Journal of Machine Learning Research 27 (2026) 1-42 Submitted 9/25; Revised 5/26; Published 6/26

Mixing times of data-augmentation Gibbs samplers for high-dimensional probit regression

Filippo Ascolani FILIPPO.ASCOLANI@DUKE.EDU
Department of Statistical Science,
Duke University
Durham, NC 27708, USA

Giacomo Zanella GIACOMO.ZANELLA@UNIBOCCONI.IT
Department of Decision Sciences and BIDSA,
Bocconi University
Milano, 20136, Italy

Editor: Anthony Lee

Abstract

We investigate the convergence properties of popular data-augmentation samplers for Bayesian probit regression. Leveraging recent results on Gibbs samplers for log-concave targets, we provide simple and explicit non-asymptotic bounds on the associated mixing times (in Kullback-Leibler divergence). The bounds depend explicitly on the design matrix and the prior precision, while they hold uniformly over the vector of responses. We specialize the results for different regimes of statistical interest, when both the number of data points n and parameters p are large: in particular we identify scenarios where the mixing times remain bounded as $n, p \to \infty$, and ones where they do not. The results are shown to be tight (in the worst case with respect to the responses) and provide guidance on choices of prior distributions that provably lead to fast mixing. An empirical analysis based on coupling techniques suggests that the bounds are effective in predicting practically observed behaviours.

Keywords: Bayesian binary regression, Markov chains, entropy contraction, random design regression.

1. Introduction

1.1 The model

Probit regression is a popular methodology when the relationship between binary data $y_i \in \{0, 1\}$ and a set of predictors is of interest. It is an instance of generalized linear model (McCullagh and Nelder, 1989) and its usual Bayesian formulation reads

$$y_i \mid \beta \sim \text{Bernoulli}(\Phi(x_i^T \beta)), \quad \beta \sim N(m, Q_0^{-1}), \qquad i = 1, \dots, n \tag{1}$$

where $x_i \in \mathbb{R}^p$ is the i-th row of a design matrix $X \in \mathbb{R}^{n \times p}$, $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian random variable and $N(m, Q_0^{-1})$ denotes the multivariate normal distribution with mean m and precision matrix Q_0. Given data

$y = (y_1, \dots, y_n) \in \{0,1\}^n$, the posterior distribution of β has density

$$\pi(\beta) \propto N(\beta \mid m, Q_0^{-1}) \prod_{i=1}^{n} \Phi(x_i^T \beta)^{y_i} (1 - \Phi(x_i^T \beta))^{1-y_i}, \tag{2}$$

where $N(\beta \mid m, Q_0^{-1})$ denotes the density of $N(m, Q_0^{-1})$ at β. Several strategies have been developed to approximate this distribution, ranging from exact rejection sampling (Botev, 2017; Durante, 2019), variational inference (Chopin and Ridgway, 2017; Fasano et al., 2022) and sampling with Markov chain Monte Carlo (MCMC) techniques (Albert and Chib, 1993; Held and Holmes, 2006), which is the focus of this article. A popular class of MCMC methods for $\pi(\beta)$ relies on a data augmentation scheme based on re-writing model (1) as

$$\begin{aligned} &y_i = \mathbb{1}(z_i > 0) && i = 1, \dots, n, \\ &z|\beta \sim N(X\beta, I_n), \quad \beta \sim N(m, Q_0^{-1}), \end{aligned} \tag{3}$$

where $z = (z_1, \dots, z_n)^T \in \mathbb{R}^n$, I_n denotes the $n \times n$ identity matrix and $\mathbb{1}$ denotes the indicator function. The joint posterior density of z and β is

$$\pi(z, \beta) \propto N(\beta \mid m, Q_0^{-1}) N(z \mid X\beta, I_n) \prod_{i=1}^{n} \mathbb{1}\left(y_i = g(z_i)\right), \tag{4}$$

where $g(z_i) = \mathbb{1}(z_i > 0)$, and its marginal density over β coincides with (2).

1.2 The algorithms

We consider two popular Gibbs Sampling schemes used to draw samples from $\pi(z, \beta)$.

1.2.1 Data Augmentation (DA)

First, we consider the two-block deterministic-scan Gibbs Sampler originally proposed in Albert and Chib (1993), which alternates the update of z from

$$\pi(z \mid \beta) \propto N(z \mid X\beta, I_n) \prod_{i=1}^{n} \mathbb{1}(y_i = g(z_i))$$

and β from

$$\pi(\beta \mid z) = N\left(\beta \mid (X^T X + Q_0)^{-1}(Q_0 m + X^T z), (X^T X + Q_0)^{-1}\right). \tag{5}$$

Equivalently, its Markov kernel P_{DA} is defined as the composition of two kernels, $P_{\mathrm{DA}} = P_\beta P_z$, with

$$P_z((z,\beta), (\mathrm{d}z', \mathrm{d}\beta')) = \delta_\beta(\mathrm{d}\beta')\pi(\mathrm{d}z' \mid \beta) \quad \text{and} \quad P_\beta((z,\beta), (\mathrm{d}z', \mathrm{d}\beta')) = \delta_z(\mathrm{d}z')\pi(\mathrm{d}\beta' \mid z) \tag{6}$$

for $z \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$. The pseudocode for P_{DA} is given in Algorithm 1. Note that the conditional distribution $\pi(z \mid \beta)$ factorizes across the n components of z, so that P_z entails sampling n independent truncated normal random variables. Thus, both P_z and P_β can be implemented in closed form, which is the main computational advantage of the latent variable representation in (3).

Algorithm 1 (Data Augmentation Gibbs sampler P_{DA})

Initialize $\beta^{(0)}$.
for $t = 1, 2, \ldots$ **do**
Sample $z_i^{(t)} \sim \pi(z_i \mid \beta^{(t-1)}) \propto N(z_i \mid x_i^T \beta^{(t-1)}, 1)\mathbb{1}(y_i = g(z_i))$ independently for $i = 1, \ldots, n$.
Sample $\beta^{(t)} \sim \pi(\beta \mid z^{(t)})$ with $\pi(\beta \mid z)$ as in (5).
end for

1.2.2 Collapsed Gibbs (CG)

Second, we consider the n-blocks random scan Gibbs sampler on $\mathcal{X} = \mathbb{R}^n$ targeting

$$\pi(z) = \int_{\mathbb{R}^p} \pi(z, \beta) d\beta \propto N(z \mid Xm, I_n + XQ_0^{-1}X^T) \prod_{i=1}^{n} \mathbb{1}(y_i = g(z_i)), \tag{7}$$

which we refer to as *Collapsed Gibbs* (CG) sampler. Its Markov kernel P_{CG} is defined as

$$P_{\mathrm{CG}}(z, \mathrm{d}z') = \frac{1}{n} \sum_{i=1}^{n} P_{\mathrm{CG},i}(z, \mathrm{d}z'), \quad \text{with } P_{\mathrm{CG},i}(z, \mathrm{d}z') = \delta_{z_{-i}}(\mathrm{d}z'_{-i}) \pi(\mathrm{d}z'_i \mid z_{-i}), \tag{8}$$

where $\pi(\mathrm{d}z_i \mid z_{-i})$ is the conditional distribution with density

$$\pi(z_i \mid z_{-i}) \propto N\left(z_i \mid (1-h_i)^{-1} x_i^T V X^T (z - Q_0 m) - h_i (1-h_i)^{-1} z_i, (1-h_i)^{-1}\right) \mathbb{1}(y_i = g(z_i)), \tag{9}$$

with $V = (X^T X + Q_0)^{-1}$ and $h_i = x_i^T V x_i$. See Section 2 in Held and Holmes (2006) for a derivation of $\pi(z)$ and its full conditionals. The pseudocode for P_{CG} is given in Algorithm 2. The kernel P_{CG} can be used to sample from $\pi(z)$ and, given a sample from $\pi(z)$, one can

Algorithm 2 (Collapsed Gibbs sampler P_{CG})

Initialize $z^{(0)} \in \mathbb{R}^n$.
for $t \geq 1$ **do**
Sample I uniformly at random from $\{1, \ldots, n\}$.
Given $I = i$, sample $z_i^{(t)} \sim \pi(z_i \mid z_{-i}^{(t-1)})$, with $\pi(z_i \mid z_{-i})$ as in (9).
end for

obtain samples from $\pi(z, \beta)$ by drawing β from $\pi(\beta \mid z)$ defined in (5). The deterministic scan version of P_{CG} was originally considered in Held and Holmes (2006). Here we consider the random scan version for theoretical convenience, since the latter is easier to analyse in our context.

