quad=v^2\left[\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2\right]-2uv\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)\right]+u^2\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2\right]\\
&\quad=\left[\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2\right]\left[v-u\frac{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2}\right]^2\\
&\qquad\qquad+u^2\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2-\frac{\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)\right\}^2}{\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2}\right].
\end{aligned}$$

As regards the first element in (98), by integrating with respect to v we get

$$\int_{u^2>A}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\,\mathrm{d}u\mathrm{d}v\le\int_{u^2>A}\frac{e^{-\frac{2\sigma_J^2\frac{1}{J}\sum_{j=1}^{J}\left[u(\nu_j-\bar{Y}_j)-v(\mu^{(J)}-\bar{Y}_j)\right]^2}{1+4\sigma_J^4u^2}}}{\left[1+4u^2\sigma_J^4\right]^{J/4}}\,\mathrm{d}u\mathrm{d}v$$
$$\le\sqrt{\frac{\pi}{2\sigma_J^2\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}}\int_A^{\infty}\frac{e^{-\frac{2\sigma_J^2}{1+4\sigma_J^4u^2}u^2\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2-\frac{\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu_J-\bar{Y}_j)\right\}^2}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right]}}{\left[1+4u^2\sigma_J^4\right]^{J/4-1/2}}\,\mathrm{d}u,$$

where

$$\sigma_J^2=\frac{1}{m\tau_0^{(J)}+\tau_1^{(J)}},\quad \nu_j=\frac{m\tau_0^{(J)}}{m\tau_0^{(J)}+\tau_1^{(J)}}\mu^{(J)}+\frac{\tau_1^{(J)}}{m\tau_0^{(J)}+\tau_1^{(J)}}\bar{Y}_j.$$

By the Law of Large Numbers we have

$$\liminf\ \frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2=\liminf\ \frac{1}{J}\sum_{j=1}^{J}(\mu^*-\bar{Y}_j)^2=c_1>0$$

$Q_{\psi^*}^{(\infty)}$-almost surely and similarly

$$\liminf\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2-\frac{\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)\right\}^2}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right\}=c_2>0,$$

by Cauchy-Schwartz inequality, $Q_{\psi^*}^{(\infty)}$-almost surely. Moreover, by Lemma C.10

$$\sigma_J^2 \in \left(\frac{1}{2}\frac{1}{m\tau_0^* + \tau_1^*}, \frac{2}{m\tau_0^* + \tau_1^*}\right) = (\sigma_1^2, \sigma_2^2)$$

$Q_{\psi^*}^{(\infty)}$-almost surely, for J high enough. Therefore

$$\begin{aligned}\lim_{A\to\infty}\lim_{B\to\infty}\limsup_{J\to\infty}\int_{u^2>A}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| \mathrm{d}u\mathrm{d}v \\ \leq \lim_{A\to\infty}\sqrt{\frac{\pi}{2\sigma_1^2 c_1}}\int_A^\infty \frac{e^{-\frac{2c_2\sigma_1^2}{1+4\sigma_2^4 u^2}u^2}}{[1+4u^2\sigma_1^4]^{J/4-1/2}}\mathrm{d}u = 0\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely. As regards the second addend in (98) we get

$$\begin{aligned}\limsup_{J\to\infty}\int_{(u^2\leq A, v^2>B)}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| \mathrm{d}u\mathrm{d}v \\ \leq \int_{(u^2\leq A, v^2>B)} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^J(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^J(\mu^{(J)}-\bar{Y}_j)^2}\right]^2}\mathrm{d}u\mathrm{d}v,\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely. Fix $A > 0$ and notice that for every u we have

$$\lim_{B\to\infty}\int_B^\infty e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^J(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^J(\mu^{(J)}-\bar{Y}_j)^2}\right]^2}\mathrm{d}v = 0.$$

Moreover

$$\int_{u^2\leq A} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^J(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^J(\mu-\bar{Y}_j)^2}\right]^2}\mathrm{d}u\mathrm{d}v < \infty,$$

so that, by Dominated Convergence Theorem we get

$$\lim_{B\to\infty}\int_{(u^2\leq A, v^2>B)} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^J(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^J(\mu^{(J)}-\bar{Y}_j)^2}\right]^2}\mathrm{d}u\mathrm{d}v = 0,$$

for every $A > 0$ and the result follows. □

Proof of Proposition 5.6. The result follows by arguments similar to the proof of Theorem 4.2, that we briefly summarize. Since by construction

$$\mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{T}, Y_{1:J}\right)$$

a direct analogue of Lemma 4.1 holds. Moreover, by Lemmas C.39 and C.43, we can use Lemma C.18 to prove that $\mathcal{L}\left(\mathrm{d}\tilde{\boldsymbol{T}}, \mathrm{d}\tilde{\psi} \mid Y_{1:J}\right)$, as in (95), converges to a Gaussian vector with non singular covariance matrix. Finally, Lemma C.6 holds for P, being a two-block Gibbs sampler. Therefore the Gibbs sampler on the limit Gaussian target has a strictly positive spectral gap: thus the result follows by Corollary 2.5. □

