Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics and Computer Science

Cycle: XXXVI PhD cycle

Disciplinary Field (code): SECS-S/01

Topics in scalable Bayesian posterior estimation

Advisor: Giacomo Zanella

Co-Advisor: Botond Szabó

PhD Thesis by

Paolo Maria Ceriani

ID number: 3072885

Year 2025

Abstract

As the complexity and dimensionality of data continue to increase, it is becoming fundamental to develop advanced strategies for statistical inference and to explore their computational properties (Bishop, 2006).

This thesis considers Bayesian models, known for their ability to frame prediction and uncertainty within a coherent probabilistic framework. However, achieving accurate estimates of posterior quantities within these models generally requires innovative techniques to accommodate the challenges of modern data analysis. We aim at developing algorithms for exact and approximate posterior estimation exhibiting linear computational cost in the number of parameters, for asymptotic settings where both the numbers of parameters and observations grow to infinity. Such performances are substantially unattainable for state of the art gradient based sampling methods, and are achieved only leveraging the hidden probabilistic structure of the models under consideration.

The first and second chapters of this document focus on couplings, a relatively simple probabilistic construction whose potential for unbiased estimation has been recently spotlighted thanks to the work of (Glynn and Rhee, 2014; Jacob et al., 2020). After a brief review on couplings and their applications for unbiased sampling and estimation in Chapter 1, we present in Chapter 2 theoretical results bounding the computational effort required by the coupling construction of Jacob et al. (2020) for certain Gibbs samplers, proving its scalability in a wide range of applications, spanning from crossed random effect to sparse graphical models. Unbiased estimation via couplings therefore presents a promising way to enhance the precision and accuracy of statistical inference, offering insights beyond traditional estimation approaches.

Turning to Chapter 3, we cover topics related to variational inference. Variational inference has captured significant attention in the past decades: essentially, it translates

the probabilistic problem of finding the posterior distribution as an optimization task (Blei et al., 2017). This chapter not only presents its theoretical foundations but also explores practical implementation and provides results on scalability of the mean field variational approximation for certain large scale hierarchical models. More in detail, assuming the data is randomly generated from a specific distribution, we characterize the rate at which the iterates produced by the coordinate ascent variational inference (CAVI) algorithm converge to a variational minimizer for large scale hierarchical models, proving dimension-free convergence under warm start assumptions. Our work builds upon (Ascolani and Zanella, 2024), highlighting the effectiveness of CAVI in efficiently approximating posterior quantities for models where Gibbs sampling has proved to be effective, given the inherent similarities between these coordinate-wise schemes (Tan and Nott, 2014).

Chapter 4 contains some recent advances developed during my visiting period at Warwick University with professor Gareth Roberts. Specifically, we study some properties of Catalytic couplings (Breyer and Roberts, 2001), a coupling procedure well suited for settings where only unnormalized distributions are available and able to couple multiple chains at once.

In summary, this dissertation aims at presenting efficient methods in the realm of Bayesian posterior estimation for models with sparse dependencies, such as hierarchical and crossed models, leveraging their probabilistic structure to obtain linear cost estimates. By investigating coupling methods and variational inference, we aim at helping bridge the gap between state-of-the-art statistical procedures and the understanding of their computational properties.

References

Ascolani, F. and Zanella, G. (2024+). Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models. *The Annals of Statistics*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Breyer, L. A. and Roberts, G. O. (2001). Catalytic perfect simulation. *Methodology And Computing In Applied Probability*, 3:161–177.

Glynn, P. and Rhee, C.-H. (2014). Exact Estimation for Markov chain equilibrium expectation. *Journal of Applied Probability*, 51A:377–389.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Tan, S. L. and Nott, D. J. (2014). Variational Approximation for Mixtures of Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 23(2):564–585.

Contents

Acknowledgements

- Les gens ont des étoiles qui ne sont pas les mêmes. Pour les uns, qui voyagent, les étoiles sont des guides. Pour d'autres, elles ne sont rien que de petites lumières. Pour d'autres, qui sont savants, elles sont des problèmes. Pour mon businessman, elles étaient de l'or. Mais toutes ces étoiles-là se taisent. Toi, tu auras des étoiles comme personne n'en a...

– Que veux-tu dire ?

– Quand tu regarderas le ciel, la nuit, puisque j'habiterai dans l'une d'elles, puisque je rirai dans l'une d'elles, alors ce sera pour toi comme si riaient toutes les étoiles. Tu auras, toi, des étoiles qui savent rire !

Le petit prince, Antoine de Saint-Exupéry

A tutti voi, che non posso più ringraziare, ma che avete riempito il mio cielo di stelle che sanno ridere. Grazie.

Chapter 1

Couplings for Monte Carlo Markov Chains

This chapter addresses existing research on couplings, which will be essential for the developments discussed in Chapter 2. We begin with formal definitions of couplings of distributions and of transition kernels in Section 1.1.1. Subsequently, in Sections 1.1.2 and 1.1.3, we explore various algorithms for sampling from maximal and optimally contractive couplings, respectively. Maximal couplings aim to maximize the probability that draws from two coupled distributions will be equal, while optimally contractive couplings, on the other hand, focus on minimizing the expected square distance between coupled processes over time. The combination of the two is especially useful in applied contexts, as later explained.

In Section 1.2, we focus on previous seminal contributions that motivate our work in Chapter 2. This includes a discussion of the "coupling from the past" method for exact simulation, as introduced by Propp and Wilson (1996). This method leverages coupling to generate samples from the exact stationary distribution of a Markov chain by running chains backwards in time until they coalesce in one point. We also present the innovative "couplings for unbiased estimation" approach developed by Jacob et al. (2020), which employs coupling techniques to produce unbiased estimates in Monte Carlo simulations. Both of these applications require the precise meeting of coupled chains, attainable only through careful coupling constructions. We will focus on those constructions involving both maximal and optimally contractive couplings, as explained in Section 1.2.5, aiming

at procedures requiring small number of iterations for exact meeting of the chains.

These applications underscore the significance of couplings in probabilistic sampling and estimation, demonstrating their broad utility in both theoretical and practical contexts. By providing a comprehensive review of existing works and methodologies, this chapter sets the stage for the advanced developments presented in the subsequent chapters.

1.1 Introduction to couplings

1.1.1 Definitions and Notation

In the following, vectors are denoted in bold, matrices in capital letters and univariate quantities in standard lowercase. We denote the space of probability measures over a space $\mathcal{X}$ by $\mathcal{P}(\mathcal{X})$. Densities are denoted by $p(\cdot)$, probabilities as $\Pr(\cdot)$ in opposition to transition kernels P. Other conventions and notational details will be introduced later in the thesis whenever needed.

Let $\mathcal{X}$ be a Polish space and $\mathcal{P}(\mathcal{X})$ the space of probability measures over $\mathcal{X}$, then

Definition 1. *A coupling of two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$, is a joint distribution on $\mathcal{X} \times \mathcal{X}$ whose first and second marginals are, respectively, P and Q.*

The class of all possible couplings of P and Q is denoted as $\Gamma(P, Q)$. In the following we will consider mainly probability measures P, Q admitting density with respect to the Lebesgue measure $\mathcal{L}$ on $\mathbb{R}^d$, respectively $p = \frac{dP}{d\mathcal{L}}$ and $q = \frac{dQ}{d\mathcal{L}}$, and we will denote the space of couplings as $\Gamma(p, q)$ for notational convenience. With a slight abuse of notation we use $\Gamma(p, q)$ to denote both the collection of distributions and that of random variables, i.e. we also write $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p, q)$ for random vectors $(\mathbf{X}, \mathbf{Y})$ such that $\mathbf{X} \sim p, \mathbf{Y} \sim q$.

Given $\mathbf{X}, \mathbf{Y}$ distributed as p, q, it holds $\Gamma(p, q) \neq \emptyset$ since trivially the *independent coupling*, i.e. a joint distribution with p and q as independent components, belongs to the set. In particular, given that two marginal distributions don't specify uniquely the joint, infinitely many different couplings can be devised for the marginals P and Q. As a simple example consider $X \sim Bern(p)$ and $Y \sim Bern(q)$, for $0 \leq p, q \leq 1$. The independent coupling amounts to sampling X and Y independently, but it is also possible to sample

$u \sim U(0,1)$ and set $(X,Y) = (I_{\{u<p\}}, I_{\{u<q\}})$, where I indicates the indicator function. In general it holds

Lemma 1 (Coupling Lemma, (Thorisson, 2000; Lindvall, 1992))**.** *Given $\boldsymbol{X}$, $\boldsymbol{Y}$ random variables defined on $\mathcal{X}$ distributed according to p and q respectively, it holds*

$$\sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\Pr(\boldsymbol{X} \in A) - \Pr(\boldsymbol{Y} \in A)| = \|p - q\|_{tv} \leq \Pr(\boldsymbol{X} \neq \boldsymbol{Y}),$$

where $\|p-q\|_{tv}$ denotes the total variation distance and $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-algebra.

Notably, the left hand side depends only on properties of the marginals, while the right concerns a relation between $\mathbf{X}$ and $\mathbf{Y}$. A coupling of $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p,q)$ maximizing the probability of equality is termed *maximal* and we will denote the collection of maximal couplings as $\Gamma_{max}(p,q) \subset \Gamma(p,q)$. We report in Section 1.1.2 a brief recap on famous algorithms for sampling from $\gamma \in \Gamma_{max}(p,q)$. By definition, for every $\gamma \in \Gamma_{max}(P,Q)$, it holds

$$\Pr_{(\mathbf{X},\mathbf{Y})\sim\gamma}(\mathbf{X} = \mathbf{Y}) = 1 - \|p - q\|_{tv}.$$

We also denote by $(p \otimes q) \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ the product measure defined as $(p \otimes q)(A \times B) = p(A)q(B)$ for all $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$.

We also introduce the notion of Wasserstein-2 optimal coupling: a coupling $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p,q)$ minimizing $\mathbb{E}[\|\mathbf{X} - \mathbf{Y}\|^2]$ among all couplings of p and q is called W_2-optimal, and we will denote the family of such optimal couplings as $\Gamma_{W_2}(p,q)$. Such a family will be central in Chapter 2 for the development of the two step strategy of Algorithm 5: aiming at coupled chains with small meeting times, the W_2 optimal couplings produce draws with minimal expected square distance, hence they generally contract the chains and allow for faster coalescence, i.e. chains with same values. Note that for all univariate distributions, it is known that the *monotone map* coupling, which consists of using same random number for inverse cdf method, is W_2-optimal (Santambrogio (2015))[1]. See Section 1.1.3 for further details on sampling.

In the following, we will often consider couplings of Markov chains. For the purpose we introduce here the notions of Markov transition kernels and couplings of them.

[1] Actually not only for the euclidean cost, i.e. $c(x,y) = \|x-y\|$, but also for every cost function of the form $c(x,y) = h(x-y)$ with h convex, see e.g. Theorem 2.9 in Santambrogio (2015).

Definition 2. *Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. A Markov transition kernel P is a map $P : (\mathcal{X}, \mathcal{F}) \to [0, 1]$ s.t.*

- *For every fixed $A_0 \in \mathcal{F}$ the map $x \mapsto P(x, A_0)$ is $\mathcal{F}$-measurable.*
- *For every fixed $x_0 \in \mathcal{X}$ the map $A \mapsto P(x_0, A)$ is a probability measure on $\mathcal{X}$.*

From the above definition we are ready to define couplings of kernels.

Definition 3. *Given a transition kernel P on $\mathcal{X}$, a coupling of P with itself, denoted by $\bar{P}[P]$ (or simply $\bar{P}$), is a kernel on $\mathcal{X} \times \mathcal{X}$ such that*

$$\bar{P}[P]((\boldsymbol{x}, \boldsymbol{y}), \cdot) \in \Gamma\left(P(\boldsymbol{x}, \cdot), P(\boldsymbol{y}, \cdot)\right) \text{ for every } (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{X}.$$

The space of couplings of P with itself will be denoted as $\Gamma[P]$. Analogously, we write $\bar{P}[P] \in \Gamma_{max}[P]$ if $\bar{P}((\mathbf{x}, \mathbf{y}), \cdot) \in \Gamma_{max}(P(\mathbf{x}, \cdot), P(\mathbf{y}, \cdot))$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Using the same notational conventions, we say that $\bar{P}$ is a W_2-optimal coupling of a Markov kernel P (with itself), and write $\bar{P}[P] \in \Gamma_{W_2}[P]$, if $\bar{P}((\mathbf{x}, \mathbf{y}), \cdot) \in \Gamma_{W_2}(P(\mathbf{x}, \cdot), P(\mathbf{y}, \cdot))$ for every $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

1.1.2 Sampling from maximal couplings

We briefly review algorithms for sampling from maximal couplings of two distributions $p, q \in \mathcal{P}(\mathcal{X})$, with $\mathcal{X} = \mathbb{R}^d$. Provided p and q admit densities, there always exists an algorithm with maximal meeting probability, referred to as *maximal rejection* (Algorithm 1); for further references see (Thorisson, 2000, Sec. 4.5 of Chap 1) or the γ-coupling in Johnson (1998). However, in the case of spherically symmetric distributions (e.g. multivariate Gaussian with the same covariance matrix), an alternative approach called *maximal reflection coupling* (see Algorithm 2 or Eberle et al. (2019); Bou-Rabee et al. (2020)) allows for sampling with a deterministic cost and yet allowing for a good contraction of the square distance between the resulting draws (see e.g. Lemma 6).

Maximal rejection coupling. Algorithm 1 reports the pseudo code for implementing a maximal rejection coupling of p, q. The computational cost of Algorithm 1 depends on the number of iterations required to accept the proposed sample. Only one sample is

Algorithm 1: Maximal rejection coupling of $p, q \in \mathcal{P}(\mathbb{R}^d)$

Input: densities p, q;
sample $\mathbf{X} \sim p$
sample $W \sim U(0,1)$
if $Wp(\boldsymbol{X}) \leq q(\mathbf{X})$ **then**
  set $\mathbf{Y} = \mathbf{X}$
else
  sample $\mathbf{Y}^* \sim q$ and $W^* \sim U(0,1)$
  while $W^* q(\mathbf{Y}^*) < p(\mathbf{Y}^*)$ **do**
    sample $\mathbf{Y}^* \sim q$ and $W^* \sim U(0,1)$
  set $\mathbf{Y} = \mathbf{Y}^*$
Output: $(\mathbf{X}, \mathbf{Y})$.

required if $\mathbf{X}$ is accepted as value for $\mathbf{Y}$, and this happens with probability $\Pr(p(\mathbf{X})\, W \leq q(\mathbf{X})) = 1 - \|p - q\|_{TV}$. If instead $\mathbf{X}$ is rejected, the number of trials before acceptance follows a Geometric variable with parameter $\Pr(q(\mathbf{Y}^*)\, W^* > p(\mathbf{Y}^*))$, the latter being equal to $\|p - q\|_{TV}$. The resulting expected number of iterations is

$$1 - \|p - q\|_{TV} + \|p - q\|_{TV}(1/\|p - q\|_{TV} + 1) = 2.$$

The variance of the expected number of iterations is equal to $\frac{2(1-\|p-q\|_{TV})}{\|p-q\|_{TV}}$, which tends to infinity as $\|p - q\|_{TV}$ approaches zero. Gerber and Lee proposed accepting the first draw with a lower probability, thereby avoiding the problem of infinite variance, but at the cost of losing maximality (see the discussion in Jacob et al. (2020)).

Maximal reflection coupling. Algorithm 2 reports an implementation of maximal reflection coupling for Gaussian distributions with same covariance matrix. Note that,

Algorithm 2: Maximal reflection coupling of $N(\boldsymbol{\xi}, \Sigma)$ and $N(\boldsymbol{\nu}, \Sigma)$

Input: means $\boldsymbol{xi}, \boldsymbol{\nu}$, variance Σ ;
set $\mathbf{z} = \Sigma^{-1/2}(\boldsymbol{\xi} - \boldsymbol{\nu})$, $\mathbf{e} = \mathbf{z}/||\mathbf{z}||$
sample $\dot{\mathbf{X}} \sim N_d(\mathbf{0}, I_d)$, $W \sim U(0,1)$
if $W \leq \exp\{-\frac{1}{2}\mathbf{z}^\top(2\dot{\mathbf{X}} + \mathbf{z})\}$ **then**
  set $\dot{\mathbf{Y}} = \dot{\mathbf{X}} + \mathbf{z}$
else
  $\dot{\mathbf{Y}} = \dot{\mathbf{X}} - \mathbf{2}(\mathbf{e}^\top \dot{\mathbf{X}})\mathbf{e}$
$\mathbf{X} = \Sigma^{1/2}\dot{\mathbf{X}} + \boldsymbol{\xi}$
$\mathbf{Y} = \Sigma^{1/2}\dot{\mathbf{Y}} + \boldsymbol{\nu}$ **Output:** $(\mathbf{X}, \mathbf{Y})$.

differently from Algorithm 1, in Algorithm 2 no rejection mechanism is required and the algorithm's runtime is deterministic. Furthermore, in high-dimensional cases, this procedure shows favourable behaviours: in addition to being maximal, the algorithm contracts the distance between chains with a good rate. Thus, when applicable, it is generally preferred to Algorithm 1.

1.1.3 Sampling from W_2-optimal couplings

For all univariate distributions, it is known that the *monotone map* coupling (i.e. using same random number for inverse cdf method) is optimal for every cost function of the form $c(x, y) = h(x - y)$ with h convex (Santambrogio, 2015, Thm.2.9), and hence is W_2-optimal. For p, q univariate distributions, let $F_p(\cdot)$ and $F_q(\cdot)$ denote their cumulative density function (cdf). We define the inverse cdf as

$$F_p^{-1}(u) := \inf_{t \in \mathbb{R}} \{t : F_p(t) \geq u\},$$

and F_q^{-1} accordingly. It is then possible to sample from the W_2 optimal coupling as in Algorithm 3.

Algorithm 3: monotone transport map for univariate distributions

Input: pseudo inverse $F_p^{-1}(\cdot), F_q^{-1}(\cdot)$;
sample $U \sim U(0, 1)$
set $X = F_p^{-1}(U)$
set $Y = F_q^{-1}(U)$
Output: (X, Y).

No universal optimality result exists for general multivariate distributions, but a natural extension of the monotone map above, called *common random number* (*crn*) coupling, is indeed optimal for multivariate Gaussians whose covariance matrices commute (Dowson and Landau, 1982; Olkin and Pukelsheim, 1982), as stated below. While the following result is known, we report a proof for self-containedness.

Lemma 2 (Optimality of *crn* coupling for Gaussian distributions)**.** *Let* $p = N(\boldsymbol{\xi}, \Sigma_1)$

and $q = N(\boldsymbol{\nu}, \Sigma_2)$ be d-dimensional Gaussian, with $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$. Define

$$\Gamma^* := N\left(\begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \Sigma_1 & FG^\top \\ GF^\top & \Sigma_2 \end{pmatrix}\right), \tag{1.1}$$

where $FF^\top = \Sigma_1, GG^\top = \Sigma_2$. Then $\Gamma^* \in \Gamma_{W_2}(p, q)$, i.e. Γ^* is the W_2-optimal coupling of p and q.

Proof of Lemma 2. Let $(\mathbf{X}, \mathbf{Y}) \sim \Gamma^*$. If $\mathbf{Z} \sim N(\mathbf{0}, 1_d)$, we have $\mathbf{X} \overset{d}{=} \boldsymbol{\xi} + F\mathbf{Z}$, $\mathbf{Y} \overset{d}{=} \boldsymbol{\nu} + G\mathbf{Z}$, where $FF^\top = \Sigma_1$ and $GG^\top = \Sigma_2$ and $\overset{d}{=}$ denotes equality in distribution. Denote the (i, j)-th entry of F and G as f_{ij} and g_{ij}, respectively. It holds

$$\begin{aligned}
\mathbb{E}[\|\mathbf{X} - \mathbf{Y}\|^2] &= \mathbb{E}[||\boldsymbol{\xi} - \boldsymbol{\nu} + (F - G)\mathbf{Z}||^2] = \sum_{i=1}^{d} \mathbb{E}\left[\left((\xi_i - \nu_i) + \sum_{j=1}^{d}(f_{ij} - g_{ij})Z_j\right)^2\right] \\
&= \sum_{i=1}^{d}(\xi_i - \nu_i)^2 + \sum_{i=1}^{d}\mathbb{E}\left[\left(\sum_{j=1}^{d}(f_{ij} - g_{ij})Z_j\right)^2\right] + 2\sum_{i=1}^{d}(\xi_i - \nu_i)\left(\sum_{j=1}^{d}(f_{ij} - g_{ij})\mathbb{E}[Z_j]\right).
\end{aligned}$$

Since $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_iZ_j] = 0$ for $i \neq j$ and $Z_j^2 \sim \chi(1)$, then

$$\begin{aligned}
\mathbb{E}[\|\mathbf{X} - \mathbf{Y}\|^2] &= ||\boldsymbol{\xi} - \boldsymbol{\nu}||^2 + \mathbb{E}[||(F - G)\mathbf{Z}||^2] \\
&= \|\boldsymbol{\xi} - \boldsymbol{\nu}\|^2 + \sum_{ij}|f_{ij} - g_{ij}|^2 = \|\boldsymbol{\xi} - \boldsymbol{\nu}||^2 + \|F - G\|_{fr}^2,
\end{aligned}$$

where $\|\cdot\|_{fr}$ denotes the Frobenius norm of matrices. The latter expression exactly equals the minimizer of the optimal transport problem in the Gaussian case, whenever the variance covariance matrices commute (Givens and Shortt, 1984). □

Thus, in order to obtain draws from Γ^* in Lemma 2, one can simply sample $\mathbf{Z} \sim N(\mathbf{0}_d, 1_d)$ and then set

$$\begin{cases} \mathbf{X} = \boldsymbol{\mu} + F\mathbf{Z}, \\ \mathbf{Y} = \boldsymbol{\nu} + G\mathbf{Z}. \end{cases} \tag{1.2}$$

We will refer to (1.2) as *crn* coupling. Recall also that the W_2-optimal map is unique (Brenier, 1991).

1.2 Historical overview

The current section is devoted to a brief overview on the different applications of couplings in the context of MCMC. We start in Section 1.2.1 by reviewing a technique allowing for exact sampling from a target distribution, trailblazer of the subsequent literature on the field. We then move to Section 1.2.2 for exact estimation with randomly truncated sum and Section 1.2.3 for the recent work of Jacob et al. (2020), that will later be used in Chapter 2 for the development of our theory.

1.2.1 Exact simulation

With the appearance of Propp and Wilson (1996)'s pioneering paper, an exciting prospect for Markov Chain Monte Carlo was unveiled: to simulate exactly from the target distribution leveraging existing Markov kernels to drive the simulations. The approach is based on the "Coupling from the past" algorithm (CFTP), i.e. a procedure that, instead of running the Markov chains forward up to convergence, runs coupled chains from one undefined point in the past up to the present. The distance into the past that one needs to go is determined during the running of the algorithm itself (see details below). In the following we briefly describe the approach of the original formulation, developed for Markov chains on finite state space or admitting a natural partial ordering.

The building block of the CFTP is the *backward simulation*, that is a method to run chains from iteration $-M$ in the past until time 0 in such a way that they have the same distribution as standard M steps forward chains, but the last draw can be interpreted as the outcome of an infinitely long chain in the past. Start by considering a Markov chain $(X^t)_{-t\in\mathbb{N}}$ on a finite state space $S = \{1, ..., n\}$, evolving according to a π-invariant kernel P. For simplicity of exposition, write the kernel as a deterministic function $f(n, u)$ of the starting point n and a random quantity u, i.e. f is such that $\Pr(f(n, \cdot) \in A) = P(n, A)$ for $A \subset S$. Let $\mathbf{u} = (..., u_{-2}, u_{-1})$ be an infinite sequence of those random quantities. Starting from time 0, one simulates the evolution of the chain from time -1 to time 0: since obviously the value of the chain at 0 is determined by the previous iteration, one should run the chains for all possible values and keep track of their evolution, i.e.

store $f(n, u_{-1})\,\forall n \in S$. Similarly, it should be done for all times t with $-M \leq t \leq -1$. Define $F_{t_1}^{t_2}(n, \mathbf{u}) = f(f(...(f(n, u_{t_1}), u_{t_1+1}), ..., u_{t_2-2}), u_{t_2-1})$, where $t_1 < t_2$. The output of the fixed time simulations will be $|S|$ trajectories, associating every starting point n to $F_{-M}^{0}(n, \mathbf{u})$, namely the endpoint of the M-step trajectory starting at $-M$ in the past. If there was nothing more, the above procedure simply run M steps of the Markov kernels with same distribution as the M steps forward simulation. The main novelty is the observation that, if somehow it is possible to make $F_{-M}^{0}(\cdot, \mathbf{u})$ almost surely constant (namely all the trajectories pointing at the same point), then the output can be seen as a result of a stationary chain run starting from infinite time in the past, and hence as draws directly from π (see Theorems 1 and 2 in Propp and Wilson (1996)).

Algorithm 4: Exact sampling

Input: function $f(\cdot, u)$, state space S;
Set $t = 0$;
Sample $\mathbf{u}$;
Set $F_t^0(n, \mathbf{u}) = n$, $\forall n \in S$;
while $F_t^0(n, \boldsymbol{u})$ *not constant* **do**
  $t = t - 1$;
  $F_t^0(n, \mathbf{u}) = F_{t+1}^0(f(n, u_t), \mathbf{u})$;
Output: one sample $F_t^0(n, \mathbf{u})$ from π.

The algorithm can be extended to an infinite but partially ordered state space S admitting lower and upper elements, i.e. $\exists \hat{0}, \hat{1} \in S$ s.t. $\hat{0} \leq x \leq \hat{1}\,\forall x \in S$. Suppose furthermore that the transition kernel satisfies the monotonicity property, i.e. if $x < y$ then $f(x, u) < f(y, u)$ almost surely. Note that if $u_{-T}, ..., u_{-1}$ for some T have the property that $F_{-T}^{0}(\hat{0}, \mathbf{u}) = F_{-T}^{0}(\hat{1}, \mathbf{u}) = y$, then by the monotonicity property it must hold $F_{-T}^{0}(x, \mathbf{u}) = y\,\forall x \in S$. Hence there is no need to consider all trajectories starting in all $|S|$ possible states; but two will be sufficient. See Propp and Wilson (1996) for further details on implementations and applications.

Following the original paper, subsequent works have alleviated many of the initial restrictive assumptions (Wilson, 2000; Murdoch, 2000; Breyer and Roberts, 2002). Despite the powerful ideas and exciting prospects introduced by Propp and Wilson (1996), couplings for general chains can typically be constructed only for Markov chains that are π-irreducible (Glynn and Rhee, 2014, Section 2) and require huge amount of computations

or significant structure within the Markov chain itself.

1.2.2 Unbiased estimation via random truncation

Although stochastic simulation from the posterior distribution exploiting CFTP is generally not feasible as made clear in the previous section, it turns out to be very commonly feasible and practical to obtain unbiased estimates for the expectation of any arbitrary functional of the distribution of interest (Glynn and Rhee, 2014; Agapiou et al., 2018) enjoying, besides unbiasedness, central limit theorems type results relating the goodness of the estimator to the computational cost, see (1.7) below. Such unbiased estimators were originally developed for two main settings: the simulation of stochastic differential equations (SDEs) (Rhee and Glynn, 2015) and the study of ergodic Markov chains targeting unknown posterior distributions. Suppose we are interested in simulating an unbiased estimator of the expectation of a real valued random variable X (or a function $f(X)$), whose simulation cost is prohibitively high. Direct Monte Carlo simulation is infeasible and hence one must resort to approximations X_i of X (later we will suppose that the computational cost for generating each X_i is increasing in i) for $i \geq 0$. Assume that the approximations X_i satisfy $\mathbb{E}[X_i] \to \mathbb{E}[X]$ as i grows to infinity, and let

$$Z := \sum_{i=0}^{N} \frac{\Delta_i}{Pr(N \geq i)}, \tag{1.3}$$

where $(\Delta_i)_{i\geq 1}$ is a sequence of random variables for which $\mathbb{E}[\Delta_i] := X_i - X_{i-1}$ and

$$\mathbb{E} \sum_{k=0}^{\infty} [|\Delta_k|] < \infty, \tag{1.4}$$

and N is an integer valued random variable independent of X_i for every i and s.t. $Pr(N \geq i) > 0$. Then Z is unbiased (Glynn, 1983), indeed

$$\mathbb{E}[Z] = \mathbb{E}\left[\sum_{i=0}^{N} \frac{\Delta_i I_{N\geq i}}{\Pr(N \geq i)}\right] = \sum_{i=0}^{\infty} \mathbb{E}[\Delta_i] = \mathbb{E}[X].$$

Clearly, in order for the estimator to be of practical use, one need its variance as well as its expected cost to be finite. Let t_i be the expected incremental effort required to

calculate X_i, then the expected work required to generate a value for Z is

$$\mathbb{E}[cost] = \mathbb{E}\left[\sum_{i=1}^{N} t_i\right] = \sum_{i=0}^{\infty} t_i \Pr(N \geq i), \tag{1.5}$$

and Rhee (2013) showed

$$var(Z) = \sum_{i=0}^{\infty} \frac{\beta_i}{Pr(N \geq i)}, \tag{1.6}$$

with $\beta_i = O(E[(X - X_i)^2])$. From (1.5) and (1.6) it is apparent that particular care must be used in order to carefully tune the rate at which $P(N \geq i)$ goes to zero as i increases, since a too slow decay will result in infinite expected cost, while a too fast one will give high variance. A sufficient condition for obtaining finite variance unbiased estimators is contained in Proposition 1.

Proposition 1 (Proposition 6, Rhee (2013))**.** *Let* $\{\Delta_i\}_{i\geq 0}$ *and* N *as in* (1.3) *and* (1.4)*. Assume that*

$$\sum_{i \leq l} \frac{\|\Delta_i\|_2 \|\Delta_l\|_2}{\Pr(N \geq i)} < \infty,$$

where $\|\Delta_i\| = \mathbb{E}[\Delta_i^2]^{\frac{1}{2}}$*. Then* $X_n := \sum_{i=0}^{n} \Delta_i \to X = \sum_{i=0}^{\infty} \Delta_i$ *in* $\mathcal{L}^2$*. Let* $\alpha = \mathbb{E}[X]$ *and suppose that for all* i*,* $\tilde{\Delta}_i$ *is a copy of* Δ_i *such that* $(\tilde{\Delta}_i)_{i\geq 1}$ *are mutually independent. Then* $\tilde{Z} := \sum_{i=0}^{N} \frac{\tilde{\Delta}_i}{P(N\geq i)}$ *is an unbiased estimator of* α *with finite second moment*

$$\mathbb{E}[\tilde{Z}^2] = \sum_{i=0}^{\infty} \frac{\tilde{\nu}_i}{Pr(N \geq i)},$$

where $\tilde{\nu}_i = var(\Delta_i) + (\alpha - \mathbb{E}[X_{i-1}])^2 - (\alpha - \mathbb{E}[X_i])^2$*.*

Note that it is fundamental for Proposition 1 that the $(\tilde{\Delta}_i)_{i\geq 1}$ are mutually independent, hence, if the $(X_i)_{i\geq 1}$ are successive draws of a Markov chain, one must compute each Δ_i from an independent run of the whole chain up to iteration i. Under the assumptions of finite variance and cost, a central limit theorem applies

$$c^{\frac{1}{2}}(\hat{\alpha}(c) - \mathbb{E}[X]) \to \sqrt{\mathbb{E}[cost] var(Z)} N(0, 1), \tag{1.7}$$

where $\hat{\alpha}(c)$ is the Monte Carlo estimator produced from independent replicates of Z that can be generated in c units of computer time, matching the optimal Monte Carlo rate.

In the context of SDEs, the whole procedure described before is directly applicable to many popular discretization schemes, provided that the $E[(X - X_i)^2]$ converges quick enough, (see (1.6)), with the notable exception of the Euler Maruyama (Bally and Talay, 1996). As for MCMC, the obvious choice of $\Delta_i = f(X_i) - f(X_{i-1})$, fails to satisfy the $\mathcal{L}^2$ convergence condition, hence the key is to find a purposely designed coupling between X_i and X_{i-1} that forces Δ_i to go to zero sufficiently fast. Glynn and Rhee (2014) proposed for uniformly recurrent, contractive and Harris Markov chains an appropriate coupling able to turn the convergence in distribution of the forward chain into the $\mathcal{L}^2$ convergence of the backward process, and then apply the methodology above. See Glynn and Rhee (2014) and Agapiou et al. (2018)(Section 1.3) for a practical implementation of the method. Agapiou et al. (2018) extended the construction of Glynn and Rhee (2014); Rhee and Glynn (2015), to cover the case that only a simulatable contracting coupling between runs of the chain started at different states is available, provided this coupling drives the two chains towards each other quickly enough in expectations. In particular their extension can also be applied without representing the chains through iterated random functions.

1.2.3 Unbiased estimation via coalescence

We now present the renowned work of Jacob et al. (2020), in a sense building upon the literature of Section 1.2.1 and 1.2.2. We are interested in approximating expectations of the form

$$\mathbb{E}_\pi[h] = \int_{\mathcal{X}} h(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}),$$

where $\pi \in \mathcal{P}(\mathcal{X})$ is the target probability measure and $h : \mathcal{X} \mapsto \mathbb{R}$ a measurable function. Following the work of Glynn and Rhee (2014) and Jacob et al. (2020), we consider unbiased estimators of $\mathbb{E}_\pi[h]$ based on a coupled pair of Markov chains that marginally evolve according to a common π-invariant transition kernel P. Let $(\mathbf{X}^t, \mathbf{Y}^t)_{t\geq 0}$ be a Markov chain on $\mathcal{X} \times \mathcal{X}$ with coupled kernel $\bar{P}[P]$, such that:

(A1) $\mathbf{X}^0$ marginally starts from a distribution $\pi_0 P$ and $\mathbf{Y}^0 \sim \pi_0$, for π_0 a distribution on $\mathcal{X}$, both evolve according to a transition kernel P and are such that $\lim_{t\to+\infty} \mathbb{E}[h(\mathbf{X}^t)] = \lim_{t\to+\infty} \mathbb{E}[h(\mathbf{Y}^t)] = \pi(h)$. Furthermore, there exist constants $\eta > 0$ and $D < +\infty$ such that $\mathbb{E}[|h(\mathbf{X}^t)|^{2+\eta}] < D$ for all $t \geq 0$.

(A2) The two chains must meet after finite time, i.e. if we define $T = \min\{t \geq 0 \mid \mathbf{X}^t = \mathbf{Y}^t\}$ it must hold $\Pr(T < \infty) = 1$, and in particular $\Pr(T > n) \leq Kn^{-\kappa}$ for some constants $0 < K < \infty$ and $\kappa > 2(2\eta^{-1} + 1)$, where η as in (A1) (Middleton et al., 2020).

(A3) After meeting the two chains stay together, i.e. $\mathbf{X}^t = \mathbf{Y}^t$ for all $t \geq T$.

Note that the initial distribution is taken to be $(\mathbf{X}^0, \mathbf{Y}^0) \sim (\pi_0 P) \otimes \pi_0$ for some π_0, meaning that we initialize $\mathbf{X}^{-1} \sim \pi_0$ and $\mathbf{Y}^0 \sim \pi_0$, with $\mathbf{X}^{-1}$ and $\mathbf{Y}^0$ independent, and $\mathbf{X}^0 | \mathbf{X}^{-1} \sim P(\mathbf{X}^{-1}, \cdot)$.

Under assumptions (A1)-(A2)-(A3) above, the random variable

$$H_k\left((\mathbf{X}^t)_{t\geq 1}, (\mathbf{Y}^t)_{t\geq 1}\right) = h\left(\mathbf{X}^k\right) + \sum_{t=k+1}^{T-1} \left(h(\mathbf{X}^t) - h(\mathbf{Y}^t)\right) \qquad k \geq 0, \tag{1.8}$$

is an unbiased estimator of $\mathbb{E}_\pi[h]$. Note that $H_k(\mathbf{X}, \mathbf{Y}) = h\left(\mathbf{X}^k\right)$ if $k+1 > T-1$. Taking the average of $H_l(\mathbf{X}, \mathbf{Y})$ for $l \in \{k, \ldots, m\}$, where k is a burn-in value and m a maximum number of iterations, leads to the unbiased estimator

$$\begin{aligned} H_{k:m}(\mathbf{X}, \mathbf{Y}) &= \frac{1}{m-k+1} \sum_{l=k}^{m} H_l(\mathbf{X}, \mathbf{Y}) \\ &= \frac{1}{m-k+1} \sum_{l=k}^{m} h(\mathbf{X}^l) + \sum_{l=k+1}^{T-1} \min\left(1, \frac{l-k}{m-k+1}\right) \left(h(\mathbf{X}^t) - h(\mathbf{Y}^t)\right) \quad 0 \leq k < m, \end{aligned}$$

which coincides with the usual MCMC ergodic average estimate plus a bias correction term. Standard guidelines in Jacob et al. (2020) suggest to choose k as a large quantile of the meeting time T and m as a multiple of k, hence for the method to be most practical, the meeting time should occur as early as possible.

Under some regularity assumptions, Jacob et al. (2020) proved that:

$$var[H_{k:m}(\mathbf{X}, \mathbf{Y})] \leq MSE_{k:m} + 2\sqrt{MSE_{k:m}} \frac{C_{\delta,\beta}\delta_\beta^k}{m-k+1} + \frac{C_{\delta,\beta}\delta_\beta^k}{(m-k+1)^2},$$

where $MSE_{k:m}$ is the mean square error of the standard MCMC estimator using draws from iterations k up to m. Therefore, the variance of the estimator is bounded by the mean square error of an MCMC estimator, plus an additive term that vanish geometrically

in k and polynomially in $m-k$. The estimator in (1.8) and in (1.2.3) has been later generalized with a generic lag L in Vanetti and Doucet (2020), showing that increasing L can actually lead to significant variance reduction. Standard guidelines in Jacob et al. (2020) suggest to choose k as a large quantile of the meeting time T and m as a multiple of k. Hence, for the method to be most practical, the meeting time should occur as early as possible.

1.2.4 Couplings for convergence diagnostics

L-lag couplings can be used to generate computable, non-asymptotic upper bound estimates for the total variation or the Wasserstein distance of general Markov chains to their target, without involving model-specific or intractable quantities (Biswas et al., 2019). The method is based on coupled chains $(\mathbf{X}^t, \mathbf{Y}^t)_{t\geq 1}$ as in Section 1.2.3, with the sole difference that $(\mathbf{X}^t)_{t\geq 1}$ is started at $\mathbf{X}^{-L}$ instead of $\mathbf{X}^{-1}$. The method applies to integral probability metric, i.e. distances between probabilities that can be written through differences of expectations, i.e.

Definition 1 (Integral probability metric)**.** *Let H be a class of real valued functions on a measurable space $\mathcal{X}$. For all probability measures $P, Q \in \mathcal{P}(\mathcal{X})$, the corresponding IPM is defined as:*

$$d_H(P, Q) := \sup_{h\in H} |\mathbb{E}_{X\sim P}[h(X)] - \mathbb{E}_{X\sim Q}[h(X)]|.$$

Note that the total variation distance can be obtained as IPM with $H := \{h : \sup_{x\in X} |h(x)| \leq 0.5\}$, and also the 1-Wasserstein distance can be obtained with $H := \{h : |h(x) - h(y)| \leq d(x, y),\, \forall x, y \in \mathcal{X}\}$ (1-Lipschitz functions). Define $\pi_t := P^t\pi_0$ (the marginal distribution of one chain at iteration t) and let π be target distribution as before. If it holds that $\sup_{h\in H} |h(x) - h(y)| \leq M_h(x, y)\, \forall x, y \in \mathcal{X}$ (e.g. $M_h = 1$ for tv), then

Theorem 1 (Biswas et al. (2019))**.** *For an IPM with function set H and upper bound M_H, let $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t\geq 1}$ be Markov chains satisfying (A1)-(A2)-(A3) of Section 1.2.3, then*

for every $L > 0$ *and any* $t \geq 0$*:*

$$d_H(\pi_t, \pi) \leq \mathbb{E}\left[\sum_{j=1}^{\lceil \frac{T-L-t}{L} \rceil} M_H(\boldsymbol{X}^{t+(j-1)L}, \boldsymbol{Y}^{t+(j-1)L})\right],$$

where $\lceil \cdot \rceil$ denotes the ceiling function. Informally, recalling that d_h is a distance (and triangular inequality holds), the proof comes directly from the inequalities

$$d_H(\pi_t, \pi) \leq \sum_{j=1}^{+\infty} d_H(\pi_{t+jL}, \pi_{t+(j-1)L}) \leq \sum_{j=1}^{+\infty} \mathbb{E}\left[M_H\left(\mathbf{X}^{t+jL}, \mathbf{Y}^{t+(j-1)L)}\right)\right].$$

For the total variation case

$$d_{TV}(\pi_t, \pi) \leq \mathbb{E}\left[\max\left(\frac{T-L-t}{L}, 0\right)\right].$$

The method is fairly general, requiring only the ability to generate coupled Markov chains that can meet exactly after a random but finite number of iterations and in principle can be applied to any chain. The tightness of the bound directly depends on the meeting times of the chosen coupling procedure, and hence it becomes even more important to have efficient couplings algorithms yielding small meeting times. A tighter bound was later obtained in Craiu and Meng (2022) minimizing a class of bounds of $d_{TV}(\pi_h, \pi)$ over the choice of control variates, finding

$$d_{TV}(\pi_t, \pi) \leq \sum_{j \geq 1} \min\left(\Pr(\max\left(0, \left\lceil \frac{\tau_L - L - t}{L} \right\rceil\right) < j), \Pr(\max\left(0, \left\lceil \frac{\tau_L - L - t}{L} \right\rceil\right) > j)\right).$$

for $\xi \sim Bern(0.5)$. The bound provided by the coupled chains can be used to obtain guidance on the choice of burn in, compare different MCMC algorithms targeting same distribution and measure mixing times of approximate and exact MCMC methods.