2. Main results

2.1 KL-mixing times

In this paper we measure distance to stationarity using the Kullback-Leibler (KL) divergence. For every $\mu, \nu \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures over $\mathcal{X}$,

let

$$\mathrm{KL}(\mu,\nu)=\int_{\mathcal{X}}\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\right)\mu(\mathrm{d}x),$$

where $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ denotes the Radon-Nikodym derivative between μ and ν. Given a π-invariant Markov kernel P and a starting distribution $\mu\in\mathcal{P}(\mathcal{X})$, we define the mixing times with respect to KL as

$$\tau_{\mathrm{mix}}(\epsilon,\mu,P):=\inf\{t\geq 1\,:\,\mathrm{KL}(\mu P^t,\pi)\leq\epsilon\},\qquad \epsilon\in[0,\infty),\tag{10}$$

where $\mu P^t(A)=\int_{\mathcal{X}}P^t(x,A)\mu(\mathrm{d}x)$ for every $A\subseteq\mathcal{X}$. In words, $\tau_{\mathrm{mix}}(\epsilon,\mu,P)$ is the number of iterations needed for the chain to be ϵ-close in KL to its stationary distribution π, when starting from μ.

2.2 Mixing time bounds

The goal of this work is to quantify the computational effort required by P_{DA} and P_{CG} to produce approximate samples from $\pi(z,\beta)$. The key task in doing so is to upper bound their mixing times. Since the cost per iteration of P_{DA} is n times larger than that of P_{CG} (see Section A of the Supplementary Material for details), we will compare a single iteration of P_{DA} to n iterations of P_{CG}. In other words, we express results in terms of $\tau_{\mathrm{mix}}(\epsilon,\mu,P_{\mathrm{DA}})$ and $\tau_{\mathrm{mix}}(\epsilon,\mu_1,P_{\mathrm{CG}}^n)$, where $\mu\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^p)$ and $\mu_1\in\mathcal{P}(\mathbb{R}^n)$ denotes the first marginal of μ. The next theorem provides an explicit upper bound to these two quantities.

Theorem 1 *For every $\mu\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^p)$ and $\epsilon>0$, we have*

$$\tau_{\mathrm{mix}}(\epsilon,\mu,P_{\mathrm{DA}})\leq(2+\lambda_{max}(XQ_0^{-1}X^T))\log\left(\frac{\mathrm{KL}(\mu,\pi)}{\epsilon}\right),\tag{11}$$

and

$$\tau_{\mathrm{mix}}(\epsilon,\mu_1,P_{\mathrm{CG}}^n)\leq\frac{1+\lambda_{max}(XQ_0^{-1}X^T)}{1+\lambda_{min}(XQ_0^{-1}X^T)}\log\left(\frac{\mathrm{KL}(\mu,\pi)}{\epsilon}\right),\tag{12}$$

where λ_{min} and λ_{max} denote the minimum and maximum (modulus) eigenvalues of the given matrix.

Proof The proof is postponed to Section D.1 of the Supplementary Material. ■

In Section 8 we provide a so-called feasible starting distribution μ such that $\log(\mathrm{KL}(\mu,\pi))$ is at most of order $\log\left(n+n\log\left(n\lambda_{max}(XQ_0^{-1}X^T)\right)\right)$, see Proposition 15. Thus, by Theorem 1, the mixing times of both P_{DA} and P_{CG}^n are upper bounded by $\lambda_{max}(XQ_0^{-1}X^T)$, up to constants and logarithmic terms.

Remark 2 *The bounds in Theorem 1 hold for every fixed X and y. The right-hand side is independent of y, which means that it considers the worst-case with respect to the data y. This will be important in interpreting the results later on.*

Remark 3 *The statements in Theorem 1 are actually a consequence of a stronger result. For example in the case of P_{CG} we upper bound the one-step entropy contraction as*

$$\frac{\mathrm{KL}(\mu_1 P_{\mathrm{CG}}, \pi_1)}{\mathrm{KL}(\mu_1, \pi_1)} \leq 1 - \frac{1}{n}\left[\frac{1+\lambda_{min}(XQ_0^{-1}X^T)}{1+\lambda_{max}(XQ_0^{-1}X^T)}\right], \tag{13}$$

for every $\mu_1 \in \mathcal{P}(\mathbb{R}^n)$ such that $\mathrm{KL}(\mu_1, \pi_1) < \infty$. This implies (12), see (24) and Theorem 24 in Appendix D. The case of P_{DA} is more complex, due to the lack of reversibility: we provide a general framework to study entropy contraction for data augmentation schemes in Section 7. Bounds as in (13) allow to study mixing times in other metrics. For example, denoting with

$$\chi^2(\mu,\nu) = \int_{\mathcal{X}} \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x) - 1\right)^2 \nu(\mathrm{d}x), \qquad \mu,\nu \in \mathcal{P}(\mathcal{X})$$

the chi-square divergence and with $\tau_{\mathrm{mix},2}$ the mixing times obtained by replacing $\mathrm{KL}(\mu P^t, \pi)$ in (10) with $\chi^2(\mu P^t, \pi)$, we have

$$\tau_{\mathrm{mix},2}(\epsilon, \mu, P_{\mathrm{DA}}) \leq (3 + 2\lambda_{max}(XQ_0^{-1}X^T)) \log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right)$$

and

$$\tau_{\mathrm{mix},2}(\epsilon, \mu_1, P_{\mathrm{CG}}^n) \leq 2\,\frac{1+\lambda_{max}(XQ_0^{-1}X^T)}{1+\lambda_{min}(XQ_0^{-1}X^T)} \log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right).$$

See Section C.3 and Theorem 23 in the Supplementary Material for details and a formal proof. Since $\chi^2(\mu,\nu) \geq \mathrm{KL}(\mu,\nu)$ the latter inequalities are not implied by (11) and (12) alone: moreover at this level of generality they are tight, see Proposition 8 for a matching lower bound in a special case.

Remark 4 *While in the main body of the paper we focus on probit regression for simplicity, P_{DA} and P_{CG} can also be used to sample from other popular models that can be expressed as partially or fully discretized Gaussian linear regression, such as multinomial probit and tobit regressions, see e.g. Anceschi et al. (2023) for a recent review. In Section B in the supplementary material we illustrate extensions of Theorem 1 to those cases.*

2.2.1 Proof technique

The main idea underlying the proofs is to recognize that the target distributions of both P_{DA} and P_{CG} can be written as

$$\pi(x) = N(x \mid \bar{m}, \bar{Q}^{-1}) \prod_{i=1}^{d} g_i(x_i), \quad x \in \mathbb{R}^d,$$

for some appropriate x, d, $\bar{m}$ and $\bar{Q}$, where $-\log(g_i(x_i))$ is a convex function for every i. Then, the results in Ascolani et al. (2026) (in particular Theorem 3.1 therein) imply that the mixing times of a random scan Gibbs sampler on π can be upper bounded by the ones of a random scan Gibbs sampler on $N(x \mid \bar{m}, \bar{Q}^{-1})$. This allows to derive explicit expressions,

since Gibbs samplers on Gaussian targets are amenable to analytical study (see e.g. Ascolani et al. (2026, Lemma 3.10), and earlier work in Amit (1996); Roberts and Sahu (1997)).

In our context, this means that the mixing time of P_{DA} is upper bounded by the one of the two-block Gibbs Sampler targeting the corresponding prior distribution, i.e. $N(\beta \mid m, Q_0^{-1})N(z \mid X\beta, I_n)$, and similarly for P_{CG} and the corresponding marginal prior on z. In other words, we can ignore the likelihood because in this context it can only speed up the convergence of both P_{DA} and P_{CG}.

The results of Ascolani et al. (2026) are limited to the random scan case: we apply them directly to P_{CG} in Section D.1 of the Supplementary Material. Extending the approach to the case of two-block deterministic-scan samplers (also called Data Augmentation samplers, such as P_{DA}), requires some technical work, which we carry out in Section 7. Before doing that, we discuss the related literature (Section 2.3) and analyze the implications of Theorem 1 in various regimes of statistical interest, namely g priors (Section 2.4), random design models (Section 2.4), models with and without the intercept (Section 3); discuss the resulting computational complexity and compare it with the one of gradient-based samplers (Section 4) and provide numerical illustrations (Section 5). Finally, we provide some guarantees on using the prior as a starting distribution in Section 8. The code to reproduce the numerical experiments is available at `https://github.com/gzanella/ProbitDA`.

2.3 Related literature

Other papers have studied the convergence properties of P_{DA}. For example, Roy and Hobert (2007) proved that P_{DA} is geometrically ergodic through drift and minorization techniques (Rosenthal, 1995). However, such bounds deteriorate with n and p, making them not informative in high-dimensional problems. Subsequent works, and in particular Qin and Hobert (2019, 2022), showed that such bounds could be substantially improved under various assumptions, such as p fixed and $n \to \infty$ with proper assumptions on the data-generating mechanism (Theorem 17 in Qin and Hobert (2019)), n fixed and $p \to \infty$ (Theorem 22 in Qin and Hobert (2019)), as well as other settings such as repeated rows in the matrix X (see in particular Qin and Hobert (2022)).