C.25 Proof of Corollary 5.7

Let $\phi = (\tau_1, \tau_0)$ and define

$$\mathcal{I}(\phi^*) = \begin{bmatrix} \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2(\tau_1^*+m\tau_0^*)^2} & \frac{m}{2(\tau_1^*+m\tau_0^*)^2} \\ \frac{m}{2(\tau_1^*+m\tau_0^*)^2} & \frac{m-1}{2(\tau_0^*)^2} + \frac{(\tau_1^*)^2}{2(\tau_0^*)^2(\tau_1^*+m\tau_0^*)^2} \end{bmatrix}, \quad C(\phi^*) = \begin{bmatrix} \frac{1}{(m\tau_0^*+\tau_1^*)^2} & \frac{m}{(m\tau_0^*+\tau_1^*)^2} \\ \frac{1}{(m\tau_0^*+\tau_1^*)^2} & \frac{m}{(m\tau_0^*+\tau_1^*)^2} \end{bmatrix}$$

and

$$V(\phi^*) = \begin{bmatrix} \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{m\tau_0^*(\tau_1^*)^{-1}}{(m\tau_0^*+\tau_1^*)^2} & -\frac{2}{(m\tau_0^*+\tau_1^*)^2} \\ -\frac{2}{(m\tau_0^*+\tau_1^*)^2} & \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{\tau_1^*(m\tau_0^*)^{-1}}{(m\tau_0^*+\tau_1^*)^2} \end{bmatrix}.$$

We have a preliminary Lemma.

Lemma C.44. *Consider the setting of Proposition 5.6. Then we have*

$$\gamma(\psi^*) = \min\left\{ \frac{1}{1+\lambda_i} \,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\phi^*)\, C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*) \right\}.$$

Proof. With the same reasoning of Corollary 4.6, $\gamma(\psi^*)$ is the spectral gap on the limiting Gaussian distribution of $\left(\tilde{\psi}, \tilde{\boldsymbol{T}}\right)$, given by by Lemmas C.39 and C.43. By inspecting $\mathcal{I}(\psi^*)$ in (84) and $C(\psi^*)$ in (96), we have that $\tilde{\mu}$ is asymptotically independent from everything else, therefore it suffices to study the Gibbs sampler that alternates updates of $(\tilde{\tau}_1, \tilde{\tau}_0)$ and $\tilde{\boldsymbol{T}}$. Then the result follows by the same arguments of Corollary 4.6. □

Proof of Corollary 5.7. By Lemma C.44 we have to study the eigenvalues of

$$V^{-1}(\phi^*)\, C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*). \tag{99}$$

Notice that

$$\mathcal{I}(\phi^*) = \frac{1}{(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2} & \frac{m}{2} \\ \frac{m}{2} & \frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{2(\tau_0^*)^2} \end{bmatrix}, \quad C(\phi^*) = \frac{1}{(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} 1 & m \\ 1 & m \end{bmatrix}$$

and

$$V(\phi^*) = \frac{1}{(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} 2+4\frac{m\tau_0^*}{\tau_1^*} & -2 \\ -2 & 2+4\frac{\tau_1^*}{m\tau_0^*} \end{bmatrix}$$

Notice that

$$\begin{aligned} \left((m\tau_0^*+\tau_1^*)^2 V(\phi^*)\right)^{-1} &= \frac{m\tau_0^*\tau_1^*}{8(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} 2+4\frac{\tau_1^*}{m\tau_0^*} & 2 \\ 2 & 2+4\frac{m\tau_0^*}{\tau_1^*} \end{bmatrix} \\ &= \frac{1}{4(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} m\tau_0^*\tau_1^*+2(\tau_1^*)^2 & m\tau_0^*\tau_1^* \\ m\tau_0^*\tau_1^* & m\tau_0^*\tau_1^*+2(m\tau_0^*)^2 \end{bmatrix} \end{aligned}$$

and

$$\left((m\tau_0^*+\tau_1^*)^2\mathcal{I}(\phi^*)\right)^{-1} = \frac{2(\tau_1^*)^2}{m^2(m-1)(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} \frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2} & -m \\ -m & \frac{(m\tau_0^*)^2}{(\tau_1^*)^2} \end{bmatrix}$$