1.2.5 Two-step couplings

Aiming at fast coalescent chains, in this work we consider coupled kernels $\bar{P}$ that follows a two-step strategy: w whenever the chains are "far away" we employ a *contractive* coupling

$\bar{P}^c$ whose aim is to bring the chains closer to each other (see e.g. Section 2.6); when the chains are "close enough" we employ a maximal coupling $\bar{P}^m$ (see e.g. Algorithm 1 or Algorithm 2 in the supplement), which maximizes the probability of the chains being exactly equal at the next step. The resulting kernel $\bar{P}$ takes the form

$$\bar{P}[P]((\mathbf{x},\mathbf{y}),\cdot) = \begin{cases} \bar{P}^c[P]((\mathbf{x},\mathbf{y}),\cdot) & \text{if } d(\mathbf{x},\mathbf{y}) > \varepsilon \\ \bar{P}^m[P]((\mathbf{x},\mathbf{y}),\cdot) & \text{if } d(\mathbf{x},\mathbf{y}) \leq \varepsilon, \end{cases} \tag{1.9}$$

where $d(\mathbf{x},\mathbf{y})$ is a measure of distance between $\mathbf{x}$ and $\mathbf{y}$, such as $d(\mathbf{x},\mathbf{y}) = \|P(\mathbf{x},\cdot) - P(\mathbf{y},\cdot)\|_{TV}$ or $d(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$, and ε is a threshold parameter. Algorithm 5 provides a pseudo-code implementing this strategy.

Algorithm 5: Two-step coupling algorithm

Input: initial distribution π_0, kernels P, $\bar{P}^c$, $\bar{P}^m$
sample $\mathbf{X}^{-1} \sim \pi_0$, $\mathbf{Y}^0 \sim \pi_0$ and $\mathbf{X}^0 \sim P(\mathbf{X}^{-1},\cdot)$ **while** $\boldsymbol{X}^t \neq \boldsymbol{Y}^t$ **do**
  if $d(\boldsymbol{X}^t, \boldsymbol{Y}^t) > \varepsilon$ **then**
    $(\mathbf{X}^{t+1}, \mathbf{Y}^{t+1}) \sim \bar{P}^c[P]((\mathbf{X}^t, \mathbf{Y}^t),\cdot)$
  else
    $(\mathbf{X}^{t+1}, \mathbf{Y}^{t+1}) \sim \bar{P}^m[P]((\mathbf{X}^t, \mathbf{Y}^t),\cdot)$
  $t \leftarrow t+1$
Output: trajectory $(\mathbf{X}^t, \mathbf{Y}^t)_{t\in\{0,\ldots,T\}}$

Two-step couplings have been previously used in the literature, see e.g. Roberts and Rosenthal (2002); Alexandros and Roberts (2005); Bou-Rabee et al. (2020); Biswas et al. (2019). The motivation behind this construction is that *one-step* couplings, which aim for exact chain meeting at each step, are generally suboptimal in terms of meeting times (Griffeath, 1975). The intuitive reason is that high meeting probability and effective contraction are typically separate qualities in couplings: when a maximal coupling fails, preserving marginals might imply sampling distant points in $\mathcal{X}$, thus reducing meeting probability in subsequent steps. Algorithm 5 avoids the previous issue by using $\bar{P}^c$ and $\bar{P}^m$ in order to, respectively, achieve optimal contraction rates and maximal meeting probabilities. In Section 2.6 of Chapter 2 we show how to design $\bar{P}^c$ and $\bar{P}^m$ in the context of Blocked Gibbs samplers for distributions with high degree of independence.

References

Agapiou, S., Roberts, G. O., and Vollmer, S. J. (2018). Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models. *Bernoulli Society for Mathematical Statistics and Probability*.

Alexandros, B. and Roberts, G. O. (2005). One-Shot CFTP, Application to a Class of Truncated Gaussian Densities. *Methodology and Computing in Applied Probability*, 76(1):407–437.

Bally, V. and Talay, D. (1996). The law of the Euler scheme for stochastic differential equations. *Probability Theory and Related Fields*, 104:43–60.

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating Convergence of Markov chains with L-Lag Couplings. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bou-Rabee, N., Eberle, A., and Zimmer, R. (2020). Coupling and convergence for Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 30(3):1209 – 1250.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417.

Breyer, L. A. and Roberts, G. O. (2002). A new method for coupling random fields. *LMS Journal of Computation and Mathematics*, 5:77–94.

Craiu, R. V. and Meng, X.-L. (2022). Double happiness: enhancing the coupled gains of L-lag coupling via control variates. *Statistica Sinica*, 32(4):pp. 1745–1766.

Dowson, D. C. and Landau, B. V. (1982). The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.*, 12(3):450–455.

Eberle, A., Guillin, A., and Zimmer, R. (2019). Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982 – 2010.

Givens, C. R. and Shortt, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231 – 240.

Glynn, P. and Rhee, C.-H. (2014). Exact Estimation for Markov chain equilibrium expectation. *Journal of Applied Probability*, 51A:377–389.

Glynn, P. W. (1983). Randomized estimators for time integrals.

Griffeath, D. (1975). A maximal coupling for Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31(2):95–106.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Johnson, V. E. (1998). A Coupling-Regeneration Scheme for Diagnosing Convergence in Markov Chain Monte Carlo Algorithms. *Journal of the American Statistical Association*, 93(441):238–248.

Lindvall, T. (1992). Lectures on the coupling method.

Middleton, L., Deligiannidis, G., Doucet, A., and Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electronic Journal of Statistics*, 14(2):2842 – 2891.

Murdoch, D. (2000). Exact sampling for bayesian inference: Unbounded state spaces. *Fields Inst Commun*, 26.

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.*, 48:257–263.

Propp, J. G. and Wilson, D. B. (1996). Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics. *Random Structures Algorithms*, 9(1-2):223–252.

Rhee, C.-H. (2013). *Unbiased estimation with biased samplers*. Stanford University.

Rhee, C.-H. and Glynn, P. W. (2015). Unbiased non with square root convergence for sde models. *Operations Research*, 63(5):1026–1043.

Roberts, G. O. and Rosenthal, J. S. (2002). One-shot coupling for certain stochastic recursive sequences. *Stochastic Processes and their Applications*, 99(2):195–208.

Santambrogio, F. (2015). Optimal transport for applied mathematicians.

Thorisson, H. (2000). *Coupling, Stationarity, and Regeneration*. Springer.

Vanetti, P. and Doucet, A. (2020). Discussion of Unbiased MCMC using Couplings by Jacob et al. *Journal of the Royal Statistical Society Series B*, 82:592–593.

Wilson, D. (2000). How to couple from the past using a read-once source of randomness. *Random Structures and Algorithms*, 16.

Chapter 2

Linear-cost unbiased posterior estimates for crossed effects and matrix factorization models

In this chapter we present a methodology for the design of unbiased MCMC procedures of blocked Gibbs samplers (BGS) with linear computational cost in the number of data points under modern asymptotic regimes, building upon the coupling estimators introduced by Jacob et al. (2020). A significant portion of the work presented in this chapter culminated in the publication of Ceriani and Zanella (2024). Specifically, in this chapter we analyze BGS couplings for Gaussian targets by providing explicit bounds on the expected number of iterations needed for coalescence, showing that practical two-step coupling BGS schemes achieve coupling times that match relaxation times up to a logarithmic factor. We further address implementation aspects of couplings of BGS with conditionally independent blocks and provide a novel BGS scheme for probabilistic matrix factorization. To illustrate the practical relevance of our methodology, we apply it to crossed random effect and probabilistic matrix factorization models, furnishing unbiased estimates for unknown quantities at a computational cost linearly proportional to the number of data points, notably matching cost of state-of-the-art procedures.

2.1 Introduction

In recent years, unbiased Markov Chain Monte Carlo via couplings (UMCMC) has emerged as a promising framework to remove bias from MCMC estimates, thus potentially allowing for early stopping, simplifying the convergence diagnostic process and facilitating parallelization (Glynn and Rhee, 2014; Jacob et al., 2020). In UMCMC, coupled chains are run for a random number of iterations (at least up to coalescence) and their values are combined to produce unbiased estimates. A natural question that arises is whether this class of estimates incurs a greater computational cost than conventional MCMC based on simple ergodic averages and to quantify this potential difference. Framing the question differently, one may ask whether it is possible to devise UMCMC methods with computational cost matching top performing MCMCs, while enjoying the above mentioned benefits.

On a different line of research, various works showed how carefully designed blocked Gibbs Samplers (BGSs), i.e. Gibbs sampling schemes that update entire blocks of coordinates jointly, can achieve state-of-the-art performances for sampling from the posterior distributions of various challenging high-dimensional Bayesian models, such as non-nested models with crossed dependencies (Papaspiliopoulos et al., 2019, 2023). In particular, BGSs achieve linear computational costs in the number of parameters and observations in asymptotic regimes where both diverge to infinity.

In this work, we seek to combine these two lines of research, aiming to design UMCMC BGS methods with linear computational cost in the aforementioned high-dimensional regimes. Specifically, we provide a theoretical contribution, i.e. the analysis of BGS couplings for Gaussian targets via explicit bounds on the expected number of iterations, showing that practical two-step BGS coupling schemes achieve coupling times that match relaxation times up to a logarithmic factor; and some methodological ones, discussing implementation aspects of couplings of BGS with conditionally independent blocks and developing a novel BGS scheme for probabilistic matrix factorization which empirically reduces the MCMC complexity to linear for those models. To illustrate the practical relevance of our methodology, we apply it to crossed random effect models (Gelman, 2005; Baayen et al., 2008), a commonly used class of additive models that connect a re-

sponse variable to categorical predictors, and to probabilistic matrix factorization (PMF) models (Salakhutdinov and Mnih, 2008; Miller and Carter, 2020), which can be seen as dimensionality reduction models based on low-rank representations.

The remaining part of the section is organized as follows: after briefly presenting the objectives of the work and the three running examples that motivate our research in Section 2.2, we review how to exploit couplings to obtain unbiased estimates in Section 2.3. The main theoretical results are presented in Section 2.4: we provide bounds on the expected number of iterations needed for coalescence of coupled chains and specialize it for different classes of Markov chains. We apply the methodology and the theoretical results to Gaussian crossed random effect models in Section 2.5, generalized linear mixed models (GLMMs) with crossed effects in Section 2.7 and to PMF models in Section 2.8. Section 2.6 discusses some methodological aspects related to couplings of BGS with conditionally independent blocks. The code to reproduce the simulations reported in this work can be found at https://github.com/paoloceriani/couplings_bgs.

2.2 Motivation and objectives

Many high-dimensional Bayesian models with a high degree of conditional independence are well-suited for BGSs. For these classes of models, BGSs often achieve state-of-the-art performance and in particular, unlike most available sampling schemes, result in a total computational cost scaling linearly in the number of observations and parameters. In this thesis we will consider the following three models as running examples motivating the methodology and theory developed later. Despite their relatively simple formulations, these models are computationally challenging to estimate: correlated errors lead to expensive GLS estimates and strong posterior dependence induced by observations results in slow mixing of standard MCMCs.

Model 1 (Gaussian crossed random effects)**.** *Cross-classified data, where each observation can be simultaneously classified according to two or more variables, are commonly found in the modern scientific literature, with applications in various domains including health and social sciences (Gelman, 2005; Baayen et al., 2008). More in detail, a univariate response variable y is assumed to depend additively on the unknown effects of*

K categorical variables, termed factors, each one with I_k different possible values, called levels, for $k = 1, ..., K$. The effect of the i-th level of the k-th factor is described by a random variable $a_i^{(k)}$, object of inference. Let $\mathbf{a}^{(k)} = (a_1^{(k)}, ..., a_{I_k}^{(k)})$ denote the I_k-dimensional vector of effects of the k-th factor, for $k = 1, ..., K$; $\mathbf{y} = (y_n)_{n=1}^N$, $\mathbf{a} = (\mathbf{a}^{(k)})_{k=1}^K$ and $\boldsymbol{\tau} = (\tau_k)_{k=0}^K$ the vectors of all data, effects and precisions respectively. Furthermore let $i_k[n]$ denote the level of the k-th factor associated to the n-th observation (see e.g. Section 1.1 and Chapter 11 of Gelman and Hill (2006) for more details on this notation). In this thesis, for clarity of exposition, we will consider the intercept-only version, although the concepts discussed extend to more general versions with covariates as well as random slopes (Gao and Owen, 2016; Papaspiliopoulos et al., 2023). The model with its standard prior can then be written as

$$\begin{aligned} y_n|\mu, \mathbf{a}, \tau_0 &\sim N\left(\mu + \sum_{k=1}^{K} a_{i_k[n]}^{(k)}, \tau_0^{-1}\right) && n = 1, ..., N, \\ a_i^{(k)}|\tau_k &\sim N(0, \tau_k^{-1}) && i = 1, ..., I_k,\ k = 1, ..., K, \\ p(\tau_k) &\propto \tau_k^{-0.5} && \text{for } k = 0, ..., K, \\ p(\mu) &\propto 1, \end{aligned} \tag{2.1}$$

where μ is a random intercept. In (2.1), $p(\cdot)$ denotes the density of the random variable inside the brackets and $N(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. We will often consider the model with $K = 2$ factors, where one can think of y_n as the rating that user $i_1[n]$ gave to film $i_2[n]$.

Model 2 (GLMMs with crossed effects)**.** *Generalized linear mixed models (GLMMs) extend the framework of linear mixed models to accommodate non-Gaussian response variables by incorporating a link function, but still retaining the same dependence structure. They are a powerful tool widely used in many academic fields, such as political science, biology and medicine; see e.g. (Wood, 2017, Ch.3) and Jiang and Nguyen (2021). Extending Model 1 to allow for general response gives*

$$\mathcal{L}(y_n|\mu, \mathbf{a}) = \mathcal{L}(y_n|\eta_n) \text{ with } \eta_n = \mu + \sum_{k=1}^{K} a_{i_k[n]}^{(k)} \text{ for } n = 1, ..., N, \tag{2.2}$$

where $\mathcal{L}(\cdot)$ denotes the law of the random variable within brackets. Different choices of the conditional distribution $\mathcal{L}(y_n|\eta_n)$ lead to different models, e.g. $\mathcal{L}(y_n|\eta_n) = N(\eta_n, \tau_0^{-1})$ is equivalent to Model 1 or $\mathcal{L}(y_n|\eta_n) = Bern(p)$ with $p = \frac{1}{1+e^{-\eta_n}}$ leads to a logit model for binary data. In Section 2.7 we will consider the case $\mathcal{L}(y_n|\mu, \mathbf{a}) = Lapl(\eta_n, 1/\sqrt{2})$, where $Lapl(\mu, b)$ denotes the univariate Laplace distribution with mean μ and scale b.

Model 3 (Probabilistic matrix factorization)**.** *Low rank matrix approximation methods provide one of the simplest and most effective approaches to collaborative filtering (Salakhutdinov and Mnih, 2008; Miller and Carter, 2020), i.e. forecasting of users' interests exploiting other users' preferences. As for Models 1 and 2 with $K = 2$, one can think of observation y_n as representing the rating that user $i[n]$ gives to film $j[n]$. Denoting by $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ respectively the d-dimensional latent user-specific and film-specific factors for $I = 1, ..., I_1$ and $j = 1, ..., I_2$, and by $\boldsymbol{u} = (\boldsymbol{u}_i)_{i=1}^{I_1} \in \mathbb{R}^{I_1 \times d}$, $\boldsymbol{v} = (\boldsymbol{v}_j)_{j=1}^{I_2} \in \mathbb{R}^{I_2 \times d}$ their collections, the model can be formulated as*

$$\begin{aligned} y_n|\rho, \boldsymbol{u}, \boldsymbol{v}, \tau_0 &\sim N(\rho \boldsymbol{u}_{i[n]}^\top \boldsymbol{v}_{j[n]}, \tau_0^{-1}) && n = 1, ..., N, \\ \boldsymbol{u}_i, \boldsymbol{v}_j &\sim N(\boldsymbol{0}, 1_d) && i = 1, ..., I_1, \;\; j = 1, ..., I_2, \\ \tau_0 &\sim Gamma(c, d), && \rho^{-\frac{1}{2}} \sim Gamma(a, b), \end{aligned} \tag{2.3}$$

where $Gamma(a, b)$ denotes a Gamma random variable with shape parameter a and scale parameter b, 1_d denotes the d-dimensional identity matrix and ρ indicates a positive quantity acting as scaling factor for the random effects. PMF models can be seen as a multiplicative extension of the models in (2.1) *and* (2.2)*, and are usually more computationally challenging to estimate (due to invariances with respect to rotations, a lower degree of linearity, etc).*

All the models above feature a small number of high-dimensional blocks whose conditional distributions are relatively easy to manage, while the joint distribution is computationally much harder to deal with. This is a common structure arising in Bayesian modelling scenarios (Gelman, 2005), and we expect the discussion below on couplings of BGS to be generally relevant to Bayesian models with sparse conditional independent structure where BGS perform well.

2.2.1 Asymptotic regimes of interest and computational cost

Models 1, 2 and 3 naturally lead to situations where both the number of observations N and parameters $p = O(\sum_{k=1}^{K} I_k)$ are large. We use the notation $(T_n)_{n\in\mathbb{N}} = O(f(n))$ if there exist constants $c, C \in \mathbb{R}$ with $0 < c < C < \infty$ such that $cf(n) \leq T_n \leq Cf(n)$ for all n. In the following we will talk about asymptotic regimes in terms of $N \to \infty$, implicitly assuming that p is a function of N that is also diverging as $N \to \infty$.

Also, the above models are commonly used in *sparse* settings, where a small fraction of the possible combinations of effects are observed, i.e. $N \ll \prod_{k=1}^{K} I_k$. For example, when $K = 2$ one often has $1 \ll p < N \ll I_1 \times I_2$ (see Gao and Owen (2016) for further discussion). Using the analogy of films and ratings for recommender systems, the above means assuming that the number of ratings, users and films is large, but only a small fraction of the film is rated by each user. Depending on the degree of sparsity in the observation design, one could have either $p = O(N)$ or $p/N \to 0$ as $N \to \infty$.

We consider the task of performing posterior inference for the above models using MCMC methods. We are interested in quantifying the computational effort needed for the posterior estimation (both in the MCMC and UMCMC context) as $N \to \infty$. In the (U)MCMC context, the total cost is defined as the product between the cost per iteration and the expected number of iterations for the convergence (coalescence) of the chains. As discussed below, recent works suggest that BGS can achieve state-of-the-art performances of $O(N)$ posterior estimation cost. Our main objective is to assess whether UMCMC methods with the same cost can be devised for this problem, as well as to provide some guidance on how to do so.

Models with crossed dependencies are computationally harder than classical Bayesian hierarchical models with nested structures. For example, even in the Gaussian case (i.e. Model 1), evaluating the marginal likelihood once (e.g. computing $\mathcal{L}(\boldsymbol{y} \mid \boldsymbol{\tau}, \mu)$ for a given $\boldsymbol{\tau}$ and μ) requires the inversion of a $O(\sum_{k=1}^{K} I_k)$-dimensional matrix. Despite the matrix being sparse, the crossed dependence structure leads to a dense Cholesky factor (Pandolfi et al., 2024, Sec.3), and more generally prevents the use of efficient sparse linear algebra tools available for, e.g., nested or spatial hierarchical model, leading to a computational cost of at least $O(N^{3/2})$ for each evaluation (Gao and Owen, 2016; Perry, 2017; Pa-

paspiliopoulos et al., 2023; Menictas et al., 2023). The situation is obviously worse in the non-Gaussian case, where analogous computation involve general $O(\sum_{k=1}^{K} I_k)$-dimensional integrals. On the other hand, the above models lend themselves naturally to block updating schemes, such as BGSs in the sampling context or block coordinate ascent (aka backfitting) for maximum a posteriori (MAP) or generalized least square (GLS) computations. For example, given the conditional independence structure of Model 1, the posterior conditional distribution of $\mathbf{a}^{(k)}$ factorizes as $\mathcal{L}(\mathbf{a}^{(k)}|\mu, \mathbf{a}^{(-k)}, \boldsymbol{\tau}, \mathbf{y}) = \otimes_{i=1}^{I_k} \mathcal{L}(a_i^{(k)}|\mu, \mathbf{a}^{(-k)}, \boldsymbol{\tau}, \mathbf{y})$, where $\otimes$ denotes the product of independent distributions. Thus a BGS with components $\mu, \boldsymbol{a}^{(1)}, ..., \boldsymbol{a}^{(K)}$ and $\boldsymbol{\tau}$, which we will call *vanilla* BGS, can be trivially implemented at $O(N)$ cost per iteration for Model 1. However, such vanilla version can mix slowly. In particular, Gao and Owen (2016) showed that for Model 1 with $K = 2$ factors, known variances and full observation designs, the vanilla BGS requires $O(\sqrt{N})$ to converge, leading to a prohibitive $O(N^{\frac{3}{2}})$ total cost. This follows from the fact that observed values create strong a posteriori dependence between unknown factors. Papaspiliopoulos et al. (2019) proposed a collapsed Gibbs Sampler (see Algorithm 6 below) which preserves the $O(N)$ cost per iteration and converges in $O(1)$ iterations under appropriate assumptions, see also (Papaspiliopoulos et al., 2023, Thm.2). Similar techniques have been employed to develop a *back-fitting* algorithm to perform GLS estimation for an analogue of Model 1 with $O(N)$ cost in Ghosh et al. (2022). A first question of interest that we consider is whether the same computational efficiency can be extended to the UMCMC context, which we answer positively in Section 2.5. The extension to the UMCMC case allows one to stop MCMC runs after few (e.g. around 10, see Section 2.5.3) iterations while still obtaining unbiased estimates, getting closer to what one could do in the GLS case (where backfitting is often reported to converge in few iteration, see e.g. the discussion in Ghosh et al. (2022, Section 7) about comparison between the cost of Bayesian and frequentist computations for those models).

2.3 Background on couplings for estimation and blocked Gibbs samplers

We now provide some background material on UMCMC and BGSs. Specifically, Section 2.3.1 introduces BGS kernels and the corresponding notation, Section 2.3.2 provides an introduction to mixing and relaxation times, while Section 2.3.3 characterizes Gaussian Gibbs samplers and their relaxation times.

2.3.1 Blocked Gibbs Sampler kernels

We now formally define BGS kernels. Let $\mathbf{x} = (\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(K)}) \sim \pi$, with $\pi \in \mathcal{P}(\mathcal{X})$ and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ partitioned in K blocks of dimension I_k for $k = 1, ..., K$, i.e. $\mathbf{x}_{(k)} \in \mathcal{X}_k \subseteq \mathbb{R}^{I_k}$. We indicate by $\mathbf{x}_{(-k)} = (\mathbf{x}_{(j)})_{j \neq k}$ the whole vector except the k-th block and by $\pi\left(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}\right)$ the so-called full conditional distribution of the k-th block. Various BGS variants can be derived depending on the chosen updating order. For example, the (deterministic-scan) *forward* version of BGS iteratively samples from $\pi\left(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}\right)$ for $k = 1, ..., K$ at each iteration. The resulting kernel, which we denote as $P^{(F)}$, can be written as the following composition of K kernels

$$P^{(F)} = P_K \cdots P_1, \tag{2.4}$$

$$P_k(\mathbf{x}, d\mathbf{x}') = \pi(d\mathbf{x}'_{(k)}|\mathbf{x}_{(-k)})\delta_{\mathbf{x}_{(-k)}}(d\mathbf{x}'_{(-k)}) \qquad k = 1, \dots, K, \ \mathbf{x} \in \mathcal{X}\,. \tag{2.5}$$

Other natural updating orders include the backward order, as well as the forward-backward or random-scan versions.

2.3.2 Mixing time and Relaxation time

Mixing time and relaxation time are closely related quantities, both providing interesting insights on convergence properties of Markov chains, see e.g. (Levin and Peres, 2017, Section 12). Informally, the mixing time measures the time required by a Markov chain for the distance to the stationary distribution to be small in some chosen sense. More in detail, given a discrete time Markov chain $(\mathbf{X}^t)_{t \geq 1}$ with state space $\mathcal{X}$ and transition

kernel P, starting distribution μ and target distribution π, for $\mu, \pi \in \mathcal{P}(\mathcal{X})$, the mixing time (with respect to the total variation distance) is defined as

$$T_{mix}(\varepsilon, \mu) = \min\{t \in \mathbb{N} : \|\mu P^t - \pi\|_{TV} < \varepsilon\}.$$

Generally, whenever $\mathcal{X}$ is finite or bounded, the mixing time for a fixed ε can be defined as the supremum over all possible initializations.

We introduce now some preliminary quantities needed for the relaxation time. Let f, g be functions, i.e. $f, g : \mathcal{X} \mapsto \mathbb{R}$, we define the inner product $\langle f, g\rangle_\pi := \int_{\mathcal{X}} f(x)g(x)\pi(dx)$ and the induced norm $\|f\|_\pi = \sqrt{\langle f, f\rangle}_\pi$. Furthermore let $\mathcal{L}^2(\pi) = \{f : \mathcal{X} \mapsto \mathbb{R}, \|f\|_\pi < \infty\}$ and $\mathcal{L}^2_0(\pi) \subset \mathcal{L}^2(\pi)$ s.t. $f \in \mathcal{L}^2_0(\pi) \leftrightarrow \int_{\mathcal{X}} f(x)\,\pi(dx) = 0$. The kernel P can act both on functions and on probability measures. Specifically for $\mu \in \mathcal{P}(\mathcal{X})$ we denote by $\mu P(dy) = \int_{\mathcal{X}} P(x, dy)\,\pi(dx)$ the probability measure resulting from the application of P on μ and for f a function we denote by $Pf(x) = \int_{\mathcal{X}} f(y)\,P(x, dy)$, the expected value of f under the measure $P(x, dy)$. We say P is π-invariant for $\pi \in \mathcal{P}$ if $\pi P = \pi$. In this case it holds that

$$\begin{aligned}
\|Pf\|^2_\pi &= \int_{\mathcal{X}} (Pf(x))^2\,\pi(dx) = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(y)\,P(x, dy)\right)^2 \pi(dx) \\
&\leq \int_{\mathcal{X}}\int_{\mathcal{X}} f^2(y)\,P(x, dy)\,\pi(dx) = \int_{\mathcal{X}} f^2(y) \int_{\mathcal{X}} P(x, dy)\,\pi(dx) \\
&= \int_{\mathcal{X}} f^2(y)\pi(dy) = \|f\|^2_\pi,
\end{aligned}$$

so that P is a bounded linear operator. We say that P is π-reversible if $\pi(dx)P(x, dy) = \pi(dy)P(y, dx)\,\forall x, y \in \mathcal{X}$. It follows the spectrum is real and bounded, so that $\sigma(P) \subset [-1, 1]$, where $\sigma(P) = \{f : \mathcal{X} \mapsto \mathbb{R} \text{ s.t. } P - fI \text{ is not invertible}\}$. Furthermore we have that P is self-adjoint i.e.

$$\begin{aligned}
\langle Pf, g\rangle &= \int_{\mathcal{X}}\int_{\mathcal{X}} f(y)P(x, dy)\,\pi(dx) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} f(y)P(dx, y)\,\pi(dy) = \langle f, Pg\rangle \quad \forall f, g \in \mathcal{L}^2(\pi).
\end{aligned}$$

Note that, if we restrict ourself to the space $\mathcal{L}_0^2(\pi)$ then we can define

$$\lambda_0^{max} := \sup_{f \in \mathcal{L}_\pi^0} \frac{\langle f, Pf \rangle}{\langle f, f \rangle}, \qquad \lambda_0^{min} := \inf_{f \in \mathcal{L}_\pi^0} \frac{\langle f, Pf \rangle}{\langle f, f \rangle}, \tag{2.6}$$

and

$$\bar{\lambda} = \max(|\lambda_0^{max}|, |\lambda_0^{min}|).$$

Let $AbsGap(P) = 1 - \bar{\lambda}$. For $\mu \ll \pi$, denote its Radon-Nikodym derivative as $\frac{d\mu}{d\pi}$ and consider the Chi-square distance between distributions defined as $\chi^2(\mu, \pi) := \| \frac{d\mu}{d\pi} - 1 \|_\pi$, then one can prove

Lemma 3. *Let $\mu, \pi \in \mathcal{P}(\mathcal{X})$ with $\mu \ll \pi$, and P a π-invariant kernel. For all $t \geq 1$ it holds that*

$$\chi^2(\mu P^t, \pi) \leq (1 - AbsGap(P))^n \, \chi^2(\mu, \pi).$$

Proof. From the definitions in (2.6) and Andrieu et al. (2022, Equation (5)) we have

$$\|P^t f_0\|_\pi^2 = \langle f_0, P^{2t} f_0 \rangle \leq \bar{\lambda}^{2t} \|f_0\|_\pi^2. \tag{2.7}$$

Setting $f_0 = \frac{d\mu}{d\pi} - 1$, so that $f_0 \in \mathcal{L}_0^2(\pi)$, by (2.7) we get

$$\begin{aligned} \|P^t f_0\|_\pi^2 = \|P^t \left(\frac{d\mu}{d\pi} - 1 \right) \|_\pi^2 &= \int_{\mathcal{X}} \left(P^t \left(\frac{d\mu}{d\pi} - 1 \right)(x) \right)^2 \pi(dx) \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} P^t(x, dy) \left(\frac{d\mu}{d\pi}(y) - 1 \right) \right)^2 \pi(dx) = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} P^t(x, dy) \left(\frac{d\mu}{d\pi}(y) \right) \right)^2 \pi(dx) - 1. \end{aligned}$$

Furthermore by π-reversibility it holds $\frac{d}{d\pi} \mu P^t = P^t \frac{d\mu}{d\pi}$, hence substituing above one gets

$$\|P^t f_0\|_\pi^2 = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} P^t(dx, y) \left(\frac{d\mu}{d\pi}(x) \right) \right)^2 \pi(dx) - 1 = \left(\chi^2(\mu P^t, \pi) \right)^2,$$

Combining the above with (2.6) we get

$$\chi^2(\mu P^t, \pi) = \|P^t f_0\| \leq (1 - AbsGap(P))^t \, \chi^2(\mu, \pi).$$

□

The relaxation time is defined as

$$T_{rel} := \frac{1}{1 - AbsGap(P)}.$$

In view of the result of Lemma 3 and the definition of T_{rel}, note that if one wants Chi square distance less than a fixed ε threshold, he should run the chains for

$$t \leq 1 + \left(\frac{\ln(\chi^2(\mu, \pi)/\varepsilon)}{\ln(T_{rel})} \right).$$

An explicit connection to the mixing time derive from the known property of total variation and Chi square distance. While total variation distance is always bounded in the interval $[0, 1]$, the Chi-square distance assumes values in $[0, +\infty]$. By Cauchy–Schwarz inequality, the Chi-square distance gives an upper bound for the total variation distance, namely (Khare and Zhou, 2009a, Section 2.1)

$$\|\mu - \nu\|_{TV} \leq \frac{1}{2}\sqrt{\chi^2(\mu, \nu)},$$

and hence

$$T_{mix}(\varepsilon, \mu) \leq 1 + \left(\frac{\ln(\chi^2(\mu, \pi)/4\varepsilon^2)}{\ln(T_{rel})} \right).$$

Another interesting asymptotic characterization is reported in the lemma below.

Lemma 4 (Rosenthal (2003))**.** *Let P be the kernel of a reversible Markov chain, then*

$$\sup_{\mu \in \mathcal{L}^2_\pi, \|\mu\|_\pi < \infty} \lim_{t \to \infty} \frac{1}{t} \log \left(\|\mu P^t - \pi\|_{TV} \right) = \log(AbsGap(P)).$$

2.3.3 Gaussian Gibbs Sampler kernels

Gaussian Gibbs Sampler kernels enjoy an equivalent representation as autoregressive chains that allows for neater convergence results. Let $(\mathbf{X}^t)_{t \geq 1}$ be a Markov chain on $\mathcal{X}$ with transition kernel P and target distribution π divided in K blocks of dimension $I_1, ..., I_K$ respectively, as in Section 2.3.1. Furthermore let π be a $d = I_1 + ... + I_K$ dimensional Gaussian $N(\boldsymbol{\mu}, \Sigma)$. In this case, $(\mathbf{X}^t)_{t \geq 1}$ can equivalently be seen as a Gaussian auto-regression process. We formally state it as a lemma below, where we also provide

an expression for the associated relaxation time, which will be useful later.

Lemma 5. *Let $\pi = N(\boldsymbol{\mu}, \Sigma)$ and P be a BGS kernel with updating order given by $(k_1, \ldots, k_s) \in \{1, \ldots, K\}^s$ for some $s \in \mathbb{N}$, i.e. $P = P_{k_s} \cdots P_{k_1}$ with P_k defined in (2.5). Then, one has*

$$P(\boldsymbol{x}, \cdot) = N\left(B\boldsymbol{x} + \boldsymbol{b}, \Sigma - B\Sigma B^\top\right), \tag{2.8}$$

where B a matrix depending on the updating order $(k_1, \ldots, k_s)$ and on the target precision matrix $Q = \Sigma^{-1}$, and $\mathbf{b} = (I - B)\boldsymbol{\mu}$. Furthermore the relaxation time of P is given by $T_{rel} = 1/(1 - \rho(B))$, where $\rho(B)$ denotes the largest modulus eigenvalue of B.

Lemma 5 is a well-known result, whose proof we omit, see e.g. Roberts and Sahu (1997, Lemma 1). Recall that for π above it holds

$$\pi(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}) = N\left(\sum_{j \neq k} A_{(k,j)}\mathbf{x}_{(j)} + \mathbf{a}_k, Q^{-1}_{(k,k)}\right) \qquad k = 1, ..., K, \tag{2.9}$$

where $A = 1_d - diag(Q^{-1}_{(1,1)}, ..., Q^{-1}_{(K,K)})Q$ and $\mathbf{a}_i = Q^{-1}_{(i,i)} \sum_{j=1}^s Q_{(i,j)}\boldsymbol{\mu}_{(j)}$. Then, given every updating order $(k_1, ..., k_s) \in \{1, ..., K\}^s$ without repetitions, the $(n+1)$-th iteration of a Gibbs update has the form

$$\begin{aligned}
\mathbf{X}^{n+1}_{(k_1)} &= \sum_{j=2}^{s} A_{(k_1,k_j)}\mathbf{X}^n_{(k_j)} + \mathbf{a}_{k_1} + Q^{-1}_{(k_1,k_1)}\mathbf{Z}_{(k_1)} \\
\mathbf{X}^{n+1}_{(k_2)} &= A_{(k_2,k_1)}\mathbf{X}^{n+1}_{(k_1)} + \sum_{j=3}^{S} A_{(k_2,k_j)}\mathbf{X}^n_{(k_j)} + \mathbf{a}_{k_2} + Q^{-1}_{(k_2,k_2)}\mathbf{Z}_{(k_2)} \\
&... \\
\mathbf{X}^{n+1}_{(k_s)} &= \sum_{j=1}^{s-1} A_{(k_s,k_j)}\mathbf{X}^{n+1}_{(k_j)} + \mathbf{a}_{k_s} + Q^{-1}_{(k_s,k_s)}\mathbf{Z}_{(k_s))}.
\end{aligned}$$

Then, solving the recursion it is possible to write explicitly the matrix B and the vector $\mathbf{b}$, see e.g. the proof of Remark 4 for a three-blocks BGS. Lemma 1 of Roberts and Sahu (1997) provides an explicit expression of B in the forward updating case, corresponding to $s = K$, $k_i = i$ for all i and $P = P^{(F)}$, namely

$$B = (1_d - L)^{-1}U, \tag{2.10}$$

where U and L are respectively upper and lower triangular matrices such that $U+L=A$, for A as in (2.9). Consider Model 1 with fixed variances. In this case it is possible to explicitly compute B for many updating orders, see e.g. the proof of Proposition 3 in Papaspiliopoulos et al. (2019) or of Proposition 2 in the Appendix of Papaspiliopoulos et al. (2023). With the same notation of Section 2.2, let $n_{ij}^{(k,l)}$ denote the number of observations with level i on the k-th factor and level j on the l-th factor, i.e. $n_{ij}^{(k,l)} = \sum_{n=1}^{N} \mathbb{I}(i_k[n]=i)\mathbb{I}(i_l[n]=j)$, let $n_i^{(k)} = \sum_j n_{ij}^{(k,l)}$, i.e. the number of observation with level i of factor k, for $i=1,\dots,I_k$, $j=1,...,I_l$ and $k,l=1,...,K$. Let C be the co-occurance matrix organized in blocks according to the factors, with $C_{ij}^{(k,l)} = n_{ij}^{(k,l)}, C_{ii}^{(k,k)} = n_i^{(k)}$ and $C_{ij}^{(k,k)} = 0$ for $i \neq j$. Furthermore assume that for each k we have $n_i^{(k)} = N/I_k$. With the priors in (2.1), the posterior distribution of the random vector $(\mu, \mathbf{a}^{(1)}, ..., \mathbf{a}^{(K)})$ is Gaussian with precision matrix of the form

$$
\begin{aligned}
&Q^{(0,0)} = N\tau_0, \qquad Q_i^{(0,k)} = \tau_0 n_i^{(k)} && \text{for } i=1,...,I_k,\ k=1,...,K,\\
&Q^{(k,k)} = \tau_k 1_{I_k} + \tau_0 C^{(k,k)}, \quad Q^{(k,l)} = \tau_0 C^{(k,l)} && \text{for } k,l=1,...,K, k\neq l
\end{aligned}
$$

On the other hand, for the collapsed algorithm (Algorithm 6), the posterior precision matrix is the one obtained by marginalizing out, i.e. collapsing, the global mean μ, hence leading to a posterior matrix of the form

$$
\begin{aligned}
&Q_{coll}^{(k,k)} = \tau_k 1_{I_k} + \tau_0 C^{(k,k)} - \frac{\tau_0}{N}\mathbf{n}^{(k)}\left(\mathbf{n}^{(k)}\right)^\top && \text{for } k=1,...,K,\\
&Q_{coll}^{(k,l)} = \tau_k C^{(k,l)} - \frac{\tau_0}{N}\mathbf{n}^{(k)}\left(\mathbf{n}^{(l)}\right)^\top && \text{for } k,l=1,...,K,\ k\neq l,
\end{aligned}
$$

where $\mathbf{n}^{(k)} = (n_1^{(k)}, ..., n_{I_k}^{(k)})$. It is then possible to exploit (2.10) to get the expressions of B for both plain vanilla and collapsed schemes and compute their spectral radii, or resort to characterizations as those in Lemma 1 and Lemma 2 of Papaspiliopoulos et al. (2019).

2.4 Bounds for couplings of Gaussian Gibbs Samplers

In this section we provide bounds on the meeting times of BGS coupled via Algorithm 5 when the target distribution is Gaussian. As discussed later in Section 2.4.2, we seek to obtain UMCMC schemes whose coupling times T are of the same order of (or not much greater than) the relaxation times of the original kernel P.

Algorithmic details 1. *For all the theoretical results of this section, we consider Algorithm 5 with* $\bar{P}^m[P]$ *being the* maximal reflection coupling *reported in Algorithm 2 in the supplement,* $\bar{P}^c[P]$ *being the common random numbers (*crn*) coupling reported in Lemma 2 and* (2.65) *in the supplement, and* $d(\boldsymbol{x}, \boldsymbol{y}) = \|P(\boldsymbol{x}, \cdot) - P(\boldsymbol{y}, \cdot)\|_{TV}$. *We defer more discussion on general specifications and implementations of Algorithm 5 to Section 2.6.*

2.4.1 Bound for reversible chains

Our first bound applies to π-reversible BGS kernels, i.e. one where the updating order satisfies $(k_1, \dots, k_s) = (k_s, \dots, k_1)$. A classical example is the forward-backward kernel, defined as

$$P^{(FB)} = P_1 \cdots P_{K-1} P_K P_{K-1} \cdots P_1 \,. \tag{2.11}$$

Algorithmically, $P^{(FB)}$ performs updates from $\pi(\mathbf{x}_{(k)} \mid \mathbf{x}_{(-k)})$ sequentially for $k = 1, 2, \dots, K-1, K, K-1, \dots, 2, 1$. If P is a π-reversible BGS kernel, it holds $\Sigma B^\top = B\Sigma$, with B as in Lemma 5 (see e.g. Proposition 4.27 of Khare and Zhou (2009b)). This allows for neater theoretical results. In Section 2.4.4 we extend the result to non-reversible Gibbs samplers, such as those generated by the forward kernel $P^{(F)}$ in (2.4), where the result requires additional technical assumptions.

Theorem 2 (Bound for reversible chains)**.** *Let* $\pi = N(\boldsymbol{\mu}, \Sigma)$ *and* $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t \geq 0}$ *be a Markov chain marginally evolving with* π*-invariant BGS kernel and coupled via Algorithm 5 with*

Algorithmic specification 1. Then $T := \min\{t \geq 0 : \boldsymbol{X}^t = \boldsymbol{Y}^t\}$ *satisfies*

$$\mathbb{E}[T|\boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 4 + \frac{1}{-\ln(\rho(B))}\left[-\frac{1}{2}\ln(1-\lambda_{min}(B)^2) + C_0 + C_\varepsilon\right], \tag{2.12}$$

where λ_{min} *denotes the minimum norm of the eigenvalues,* B *is as in* (2.8), $C_0 := \ln(\|L^{-1}(\boldsymbol{X}^0 - \boldsymbol{Y}^0)\|)$ *with* L *such that* $LL^\top = \Sigma$, *and* $C_\varepsilon \leq 6\, erf^{-1}(\varepsilon) - \ln(erf^{-1}(\varepsilon))$, *for* erf^{-1} *the inverse error function and* $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t$.