The most complete bounds up to now are given in the recent work Lee and Zhang (2024), where the authors study the mixing times in Total Variation distance through the profile conductance of P_{DA}. By Pinsker inequality, the KL bound in (11) also implies a corresponding bound in Total Variation: in this context the results in Theorems 3.2 and 3.6 in Lee and Zhang (2024) are similar to ours in terms of overall resulting complexity (as a function of n and p), but we employ an arguably easier proof technique (based on Ascolani et al. (2026)) that leads to more explicit and smaller constants. Moreover Theorem 1 depends on the starting distribution μ through $\mathrm{KL}(\mu, \pi)$, without requiring the stronger assumption that μ is warm, i.e. $\sup_x \frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x) < \infty$. In addition, the KL mixing times bounds in Theorem 1 follow by the stronger result we prove on the one-step entropy contraction: see Section 7 for details. This implies a lower bound on the spectral gap (see Caputo (2023) below Lemma 2.15) and upper bounds on χ^2-mixing times (see Remark 3).

On the contrary, we found no explicit results on the convergence of P_{CG}. Thus the bound in (12), again based on Ascolani et al. (2026) (see Section D.1 of the Supplementary Material), is novel to the best of our knowledge.

2.4 Implications

We now consider two popular choices of Q_0 and investigate the implications of Theorem 1.

2.4.1 g prior

If X^TX is invertible, then the so-called g prior (Zellner, 1986; Liang et al., 2008) is given by $Q_0^{-1} = g(X^TX)^{-1}$, with $g \in (0, \infty)$ being a multiplicative scalar. The g prior requires X^TX to be invertible, which for example cannot hold when $p > n$. We thus consider the more general case $Q_0^{-1} = (X^TX/g + cI_p)^{-1}$ with $c \geq 0$, which is always well defined if $c > 0$ and reduces to the standard g prior if $c = 0$. In the next corollary we obtain upper bounds on the mixing times for those cases.

Corollary 5 *Let $Q_0^{-1} = (X^TX/g + cI_p)^{-1}$, with $c \geq 0$. Then*

$$\tau_{\text{mix}}(\epsilon, \mu, P_{\text{DA}}) \leq (2+g)\log\left(\text{KL}(\mu,\pi)/\epsilon\right), \quad \tau_{\text{mix}}(\epsilon, \mu_1, P^n_{\text{CG}}) \leq (1+g)\log\left(\text{KL}(\mu,\pi)/\epsilon\right).$$

Proof We have $XQ_0^{-1}X^T = X(X^TX/g + cI_p)^{-1}X^T = gX(cgI_p + X^TX)^{-1}X^T$. By Woodbury's identity

$$X(cgI_p + X^TX)^{-1}X^T = I_n - (I_n + XX^T/(cg))^{-1}.$$

The above equalities imply $\lambda_{max}(XQ_0^{-1}X^T) \leq g$ and the bounds follow from Theorem 1. ■

Interestingly the upper bounds in Corollary 5 do not depend on X, nor on n and p. This implies that convergence speed does not deteriorate in high dimensions if g is held fixed. On the other hand, the bounds increase with g, i.e. as the prior becomes less informative. These features are confirmed by the simulations, and they occur not only for worst-case data y but also under randomly generated y, see Section 5.

2.4.2 Diagonal precision under random design

We now consider the case of isotropic prior and random design matrix $X = (X_{ij})_{ij} \in \mathbb{R}^{n\times p}$, as specified in the next assumption.

Assumption A *Assume either:*
(a) $Q_0^{-1} = cI_p$ and $X_{ij} = G_{ij}/\sqrt{p}$ or
(b) $Q_0^{-1} = (c/p)I_p$ and $X_{ij} = G_{ij}$,
with $c > 0$ and $G_{ij} \overset{i.i.d.}{\sim} F$ for $(i,j) \in \{1,\dots,n\} \times \{1,\dots,p\}$, where $F \in \mathcal{P}(\mathbb{R})$ has zero mean, unit variance and finite fourth moment.

Rescaling either Q_0^{-1} or X_{ij}^2 by $1/p$, as in Assumption A, is a standard practice in high-dimensional Bayesian regression (Simpson et al., 2017; Fuglstad et al., 2020), which ensures that the variance of linear predictors $x_i^T\beta$ under the prior remains roughly constant as p increases, since $\text{Var}(x_i^T\beta) = \frac{c}{p}\sum_{j=1}^p G_{ij}^2 \to c$ almost surely as $p \to \infty$.

The random design assumption allows us to use random matrix theory to obtain high-probability bounds on $\lambda_{max}(XX^T)$ and $\lambda_{min}(XX^T)$, and thus on mixing times, as detailed in the next corollary.

Corollary 6 *Under Assumption A we have that*

$$\tau_{\text{mix}}(\epsilon, \mu, P_{\text{DA}}) \leq \left[2 + c(1+\sqrt{r})^2 + \delta\right]\log\left(\text{KL}(\mu,\pi)/\epsilon\right)$$

and

$$\tau_{\text{mix}}(\epsilon, \mu_1, P^n_{\text{CG}}) \leq \frac{1 + c(1+\sqrt{r})^2 + \delta}{1 + c(1 - \sqrt{\min\{1, r\}})^2} \log\left(\text{KL}(\mu, \pi)/\epsilon\right),$$

almost surely as $n \to \infty$ *and* $n/p \to r \in (0, \infty)$*, for any arbitrarily small positive constant* $\delta > 0$*.*

Proof Denoting $G = (G_{ij})_{ij} \in \mathbb{R}^{n \times p}$, Theorem 2 in Bai et al. (1993) implies that

$$\lambda_{max}(XQ_0^{-1}X^T) = \frac{c}{p}\lambda_{max}(GG^T) \to c(1+\sqrt{r})^2, \tag{14}$$

almost surely as $n \to \infty$ and $n/p \to r \in (0, \infty)$, and, if $r < 1$, also $\lambda_{min}(XQ_0^{-1}X^T) \to c(1-\sqrt{r})^2$ almost surely. Combining those with Theorem 1 gives the result. ■

The results of Corollary 6 provide various insights, namely:

(i) Both P_{DA} and P_{CG} mix fast (i.e. in $\mathcal{O}(1)$ iterations) in high-dimensional scenarios where p is comparable to (or larger than) n and c is small.

(ii) When $p < n$ the bound on P_{CG} is similar to the one on P_{DA}. When $p > n$ and c is large, the bound on P_{CG} can be better. This dependence on c and r is also confirmed by numerical experiments, see Section E in the Supplementary Material.

(iii) If c is fixed, both bounds deteriorate when n/p grows.

These features are empirically confirmed by the simulation study in Section 5.

3. The role of the intercept (and unbalanced data)

Contrarily to Assumption A, where each column is rescaled by a factor of $1/\sqrt{p}$, we now assume that the first column of X has all entries equal to 1, i.e. that β_1 is the intercept. It is indeed common practice not to rescale the intercept (see e.g. (Sardy, 2008, Section 2)), or equivalently not to strongly shrink it towards 0, in order to allow the intercept to account for the average proportion of ones in y marginally over the covariates. We state the assumption only in terms of rescaling covariates for the sake of brevity.

Assumption B $Q_0^{-1} = cI_p$*,* $X_{i1} = 1$ *for* $i \in \{1, \dots, n\}$*,* $X_{ij} = G_{ij}/\sqrt{p}$ *and* $G_{ij} \overset{i.i.d.}{\sim} F$ *for* $i \in \{1, \dots, n\}$ *and* $j \in \{2, \dots, p\}$*, where* $c > 0$ *and* $F \in \mathcal{P}(\mathbb{R})$ *has zero mean, unit variance and finite fourth moment.*

Including the intercept has a major impact on the mixing of P_{DA} and P_{CG}, as shown below.