Therefore

$$\begin{aligned} \frac{m^2(m-1)(m\tau_0^*+\tau_1^*)^4}{2(\tau_1^*)^2} C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*) &= \begin{bmatrix} -m^2+\frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2} & \frac{m^3(\tau_0^*)^2}{(\tau_1^*)^2}-m \\ -m^2+\frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2} & \frac{m^3(\tau_0^*)^2}{(\tau_1^*)^2}-m \end{bmatrix}\begin{bmatrix} 1 & 1 \\ m & m \end{bmatrix} \\ &= \left(\frac{m^4(\tau_0^*)^2}{(\tau_1^*)^2} - 2m^2 + \frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2}\right)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \left(\frac{m^4(\tau_0^*)^4 - 2m^2(\tau_0^*)^2(\tau_1^*)^2 + (m-1)(\tau_1^*)^2(m\tau_0^*+\tau_1^*)^2 + (\tau_1^*)^4}{(\tau_0^*)^2(\tau_1^*)^2}\right)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} V^{-1}(\phi^*)\, C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*) = &\left(\frac{m^4(\tau_0^*)^4 - 2m^2(\tau_0^*)^2(\tau_1^*)^2 + (m-1)(\tau_1^*)^2(m\tau_0^*+\tau_1^*)^2 + (\tau_1^*)^4}{2m^2(m-1)(\tau_0^*)^2(m\tau_0^*+\tau_1^*)^4}\right) \\ &\begin{bmatrix} 2m\tau_0^*\tau_1^*+2(\tau_1^*)^2 & 2m\tau_0^*\tau_1^*+2(\tau_1^*)^2 \\ 2m\tau_0^*\tau_1^*+2(m\tau_0^*)^2 & 2m\tau_0^*\tau_1^*+2(m\tau_0^*)^2 \end{bmatrix} \end{aligned}$$

Notice that the matrix on the right hand side admits 0 as an eigenvalue, so that the highest eigenvalue in absolute value is given by its trace, that is

$$4m\tau_0^*\tau_1^* + 2(\tau_1^*)^2 + 2(m\tau_0^*)^2 = 2(m\tau_0^* + \tau_1^*)^2,$$

so that the highest eigenvalue of (99) is given by

$$\frac{m^4(\tau_0^*)^4 - 2m^2(\tau_0^*)^2(\tau_1^*)^2 + (m-1)(\tau_1^*)^2(m\tau_0^* + \tau_1^*)^2 + (\tau_1^*)^4}{m^2(m-1)(\tau_0^*)^2(m\tau_0^* + \tau_1^*)^2}.$$

The result follows by noticing

$$\begin{aligned} m^4(\tau_0^*)^4 - 2m^2(\tau_0^*)^2(\tau_1^*)^2 + (\tau_1^*)^4 &= \left[m^2(\tau_0^*)^2 - (\tau_1^*)^2\right]^2 \\ &= (m\tau_0^* - \tau_1^*)^2(m\tau_0^* + \tau_1^*)^2. \end{aligned}$$

□

C.26 Proof of Lemma 5.8

Proof. The proof follows the same lines of Lemma 4.1, that we briefly summarize. Since

$$\mathcal{L}\left(\mathrm{d}\theta, \mathrm{d}\tau_\beta \mid \boldsymbol{\beta}, Y^{(n)}\right) = \mathcal{L}(\mathrm{d}\theta, \mathrm{d}\tau_\beta \mid \boldsymbol{T}(\boldsymbol{\beta}), Y^{(n)}\Big) \tag{100}$$

holds by definition of $\boldsymbol{T}$, reasoning as in (59) we can conclude

$$\begin{aligned} \mathcal{L}\Big(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\theta^{(t)}, \mathrm{d}\tau_\beta^{(t)} \mid \boldsymbol{T}^{(t-1)}, \theta^{(t-1)}, \tau_\beta^{(t-1)}\Big) \\ = \hat{\pi}_n\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \theta^{(t-1)}, \tau_\beta^{(t-1)}\right)\hat{\pi}_n\left(\mathrm{d}\theta^{(t)}, \mathrm{d}\tau_\beta^{(t)} \mid \boldsymbol{T}^{(t)}\right), \end{aligned}$$

which proves that the transition kernel of the induced chain $\left(\boldsymbol{T}^{(t)}, \theta^{(t)}, \tau_\beta^{(t)}\right)_{t\geq 1}$ coincides with $\hat{P}_n$. The second part of the Lemma follows by the same reasoning used in (60). □

C.27 Proof of Corollary 5.9

Proof. By Lemma 5.8 we have

$$t_{mix}^{(n)}(\epsilon, M) = \sup_{\nu \in \mathcal{N}(\hat{\pi}_n, M)} \hat{t}_{mix}^{(n)}(\epsilon, \nu).$$

The result then follows by Corollary 2.5, whose conditions hold by assumption. □

C.28 Proof of Corollary 5.10

Proof. It is easy to show that an analogue of Lemma 5.8 holds, with $\psi = (\theta, \tau_\beta, \tau_\epsilon)$ and $\boldsymbol{T} = \left(T_\theta, T_{\tau_\beta}, T_{\tau_\epsilon}\right)$. Thus the result follows with the same reasoning of Corollary 5.9. □

C.29 Proof of Theorem 6.1

Denote with $\tilde{\mu}_J$ the push-forward measure of μ_J according to transformations (17) and (19). The next theorem shows that the rescaled version of μ_J is a warm start for the limiting distribution in Proposition 4.5.