Remark 1. *Both the distribution of* T *and the bound in* (2.12) *are invariant under block diagonal linear transformations that preserve the* K*-partite block structure of* $(\boldsymbol{x}_{(1)}, ..., \boldsymbol{x}_{(K)})$. *See Section 2.12.1 in the Supplement for a more detailed statement and proof.*

Remark 2. *We report in Figure 2.1 the behaviour of* C_ε *and its bound for* $\varepsilon \in (0, 1)$. *As visible in Figure 2.1, the bound degenerates to infinity for* ε *close to 0 and 1. On*

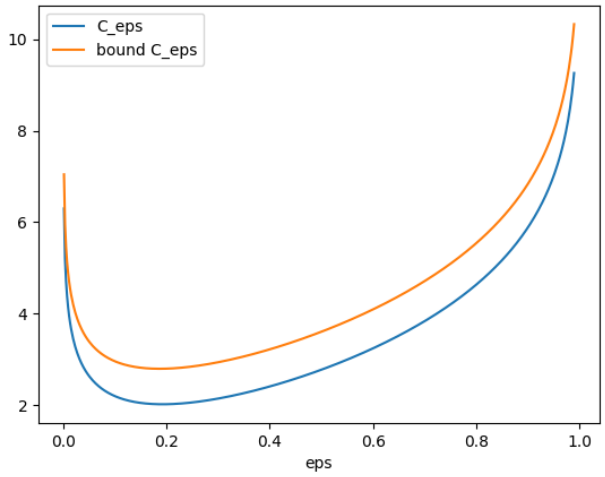

Figure 2.1: C_ε and its bound for $\varepsilon \in (0, 1)$.

the one hand, as ε *goes to* 0, *the two chains must be brought closer and closer before attempting coalescence, hence the meeting times and our bound increase to infinity. On the other hand, as* ε *approaches 1, the coupling strategy resembles the "one step coupling", i.e. coalescence is attempted no matter the distance between chains, but the bounding technique of Theorem 2 cannot be applied since the bound for the second part of* T *in* (2.29) *becomes too lose. A promising approach is generalizing the results presented in Section E of the appendix of Douc et al. (2024) for univariate autoregressive processes.*

For the proof of Theorem 2 we exploited the following bound for the expected squared distance between Gaussian distributions coupled via reflection maximal coupling which can be of independent interest.

Lemma 6. *Let $p = N(\boldsymbol{\xi}, \Sigma)$ and $q = N(\boldsymbol{\nu}, \Sigma)$ be d-dimensional Gaussians, and $(\boldsymbol{X}, \boldsymbol{Y}) \in \Gamma_{max}(p, q)$ coupled via maximal reflection coupling (see e.g. Algorithm 2). If $\|\Sigma^{-1/2}(\boldsymbol{\xi} - \boldsymbol{\nu})\|^2 \leq 1$, then for every $A \in \mathbb{R}^{d\times d}$ it holds*

$$\mathbb{E}[\|A(\mathbf{X} - \mathbf{Y})\|^2 | \mathbf{X} \neq \mathbf{Y}] \leq \frac{\|A(\boldsymbol{\xi} - \boldsymbol{\nu})\|^2}{\|\Sigma^{-1/2}(\boldsymbol{\xi} - \boldsymbol{\nu})\|^4} \left(12 + 8\sqrt{\frac{2}{\pi}}\right).$$

Note that, for fixed Σ and A, the bound in Lemma 6 scales as $O(\|\boldsymbol{\xi} - \boldsymbol{\nu}\|^{-2})$ as $\|\boldsymbol{\xi} - \boldsymbol{\nu}\| \to 0$, which can be easily checked to be the correct rate for $d = 1$.

2.4.2 Connection to relaxation times

Combining Theorem 2 with Lemma 5 leads to the following result.

Corollary 1. *Under the same assumptions of Theorem 2, we have*

$$\mathbb{E}[T | \boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 4 + T_{rel}\left[\frac{1}{2}\ln(T_{rel}) + C_0 + C_\varepsilon\right], \tag{2.13}$$

where $T_{rel} := 1/(1 - \rho(B^{(FB)}))$.

Proof of Corollary 1. The result follows from (2.12), noting that $\frac{1}{-\ln(x)} < \frac{1}{1-x}$, $-\ln(1 - x^2) < \ln(1/(1 - x))$ for $x \in (0, 1)$ and the latter is monotonically increasing in x, hence allowing to substitute $\lambda_{min}(B^{(FB)})$ with the quantity $\rho(B^{(FB)})$, greater by definition. Then since $T_{rel} = 1/(1 - \rho(B^{(FB)}))$ by Lemma 5 we get the result. □

Corollary 1 provides interesting insights and implications. In particular, interpreting T_{rel} as the number of iterations required for $\mathbf{X}^t$ to converge, it suggests that in this context UMCMC provides unbiased estimates with an average number of iterations (and an overall computational cost) that is comparable to the minimal number of iterations required by standard MCMC to converge (up to a logarithmic factor). Also, from an high-dimensional asymptotics perspective, it also implies that whenever the relaxation

time of BGS is bounded as the number of data point and parameter grows (see e.g. Section 2.5.2), then also the meeting time is bounded in expectation, meaning that UMCMC does not increase the overall complexity (while allowing for e.g. early stopping and parallelization). On the other hand, (2.13) implies that whenever the meeting times of the *two-step* strategy diverges for a chosen BGS scheme, also the respective T_{rel} diverges.

One could interpret Corollary 1 as a best-case result for UMCMC. The underlying assumption would be that obtaining coupling times that are of smaller order than T_{rel} is typically unfeasible. While we are not aware of rigorous results in this direction (i.e. showing that $\mathbb{E}[T]$ cannot be much smaller T_{rel} under appropriate conditions), this seems reasonable to assume given that, for example, T_{rel} provides lower bounds on total variation mixing times (Levin and Peres, 2017, Section 12) and that the quantiles of T can be used to derive non-asymptotic upper bounds on those (see e.g. equation (4) in (Biswas et al., 2019)). While interesting, we leave a more detailed and rigorous exploration of lower bounds to $\mathbb{E}[T]$ to future works.

2.4.3 Bound for two-block Gibbs samplers

We now consider the two-block case. In this context, it is possible to find a block-diagonal transformation, in the spirit of Remark 1, which allows for a direct extension of the bound in Theorem 2 also to the non-reversible kernel $P^{(F)}$.

Theorem 3. *Let $\pi = N(\boldsymbol{\mu}, \Sigma)$, $K = 2$ and $\left(\boldsymbol{X}^t, \boldsymbol{Y}^t\right)_{t\geq 0}$ be a Markov chain marginally evolving with $P^{(F)}$, coupled via Algorithm 5 with Algorithmic specification 1. Then*

$$\mathbb{E}[T|\boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 5 + T_{rel}\left[C_0 + C_\varepsilon\right], \tag{2.14}$$

where $T_{rel} = 1/(1-\rho(B^{(F)}))$, $B^{(F)}$ as in (2.8), and (T, C_0, C_ε) as in Theorem 2.

Note that in contrast to Corollary 1, the bound has the same order of magnitude of T_{rel} without additional logarithmic terms.

2.4.4 Bound non-reversible case Gibbs samplers

We have the following result for the case of BGS with general updating order.

Theorem 4. *Let $\pi = N(\boldsymbol{\mu}, \Sigma)$ and $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t\geq 0}$ be a Markov chain evolving with kernel P as in (2.8), coupled via Algorithm 5, with Algorithmic specification 1. For any $\delta > 0$, it holds that*

$$\mathbb{E}[T|\boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 4 + 3\max\left(n^*_\delta, (1+\delta)T_{rel}\left[-\frac{1}{2}\ln(1-\lambda_{\min}(NN^\top)) + C_0 + C_\varepsilon\right]\right), \quad (2.15)$$

with $T_{rel} = 1/(1-\rho(B))$, $N = L^{-1}BL$, $(T, C_0, C_\varepsilon, L)$ as in Theorem 2, and

$$n^*_\delta := \inf\left\{n_0 \geq 1 : \forall n \geq n_0 \;\; 1 - \|N^n\|_2^{\frac{1}{n}} \geq \frac{1-\rho(N)}{1+\delta}\right\}.$$

The bound in (2.15) features the additional term n^*_δ. The reason for it is that, due to the non-reversibility of P, N is generally not symmetric. Thus $\|N^n\|^{\frac{1}{n}} \to \rho(N)$ from above (Gelfand, 1941), but in general $\|N^n\|^{\frac{1}{n}} \neq \rho(N)$ for finite n. In order to make the result in (2.15) fully informative, like the ones in the reversible and two-block cases, one would need to provide an explicit bound on n^*_δ for the given matrix N under consideration. In all our numerical experiments we observed n^*_δ to be smaller than the second term and never the leading term of the bound, and we expect it to be well-behaved in our contexts of interest. On the other hand, we are not aware of general tight bounds for n^*_δ and we thus left it as an explicit term in the bound.

Remark 3. *Since the focus of this section is primarily to provide explicit and sharp bounds on $\mathbb{E}[T]$ for Gaussian BGS schemes, we did not specifically address issues related to the regularity assumptions necessary for the validity and finite variance of the unbiased estimator. However, we expect that the proofs of this section can be extended quite naturally to control higher moments of T, e.g. proving $\mathbb{E}[T^k] < +\infty$ for $k > 1$. In view of Theorem 2.1 in Atchadé and Jacob (2024), this would imply finite variance of the unbiased estimator.*

2.5 Application to Gaussian crossed random effect models

In this section we combine the findings of Section 2.4 with existing results on Model 1. We highlight that all the theoretical results we will derive hold under the assumption of fixed $\boldsymbol{\tau}$ in (2.1) of Model 1. We first describe in Section 2.5.1 state-of-the-art marginal algorithms for Model 1, then present in Section 2.5.2 the resulting bound for meeting times if a two-step coupling is implemented and finally report numerical simulations in Section 2.5.3.

2.5.1 Collapsed Gibbs sampler for Model 1

Despite the favourable cost per iteration of the vanilla Gibbs sampler for Model 1 presented in Section 2.2.1, there are many settings of interest where its mixing is provably poor, often leading to a super linear overall computational cost. Papaspiliopoulos et al. (2019) noted that integrating out the global mean μ while updating the remaining regression parameters in K blocks, leads to a much more efficient (i.e faster mixing) updating scheme, while preserving the same $O(N)$ cost per iteration of vanilla BGS. The resulting algorithm is called *collapsed Gibbs sampler* (Papaspiliopoulos et al., 2019) and reported in Algorithm 6: at every iteration we first sample from $\mathcal{L}(\mu|\mathbf{a}^{(-k)}, \boldsymbol{\tau}, \mathbf{y})$ and then iteratively update the factor effect block from $\mathcal{L}(\mathbf{a}^{(k)}|\mu, \mathbf{a}^{(-k)}, \boldsymbol{\tau}, \mathbf{y})$, repeating the procedure for $k = 1, .., K$. The exact form of the full conditionals is reported in Section 2.11.1 of the supplement.

Algorithm 6: One iteration of the collapsed Gibbs sampler for Model 1

for *k= 1,...,K* **do**
  draw $\mu \sim \mathcal{L}(\mu|\mathbf{a}_{(-k)}, \boldsymbol{\tau}, \mathbf{y})$
  for i=1, ..., I_k **do**
    draw $a_i^{(k)} \sim \mathcal{L}(a_i^{(k)}|\mathbf{a}_{(-k)}, \mu, \boldsymbol{\tau}, \mathbf{y})$
  draw $\tau_k \sim \mathcal{L}(\tau_k|\mathbf{a}, \mu, \boldsymbol{\tau}_{-k}, \mathbf{y})$
draw $\tau_0 \sim \mathcal{L}(\tau_0|\mathbf{a}, \mu, \boldsymbol{\tau}_{-0}, \mathbf{y})$

2.5.2 Bound on meeting times under random design assumptions

For Model 1 with $K = 2$ factors, balanced level designs (i.e. the same number of observations is observed for every level of each factor) and fixed $\boldsymbol{\tau}$, Papaspiliopoulos et al. (2019) show that the relaxation time of the collapsed algorithm, denoted by T_{cg}, is upper bounded by $T_{cg} \leq C\, T_{aux}$, where $C = 1 + \frac{\tau_0}{\min\{\tau_1,\tau_2\}}$ is constant with respect to N and p, and T_{aux} is the relaxation time of the auxiliary two-block Gibbs sampler on the discrete space $\{1, ..., I_1\} \times \{1, ..., I_2\}$ with invariant distribution $\Pr((i, j)) = n_{ij}/N$, where $n_{ij} = \sum_{n=1}^{N} \mathbb{I}(i_1[n] = i)\mathbb{I}(i_2[n] = j)$ denotes the number of observations of level i of factor 1 and j of factor 2. Under random design assumptions, the quantity T_{aux} can be bounded using random graph theory results, as done in Papaspiliopoulos et al. (2023). In particular, for N multiple of d_1 and d_2, denote by $\mathcal{D}(N, d_1, d_2)$ the collection of all the possible observation patterns with exactly N observations, $I_1 = N/d_1$, $I_2 = N/d_2$ and binary balanced levels (i.e. there must be exactly d_1 and d_2 observations for each level of factor 1 and 2 respectively and $n_{ij} \in \{0, 1\}$ for all $i = 1, ..., I_1$ and $j = 1, ...I_2$). Then, supposing uniformly at random designs among $\mathcal{D}(n, d_1, d_2)$ with $d_1, d_2 > 4$, one has

$$T_{aux} \leq 1 + \frac{2}{\sqrt{\min\{d_1, d_2\} - 2}} + \gamma,$$

asymptotically almost surely as $N \to +\infty$, for every $\gamma > 0$. The result follows from relating T_{aux} to the spectrum of a random bipartite bi-regular graph, and then applying an extension of the Friedman's second largest eigenvalue theorem to bipartite graphs developed in Brito et al. (2018). Combining the above with Theorem 3, we obtain the following bound for the expected meeting times.

Corollary 2. *Let $(\boldsymbol{X}^t, \boldsymbol{Y}^t)_{t\geq 0}$ be as in Theorem 3, where P is the collapsed Gibbs kernel (Algorithm 6) and let $\pi = N(\boldsymbol{\mu}, \Sigma)$ be the posterior distribution of Model 1 with $K = 2$ factors, fixed $\boldsymbol{\tau}$ and design $(n_{ij})_{i,j}$ picked uniformly at random from $\mathcal{D}(n, d_1, d_2)$. Then*

$$\Pr\left(\mathbb{E}[T|\boldsymbol{X}^0, \boldsymbol{Y}^0] \leq 5 + \left(1 + \frac{\tau_0}{\min\{\tau_1, \tau_2\}}\right)\left(1 + \frac{2}{\sqrt{\min\{d_1, d_2\} - 2}} + \gamma\right)[C_0 + C_\varepsilon]\right) \to 1,$$

as $N \to +\infty$, with (T, C_ε, C_0) as in Theorem 3, and the probability is with respect to the randomness of the design.

Interestingly, Corollary 2 provides an upper bound on the average coupling time that remains bounded as N (and p) diverge. In the next section we explore numerically the tightness of the bound, and its robustness to the specific assumptions used to derive it.

2.5.3 Numerics

We compare the bounds of Theorems 2, 3 and 4 with the average meeting times of simulated coupled chains for both the vanilla and collapsed Gibbs samplers of Sections 2.2.1 and 2.5.1, for synthetic and real data in Section 2.5.3 and 2.5.3 respectively. Although the bounds are valid only for Gaussian chains (hence for Model 1 with fixed $\boldsymbol{\tau}$), we compare them with the average meeting times of coupled chains with known (i.e. fixed) as well as unknown (i.e. assigning to it a prior and including it into the Bayesian model) $\boldsymbol{\tau}$, yielding similar behaviours. The results support the intuition that, for the models under consideration, the convergence properties of the known and unknown variances case are very similar and thus the bounds are reasonably predictive also of the behaviour in the practically-used unknown variance case.

Algorithmic details 2. *We implement the two-step coupling procedure of Algorithm 5 for both the vanilla and the collapsed Gibbs sampler of Algorithm 6, with the standard priors in* (2.1)*. More precisely, we use the block-wise coupling reported* (2.16) *and* (2.17) *(see Section 2.6 for further details), with maximal reflection coupling of Algorithm 2 as $\bar{P}_{max}[P_k]$ whenever implementable (e.g. Gaussian with same variance), or Algorithm 1 otherwise; $\bar{P}_{W_2}[P_k]$ is the* crn *coupling of Lemma 2. The distance in Algorithm 5 is set to $d(\boldsymbol{X}^t, \boldsymbol{Y}^t) = \|\boldsymbol{X}^t - \boldsymbol{Y}^t\|$ and the threshold parameter $\varepsilon = O((K\,I)^{-1})$.*

Remark 4. *The bounds of Theorems 2, 3, 4 are derived for strategies implementing $\bar{P}_{max}[P]$ and $\bar{P}_{W_2}[P]$, respectively maximal reflection and W_2 optimal couplings of the kernel P in* (2.8)*. For ease of implementation, in this section we run simulations with successive composition of $\bar{P}_{max}[P_k]$ and $\bar{P}_{W_2}[P_k]$for $k = 1, ..., K$ as in* (2.16) *and* (2.17)*, hence obtaining in principle worse performing couplings. We highlight that for Gaussian*

distributions (2.17) *is actually equivalent to the* W_2 *optimal and* $\bar{P}_{max}[P]$ *is implementable with a computational cost of the same order of* (2.16)*. A detailed proof of the equivalence of the costs is deferred to Section 2.12.4 of the supplement.*

Simulated data

We simulate data according to Model 1, with $\tau_0 = \tau_1 = ... = \tau_k = 1$, $I_1 = ... = I_K = I$ for fixed I, and different number of factors K. Observations are generated according to two different asymptotic regimes, with completely missing at random designs.

Asymptotic regime 1. *Each combination of factor levels is either observed once or not with probability* $p = 0.1$ *independently from the rest, i.e.* $n_{ij}^{(s,l)} \overset{iid}{\sim} Bern(p)$ *for* $i = 1, \ldots, I_s$, $j = 1, \ldots, I_l$ *and* $s \neq l \in \{1, ..., K\}$*, where* $n_{ij}^{(s,l)} = \sum_{n=1}^{N} \mathbb{I}(i_s[n] = i)\mathbb{I}(i_l[n] = j)$ *denotes the number of observations of level* i *of factor* s *and* j *of factor* l*. In this regime, one has* $I = O(N^{1/K})$*.*

Asymptotic regime 2. *Same as Regime 1 but with* $p = 10/I^{K-1}$*. This regime induces more sparsity, e.g. one has* $I = O(N)$*.*

We plot the average of the meeting times as a function of the total number of parameters of the model, i.e. $1 + KI$ plus the number of scale parameters if any. Figure 2.2 reports results for the collapsed Gibbs sampler coupled with Algorithmic specification 2, for $K = 2$, $I_1 = I_2 = I \in \{50, 100, 250, 500, 1000\}$ levels, fixed and free variances. The bound of Theorem 3 for the two blocks collapsed Gibbs sampler with fixed variances (using the true generating values) is also reported. The left and right panel corresponds, respectively, to Regime 1 and Regime 2. The results yield remarkably low meeting times and highlight a close resemblance of the meeting time behaviour with that of the bound. As a comparison, we report in Figure 2.3 the results for the vanilla algorithm on the same model. As expected, the provably higher relaxation time of the vanilla Gibbs scheme results in an higher bound and, more importantly, higher on average meeting times of the coupled chains.

In Figure 2.4 we report the average meeting times for $K = 3$ (left) and $K = 4$ (right) factors, only for Regime 2, and the bound of Theorem 4. For these models the relaxation time is not computable explicitly even under the usual simplifying assumptions (fixed

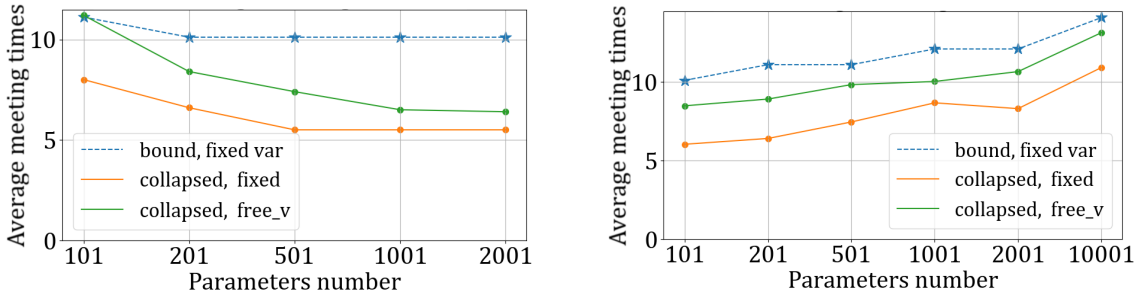

Figure 2.2: Estimated meeting times and bounds for $K = 2$, $I_1 = I_2 = I \in \{50, 100, 250, 500, 1000\}$, $\tau_0 = \tau_1 = \tau_2 = 1$ for Algorithm 6. Left: Regime 1, right: Regime 2.

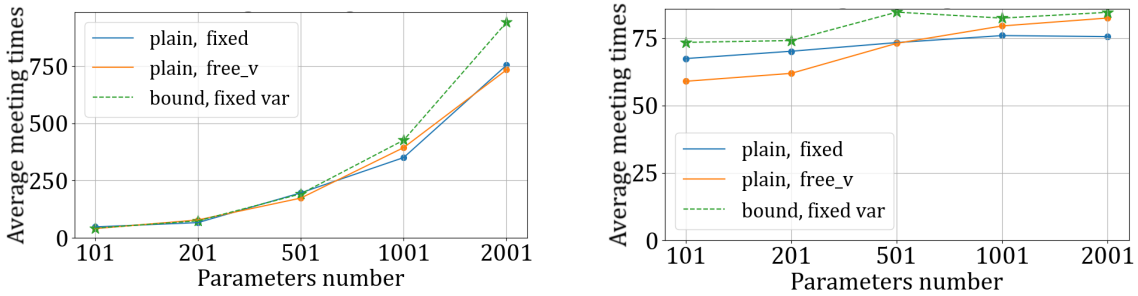

Figure 2.3: Estimated meeting times and bounds for $K = 2$, $I_1 = I_2 = I \in \{50, 100, 250, 500, 1000\}$, $\tau_0 = \tau_1 = \tau_2 = 1$, for plain vanilla algorithm. Left: Regime 1, right: Regime 2.

variances and balanced levels or cells), see e.g. Papaspiliopoulos et al. (2019). Thus proving scalability of the meeting times (or lack thereof), in light of Theorem 4, provide interesting insights on the mixing properties of the single chains themselves.

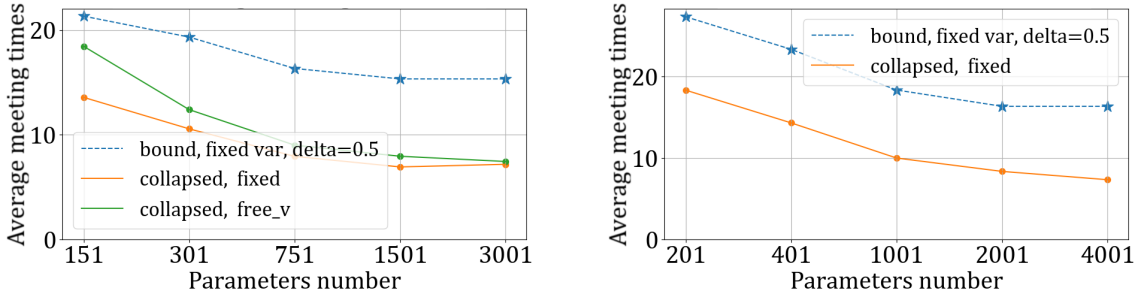

Figure 2.4: Estimated mean number of iterations and bounds for $K = 3$ (left), $K = 4$ (right), $I_1 = ... = I_4 = I \in \{50, 100, 250, 500, 1000\}$, $\tau_k = 1$ for $k \in \{0, 1, 2, 3, 4\}$. Regime 2, Algorithm 6.

One-step vs two-step couplings

In Figure 2.5 we report the estimated distribution of meeting times for the collapsed Gibbs scheme with Algorithmic specification 2, when traditional or *two-step* coupling is implemented on a synthetic dataset with $K = 3$ factors and Regime 2. As can be

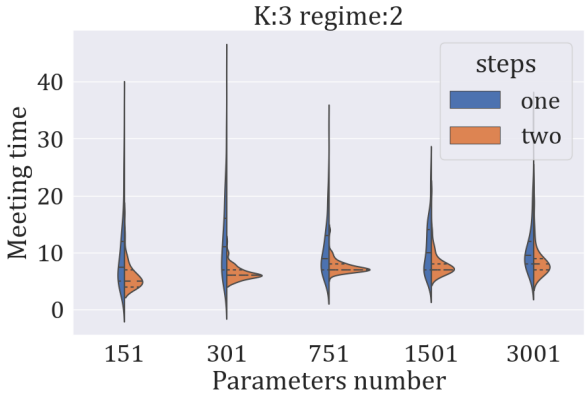

Figure 2.5: Estimated distribution of meeting times for $K = 3$, $I_1 = ... = I_3 = I \in \{50, 100, 250, 500, 1000\}$, $\tau_k = 1$ for $k \in \{0, 1, 2, 3\}$ one vs two-step. Regime 2, Algorithm 6.

seen from the above, the distribution of meeting times for the *two-step* strategy is more concentrated on smaller values with considerably lighter tails, thus supporting the choice of a two-step coupling. On the other hand, we see that, in our context, using one- versus two- step coupling is less influential than, for example, in the one of (Biswas et al., 2022, Fig.3).

Real data example

We now consider a real dataset, containing university lecture evaluations by students at ETH Zurich. The dataset is freely available from the R package **lme4** (Bates et al., 2015) under the name "InstEval". Each observation includes a score ranging from 1 to 5, assigned to a lecture, along with 6 factors that may potentially impact the score, including the identity of the student giving the rating or department that offers the course. Following the notation in (2.1), we have $N = 73421, K = 6$ and $(I_1, ..., I_6) = (2972, 1128, 4, 6, 2, 14)$. We implement the two-step coupling with Algorithmic specification 2. We compute the estimated meeting times for different numbers and combinations of factors for Model 1 with fixed variances (estimated via standard MCMC and plugged

in the coupling procedure). We numerically computed the bound for each combination using the MCMC variance estimates. The results of the experiment are shown in Table 2.1.

Algorithm	Factor number	mean #iter	bound for fixed $\boldsymbol{\tau}$
collapsed	[1,2]	10.1	15
collapsed	[1,6]	9.3	17
vanilla	[1,2]	50.7	73
vanilla	[1,6]	127.6	69

Table 2.1: Average meeting times for InstEval Dataset

The fact that for the vanilla scheme the bound is smaller than the estimated meeting time is not undermining the validity of our theory: as highlighted in Remark 4 of Section 4, the bound is derived for strategies coupling directly the kernel P of (2.8), while for ease of implementation we used (2.16) and (2.17). Furthermore, the bound holds for $d(\mathbf{x}, \mathbf{y}) = \|P(\mathbf{x}, \cdot) - P(\mathbf{y}, \cdot)\|_{TV}$ while we used $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

2.6 Coupling strategies for blocked Gibbs samplers

The current section discusses alternative coupling strategies for BGSs with high degrees of conditional independence, providing some justification to the ones employed in Sections 2.4 and 2.5.3. Specifically, Section 2.6.1 focuses on aspects related to the general structure of Gibbs samplers, namely the necessity to couple successively composed kernels, while Section 2.6.2 focuses on those related to the updates of conditionally independent blocks.

2.6.1 BGS and compositions of couplings

In order to implement Algorithm 5, we need to specify $\bar{P}^c[P]$ and $\bar{P}^m[P]$, where P is a BGS kernel. Since P is defined as a composition of kernels, i.e. $P = P_{k_s} \cdots P_{k_1}$ with $s \in \mathbb{N}$ and $(k_1, \ldots, k_s) \in \{1, \ldots, K\}^s$ as in Lemma 5, a natural strategy is to sequentially compose maximal or optimally contractive couplings of P_{k_i} for $i = 1, \ldots, s$. We denote

the resulting coupling kernels as

$$\bar{P}^{m*}((\mathbf{x},\mathbf{y}),\cdot) := \bar{P}_{max}[P_{k_s}]\cdots\bar{P}_{max}[P_{k_1}]\left((\mathbf{x},\mathbf{y}),\cdot\right) \qquad \forall \mathbf{x},\mathbf{y}\in\mathcal{X}, \tag{2.16}$$

$$\bar{P}^{c*}((\mathbf{x},\mathbf{y}),\cdot) := \bar{P}_{W_2}[P_{k_s}]\cdots\bar{P}_{W_2}[P_{k_1}]\left((\mathbf{x},\mathbf{y}),\cdot\right) \qquad \forall \mathbf{x},\mathbf{y}\in\mathcal{X}, \tag{2.17}$$

where $\bar{P}_{max}[P_k] \in \Gamma_{max}[P_k]$ and $\bar{P}_{W_2}[P_k] \in \Gamma_{W_2}[P_k]$ for all $k = 1, ..., K$. By construction, both $\bar{P}^{m*}((\mathbf{x},\mathbf{y}),\cdot)$ and $\bar{P}^{c*}((\mathbf{x},\mathbf{y}),\cdot)$ belong to $\Gamma[P]$. In all implementations of Algorithm 5, we employ $\bar{P}^{m}[P] = \bar{P}^{m*}$ and $\bar{P}^{c}[P] = \bar{P}^{c*}$. The appeal of $\bar{P}^{m*}$ and $\bar{P}^{c*}$ is that, in order to implement them, one needs to work only with the individual full conditionals involved in the original BGS scheme, which are often available in closed form, while the joint distribution $P(\mathbf{x},\cdot)$ might be harder to work with. Note that these strategies are not guaranteed to be optimal since in general one has $\bar{P}^{m*} \notin \Gamma_{max}[P]$ and $\bar{P}^{c*} \notin \Gamma_{W_2}[P]$. For example, in the case $P = P^{(F)}$, $\bar{P}^{c*}$ coincides with the so-called Knothe-Rosenblatt map (Rosenblatt, 1952; Knothe, 1957) of $P^{(F)}(\mathbf{x},\cdot)$ and $P^{(F)}(\mathbf{y},\cdot)$, which is in general different from the optimal transport one (Santambrogio, 2015, Section 2.3). Nonetheless, we still observe very fast contraction of $\bar{P}^{c*}$ in our numerics, which might be partly explained by the fact that in the Gaussian case one indeed has $\bar{P}^{c*} \in \Gamma_{W_2}[P]$, as shown below.

Lemma 7 (Optimality of composition of W_2 couplings for Gaussians)**.** *Let $\pi = N(\boldsymbol{\mu}, \Sigma)$ and $\bar{P}^{c*}$ as in* (2.17)*, with $s = K$ and $(k_1, \ldots, k_K)$ being a permutation of $(1, \ldots, K)$. Then for all integers $n \geq 1$ it holds $\left(\bar{P}^{c*}\right)^n \in \Gamma_{W_2}[P^n]$.*

The proof of Lemma 7 builds upon well known results about contractive couplings of Gaussian distributions. Firstly we exploit the fact that the optimal transport map between Gaussian distributions whose variance covariance matrices commute is the *crn* coupling (see Lemma 2 in Section 1.1.3 and Dowson and Landau (1982); Olkin and Pukelsheim (1982)). Then, given the autoregressive form of Gaussian Gibbs samplers of Lemma 5, it is possible to write explicitly the W_2 optimal coupling for such kernel (see (2.65) in the supplement). Lastly, it is left to prove that such coupling is indeed equivalent to $\bar{P}^{c*}$.

2.6.2 Couplings of conditionally independent blocks

We now discuss how to implement $\bar{P}_{max}[P_k]$ and $\bar{P}_{W_2}[P_k]$ in cases where the associated full conditional factorizes as

$$\pi\left(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}\right) = \otimes_{i=1}^{I_K}\pi\left(x_{(k),i}|\mathbf{x}_{(-k)}\right), \tag{2.18}$$

for $x_{(k),i}$ denoting the i-th component of the vector $\mathbf{x}_{(k)}$, i.e. $\mathbf{x}_{(k)} = (x_{(k),1}, ..., x_{(k),I_k})$, and I_k is large.

By (2.18), independently sampling from the univariate distributions $\pi\left(x_{(k),i}|\mathbf{x}_{(-k)}\right)$ is equivalent to sampling directly from the entire block $\pi\left(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}\right)$. The same intuition extends to W_2 optimal couplings but not to maximal ones. In particular, one has that the product of independent W_2-optimal couplings of $\pi\left(x_{(k),i}|\mathbf{x}_{(-k)}\right)$ for $i = 1, \ldots, I_k$ is W_2-optimal for $\pi\left(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}\right)$ while the same is not true for maximal ones. In particular, when p and q are two product measures, one has the following well-known facts, which we collect in a lemma whose proof we omit for brevity.

Lemma 8. *Let $p, q \in \mathcal{P}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_d)$ with $p = \bigotimes_{i=1}^{d} p_i$ and $q = \bigotimes_{i=1}^{d} q_i$. Then $\mu_i \in \Gamma_{W_2}(p_i, q_i)$ for all $i = 1, \ldots, d$ implies $(\otimes_{i=1}^{d}\mu_i) \in \Gamma_{W_2}(p, q)$. On the contrary, $\mu_i \in \Gamma_{max}(p_i, q_i)$ for all $i = 1, \ldots, d$ does not imply $(\otimes_{i=1}^{d}\mu_i) \in \Gamma_{max}(p, q)$ in general. In particular, one has*

$$\min_{i=1,\ldots,d} \mathrm{Pr}_{max}(p_i, q_i) \geq \mathrm{Pr}_{max}(p, q) \geq \prod_{i=1}^{d} \mathrm{Pr}_{max}(p_i, q_i), \tag{2.19}$$

where we use the notation $\mathrm{Pr}_{max}(p, q) := 1 - \|p - q\|_{TV}$, and all the inequalities can be strict.

Lemma 8 implies that, under (2.18), we can simply take a contractive coupling $\bar{P}_{W_2}[P_k]$ which factorizes across coordinates. On the contrary, joint maximal couplings of $\pi\left(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}\right)$ do not factorize across coordinates under (2.18). The lower bound $\mathrm{Pr}_{max}(p, q) \geq \prod_{i=1}^{d} \mathrm{Pr}_{max}(p_i, q_i)$ in (2.19) implies that setting $\|p_i, q_i\|_{TV} = O(d^{-1})$ ensures $\mathrm{Pr}_{max}(p, q)$ is bounded away from 0. Lemma 9 quantifies the tightness of such lower bound for d-dimensional Gaussian distributions with same variance covariance matrix.

We consider the regime where d goes to infinity and the distance between each rescaled mean decreases with d, which is arguably descriptive of what happens when using the two-step algorithm (Algorithm 5) of Section 1.2.5 in high dimensions.

Lemma 9. *Consider* $p = N(\boldsymbol{\mu}, diag(\boldsymbol{\sigma}))$ *and* $q = N(\boldsymbol{\nu}, diag(\boldsymbol{\sigma}))$, *d-dimensional Gaussian distribution with* $\boldsymbol{\sigma} = (\sigma_1^2, ..., \sigma_d^2)$ *such that* $\frac{\mu_i - \nu_i}{\sigma_i} = c_i d^{-\alpha}, i = 1, ..., d$, *with* $0 < \inf_i |c_i| \leq \sup_i |c_i| < +\infty$ *and* $\alpha > 0$, *then:*

$$\Pr_{max}(p, q) \asymp \begin{cases} \frac{d^{\alpha - \frac{1}{2}}}{\sqrt{\pi}\bar{c}_d} \exp\left(-\frac{\bar{c}_d^2}{\sqrt{2}} d^{-2\alpha+1}\right) & \text{for } 0 < \alpha \leq \frac{1}{2} \\ 1 - \frac{2\bar{c}_d}{\sqrt{\pi}} d^{-\alpha+\frac{1}{2}} & \text{for } \alpha > \frac{1}{2} \end{cases} \qquad \text{as } d \to \infty,$$

$$\prod_{i=1}^{d} \Pr_{max}(p_i, q_i) \asymp \exp(-d^{1-\alpha}\tilde{c}_d) \qquad \text{as } d \to \infty,$$

where $\bar{c}_d := \sqrt{\frac{\sum_{i=1}^d c_i^2}{8d}}$, $\tilde{c}_d = \frac{\sum_{i=1}^d |c_i|}{d\sqrt{2\pi}}$ *and we write* $f(x) \asymp g(x)$ *whenever* $\lim_{x \to +\infty} \frac{f(x)}{g(x)} = 1$

It follows that both probabilities go to zero for $0 < \alpha < 0.5$ as $d \to +\infty$, while for $0.5 < \alpha < 1$ only $\Pr_{max}(p, q)$ goes to 1. For $\alpha > 1$, both probabilities converge to 1 (although with different regimes). In Figure 2.6 we report the ratio of the blocked and the component-wise meeting probabilities, i.e. $\Pr_{max}(p, q)/(\prod_{i=1}^d \Pr_{max}(p_i, q_i))$, for a d-dimensional Gaussian distribution with independent components, where $c_i = 1$ for all i and different values of α, along with a dotted line representing the value 1. Figure 2.6

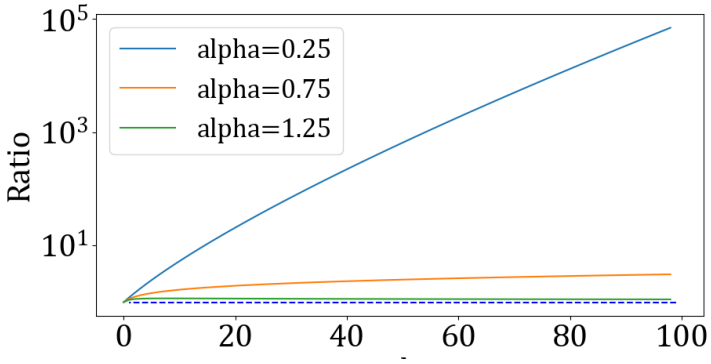

Figure 2.6: Ratio of blocked and component-wise meeting probability for d-dimensional Gaussian, different α values. Dimension on x-axis, logarithmic scale on y-axis.

shows that for $\alpha > 1$, the blocked maximal coupling has meeting probabilities comparable to that of the independent counterpart.

2.7 High-dimensional GLMMs with crossed effects

We now consider applications to Model 2. First, Section 2.7.1 reviews state-of-the-art samplers and their computational cost for this class of models, and briefly discusses our coupling strategy, which requires to extend some of the methodologies of Section 2.6 to the case of Metropolis-within-Gibbs algorithms. Then Section 2.7.2 reports experimental results on simulated data.

2.7.1 Algorithms for Model 2

Similarly to what seen for Model 1 in Section 2.2.1, also for Model 2 the vanilla Gibbs procedure, i.e. the one updating from the full conditionals $\mu, \mathbf{a}^{(1)}, ..., \mathbf{a}^{(K)}, \boldsymbol{\tau}$, suffers from slow mixing because of the posterior correlation among the unknown random effects. Given the impossibility of analytically integrating out the global mean, Papaspiliopoulos et al. (2023) propose to perform at each iteration a $\mathcal{L}\left(\mu, \mathbf{a}^{(k)} | \mathbf{y}, \boldsymbol{\tau}, \mathbf{a}^{(-k)}\right)$-invariant update using local centering within each block, hence updating a new pair of variables $(\mu, \boldsymbol{\xi}^{(k)})$, where $\boldsymbol{\xi}^{(k)} := \mu + \mathbf{a}^{(k)}$. For the re-parametrized model it holds that $\mathcal{L}(\mu | \mathbf{a}^{(-k)}, \boldsymbol{\xi}^{(k)}, \boldsymbol{\tau}, \mathbf{y}) = \mathcal{L}(\mu | \boldsymbol{\xi}^{(k)})$ and that the $\boldsymbol{\xi}^{(k)}$ are conditionally independent, although their full conditional might not be available in closed form. In such cases one can replace exact Gibbs updates of $\boldsymbol{\xi}^{(k)}$ with more general invariant Markov updates. The resulting scheme is described in Algorithm 7.

Algorithm 7: One iteration of Metropolis-within-Gibbs sampler with local centering for Model 2

for *k= 1,...,K* **do**
- reparametrize $(\mu, \mathbf{a}^{(k)}) \to (\mu, \boldsymbol{\xi}^{(k)})$
- draw μ from $\mathcal{L}(\mu | \boldsymbol{\xi}^{(k)})$
- **for** i=1, ..., I_k **do**
  - update $\xi_i^{(k)}$ with a $\mathcal{L}(\xi_i^{(k)} | \mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}, \mathbf{y})$-invariant Markov kernel
- reparametrize $(\mu, \boldsymbol{\xi}^{(k)}) \to (\mu, \mathbf{a}^{(k)})$
- draw τ_k from $\mathcal{L}(\tau_k | \mu, \mathbf{a}, \mathbf{y})$

For Gaussian targets and fixed $\boldsymbol{\tau}$, Corollary 1 in Papaspiliopoulos et al. (2023) shows that

$$T_{cg} < T_{lc} < T_{cg} + C, \tag{2.20}$$

where T_{cg} and T_{lc} denote the relaxation times of Algorithms 6 and Algorithm 7, respectively, and C is a constant depending only on $\boldsymbol{\tau}$. The previous inequality allows to directly extend the results developed in Section 2.5.2 for the collapsed Gibbs scheme to the local centering version in Algorithm 7. Although the inequality holds only for Gaussian targets, numerical results in Papaspiliopoulos et al. (2023) show that also in the non conjugate case, where sampling of $\xi_i^{(k)}$ is done through Metropolis-Hastings updates, the convergence speed remains bounded as N and the number of parameters increase.