Corollary 7 *Under Assumption B we have that*

$$\tau_{\text{mix}}(\epsilon, \mu, P_{\text{DA}}) \leq (2 + (c+\delta)n) \log\left(\text{KL}(\mu, \pi)/\epsilon\right)$$

and

$$\tau_{\text{mix}}(\epsilon, \mu_1, P^n_{\text{CG}}) \leq (1 + (c+\delta)n) \log\left(\text{KL}(\mu, \pi)/\epsilon\right),$$

almost surely as $n \to \infty$ *and* $n/p \to r \in (0, \infty)$*, for any arbitrarily small positive constant* $\delta > 0$*.*

Proof By construction the matrix X^TX, whose non-zero eigenvalues are the same of XX^T, has the form

$$X^TX = \begin{bmatrix} n & \sum_{i=1}^n X_{i2} & \dots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i2} & 0 & \dots & 0 \\ \vdots & & & \\ \sum_{i=1}^n X_{ip} & 0 & \dots & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \tilde{X}^T\tilde{X} & \\ 0 & & & \end{bmatrix}$$

where $\tilde{X} \in \mathbb{R}^{n \times p-1}$ is the matrix X without the first column. Then by Weyl's inequality, we have that $\lambda_{max}(X^TX) \leq n + \lambda_{max}(\tilde{X}^T\tilde{X})$ and $\lambda_{max}(\tilde{X}^T\tilde{X}) \to (1+\sqrt{r})^2$ almost surely by Theorem 2 in Bai et al. (1993). Thus $\lambda_{max}(XQ_0^{-1}X^T) \leq (c+\delta)n$ almost surely as $n \to \infty$ and $n/p \to r$ for every $\delta > 0$, and the result follows by Theorem 1. ■

The bounds in Corollary 7 deteriorate linearly with n, implying that $\mathcal{O}(n)$ iterations are sufficient for convergence. The next proposition, in the intercept-only case (i.e. $p = 1$), shows that this is not improvable in general and it corresponds to the case of highly imbalanced data. The proof can be found in Section D.2 of the Supplementary Material.

Proposition 8 *Let $p = 1$, $m = 0$, $Q_0^{-1} = c > 0$, $x_i = 1$ for every $i \in \{1, \dots, n\}$ and $y_i = 1$ for every i (or $y_i = 0$ for every i). Then, for every $n \geq 8/c$*

(i) we have that

$$\sup_{\mu} \frac{\mathrm{KL}(\mu P_{\mathrm{DA}}, \pi)}{\mathrm{KL}(\mu, \pi)} \geq 1 - \frac{\log(cn)}{d(1+cn)}$$

for a universal constant $d > 0$, where the supremum is taken over every $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R})$ such that $\mathrm{KL}(\mu, \pi) < \infty$.

(ii) there exists $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R})$ such that

$$(3+2cn)\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right) \geq \tau_{\mathrm{mix},2}(\epsilon, \mu, P_{\mathrm{DA}}) \geq d'\left(\frac{1+cn}{\log(cn)}\right)\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right) \quad (15)$$

for a universal constant $d' > 0$ and all $\epsilon > 0$.

Remark 9 *Proposition 8 is inspired by results in Johndrow et al. (2019), which prove that P_{DA} with imbalanced data and intercept only mixes slowly as n increases. In particular Theorem 3.2 therein implies a lower bound of order $\sqrt{n}$ on the mixing times in Total variation distance. Here we improve this to order n (ignoring logarithmic terms) by passing to the χ^2 divergence.*

The proof of Proposition 8 relies on showing that $\mathrm{Var}(\beta_1 \mid z) = \mathcal{O}(n^{-1})$ for every z while $\mathrm{Var}_\pi(\beta_1) = \mathcal{O}(1)$ in the imbalanced case as $n \to \infty$. Part (i) then shows that in the worst case P_{DA} reduces the entropy by a factor of $1/n$ in one iteration, suggesting that $\mathcal{O}(n)$ iterations may be needed to reach stationarity; this is confirmed in part (ii), which provides matching lower and upper bounds on mixing times in χ^2 divergence.

In order to solve this issue, we consider a simple modification of Algorithm 1, where an additional Metropolis update of β_1 from $\pi(\beta_1 \mid \beta_{-1})$ is included before updating z: see

Algorithm 3 for the pseudocode. The resulting algorithm is still $\pi(z, \beta)$-invariant (as, for example, it can be interpreted as an instance of partially-collapsed Gibbs sampler (Van Dyk and Park, 2008)). Note that the additional update of β_1 is invariant with respect to $\pi(\beta_1 \mid \beta_{-1})$, which we expect to have $\mathcal{O}(1)$ variance in the imbalanced case, instead of $\pi(\beta_1 \mid z)$, which has always $\mathcal{O}(n^{-1})$ variance. This modification (which recalls the one proposed in Johndrow et al. (2019) for $p = 1$) is shown empirically to mix fast in both balanced and imbalanced settings in Section 5. Alternative solutions have been explored in the literature, such as interweaving strategies (Yu and Meng, 2011) and parameter expansions (Zens et al., 2024), which we also expect to be effective in solving the same issue.

Algorithm 3 (Modified data augmentation Gibbs sampler $P_{\mathrm{DA,mod}}$)

Initialize $(z^{(0)}, \beta^{(0)})$.
for $t \geq 1$ **do**
  Set $\tilde{\beta} = \beta^{(t-1)}$.
  Sample $\beta_1 \sim N(\tilde{\beta}_1, \sigma^2)$ and set $\tilde{\beta} = (\beta_1, \tilde{\beta}_{-1})$ w.p. $\min\left\{1, \pi(\beta_1, \tilde{\beta}_{-1})/\pi(\tilde{\beta})\right\}$.
  Sample $z_i^{(t)} \sim \pi(z_i \mid \tilde{\beta}) \propto N(z_i \mid x_i^T \tilde{\beta}, 1)\mathbb{1}(y_i = g(z_i))$, for $i = 1, \dots, n$.
  Sample $\beta^{(t)} \sim \pi(\beta \mid z^{(t)})$, with $\pi(\beta \mid z)$ as in (5).
end for

Let us stress again that the results in Theorem 1 are worst-case with respect to y, and may substantially differ from the average case: for example, we expect that if the dataset is balanced (i.e. if $n^{-1}\sum_{i=1}^n y_i$ is far from 0 and 1) then both P_{DA} and P_{CG} converge fast in the setting of Proposition 8: indeed also $\mathrm{Var}_\pi(\beta_1) = \mathcal{O}(n^{-1})$ is expected in that case. This is coherent with the findings in Qin and Hobert (2019), which show that if p is fixed and $n \to \infty$ with data generated according to model (3) the mixing times remain bounded with n, and with the simulations in Section 5.

4. Computational cost and comparisons

We now complement the above mixing time bounds with a discussion on the overall computational cost of P_{DA} and P_{CG}, and a brief comparison with gradient-based samplers.

4.1 Cost per iteration

Running either P_{DA} or P_{CG}^n requires $\mathcal{O}(np\min\{n, p\})$ pre-computation cost[1] to compute and factorize the covariance matrix of β, which needs to be done once, and $\mathcal{O}(np)$ cost per iteration. When $p > n$ the cost per iteration can be reduced to $\mathcal{O}(n^2)$ in some cases, see Section A of the Supplementary Material for more details on this and the actual implementation.

1. Note that the pre-computation cost refers to a *single* matrix factorization or inversion which, while being nominally of order $\mathcal{O}(np\min\{n, p\})$, is usually not the dominant cost in practice. Moreover, in settings where this becomes the dominant cost, there is a large body of well-established tools that could be used to reduce this cost at the price of some small approximation error, see e.g. Pandolfi et al. (2024) for examples of using conjugate gradient solvers to avoid full matrix inversions in Gibbs Samplers with Gaussian conditionals. To make these arguments complete, one should then quantify how the approximation error transfer into the posterior one, which we leave to future work.

4.2 Comparison with gradient-based schemes

An alternative to P_{DA} or P_{CG} is to target directly $\pi(\beta)$ in (2) with a gradient-based MCMC algorithm, such as Langevin or Hamiltonian Monte Carlo, without relying on the data augmentation structure in (4). Indeed $\pi(\beta)$ is log-concave and a large literature is available on the resulting performances of gradient-based samplers in this setting (see e.g. Chewi (2023) for an overview). Computing the gradient of $\log \pi(\beta)$ requires $\mathcal{O}(np)$ cost, which matches the cost per iteration of P_{DA} and P_{CG}^n. To the best of our knowledge, currently available upper bounds on the mixing times of gradient-based MCMC schemes which apply to $\pi(\beta)$ are of the form $\mathcal{O}(\kappa^a p^b)$, where κ is the condition number of $\pi(\beta)$ and both $a \geq 1$ and $b > 0$ depend on the specific algorithm. For example, Wu et al. (2022); Altschuler and Chewi (2024) proved that the mixing time of the Metropolis-Adjusted Langevin Algorithm (MALA), possibly after an algorithmic warm start, is of order $\mathcal{O}(\kappa p^{1/2})$. Proposition 19 in the Supplementary Material shows that, after pre-conditioning with the prior variance Q_0^{-1}, the condition number of $\pi(\beta)$ satisfies $\kappa \leq 1 + \lambda_{max}(XQ_0^{-1}X^T)$. This implies an upper bound of order $\mathcal{O}\big(p^{1/2}\big(1+\lambda_{max}(XQ_0^{-1}X^T)\big)\big)$ for the mixing times for MALA, which is strictly worse than the upper bounds for P_{DA} and P_{CG} in Theorem 1 by a factor of $p^{1/2}$. The latter can be interpreted as the additional cost due to the discretization of the Langevin diffusion, see e.g. Ascolani et al. (2026, Section 4.2) for more discussion on this.