Lemma C.45. *Let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ be as in (35). Then under assumptions $(B1) - (B3)$ there exists a positive constant $M = M(c)$ such that*

$$Q_{\psi^*}^{(J)}\left(\tilde{\mu}_J \in \mathcal{N}\left(N(\mathbf{0}, \Sigma), M\right)\right) \quad \to \quad 1,$$

as $J \to \infty$, with Σ as in Proposition 4.5.

Proof. According to transformations (17), we have

$$\tilde{\mu}_J^{(-1)} = \mathrm{Unif}\left(\sqrt{J}\left(\hat{\psi}_J - \psi^*\right) - \Delta_J, c\right).$$

Denote with $B_r(\mathbf{x})$ the closed ball of radius $r > 0$ and center $\mathbf{x} \in \mathbb{R}^D$. By Theorem 5.39 in [64] it holds

$$Q_{\psi^*}^{(J)}\left(\left(\sqrt{J}\left(\hat{\psi}_J - \psi^*\right) - \Delta_J\right) \in B_1(\mathbf{0})\right) \quad \to \quad 1, \tag{101}$$

as $J \to \infty$. Define now

$$M = \max_{\mathbf{x} \in B_{c+1}(\mathbf{0})} \frac{\mathrm{Vol}\left(B_{c+1}(\mathbf{0})\right)}{N(\mathbf{x} \mid \mathbf{0}, \Sigma_D)}, \tag{102}$$

where $\mathrm{Vol}(A)$ is the volume of set A and $N(\mathbf{0}, \Sigma_D)$ is the marginal distribution of $N(\mathbf{0}, \Sigma)$ over the last D components. It is easy to see that $M < \infty$ and it does not depend on J. Therefore, by (101), we conclude

$$\begin{aligned} Q_{\psi^*}^{(J)}\left(\tilde{\mu}_J \in \mathcal{N}\left(N(\mathbf{0}, \Sigma), M\right)\right) &\leq Q_{\psi^*}^{(J)}\left(\max_{\mathbf{x} \in B_{c+1}(\mathbf{0})} \frac{\mathrm{d}\tilde{\mu}_J^{(-1)}}{\mathrm{d}N(\mathbf{0}, \Sigma_D)}(\mathbf{x}) \leq M\right) \\ &\leq Q_{\psi^*}^{(J)}\left(\left(\sqrt{J}\left(\hat{\psi}_J - \psi^*\right) - \Delta_J\right) \in B_1(\mathbf{0})\right) \quad \to \quad 1, \end{aligned}$$

as $J \to \infty$. □

Proof of Theorem 6.1. Let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ be as in (35). Thus, by Lemma C.45 the event $\{\tilde{\mu}_J \in \mathcal{N}\left(\tilde{\pi}, M\right)\}$ with M as in (102) holds with probability converging to 1, with respect to the law $Q_{\psi^*}^{(J)}$. Then, by Lemma 2.3, there exists $\tilde{\nu}_J \in \mathcal{N}(\tilde{\pi}_J, M)$ such that

$$\left\|\tilde{\nu}_J - \tilde{\mu}_J\right\|_{TV} \leq M \left\|\tilde{\pi}_J - \tilde{\pi}\right\|_{TV}.$$

Therefore, by the above facts, the triangle inequality and Lemma 4.1 we have

$$\begin{aligned} \left\|\mu_J P_J^t - \pi_J\right\|_{TV} &= \left\|\tilde{\mu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &\leq \left\|\tilde{\mu}_J \tilde{P}_J^t - \tilde{\nu}_J \tilde{P}_J^t\right\|_{TV} + \left\|\tilde{\nu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &\leq \left\|\tilde{\mu}_J - \tilde{\nu}_J\right\|_{TV} + \left\|\tilde{\nu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &\leq M \left\|\tilde{\pi}_J - \tilde{\pi}\right\|_{TV} + \sup_{\tilde{\nu}_J \in \mathcal{N}(\tilde{\pi}_J, M)} \left\|\tilde{\nu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &= M \left\|\tilde{\pi}_J - \tilde{\pi}\right\|_{TV} + \sup_{\nu_J \in \mathcal{N}(\pi_J, M)} \left\|\nu_J P_J^t - \pi_J\right\|_{TV}. \end{aligned}$$

Thus the result follows by Theorem 4.2. □