Remark 5 (Couplings of Metropolis-Hastings)**.** *Analogously to* (2.18)*,* $\mathcal{L}(\boldsymbol{\xi}^{(k)}|\mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}, \mathbf{y})$ *factorizes as* $\prod_{i=1}^{I_k} \mathcal{L}(\xi_i^{(k)}|\mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}, \mathbf{y})$*. The difference is that* $\mathcal{L}(\xi_i^{(k)}|\mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}, \mathbf{y})$ *are not available in closed form and thus we update each* $\xi_i^{(k)}$ *with a MH step in Algorithm 7. Similarly to Section 2.6.2, leveraging conditional independence in the coupling of the MH kernels allows to have meeting times that grow at most logarithmically with* I_k*, see Section 2.10 in the supplement for details. In the MH case, however, there is additional flexibility (Wang et al., 2021), such as deciding how to couple both the proposal and acceptance steps as well as which of those to factorize. In Section 2.10 in the supplement, we discuss these aspects in some detail, suggesting to use a fully factorized strategy both on the proposal and acceptance. Also, due to the computational difficulty of efficiently implementing* W_2*-optimal couplings of MH, we avoid the two step strategy of Algorithm 5 and rather use a one-step approach for those updates, see again Section 2.10 in the supplement.*

2.7.2 Numerical results

We apply the methodology discussed above to Model 2 and compare its performances with state-of-the-art black-box sampling algorithms such as the No-U-Turn-Sampler (NUTS) (Hoffman and Gelman, 2014) implemented in the popular software STAN (Carpenter et al., 2017). Note that the latter approach does not specifically use the structure of Model 2 and thus might be expected to be sub-optimal for that reason.

We simulate data from Model 2 with $K = 2$ factors, for Laplace response $\mathcal{L}(y_n|\mu, \mathbf{a}) = Lapl(\mu + a^{(1)}_{i_1[n]} + a^{(2)}_{i_2[n]}, 1)$ and $\tau_1 = \tau_2 = 1$, for Regime 2 of Section 2.5.3. We implement the Metropolis-within-Gibbs (MwG) scheme of Algorithm 7 with Random Walk Metropolis

(RWM) updates. As for the coupling, in light of the results of Section 2.10 in the supplement, we implement Algorithm 9, with Gaussian proposals $Q(\mathbf{x}, d\mathbf{x}')$ coupled via maximal independent coupling (Algorithm 1), and paired acceptance. We report below the graph of the resulting average of the meeting times for different numbers of RWM steps within each iteration ($S = 1, 3$). It can be noted from Figure 2.7 that the average of the meeting

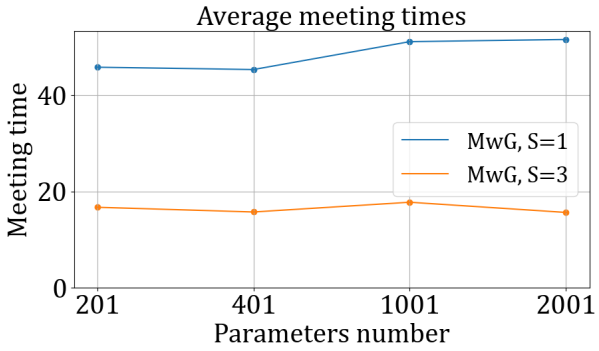

Figure 2.7: Estimated mean meeting times for $K = 2$, $I_1 = I_2 = I \in \{100, 200, 500, 1000\}$, $\tau_1 = \tau_2 = 1, b = 1$ with Laplace response. Algorithm 7 with different number of Metropolis steps S.

times remains almost constant as the number of parameters grows. Furthermore, the higher the number of Metropolis steps S within each iteration, the smaller the meeting time. Intuitively, as S grows, the conditional updates get closer to exact Gibbs updates and the resulting chain converges faster, leading to smaller meeting times. To put into context the performances of our estimation procedure, we illustrate in Figure 2.8 the convergence speed of the STAN implementation (Carpenter et al., 2017) of NUTS with the default setting, for estimating the same model. Specifically, we report the average number of gradient evaluations per Effective Sample Size (ESS), considering the minimum ESS across parameters. As Figure 2.7 and Figure 2.8 show, the $O(1)$ scaling of the meeting times of the MwG sampler with local centering as N and p diverge results in a $O(1)$ number of full likelihood evaluations per unbiased estimate, while a black-box method such as NUTS requires a number of gradient evaluations growing with p.

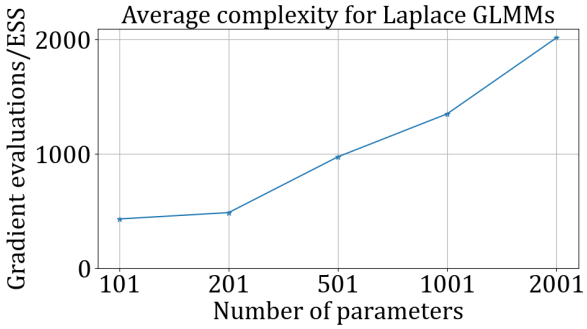

Figure 2.8: Average number of gradient evaluations divided by ESS (minimum across parameters) for NUTS, for $K = 2$, $I_1 = I_2 = I \in \{50, 100, 250, 500, 1000\}$, $\tau_1 = \tau_2 = 1$, $b = 1$ with Laplace response.

2.8 Probabilistic matrix factorization

Finally, we consider Model 3. A well-known feature of such model is that, for fixed $\mathbf{u}$, the model reduces to a standard linear regression with coefficients $\mathbf{v}$, and viceversa for fixed $\mathbf{v}$. On the contrary, the likelihood is not analytically tractable if $\mathbf{u}$ and $\mathbf{v}$ vary jointly. As mentioned in Section 2.2, this structure naturally lends itself to conditional updating schemes, such as BGS for sampling or block coordinate ascent (usually called alternating least squares whenever applied to Model 3) for optimization. However, the vanilla BGS scheme, which updates $\mathbf{u}, \mathbf{v}, \rho, \boldsymbol{\tau}$ from their full conditionals at every iteration, often results in slow mixing of the chain, for reasons analogous to the ones discussed for Models 1 and 2 in previous sections (see also numerics in Figure 2.9). We thus propose a "local centering" version of BGS for Model 3, where we reparametrise the random effects using ρ, i.e. at each iteration we update $(\rho, \bar{\mathbf{u}})$, with $\bar{\mathbf{u}} := \rho\mathbf{u}$, and then $(\rho, \bar{\mathbf{v}})$, for $\bar{\mathbf{v}} := \rho\mathbf{v}$ analogously. Algorithm 8 provides high-level pseudo-code for one iteration of the resulting scheme, while Algorithm 10 in the supplement provides full implementation details.

Algorithm 8: One iteration of BGS with local centering for Model 3

for $(\boldsymbol{r}, \boldsymbol{s}) \in \{(\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{v}, \boldsymbol{u})\}$ **do**
- reparametrize $(\rho, \mathbf{r}) \to (\rho, \bar{\boldsymbol{r}})$, with $\bar{\boldsymbol{r}} := \rho\boldsymbol{r}$
- draw $\rho \sim \mathcal{L}(\rho|\bar{\boldsymbol{r}}, \boldsymbol{s}, \tau_0, \mathbf{y})$
- draw $\bar{\boldsymbol{r}} \sim \mathcal{L}(\bar{\boldsymbol{r}}|\rho, \boldsymbol{s}, \tau_0, \mathbf{y})$
- reparametrize $(\rho, \bar{\boldsymbol{r}}) \to (\rho, \boldsymbol{r})$

draw $\tau_0 \sim \mathcal{L}(\tau_0|\rho, \mathbf{u}, \mathbf{v}, \mathbf{y})$

Regarding the UMCMC version of Algorithm 8, since the high-dimensional full conditionals involved are multivariate Gaussian, we can implement the same coupling strategy as for Model 1, namely Algorithmic specification 2. In particular, joint maximal couplings for the high-dimensional updates of $\bar{\boldsymbol{r}}$ can be implemented efficiently. Below we provide numerical illustrations of the performances of the UMCMC version of Algorithm 8 and vanilla BGS. As discussed in more details in Remark 7, we restrict ourselves to the case $d = 1$.

Remark 6 (Related literature on Bayesian factor models)**.** *Model 3 is closely related to Bayesian factor analysis. With the same notation as in* (2.3)*, a factor model (Gorsuch, 2014) can be written as*

$$y_n|\mu, \boldsymbol{F}, \Lambda, \boldsymbol{\tau} \sim N(\mu_{i[n]} + \Lambda_{j[n],:}\boldsymbol{F}_{i[n]}, \tau_0^{-1}), \tag{2.21}$$

for $\boldsymbol{\mu} \in \mathbb{R}^{I_1}$, $\boldsymbol{F} = (\boldsymbol{F}_i)_{i=1}^{I_1}$ being the collection of unknown factors, $\boldsymbol{F}_i \in \mathbb{R}^d$, and $\Lambda \in \mathbb{R}^{I_2 \times d}$ the factor loading matrix, with d being the latent dimension. Indeed, factor models exhibit the same structure of Model 3, and BGS schemes are also widely used in that context (Conti et al., 2014; Papastamoulis and Ntzoufras, 2022), even if there the focus is usually on the full design case (where all the combinations users/films are observed) and on regimes where $I_2 \ll I_1$ and I_1 grows.

Remark 7 (Issues related to rotational invariance)**.** *A well-known issue of Model 3 is the invariance with respect to joint rotations of $\boldsymbol{u}$ and $\boldsymbol{v}$. This creates multimodality in the posterior, thus inducing slow convergence and lack of posterior interpretability. Many ad hoc methodologies have been developed to deal with such issue, including constraining a priori the matrix of factor loadings or post-processing (Conti et al., 2014; Papastamoulis and Ntzoufras, 2022). Although of interest, these issues and techniques are somehow orthogonal to our focus here, and thus we restrict to the case $d = 1$ for simplicity and leave further exploration to future work.*

2.8.1 Numerical results

We simulate data coming from Model 3 for different asymptotic regimes and parameter specifications. We consider $I_1 = I_2 = I \in \{100, 200, 500, 1000\}$ levels and data coming from Regime 1 and 2. In Figure 2.9 we report the average meeting times for both vanilla BGS (PV in the legend) and Algorithm 8 (Local Centering) with Algorithmic specification 2 as discussed above. As expected, the slow mixing of vanilla BGS results in exploding

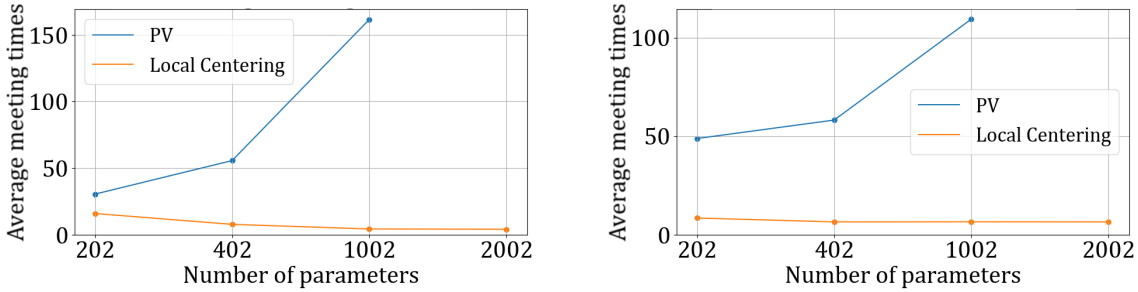

Figure 2.9: Estimated mean meeting times for Probabilistic Matrix Factorization model. $I_1 = I_2 = I \in \{100, 20, 500, 1000\}$. Left: Regime 1, right: Regime 2.

meeting times of the coupled chains, which are not reported for I greater than 500 for visual convenience. Remarkably, as few as 10 iterations are on average sufficient for the coupled chains evolving according to Algorithm 8 to meet even in the high-dimensional cases.

2.9 Discussion

Building on the recent advancements presented in Jacob et al. (2020) and Papaspiliopoulos et al. (2019, 2023), we propose a UMCMC procedure specifically tailored for high-dimensional BGS with high degree of conditional independence. In the Gaussian case, we show the resulting expected meeting times are bounded by a quantity that depends on the relaxation time of the individual chains, multiplied by its logarithm. In several applications, this results in procedures that require as few as a dozen iterations to obtain unbiased estimates for complex models with over 10,000 parameters, thus enhancing the appeal of efficient and cost-effective parallelization.

Interestingly, unlike many sampling algorithms (such as gradient-based ones), no

adaptation of tuning parameters is required for BGS. This couples particularly well with the UMCMC framework: specifically, it avoids the need for potentially long preliminary runs or adaptation phases, thus genuinely allowing for parallelizable short runs and "early stopping" in case of fast mixing chains, see also Biswas et al. (2022) for analogous examples.

We conclude by mentioning some possible directions for future research. First, while challenging, it would be valuable to extend the theoretical results of Section 2.4 beyond the Gaussian case. A possible approach would be to leverage the recent results on BGS for log-concave distributions in Ascolani et al. (2024). More generally, it is relevant to assess which coupling methodologies lead to meeting times that are of the same order of mixing or relaxation times of the original chain, similarly to Theorems 2, 3 and 4. As discussed in Remark 3, we expect the results in Section 2.4 to extend relatively directly to control higher moments of T, thus implying finite variance of the unbiased estimator (Atchadé and Jacob, 2024, Thm.2.1). Also, it would be interesting to derive lower bounds for the average meeting times of coupled chains explicitly depending on the convergence properties of the original chain, in order to make the arguments of Section 2.4.2 rigorous. Finally, on a different line of research, an interesting direction could be to provide a broader analysis and development of local centering schemes for probabilistic matrix factorization (PMF).

Annex

2.10 Couplings for Metropolis-Hastings algorithms for product targets

In this section we discuss procedures for efficient coupling of Metropolis-Hastings (MH) kernels for targets with independent components. The motivation for such a construction stems from Algorithm 7 applied to Model 2, where each iteration consists of updating K blocks of (conditionally) independent coordinates, since each $\mathcal{L}(\boldsymbol{\xi}^{(k)}|\mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}, \mathbf{y})$ factorizes in $\prod_{i=1}^{I_k} \mathcal{L}(\xi_i^{(k)}|\mu, \mathbf{a}^{-(k)}, \boldsymbol{\tau}, \mathbf{y})$ for $k = 1, ..., K$, whose distribution might be known only up to constants. Exploiting such independence in the coupling construction, one can derive coupling strategies whose meeting times grows logarithmically with I_k, see below. Previous works on couplings for MH kernels include Johnson (1998), where the author first proposed to use maximal couplings on proposal distributions for Random Walk Metropolis, Wang et al. (2021), where among other things the authors suggest employing a maximal reflection coupling on Gaussian proposals and paired acceptance, and Papp and Sherlock (2022), where the authors focus on asymptotically optimally contractive couplings.

Consider a target distribution ν on $\mathcal{X} = \mathbb{R}^{I_k}$ with independent components, i.e. $\nu = \otimes_{i=1}^{I_k} \nu_i$. The general Metropolis kernel targeting ν has the form

$$P_b^{MH}(\mathbf{x}, d\mathbf{x}') = Q_b(\mathbf{x}, d\mathbf{x}')a_b(\mathbf{x}, \mathbf{x}') + \delta_{\mathbf{x}}(d\mathbf{x}')r_b(\mathbf{x}), \tag{2.22}$$

where $Q_b(\mathbf{x}, d\mathbf{x}')$ denotes the proposal distribution on $\mathcal{X} = \mathbb{R}^{I_k}$, $a_b(\mathbf{x}, \mathbf{x}')$ is the Metropolis acceptance ratio, i.e. $a_b(\mathbf{x}, \mathbf{x}') = 1 \wedge \frac{\nu(\mathbf{x}')}{\nu(\mathbf{x})} \frac{Q_b(\mathbf{x}', \mathbf{x})}{Q_b(\mathbf{x}, \mathbf{x}')}$, and $r_b(\mathbf{x}) = 1 - \int_{\mathcal{X}} Q_b(\mathbf{x}, d\mathbf{x}')a_b(\mathbf{x}, d\mathbf{x}')$. The standard way to sample from P_b^{MH} in (2.22) is sampling a proposal $\mathbf{x}'$ from $Q_b(\mathbf{x}, \cdot)$, compute $a_b(\mathbf{x}, \mathbf{x}')$ and accept if $U \sim U(0, 1)$ is smaller than the acceptance ratio. Given the known independence structure of the target, however, it is possible to propose and accept/reject each component individually, leading to much higher acceptance rates and better dimensionality scaling. The resulting kernel, which is a product of univariate MH

kernels, can be written as

$$P_f^{MH}(\mathbf{x}, d\mathbf{x}') = \otimes_{i=1}^{I_k} P_i^{MH}(x_i, dx_i') = \otimes_{i=1}^{I_k} \left(Q_f(x_i, dx_i') a_f(x_i, x_i') + \delta_{x_i}(dx_i') r_f(x_i) \right), \tag{2.23}$$

where $Q_f(x_i, dx_i')$ is a proposal kernel on $\mathbb{R}$, $a_f(x_i, x_i') = 1 \wedge \frac{\nu_i(x_i')}{\nu_i(x)} \frac{Q_f(x_i', x_i)}{Q_f(x_i, x_i')}$, and analogously $r_f(x) = 1 - \int_{\mathbb{R}} Q_f(x_i, dx_i') a_f(x_i, dx_i')$. Note that (2.22) proposes and accept jointly all the components at once, while (2.23) does it component-wise. The coupling strategy can exploit such independence in two different ways, namely factorizing both the proposal and acceptance step or only the acceptance step.

Differently from models with conjugate full-conditional distributions (such as e.g. Models 1 and 3), (optimally) contractive couplings for MH kernels are difficult to implement, requiring numerical integration, and in our simulations they did not provide significant enough decrease in distance within subsequent steps to justify their use. Similarly, simple *crn* couplings of the MH kernels, i.e. using same random number for the proposal distributions (amounting at implementing the W_2 optimal coupling on the proposals whenever Gaussians) and acceptance steps, were also not effective in contracting efficiently the chains (specifically they typically soon reach a plateau distance not small enough to provide high chances of coalescence). For the above reasons, when using MH steps to update from high-dimensional and conditionally independent blocks, we avoid the two step strategy of Algorithm 5 and instead concentrate on one-step, maximal-only strategies.

Following guidelines in Wang et al. (2021), we consider kernels with synchronous acceptance, i.e. using same uniform for accept-reject in the $\mathbf{x}$ and $\mathbf{y}$ chain. For a_b, a_f, r_b and r_f as in (2.22) and (2.23), we define

$$\bar{a}_b = \begin{pmatrix} a_b(\mathbf{x}, \mathbf{x}') \cdot \mathbf{1}_{I_k} \\ a_b(\mathbf{y}, \mathbf{y}') \cdot \mathbf{1}_{I_k} \end{pmatrix} \in \mathbb{R}^{2I_k}, \qquad \bar{a}_f = \begin{pmatrix} a_f(x_i, x_i') \\ a_f(y_i, y_i') \end{pmatrix} \in \mathbb{R}^2,$$

$$\Delta_b = \begin{pmatrix} \delta_{\mathbf{x}}(d\mathbf{x}') r_b(\mathbf{x}) \\ \delta_{\mathbf{y}}(d\mathbf{y}') r_b(\mathbf{y}) \end{pmatrix} \in \mathbb{R}^{2I_k}, \qquad \Delta_f = \begin{pmatrix} \delta_{x_i}(dx_i') r_f(x_i) \\ \delta_{y_i}(dy_i') r_b(y_i) \end{pmatrix} \in \mathbb{R}^2,$$

where $\mathbf{1}_{I_k}$ denotes the vector of ones of length I_k. Below we illustrate numerically the

performances of the following list of possible coupled kernels:

1. *Blocked reflection*: $\bar{P}_{b,r} := \bar{P}_{max}[Q_b] \odot \bar{a}_b + \Delta_b$, where $\bar{P}_{max}[Q_b]$ is Algorithm 2 and $\odot$ denotes the Hadamard product, i.e. component-wise product.

2. *Blocked maximal*: $\bar{P}_{b,m} := \bar{P}_{max}[Q_b] \odot \bar{a}_b + \Delta_b$, where $\bar{P}_{max}[Q_b]$ is Algorithm 1.

3. *Blocked factorized reflection*: $\bar{P}_{bf,r} := \otimes_{i=1}^{I_k} \left(\bar{P}_{max}[Q_b]_{[i]} \odot \bar{a}_f + \Delta_f \right)$, where, if $(\mathbf{x}, \mathbf{y}) \sim \bar{P}_{max}[Q_b]$, the symbol $\bar{P}_{max}[Q_b]_{[i]}$ indicates the vector (x_i, y_i), and $\bar{P}_{max}[Q_b]$ is Algorithm 2.

4. *Blocked factorized maximal*: $\bar{P}_{bf,m} := \otimes_{i=1}^{I_k} \left(\bar{P}_{max}[Q_b]_{[i]} \odot \bar{a}_f + \Delta_f \right)$, where $\bar{P}_{max}[Q_b]$ is Algorithm 1.

5. *Fully factorized reflection*: $\bar{P}_{ff,r} := \otimes_{i=1}^{I_k} \left(\bar{P}_{max}[Q_f] \odot \bar{a}_f + \Delta_f \right)$, where $\bar{P}_{max}[Q_b]$ is Algorithm 2.

6. *Fully factorized maximal*: $\bar{P}_{ff,m} := \otimes_{i=1}^{I_k} \left(\bar{P}_{max}[Q_f] \odot \bar{a}_f + \Delta_f \right)$, where $\bar{P}_{max}[Q_b]$ is Algorithm 1.

We report in Algorithm 9 the pseudo-code for one iteration of either $\bar{P}_{bf,r}$, $\bar{P}_{bf,m}$, $\bar{P}_{ff,r}$ or $\bar{P}_{ff,m}$, depending on the specification of $\bar{P}[Q]$.

Algorithm 9: Coupling strategy for MH with independent target

Input: $(\mathbf{X}^t, \mathbf{Y}^t)$, target ν, proposal Q, desired coupling $\bar{P}$;
sample $(\mathbf{X}', \mathbf{Y}') \sim \bar{P}[Q]((\mathbf{X}^t, \mathbf{Y}^t), \cdot)$ **for** $i = 1, ..., I_k$ **do**
  sample $U \sim U(0,1)$ **if** $U \leq \frac{\nu(X_i')}{\nu(X_i^t)} \frac{Q(X_i', X_i^t)}{Q(X_i^t, X_i')}$ **then**
    set $X_i^{t+1} = X_i'$
  else
    set $X_i^{t+1} = X_i^t$
  if $U \leq \frac{\nu(Y_i')}{\nu(Y_i^t)} \frac{Q(Y_i', \boldsymbol{Y}_i^t)}{Q(Y_i^t, Y_i')}$ **then**
    set $Y_i^{t+1} = Y_i'$
  else
    set $Y_i^{t+1} = Y_i^t$
$t = t + 1$
Output: $(\mathbf{X}^{t+1}, \mathbf{Y}^{t+1})$.

We provide a numerical illustration where ν is taken to be a product of independent Laplace distributions, i.e. $\nu = \otimes_{i=1}^{d} Lapl(0, 1/\sqrt{2})$. Consider $(\mathbf{X}^t, \mathbf{Y}^t)_{t \geq 0}$ coupled

chains marginally evolving via the kernels 1 to 6 above with $Q_b(\mathbf{x}, \cdot) = N(\mathbf{x}, \sqrt{2}I_d)$ or $Q_f(x_i, dx_i) = N(x_i, \sqrt{2})$, where step-size are chosen following the guidance in Roberts et al. (1997) for univariate Metropolis steps. We plot in Figure 2.10 the average meeting times for coupled chains with different strategies, as the target dimension d grows. As

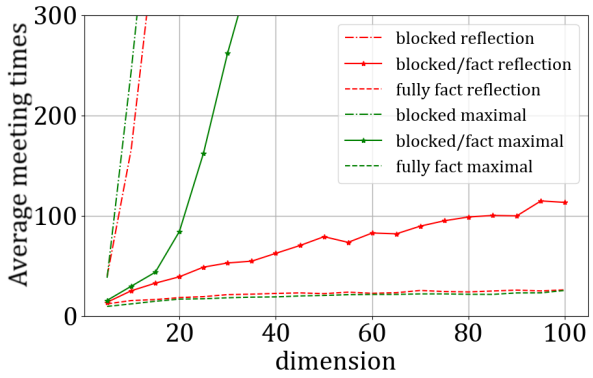

Figure 2.10: Average meeting times for different dimensions, Laplace target.

one might expect, Figure 2.10 shows that the strategies yielding smaller meeting times are those leveraging the independence structure of the target, proposing and accepting the components independently. *Blocked* strategies instead generally perform worse, with the sole exception of *block/fact* reflection, due to the intrinsic contraction properties of Algorithm 2.

To better illustrate the phenomenon, Figure 2.11 plots the proportion of components that did not meet for the same chains as in Figure 2.10, for $d = \{3, 100\}$.

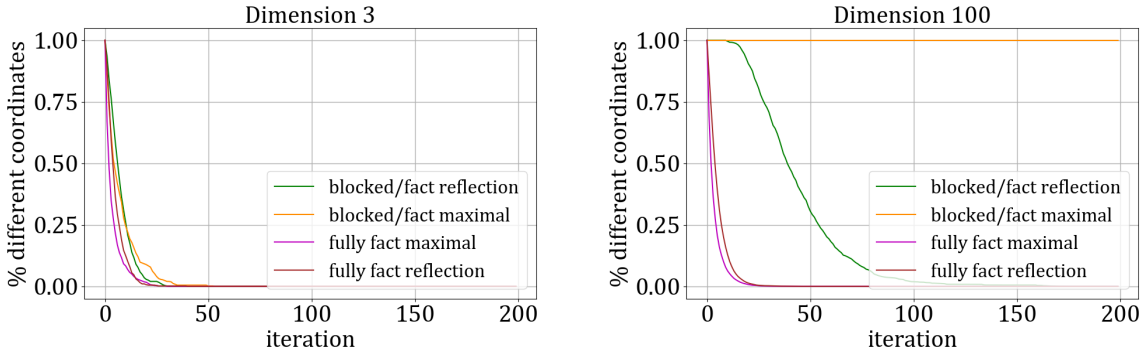

Figure 2.11: Estimated percentage of non coalesced components for blocked/component-wise proposals and component-wise acceptance; $d = 3, 100$.

The above examples suggest that, in case of conditionally independent blocks, fully-factorized couplings of the MH updates are preferable. This is coherent with the consider-

ation that, in the fully-factorized case, the overall meeting time, which coincides with the one of the slowest component that coalesces, is simply the supremum of d independent random variables, hence typically growing logarithmically as the dimensionality grows (or at least sub-linearly, see Correa and Romero (2021)).

2.11 Algorithmic implementation details

In this section, we report the explicit expressions for the full-conditional distributions required to implement the proposed algorithms for Models 1 and 3.

2.11.1 Full conditionals for Model 1

Under (2.1), the conditional distributions required to implement Algorithm 6 are

$$\mathcal{L}(\mu|\mathbf{a}^{(-k)},\boldsymbol{\tau},\mathbf{y}) = N\left(\frac{1}{\sum_j s_j^{(k)}}\sum_j s_j^{(k)}\left(\tilde{y}_j^{(k)} - \frac{\sum_{l\neq k}\sum_i a_i^{(l)} n_{ji}^{(k,l)}}{n_j^{(k)}}\right), \frac{1}{\tau_k \sum_j s_j^{(k)}}\right),$$

$$\mathcal{L}(a_i^{(k)}|\mathbf{a}^{(-k)},\mu,\boldsymbol{\tau},\mathbf{y}) = N\left(\frac{n_i^{(k)}\tau_0}{n_i^{(k)}\tau_0+\tau_k}\left(\tilde{y}_i^{(k)} - \mu - \frac{\sum_{l\neq k,0}\sum_{j=1}^{I_l} a_j^{(l)} n_{ij}^{(k,l)}}{n_i^{(k)}}\right), \frac{1}{n_i^{(k)}\tau_0+\tau_k}\right),$$

$$\mathcal{L}(\tau_k|\mathbf{a},\mu,\boldsymbol{\tau}^{-k},\mathbf{y}) = Gamma\left(\frac{I_k-1}{2}, \frac{2}{\sum_{i=1}^{I_k}(a_i^{(k)})^2}\right),$$

where $n_i^{(k)} = \sum_{n=1}^N \mathbb{I}(i_k[n] = i)$, $s_j^{(k)} = n_j^{(k)}\tau_0/(\tau_k + n_j^{(k)}\tau_0)$, $n_{ji}^{(k,l)} = \sum_{n=1}^N \mathbb{I}(i_k[n] = j)\mathbb{I}(i_l[n] = i)$ denotes the number of observations of level j of factor k and i of factor l and finally $\tilde{y}_i^{(k)} = \sum_{n:i_k[n]=i} y_n/|\{n : i_k[n] = i\}|$ is the average of all observations with level i on factor k. See also (Papaspiliopoulos et al., 2019, Eq. 4 and Prop. 2) for similar expressions.

2.11.2 Local centering algorithm for Model 3

Under (2.3), the conditional distributions required to implement Algorithm 8 are

$$\mathcal{L}(\mathbf{u}_i|\mathbf{v},\rho,\tau_0,\mathbf{y}) = N(\boldsymbol{\mu}_{\mathbf{u}_i}, Q_{\mathbf{u}_i}^{-1}), \qquad \mathcal{L}(\bar{\mathbf{u}}_i|\mathbf{v},\rho,\tau_0,\mathbf{y}) = \frac{1}{\rho}N(\boldsymbol{\mu}_{\mathbf{u}_i}, Q_{\mathbf{u}_i}^{-1}),$$
$$\mathcal{L}(\mathbf{v}_j|\mathbf{u},\rho,\tau_0,\mathbf{y}) = N(\boldsymbol{\mu}_{\mathbf{v}_j}, Q_{\mathbf{v}_j}^{-1}), \qquad \mathcal{L}(\bar{\mathbf{v}}_j|\mathbf{u},\rho,\tau_0,\mathbf{y}) = \frac{1}{\rho}N(\boldsymbol{\mu}_{\mathbf{v}_j}, Q_{\mathbf{v}_j}^{-1}),$$

where $\boldsymbol{\mu}_{\mathbf{u}_i}$, $\boldsymbol{\mu}_{\mathbf{v}_j}$, $Q_{\mathbf{u}_i} = (q_{rs})_{r,s=1}^d$ and $Q_{\mathbf{v}_j} = (p_{rs})_{r,s=1}^d$ are given by

$$q_{rr} = 1 + \tau_0\rho^2 \sum_{n:i[n]=i} v_{j[n],r}^2, \qquad p_{rr} = 1 + \tau_0\rho^2 \sum_{n:j[n]=j} u_{i[n],r}^2 \text{ for } r = 1,...,d,$$
$$q_{rs} = \tau_0\rho^2 \sum_{n:i[n]=i} v_{j[n],r}v_{j[n],s}, \qquad p_{rs} = \tau_0\rho^2 \sum_{n:j[n]=j} u_{i[n],r}u_{i[n],s}, \text{ for } r,s = 1,...,d,$$
$$\boldsymbol{\mu}_{\mathbf{u}_i} = Q_{\mathbf{u}_i}^{-1}\left(\tau_0\rho \sum_{n:i[n]=i} \mathbf{v}_{j[n]}y_n\right), \qquad \boldsymbol{\mu}_{\mathbf{v}_j} = Q_{\mathbf{v}_j}^{-1}\left(\tau_0\rho \sum_{n:j[n]=j} \mathbf{u}_{i[n]}y_n\right),$$

and

$$\mathcal{L}(\rho^{-2}|\bar{\mathbf{u}},\mathbf{v},\tau_0,\mathbf{y}) = Gamma\left(a + \frac{dI_1}{2}, \left(\frac{1}{b} + \sum_{i=1}^{I_1}\frac{\|\bar{\mathbf{u}}_i\|^2}{2}\right)^{-1}\right),$$
$$\mathcal{L}(\rho^{-2}|\mathbf{u},\bar{\mathbf{v}},\tau_0,\mathbf{y}) = Gamma\left(a + \frac{dI_2}{2}, \left(\frac{1}{b} + \sum_{j=1}^{I_2}\frac{\|\bar{\mathbf{v}}_j\|^2}{2}\right)^{-1}\right),$$
$$\mathcal{L}(\tau_0|\mathbf{u},\mathbf{v},\rho,\mathbf{y}) = Gamma\left(c + \frac{N}{2}, \left(\frac{1}{d} + \sum_{n=1}^{N}\frac{(y_n - \rho\mathbf{u}_{i[n]}\mathbf{v}_{j[n]})^2}{2}\right)^{-1}\right).$$

For the vanilla scheme with improper prior $p(\rho) \propto 1$, then

$$\mathcal{L}(\rho^{-2}|\mathbf{u},\mathbf{v},\tau_0,\mathbf{y}) = TG\left(\frac{\sum_{n=1}^N \mathbf{u}_{i[n]}^\top\mathbf{v}_{j[n]}y_n}{\sum_{n=1}^N(\mathbf{u}_{i[n]}^\top\mathbf{v}_{j[n]})^2}, \frac{1}{\tau_0\sum_{n=1}^N(\mathbf{u}_{i[n]}^\top\mathbf{v}_{j[n]})^2}; 0, +\infty\right). \tag{2.24}$$

We report in Algorithm 10 a more detailed pseudo-code for implementing the local centering approach described in Algorithm 8.

Algorithm 10: One iteration of BGS with local centering for Model 3

$\bar{\mathbf{u}} = \rho \cdot \mathbf{u}$

$\rho = \left(Gamma\left(a + \frac{d\, I_1}{2}, \left(\frac{1}{b} + \sum_{i=1}^{I_1} \frac{\|\bar{\mathbf{u}}_i\|^2}{2} \right)^{-1} \right) \right)^{-\frac{1}{2}}$

for $i = 1, ..., I_1$ **do**

  $\bar{\mathbf{u}}_i \sim N\left(\boldsymbol{\mu}_{\bar{\mathbf{u}}_i}, Q_{\bar{\mathbf{u}}_i}^{-1} \right)$

$\mathbf{u} = \bar{\mathbf{u}}/\rho$

$\bar{\mathbf{v}} = \rho \cdot \mathbf{v}$

$\rho = \left(Gamma\left(a + \frac{d\, I_2}{2}, \left(\frac{1}{b} + \sum_{i=1}^{I_2} \frac{\|\bar{\mathbf{v}}_i\|^2}{2} \right)^{-1} \right) \right)^{-\frac{1}{2}}$

for $i = 1, ..., I_2$ **do**

  $\bar{\mathbf{v}}_i \sim N\left(\boldsymbol{\mu}_{\bar{\mathbf{v}}_i}, Q_{\bar{\mathbf{v}}_i}^{-1} \right)$

$\mathbf{v} = \bar{\mathbf{v}}/\rho$

$\tau_0 \sim Gamma\left(c + \frac{N}{2}, \left(\frac{1}{d} + \sum_{n=1}^{N} \frac{(y_n - \rho \mathbf{u}_{i[n]} \mathbf{v}_{j[n]})^2}{2} \right)^{-1} \right)$

2.12 Proofs

2.12.1 Proofs of the results in Section 2.4.1

Proof of Theorem 2

The proof of Theorem 2 builds upon Lemma 10 and Lemma 11, whose statements and proofs are deferred after the end of the former.

Proof of Theorem 2. In the following, we will state the results assuming that $\mathbf{X}^0, \mathbf{Y}^0$ are fixed, or equivalently conditioning on their values, omitting the explicit conditioning in the notation for brevity.

Define $(\mathcal{D}^t)_{t \geq 0}$ as

$$\mathcal{D}^t := \|\mathcal{L}(\mathbf{X}^{t+1}|\mathbf{X}^t) - \mathcal{L}(\mathbf{Y}^{t+1}|\mathbf{Y}^t)\|_{TV} \qquad t \geq 0, \tag{2.25}$$

Denote by $(t_k)_{k \geq 1}$ as the sequence of times at which $\mathcal{D}^t < \varepsilon$, i.e.

$$t_k := \min\{t > t_{k-1} : \mathcal{D}^t < \varepsilon\} \qquad k \geq 1, \tag{2.26}$$

with $t_0 := -1$ by convention. Note that by the form of Algorithm 5, the t_k's are exactly the iterations at which a maximal coupling is implemented. Also, let A_k be a binary variable indicating whether the maximal coupling attempt at t_k is successful, i.e.

$$A_k := \begin{cases} 1 & \text{if } \mathbf{X}^{t_k+1} = \mathbf{Y}^{t_k+1} \\ 0 & \text{otherwise} \end{cases} \qquad k \geq 1. \tag{2.27}$$

By faithfulness, $A_k = 1$ implies that $\mathbf{X}^t = \mathbf{Y}^t$, for all $t \geq t_k$ and by convention $A_{k'} = 1$ for all $k' > k$. Note also that from (2.26) and (2.27), one has

$$\mathbb{E}[1 - A_k | \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}, A_{k-1} = 0] = \Pr(A_k = 0 | \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}, A_{k-1} = 0) = \mathcal{D}^{t_k}. \tag{2.28}$$

Thus, T can be written as

$$T = t_1 + 1 + \sum_{k=1}^{+\infty} (1 - A_k)(t_{k+1} - t_k). \tag{2.29}$$

We bound $\mathbb{E}[T]$ using the form of (2.29). In particular, by Lemma 10, we have

$$t_1 + 1 \leq f_1(\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|, \varepsilon, B), \tag{2.30}$$

for f_1 defined therein. Note that, conditionally on $(\mathbf{X}^0, \mathbf{Y}^0)$, the bound is deterministic: provided the chains evolve via *crn* coupling from iteration 1 up to t_1, $\mathcal{D}^{t_1}$ is a deterministic function of the starting points $(\mathbf{X}^0, \mathbf{Y}^0)$ and matrices B, N, L (see (2.35) and (2.36) for further details).