5. Simulations

5.1 Practical considerations: coupling-based upper bounds

In order to empirically assess the above bounds, we rely on recent couplings methodologies (Jacob et al., 2020; Biswas et al., 2019), which allow to generate numerical upper bounds to the total variation (TV) distance $\big\|\mu P^t - \pi\big\|_{TV} = \sup_A |\mu P^t(A) - \pi(A)|$. In particular, we consider the methodology introduced in Biswas et al. (2019), which we briefly describe. Consider a bivariate Markov chain $(X_1^{(t)}, X_2^{(t)})_t$ with operator $K((x_1, x_2), \cdot)$ such that marginally $(X_i^{(t)})_t$ is a Markov chain with kernel P for $i = 1, 2$. The kernel K is called a coupling and it is usually chosen so that the meeting time $\tau^{(L)} = \inf\{t > L \mid X_1^{(t)} = X_2^{(t-L)}\}$ is almost surely finite and $X_1^{(t)} = X_2^{(t-L)}$ for every $t > \tau^{(L)}$, where $L > 0$ is a suitable lag. The pseudocode to sample a realization of $\tau^{(L)}$ (which corresponds to Algorithm 1 in Biswas et al. (2019)), is given in Algorithm 4.

Algorithm 4 (Sampling meeting times $\tau^{(L)}$)

Initialize $X_2^{(0)} \sim \mu$, $X_1^{(0)} \sim \mu$ and $X_1^{(t)} \mid X_1^{(t-1)} \sim P(X_1^{(t-1)}, \cdot)$ for $t = 1, \dots, L$.
for $t > L$ **do**
  Sample $\left(X_1^{(t)}, X_2^{(t-L)}\right) \sim K\left((X_1^{(t-1)}, X_2^{(t-L-1)}), \cdot\right)$
  If $X_1^{(t)} = X_2^{(t-L)}$, return $\tau^{(L)} = t$ and exit the for loop.
end for

Theorem 1 in Biswas et al. (2019) shows that $\left\|\mu P^t - \pi\right\|_{TV} \leq \bar{d}(t)$ with

$$\bar{d}(t) = \mathbb{E}\left[\max\left\{0, \left\lceil \frac{\tau^{(L)} - L - t}{L} \right\rceil\right\}\right], \tag{16}$$

for every t, where the bound is tighter as L increases. Thus, we can use Algorithm 4 to obtain N independent realizations of $\tau^{(L)}$ and approximate the right hand side of (16) with their empirical average. As regards the choice of K, we consider the two-step coupling described in Ceriani et al. (2026) (see also references therein): when the two chains are far (i.e. $d(X_t^{(1)}, X_{t-L}^{(2)}) > \epsilon$, for some suitable distance function d) then a contractive coupling is used in order to make the chains closer, otherwise a maximal coupling (maximizing the probability of the chains being exactly equal in one step) is employed. Section C in the Supplementary Material provides more details and a pseudocode for the couplings we employ. The results is a Monte Carlo estimate of $\bar{d}(t)$, which we can either plot as a function of t to monitor convergence (as in Figure 1) or use to upper bound the TV-mixing times, $\tau_{\text{mix}}^{TV}(\epsilon, \mu, P) = \inf\{t \geq 1 : \left\|\mu P^t - \pi\right\|_{TV} \leq \epsilon\}$, as

$$\tau_{\text{mix}}^{TV}(\epsilon, \mu, P) \leq \inf\left\{t \geq 1 : \bar{d}(t) \leq \epsilon\right\}, \tag{17}$$

which follows from $\left\|\mu P^t - \pi\right\|_{TV} \leq \bar{d}(t)$.

Recall that, while our results in Theorem 1 bound mixing times in KL, by the Pinsker inequality they also provide bounds to mixing times in TV that are equivalent up to small multiplicative constants (see e.g. Remark 3.6 in Ascolani et al. (2026)). In this sense, looking at TV-mixing times can also be seen as a way to validate the tightness of the bounds in Theorem 1.

5.2 Simulation studies for various priors

In Tables 1 and 2 we report the upper bounds to $\tau_{\text{mix}}^{TV}(\epsilon, \mu, P)$ based on (17), with μ as in Section 8 below. The rows refer to different choices of the design matrix X and prior precision Q_0, while different columns feature distinct combinations of n and p. Table 1 considers the imbalanced case with $y_i = 1$, while for Table 2 the responses are generated from the model itself, as defined in (1).

The first two rows refer to g priors, with different values of the parameter g which measures the amount of prior information. Coherently with Corollary 5, the estimates of the mixing times are always small, regardless of n, p and the data generation mechanism: interestingly, there seems to be very little variation for different choices of n and p with the same ratio n/p. Moreover, as expected, the mixing times increase with g (i.e. passing from the first to the second row) in all the scenarios.

Third and fourth rows consider isotropic priors with random X and no intercept, i.e. Assumption A. Also here the estimates are quite small, with an increasing trend in r as suggested by Corollary 6: the latter phenomenon is more apparent with data generated from the model and c is large. Also, the mixing times increases with c, going from the third to the fourth row. Such increase is negligible for P_{CG} when $r < 1$, which is coherent with the bound in Corollary 6.

The last row considers the case with intercept, as in Assumption B. Here the two tables yield very different behaviour: with imbalanced data (Table 1) the mixing times grow

Imbalanced data: $y_i = 1$ for all i	Method	n/p=0.2			n/p=1.25			n/p=3		
		p=50	p=100	p=200	p=50	p=100	p=200	p=50	p=100	p=200
g prior	P_{DA}	11	11	11	6	7	7	6	6	6
$(g = 1, c = 0.001)$	P^n_{CG}	8	10	11	20	23	24	24	25	27
g prior	P_{DA}	57	62	65	38	40	43	27	29	29
$(g = 10, c = 0.001)$	P^n_{CG}	11	13	15	34	36	38	46	48	49
Assumption A	P_{DA}	6	7	7	7	8	8	9	9	9
$(c = 1)$	P^n_{CG}	14	16	18	22	24	26	26	27	30
Assumption A	P_{DA}	38	43	44	39	44	54	33	24	20
$(c = 10)$	P^n_{CG}	16	19	22	36	38	45	52	42	39
Assumption B	P_{DA}	21	35	56	81	143	247	160	302	591
$(c = 1)$	P^n_{CG}	26	40	62	102	169	301	244	416	724

Table 1: Upper bounds on the mixing times $\tau^{TV}_{\mathrm{mix}}(\epsilon, \mu, P_{\mathrm{DA}})$ and $\tau^{TV}_{\mathrm{mix}}(\epsilon, \mu, P^n_{\mathrm{CG}})$, for $\epsilon = 0.1$ and $\mu(\mathrm{d}z, \mathrm{d}\beta) = N(\mathrm{d}\beta \mid 0, Q_0^{-1})\pi(\mathrm{d}z \mid \beta)$, obtained from (17), taking $L = 200$ and estimating $\bar{d}(t)$ with $N = 500$ independent simulations of $\tau^{(L)}$. X is generated either as in Assumption B (rows 1, 2 and 5) or as in Assumption A (rows 3 and 4), with $F = N(0, 1)$. In all cases, $y_i = 1$ for $i \in \{1, \ldots, n\}$.

Well-specified data: $y_i \sim \mathrm{Bern}(\Phi(x_i^T\beta))$	Method	n/p=0.2			n/p=1.25			n/p=3		
		p=50	p=100	p=200	p=50	p=100	p=200	p=50	p=100	p=200
g prior	P_{DA}	11	11	11	6	7	7	5	6	6
$(g = 1, c = 0.001)$	P_{CG}	8	10	11	20	22	24	23	24	26
g prior	P_{DA}	58	63	64	40	41	44	25	28	30
$(g = 10, c = 0.001)$	P_{CG}	11	13	15	34	36	37	43	46	49
Assumption A	P_{DA}	6	7	7	8	9	9	11	11	11
$(c = 1)$	P_{CG}	13	16	18	22	25	26	28	29	31
Assumption A	P_{DA}	38	43	47	58	71	69	106	104	90
$(c = 10)$	P_{CG}	16	19	21	48	55	54	130	133	118
Assumption B	P_{DA}	21	9	10	30	10	34	27	13	17
$(c = 1)$	P^n_{CG}	26	18	20	45	25	52	49	31	36

Table 2: Same as Table 1 with data generated from the correct model, i.e. $y_i \sim \mathrm{Bern}(\Phi(x_i^T\beta))$ for all $i \in \{1, \ldots, n\}$, with $\beta \sim N(0, Q_0^{-1})$.

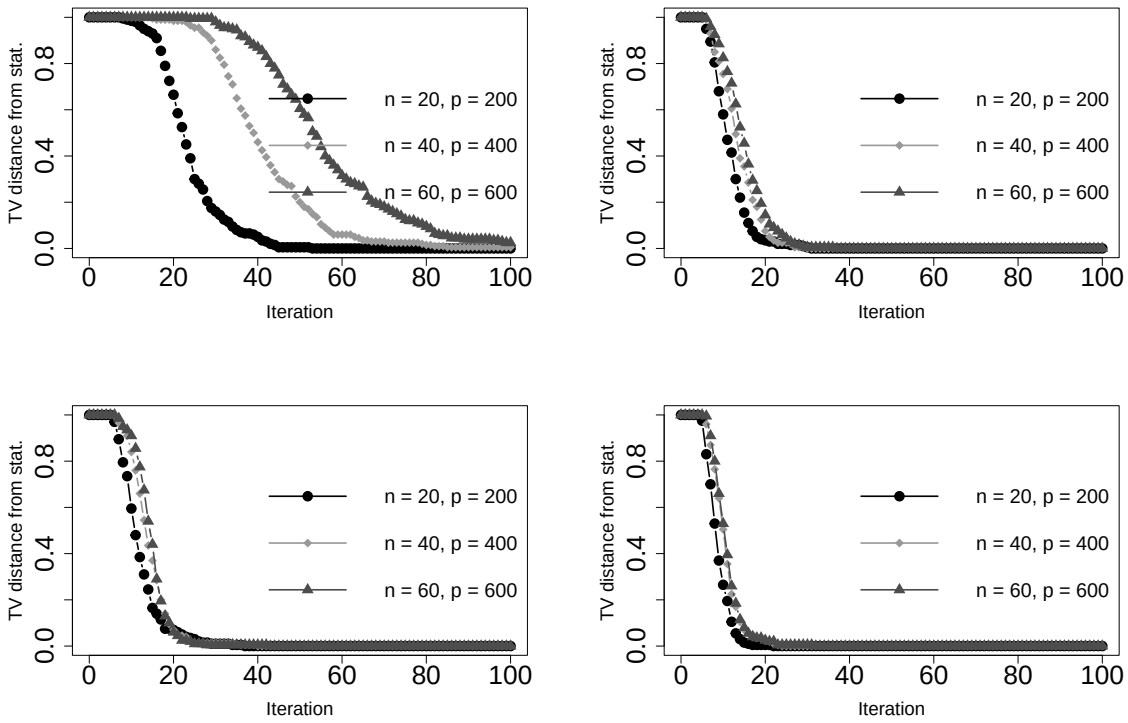

Figure 1: Upper bounds on $\|\mu P_{\mathrm{DA}}^t - \pi\|_{TV}$ (left column) and $\left\|\mu P_{\mathrm{DA,mod}}^t - \pi\right\|_{TV}$ (right column) as a function of t, with $P_{\mathrm{DA,mod}}$ defined in Algorithm 3, $\mu(\mathrm{d}z, \mathrm{d}\beta) = N(\mathrm{d}\beta \mid 0, I_p)\pi(\mathrm{d}z \mid \beta)$, $Q_0^{-1} = I_p$ and X generated according to Assumption B with $F = N(0,1)$. Bounds are obtained from (16), taking $L = 500$ and estimating $\bar{d}(t)$ with $N = 500$ independent simulations of $\tau^{(L)}$. Observations are generated as $y_i = 1$ (top row) and according to model (3) (bottom row).

quickly with n regardless of p, while for random data (Table 2) they do not. This empirically confirms Corollary 7, which suggests that mixing times increases with n in the worst case. The behaviour for random data was also expected given previous findings (Qin and Hobert, 2019). To confirm that the problem is given only by the intercept, as suggested by Proposition 8, Figure 1 reports the Total Variation bounds for P_{DA} (first column) and $P_{\mathrm{DA,mod}}$ (second column) defined in Algorithm 3. When the data are imbalanced (first row), the former quickly deteriorates with n while the latter remain unaffected. With random data (second row) no notable effect is visible for both algorithms.

6. Discussion

6.1 Some practical takeaways (and a computationally convenient prior)

The results of this paper provide guidance on which choices of Q_0 are expected to lead to fast mixing of P_{DA} and P_{CG}^n, or lack thereof. In particular, Corollary 6 implies that, under Assumption A, choosing $c = b/(1+r)$, with $r = n/p$ and $b > 0$ fixed, leads to mixing times that are bounded with respect to both n and p, as formalized below.

Corollary 10 *Under Assumption A and $c = b/(1+r)$ with $b > 0$, we have $\tau_{\mathrm{mix}}(\epsilon, \mu, P_{\mathrm{DA}}) \leq (2 + 2b) \log(\mathrm{KL}(\mu, \pi)/\epsilon)$ and $\tau_{\mathrm{mix}}(\epsilon, \mu_1, P_{\mathrm{CG}}^n) \leq (1 + 2b) \log(\mathrm{KL}(\mu, \pi)/\epsilon)$ almost surely as $n \to \infty$ and $n/p \to r \in (0, \infty)$.*

Proof Follows from Corollary 6 and $c(1+\sqrt{r})^2 + \delta = b(1+\sqrt{r})^2/(1+r) + \delta < 2b$ for small enough $\delta > 0$, which follows from $(1+\sqrt{r})^2 < 2(1+r)$ for $r > 0$. ■

When covariates are standardized to unit variance, the choice $c = b/(1+r)$ corresponds to setting

$$Q_0^{-1} = \frac{c}{p} I_p = \frac{b}{p+n} I_p \,. \tag{18}$$

We can interpret (18) as follows: while rescaling Q_0^{-1} by p^{-1} is natural for statistical reasons (see e.g. discussion and references after Assumption A), rescaling it also by n^{-1} is computationally convenient since it guarantees fast convergence of P_{DA} and P_{CG}^n. When $r = n/p$ is small, the latter modification (18) is equivalent to the standard p^{-1} scaling of Q_0^{-1}, while when $r \gg 1$ it increases the amount of shrinkage or informativity of the prior (roughly speaking keeping it as a constant, even if possibly small, fraction of the data informativity). Exploring the statistical properties of such a choice is beyond the scope of this work, and we leave it to future research.

Combined with the results of Section 3, (18) leads to the following recipe:

(i) Standardize covariates so that $\sum_{i=1}^n X_{ij} = 0$ and $n^{-1} \sum_{i=1}^n X_{ij}^2 = 1$ for all j (excluding the intercept)

(ii) Set $Q_0^{-1} = \frac{b}{p+n} I_n$ for some fixed $b > 0$, e.g. $b = 10$

(iii) If X contains an intercept, sample from $\pi(\beta)$ using Algorithm 3, otherwise using Algorithm 1.

Our results suggest that, if the design matrix X is not too far from a random one as in Assumption A or B, the above recipe leads to mixing times that remain bounded (and fairly small) for all ranges of n and p, thus resulting in computational robustness and efficiency of P_{DA} and P_{CG}.

6.2 Open questions

We now briefly discuss some questions and open problems arising from the above exploration:

1. Even if the above findings show that the mixing times of P_{DA} and P_{CG} remain bounded as $n, p \to \infty$ in various settings, they also identify situations where this does not happen. For example, the upper bound in Corollary 6 suggests that the mixing times may increase as n/p increases, when $Q_0 = cI_p$ with fixed c and X contains no intercept, which is what we observe in the simulations. Similarly, Proposition 8, and the corresponding empirical study, shows that the mixing time of P_{DA} increases linearly with n in the case of fully imbalanced data and presence of the intercept (see also Johndrow et al. (2019)). While these issues can be solved as discussed in Section 6.1, this requires adapting the prior to n. It is thus natural to wonder whether it is possible to find a π-invariant Markov operator P whose mixing times are provably bounded uniformly over y, n and p under Assumptions A or B with fixed c. For example, one might look for P such that

$$\inf_y \lim_{n\to\infty} \rho_{EC}(P, \pi) > 0$$

almost surely, both when p is fixed and when it grows with n, where ρ_{EC} is the entropy contraction coefficient defined in (20). For example, a good candidate for P is given by the interweaving strategy proposed in Yu and Meng (2011), which alternates a centered and non-centered step. However our proof strategy is not applicable any more, since the results of Ascolani et al. (2026) do not apply for the non-centered step of the algorithm.