Considering the third addend in (2.29), by the definitions of $\mathcal{D}^t$, $(t_k)_{k\geq 1}$, $(A_k)_{k\geq 1}$ and the form of Algorithm 5, it follows that $\{A_k = 0\}$ implies $\{A_i = 0\}$ for $i \leq k$, then, exploiting the equality in (2.28), we get

$$\begin{aligned} &\Pr(A_k = 0 | \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}) = \Pr(A_1 = 0, ..., A_{k-1} = 0, A_k = 0 | \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}) \\ &= \Pr(A_1 = 0, ..., A_{k-1} = 0 | \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}) \Pr(A_k = 0 | \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}, A_{k-1} = 0) \\ &\leq \varepsilon^{k-1} \mathcal{D}^{t_k} \qquad\qquad k \geq 1, \end{aligned} \tag{2.31}$$

where the last inequality follows from the fact that $\Pr(A_j = 0|\mathbf{X}^{t_k}, \mathbf{Y}^{t_k}) \leq \varepsilon$ for all $j = 1, ..., k$ given that, by the form of Algorithm 5, coalescence is attempted only if it has probability greater than ε. Combining the last equality with the Monotone Convergence Theorem and Lemma 10 we can rewrite the third term of (2.29) as

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{+\infty}(1-A_k)(t_{k+1}-t_k)\right] &= \sum_{k=1}^{+\infty}\mathbb{E}\left[\mathbb{E}[(1-A_k)(t_{k+1}-t_k)|\mathbf{X}^{t_k},\mathbf{Y}^{t_k}]\right] \\
&= \sum_{k=1}^{+\infty}\mathbb{E}\left[\Pr(A_k=0|\mathbf{X}^{t_k},\mathbf{Y}^{t_k})\mathbb{E}[(t_{k+1}-t_k)|\mathbf{X}^{t_k},\mathbf{Y}^{t_k},A_k=0]\right] \\
&\leq \sum_{k=1}^{+\infty}\varepsilon^{k-1}\mathbb{E}\left[\mathcal{D}^{t_k}\mathbb{E}[(t_{k+1}-t_k)|\mathbf{X}^{t_k},\mathbf{Y}^{t_k},A_k=0]\right]. \qquad (2.32)
\end{aligned}$$

By Lemma 11, we have

$$\mathcal{D}^{t_k}\ \mathbb{E}[t_{k+1}-t_k|A_k=0,\mathbf{X}^{t_k},\mathbf{Y}^{t_k}] \leq f_2(\varepsilon,B),$$

for f_2 defined therein. Note that the inequality above holds almost surely, i.e. with probability one. Crucially, the bound does not depend on the exact distance between the chains at time t_k, but only on ε. Substituting in (2.32) we obtain

$$\mathbb{E}\left[\sum_{k=1}^{+\infty}(1-A_k)(t_{k+1}-t_k)\right] \leq f_2(\varepsilon,B)\sum_{k=1}^{+\infty}\varepsilon^{k-1} = f_2(\varepsilon,B)\,(1-\varepsilon)^{-1}.$$

It then follows

$$\mathbb{E}[T] \leq 1 + f_1(\|L^{-1}(\mathbf{X}^0-\mathbf{Y}^0)\|,\varepsilon,B) + (1-\varepsilon)^{-1}f_2(\varepsilon,B).$$

After explicit computations

$$\begin{aligned}
\mathbb{E}[T] \leq 2 + \frac{1}{-\ln(\rho(B))}\Bigg(&\ln(\|L^{-1}(\mathbf{X}^0-\mathbf{Y}^0)\|) - \frac{1}{2}\ln(1-\lambda^2_{min}(B)) - \ln(2\sqrt{2}erf^{-1}(\varepsilon)) \\
&+ \frac{1}{1-\varepsilon}\cdot\left(1+\ln(12+8\sqrt{2/\pi})erf^{-1}(\varepsilon)/\sqrt{\pi}+\sqrt{2}(\sqrt{\pi}e)^{-1}\right)\Bigg).
\end{aligned}$$

It is possible to further simplify the bound provided $\varepsilon < 0.5$, thus getting the expression

in Theorem 2, where we set $C_0 := \ln(\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|)$ and

$$C_\varepsilon := -\ln(erf^{-1}(\varepsilon)2\sqrt{2}) + 2\ln(12 + 8\sqrt{2/\pi})erf^{-1}(\varepsilon) + \sqrt{2}(\sqrt{\pi}e)^{-1} \leq -\ln(erf^{-1}(\varepsilon)) + 6erf^{-1}(\varepsilon)\,.$$

□

Lemma 10. *Under the assumptions of Theorem 2 we have*

$$t_1 + 1 \leq f_1(\|L^{-1}(\boldsymbol{X}^0 - \boldsymbol{Y}^0)\|, \varepsilon, B),$$

with

$$f_1(r, \varepsilon, B) = \left\lceil \frac{\ln(r) - \frac{1}{2}\ln(1 - \lambda^2_{min}(B)) - \ln(erf^{-1}(\varepsilon)2\sqrt{2})}{-\ln(\rho(B))} \right\rceil \quad r \in (0, \infty),\ \varepsilon \in (0, 1).$$

Proof of Lemma 10. Recall that given $p = N(\boldsymbol{\mu}, \Sigma)$ and $q = N(\boldsymbol{\nu}, \Sigma)$, it holds

$$\|p - q\|_{TV} = erf\left(\sqrt{\frac{(\boldsymbol{\mu} - \boldsymbol{\nu})^\top \Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu})}{8}}\right). \tag{2.33}$$

From (2.33) above, the autoregressive form of Gaussian chain in (2.8) and the equation of the *crn* coupling in (1.2), for all $t < t_1$ we have

$$\mathcal{D}^t = erf\left(\sqrt{\frac{(\mathbf{X}^t - \mathbf{Y}^t)^\top B^\top (\Sigma - B\Sigma B^\top)^{-1} B(\mathbf{X}^t - \mathbf{Y}^t)}{8}}\right). \tag{2.34}$$

We are interested in sufficient conditions for having $\mathcal{D}_t \leq \varepsilon$. Set $N := L^{-1}BL$, with $LL^\top = \Sigma$, and also

$$\begin{aligned} \mathbf{d}^t &:= L^{-1}B(\mathbf{X}^t - \mathbf{Y}^t) = L^{-1}B^{t+1}(\mathbf{X}^0 - \mathbf{Y}^0) \\ &= L^{-1}B^{t+1}LL^{-1}(\mathbf{X}^0 - \mathbf{Y}^0) = N^{t+1}L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0), \end{aligned} \tag{2.35}$$

where the first equality follows since the draws are paired via *crn* coupling up to iteration

t (see Algorithmic specification 1). Then (2.34) becomes

$$\mathcal{D}^t = erf\left(\sqrt{\frac{(\mathbf{d}^t)^\top(1_d - NN^\top)^{-1}\mathbf{d}^t}{8}}\right), \tag{2.36}$$

where we used $(\Sigma - B\Sigma B^\top)^{-1} = (LL^\top - BLL^\top B^\top)^{-1} = L^{-\top}(1 - NN^\top)^{-1}L^{-1}$. Given π reversibility of P, one has $\Sigma B^\top = B\Sigma$ (Khare and Zhou, 2009b, Proposition 4.27), implying $N = N^\top$. Hence by properties of the spectral radius and symmetric matrices

$$\begin{aligned}(\mathbf{d}^t)^\top(1 - NN^\top)^{-1}\mathbf{d}^t &\leq \|\mathbf{d}^t\|^2\rho(1_d - NN^\top)^{-1} \\ &= \|N^{t+1}L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2\rho(1_d - NN^\top)^{-1} \\ &\leq \|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2\frac{\|N^{t+1}\|_2^2}{\rho(1_d - NN^\top)} \\ &\leq \|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2\frac{\rho(N)^{2(t+1)}}{1 - \lambda_{min}(NN^\top)},\end{aligned} \tag{2.37}$$

where for $A \in \mathbb{R}^{m,n}$ we denote by $\|A\|_2 = \sup_{\mathbf{x}\neq 0, \mathbf{x}\in\mathbb{R}^n}\frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$ the induced 2 norm. Combining (2.36) and (2.37), a sufficient condition for $\mathcal{D}_t \leq \varepsilon$ is to have

$$\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2\frac{\rho(N)^{2(t+1)}}{1 - \lambda_{min}(NN^\top)} < 8\left(erf^{-1}(\varepsilon)\right)^2. \tag{2.38}$$

Note that this also implies that

$$\|\mathbf{d}^t\|^2 \leq 8\left(erf^{-1}(\varepsilon)\right)^2\left(1 - \lambda_{\min}(NN^\top)\right). \tag{2.39}$$

Again by π-reversibility of P, since $N = N^\top$, one has $\lambda_{min}(NN^\top) = \lambda_{min}(N)^2 = \lambda_{min}(B)^2$ and also $\rho(B) = \rho(N)$. Substituting into (2.38) and solving for t we get the result. □

Lemma 11. *Under the assumptions of Theorem 2 we have*

$$\mathcal{D}^{t_k}\,\mathbb{E}[t_{k+1} - t_k | A_k = 0, \boldsymbol{X}^{t_k}, \boldsymbol{Y}^{t_k}] \leq f_2(\varepsilon, B) \qquad a.s.,$$

where

$$f_2(\varepsilon, B) = \left\lceil \frac{\ln(12 + 8\sqrt{2/\pi})erf^{-1}(\varepsilon)/\sqrt{\pi} + 2\sqrt{2}(\sqrt{\pi}e)^{-1}}{-\ln(\rho(B))} \right\rceil.$$

Proof of Lemma 11. The proof of Lemma 11 relies on two different parts. In the first we bound the expected square distance between $\mathbf{X}^{t_k+1}$ and $\mathbf{Y}^{t_k+1}$ conditionally on $A_k = 0$, and this is achieved controlling the first and second moments of truncated Gaussians (see Lemma 6 in the main body of the Chapter). Then we bound the product of the former (possibly growing to $+\infty$ as ε goes to zero) times the total variation distance between the chains themselves.

We start by noting that the result of Lemma 10 can be extended for every $t_k - t_{k-1}$ with $k \geq 2$, since the arguments rely only on the form of Algorithm 5 and on the expression in (2.33). It follows that

$$\mathbb{E}[t_{k+1} - t_k | A_k = 0, \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}] \leq \mathbb{E}[f_1(\|L^{-1}(\mathbf{X}^{t_k+1} - \mathbf{Y}^{t_k+1})\|, \varepsilon, B) | A_k = 0, \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}].$$

Thus one can write

$$\begin{aligned}
&\mathcal{D}^{t_k}\mathbb{E}\left[t_{k+1} - t_k | A_k = 0, \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}\right] \\
&\quad \leq \mathcal{D}^{t_k}\mathbb{E}\left[f_1(\|L^{-1}(\mathbf{X}^{t_k+1} - \mathbf{Y}^{t_k+1})\|, \varepsilon, B) | A_k = 0, \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}\right] \\
&\quad \leq \mathcal{D}^{t_k} f_1\left(\mathbb{E}\left[\|L^{-1}(\mathbf{X}^{t_k+1} - \mathbf{Y}^{t_k+1})\| | A_k = 0, \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}\right], \varepsilon, B\right),
\end{aligned} \tag{2.40}$$

where the last inequality follows from Jensen applied to $f_1(\cdot, \varepsilon, B)$, and in particular to the logarithmic function in its expression. Define $z^{t_k} := \sqrt{(\mathbf{d}^{t_k})^\top (1 - NN^\top)^{-1} \mathbf{d}^{t_k}}$, with $\mathbf{d}^t$ as in (2.35). Since at iteration t_k we implement maximal reflection coupling, we can apply Lemma 6 on the argument of f_1, provided that $z_{t_k} \leq 1$, thus getting

$$\mathbb{E}\left[\|L^{-1}(\mathbf{X}^{t_k+1} - \mathbf{Y}^{t_k+1})\|^2 | A_k = 0\right] \leq \frac{\|L^{-1}B(\mathbf{X}^{t_k} - \mathbf{Y}^{t_k})\|^2}{(z^{t_k})^4} \left(12 + 8\sqrt{\frac{2}{\pi}}\right).$$

Hence substituting the bound in the expression of f_1 in (2.40), one gets

$$\begin{aligned}
&\mathcal{D}^{t_k}\mathbb{E}[(t_{k+1} - t_k) | A_k = 0, \mathbf{X}^{t_k}, \mathbf{Y}^{t_k}] \\
&\quad \leq \frac{\mathcal{D}^{t_k}}{-\ln(\rho(B))} \left(\frac{1}{2}\ln(\|\mathbf{d}^{t_k}\|^2) - \frac{1}{2}\ln(1 - \lambda_{min}^2(B)) - \ln(2\sqrt{2} erf^{-1}(\varepsilon))\right) &(2.41)\\
&\quad \leq \frac{\mathcal{D}^{t_k}}{-2\ln(\rho(B))} \left(-2\ln(z_{t_k}) + \ln(12 + 8\sqrt{2/\pi})\right) &(2.42)
\end{aligned}$$

$$\leq \frac{z^{t_k}}{-\sqrt{2\pi}\ln(\rho(B))}\left(-4\ln(z_{t_k})+\ln(12+8\sqrt{2/\pi})\right) \tag{2.43}$$

$$\leq \frac{\ln(12+8\sqrt{2/\pi})erf^{-1}(\varepsilon)/\sqrt{\pi}+\sqrt{2}(\sqrt{\pi}e)^{-1}}{-\ln(\rho(B))}, \tag{2.44}$$

where from (2.41) to (2.42) we used the condition in (2.39) and subsequent simplifications, and from (2.42) to (2.43) we instead used that, by construction and (2.34), one has $\mathcal{D}^{t_k} = erf(z^{t_k}/\sqrt{8})$. Finally it holds $erf(x) < \frac{2}{\sqrt{\pi}}x$ and $-\ln(x)x \leq 1/e$ for $x > 0$, and so $z^{t_k} \leq 2\sqrt{2}erf^{-1}(\varepsilon)$ by (2.35). □

Proof of Lemma 6

In order to prove Lemma 6, we use an instrumental lemma, namely Lemma 12. In the following, we denote $TG(\mu, \sigma^2; a, b)$ a truncated Gaussian with mean parameter μ, variance parameter σ^2, and constrained between a and b.

Lemma 12. *Let $\sigma \in (0, \infty)$ and $\alpha \in \mathbb{R}$ and $X \sim TG(0, \sigma^2; \alpha, +\infty)$. It holds that*

$$\mathbb{E}[X] \leq \max(0, \alpha) + \sigma\sqrt{\frac{2}{\pi}}, \quad \mathbb{E}[X^2] \leq \sigma^2 + \alpha^2 + \sqrt{\frac{2}{\pi}}\alpha\sigma.$$

Proof of Lemma 12. For $T \sim TG(\mu, \sigma^2; \alpha, +\infty)$, we know

$$\mathbb{E}[T] = \mu + \frac{\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}\sigma, \tag{2.45}$$

$$\mathbb{E}[T^2] = \sigma^2 + \sigma^2\frac{\frac{\alpha-\mu}{\sigma}\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)} + \mu^2 + 2\mu\sigma\frac{\phi\left(\frac{\alpha-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\alpha-\mu}{\sigma}\right)}, \tag{2.46}$$

where $\phi(\cdot), \Phi(\cdot)$ denote respectively the density and the cumulative functions of the standard normal distribution. We divide the proof in the cases $\alpha < 0$ and $\alpha \geq 0$.

Consider $\alpha < 0$. Denote by $c_{\mu,\sigma^2;\alpha}$ the normalizing constant $c_{\mu,\sigma^2;\alpha} = \int_\alpha^{+\infty} e^{\frac{-(x-\mu)^2}{2\sigma^2}}dx$. Since $X \sim TG(0, \sigma^2; \alpha, +\infty)$ and $\alpha < 0$, we have

$$\mathbb{E}[X] = \int_\alpha^{+\infty} x\frac{e^{\frac{-x^2}{2\sigma^2}}}{c_{0,\sigma^2;\alpha}}dx \leq \int_0^{+\infty} x\frac{e^{\frac{-x^2}{2\sigma^2}}}{c_{0,\sigma^2;\alpha}}dx.$$

Multiplying and dividing by $c_{0,\sigma^2;0}$ and recalling that for $Y \sim TG(0, \sigma^2; 0, +\infty)$ one has

$\mathbb{E}[Y] = \sqrt{\frac{2}{\pi}}\sigma$ from (2.45), then

$$\int_0^{+\infty} x \frac{e^{\frac{-x^2}{2\sigma^2}}}{c_{0,\sigma^2;\alpha}} dx \leq \frac{c_{0,\sigma^2;0}}{c_{0,\sigma^2;\alpha}} \sqrt{\frac{2}{\pi}}\sigma.$$

Furthermore since $\frac{c_{0,\sigma^2;0}}{c_{0,\sigma^2;\alpha}} = \frac{c_{0,\sigma^2;0}}{\int_\alpha^0 e^{-\frac{t^2}{2\sigma^2}} dx + c_{0,\sigma^2;0}} < 1$, it follows that

$$\mathbb{E}[X] \leq \sqrt{\frac{2}{\pi}}\sigma. \tag{2.47}$$

Consider now $\alpha \geq 0$. We prove

$$\mathbb{E}[X] \leq \alpha + \sqrt{\frac{2}{\pi}}\sigma = \mathbb{E}[Y], \tag{2.48}$$

with $Y \sim TG(\alpha, \sigma^2; \alpha, +\infty)$. We exploit stochastic ordering: if there exists a coupling between $X \sim TG(0, \sigma^2; \alpha, +\infty)$ and $Y \sim TG(\alpha, \sigma^2; \alpha, +\infty)$ such that $\Pr(X < Y) = 1$, then the desired result follows. Given that the Gaussian distribution belongs to the exponential family, it has monotone likelihood ratio in its canonical statistics, that is x, hence implying stochastic ordering.

For the second moment, by (2.46) and the bound just found in (2.47) and (2.48), we get

$$\mathbb{E}[X^2] = \sigma^2 + \alpha\sigma\mathbb{E}[Y] \leq \sigma^2 + \alpha^2 + \sqrt{\frac{2}{\pi}}\alpha\sigma,$$

for $Y \sim TG(0, 1; \frac{\alpha}{\sigma}, +\infty)$. □

Proof of Lemma 6. We first prove the bound for $d = 1$ and then generalize. The explicit form of Algorithm 2 will be exploited repeatedly throughout the proof.

Suppose $X \sim N(\xi, \sigma^2), Y \sim N(\nu, \sigma^2)$ where, without loss of generality, $\xi > \nu$, and define $z := \frac{\xi - \nu}{\sigma} > 0$, $W \sim U(0, 1)$, $\dot{X} \sim N(0, 1)$ as in Algorithm 5. Coalescence is not reached only whenever $W > \frac{s(\dot{X}+z)}{s(\dot{X})}$, or equivalently

$$\dot{X} > -\frac{z}{2} - \frac{\ln(W)}{z}.$$

Then it holds $\dot{X}|W, X \neq Y \sim TN\left(0, 1; -\frac{z}{2} - \frac{\ln(W)}{z}, +\infty\right)$. Furthermore, whenever coalescence is not reached we have $X = \sigma\dot{X} + \mu$ and $Y = -\sigma\dot{X} + \nu$, then, for $a > 0$

$$\begin{aligned}\mathbb{E}[a^2(X-Y)^2|W, X \neq Y] &= a^2\mathbb{E}[(2\sigma\dot{X} + \xi - \nu)^2|W, X \neq Y] \\ &= a^2\sigma^2\mathbb{E}[(2\dot{X} + z)^2|W, X \neq Y] \\ &= a^2\sigma^2\left[4\mathbb{E}[\dot{X}^2|W, X \neq Y] + z^2 + 4z\mathbb{E}[\dot{X}|W, X \neq Y]\right].\end{aligned}$$

Therefore, one can apply the bounds of Lemma 12 to get bounds for the squared distance among the two distributions.

We now extend for the multivariate case, leading to the result. Again denote by $\mathbf{z} := \Sigma^{-\frac{1}{2}}(\boldsymbol{\xi} - \boldsymbol{\nu})$, $\dot{X} \sim N_d(\mathbf{0}, 1_d)$, $W \sim U(0,1)$ and $\mathbf{e} := \frac{\mathbf{z}}{\|\mathbf{z}\|}$ as in the formulation of Algorithm 2. Whenever coupling is not reached it holds that

$$\mathbf{z}^\top\dot{\mathbf{X}} \geq -\frac{\|\mathbf{z}\|^2}{2} - \ln(W).$$

It is possible to find an orthonormal matrix R, i.e. a rotation matrix, such that the first coordinate of $\hat{\mathbf{X}} := R\dot{\mathbf{X}}$ becomes $\mathbf{z}$ and has squared norm exactly $\mathbf{z}^\top\dot{\mathbf{X}}$. It then follows from orthonormality and symmetry of $\dot{\mathbf{X}}$ that $\hat{\mathbf{X}} \sim N_d(\mathbf{0}, 1_d)$. Whenever coupling is not reached, only the first coordinate $\hat{X}_1 = \mathbf{z}^\top\dot{\mathbf{X}}$ is constrained to be greater or equal than $-\frac{\|\mathbf{z}\|^2}{2} - \ln(W)$, independently on the other coordinates, i.e. $\mathbf{z}^\top\dot{\mathbf{X}}|\mathbf{X} \neq \mathbf{Y}, W \sim TN(0, 1; -\frac{\|\mathbf{z}\|^2}{2} - \ln(W))$. For this coordinate, the bounds in Lemma 12 still hold, giving

$$\mathbb{E}[\mathbf{z}^\top\dot{\mathbf{X}}|\mathbf{X} \neq \mathbf{Y}, W] \leq \max\left(0, -\frac{\|\mathbf{z}\|^2}{2} - \ln(W)\right) + \sqrt{\frac{2}{\pi}}, \tag{2.49}$$

$$\mathbb{E}[(\mathbf{z}^\top\dot{\mathbf{X}})^2|\mathbf{X} \neq \mathbf{Y}, W] \leq 1 + \left(-\frac{\|\mathbf{z}\|^2}{2} - \ln(W)\right)^2 + \sqrt{\frac{2}{\pi}}\left(-\frac{\|\mathbf{z}\|^2}{2} - \ln(W)\right). \tag{2.50}$$

Then, leveraging the expression of $\mathbf{X} - \mathbf{Y}|\mathbf{X} \neq \mathbf{Y}$ and $\mathbf{e}$, we get

$$\begin{aligned}\mathbb{E}[\|A(\mathbf{X}-\mathbf{Y})\|^2|\mathbf{X} \neq \mathbf{Y}, W] &= \mathbb{E}\left[\|A(\boldsymbol{\xi} - \boldsymbol{\nu} + 2(\mathbf{e}^T\dot{\mathbf{X}})\Sigma^{\frac{1}{2}}\mathbf{e})\|^2|\mathbf{X} \neq \mathbf{Y}, W\right] \\ &= \mathbb{E}\left[\left\|A(\boldsymbol{\xi} - \boldsymbol{\nu})\left(1 + \frac{2}{\|\mathbf{z}\|^2}(\mathbf{z}^\top\dot{\mathbf{X}})\right)\right\|^2|\mathbf{X} \neq \mathbf{Y}, W\right]\end{aligned}$$

$$
\begin{aligned}
&= \|A(\boldsymbol{\xi}-\boldsymbol{\nu})\|^2\,\mathbb{E}\left[1+\frac{4}{\|\mathbf{z}\|^4}(\mathbf{z}^\top\dot{\mathbf{X}})^2+\frac{4}{\|\mathbf{z}\|^2}\mathbf{z}^\top\dot{\mathbf{X}}|\mathbf{X}\neq\mathbf{Y},W\right]\\
&\leq \|A(\boldsymbol{\xi}-\boldsymbol{\nu})\|^2\left[2+\frac{4}{\|\mathbf{z}\|^4}\left(1+\ln^2(W)-\sqrt{\frac{2}{\pi}}\ln(W)\right)+\frac{4}{\|\mathbf{z}\|^2}\left(\ln(W)+1/\sqrt{2\pi}\right)\right.\\
&\left.+\max\left(0,-2-\frac{4}{\|\mathbf{z}\|^2}\ln(W)\right)\right]. \qquad (2.51)
\end{aligned}
$$

By law of iterated expectations, one has

$$
\mathbb{E}[\|A(\mathbf{X}-\mathbf{Y})\|^2|\mathbf{X}\neq\mathbf{Y}]=\mathbb{E}\left[\mathbb{E}[\|A(\mathbf{X}-\mathbf{Y})\|^2|\mathbf{X}\neq\mathbf{Y},W]\right]. \qquad (2.52)
$$

Hence we compute

$$
\begin{aligned}
\mathbb{E}\left[\max\left(0,-2-\frac{4}{\|\mathbf{z}\|^2}\ln(W)\right)\right]&=\int_0^1\max\left(0,-2-\frac{4}{\|\mathbf{z}\|^2}\ln(W)\right)dw\\
&=\frac{4}{\|\mathbf{z}\|^2}\int_0^{e^{-\frac{\|\mathbf{z}\|^2}{2}}}\left(-\frac{\|\mathbf{z}\|^2}{2}-\ln(w)\right)dw=\frac{4}{\|\mathbf{z}\|^2}e^{-\frac{\|\mathbf{z}\|^2}{2}}.
\end{aligned}
$$

Knowing $\mathbb{E}[\ln(W)]=-1$, $\mathbb{E}[\ln^2(W)]=2$, plugging (2.51) in (2.52), one gets

$$
\begin{aligned}
&\mathbb{E}[\|A(\mathbf{X}-\mathbf{Y})\|^2|\mathbf{X}\neq\mathbf{Y}]=\mathbb{E}_W\left[\mathbb{E}[\|A(\mathbf{X}-\mathbf{Y})\|^2|\mathbf{X}\neq\mathbf{Y},W]\right]\\
&\leq\|A(\boldsymbol{\xi}-\boldsymbol{\nu})\|^2\left(2+\frac{4}{\|\mathbf{z}\|^4}(3+\sqrt{2/\pi})+\frac{4}{\|\mathbf{z}\|^2}(e^{\frac{-\|\mathbf{z}\|^2}{2}}-1+1/\sqrt{2\pi})\right).
\end{aligned}
$$

Given that $\|\mathbf{z}\|\leq 1$ by hypothesis, then

$$
\mathbb{E}[\|A(\mathbf{X}-\mathbf{Y})\|^2|\mathbf{X}\neq\mathbf{Y}]\leq\|A(\boldsymbol{\xi}-\boldsymbol{\nu})\|^2\left(\frac{12+8\sqrt{\frac{2}{\pi}}}{\|\mathbf{z}\|^4}\right).
$$

□

Proof of the claim in Remark 1

Let $\mathbf{x}\sim\pi=N(\boldsymbol{\mu},\Sigma)$ divided in K blocks as in Section 2.3.1. Consider $D\in\mathbb{R}^{d\times d}$ block-diagonal matrix with same blocking structure and denote by $\mathbf{x}_D:=D\mathbf{x}$ and by π_D the transformed random variable and the induced distribution respectively. We now show

that both the distribution of the meeting time induced by Algorithm 5 and the bound in (2.12) will not change.

From well known Gaussian properties, it follows $\pi_D = N(D\boldsymbol{\mu}, D\Sigma D^\top)$, furthermore simple calculations point that $B_D = DBD^{-1}$ and $L_D = DL$. Let $(\mathbf{X}^t, \mathbf{Y}^t)_{t\geq 1}$ as in Theorem 2 and $(\mathbf{X}_D^t, \mathbf{Y}_D^t)_{t\geq 1}$ the chain targeting π_D starting at $(\mathbf{X}_D^0, \mathbf{Y}_D^0) := (D\mathbf{X}^0, D\mathbf{Y}^0)$. It follows that $\|\mathcal{L}(\mathbf{X}^{t+1}|\mathbf{X}^t) - \mathcal{L}(\mathbf{Y}^{t+1}|\mathbf{Y}^t)\|_{TV} = \|\mathcal{L}(\mathbf{X}_D^{t+1}|\mathbf{X}_D^t) - \mathcal{L}(\mathbf{Y}_D^{t+1}|\mathbf{Y}_D^t)\|_{TV}$ for all $t \geq 1$, hence the equal behaviour of Algorithm 5.

2.12.2 Proofs of Theorem 3

The proof of Theorem 3 builds upon two results presented in Lemma 13 and Lemma 14 below. In this section, under the assumption of Theorem 3, we will always assume $P^{(F)} = P_2P_1$ and $B^{(F)}$ as in Lemma 5 accordingly. Let also L be the block triangular matrix such that $LL^\top = \Sigma$.

Lemma 13. *Under the assumption of Theorem 3, for $N^{(F)} := L^{-1}B^{(F)}L$, it holds*

$$\lambda_{min}\left(N^{(F)}(N^{(F)})^\top\right) = 0.$$

Proof of Lemma 13. Leveraging the results in Remark 1, we find a suitable block diagonal linear transformation D of the chain and compute for the transformed chain $\lambda_{min}(N_D^{(F)}(N_D^{(F)})^\top)$ (using the same notation as in Remark 1), then, since $N_D^{(F)} = L_D^{-1}B_D^{(F)}L_D = L^{-1}D^{-1}DB^{(F)}D^{-1}DL = L^{-1}B^{(F)}L = N^{(F)}$, we get the result. Consider $D = diag(Q_{(1,1)}, Q_{(2,2)})^{\frac{1}{2}}$. It follows that the precision matrix of π_D is

$$Q_D = \left(\begin{array}{c|c} 1 & M \\ \hline M^\top & 1 \end{array}\right), \tag{2.53}$$

where $M = Q_{(1,1)}^{-\frac{1}{2}}Q_{(1,2)}Q_{(2,2)}^{-\frac{1}{2}}$. By Lemma 5, one gets

$$B_D^{(F)} = \left(\begin{array}{c|c} 0 & -M \\ \hline 0 & M^\top M \end{array}\right).$$

If $L_D L_D^\top = \Sigma_D$ (with Σ_D full rank), then it follows ${L_D}^{-\top} {L_D}^{-1} = Q_D$. Suppose now

$$L_D^{-1} = \left(\begin{array}{c|c} A & 0 \\ \hline B & C \end{array}\right), \tag{2.54}$$

for A, B, C matrices of suitable dimensions, one has

$$\begin{cases} A^\top A + B^\top B = 1 \\ B^\top C = M \\ C^\top C = 1. \end{cases} \tag{2.55}$$

And therefore, solving for $N_D^{(F)}$

$$N_D^{(F)} = \left(\begin{array}{c|c} 1 - AA^\top & -AB^\top \\ \hline 0 & 0 \end{array}\right).$$

From which $\lambda_{min}\left(N_D^{(F)}(N_D^{(F)})^\top\right) = \lambda_{min}(NN^\top) = 0.$ □

Lemma 14. *Let $B^{(F)}$ and $B^{(FB)}$ be respectively the auto-regressive matrices induced by $P^{(F)}$ of (2.4) and $P^{(FB)}$ of (2.11) for $\pi = N(\boldsymbol{\mu}, \Sigma)$, with $K = 2$ blocks. Let $N^{(F)} = L^{-1}B^{(F)}L$ and $N^{(FB)} = L^{-1}B^{(FB)}L$. For all $t > 1$ one has*

$$(B^{(F)})^t = A_2 (B^{(FB)})^{t-1},$$

with A_2 as in (2.58). If furthermore

$$Q = \left(\begin{array}{c|c} 1 & M \\ \hline M^\top & 1 \end{array}\right), \tag{2.56}$$

it holds

$$\| \left(N^{(F)}\right)^t \|_2 \leq \rho\left(\left(B^{(FB)}\right)^{t-1}\right) = \rho\left(B^{(F)}\right)^{t-1} = \rho\left(M^\top M\right)^{t-1},$$

where $\| \cdot \|_2$ is the induced 2-norm.

Proof. For a two block Gaussian it holds $\mathbb{E}[\mathbf{x}_{(i)}|\mathbf{x}_{(j)}] = A_{ij}\mathbf{x}_{(j)} + \mathbf{a}_{(i)}$ for $i \neq j \in \{1, 2\}$. So

$$B^{(F)} = \left(\begin{array}{c|c} 0 & A_{12} \\ \hline 0 & A_{21}A_{12} \end{array}\right), \qquad B^{(FB)} = \left(\begin{array}{c|c} 0 & A_{12}A_{21}A_{12} \\ \hline 0 & A_{21}A_{12} \end{array}\right). \tag{2.57}$$

Note that from the above $\rho\left(B^{(F)}\right) = \rho(A_{21}A_{12}) = \rho\left(B^{(FB)}\right)$. One can rewrite (2.57) as $B^{(F)} = A_2A_1$ and $B^{(FB)} = A_1A_2A_1$ for

$$A_1 = \left(\begin{array}{c|c} 1 & 0 \\ \hline A_{21} & 0 \end{array}\right), \qquad A_2 = \left(\begin{array}{c|c} 0 & A_{12} \\ \hline 0 & 1 \end{array}\right). \tag{2.58}$$

Simple algebra shows that $A_1^2 = A_1, A_2^2 = A_2$ and $(B^{(F)})^t = A_2(B^{(FB)})^{t-1}$. Furthermore:

$$(N^{(F)})^t = L^{-1}(B^{(F)})^tL = L^{-1}A_2LL^{-1}(B^{(FB)})^{t-1}L = \tilde{A}_2(N^{(FB)})^{t-1}, \tag{2.59}$$

where we defined $\tilde{A}_2 := L^{-1}A_2L$. By symmetry of $N^{(FB)}$ and submultiplicativity of the matrix norm, it follows:

$$\|(N^{(F)})^t\|_2 = \|\tilde{A}_2(N^{(FB)})^{t-1}\|_2 \leq \|\tilde{A}_2\|_2\rho(N^{(FB)})^{t-1}.$$

If furthermore Q has the form in (2.56), then one has $A_{12} = M, A_{21} = -M^\top$. Let L^{-1} be such that

$$L^{-1} = \left(\begin{array}{c|c} A & 0 \\ \hline B & C \end{array}\right), \tag{2.60}$$

It follows

$$\tilde{A}_2 = L_{-1}A_2L = \left(\begin{array}{c|c} 1 & 0 \\ \hline (B - CM^\top)A^{-1} & 0 \end{array}\right),$$

and from (2.55) it is easy to see that $B = C^{-\top}M = CM$ and so $\|\tilde{A}_2\|_2 = 1$. Plugging the result in (2.59) gives

$$\| \left(N^{(F)}\right)^t \|_2 \leq \rho\left(\left(N^{(FB)}\right)^{t-1}\right) = \rho\left(M^\top M\right)^{t-1}.$$

□

Proof of Theorem 3. The result follows from an adaptation of Lemma 10 and Lemma 11 to the case of $K = 2$ blocks, combined with Lemma 13 and Lemma 14 above. Specifically by Lemma 13 and Lemma 14 we have

$$\frac{\|(N^{(F)})^t\|^2}{1 - \lambda_{min}\left(N^{(F)}(N^{(F)})^\top\right)} = \rho(N^{(F)})^{2(t-1)}.$$

And then the sufficient condition in (2.38) becomes

$$\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2 \rho(N^{(F)})^{2(t-1)} < 8\left(erf^{-1}(\varepsilon)\right)^2.$$

Since then $\rho(N^{(BF)}) = \rho(B^{(FB)}) = \rho(B^{(F)})$, it leads to

$$t > \frac{\ln(\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|) - \ln(2\sqrt{2}\, erf^{-1}(\varepsilon))}{-\ln(\rho(B^{(F)}))}.$$

Hence one can rewrite the formula for f_1 of Lemma 10 as

$$f_1(r, \varepsilon, B) = 1 + \left\lceil \frac{\ln(r) - \ln(erf^{-1}(\varepsilon)2\sqrt{2})}{-\ln(\rho(B^{(F)}))} \right\rceil.$$

The final formula is then obtained by plugging the above in the proof of Lemma 11 for f_2. □

2.12.3 Proof of Theorem 4

Proof of Theorem 4. With the same reasoning as in Theorem 2, one can show that

$$T \leq 1 + \tilde{f}_1(\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|, \varepsilon, B, \delta) + (1 - \varepsilon)^{-1}\tilde{f}_2(\varepsilon, B, \delta),$$

where

$$\tilde{f}_1(r, \varepsilon, B, \delta) = \max\left(n^*_\delta, \left\lceil \frac{\ln(r) - \frac{1}{2}\ln(1 - \rho(NN^\top)) - \ln(erf^{-1}(\varepsilon)2\sqrt{2})}{1 - \rho(B)}(1 + \delta) \right\rceil\right),$$

and

$$\tilde{f}_2(\varepsilon, B, \delta) = \max\left(n^*_\delta, \left\lceil \frac{\ln(12 + 8\sqrt{2/\pi})erf^{-1}(\varepsilon)/\sqrt{\pi} + \sqrt{2}(\sqrt{\pi}e)^{-1}}{1 - \rho(B)}(1+\delta)\right\rceil\right).$$

Then combining the results with the same reasoning as in the proof of Theorem 2 gives the result.

The form of $\tilde{f}_1$ comes from a generalization of f_1 in Lemma 10. Given (2.34), a sufficient condition for $\mathcal{D}^t < \varepsilon$ is

$$\|\mathbf{d}^t\|^2 < 8(1 - \lambda_{min}(NN^\top))(erf^{-1}(\varepsilon))^2,$$

where as before $N = L^{-1}BL$ (no longer symmetric) and $\mathbf{d}^t = L^{-1}B(\mathbf{X}^t - \mathbf{Y}^t)$. By properties of matrix norm one has

$$\|\mathbf{d}^t\|^2 = \|N^{t+1}L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2 \leq \|N^{t+1}\|_2^2\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2.$$

The definition of n^*_δ implies that for all $t \geq n^*_\delta$:

$$\|N^t\|_2 \leq \left(1 - \frac{1-\rho(N)}{1+\delta}\right)^t = \left(1 - \frac{1-\rho(B)}{1+\delta}\right)^t \leq e^{-t\frac{1-\rho(B)}{1+\delta}}.$$

Hence if t is bigger than n^*_δ we have

$$\|\mathbf{d}^t\|^2 \leq e^{-(t+1)\frac{1-\rho(B)}{1+\delta}}\|L^{-1}(\mathbf{X}^0 - \mathbf{Y}^0)\|^2.$$

Imposing the former to be smaller than $8(1 - \lambda_{min}(NN^\top))(erf^{-1}(\varepsilon))^2$ and solving for $t+1$ leads to the result.

As for $\tilde{f}_2$, the result follows from substituting f_1 with $\tilde{f}_1$ in the proof of Lemma 11. □

2.12.4 Proof of the claim in Remark 4

Proof of the claim in Remark 4. Consider a general d-dimensional Gaussian $\pi = N(\boldsymbol{\mu}, Q^{-1})$ divided in K blocks of dimensions $I_1, ..., I_K$, for $d = \sum_k I_k$. For π above it holds

$$\pi(\mathbf{x}_{(k)}|\mathbf{x}_{(-k)}) = N\left(\sum_{j\neq k} A_{(k,j)}\mathbf{x}_{(k)} + \mathbf{a}_i, Q^{-1}_{(k,k)}\right), \tag{2.61}$$

where $A = 1_d - diag(Q^{-1}_{(1,1)}, ..., Q^{-1}_{(K,K)})Q$ and $\mathbf{a}_i = Q^{-1}_{(i,i)} \sum_{j=1}^{s} Q_{(i,j)}\boldsymbol{\mu}_{(j)}$. Lemma 7 of Section 2.6 shows the equivalence between $\bar{P}_{W_2}[P]$ and $\bar{P}^{c*}$ of (2.17). As for the maximal coupling, note that the leading term in the computational cost of Algorithm 2 is the cost of the Cholesky decomposition of Q necessary for computing $\mathbf{z}$, known to be $O(d^3)$. It follows that implementing naively $\bar{P}_{max}[P]$ has a cost of $O((\sum_k I_k)^3)$, while implementing $\bar{P}^{c*}$ of $O(\max(I_1^3, ..., I_K^3))$ (for fixed K), since composition of K successive maximal reflection couplings. We show it is possible to implement $\bar{P}_{max}[P]$ at the same cost. In the following we consider the case of $K = 3$ blocks and forward kernel, although the procedure can be extended straightforwardly to other specifications. A sweep of the Gibbs kernel P of (2.4) can be written as

$$\begin{aligned}
\mathbf{X}^{t+1}_{(1)} &= A_{(1,2)}\mathbf{X}^{t}_{(2)} + A_{(1,3)}\mathbf{X}^{t}_{(3)} + \mathbf{a}_1 + Q^{-\frac{1}{2}}_{(1,1)}\ \mathbf{Z}_1 \\
\mathbf{X}^{t+1}_{(2)} &= A_{(2,1)}\mathbf{X}^{t+1}_{(1)} + A_{(2,3)}\mathbf{X}^{t}_{(3)} + \mathbf{a}_2 + Q^{-\frac{1}{2}}_{(2,2)}\mathbf{Z}_2 \\
&= A_{(2,1)}A_{(1,2)}\mathbf{X}^{t}_{(2)} + (A_{(2,1)}A_{(1,3)} + A_{(2,3)})\mathbf{X}^{t}_{(3)} + A_{(2,1)}Q^{-\frac{1}{2}}_{(1,1)}\mathbf{Z}_1 + Q^{-\frac{1}{2}}_{(2,2)}\mathbf{Z}_2 + \mathbf{c}_2 \\
\mathbf{X}^{t+1}_{(3)} &= A_{(3,1)}\mathbf{X}^{t+1}_{(1)} + A_{(3,2)}\mathbf{X}^{t+1}_{(2)} + \mathbf{a}_3 + Q^{-\frac{1}{2}}_{(3,3)}\mathbf{Z}_3 \\
&= \left(A_{(3,1)}A_{(1,2)} + A_{(3,2)}A_{(2,1)}A_{(1,2)}\right)\mathbf{X}^{t}_{(2)} + \left(A_{(3,1)}A_{(1,3)} + A_{(3,2)}A_{(2,1)}A_{(1,3)} + A_{(3,2)}A_{(2,3)}\right)\mathbf{X}^{t}_{(3)} + \\
&+ \left(A_{(3,1)} + A_{(3,2)}A_{(2,1)}\right)Q^{-\frac{1}{2}}_{(1,1)}\mathbf{Z}_1 + A_{(3,2)}Q^{-\frac{1}{2}}_{(2,2)}\mathbf{Z}_2 + Q^{-\frac{1}{2}}_{(3,3)}\mathbf{Z}_3 + \mathbf{c}_3,
\end{aligned} \tag{2.62}$$

where $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3$ are Gaussian of dimensions I_1, I_2, I_3 respectively and $\mathbf{c}_2, \mathbf{c}_3$ vectors depending solely on A, $\boldsymbol{\mu}$ and Σ. That is

$$\begin{pmatrix} \mathbf{X}_{(1)}^{t+1} \\ \mathbf{X}_{(2)}^{t+1} \\ \mathbf{X}_{(3)}^{t+1} \end{pmatrix} = B \begin{pmatrix} \mathbf{X}_{(1)}^{t} \\ \mathbf{X}_{(2)}^{t} \\ \mathbf{X}_{(3)}^{t} \end{pmatrix} + \begin{pmatrix} Q_{(1,1)}^{-\frac{1}{2}} & 0 & 0 \\ A_{(2,1)}Q_{(1,1)}^{-\frac{1}{2}} & Q_{(2,2)}^{-\frac{1}{2}} & 0 \\ \left(A_{(3,1)} + A_{(3,2)}A_{(2,1)}\right) Q_{(1,1)}^{-\frac{1}{2}} & A_{(3,2)}Q_{(2,2)}^{-\frac{1}{2}} & Q_{(3,3)}^{-\frac{1}{2}} \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \mathbf{Z}_3 \end{pmatrix} + \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{pmatrix}, \tag{2.63}$$

for B the same of Lemma 5. Since by Lemma 5 we have

$$\begin{pmatrix} \mathbf{X}_{(1)}^{t+1} \\ \mathbf{X}_{(2)}^{t+1} \\ \mathbf{X}_{(3)}^{t+1} \end{pmatrix} = B\mathbf{X}^t + (\Sigma - B\Sigma B^\top)^{\frac{1}{2}}\mathbf{Z} + \mathbf{b}, \tag{2.64}$$

then equating (2.63) and (2.64), it must hold

$$\begin{pmatrix} Q_{(1,1)}^{-\frac{1}{2}} & 0 & 0 \\ A_{(2,1)}Q_{(1,1)}^{-\frac{1}{2}} & Q_{(2,2)}^{-\frac{1}{2}} & 0 \\ \left(A_{(3,1)} + A_{(3,2)}A_{(2,1)}\right) Q_{(1,1)}^{-\frac{1}{2}} & A_{(3,2)}Q_{(2,2)}^{-\frac{1}{2}} & Q_{(3,3)}^{-\frac{1}{2}} \end{pmatrix} = (\Sigma - B\Sigma B^\top)^{\frac{1}{2}}.$$

Implementing Algorithm 2 for iteration of teh chains $(\mathbf{X}^t, \mathbf{Y}^t)_{t\geq 1}$, one has $\mathbf{z} := (\Sigma - B\Sigma B^\top)^{-\frac{1}{2}}B(\mathbf{X}^t - \mathbf{Y}^t)$, hence computing $\mathbf{z}$ is actually equivalent to solving the triangular system

$$(\Sigma - B\Sigma B^\top)^{\frac{1}{2}}\mathbf{z} = B(\mathbf{X}^t - \mathbf{Y}^t).$$

Starting from the first coordinate, it can be proved the solution is

$$\mathbf{z}_{(1)} = Q_{(1,1)}^{\frac{1}{2}}\left(A_{(1,2)}(\mathbf{X}_{(2)}^t - \mathbf{Y}_{(2)}^t) + A_{(1,3)}(\mathbf{X}_{(3)}^t - \mathbf{Y}_{(3)}^t)\right).$$

Then

$$\begin{aligned} \mathbf{z}_{(2)} &= Q_{(2,2)}^{\frac{1}{2}}\left(A_{(2,1)}A_{(1,2)}(\mathbf{X}_{(2)}^t - \mathbf{Y}_{(2)}^t) + (A_{(2,1)}A_{(1,3)} + A_{(2,3)})(\mathbf{X}_{(3)}^t - \mathbf{Y}_{(3)}^t) - A_{(2,1)}Q_{(1,1)}^{-\frac{1}{2}}\mathbf{z}_1\right) \\ &= Q_{(2,2)}^{\frac{1}{2}}A_{(2,3)}(\mathbf{X}_{(3)}^t - \mathbf{Y}_{(3)}^t), \end{aligned}$$

and lastly

$$\mathbf{z}_{(3)} = 0.$$

From the above it follows that the computational cost of solving for $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$ is exactly $O(\max(I_1^3, I_2^3, I_3^3))$. □

2.12.5 Proof of Lemma 7

Proof of Lemma 7. We first show that the distribution induced by $\bar{P}^{c*}$ of (2.17) is the same as the one induced by $\bar{P}_{W_2}[P]$ of Lemma 2. We then extend the result for the composition of n kernels.