2. The results of Theorem 1 hold uniformly over y, which means that they are worst-case with respect to y. As we highlighted in the simulation studies of Section 5 the latter can significantly differ from the average case, which can be a reasonable assumption in many scenarios: it would be interesting to develop upper bounds that capture the dependence on y and, e.g., differ for well-specified data as opposed to worst-case ones.

3. Finally, another open question is to analyze mixing times with other priors on β. In this paper we focused on the normal distribution (which corresponds to a ridge penalization), but in high-dimensional scenarios a sparsity inducing prior, like the spike and slab (George and McCulloch, 1993) or Horseshoe (Carvalho et al., 2009), might be preferred. While it would be of great interest to extend our results in this direction, our proof technique heavily relies on Gaussianity (and more generally log-concavity), see Section 2 for more details. Extensions to non log-concave priors will likely require different and novel mathematical tools.

7. General theory about entropy contraction of two-block Gibbs Samplers

In this section we provide some general results on the entropy contraction coefficients of two-block deterministic-scan Gibbs Samplers (aka Data Augmentation kernels), of which the kernel P_{DA} defined in (6) is a special case. Since the results of this section apply more generally than model (4) and can be of independent interest, we use a general notation defined as follows.

7.1 Data augmentation and marginal kernels

Let $\pi \in \mathcal{P}(\mathcal{X})$ with $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. We denote by $\pi_i \in \mathcal{P}(\mathcal{X}_i)$ its i-th marginal and by $\Pi_{i\to j}$ the conditional distributions of x_j given x_i under π, i.e. $\Pi_{i\to j}$ is a Markov kernel from $\mathcal{X}_i$ to $\mathcal{X}_j$ defined as $\Pi_{i\to j}(x_i, A_j) = \mathbb{P}_{(X_1,X_2)\sim\pi}(X_j \in A_j | X_i = x_i)$ for every $x_i \in \mathcal{X}_i$ and $A_j \subseteq \mathcal{X}_j$. The two-block deterministic-scan Gibbs Sampler on π is a Markov transition kernel on $\mathcal{X}$ defined as $P_{\mathrm{DA}} = P_2 P_1$, where

$$P_1(x, \mathrm{d}x') = \delta_{x_2}(\mathrm{d}x'_2)\Pi_{2\to 1}(x_2, \mathrm{d}x_1), \text{ and } P_2(x, \mathrm{d}x') = \delta_{x_1}(\mathrm{d}x'_1)\Pi_{1\to 2}(x_1, \mathrm{d}x_2)\,. \tag{19}$$

If $\{(X_1^{(t)}, X_2^{(t)})\}_t$ is a Markov chain with kernel P_{DA}, then $\{X_2^{(t)}\}_t$ is also a Markov chain and it has kernel $P_{\mathrm{MG}} = \Pi_{1\to 2}\Pi_{2\to 1}$, see e.g. (Roberts and Rosenthal, 2001, Section 3.3). The convergence properties of P_{DA} are closely related to the ones of P_{MG}, as shown in the following proposition.

Proposition 11 *Let $\pi, \mu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ and $t \geq 1$. Then*

$$\mathrm{KL}\left(\mu P_{\mathrm{DA}}^{t+1}, \pi\right) \leq \mathrm{KL}\left(\mu_2 P_{\mathrm{MG}}^{t}, \pi_2\right) \leq \mathrm{KL}\left(\mu P_{\mathrm{DA}}^{t}, \pi\right)\,.$$

Proof By definition of P_{MG}, we have $\int_{\mathcal{X}_1} \mu P_{\mathrm{DA}}^t(\mathrm{d}x_1, A) = \mu_2 P_{\mathrm{MG}}^t(A)$, for every $A \subset \mathcal{X}_2$. By the chain rule for the KL, this implies $\mathrm{KL}\left(\mu_2 P_{\mathrm{MG}}^t, \pi_2\right) \leq \mathrm{KL}\left(\mu P_{\mathrm{DA}}^t, \pi\right)$. To prove the other inequality, consider the Markov kernel from $\mathcal{X}_2$ to $\mathcal{X}$ defined as $\Pi_{2\to(1,2)}(x_2, \mathrm{d}x') = \Pi_{2\to 1}(x_2, \mathrm{d}x'_1)\Pi_{1\to 2}(x'_1, \mathrm{d}x'_2)$, so that $\mu P_{\mathrm{DA}}^{t+1} = \mu_2 P_{\mathrm{MG}}^t \Pi_{2\to(1,2)}$ for all $t \geq 0$. Combining the latter with $\pi_2 \Pi_{2\to(1,2)} = \pi$, and the chain rule we have

$$\mathrm{KL}\left(\mu P_{\mathrm{DA}}^{t+1}, \pi\right) = \mathrm{KL}\left(\mu_2 P_{\mathrm{MG}}^t \Pi_{2\to(1,2)}, \pi_2 \Pi_{2\to(1,2)}\right) \leq \mathrm{KL}\left(\mu_2 P_{\mathrm{MG}}^t, \pi_2\right),$$

as desired. ■

Proposition 11 implies that $\tau_{\mathrm{mix}}(\epsilon, \mu, P_{\mathrm{DA}}) \leq 1 + \tau_{\mathrm{mix}}(\epsilon, \mu_2, P_{\mathrm{MG}})$, which allows us to focus on P_{MG}, which enjoys more analytical tractability.

7.2 Entropy contraction

For a Markov kernel P from $\mathcal{X}$ to $\mathcal{Y}$ and a distribution $\nu \in \mathcal{P}(\mathcal{X})$, define the entropy contraction coefficient of P relative to ν as

$$\rho_{EC}(P, \nu) = \sup_{\mu \in \mathcal{M}} \frac{\mathrm{KL}(\mu P, \nu P)}{\mathrm{KL}(\mu, \nu)}\,. \tag{20}$$

where $\mathcal{M} = \{\mu \in \mathcal{P}(\mathcal{X}) \mid \mathrm{KL}(\mu, \nu) < \infty\}$. The next lemma shows that ρ_{EC} is sub-multiplicative.

Lemma 12 *Let P be a kernel from $\mathcal{X}$ to $\mathcal{Y}$ and Q a kernel from $\mathcal{Y}$ to $\mathcal{Z}$. Then, for every $\nu \in \mathcal{P}(\mathcal{X})$, we have*

$$\rho_{EC}(QP, \nu) \leq \rho_{EC}(P, \nu)\rho_{EC}(Q, \nu P)$$

Proof Fix $\nu \in \mathcal{P}(\mathcal{X})$ and note that $\nu P \in \mathcal{P}(\mathcal{Y})$. For every $\mu \in \mathcal{P}(\mathcal{X})$ we have $\mu(QP) = (\mu P)Q$ and

$$\begin{aligned}\frac{\mathrm{KL}(\mu(QP), \nu(QP))}{\mathrm{KL}(\mu,\nu)} &= \frac{\mathrm{KL}((\mu P)Q, (\nu P)Q)}{\mathrm{KL}(\mu,\nu)} \\ &= \frac{\mathrm{KL}((\mu P)Q, (\nu P)Q)}{\mathrm{KL}(\mu P, \nu P)} \frac{\mathrm{KL}(\mu P, \nu P)}{\mathrm{KL}(\mu,\nu)} \leq \rho_{EC}(Q, \nu P)\rho_{EC}(P,\nu).\end{aligned}$$

The result follows since μ is arbitrary. ■

Interestingly, the so-called approximate tensorization of the entropy for π, i.e. the inequality in (21) below, controls both $\rho_{EC}(\Pi_{2\to 1})$ and $\rho_{EC}(\Pi_{1\to 2})$, as shown below.

Theorem 13 *Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and $\pi \in \mathcal{P}(\mathcal{X})$. If*

$$\frac{\mathrm{KL}(\mu_1, \pi_1) + \mathrm{KL}(\mu_2, \pi_2)}{2} \leq \left(1 - \frac{1}{2\kappa^*}\right) \mathrm{KL}(\mu, \pi) \tag{21}$$

for all $\mu \in \mathcal{P}(\mathcal{X})$, then

$$\max\{\rho_{EC}(\Pi_{2\to 1}, \pi_2), \rho_{EC}(\Pi_{1\to 2}, \pi_1)\} \leq \left(1 - \frac{1}{\kappa^*}\right) \tag{22}$$

and $\rho_{EC}(P_{\mathrm{MG}}, \pi_2) \leq (1 - 1/\kappa^*)^2$.