Given any updating order $(k_1, ..., k_K)$, let σ be the permutation of $(1, ..., K)$, such that $(k_1, ..., k_K) = (\sigma(1), ..., \sigma(K))$. Define $A = 1_d - diag(Q_{(1,1)}^{-1}, ..., Q_{(K,K)}^{-1})\, Q$, A^* the matrix whose blocks are $A^*_{(i,j)} = A_{(k_i,k_j)}$ and also $B^* = (I - L^*)^{-1}U^*$, for U^* and L^* upper and lower decomposition of A^*, i.e. $U^* + L^* = A^*$. Lastly define the matrix $B^{(\sigma)}$ as $B^{(\sigma)}_{(k_i,k_j)} = B^*_{(i,j)}$. From Lemma 2 and Lemma 5 of Section 2.2.1, the W_2-optimal coupling of P is

$$\bar{P}_{W_2}[P]((\mathbf{x}, \mathbf{y}), \cdot) = N\left(\begin{pmatrix} B^{(\sigma)}\mathbf{x} + \mathbf{b}^{(\sigma)} \\ B^{(\sigma)}\mathbf{y} + \mathbf{b}^{(\sigma)} \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes \left(\Sigma - B^{(\sigma)}\Sigma(B^{(\sigma)})^\top\right)\right), \quad (2.65)$$

where $\mathbf{b}^{(\sigma)} = (I - B^{(\sigma)})\boldsymbol{\mu}$. It follows from Lemma 2 and (2.61) that $\bar{P}_{W_2}[P_k]$ is

$$\bar{P}_{W_2}[P_k]\left((\mathbf{x}, \mathbf{y}), (d\mathbf{x}, d\mathbf{y})\right) = \bar{\nu}[\nu](d\mathbf{x}_{(k)}, d\mathbf{y}_{(k)})\delta_{(\mathbf{x}_{(-k)}, \mathbf{y}_{(-k)})}(d\mathbf{x}_{(-k)}, \mathbf{y}_{(-k)}), \quad (2.66)$$

with

$$\bar{\nu}[\nu](d\mathbf{x}_{(k)}, d\mathbf{y}_{(k)}) = N\left(\begin{pmatrix} \boldsymbol{\mu}_{(k)} + A_{(k,:)}(\mathbf{x} - \boldsymbol{\mu}) \\ \boldsymbol{\mu}_{(k)} + A_{(k,:)}(\mathbf{y} - \boldsymbol{\mu}) \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \otimes Q^{-1}_{(k,k)}\right).$$

Given that $\bar{P}_{W_2}[P_k] \in \Gamma[P_k]$ for all $k = 1, ..., K$, it directly follows from the definition of couplings that $\mathbb{E}[\bar{P}^{c*}] = \mathbb{E}[\bar{P}_{W_2}[P]]$, and also the diagonal elements of the variance

covariance matrix of $\bar{P}^{c*}$ and $\bar{P}_{W_2}[P]$ must be equal. As for the covariances, if $(d\mathbf{x}, d\mathbf{y}) \sim \bar{P}^{c*}((\mathbf{x}, \mathbf{y}), \cdot)$, for $s = 1, ..., K$ we have

$$
\begin{aligned}
&d\mathbf{y}_{(s)} - \mathbb{E}[d\mathbf{y}_{(s)}] = d\mathbf{y}_{(s)} - \boldsymbol{\mu}_{(s)} - \sum_{l=1}^{k_s-1} A^*_{(s,l)}(d\mathbf{y}_{(k_l)} - \boldsymbol{\mu}_{(k_l)}) - \sum_{l=k_s+1}^{K} A^*_{(s,l)}(\mathbf{y}_{(k_l)} - \boldsymbol{\mu}_{(k_l)}) \\
&= F_s Z_s \\
&= d\mathbf{x}_{(s)} - \boldsymbol{\mu}_{(s)} - \sum_{l=1}^{k_s-1} A^*_{(s,l)}(d\mathbf{x}_{(k_l)} - \boldsymbol{\mu}_{(k_l)}) - \sum_{l=k_s+1}^{K} A^*_{(s,l)}(\mathbf{x}_{(k_l)} - \boldsymbol{\mu}_{(k_l)}) \\
&= d\mathbf{x}_{(r)} - \mathbb{E}[d\mathbf{x}_{(s)}],
\end{aligned}
$$

where F_s such that $F_s F_s^\top = (Q^{-1})_{(s,s)}$, $Z_s \sim N(\mathbf{0}_{I_s}, 1_{I_s})$. For every $1 \leq r < s \leq K$, it follows

$$
\begin{aligned}
cov\left(d\mathbf{x}_{(r)}, d\mathbf{y}_{(s)}\right) &= \mathbb{E}\left[(d\mathbf{x}_{(r)} - \mathbb{E}[d\mathbf{x}_{(r)}])(d\mathbf{y}_{(s)} - \mathbb{E}[d\mathbf{y}_{(s)}])^\top\right] \\
&= \mathbb{E}\left[(d\mathbf{x}_{(r)} - \mathbb{E}[d\mathbf{x}_{(r)}])(d\mathbf{x}_{(s)} - \mathbb{E}[d\mathbf{x}_{(s)}])^\top\right] \\
&= cov\left(d\mathbf{x}_{(r)}, d\mathbf{x}_{(s)}\right) \\
&= \left(\Sigma - B^{(\sigma)}\Sigma(B^{(\sigma)})^\top\right)_{(r,s)},
\end{aligned}
$$

hence the result for $n = 1$.

We now prove the result for $n \geq 2$. Given $\bar{P}^{c*} = \bar{P}_{W_2}[P]$ as proved above and leveraging the equivalent formulation in (2.8) along with properties of Gaussian distribution, it follows that iterating n times $P(\mathbf{x}, \cdot)$ is the same as

$$
P^n(\mathbf{x}, \cdot) \overset{d}{=} N\left((B^{(\sigma)})^n\mathbf{x} + \left(\sum_{j=0}^{n-1}(B^{(\sigma)})^j\right)\mathbf{b}^{(\sigma)}, \left(\sum_{j=0}^{n-1}(B^{(\sigma)})^j\right)\left(\Sigma - (B^{(\sigma)})\Sigma(B^{(\sigma)})^\top\right)\left(\sum_{j=0}^{n-1}(B^{(\sigma)})^j\right)^\top\right).
$$

Suppose that $(d\mathbf{x}', d\mathbf{y}') \sim \left(\bar{P}_{W_2}[P]\right)^n((\mathbf{x}, \mathbf{y}), \cdot)$, then

$$
\mathbb{E}[\|\mathbf{x}' - \mathbf{y}'\|^2] = \|(B^{(\sigma)})^n\mathbf{x} - (B^{(\sigma)})^n\mathbf{y}\|^2 = W_2\left(P^n(\mathbf{x}, \cdot), P^n(\mathbf{y}, \cdot)\right).
$$

□

2.12.6 Proof of Lemma 9

Proof of Lemma 9. By (2.33), and the definition of $\bar{c}_d$, it holds

$$\|p - q\|_{TV} = erf\left(\bar{c}_d d^{-\alpha+\frac{1}{2}}\right). \tag{2.67}$$

If $\alpha > \frac{1}{2}$, as $d \to +\infty$, Taylor expanding the erf function around 0 gives

$$\Pr_{max}(p, q) = 1 - \|p - q\|_{TV} \asymp 1 - \frac{2\bar{c}_d}{\sqrt{\pi}} d^{-\alpha+\frac{1}{2}}. \tag{2.68}$$

If instead $0 < \alpha < \frac{1}{2}$, the argument of the erf function goes to $+\infty$. We exploit Gaussian tail bounds to characterize the behaviour. Recall indeed that

$$erf(x) = 2\Phi(\sqrt{2}x) - 1,$$

$$\Pr_{max}(p, q) = 1 - erf\left(\bar{c}_d d^{-\alpha+\frac{1}{2}}\right) = 2\left(1 - \Phi\left(\sqrt{2} d^{-\alpha+\frac{1}{2}} \bar{c}_d\right)\right),$$

where $\Phi(\cdot)$ indicates the standard Gaussian cumulative. Furthermore for x going to infinity, it holds $1 - \Phi(x) \asymp \frac{\phi(x)}{x}$, where ϕ denotes the density function of the standard Gaussian, it follows

$$\Pr_{max}(p, q) \asymp \frac{d^{\alpha-\frac{1}{2}}}{\sqrt{\pi}\bar{c}_d} e^{-\frac{\bar{c}_d^2}{\sqrt{2}} d^{-2\alpha+1}}.$$

On the other hand, considering the product of independent maximal couplings, the argument of each erf function goes to 0 as $d \to +\infty$ and hence we exploit the same expansion as in (2.68), getting

$$\begin{aligned}\prod_{i=1}^{d} \Pr_{max}(p_i, q_i) &= \prod_{i=1}^{d} \left(1 - erf\left(d^{-\alpha}\sqrt{\frac{c_i^2}{8}}\right)\right) \\ &= \prod_{i=1}^{d} \left(1 - \frac{|c_i|}{\sqrt{2\pi}} d^{-\alpha} + o(d^{-\alpha})\right) \asymp e^{-d^{1-\alpha}\tilde{c}_d},\end{aligned}$$

where $\tilde{c}_d = \frac{\sum_{i=1}^{d} |c_i|}{d\sqrt{2\pi}}$. □

References

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

Ascolani, F., Lavenant, H., and Zanella, G. (2024). Entropy contraction of the Gibbs sampler under log-concavity. *ArXiv preprint at: 2410.00858*.

Atchadé, Y. F. and Jacob, P. E. (2024). Unbiased Markov Chain Monte Carlo: what, why, and how. *Arxiv preprint at: 2406.06851v1*.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Biswas, N., Bhattacharya, A., Jacob, P., and Johndrow, J. (2022). Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84.

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). Estimating convergence of Markov chains with L-lag couplings. *Advances in neural information processing systems*, pages 7389–7399.

Brito, G., Dumitriu, I., and Harris, K. D. (2018). Spectral gap in random bipartite biregular graphs and applications. *Comb. Probab. Comput.*, 31:229–267.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32.

Ceriani, P. M. and Zanella, G. (2024). Linear-cost unbiased posterior estimates for crossed effects and matrix factorization models via couplings. *ArXiv preprint at: 2410.08939*.

Conti, G., Fruhwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). Bayesian exploratory factor analysis. *Journal of Econometrics*, 183(1):31–57.

Correa, J. R. and Romero, M. (2021). On the asymptotic behavior of the expectation of the maximum of i.i.d. random variables. *Operations Research Letters*, 49(5):785–786.

Douc, R., Jacob, P. E., Lee, A., and Vats, D. (2024). Solving the Poisson equation using coupled Markov chains. *ArXiv preprint at: 2206.05691*.

Dowson, D. C. and Landau, B. V. (1982). The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.*, 12(3):450–455.

Gao, K. and Owen, A. (2016). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics*, 11.

Gelfand, I. (1941). Normierte ringe. *Rec. Math. [Mat. Sbornik] N.S.*, 9(51)(1):3–24.

Gelman, A. (2005). Analysis of Variance: Why It Is More Important than Ever. *Annals of Statistics*, pages 1–31.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Ghosh, S., Hastie, T. J., and Owen, A. B. (2022). Backfitting for large scale crossed random effects regressions. *The Annals of Statistics*.

Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389.

Gorsuch, R. L. (2014). *Factor Analysis: Classic Edition (2nd ed.)*. Routledge.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600.

Jiang, J. and Nguyen, T. (2021). *Linear and Generalized Linear Mixed Models and Their Applications*.

Johnson, V. E. (1998). A Coupling-Regeneration Scheme for Diagnosing Convergence in Markov Chain Monte Carlo Algorithms. *Journal of the American Statistical Association*, 93(441):238–248.

Khare, K. and Zhou, H. (2009a). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Annals of Applied Probability*, 19.

Khare, K. and Zhou, H. (2009b). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Annals of Applied Probability*, 19:737–777.

Knothe, H. (1957). Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Menictas, M., Di Credico, G., and Wand, M. P. (2023). Streamlined variational inference for linear mixed models with crossed random effects. *Journal of Computational and Graphical Statistics*, 32(1):99–115.

Miller, J. W. and Carter, S. L. (2020). Inference in generalized bilinear models. *Arxiv preprint at: 2010.04896*.

Olkin, I. and Pukelsheim, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.*, 48:257–263.

Pandolfi, A., Papaspiliopoulos, O., and Zanella, G. (2024). Conjugate gradient methods for high-dimensional GLMMs. *Arxiv preprint at: 2411.04729*.

Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2019). Scalable inference for crossed random effects models. *Biometrika*, 107(1):25–40.

Papaspiliopoulos, O., Stumpf-Fétizon, T., and Zanella, G. (2023). Scalable Bayesian computation for crossed and nested hierarchical models. *Electronic Journal of Statistics*, 17(2):3575 – 3612.

Papastamoulis, P. and Ntzoufras, I. (2022). On the identifiability of Bayesian factor analytic models. *Statistics and Computing*, 32(23).

Papp, T. P. and Sherlock, C. (2022). A new and asymptotically optimally contracting coupling for the random walk Metropolis. *Arxiv preprint at: 2211.12585*.

Perry, P. O. (2017). Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):267–291.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):291–317.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Rosenthal, J. S. (2003). Asymptotic Variance and Convergence Rates of Nearly-Periodic Markov Chain Monte Carlo Algorithms. *Journal of the American Statistical Association*, 98(461):169–177.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. volume 25, pages 880–887.

Santambrogio, F. (2015). Optimal transport for applied mathematicians.

Wang, G., O'Leary, J., and Jacob, P. (2021). Maximal Couplings of the Metropolis-Hastings Algorithm. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR*, pages 1225–1233.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R, Second Edition (2nd ed.)*. Chapman and Hall/CRC.

Chapter 3

Dimension-free convergence of coordinate-ascent variational inference algorithms for large hierarchical models

Chapter 3 is devoted to the study of the computational cost associated to the estimation of variational approximations for large scale Bayesian model. We will focus on Mean Field Variational Inference (MF-VI), a popular alternative to MCMC for approximating posterior distribution in complex models. The associated optimization task of finding the KL-optimal distribution within the family of fully factorized distributions is often performed via the Coordinate Ascent VI algorithm (CAVI). Despite its popularity, there are still few quantitative results on the convergence properties of the algorithm itself. We investigate how the speed of convergence of CAVI is affected by the dimensionality of the latent space for two-level hierarchical models (also called global-local models). In particular, we obtain dimension-free convergence results under random data generating assumptions in asymptotic settings where both the number of data and parameters grows, with minimal assumptions on the model. Such results are substantiated by extensive simulations on synthetic and real datasets.

After a brief review on the main contributions on the theoretical properties of VI in

Section 3.1, we provide basic knowledge on variational methods in Section 3.2, focusing on the Mean Field family and CAVI algorithm in Section 3.2.1 and Section 3.2.2 respectively. The bulk of our contributions are contained in Section 3.3, presenting our ongoing research on the scalability of CAVI for large scale hierarchical models.

3.1 Introduction

Variational inference (Bishop, 2006; Jordan et al., 1999) emerged in last decades as a promising alternative framework for posterior estimation in the Bayesian setting with large scale models. In the Bayesian context, it seeks to approximate difficult to compute posterior densities with the closest approximating density belonging to a pre-defined family (Blei et al., 2017). Shifting the problem from stochastic approximation to deterministic optimization, variational methods benefit from the advances in stochastic and distributed optimization (Kingma and Ba, 2014; Kushner and Yin, 2003). Differently from standard MCMC, that under mild regularity assumptions are guaranteed to converge to the invariant distribution (Meyn et al., 2009; Roberts and Rosenthal, 2004; Roberts and Smith, 1994), VI has been studied less rigorously than MCMC, and its statistical properties are less well understood.

Seminal contributions were mainly focusing on investigating large sample properties of the variational optimizer for fixed dimensional target. Specifically, the majority of the works developed in the past decade were model specific results valid for a single variational family: their approach was to consider the VI posterior means as point estimates and show they enjoy the usual frequentist asymptotics (Hall et al., 2011; Celisse et al., 2012; Bickel et al., 2012). Recently, Yang et al. (2020) produced variational inequalities for a new class of variational approximations, termed α-VB, implying that point estimates constructed from the α-VB procedure converge at an optimal rate to the true parameter in a wide range of problems (i.e. the same rate as those constructed from the actual posterior).

Among different possible approximation families, the mean field (MF) family is one of the most widely used. Originating from statistical physics for physical models of ferromagnetism (Parisi, 1988; Kadanoff, 2009), it became soon popular as approximating

family due to its computational advantages. For this variational family, Wang and Blei (2019a) established the consistency and asymptotic normality of the VB posterior (i.e. the minimizer of the KL with the posterior, among the MF variational family), assuming data are generated from the model with a true parameter, and in particular they showed that the VB posterior converges to the Kullback-Leibler (KL) minimizer of a normal distribution, centred at the truth and that the corresponding variational expectation of the parameter is consistent and asymptotically normal. The result has been later generalized to the misspecified setting in Wang and Blei (2019b), where the authors proved that the VB posterior is asymptotically normal and centers at the value that minimizes the Kullback-Leibler (KL) divergence to the true data-generating distribution and moreover that the VB posterior mean centers at the same value and is also asymptotically normal. Ray and Szabó (2022) considered the case of high dimensional linear regression with sparse priors and provided for this model oracle inequalities for the mean-field VB approximation, implying that it converges to the sparse truth at the optimal rate and gives optimal prediction of the response vector.

In many contexts the MF minimizer is computed using the Coordinate Ascent Variational Inference (CAVI) algorithm (Bishop, 2006), a coordinate wise algorithm well suited for distributions belonging to the exponential family (Barndorff-Nielsen, 2014). A first specific study of the induced asymptotic properties has been done in Zhang and Gao (2020): they considered the mean field method for community detection under the stochastic block model and showed that an iterative batch CAVI algorithm shows linear convergence rate and converges to the minimax rate within log n iterations. More recently Bhattacharya et al. (2023), focusing on the two-block case, analysed the convergence of CAVI providing general conditions for certifying global or local exponential convergence. From a more mathematical standpoint, under (strong) log concavity of the approximating density, Jiang et al. (2024) proved the convergence of a restricted family of distributions, while under the same assumptions Arnese and Lacker (2024); Lavenant and Zanella (2024) proved the convergence of CAVI respectively for the deterministic and random scan version, and provided conditions for linear and exponential convergence of the approximating densities.

Along a different line of research, the exceptional work of Ascolani and Zanella (2024)

proved the dimension free convergence of the Gibbs sampling for a broad class of hierarchical models with growing number of parameter and observations under random data generating assumptions. Leveraging the known connection between Gibbs samplers and CAVI algorithm for MF-VI, in this Chapter we compute explicit convergence rate for the CAVI algorithm and prove dimension free convergence of the CAVI iterates to the CAVI solution for the same class of models, under random data generating assumptions. Our results suggest a strong agreement with the existing literature on convergence of other coordinate-wise methods on the same class of models, specifically with the work of Liu (1994) on the Bayesian Fraction of missing information for Gibbs samplers and Meng and Rubin (1994) on the Fraction of missing information for the study of Expectation Maximization algorithms Dempster et al. (1977), with the notable difference that our asymptotic characterization allows for explicit expression of the limiting convergence rate.

Consequently, CAVI offers an efficient and accurate means of approximating posterior distributions for this vast class of models.

3.2 Variational Inference

Consider $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ a set of observations and let $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_d)$ be the correspondent latent variables. A Bayesian model specifies a prior distribution $p(\mathbf{z})$, a conditional likelihood $p(\mathbf{x}|\mathbf{z})$, and then compute the posterior

$$p(\mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{z})p(\mathbf{z}). \tag{3.1}$$

Whenever the latter is not available in closed form, variational inference comes at hand: once chosen a family of approximating distributions $\mathcal{Q}$, it suffices to find $q^*(\mathbf{z})$ s.t.

$$q^*(\mathbf{z}) = argmin_{q(\mathbf{z})\in\mathcal{Q}}\, d\left(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})\right), \tag{3.2}$$

where d is some notion of distance between distributions. The most popular choice in practice is the "backward" Kullback-Leibler divergence, defined as

$$KL(q\|p) = \int \log\left(\frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}) d\mathbf{z}. \tag{3.3}$$

$KL(q\|p)$ basically measures the discrepancy between the approximating and the true log densities. Other valid choices include the "forward" $KL(p\|q)$, used in the context of expectation propagation (Vehtari et al., 2020), the Renyi or α-divergences $R_\alpha(p\|q)$ (Minka, 2005; Li and Turner, 2016), which interpolate between forward and backward KL when $\alpha \in (0, 1)$, and score-based (or Fisher) divergences (Cai et al., 2024), which measure the discrepancy between $\nabla \log(p)$ and $\nabla \log(q)$.

Variational inference hence turns a probabilistic problem (the determination of the posterior distribution) into an optimization one. The most popular techniques are the mean field approximation (Bishop, 2006, Chapter 10.1) and stochastic variational inference (Hoffman et al., 2013). For the effects of different divergence choices on uncertainty quantification, we refer the reader to Margossian et al. (2024).

In practice, instead of minimizing the KL, most methods maximize a tightly related quantity, namely the *evidence lower bound* (ELBO). Given a model as in (3.1), the ELBO is defined as

$$ELBO(q) := \int q(\mathbf{z}) \ln\left(\frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})}\right) d\mathbf{z}. \tag{3.4}$$

Differently from the KL, the expression in (3.4) only depends explicitly on the approximating density and quantities specified by the model itself. Furthermore it is directly linked to the expression of the KL in (3.3): the log marginal probability of (3.1) can be decomposed as

$$\ln(p(\mathbf{x})) = \int \ln(p(\mathbf{x})) q(\mathbf{z}) d\mathbf{z} = \int \ln\left(\frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right) q(\mathbf{z}) d\mathbf{z}$$

$$= ELBO(q) + KL(q\|p).$$

Given that the log likelihood depends only on the observed data and not affected by q,

minimizing the KL divergence is exactly equivalent to maximizing the ELBO, that only depends on computable quantities.

3.2.1 Mean Field Approximation

In Mean Field Variational Inference (MF-VI) (Bishop, 2006; Blei et al., 2017) the choice of the approximating family $\mathcal{Q}$ is the one of completely factorized distributions. Specifically, suppose $\mathbf{z} \in \mathcal{X}_1 \times ... \times \mathcal{X}_I$, with $X_i \subset \mathbb{R}^{d_i}$ and $\sum_{i=1}^{I} d_i = d$, then define

$$\mathcal{Q}_{MF} := \{q : q(\mathbf{z}) = q_1 \otimes ... \otimes q_d(\mathbf{z}) : q(\mathbf{z}) \ll p(\mathbf{z}|x) \text{ and } KL(q\|p) < \infty\}. \tag{3.5}$$

Note that in the present variational formulation no further assumptions are made about the distributions, in particular, no restrictions on the functional forms of the single factors are assumed.

Even with the restrictions on $\mathcal{Q}$, the optimization problem is not easy to solve in general: the objective function is convex, but we seek to minimize over the space of factorized distribution, that is not convex since a mixture of product distributions is not a product distribution in general. It follows that multiple local minima (Wainwright and Jordan, 2008, Section 5.4) might exist. In order to solve the optimization problem, one usually resorts to iterative methods such as coordinate ascent (descent), i.e. maximizing (minimizing) the objective function along one coordinate, keeping all the others fixed, and iterate along all axes. The procedure is called "Coordinate Ascent Variational Inference" (CAVI), and can be implemented sequentially on all coordinates in a predetermined or random order. Either way, it is guaranteed to converge to a local optimum under mild regularity assumptions, since intuitively each update step in CAVI is designed to increase the ELBO in (3.4) which is bounded from above, hence the procedure must land sooner or later in an optimum. The main appeal of CAVI is that the coordinate updates are explicit, indeed it is easy to show that each update has the form reported in Proposition 2.

Proposition 2. *Given a model as in* (3.1) *and the variational family* $\mathcal{Q}$ *of* (3.5)*, fix*

$(q_j^*)_{j\neq i}$, *the KL minimizer for the i-th coordinate*

$$q_i^*(\boldsymbol{z}_i) = argmin_{q_i\in\mathcal{P}(\mathcal{X}_i)} KL\left(q_i^*(\boldsymbol{z}_i)\prod_{j\neq i} q_j(\boldsymbol{z}_j)||p(\boldsymbol{z}|\boldsymbol{x})\right),$$

satisfies

$$q_i^*(\boldsymbol{z}_i) \propto \exp \mathbb{E}_{\boldsymbol{z}_{-i}\sim q_{-i}}\Big[\log\left(p(\boldsymbol{z}_i|\boldsymbol{z}_{-i},\boldsymbol{x})\right)\Big], \tag{3.6}$$

where the expectation in (3.6) *is with respect to all components except the i-th.*

However note that since the denominator of the conditional does not directly depend on z_i, one can also write

$$q_i^*(\mathbf{z}_i) \propto \exp \mathbb{E}_{q_{-i}}\Big[\log\left(p(\mathbf{z},\mathbf{x})\right)\Big].$$

We report in Algorithm 11 a general implementation of the scheme.

Algorithm 11: CAVI algorithm for general densities

Input: densities $p(\mathbf{z}|\mathbf{x})$ and $(q_i)_{i=1}^d$;
while *not converged* **do**
  for $i = 1, ..., I$ **do**
    update $q_i^*(\mathbf{z}_i) \propto \exp \mathbb{E}_{q_{-i}}\Big[\log\left(p(\mathbf{z}_i|\mathbf{z}_{-i},\mathbf{x})\right)\Big]$;
Output: approximating densities $(q_i)_{i=1}^d$;

Proposition 2 and (3.6) highlight a direct connection with Gibbs sampling: both methods involve full conditionals, but while the latter actually requires sampling from them, CAVI at each iteration only retains the form of the full conditional and optimize on it. It is then reasonable to expect that CAVI approach leads to tractable solutions when the Gibbs does and with similar convergence properties, as we later show in Section 3.3.

3.2.2 CAVI for Exponential family

The CAVI algorithm of Section 3.2.1 becomes even more amenable when approximating distributions with full conditionals belonging to the exponential family, that is $p(\mathbf{z}_i|\mathbf{z}_{-i},\mathbf{x})$ belongs to the exponential family for all i. To set notation, let $\mathbf{z} \in \mathbb{R}^d$ be a random vector,

we say it belongs to the exponential family (in its canonical form) if its density function can be written as

$$f(\mathbf{z}) = h(\mathbf{z}) \exp\{\sum_{s=1}^{S} \eta_s \cdot T_s(\mathbf{z}) - A(\boldsymbol{\eta})\},$$

where $h : \mathbb{R}^d \mapsto \mathbb{R}$ is a non negative function, $\boldsymbol{\eta} \in \mathbb{R}^S$ is the *canonical parameter* of the family, $\mathbf{T} = (T_1, ..., T_S)$ for $T_i : \mathbb{R}^d \mapsto \mathbb{R}$ $i = 1, ..., S$ is called *sufficient statistics* and $A : \mathbb{R}^S \to \mathbb{R}$ is the *log normalizer*. Recall, by standard properties of exponential family (Barndorff-Nielsen, 1978; Diaconis and Ylvisaker, 1979), it holds

$$\mathbb{E}[\mathbf{T}(\mathbf{z})] = \nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}), \qquad var\left(\mathbf{T}(\mathbf{z})\right) = \mathcal{H}\left(A(\boldsymbol{\eta})\right), \tag{3.7}$$

where $\mathcal{H}_{ij}(A(\boldsymbol{\eta})) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\boldsymbol{\eta})$.

Suppose that given a probabilistic model, each full conditional $p(\mathbf{z}_i|\mathbf{z}_{-i}, \mathbf{x})$ belongs to the exponential family. It holds that the updates of the CAVI algorithm preserve the form of the conditional and hence result in an even more appealing parametric update. Indeed for $S = 1$ (i.e. single parameter exponential family), one has

$$\mathcal{L}(\mathbf{z}_i|\mathbf{z}_{-i}) = h(\mathbf{z}_i) \exp\{\eta T(\mathbf{z}_i) - A(\eta)\},$$

where $\eta = \eta(\mathbf{z}_{-i})$, since the canonical parameter will depend on the conditioning variables. Then by Proposition 2:

$$q_i(\mathbf{z}_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\eta] T(\mathbf{z}_i) - A(\mathbb{E}_{q_{-i}}[\eta])\}, \tag{3.8}$$

i.e. the variational distribution for $\mathbf{z}_i$ is the full conditional with updated parameter being equal to the expectation, with respect to the variational distribution, of all the other variables. We report the resulting updating scheme in Algorithm 12.

Algorithm 12: CAVI algorithm for exponential family distribution

Input: $\boldsymbol{\eta}$;
for *t=1,...,T* **do**
  for *i=1,...,I* **do**
    update $\eta_i = \mathbb{E}_{\mathbf{z}_{-i} \sim q_{\eta_{-i}}}[\eta_i]$;
Output: updated $\boldsymbol{\eta}$.

3.3 Dimension-free convergence of CAVI for large hierarchical models

Consider models where the observed data $\mathbf{y}_{1:J}$ are divided in J groups of dimension M, each one characterized by local variables $\boldsymbol{\theta}_j \in \mathbb{R}^{d'}, j = 1, ..., J$ whose behaviour is governed by an hyperprior $\boldsymbol{\psi} \in \mathbb{R}^d$. In symbols

$$\begin{aligned} \mathbf{y}_{jm}|\boldsymbol{\theta}_j &\overset{\text{iid}}{\sim} f(\cdot|\boldsymbol{\theta}_j) \qquad m = 1, ..., M \; j = 1, ..., J, \\ \boldsymbol{\theta}_j|\boldsymbol{\psi} &\overset{\text{iid}}{\sim} p(\cdot|\boldsymbol{\psi}) \qquad j = 1, ..., J, \\ \boldsymbol{\psi} &\sim p_0(\cdot), \end{aligned} \tag{3.9}$$

for f, p, p_0 respectively some likelihood, local and global prior. Models as in (3.9) are workhorse of Bayesian Statistics and are commonly employed in many applied contexts (Gelman and Hill, 2006; Gelman et al., 2024). We consider variational approximations of the posterior density belonging to the mean field family $\mathcal{Q}_{MF}$ of (3.5), that becomes

$$\mathcal{Q}_{MF} = \{q : q(\boldsymbol{\psi}, \boldsymbol{\theta}) = q(\boldsymbol{\psi}) \prod_j q_j(\boldsymbol{\theta}_j)\}, \tag{3.10}$$

for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_J)$. We are interested in studying how the speed of convergence of the CAVI algorithm of Section 3.2.1 to the CAVI minimizer is affected by the number of data and parameter J going to infinity for models of (3.9), under random data generating assumptions, that is, supposing that there exists a distribution P s.t. $\mathbf{y}_j \overset{\text{iid}}{\sim} P$.

Review of the main contributions We first give quantitative results on the rate at which the variational means provided by the algorithm at iteration t converge to their limits for sufficiently large J. Given the recursive form of the CAVI updates in (3.6), the former amounts to studying an $O(J)$-dimensional recursion. Leveraging a dimension reduction trick similar to that of Ascolani and Zanella (2024), we show that the former characterization is sufficient to directly bound the KL divergence between the CAVI iterates and the optimizer as the number of data and parameters grow to infinity. The rest of the section is divided as follows: Section 3.3.1 contains notational

choices to facilitate the exposition of the subsequent results, Section 3.3.2 is devoted to the analysis of a simple conjugate model that already retains all the properties we will later investigate, while in Section 3.3.3 we briefly review some of the results in Ascolani and Zanella (2024) relevant for our work. We then report our novel results for univariate hyperprior in Section 3.3.4, both for Gaussian priors and more general ones. We finally extend the results to multivariate priors in Section 3.3.5, where some results need further investigations. Proofs are deferred to Appendix 3.5.

3.3.1 Notation

For models as in (3.9) with J latent variables and the variational family in (3.10), we denote by $q^t(\boldsymbol{\theta}, \boldsymbol{\psi}) = q^t(\boldsymbol{\psi}) \prod_{j=1}^J q_j^t(\theta_j)$ the variational distribution at iteration t of Algorithm 11, where the factorization follows by the definition of $\mathcal{Q}_{MF}$. The variational optimizer, defined as the distribution to which CAVI converges to (under mild regularity assumption it always exists) will be denoted as $q_J^*(\boldsymbol{\theta}, \boldsymbol{\psi}) = q^*(\boldsymbol{\psi}) \prod_{j=1}^J q_j^*(\boldsymbol{\theta}_j)$. Often we will denote $\mathcal{L}_v(\cdot)$ the variational law of the variable within brackets and $\mathbb{E}_v[\cdot]$ expectations with respect to it. For models with J groups, we denote by $\boldsymbol{\lambda}_J(t) = \mathbb{E}_v[\boldsymbol{\psi}]$ the mean value of $q^t(\boldsymbol{\psi})$, and write equivalently $q_{\boldsymbol{\lambda}_J(t)}(\boldsymbol{\psi}) = q^t(\boldsymbol{\psi})$ and $q^j_{\boldsymbol{\lambda}_J(t)}(\theta_j) = q_j^t(\theta_j)$, highlighting the dependence to the global parameter (see Section 3.3.2 for further details). The mean parameter of the CAVI solution is instead denoted as $\boldsymbol{\lambda}_J^*$. Lastly, we write

$$KL_J(t) := KL\left(q^t(\boldsymbol{\psi}) \prod_{j=1}^J q_j^t(\theta_j) \| q_J^*(\boldsymbol{\theta}, \boldsymbol{\psi})\right). \tag{3.11}$$

The statements of the following sections will be derived supposing that there exists either a true measure P with density p s.t. $\mathbf{y}_j \overset{\text{iid}}{\sim} P$ not necessarily belonging to the model in (3.9), i.e. the misspecified setting, or that there exists a true global parameter $\boldsymbol{\psi}^*$ and a true probability measure $G_{\boldsymbol{\psi}^*}$ s.t. $\mathbf{y}_j \overset{\text{iid}}{\sim} G_{\boldsymbol{\psi}^*}$ with density

$$g(\mathbf{y}|\boldsymbol{\psi}^*) = \int f(\mathbf{y}|\boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j|\boldsymbol{\psi}^*) d\boldsymbol{\theta}_j. \tag{3.12}$$

In either case, we denote the "true" marginal mean by $\boldsymbol{\mu}_P = \mathbb{E}_P[\mathbf{y}_j]$ and the "true" marginal variance covariance matrix as $\Sigma_P^2 = \mathbb{E}_P[(\mathbf{y}_j - \boldsymbol{\mu}_P)(\mathbf{y}_j - \boldsymbol{\mu}_P)^\top]$ for $\mathbf{y}_j \overset{\text{iid}}{\sim} P$.

3.3.2 Illustrative example for univariate hyperprior

We report below explicit results for a conjugate model for which the computations admit closed form solutions, but already exhibit the properties that will be central afterwards. Consider a hierarchical model as in (3.9), where f, p, p_0 are supposed to be Gaussian. In this particular case it is possible to explicitly characterize for every fixed J the approximating distribution at each iteration of the algorithm, the CAVI solution, the speed of convergence and also the limiting behaviour as J grows to infinity. Nonetheless this simplified setting provides useful insights for the proof techniques we will develop later. The model can be written as

$$\begin{aligned} y_{jm}|\theta_j &\overset{\text{iid}}{\sim} N(\theta_j, \tau_0^{-1}) \qquad m = 1, ..., M \; j = 1, ..., J, \\ \theta_j|\mu, \tau_1 &\overset{\text{iid}}{\sim} N(\mu, \tau_1^{-1}) \qquad j = 1, .., J, \\ \mu &\sim N(\mu_0, \tau_{00}^{-1}), \end{aligned} \tag{3.13}$$

where $\tau_0, \tau_1, \mu_0, \tau_{00}^{-1}$ are supposed to be fixed. Note that each full conditional belongs to the exponential family, and the results in Section 3.2.2 can be applied. Let $\pi_J(\mu, \theta_{1:J}|y_{1:J})$ be the posterior distribution of the model in (3.13). We are interested in studying the rate of convergence of $KL_J(t)$, as in (3.11), showing exponential convergence as J grows. In particular we first study the convergence of the variational parameter and then compose the result to bound the KL divergence. Set

$$\lambda_J(t) := \int \mu \; dq^t(\mu), \qquad\qquad \phi_j(t) := \int \theta_j \; dq_j^t(\theta_j) \;\text{ for } j = 1, ..., J,$$

where λ_J will be called *global* variational parameter in contrast with the *local* ϕ_j's. Initialize $\lambda_J(0)$, then sequentially update the ϕ_j's. Exploiting the formula in Proposition 2, it follows

$$q_j^{t+1}(\theta_j) = N\left(\frac{m\tau_0}{m\tau_0 + \tau_1}\bar{y}_j + \frac{\tau_1}{m\tau_0 + \tau_1}\lambda_J(t), (m\tau_0 + \tau_1)^{-1}\right) \qquad j = 1, ..., J, \tag{3.14}$$

$$q^{t+1}(\mu) = N\left(\frac{\tau_{00}\mu_0 + \tau_1 \sum_j \phi_j(t+1)}{\tau_{00} + J\tau_1}, (\tau_{00} + J\tau_1)^{-1}\right), \tag{3.15}$$

where $\bar{y}_j := (\sum_m y_{jm})/m$. Combining (3.14) and (3.15), it is possible to write a recursion solely in term of the global variation parameter λ_J, indeed

$$\begin{aligned}\lambda_J(t+1) &= g_J\left(\lambda_J(t), \mathbf{y}_{1:J}\right) \\ &:= \frac{\tau_{00}\mu_0}{\tau_{00}+J\tau_1} + \frac{J\tau_1}{(\tau_{00}+J\tau_1)(m\tau_0+\tau_1)}\left(m\tau_0\frac{\sum_j \bar{y}_j}{J} + \tau_1\lambda_J(t)\right),\end{aligned} \tag{3.16}$$

where $g_J : \mathbb{R} \times \mathbb{R}^{m\times J} \to \mathbb{R}$ is a function depending explicitly on J, the variational parameter λ and the observed data $\mathbf{y}_{1:J}$. The form of (3.16) will be central for the theory on recursions later developed in Section 3.3.4 and 3.3.5. Define

$$\rho_{cavi} := \partial_\lambda g_J(\lambda, \mathbf{y}_{1;J}) = \frac{J\tau_1}{(\tau_{00}+J\tau_1)}\frac{\tau_1}{(m\tau_0+\tau_1)}. \tag{3.17}$$

Then, since

$$\rho_{cavi} < 1 \quad \forall J \geq 1,$$

by the Banach fixed point theorem there exists a unique fixed point λ_J^* for (3.16) to which the iterations converge, satisfying

$$\lambda_J^* = g_J(\lambda_J^*, \mathbf{y}_{1:J}).$$

Solving the equation leads to

$$\begin{aligned}\lambda_J^* = \frac{m\tau_0+\tau_1}{m\tau_0(\tau_{00}+J\tau_1)+\tau_{00}\tau_1}&\left[\tau_{00}\mu_0 + \frac{m\tau_0\tau_1 J(\sum_j \bar{y}_j)/J}{m\tau_0+\tau_1}\right] \\ &\to \mu_P := \lambda^* \qquad\qquad P^{(\infty)} - a.s.,\end{aligned}$$

where the limiting result is obtained applying the strong law of large numbers on the $\mathbf{y}$'s, meaning $\frac{\sum_j \bar{y_j}}{J} \to \mu_P$. Furthermore (3.16) allows to characterise the speed of convergence to the fixed point. Note indeed that

$$\begin{aligned}\lambda_J(t) - \lambda_J^* &= \prod_{k=0}^{t-1}\frac{\lambda_J(k+1)-\lambda_J^*}{\lambda_J(k)-\lambda_J^*}(\lambda_J(0)-\lambda_J^*) \\ &= \left(\frac{J\tau_1^2}{(\tau_{00}+J\tau_1)(m\tau_0+\tau_1)}\right)^t(\lambda_J(0)-\lambda_J^*) < \rho^t(\lambda_J(0)-\lambda_J^*).\end{aligned} \tag{3.18}$$

where $\rho := \frac{\tau_1}{\tau_1 + m\tau_0}$, independently of J. The equation above certifies an exponential convergence of the global parameter to the fixed point. We now turn to the study of the convergence of the KL divergence. Recall that $KL\left(N(\mu_1, \sigma_1^2) \| N(\mu_2, \sigma_2^2)\right) = \frac{1}{2}\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) - \frac{1}{2} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2}$, then substituting with the expressions in (3.14) and (3.15), we obtain

$$\begin{aligned}
KL_J(t) &= KL\left(q^t(\mu) \| q^*(\mu)\right) + \sum_j KL\left(q_j^t(\theta_j) \| q_j^*(\theta_j)\right) \\
&= \frac{\tau_{00} + J\tau_1}{2}(\lambda_J(t) - \lambda_J^*)^2 + \sum_{j=1}^{J} \frac{\tau_1}{2}(\lambda_J(t) - \lambda_J^*)^2 \\
&= \left(\frac{\tau_{00} + J\tau_1}{2} + \frac{J\tau_1}{2}\right)(\lambda_J(t) - \lambda_J^*)^2 \\
&\leq \left(\frac{\tau_{00} + J\tau_1}{2} + \frac{J\tau_1}{2}\right)\rho^{2t}(\lambda_J(0) - \lambda_J^*)^2,
\end{aligned}$$

where the last inequality is obtained by leveraging (3.18). From the above, supposing $y_j \overset{\text{iid}}{\sim} P$, two claims are possible.

Proposition 3.3.1. *Consider the model in* (3.13)*, and define q_J^t and q_J^* as in Section 3.3.1. Suppose $\exists P$ s.t. $y_j \overset{iid}{\sim} P$, then under* warm start *assumptions, i.e. $|\lambda_J(0) - \lambda_J^*| < \frac{c}{\sqrt{J}}$, $\forall \varepsilon > 0$, $\exists T(\varepsilon)$ s.t.*

$$\lim_{J \to +\infty} KL(q_J^{T(\varepsilon)} \| q_J^*) < \varepsilon.$$

If instead $|\lambda_J(0) - \lambda_J^*| < c$, $\forall \varepsilon > 0$, $\exists T(\varepsilon, J)$ s.t.

$$\lim_{J \to +\infty} KL(q_J^{T(\varepsilon, J)} \| q_J^*) < \varepsilon,$$

and $T(\varepsilon, J) = O(\log(J))$.

Note that the above statement is completely deterministic since the rate of convergence does not depend on the observations. In the following sections instead the results will be under $P^{(\infty)}$ or $G_{\psi^*}^{(\infty)}$ probability, that is the one generated by an infinite sequence of $\mathbf{y} \overset{\text{iid}}{\sim} P$ or $\mathbf{y} \overset{\text{iid}}{\sim} G_{\psi^*}$, as in (3.12).

To illustrate the phenomenon, we report below the average number of iterations needed for convergence of CAVI on the model in (3.13), under both warm and cold start. For this model, the CAVI solutions and the expression of the KL divergence are available in closed form and we considered the algorithm as converged when the KL was

smaller than a chosen threshold. In the following section, we generalize the results to fam-

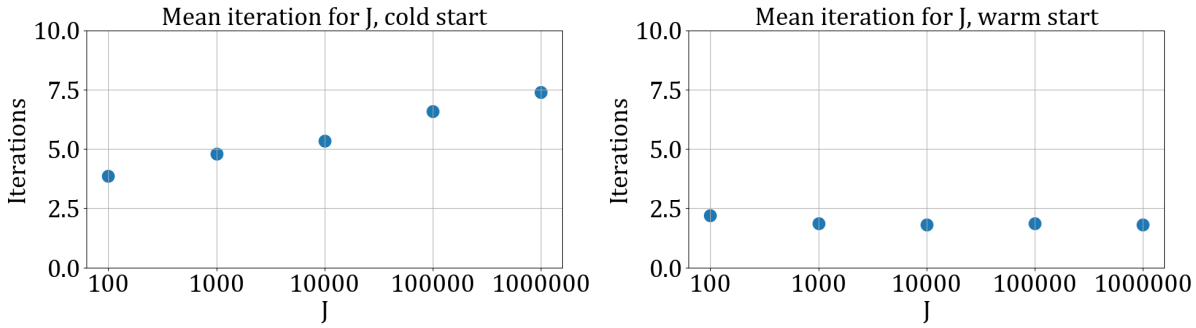

Figure 3.1: Iterations to convergence for hierarchical Gaussian as the number of latent varibles J grows, cold and warm start, log scale on the x axis.

ily of distributions for which it is prohibitively difficult to solve the fixed point equation in (3.16) and derive the inherent properties.

Remark 8 (Connection to alternative coordinate wise schemes)**.** For the model in (3.13) with fixed variances, Ascolani and Zanella (2024) characterized the convergence behaviour of a deterministic scan Gibbs samplers as J grows to infinity.

Lemma 15 (Ascolani and Zanella (2024) Section 4.3)**.** *Under random data generating assumptions, the convergence of the Gibbs algorithm for J large enough is characterized by*

$$\rho_{gibbs} = \frac{\mathbb{E}[var(T(\theta_j)|\mu^*, y_j)]}{var\left(\mathbb{E}[T(\theta_j)|\mu^*, y_j]\right) + \mathbb{E}[var(T(\theta_j)|\mu^*, y_j)]}.$$

In our notation, the result yields $\rho_{gibbs} = \frac{\tau_1}{m\tau_0+\tau_1}$. The previous expression coincides exactly with the convergence speed ρ_{cavi} of (3.17). This does not come as a surprise: Tan and Nott (2014) proved, exploiting the autoregressive formulation of Gaussian Gibbs sampler in Roberts and Sahu (1997), that for Gaussian targets CAVI and Gibbs sampler converge at the same rate. For more general hierarchical models, the convergence of Gibbs sampler is characterized by the maximal fraction of missing information , quantified as the extra variation caused by missing θ (Liu (1994))

$$\gamma_B = \inf_h \frac{var\left(\mathbb{E}[h(\boldsymbol{\mu})|\mathbf{Y}_{1:J}, \boldsymbol{\theta}_{1:J}]|\mathbf{Y}_{1:J}\right)}{var(h(\boldsymbol{\mu})|\mathbf{Y}_{1:J})} = \inf_h \frac{var\left(\mathbb{E}[h(\boldsymbol{\theta}_{1:J})|\mathbf{Y}_{1:J}, \boldsymbol{\mu}]|\mathbf{Y}_{1:J}\right)}{var(h(\boldsymbol{\theta}_{1:J})|\mathbf{Y}_{1:J})}.$$

The result of Ascolani and Zanella (2024) notably avoid the computation of an infimum over a set of test functions, but states that the infimum is attained exactly at the suffi-

cient statistic of $T(\theta_j)$. Considering instead the EM algorithm, (Meng and Rubin, 1994; Dempster et al., 1977) showed that the "local" rate of convergence is governed by the largest eigenvalue of the matrix of fractions of missing information due to incomplete data, namely

$$\gamma_l = I_{com}(\boldsymbol{\mu}^*)^{-1} I_{miss}(\boldsymbol{\mu}^*) \approx \frac{\tau_1}{\tau_1 + m\tau_0},$$

where the last approximate equality is obtained for J large enough. Sahu and Roberts (1999) showed indeed that the approximate rate of convergence of the Gibbs sampler by Gaussian approximation is equal to that of the corresponding EM-type algorithm.

3.3.3 Review on dimension-free mixing times of Gibbs sampler

Our approach for the analysis on CAVI on hierarchical models builds on the exceptional findings of Ascolani and Zanella (2024): although the theory we develop is in nature different from the one presented in the aforementioned work, the results we expect to achieve are similar.

The paper deals with application of Gibbs samplers to hierarchical models as those reported in (3.9) of the following section, and in particular shows dimension-free convergence results under mild assumptions on the likelihood function as the number of groups goes to infinity. Central point of the theory is that studying the properties of the Gibbs sampler as the number of parameters goes to infinity is equivalent to study the properties of the Gibbs scheme targeting the limiting distribution. Obviously, the latter has dimension that goes to infinity, but for hierarchical models with $p(\cdot|\psi)$ belonging to the exponential family, it is equivalent to study the Gibbs sampler on the fixed dimensional target $\hat{\pi}_J(d\mathbf{T}, d\psi) := \mathcal{L}(d\mathbf{T}, d\psi|Y_{1:J})$, where $\mathbf{T}$ is a set of sufficient statistics $\mathbf{T} = \mathbf{T}(\boldsymbol{\theta})$, whose dimensionality does not depend on J, such that $\mathcal{L}(d\psi|\boldsymbol{\theta}, \mathbf{Y}_{1:J}) = \mathcal{L}(d\psi|\mathbf{T}(\boldsymbol{\theta}), \mathbf{Y}_{1:J})$ (Lemma 4.1 Section 4.1). Usually the target distribution $\hat{\pi}_J$ is not available in closed form, and the corresponding two-block Gibbs sampler P^J cannot be implemented directly, but is possible to characterise the asymptotic distributions of the involved full conditionals (that we will later exploit in our proofs), namely $\mathcal{L}(\mathbf{T}|\psi, \mathbf{Y}_{1:J})$ and $\mathcal{L}(\psi|\mathbf{T}, \mathbf{Y}_{1:J})$, and the mixing properties of the related Gibbs. In particular, Proposition 4.5 shows that a suitable rescaling of $(\mathbf{T}, \psi)$ converges to a multivariate Gaussian distribution in total

variation distance. The proof combines a BvM for ψ with a less standard Central Limit Theorem for $\mathbf{T}$ conditional on ψ (Lemma 4.4).

3.3.4 Theoretical results for univariate hyperprior

For model (3.9) with univariate hyper priors, the iterative equations of CAVI have a form particularly amenable for the theoretical analysis we have in mind. We start by proving the theory for the case of Gaussian prior and hyper-prior, and general likelihood. We later extend the result for prior belonging to exponential family and conjugate hyperprior.

Gaussian prior and hyperprior

In the following we consider the estimation of a Bayesian model of the form

$$\begin{aligned} \mathbf{y}_j \mid \theta_j &\overset{\text{iid}}{\sim} f(y \mid \theta_j) && j = 1, \dots, J, \\ \theta_j \mid \mu &\overset{\text{iid}}{\sim} N(\theta \mid \mu, \tau_0^2) && j = 1, \dots, J, \\ \mu &\sim N(\mu \mid \mu_0, \sigma_0^2). \end{aligned} \tag{3.19}$$

The CAVI recursive equations can be written as

$$\lambda_J(t+1) = \frac{a_J}{J} + \frac{1}{J} \sum_{j=1}^{J} T_J \left(\mathbf{y}_j, \lambda_J(t) \right), \tag{3.20}$$

where $J, t = 1, 2, \dots$ and $T_J : \mathbf{Y} \times \mathbb{R} \to \mathbb{R}$ measurable function for all J. The form of (3.20) is inspired by (3.16). We now state the necessary assumptions for deriving our result.

($A1$) there exists a measure P s.t. $\mathbf{y}_j \overset{\text{iid}}{\sim} P$.

($A2$) For every $J \geq 1$ there exists a fixed point λ_J^* of equation (3.20). Moreover, there exists λ^* such that $\lambda_J^* \to \lambda^* \in \mathbb{R}$ P-almost surely as $J \to \infty$.

($A3$) Let $B_\epsilon^* = \{\lambda \mid |\lambda - \lambda^*| < \epsilon\}$, for $\epsilon > 0$. Then there exist $\delta > 0$ and $\rho_J < 1$ such that

$$\mathbb{E}_P \left[\sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(\mathbf{y}_j, \lambda)| \right] < \rho_J.$$

and suppose that $\exists \rho < 1$ s.t.

$$\lim_{J \to +\infty} \rho_J = \rho. \tag{3.21}$$

$(A4)$ define $\mu_J = \mathbb{E}_P[\sup_{B^*_\delta} |\partial_\lambda T_J(\mathbf{y}_j, \lambda)|] < \rho_J$ by (A3), and suppose:

$$\mathbb{E}_P[|\sup_{B^*_\delta} |\partial_\lambda T_J(\mathbf{y}_j, \lambda_J(t))| - \mu_J|^4] \leq b \qquad \forall J.$$

Note that assumptions (A1)-(A2) are the natural setting for the kind of analysis we develop, (A3) is needed to ensure convergence of the algorithm for big J and we will prove that it holds for models as those in (3.13), (A4) is instead a technical assumption required for the application of a strong law of large numbers type result on the T_Js of (3.20). Then we can prove

Lemma 16. *Consider equation* (3.20) *and assume* $(A1) - (A4)$ *are satisfied. Thus, there exists* $c > 0$ *such that, if* $|\lambda_J(0) - \lambda^*_J| < c$*, for almost every sequence* $\boldsymbol{y} = \boldsymbol{y}_{1:\infty}$ *there exists* $J^* = J^*(\boldsymbol{y})$ *such that*

$$|\lambda_J(t+1) - \lambda^*_J| \leq \tilde{\rho}^{t+1} |\lambda_J(0) - \lambda^*_J|.$$

for every $J \geq J^*$ *and* $t \geq 1$ *for* $\tilde{\rho}$ *as close as desired to* ρ *(as defined in* (3.21)*).*

Lemma 16 above provides insights on the limiting behaviour of CAVI iterates, bounding the distance to the local minimizer. We now exploit this notion to directly study the behaviour of the KL divergence. Before stating the main theorem we need some additional assumptions. With the same notation as in Section 3.3.1, let G_μ denote the marginal likelihood of the model in (3.19) and denote its density, obtained by integrating out the group specific parameter θ, as

$$g(\mathbf{y} \mid \mu) = \int_{\mathbb{R}} f(\mathbf{y} \mid \theta) p(\theta \mid \mu) \, \mathrm{d}\theta \, . \tag{3.22}$$

Suppose

$(B1)$ There exists μ^* s.t. $\mathbf{y}_j \overset{\text{iid}}{\sim} G_{\mu^*}$, for G_{μ^*} with density as in (3.12).

$(B2)$ For every J and $\mathbf{y}_{1:J}$, the λ^* of (A2) is such that $\lambda^* = \mu^*$. Moreover, there exists a finite random variable X such that $\sqrt{J}\,(\lambda^*_J - \mu^*) \to X$ in distribution, as $J \to \infty$.

($B3$) The function $\mathbb{E}\left[\theta_j \mid \mathbf{y}_j, \mu\right]$ is continuously differentiable with respect to μ, for every $\mathbf{y}_j$.

($B4$) There exist a neighbourhood Ψ^* of λ^* and constants $C_j = C(\mathbf{y}_j)$ such that the functions $KL_j(\lambda\|\lambda_J^*) := KL(q_J^{t+1}(\theta_j)\|q_J^*(\theta_j))$ are twice differentiable with respect to λ and $\partial_\lambda^2 KL_j\left(\lambda\|\lambda_J^*\right) < C$ for every $\lambda \in \Psi^*$. Moreover $\frac{1}{J}\sum_{j=1}^J C_j \leq \bar{C} < \infty$, $G_{\mu^*}^{(\infty)}$-almost surely.

Theorem 1. *Under (A1)-(A4) and (B1)-(B4), under* warm start *assumptions, i.e. provided* $|\lambda_J(0) - \lambda_J^*| < \frac{c}{\sqrt{J}}$, *for hierarchical models of* (3.19), *then* $\forall \varepsilon > 0$, $\exists T(\varepsilon)$ *(independent of* J*) s.t.*

$$\Pr\left(KL(q_J^{T(\varepsilon)}\|q_J^*) < \varepsilon\right) \to 1 \qquad G_{\mu^*}^{(\infty)} - a.s., \ as\ J \to +\infty$$

If instead one supposes cold start, *i.e.* $|\lambda_J(0) - \lambda_J^*| < c$, $\forall \varepsilon > 0$, $\exists T(\varepsilon, J)$ *s.t.:*

$$\Pr\left(KL(q_J^{T(\varepsilon,J)}\|q_J^*) < \varepsilon\right) \to 1 \qquad G_{\mu^*}^{(\infty)} - a.s.$$

where $G_{\mu^*}^{(\infty)}$ *is the law of the infite sequence* $\boldsymbol{y} = \boldsymbol{y}_{1:\infty}$.

Theorem 1 characterizes completely the behaviour of CAVI algorithm for certain large hierarchical models, it also guarantees the convergence to the minimizer in a finite number of iterations, independent of J under warm starts, or increasing logarithmically with J for standard initializations. In Section 3.3.4 we present ongoing research on the generalization of models to incorporate general hyper priors.

The theorem above can actually be extended to the case of misspecified likelihood, indeed suppose

($B1^*$) there exists P s.t. $\mathbf{y} \overset{\text{iid}}{\sim} P$ and define μ^* as the model parameter s.t. $\mu^* = \text{argmin}_\mu KL\left(P\|G_\mu\right)$, where G_μ admits density $g(\mathbf{y} \mid \mu)$ as in (3.12).

Then one can prove

Corollary 3. *Under (A1)-(A4) and (B1*)-(B4) for the same setting as in Theorem 1,*

under warm start *assumptions,* $\forall \varepsilon > 0$, $\exists T(\varepsilon)$ *(independent of* J*) s.t.*

$$\Pr\left(KL(q_J^{T(\varepsilon)} \| q_J^*) < \varepsilon\right) \to 1 \qquad P^{(\infty)} - a.s.$$

Under cold start*,* $\forall \varepsilon > 0$, $\exists T(\varepsilon, J)$ *s.t.:*

$$\Pr\left(KL(q_J^{T(\varepsilon,J)} \| q_J^*) < \varepsilon\right) \to 1 \qquad P^{(\infty)} - a.s.$$

Extension to General likelihood

We consider now an extension of the model in (3.19), allowing for general univariate hyperprior, i.e.

$$\begin{aligned} y_j \mid \theta_j &\sim f(y \mid \theta_j) && j = 1, \dots, J, \\ \mathcal{L}(\theta_j \mid \mu) &= h(\theta_j) \exp\{\mu T(\theta_j) - A(\mu)\} && j = 1, \dots, J, \\ \mu &\sim p(\mu). \end{aligned} \tag{3.23}$$

It is possible to show that the CAVI updates in this case can again be written as a difference equation on the scalar parameter $\lambda_J(t)$. We defer to later works the proof of the KL convergence, although it comes from a straightforward generalization of Theorem 1. In detail

$$\lambda_J(t) = g_J\left(\lambda_J(t-1), \mathbf{y}_{1:J}\right), \tag{3.24}$$

where g_J is a real-valued function depending on random elements $y_j \overset{\text{iid}}{\sim} P$ defined on the same probability space. We report in Lemma 16 a proof of the exponential convergence for such recursion. The result could then be applied to the recursion originated from the model in (3.28), the proof is already available up to a necessary Bernstein-von Mises like theorem for the variational distribution, we defer such a proof to next works.

Consider the assumptions (A1)-(A2) of Lemma 16, furthermore assume

($C1$) The function $g_J(\lambda, \mathbf{y}_{1:J})$ is continuously differentiable as a function of λ, for every J and $\mathbf{y}_{1:J}$.

($C2$) Let $B_\varepsilon^* = \{\lambda : |\lambda - \lambda^*| < \varepsilon\}$, there exists $\delta > 0$ such that

($C2.1$) There exists a real-valued function h such that $\partial_\lambda g_J(\lambda, \mathbf{y}_{1:J}) \to h(\lambda)$ uniformly $P^{(\infty)}$-almost surely as $J \to \infty$, for every $\lambda \in B_\delta^*$.

($C2.2$) There exists a constant $\rho \in (0,1)$ such that $h(\lambda) \leq \rho$ for every $\lambda \in B_\delta^*$.

($C2.3$) There exists a constant $C < \infty$ such that $\partial_\lambda g_J(\lambda, \mathbf{y}_{1:J}) \leq C$ for every $\lambda \in B_\delta^*$ and $\mathbf{y}_{1:J}$.

Proposition 3.3.2. *Consider a recursion as in* (3.24)*, then under* $(A1) - (A2)$ *and* $(C1) - (C2)$*, there exists* $c > 0$ *such that, if* $|\lambda_J(0) - \lambda_J^*| < c$*, for almost every sequence* $\boldsymbol{y} = \boldsymbol{y}_{1:\infty}$ *there exists* $J^* = J^*(\boldsymbol{y})$ *such that*

$$|\lambda_J(t+1) - \lambda_J^*| \leq \rho^{t+1} |\lambda(0) - \lambda_J^*|,$$

for every $J \geq J^*$ *and* $t \geq 1$*.*

3.3.5 Multivariate hyperprior

We now seek to extend the construction of the previous sections to the case of multivariate priors. As before, we start with a conjugate model where results are available in closed form, providing insights of the theory later developed. In particular, from the example below we derive the necessity to express the prior on the local level as a distribution belonging to the exponential family, written in its canonical form: this allows to write the CAVI updates as a linear recursion and hence exploit similar theory as that developed for the univariate recursions. Furthermore, the illustrative example shows the necessity to develop auxiliary results that allows neat explicit computations, momentarily used without theoretical justifications, namely the fact that under regularity assumptions the limiting fixed points are exactly the fixed points of the limiting distributions, and the limits of the derivative (or spectral radii of the Hessians) are equal to the derivative (Hessians) of the limiting recursion. For the moment, we only characterize the convergence of the variational parameter.

Illustrative example

Consider a model of the form:

$$\begin{aligned} y_{jm}|\theta_j &\overset{\text{iid}}{\sim} N(\theta_j, \sigma_r^2) && m = 1, ..., M \; j = 1, .., J, \\ \theta_j|\mu, \tau_0^2 &\overset{\text{iid}}{\sim} N(\mu, \tau_0^2) && j = 1, ..., J, \\ \mu, \tau_0^2|\mu_0, \sigma_0^2, \alpha_0, \beta_0 &\sim NIG(\mu_0, 1/\sigma_0^2, \alpha_0/2, \beta_0/2). \end{aligned} \tag{3.25}$$

Let

$$\lambda_J(t) := \int \mu \; dq^t(\mu), \qquad \nu_J(t) := \int \tau_0^2 \; dq^t(\tau_0^2), \qquad \xi_J(t) := \int \sigma_r^2 \; dq^t(\sigma_r^2). \tag{3.26}$$

We report the main computations in the Appendix. The limiting recursions are

$$\begin{aligned} \lambda^* &= \frac{\nu^*}{M\nu^* + \xi^*}\mu_Q + \frac{\lambda^*\xi^*}{M\nu^* + \xi^*}; \\ \xi^* &= \frac{\xi^*\nu^*}{(\xi^* + M\nu^*)} + \frac{1}{(\xi^* + M\nu^*)^2}\left(\nu^2 M(M-1)\sigma_y^2 + \xi^2(\sigma_y^2 + \tau_0^2) + 2\nu\xi\sigma_y^2(M-1)\right); \\ \nu^* &= \frac{M(\nu^*)^2(\sigma_y^2 + M\tau_0^2)}{(M\nu^* + \xi^*)^2} + \frac{\nu^*\xi^*}{M\nu^* + \xi^*}; \end{aligned}$$

where we derived from (3.46)

$$\mathbb{E}_Q[\sum_m (\sum_{\tilde{m}} y_{j\tilde{m}} - My_{jm})^2] = M^2(M-1)\sigma_y^2,$$

$$\mathbb{E}_Q[\sum_m (y_{jm} - \mu_Q)^2] = M(\sigma_y^2 + \tau_0^2),$$

and exploited the fact that from the first recursion we get

$$\lambda^* = \mu_Q.$$

Solving the system of fixed point equations for $M > 1$ we get

$$\lambda^* = \bar{y}, \qquad \xi^* = \sigma_y^2, \qquad \nu^* = \tau_0^2,$$

or

$$\lambda^* = \bar{y}, \qquad \xi^* = \sigma_y^2 + \tau_0^2, \qquad \nu^* = 0,$$

While for $M = 1$:

$$\lambda^* = \bar{y}, \qquad \xi^* + \nu^* = \sigma_y^2 + \tau_0^2.$$

As for the KL divergence, exploiting the factorization of the mean field solution, we have

$$KL(q^t(\mu,\tau_0^2,\sigma_r^2)\prod_j q_j^t(\theta_{1:J})\|q_J^*(\mu,\tau_0^2,\theta_{1:J})) = \\ KL\left(q^t(\mu,\tau_0^2)\|q_J^*(\mu,\tau_0^2)\right) + KL\left(q^t(\sigma_r^2)|q_J^*(\sigma_r^2)\right) + \sum_j KL\left(q_j^t(\theta_j)\|q_j^*(\theta_j)\right). \tag{3.27}$$

And taking the limit as $J \to +\infty$ it is easy to see that it displays the same rate of convergence to zero as the slowest between the hyperparameters. In Figure 3.2 we report the distance in log scale between the variational parameters and the variational minimizer, left panel: cold start, right panel: warm start.

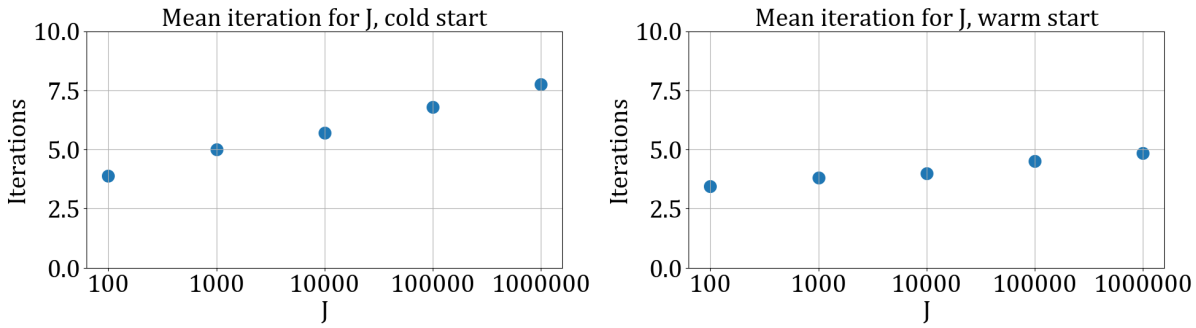

Figure 3.2: Iterations to convergence for hierarchical Gaussian models, different data and parameter number, cold and warm start, M=10.

Note that for the computations above we solved the recursion in (3.48), whose form differs from that of the previous sections because of the cross terms in the $\mathbf{y}_j$'s. Such a complication can be avoided expressing the prior for the local parameters in the canonical form of the corresponding exponential family distribution, as made clear in the following section.

Multivariate hyperprior: theory

Consider hierarchical models of the form

$$\begin{aligned} \mathbf{y}_j \mid \boldsymbol{\theta}_j &\overset{\text{iid}}{\sim} f(\mathbf{y} \mid \boldsymbol{\theta}_j) && j = 1, \dots, J, \\ \mathcal{L}(\boldsymbol{\theta}_j \mid \boldsymbol{\mu}) &= h(\boldsymbol{\theta}_j) \exp\{\boldsymbol{\mu}^\top T(\boldsymbol{\theta}_j) - A(\boldsymbol{\mu})\} && j = 1, \dots, J, \\ \mathcal{L}(\boldsymbol{\mu}|\mathbf{t}_0, n_0) &\propto \exp\{\boldsymbol{\mu}^\top \mathbf{t}_0 n_0 - n_0 A(\boldsymbol{\mu})\}, \end{aligned} \tag{3.28}$$

for $\mathbf{t}_0, n_0$ fixed parameters, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\theta}_j \in \mathbb{R}^{d'} \, \forall j = 1, ..., J$. It is possible to show that, under assumption on the model, the recursive updates induced by the CAVI algorithm have the form:

$$\boldsymbol{\lambda}(t+1) = \frac{\mathbf{a}_J}{J} + \frac{1}{J} \sum_{j=1}^{J} T\left(\mathbf{y}_j, \boldsymbol{\lambda}(t)\right), \tag{3.29}$$

for $\boldsymbol{\lambda} \in \mathbb{R}^d$ where $J, t = 1, 2, \dots$ and $T : \mathbb{Y} \times \mathbb{R}^d \to \mathbb{R}^d$ is a measurable function, i.e.

$$T(\mathbf{y}_j, \boldsymbol{\lambda}) = \left(T_1(\mathbf{y}_j, \boldsymbol{\lambda}), \dots, T_d(\mathbf{y}_j, \boldsymbol{\lambda})\right).$$

If $A \in \mathbb{R}^{d \times d}$, we define with $[A]_{mn}$ its (m, n)-entry and with $\rho(A)$ the spectral radius of A. Moreover we define the following $d \times d$ matrix as

$$[\Psi(\boldsymbol{\lambda})]_{mn} = E\left[\partial_{\lambda_m} T_n(\mathbf{y}_j, \boldsymbol{\lambda})\right], \tag{3.30}$$

with $m, n = 1, \dots, d$.

We make the following assumptions:

($A1$) $Y_j \overset{\text{iid}}{\sim} P$, where P is a probability distribution over $\mathbb{Y}$.

($A2$) For every $J \geq 1$ there exists a fixed point $\boldsymbol{\lambda}_J^*$ of equation (3.29). Moreover, there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^d$ such that $\boldsymbol{\lambda}_J^* \to \boldsymbol{\lambda}^*$ $P^{(\infty)}$-almost surely as $J \to \infty$.

($A3$) Let $B_\epsilon^* = \{\boldsymbol{\lambda} \mid \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\| < \epsilon\}$, for $\epsilon > 0$. Then there exist $\rho < 1$, $M < \infty$ and $\delta > 0$ such that

$$\rho\left(\Psi(\boldsymbol{\lambda}^*)\right) < \rho \quad \text{and} \quad E\left[\sup_{\boldsymbol{\lambda} \in B_\delta^*} |\partial^2_{\lambda_{mn}} T_k(Y_j, \lambda)|\right] < M,$$

for every $m, n, k = 1, \dots, d$.

We need a preliminary result.

Lemma 1. *Let $A \in \mathbb{R}^{d\times d}$ and assume $\rho(A) < \rho$. Then there exists $\epsilon = \epsilon(A) > 0$ such that for every $B \in \mathbb{R}^{d\times d}$ with $\max_{m,n} |[B]_{mn}| < \epsilon$ it holds that $\rho(A + B) < \rho$.*

Then we can prove the following lemma.

Lemma 2. *Consider equation* (3.29) *and assume* $(A1) - (A3)$ *are satisfied. Thus, there exists $c > 0$ such that, if $\|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*\| < c$, for almost every sequence $\boldsymbol{y} = \boldsymbol{y}_{1:\infty}$ there exists $J^* = J^*(Y)$ such that*

$$\|\boldsymbol{\lambda}(t+1) - \boldsymbol{\lambda}_J^*\| \leq \rho^{t+1} \|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*\|,$$

for every $J \geq J^*$ and $t \geq 1$.

3.4 Discussion

As the complexity and availability of data continue to expand, providing a clear understanding of the computational properties of algorithms for posterior inference has become increasingly critical. This need is especially pronounced in contexts such as variational inference, where direct, reliable uncertainty quantification is often unavailable, and robust convergence diagnostics are lacking.

In this chapter, we made an initial contribution to the understanding of the convergence properties of the Coordinate Ascent Variational Inference (CAVI) algorithm when applied to large scale hierarchical models. By leveraging the probabilistic structure of these models and asymptotic results, we were able to rigorously characterize the speed of convergence of the CAVI iterates to their solution in terms of the model characteristics. This characterization was done both for the variational parameters and the Kullback-Leibler (KL) divergence.

Our findings reveal strong agreements with existing literature on coordinate-wise schemes, confirming the efficacy of CAVI in various hierarchical settings. The results underscore the importance of model characteristics in determining the convergence behaviour, providing insights that could be valuable for practitioners in selecting and tuning variational inference algorithms.

Looking ahead, we are currently developing extensions of our theoretical results to encompass general global/local models, such as those described in Hoffman et al. (2013). This will involve a more comprehensive analysis of the statistical properties induced by CAVI.

By extending our theoretical framework, we aim to offer a deeper understanding of the convergence properties and other statistical attributes of CAVI across a wider range of models.

3.5 Annex

3.5.1 Proofs of Section 3.3.4

We state the following instrumental lemma:

Lemma 3. *Consider a triangular sequence* $(X_{n1}, ..., X_{nn})_{n=1}^{+\infty}$, *s.t.* $X_{nj} \perp X_{ni}$ *for* $i \neq j$ *and let* $\mu_n = \mathbb{E}[X_{nj}] =$ *for all* j. *If* $\mathbb{E}[|X_{nj} - \mu_n|^4] < b \forall j, n$ *then:*

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{j=1}^{n} X_{nj} - \mu_n = 0 \quad a.s.$$

Proof. Without loss of generality consider the centered sequence s.t. $\mathbb{E}[X_{nj}] = 0$. It follows:

$$\Pr\left(|\frac{1}{n} \sum_{j=1}^{n} X_{nj}| > \varepsilon \right) = \Pr(|\sum_{j=1}^{n} X_{nj}| > n\varepsilon) \leq \frac{\mathbb{E}[(\sum_{j=1}^{n} X_{nj})^4]}{n^4 \varepsilon^4}.$$

By independence and since $b > \mathbb{E}[X_{nj}^4] \geq \mathbb{E}[X_{nj}^2]^2$ by Jensen:

$$\mathbb{E}[(\sum_{j=1}^{n} X_{nj})^4] = \sum_{j=1}^{n} \mathbb{E}[X_{nj}^4] + 6 \sum_{i \neq j} \mathbb{E}[X_{nj}^2] \mathbb{E}[X_{ni}^2] \leq nb + 6n^2 b.$$

Substituting in the above

$$\sum_{n=1}^{+\infty} \Pr(|\frac{1}{n} \sum_{j=1}^{n} X_{nj}| > \varepsilon) \leq \sum_{n=1}^{+\infty} \frac{nb + 6n^2 b}{n^4 \varepsilon^4} < +\infty.$$

□

Proof of Lemma 16. Fix $t \geq 1$. By a Taylor expansion and using the definition of λ_J^* in $(A2)$ we have

$$\begin{aligned}
\lambda_J(t+1) - \lambda_J^* &= \frac{a_J}{J} + \frac{1}{J} \sum_{j=1}^{J} T_J\left(y_j, \lambda_J(t)\right) - \lambda_J^* \\
&= \frac{a_J}{J} + \frac{1}{J} \sum_{j=1}^{J} T_J\left(y_j, \lambda_J^*\right) - \lambda_J^* + (\lambda_J(t) - \lambda_J^*) \frac{1}{J} \sum_{j=1}^{J} \int_0^1 \partial_\lambda T_J\left(Y_j, \lambda_J^* + x(\lambda_J(t) - \lambda_J^*)\right) \mathrm{d}x \\
&= (\lambda_J(t) - \lambda_J^*) \frac{1}{J} \sum_{j=1}^{J} \int_0^1 \partial_\lambda T_J\left(y_j, \lambda_J^* + x(\lambda_J(t) - \lambda_J^*)\right) \mathrm{d}x,
\end{aligned}$$

for every $J \geq 1$. This implies that

$$|\lambda_J(t+1) - \lambda_J^*| \leq |\lambda_J(t) - \lambda_J^*| \frac{1}{J} \sum_{j=1}^{J} \int_0^1 \left| \partial_\lambda T_J \left(y_j, \lambda_J^* + x(\lambda_J(t) - \lambda_J^*) \right) \right| \mathrm{d}x, \tag{3.31}$$

for every $J, t \geq 1$. By assumption $(A2)$, for almost every sequence $\mathbf{y} = (y_1, y_2, ...)$ there exists $J_1^* = J_1^*(\mathbf{y})$ such that $|\lambda_J^* - \lambda^*| < \delta/2$ for every $J \geq J_1^*$. Choose now c in the assumptions s.t. $c = \delta/2$, so that we can conclude

$$|\lambda_J^* + x(\lambda_J(0) - \lambda_J^*) - \lambda^*| < \delta, \tag{3.32}$$

for every $x \in (0, 1)$ and $J \geq J_1^*$. Choose now ε s.t. $\rho + \varepsilon < 1$. By Lemma 3, which can be applied thanks to (A1), (A3), and (A4), we have

$$\lim_{J \to +\infty} \frac{1}{J} \sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(y_j, \lambda)| - \mu_J \leq \lim_{J \to +\infty} |\frac{1}{J} \sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(y_j, \lambda)| - \mu_J| = 0. \tag{3.33}$$

From the above it must exist $J_2^* = J_2^*(\mathbf{y})$ such that $\forall J > J_2^*$

$$|\frac{1}{J} \sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(y_j, \lambda)| - \mu_J| < \frac{\varepsilon}{4},$$

and so

$$|\frac{1}{J} \sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(y_j, \lambda)|| < \rho_J + \frac{\varepsilon}{2},$$

Then by (A3)-bis, $\exists J_3^*$ s.t. $|\rho_J - \rho| < \frac{\varepsilon}{2}$, so:

$$|\frac{1}{J} \sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(y_j, \lambda)| \leq \rho + \varepsilon, \tag{3.34}$$

for all $J > \max\{J_2^*, J_3^*\}$. If furthermore $J > J^* = \max\{J_1^*, J_2^*, J_3^*\}$ by (3.32)

$$\frac{1}{J} \sum_{j=1}^{J} \int_0^1 \left| \partial_\lambda T_J \left(y_j, \lambda_J^* + x(\lambda(0) - \lambda_J^*) \right) \right| \mathrm{d}x \leq \frac{1}{J} \sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_\lambda T_J(y_j, \lambda)|. \tag{3.35}$$

Combining (3.33) with (3.34), we obtain

$$\frac{1}{J}\sum_{j=1}^{J}\int_0^1 \left|\partial_\lambda T_J\left(y_j, \lambda_J^* + x(\lambda(0) - \lambda_J^*)\right)\right| \mathrm{d}x \leq \rho + \varepsilon,$$

which implies

$$|\lambda_J(1) - \lambda_J^*| \leq \tilde{\rho}|\lambda(0) - \lambda_J^*|.$$

for $\tilde{\rho} = \rho + \varepsilon < 1$. The rest by induction. □

Proof of Theorem 1. We begin by showing that for models as in (3.19) under (B1)-(B4), Lemma 16 applies. Setting again

$$\lambda_J(t) := \int \mu \; dq^t(\mu), \qquad \phi_j(t) := \int \theta_j \; dq_j^t(\theta_j) \;\text{ for } j = 1, ..., J,$$

we get

$$q_j^{t+1}(\theta_j) \propto f(y_j|\theta_j) \exp\{\lambda_J(t)\frac{\theta_j}{\tau_0^2} - \frac{\theta_j}{2\tau_0^2}\},$$

$$q^{t+1}(\mu) \propto h(\mu) \exp\left\{\mu\left(\frac{\mu_0\tau_0^2}{\sigma_0^2} + \sum \phi_j(t+1)\right) - \frac{J}{2}\mu^2\tau_0^2\right\}.$$

The recursion becomes

$$\lambda_J(t+1) = \frac{\mu_0\tau_0^2}{J\sigma_0^2 + \tau_0^2} + \frac{J\sigma_0^2}{J\sigma_0^2 + \tau_0^2}\frac{\sum_j \phi_j(t+1)}{J},$$

and satisfy the same form as in (3.20), with $T_J(y_j, \lambda) = \frac{J\sigma_0^2}{J\sigma_0^2+\tau_0^2}\phi_j(\lambda)$. We need to show that assumption (A3) is satisfied. Given assumptions (B1)-(B3), it follows by the law of iterated expectations

$$\mathbb{E}_{G_\mu}[\mathbb{E}[\theta_j|y_j, \mu]]|_{\mu=\mu^*} = \mu^*, \tag{3.36}$$

and that

$$\begin{aligned}\mathbb{E}_{G_\mu}[|\partial_\lambda \phi_j(\lambda)|] &= \mathbb{E}_{G_\mu}\left[\frac{var(\theta_j|y_j, \mu)}{var(\theta_j|\mu)}\right] \\ &= \frac{var(\theta_j|\mu) - var(\mathbb{E}_{G_\mu}[\theta_j|y_j, \mu])}{var(\theta_j|\mu)}|_{\mu=\mu^*} < 1,\end{aligned} \tag{3.37}$$

therefore there exists a neighborhood $B^*_{\mu^*}$ of μ^* such that for $\lambda \in B^*_{\mu^*}$, one has

$$\mathbb{E}_{G_{\mu^*}}\left[|\partial_\lambda \phi_j|\right]|_\mu < \rho < 1 \qquad\qquad j = 1, ..., J, \tag{3.38}$$

and hence:

$$\begin{aligned}\mathbb{E}_{G_{\mu^*}}[\sup_{B^*_\delta}|\partial_\lambda T_J(y,\lambda)|] &= \mathbb{E}_{G_{\mu^*}}\left[\sup_{B^*_\delta}\frac{J\sigma_0^2}{J\sigma_0^2+\tau_0^2}|\partial_\lambda \phi_j(\lambda)|\right] \\ &< \frac{J\sigma_0^2}{J\sigma_0^2+\tau_0^2}\rho =: \rho_J,\end{aligned}$$

and $\rho_J \to \rho < 1$. By independence assumption of the mean field family we have:

$$KL_J(t) = KL(q^t_\lambda(\mu)\|q_{\lambda^*_J}(\mu)) + \sum_j KL(q^t_j(\theta_j)\|q^*_j(\theta_j)). \tag{3.39}$$

For the first summand, being the full conditionals Gaussian and exploiting the result in Lemma 16, there exist a $J^* = J(\mathbf{y}_{1:J})$ s.t.

$$KL(q^t_\lambda(\mu)\|q_{\lambda^*_J}(\mu)) = \frac{\sigma_0^2 J + \tau_0^2}{2\sigma_0^2}(\lambda_J(t) - \lambda^*_J)^2 \tag{3.40}$$

$$\leq \frac{\sigma_0^2 J + \tau_0^2}{2\sigma_0^2}(\lambda_J(0) - \lambda^*_J)^2\rho^{2t}, \tag{3.41}$$

for ρ as in (3.38). As regards the second summand in (3.39), note firstly that every term can be seen as a function of $\lambda_J(t)$, indeed, by the recursive updates of Algorithm 12, $\phi_j(t+1) = h(\lambda_J(t))$ for some (possibly unknown) function h. Hence define:

$$KL_j(\lambda(t)\|\lambda^*_J) = KL\left(q^{t+1}_J(\theta_j)\|q^*_J(\theta_j)\right).$$

Under $(B3)$ we can apply Taylor formula to get

$$\begin{aligned}&KL(q^t_j(\theta_j)\|q^*_j(\theta_j)) = KL_j\left(\lambda_J(t-1)\|\lambda^*_J\right) \\ &\quad = KL_j\left(\lambda^*_J\|\lambda^*_J\right) + (\lambda_J(t-1) - \lambda^*_J)\partial_\mu KL_j\left(\mu\|\lambda^*_J\right)|_{\lambda^*_J} + (\lambda_J(t-1) - \lambda^*_J)^2\partial^2_\mu KL_j\left(\tilde\lambda\|\lambda^*_J\right),\end{aligned}$$

with $|\tilde\lambda - \lambda^*_J| \leq |\lambda_J(t-1) - \lambda^*_J|$. It is clear that $KL_j\left(\lambda^*_J\|\lambda^*_J\right) = \partial_\lambda KL_j\left(\lambda\|\lambda^*_J\right)|_{\lambda=\lambda^*_J} = 0$,

so that

$$KL\left(q_j^t \| q_j^*(\theta_j)\right) = (\lambda_J(t-1) - \lambda_J^*)^2 \partial_\mu^2 KL_j\left(\tilde{\lambda} \| \lambda_J^*\right),$$

with $|\tilde{\lambda} - \lambda_J^*| \leq |\lambda_J(t-1) - \lambda_J^*|$. Then, under the event $\left\{J\left(\lambda_J(t-1) - \lambda_J^*\right)^2 \leq \rho^{2(t-1)}c^2\right\}$, by $(B3)$ we have

$$KL_j\left(q_j^t(\theta_j) \| q_j^*(\theta_j)\right) \leq c^2 \rho^{2(t-1)} \frac{C_j}{J},$$

so that with J high enough

$$\sum_{j=1}^{J} KL\left(q_j^t(\theta_j) \| q_j^*(\theta_j)\right) \leq \rho^{2(t-1)} \bar{C} c^2.$$

Thus

$$\Pr\left(\sum_{j=1}^{J} KL\left(q_j^t(\theta_j) \| q_j^*(\theta_j)\right) \leq \epsilon/2\right) \geq \Pr\left(\bar{C} J\left(\lambda_J(t-1) - \lambda_J^*\right)^2 \leq \epsilon/2\right). \tag{3.42}$$

By combining (3.40) and (3.42), we conclude

$$\begin{aligned}
&\Pr\left(KL\left(q_J^t \mid q_J^*\right) < \epsilon\right) \\
&\quad \geq \Pr\left(\left\{\sigma_0^2 \rho J (\lambda_J(t-1) - \lambda_J^*)^2 \leq \epsilon/2\right\} \cap \left\{\bar{C} J\left(\lambda_J(t-1) - \lambda_J^*\right)^2 \leq \epsilon/2\right\}\right).
\end{aligned}$$

It is possible to choose t such that the right-hand side goes to 1, since by Theorem 16 we have

$$\Pr\left(J(\lambda_J(t) - \lambda_J^*)^2 \leq \rho^t c^2\right) \geq \Pr\left(|\lambda_J(t) - \lambda_J^*| \leq \rho^t |\lambda_J(0) - \lambda_J^*|\right) \to 1,$$

as $J \to \infty$, since $|\lambda_J(0) - \lambda_J^*| \leq \frac{c}{\sqrt{J}}$ by assumption.

□

Proof of Corollary 3. Instead of using the law of iterated expectation and variance, one can state similar results exploiting information in a neighbourhood of the KL minimizer and the result follows. Instead of (3.36) and (3.37), from Lemma 17 proved below we have:

$$\mathbb{E}_P[\mathbb{E}[\theta_j | y_j, \mu]]|_{\mu=\mu^*} = \mu^*,$$

and

$$\mathbb{E}_P[|\partial_\lambda \phi_j|]|_{\mu=\mu^*} < \rho < 1.$$

Since the variational distributions $q_j(\theta_j)$ has the same distribution of $\mathcal{L}(\theta_j|y,\mu)$ (by CAVI properties applied to exponential family distributions), then the proof follows. □

Lemma 17. *For the model 3.19, under assumptions (B1*), (B2) and (B3), for every $j = 1, ..., J$ it holds*

$$\mathbb{E}_P[\mathbb{E}[\theta_j|y_j,\mu]]|_{\mu=\mu^*} = \mu^* \qquad\qquad \mathbb{E}_P[|\partial_\mu \mathbb{E}[\theta_j|y_j,\mu]|]|_{\mu=\mu^*} < 1.$$

Moreover, there exists a neighborhood $B^*_{\mu^*}$ of μ^* and a ρ such that for $\mu \in B^*_{\mu^*}$, one has

$$\mathbb{E}_P[|\partial_\mu \mathbb{E}[\theta_j|y_j,\mu]|]|_{\mu=\mu^*} < \rho < 1.$$

Proof. The proof strongly relies on the formula of the KL divergence and the densities of (3.19). Given the definition of μ^* in assumption $(B1^*)$ we have $\mu^* = argmin KL(P\|G_\mu)$, hence

$$\begin{aligned} \partial_\mu KL(P\|G_\mu)|_{\mu=\mu^*} &= 0 \\ \partial^2_{\mu^2} KL(P\|G_\mu)|_{\mu=\mu^*} &> 0. \end{aligned} \tag{3.43}$$

Leveraging the above equations

$$\begin{aligned} \partial_\mu KL(P\|G_\mu) = \partial_\mu \int \ln \frac{p(y)}{g(y|\mu)} p(y)dy &= -\int \partial_\mu \ln(g(y|\mu))p(y)dy \\ &= -\int \frac{f(y|\theta)p(\theta|\mu)}{\int f(y|\theta)p(\theta|\mu)d\theta} p(y)dy = -\left(\mathbb{E}_P\left[\frac{\mathbb{E}[\theta|y,\mu]}{\tau_0^2}\right] - A'(\mu)\right), \end{aligned}$$

where $A(\mu)$ denotes the log partition function of the prior (in this case, being Gaussian, $A(\mu) = \frac{\mu^2}{2\tau_0^2}$), and we exploited the definition of $g(y|\mu)$. Substituting with (3.43), one gets

$$\mathbb{E}_P[\mathbb{E}[\theta|y,\mu^*]] = A'(\mu^*)\tau_0^2.$$

Furthermore note that

$$\partial_\mu \mathbb{E}_P[\theta_j|y_j,\mu] = \frac{var(\theta_j|y_j,\mu)}{var(\theta_j|\mu)} > 0,$$

by the formulation of the model. As for the second derivative we get

$$\partial^2_{\mu^2} KL(P\|G_\mu) = -\left(\mathbb{E}_P[\partial_\mu \mathbb{E}_P[\theta_j|y_j,\mu]] - A''(\mu)\right) = -\left(\mathbb{E}_P[|\partial_\mu \mathbb{E}[\theta_j|y_j,\mu]|] - A''(\mu)\right).$$

Again by (3.43) we have

$$\begin{aligned}\partial^2_{\mu^2} KL(P\|G_\mu)|_{\mu=\mu^*} > 0 &\leftrightarrow \mathbb{E}_P[|\partial_\mu \mathbb{E}[\theta_j|y_j,\mu]|]|_{\mu=\mu^*} < A''(\mu^*)\tau_0^2 \\ &\leftrightarrow \mathbb{E}_P[|\partial_\mu \mathbb{E}[\theta_j|y_j,\mu]|] < 1,\end{aligned}$$

where we used that $A''(\mu) = \frac{1}{\tau_0^2}$. The last statement follows by continuity of $\mathbb{E}_P[\partial_\mu \mathbb{E}[\theta_j|y_j,\mu]]$ implied by (B3). □

Proof of Lemma 3.3.2. Let $z_J(t) = \lambda_J(t) - \lambda_J^*$, with λ_J^* as $(C1)$. By formula (3.24) and definition of λ_J^* we can write

$$z_J(t) = g_J\left(z_J(t-1) + \lambda_J^*, \mathbf{y}_{1:J}\right) - \lambda_J^*.$$

By $(C1)$ we can apply Taylor formula to get

$$\begin{aligned} z_J(t) &= g_J\left(\lambda_J^*, \mathbf{y}_{1:n}\right) - \lambda_J^* + z_J(t-1)\int_0^1 (1-x)\partial_\lambda\, g_J\left(\lambda_J^* + x z_J(t-1), \mathbf{y}_{1:n}\right)\,\mathrm{d}x \\ &= z_J(t-1)\int_0^1 (1-x)\partial_\lambda\, g_J\left(\lambda_J^* + x z_J(t-1), \mathbf{y}_{1:J}\right)\,\mathrm{d}x, \end{aligned} \tag{3.44}$$

by definition of λ_J^*. By assumption $(A1)$ and $(A2)$, for almost every sequence $\mathbf{y} = (y_1, y_2, ...)$ there exists $J_1^* = J_1^*(Y)$ such that $|\lambda_J^* - \lambda^*| < \delta/2$ for every $J \geq J_1^*$. Choose now c in the assumptions s.t. $c = \delta/2$, so that we can conclude

$$|\lambda_J^* + x z_J(t-1) - \lambda^*| < \delta, \tag{3.45}$$

for every $x \in (0,1)$ and $J \geq J_1^*$. Choose ε s.t. $\rho + \varepsilon < 1$. Therefore, taking (3.44) with

$t = 1$, by $(C2)$ we have that $\exists J_2^* = J_2^*(\mathbf{y})$ s.t. $\forall J > J_2^*$

$$\int_0^1 (1-x)|\partial_\lambda\, g_J\left(\lambda_J^* + xz_J(t-1), \mathbf{y}_{1:J}\right)|\,\mathrm{d}x$$
$$\leq \int_0^1 (1-x)|\partial_\lambda\, h\left(\lambda_J^* + xz_J(t-1), \mathbf{y}_{1:J}\right) + \varepsilon|\,\mathrm{d}x \leq (\rho + \varepsilon),$$

$P^{(\infty)}$-almost surely. Notice that limit and integral can be exchanged thanks to $(B3.3)$ combined with Dominated Convergence Theorem. Therefore for $J \geq \max(J_1^*, J_2^*)$:

$$|\lambda_J(1) - \lambda_J^*| < |\lambda_J(0) - \lambda_J^*|\tilde{\rho},$$

$P^{(\infty)}$-almost surely, for $\tilde{\rho} = \rho + \varepsilon$. Iterating the argument can be reiterated t times to get the final result. □

3.5.2 Proofs and computations of Section 3.3.5

From the formulation of the model in (3.25), we get $\tau_0^2|\alpha_0, \beta_0 \sim \Gamma^{-1}(\alpha_0/2, \beta_0/2)$, $\mu|\tau_0^2 \sim N(\mu, \tau_0^2\sigma_0^2)$, and $\mathbb{E}[\mu] = \mu_0, \mathbb{E}[\tau_0^2] = \frac{\beta_0}{\alpha_0 - 2}$. Standard computations on the model show that

$$(\mu, \tau_0^2)|\cdot \sim NIG\left(\frac{\mu_0 + \sigma_0^2 \sum_j \theta_j}{1 + \sigma_0^2 J}, \frac{1 + \sigma_0^2 J}{\sigma_0^2}, \frac{\alpha_0 + J}{2}, \frac{1}{2}\left(\beta_0 + \sum_j (\theta_j - \bar{\theta})^2 + \frac{J}{1 + \sigma_0^2 J}(\bar{\theta} - \mu_0)^2\right)\right),$$

$$\theta_j|\cdot \sim N\left(\frac{\tau_0^2 \sum_m y_{jm} + \mu\sigma_r^2}{\sigma_r^2 + M\tau_0^2}, \frac{\tau_0^2\sigma_r^2}{\sigma_r^2 + M\tau_0^2}\right),$$

$$\sigma_r^2|\cdot \sim IG\left(\frac{\alpha + MJ}{2}, \frac{\beta + \sum_j \sum_m (y_{jm} - \theta_j)^2}{2}\right).$$

And also

$$g(\mathbf{y}_j|\mu, \tau_0^2, \sigma_r^2) = N(\mu, \sigma_r^2 I_m + \tau_0^2 \mathcal{H}) \qquad j = 1, ..., J, \tag{3.46}$$

where I_m and $\mathcal{H}$ denote respectively the m dimensional identity matrix and matrix of ones. Using the same notation of Section 3.3.5, we get

$$
\begin{aligned}
q_J^t((\mu,\tau_0^2)) &= NIG\left(\frac{\mu_0+\sigma_0^2\sum_j \frac{\nu_J\tilde{a}\sum_m y_{jm}+\lambda_J\xi_J a(J)}{\xi_J a+M\nu_J\tilde{a}}}{1+\sigma_0^2 J}, \frac{1+\sigma_0^2 J}{\sigma_0^2}, \frac{\alpha_0+J}{2}, \frac{\beta_0+\tilde{\beta}_J}{2}\right),\\
q_J^t(\sigma_r^2) &= IG\left(\frac{\alpha+MJ}{2}, \frac{1}{2}\left(\beta+\sum_j\sum_m\left(\left(\frac{\nu_J\sum_{\tilde{m}} y_{j\tilde{m}}+\lambda_J\xi_J}{M\nu_J+\xi_J}-y_{jm}\right)^2+\frac{\xi_J\nu_J}{M\nu_J+\xi_J}\right)\right)\right),\\
q_j^t(\theta_j) &= N\left(\frac{\nu_J\sum_m y_{jm}\tilde{a}+\lambda_J\xi_J a}{a\xi_J+M\tilde{a}\nu_J}, \frac{\nu_J\xi_J}{a\xi_J+M\tilde{a}\nu_J}\right) \qquad j=1,...,J,
\end{aligned}
$$

where $a(J) := \frac{\alpha_0+J}{\alpha_0+J-2}$, $\tilde{a}(J) := \frac{\alpha+J}{\alpha+J-2}$ and

$$
\begin{aligned}
\tilde{\beta}_J = \frac{J\xi_J\nu_J}{(a\xi_J+\tilde{a}M\nu_J)}\left(1-\frac{1}{J}\right)+\frac{1}{(a\xi_J+\tilde{a}M\nu_J)^2}\left[\nu_J^2\tilde{a}^2\left(\sum_j(\sum_m y_{jm})^2-\frac{(\sum_j\sum_m y_{jm})^2}{J}\right)\right]+\\
+\frac{J}{1+\sigma_0^2 J}\left(\left(\frac{\nu_J\bar{y}\tilde{a}+\lambda_J a\xi_J}{a\xi_J+M\tilde{a}\nu_J}-\mu_0\right)^2+\frac{\nu_J\xi_J}{J(a\xi_J+M\nu_J\tilde{a})}\right). \qquad (3.47)
\end{aligned}
$$

The corresponding recursions is

$$
\begin{aligned}
\lambda_J(t+1) &= \frac{\mu_0+\sigma_0^2\sum_j \frac{\nu_J\tilde{a}\sum_m y_{jm}+\lambda\xi_J a(J)}{\xi_J a+M\nu_J\tilde{a}}}{1+\sigma_0^2 J},\\
\nu_J(t+1) &= \frac{1}{\alpha_0+J-2}\cdot(\beta_0+\tilde{\beta}_J),\\
\xi_J(t+1) &= \frac{1}{\alpha+MJ-2}\\
&\quad\cdot\left(\beta+\sum_j\sum_m\left(\left(\frac{\nu_J\sum_{\tilde{m}} y_{j\tilde{m}}+\lambda_J\xi_J}{M\nu_J+\xi_J}-y_{jm}\right)^2+\frac{\xi_J\nu_J}{M\nu_J+\xi_J}\right)\right),
\end{aligned} \qquad (3.48)
$$

with $\tilde{\beta}_J$ as in (3.47).

As for the KL divergence, exploiting the decomposition in (3.27), we now analyse the behaviour of the summands.

For the first summand, recall that if $p = NIG(\mu_0^p,\lambda,\alpha,\beta_p), q = NIG(\mu_0^q,\lambda,\alpha,\beta_q)$,

then

$$\begin{aligned} KL(p(\mu,\tau_0^2)\|q(\mu,\tau_0^2)) &= E_p[KL(p(\mu|\tau_0^2)\|q(\mu|\tau_0^2))] + KL(p(\tau_0^2)\|q(\tau_0^2)) \\ &= E_{p(\tau_0^2)}[KL(N(\mu_0^p,\tau_0^2/\lambda)\|N(\mu_0^q,\tau_0^2/\lambda))] + KL(IG(\alpha,\beta_p)\|IG(\alpha,\beta_q)) \\ &= \frac{1}{2}\frac{\alpha}{\beta_p}\lambda^2(\mu_0^p-\mu_0^q)^2 + \alpha\ln\left(\frac{\beta_p}{\beta_q}\right) + \alpha\frac{\beta_q-\beta_p}{\beta_p}. \end{aligned}$$

hence

$$\begin{aligned} KL(q^t(\mu,\tau_0^2)\|q_J^*(\mu,\tau_0^2)) = \frac{1}{2}\frac{\alpha_0+J}{\beta_0+\tilde{\beta}_J}\left(\sum_j \mathbb{E}_{v(t)}[\theta_j] - \mathbb{E}_{v^*}[\theta_j]\right)^2 \\ + \frac{\alpha_0+J}{2}\ln\left(\frac{\beta_0+\tilde{\beta}_J}{\beta_0+\beta_J^*}\right) + \frac{\alpha_0+J}{2}\frac{\beta_J^*-\tilde{\beta}_J}{\beta_0+\tilde{\beta}_J}, \end{aligned}$$

where β_J^* as $\tilde{\beta}_J$ but with the optimal parameters. For the second summand, since $KL(N(\mu_1,\sigma_1^2)|N(\mu_2,\sigma_2^2)) = \frac{1}{2}\ln\left(\frac{\sigma_2^2}{\sigma_1^2}\right) - \frac{1}{2} + \frac{\sigma_1^2+(\mu_1-\mu_2)^2}{2\sigma_2^2}$, one has

$$KL(q_j^t(\theta_j)|q_j^*(\theta_j)) = \frac{1}{2}\ln\left(\frac{\frac{\nu_J\xi_J}{a\xi_J+M\tilde{a}\nu_J}}{\frac{\nu_J^*\xi_J^*}{a\xi_J^*+M\tilde{a}\nu_J^*}}\right) - \frac{1}{2} + \frac{\frac{\nu_J\xi_J}{a\xi_*+M\tilde{a}\nu_J} + \left(\frac{\nu_J\sum_m y_{jm}\tilde{a}+\lambda_J\xi_J^*a}{a\xi+M\tilde{a}\nu_J} - \frac{\nu_J^*\sum_m y_{jm}\tilde{a}+\lambda_J^*\xi_J^*a}{a\xi_J^*+M\tilde{a}\nu_J^*}\right)^2}{2\frac{\nu_J^*\xi_J^*}{a\xi_J^*+M\tilde{a}\nu_J^*}}.$$

For the third, recalling $KL(\Gamma^{-1}(\alpha,\beta_1)\|\Gamma^{-1}(\alpha,\beta_2)) = \alpha\log\frac{\beta_1}{\beta_2} + \alpha\frac{\beta_2-\beta_1}{\beta_1}$, hence

$$KL(q^t(\sigma_r^2)\|q_J^*(\sigma_r^2)) = \frac{\alpha+MJ}{2}\log\frac{\beta+\hat{\beta}_J}{\beta+\hat{\beta}_J^*} + \frac{\alpha+MJ}{2}\frac{\hat{\beta}_J^*-\hat{\beta}_J}{\hat{\beta}_J^*+\beta}.$$

with $\hat{\beta}_J = \frac{1}{2}\left(\beta + \sum_j\sum_m\left(\left(\frac{\nu_J\sum_{\tilde{m}} y_{j\tilde{m}}+\lambda_J\xi_J}{M\nu_J+\xi_J} - y_{jm}\right)^2 + \frac{\xi_J\nu_J}{M\nu_J+\xi_J}\right)\right)$, and $\hat{\beta}_J^*$ as $\hat{\beta}_J$ but with the limiting variances.

Proof of Lemma 2. Fix $t \geq 1$. By a Taylor expansion and using the definition of $\boldsymbol{\lambda}_J^*$ in

$(A2)$ we have

$$\begin{aligned}\boldsymbol{\lambda}(t+1) - \boldsymbol{\lambda}_J^* &= \frac{a_J}{J} + \frac{1}{J}\sum_{j=1}^{J} T\left(\mathbf{y}_j, \boldsymbol{\lambda}(t)\right) - \boldsymbol{\lambda}_J^* \\ &= \frac{a_J}{J} + \frac{1}{J}\sum_{j=1}^{J} T\left(\mathbf{y}_j, \boldsymbol{\lambda}_J^*\right) - \boldsymbol{\lambda}_J^* + A_J(t)\left(\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_J^*\right) \\ &= A_J(t)\left(\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_J^*\right),\end{aligned}$$

for every $J \geq 1$, where $A_J(t) \in \mathbb{R}^{d\times d}$ with

$$[A_J(t)]_{mn} = \frac{1}{J}\sum_{j=1}^{J}\int_0^1 \partial_{\lambda_n} T_m\left(\mathbf{y}_j, \boldsymbol{\lambda}_J^* + x(\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_J^*)\right) \mathrm{d}x. \tag{3.49}$$

Thus it suffices to prove that for almost every sequence $\mathbf{y} = \mathbf{y}_{1:\infty}$ there exists $J^* = J^*(Y)$ such that $\rho(A_J(t)) < \rho$ for every $J \geq J^*$ and $t \geq 1$.

Apply Lemma 1 with $A = \Psi(\boldsymbol{\lambda}^*)$ and fix the corresponding $\epsilon > 0$. By assumption $(A2)$, for almost every sequence $\mathbf{y}$ there exists $J_1^* = J_1^*(Y)$ such that $\|\boldsymbol{\lambda}_J - \boldsymbol{\lambda}^*\| < \min\{\frac{\delta}{2}, \frac{\epsilon}{4M}\}$ for every $J \geq J_1^*$. Similarly, choose $c < \min\{\frac{\delta}{2}, \frac{\epsilon}{4M}\}$, so that for every $J \geq J_1^*$ and $x \in (0,1)$ we can conclude

$$\|\boldsymbol{\lambda}_J^* + x(\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*) - \boldsymbol{\lambda}^*\| < \delta, \tag{3.50}$$

and

$$M\left(\|\boldsymbol{\lambda}_J^* - \boldsymbol{\lambda}^*\| + \|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*\|\right) < \frac{\epsilon}{2}. \tag{3.51}$$

Moreover, by the Law of Large Numbers, which can be applied thanks to (A1), and (A3), there exists $J_2^* = J_2^*(Y)$ such that

$$\left|\frac{1}{J}\sum_{j=1}^{J} \partial_{\boldsymbol{\lambda}_n} T_m\left(\mathbf{y}_j, \boldsymbol{\lambda}_J^*\right) - [\Psi(\boldsymbol{\lambda}^*)]_{dd'}\right| < \frac{\epsilon}{2} \quad \text{and} \quad \frac{1}{J}\sum_{j=1}^{J} \sup_{\boldsymbol{\lambda}\in B_\delta^*} |\partial^2_{\boldsymbol{\lambda}_{nk}} T_d(\mathbf{y}_j, \boldsymbol{\lambda})| < M, \tag{3.52}$$

for every $J \geq J_2^*$ and for every $m, n, k = 1, \ldots, d$.

Define now $J^* = \max\{J_1^*, J_2^*\}$ and fix $J \geq J^*$. Let $n, k = 1, \ldots, d$. If $x \in (0,1)$ again

by a Taylor expansion we have

$$\partial_{\boldsymbol{\lambda}_n} T_m\left(\mathbf{y}_j, \boldsymbol{\lambda}_J^* + x(\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*)\right) = \partial_{\lambda_n} T_m\left(\mathbf{y}_j, \boldsymbol{\lambda}^*\right) + \sum_{k=1}^{m} \left[\lambda_{J,k}^* - \lambda_k^* + x(\lambda_k(0) - \lambda_{J,k}^*)\right] \partial_{\lambda_{nk}}^2 T_m(\mathbf{y}_j, \boldsymbol{\lambda}_J(x)), \tag{3.53}$$

where $\boldsymbol{\lambda}_J(x)$ is such that $\|\boldsymbol{\lambda}_J(x) - \boldsymbol{\lambda}^*\| < \delta$. Thus, combining (3.49) and (3.53) we have that

$$A_J(0) = \Psi(\boldsymbol{\lambda}^*) + B_J,$$

where

$$[B_J]_{mn} = \frac{1}{J}\sum_{j=1}^{J} \partial_{\lambda_n} T_m\left(\mathbf{y}_j, \boldsymbol{\lambda}^*\right) - [\Psi(\boldsymbol{\lambda}^*)]_{mn} + \frac{1}{J}\sum_{j=1}^{J}\int_0^1 \sum_{k=1}^{m} \left[\lambda_{J,k}^* - \lambda_k^* + x(\lambda_k(0) - \lambda_{J,k}^*)\right] \partial_{\lambda_{nk}}^2 T_m(\mathbf{y}_j, \boldsymbol{\lambda}_J(x))\,\mathrm{d}x.$$

By (3.50) and by the definition of δ and $\boldsymbol{\lambda}_J(x)$ we have that

$$\left|\frac{1}{J}\sum_{j=1}^{J}\int_0^1 \sum_{k=1}^{n} \left[\lambda_{J,k}^* - \lambda_k^* + x(\lambda_k(0) - \lambda_{J,k}^*)\right] \partial_{\lambda_{nk}}^2 T_m(\mathbf{y}_j, \boldsymbol{\lambda}_J(x))\,\mathrm{d}x\right| \leq \left(\frac{1}{J}\sum_{j=1}^{J} \sup_{\lambda \in B_\delta^*} |\partial_{\lambda_{nk}}^2 T_m(\mathbf{y}_j, \boldsymbol{\lambda})|\right) \left(\|\boldsymbol{\lambda}_J^* - \boldsymbol{\lambda}^*\| + \|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*\|\right),$$

for every m, n, k. Thus, combining (3.51) and (3.52) we obtain

$$\max_{m,n} [B_J]_{m,n} < \epsilon,$$

which by definition of ϵ implies $\rho(A_J(0)) < \rho$ and thus

$$\|\boldsymbol{\lambda}(1) - \boldsymbol{\lambda}_J^*\| \leq \rho \|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*\|,$$

for every $J \geq J^*$. Assume now that

$$\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_J^*\| \leq \rho^t \|\boldsymbol{\lambda}(0) - \boldsymbol{\lambda}_J^*\|, \tag{3.54}$$

for every $J \geq J^*$. Then it suffices to show that $\rho(A_J(t)) < \rho$ for every $J \geq J^*$, with $A_J(t)$ as in (3.49). Notice that (3.50) and (3.54) imply

$$\|\boldsymbol{\lambda}_J^* + x(\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}_J^*) - \boldsymbol{\lambda}^*\| < \delta,$$

for every $J \geq J^*$ and $x \in (0, 1)$. Thus the result follows as above, by definition of J^*. $\square$

References

Arnese, M. and Lacker, D. (2024). Convergence of coordinate ascent variational inference for log-concave measures via optimal transport.

Ascolani, F. and Zanella, G. (2024+). Dimension-free mixing times of Gibbs samplers for Bayesian hierarchical models. *The Annals of Statistics*.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in statistical theory*. Wiley Series in Probability and Mathematical Statistics.

Barndorff-Nielsen, O. (2014). Information and exponential families: In statistical theory. *Information and Exponential Families in Statistical Theory*.

Bhattacharya, A., Pati, D., and Yang, Y. (2023). On the Convergence of Coordinate Ascent Variational Inference.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2012). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Cai, D., Modi, C., Pillaud-Vivien, L., Margossian, C. C., Gower, R. M., Blei, D. M., and Saul, L. K. (2024). Batch and match: black-box variational inference with a score-based divergence.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6(none):1847 – 1899.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics*, 7(2):269 – 281.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2024). Bayesian data analysis. *The SAGE Encyclopedia of Research Design*.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression And Multilevel/Hierarchical Models*, volume 3.

Hall, P., Pham, T. H., Wand, M. P., and Wang, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Annals of Statistics*, 39:2502–2532.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(40):1303–1347.

Jiang, Y., Chewi, S., and Pooladian, A.-A. (2024). Algorithms for mean-field variational inference via polyhedral optimization in the wasserstein space.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kadanoff, L. P. (2009). More is the same; phase transitions and mean field theories. *Journal of Statistical Physics*, 137:777–797.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kushner, H. J. and Yin, G. (2003). Stochastic approximation and recursive algorithms and applications.

Lavenant, H. and Zanella, G. (2024). Convergence rate of random scan Coordinate Ascent Variational Inference under log-concavity.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Liu, J. (1994). Fraction of missing information and convergence rate of data augmentation.

Margossian, C. C., Pillaud-Vivien, L., and Saul, L. K. (2024). Variational inference for uncertainty quantification: an analysis of trade-offs.

Meng, X.-L. and Rubin, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra and its Applications*, 199:413–425. Special Issue Honoring Ingram Olkin.

Meyn, S., Tweedie, R. L., and Glynn, P. W. (2009). *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition.

Minka, T. P. (2005). Divergence measures and message passing.

Parisi, G. (1988). *Statistical field theory*. Frontiers in Physics. Addison-wesley.

Ray, K. and Szabó, B. (2022). Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281.

Roberts, G. and Rosenthal, J. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1.

Roberts, G. and Smith, A. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and their Applications*, 49(2):207–216.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):291–317.

Sahu, S. K. and Roberts, G. O. (1999). On convergence of the EM algorithmand the Gibbs sampler. *Statistics and Computing*, 9(1):55–64.

Tan, S. L. and Nott, D. J. (2014). Variational approximation for mixtures of linear mixed models. *Journal of Computational and Graphical Statistics*, 23(2):564–585.

Vehtari, A., Gelman, A., Sivula, T., Jylänki, P., Tran, D., Sahai, S., Blomstedt, P., Cunningham, J. P., Schiminovich, D., and Robert, C. P. (2020). Expectation propagation as a way of life: A framework for bayesian inference on partitioned data. *Journal of Machine Learning Research*, 21(17):1–53.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1:1–305.

Wang, Y. and Blei, D. M. (2019a). Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wang, Y. and Blei, D. M. (2019b). Variational Bayes under Model Misspecification. *ArXiv*, abs/1905.10859.

Yang, Y., Pati, D., and Bhattacharya, A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886 – 905.

Zhang, F. and Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207.

Chapter 4

Exact simulation in non-conjugate models via Catalytic couplings and Adaptive Rejection samplers

In this last chapter we present a recent work started during the visited period at Warwick university in collaboration with professor Gareth Roberts. In the following we seek to provide theoretical insights on Catalytic couplings (Breyer and Roberts, 2001), a coupling procedure originally developed for CFTP of Section 1.2.1, and novel applications for Adaptive rejection samplers (Gilks and Wild, 1992), to provide an effective methodology for sampling of coupled chains for non conjugate models, such as those analyzed in Section 2.7.

4.1 Catalytic couplings

Catalytic couplings are a coupling strategy first introduced in the work of Breyer and Roberts (2001), specifically designed for Markov chains. Although not maximal, the algorithm is implementable also whenever the marginal distributions are known up to their normalizing constants and allows in principle to couple more than two chains at once. In detail, let $(\mathbf{X}^t)_{t\geq 1}$ be a Markov chain evolving according to some π invariant kernel P, admitting density $p(\mathbf{x}, \cdot)\, \forall \mathbf{x} \in \mathbb{R}^d$, and also suppose that the kernel update can

be written as a random function f, i.e.

$$\int \pi(\mathbf{x}) \Pr(f(\mathbf{x}) \in d\mathbf{y}) d\mathbf{x} = \pi(\mathbf{y}) d\mathbf{y},$$

$$\Pr(f(\mathbf{x}) \in d\mathbf{y}) = p(\mathbf{x}, \mathbf{y}) d\mathbf{y}.$$

Furthermore choose an easy-to-sample proposal distribution Q with density q. Then a sample from the catalytic coupling algorithm is:

$$C_{\mathbf{Y}}(f)(\mathbf{x}) := \begin{cases} \mathbf{Y} & \text{if } p(\mathbf{x}, \mathbf{Y}) q(f(\mathbf{x})) > \xi p(\mathbf{x}, f(\mathbf{x})) q(\mathbf{Y}) \\ f(\mathbf{x}) & \text{otherwise,} \end{cases} \tag{4.1}$$

where $\xi \sim U(0,1)$ and $\mathbf{Y} \sim Q$. In general terms, the idea is to sample from a modification of the original transition kernel P, and use the same procedure to other chains in such a way that they have a positive probability to coalesce at the value $\mathbf{Y}$. The proposed modification does not affect the marginal distribution, i.e.

$$\Pr(C_{\mathbf{Y}}(f)(\mathbf{x}) \in d\mathbf{y}) = p(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \Pr(f(\mathbf{x}) \in d\mathbf{y}),$$

and this follows directly from the form of $C_{\mathbf{Y}}(f)(\mathbf{x})$, that can be viewed as a Metropolis step with target $p(\mathbf{x}, \cdot)$ and proposal distribution q. The procedure in (4.1) requires the knowledge of p and $q(\cdot)$ only up to the normalizing constant and furthermore exhibits a fixed computational cost, from which it derives the appeal of such method.

For $(\mathbf{X}, \mathbf{Y}) \in \Gamma(p, q)$, a naive implementation of the Catalytic coupling is reported in Algorithm 13. where p_u, q_u denote the unnormalized versions of p, q respectively. The

Algorithm 13: Catalytic (independent) coupling of $\mathbf{X} \sim p, \mathbf{Y} \sim q$

Input: densities p, q;
Sample $\mathbf{X} \sim p$
Sample $\mathbf{Y} \sim q$
Sample $\xi \sim Unif[0, q_u(\mathbf{Y}) p_u(\mathbf{X})]$
if $\xi < q_u(\boldsymbol{X}) p_u(\boldsymbol{Y})$ **then**
  ∟ set $\mathbf{Y} = \mathbf{X}$
Output: $(\mathbf{X}, \mathbf{Y})$.

idea is to use a sample from p as proposed modification for q and accept it with the

Metropolis rate. It is possible to quantify the loss in coalescence probability between Algorithm 13 and a maximal coupling, as made clear in Lemma 18

Lemma 18. *Let $p, q \in \mathcal{P}(\mathcal{X})$, $\boldsymbol{X} \sim p$, $\boldsymbol{Y} \sim q$, coupled through Algorithm 13, then*

$$\Pr(\boldsymbol{X} = \boldsymbol{Y}) \geq \left(\Pr_{max}(p, q)\right)^2 \geq 1 - 2\|p - q\|_{tv},$$

where $\Pr_{max}(p, q)$ denotes the probability of coalescence under a maximal coupling.

Proof. From the form of Algorithm 13, it follows that

$$\Pr(\mathbf{X} = \mathbf{Y}|\mathbf{X}) = \mathbb{E}\left[\min\left(1, \frac{q_u(\mathbf{X})}{p_u(\mathbf{X})}\frac{p_u(\mathbf{Y})}{q_u(\mathbf{Y})}\right)\right] = \mathbb{E}\left[\min\left(1, \frac{q(\mathbf{X})}{p(\mathbf{X})}\frac{p(\mathbf{Y})}{q(\mathbf{Y})}\right)\right],$$

where the expectation is wrt q, hence the coupling probability

$$\begin{aligned}
\Pr(\mathbf{X} = \mathbf{Y}) &= \int \mathbb{E}\left[\min\left(1, \frac{q_u(\mathbf{x})}{p_u(\mathbf{x})}\frac{p_u(\mathbf{y})}{q_u(\mathbf{y})}\right)\right] p(d\mathbf{x}) \\
&= \iint \min\left(p(\mathbf{x})q(\mathbf{y}), q(\mathbf{x})p(\mathbf{y})\right)\, d\mathbf{y}\, d\mathbf{x} \\
&= 1 - \|p \otimes q - q \otimes p\|_{tv} \\
&\geq (1 - \|p - q\|_{tv})^2.
\end{aligned}$$

□

From the above it follows that implementing catalytic coupling on distributions that are close in total variation distance results in a nearly maximal coupling.

Algorithm 13 is suboptimal in many ways: using draws from q as Metropolis proposals might result in low acceptance rates and hence low coupling probabilities. A more tailored proposal can be used: one could draw $Z \sim h$ for h some carefully chosen distribution, and then use it as catalyst for both $\mathbf{X}$ and $\mathbf{Y}$. Coalescence happens if both accept. A possible implementation using same random number is presented in Algorithm 14. Then

Algorithm 14: Catalytic coupling of $\mathbf{X} \sim p, \mathbf{Y} \sim q$

Input: densities p, q, h;
Sample $\mathbf{X}_0 \sim p$, $\mathbf{Y}_0 \sim q$, $\mathbf{Z} \sim h$
Sample $\xi \sim Unif[0,1]$
if $\xi < \frac{h_u(\boldsymbol{X})p_u(\boldsymbol{Z})}{h_u(\boldsymbol{Z})p_u(\boldsymbol{X})}$
then
  set $\mathbf{X} = \mathbf{Z}$
else
  set $\mathbf{X} = \mathbf{X}_0$
if $\xi < \frac{h_u(\boldsymbol{Y})q_u(\boldsymbol{Z})}{h_u(\boldsymbol{Z})q_u(\boldsymbol{Y})}$
then
  set $\mathbf{Y} = \mathbf{Z}$
else
  set $\mathbf{Y} = \mathbf{Y}_0$
Output: $(\mathbf{X}, \mathbf{Y})$.

the acceptance probability becomes

$$\begin{aligned}\Pr(\mathbf{X} = \mathbf{Y} | \mathbf{X}_0, \mathbf{Y}_0, \mathbf{Z}) &= \Pr(\mathbf{X} = \mathbf{Z}, \mathbf{Y} = \mathbf{Z} | \mathbf{X}_0, \mathbf{Y}_0, \mathbf{Z}) \\ &= min\left(1, \frac{h(\mathbf{X}_0)}{p(\mathbf{X}_0)}\frac{p(\mathbf{Z})}{h(\mathbf{Z})}, \frac{h(\mathbf{Y}_0)}{q(\mathbf{Y}_0)}\frac{q(\mathbf{Z})}{h(\mathbf{Z})}\right).\end{aligned}$$

In an ongoing work, we are currently investigating the links between the formulations Catalytic couplings and the pseudo marginal approach for MCMC simulations (Andrieu and Roberts, 2009): the acceptance ratio for Algorithm 13 equals that of a maximal coupling times an estimate of the normalizing constant, as for general implementations of marginal algorithms.

4.2 Adaptive Rejection sampling

Adaptive rejection sampling (Gilks and Wild, 1992) is an efficient method for rejection sampling from any univariate log-concave probability density function. It requires a minimal number of evaluations of the density function, leveraging the construction of upper and lower envelopes used for sampling and compute acceptance probabilities, hence well suited for situations where the evaluation of the density is computationally expensive. In brief, let $f(x)$ be the density to sample from, suppose it is known up to normalizing

constants, i.e. $g(x) = cf(x)$ is known. Let $h(x) = \ln g(x)$, and let D be the domain of h, assumed to be connected. Suppose that $h(x)$ and $h'(x)$ have been evaluated in the points of $T_k := \{x_1 < ... < x_k \in \mathbb{R}\}$. The upper rejection envelope on T_k is defined as $\exp u_k(x)$, where $u_k(x)$ is a piecewise linear function formed by the tangents to $h(x)$ in the points of T_k, namely:

$$\begin{aligned} u_k(x) &= h(x_j) + h'(x_j)(x - x_j) && \text{for } x \in [z_{j-1}, z_j],\ j = 1, ..., k; \\ z_j &= \frac{h(x_{j+1}) - h(x_j) - x_{j+1}f'(x_{j+1}) + x_j h'(x_j)}{h'(x_{j+1}) - h'(x_j)} && j = 1, ..., k-1. \end{aligned}$$

The resulting proposal distribution is

$$s_k(x) = \frac{\exp u_k(x)}{\int_D \exp u_k(x')dx'} \qquad \text{for } x \in [z_{j-1}, z_j],\ j = 1, ..., k.$$

The lower hull is formed from the chords between adjacent abscissae in T_k, and is used mainly for computing the rejection probabilities without evaluating each time the expensive density $g(x)$. The sampling procedure is reported in Algorithm 15.

Algorithm 15: Adaptive rejection sampling for univariate log concave density f

Input: log density $h(x)$, $h(x_j)$ and $h'(x_j)$ for $j = 1, ..., k$;
while *not accept* **do**
  Sample $x \sim S_k$ with density $s_k(x)$
  Sample $w \sim U(0, 1)$
  if $w < \exp\{l_k(x) - u_k(x)\}$ **then**
    accept x
  else
    Evaluate $h(x)$ and $h'(x)$
    if $w < \exp\{h(x) - u_k(x)\}$ **then**
      accept x
Output: sample $x \sim f$.

Suppose one wants to couple Markov chains where the update of some coordinate is done via rejection sampling: implementing Algorithm 1 and Algorithm 2 would be impossible since the transition density is known up to a normalising constant. Leveraging Catalytic couplings of Section 4.1, is still possible to carry on the coupling procedure.

4.2.1 Applications to GLMMs with crossed effects

We consider the same setting and models of Section 2.7.2. We coupled chains marginally evolving according to Algorithm 7. The advances of Section 4.1 and 4.2 allow us to simulate directly from a coupling of the exact full conditionals without the need to resort to Metropolis within Gibbs schemes. Recent findings of Ascolani et al. (2024) showed that the performances of a generic coordinate wise scheme (measured through the conductance of the associated operator) differ from the ones of a random scan Gibbs sampler by a multiplicative factor controlled through the goodness of the conditional update, measured via the notion of conditional conductance (see Corollary 1, Proposition 1 and 3). Hence implementing adaptive rejection sampling on the updates of $\xi_i^{(k)}$ for $i = 1, ..., I_k; k = 1, .., K$ should result in a speed up of the mixing of the chains and hence, in view of the results of Theorem 4, in smaller meeting times of the chains. We report in Figure 4.1 the average meeting time for the same chains of Section 2.7.2 and the one obtained implementing catalytic coupling and adaptive rejection sampling ("ARS & catalytic" in the legend). Although generally more expensive than a Metropolis within Gibbs with $S =$

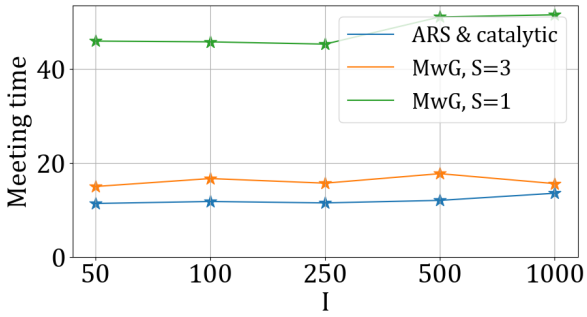

Figure 4.1: Estimated mean meeting time for $K = 2, I_1 = I_2 = \{50, 100, 250, 500, 1000\}$, $\tau_1 = \tau_2 = 1, b = 1$ with Laplace response.

1 (as those implented in Chapter 2), adaptive rejection samplers provide nearly perfect samples and hence result in the smallest meeting times among the explored options, substantiating our heuristic that faster mixing chains results in lower meeting times.

References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697 – 725.

Ascolani, F., Roberts, G. O., and Zanella, G. (2024). Scalability of Metropolis-within-Gibbs schemes for high-dimensional Bayesian models.

Breyer, L. A. and Roberts, G. O. (2001). Catalytic perfect simulation. *Methodology And Computing In Applied Probability*, 3:161–177.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 41(2):337–348.