Proof Let $\mu_1 \in \mathcal{P}(\mathcal{X}_1)$. Applying (21) to the measure $\mu(\mathrm{d}x) = \mu_1(\mathrm{d}x_1)\Pi_{1\to 2}(x_1, \mathrm{d}x_2) \in \mathcal{P}(\mathcal{X})$ we obtain

$$\frac{\mathrm{KL}\left(\mu_1, \pi_1\right) + \mathrm{KL}\left(\mu_1\Pi_{1\to 2}, \pi_2\right)}{2} \leq \left(1 - \frac{1}{2\kappa^*}\right) \mathrm{KL}(\mu, \pi)\,. \tag{23}$$

Since $\mathrm{KL}(\mu, \pi) = \mathrm{KL}(\mu_1, \pi_1)$, which follows by definition of μ and the chain rule for the KL, (23) can be written as

$$\frac{\mathrm{KL}\left(\mu_1\Pi_{1\to 2}, \pi_2\right)}{2} \leq \left(1 - \frac{1}{2\kappa^*} - \frac{1}{2}\right) \mathrm{KL}(\mu_1, \pi_1) = \frac{1}{2}\left(1 - \frac{1}{\kappa^*}\right) \mathrm{KL}(\mu_1, \pi_1),$$

which, together with $\pi_1 = \pi_2\Pi_{2\to 1}$, implies $\rho_{EC}(\Pi_{2\to 1}, \pi_2) \leq (1 - 1/\kappa^*)$. Also $\rho_{EC}(\Pi_{1\to 2}, \pi_1) \leq (1 - 1/\kappa^*)$, and thus (22), follows by symmetry. Finally $\rho_{EC}(P_{\mathrm{MG}}, \pi_2) \leq (1 - 1/\kappa^*)^2$ follows from (22) and $P_{\mathrm{MG}} = \Pi_{1\to 2}\Pi_{2\to 1}$ by Lemma 12. ■

Remark 14 *Recall that $\Pi_{1\to 2}$ and $\Pi_{2\to 1}$ are adjoint of each other with respect to the inner products of $L^2(\pi_1)$ and $L^2(\pi_2)$, and thus have the same operator norm in L^2, i.e. $\|\Pi_{1\to 2}\|_{L^2} = \|\Pi_{2\to 1}\|_{L^2}$. Nonetheless, $\rho_{EC}(\Pi_{1\to 2}) \neq \rho_{EC}(\Pi_{2\to 1})$ in general and the ratio of the two can be arbitrarily large (see e.g. Example 16 in Caputo et al. (2024)). More generally, the connection between the entropy contraction coefficients of $\Pi_{1\to 2}$, $\Pi_{2\to 1}$, P_{MG} and P_{DA} is more subtle than for their L^2-norms (or equivalently spectral Gaps), see Caputo et al. (2024) for a detailed review, as well as Andrieu (2016); Qin and Jones (2022); Chlebicka et al. (2025) for results in the L^2 context.*

In Section D.1 of the Supplementary Material we combine the results of this section with the ones in Ascolani et al. (2026) to prove (11). More generally, the proof of Theorem 1 relies on bounds of the form

$$\tau_{\text{mix}}(\epsilon, \mu, P) \leq \frac{1}{1 - \rho_{EC}(P, \pi)} \log\left(\frac{\text{KL}(\mu, \pi)}{\epsilon}\right), \tag{24}$$

which directly follows from $\text{KL}(\mu P^t, \pi) \leq \rho_{EC}(P, \pi)^t \text{KL}(\mu, \pi)$ and $\log(1 - 1/c) \leq -1/c$ for $c > 1$.

8. Starting distribution

To be fully informative, the results of Theorem 1 require to find a starting distribution $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^p)$ such that $\log(\text{KL}(\mu, \pi))$ can be suitably controlled, which is what we do on this section. Since sampling from $\pi(\mathrm{d}z \mid \beta)$ is feasible, we take starting distributions $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^n)$ of the form

$$\mu(\mathrm{d}z, \mathrm{d}\beta) = \pi(\mathrm{d}z \mid \beta)\mu_2(\mathrm{d}\beta). \tag{25}$$

for some $\mu_2 \in \mathcal{P}(R^n)$, so that $\text{KL}(\mu, \pi) = \text{KL}(\mu_2, \pi_2)$. In (Lee and Zhang, 2024, Sec.3.2.1) it is proven that a Gaussian distribution centered around the mode of $\pi(\beta)$, and with a suitable variance, is a feasible starting distribution with good control in KL. Here instead we assume to start from the prior, i.e. take $\mu_2(\beta) = N(\beta \mid m, Q_0^{-1})$. This is arguably an easier choice, which does not require additional computations to find the mode, nor knowledge of smoothness constant to tune variances. The next proposition shows that such starting distribution is also close enough in KL to lead to good mixing times bounds. We consider the case of prior with zero mean for simplicity, even if the result could be generalized at the price of slightly more complicated bounds.

Proposition 15 *Let $\mu \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^n)$ be as in (25) with $\mu_2(\beta) = N(\beta \mid m, Q_0^{-1})$. Let $m = (0, \dots, 0)^T \in \mathbb{R}^p$. Then, for every y, we have*

$$\log(\text{KL}(\mu, \pi)) \leq \log\left(2n + n \log\left(2(1 + n\lambda_{max}(X^T Q_0^{-1} X)\right)\right).$$

Acknowledgments

GZ acknowledges support from the European Research Council (ERC), through the Starting Grant (StG) 'PrSc-HDBayLe', project number 101076564.

References

Alan Agresti. *Analysis of ordinal categorical data*. John Wiley & Sons, 2010.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Emanuele Aliverti. Approximate bayesian inference for cumulative probit regression models. *arXiv preprint arXiv:2511.06967*, 2025.

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024.

Yali Amit. Convergence properties of the Gibbs sampler for perturbations of Gaussians. *The Annals of Statistics*, 24(1):122–140, 1996.

Niccolò Anceschi, Augusto Fasano, Daniele Durante, and Giacomo Zanella. Bayesian conjugacy in probit, tobit, multinomial probit and extensions: A review and new results. *Journal of the American Statistical Association*, 118(542):1451–1469, 2023.

Christophe Andrieu. On random-and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016.

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q Wang. Explicit convergence bounds for Metropolis Markov chains: Isoperimetry, spectral gaps and profiles. *The Annals of Applied Probability*, 34(4):4022–4071, 2024.

Filippo Ascolani, Hugo Lavenant, and Giacomo Zanella. Entropy contraction of the Gibbs sampler under log-concavity. *arXiv preprint arXiv:2410.00858*, 2026.

Zhi-Dong Bai, Yong-Qua Yin, et al. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab*, 21(3):1275–1294, 1993.

Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of Markov chains with L-lag couplings. *Advances in Neural Information Processing Systems*, 32, 2019.

Zdravko I Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):125–148, 2017.

Pietro Caputo. Lecture notes on entropy and Markov chains. *Preprint, available from: http://www.mat.uniroma3.it/users/caputo/entropy.pdf*, 2023.

Pietro Caputo, Zongchen Chen, Yuzhou Gu, and Yury Polyanskiy. Entropy contractions in markov chains: Half-step, full-step and continuous-time. *arXiv preprint arXiv:2409.07689*, 2024.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pages 73–80. PMLR, 2009.

Paolo Maria Ceriani, Andrea Pandolfi, and Giacomo Zanella. Linear-cost unbiased posterior estimates for crossed effects and matrix factorization models via couplings. *arXiv preprint arXiv:2410.08939*, 2026.

Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho.github.io*, 2023.

Siddhartha Chib. Bayes inference in the tobit censored regression model. *Journal of Econometrics*, 51(1-2):79–99, 1992.

Iwona Chlebicka, Krzysztof Łatuszyński, and Błażej Miasojedow. Solidarity of Gibbs Samplers: the spectral gap. *The Annals of Applied Probability*, 35(1):142–157, 2025.

Nicolas Chopin and James Ridgway. Leave pima indians alone: Binary regression as a benchmark for bayesian computation. *Statistical Science*, 32(1):64–87, 2017.

Daniele Durante. Conjugate bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779, 2019.

Augusto Fasano and Daniele Durante. A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Journal of Machine Learning Research*, 23(30):1–26, 2022.

Augusto Fasano, Daniele Durante, and Giacomo Zanella. Scalable and accurate variational bayes for high-dimensional binary regression models. *Biometrika*, 109(4):901–919, 2022.

Geir-Arne Fuglstad, Ingeborg Gullikstad Hem, Alexander Knight, Håvard Rue, and Andrea Riebler. Intuitive joint priors for variance parameters. *Bayesian Analysis*, 15:1109—-1137, 2020.

Edward I George and Robert E McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Max Goplerud, Omiros Papaspiliopoulos, and Giacomo Zanella. Partially factorized variational inference for high-dimensional mixed models. *Biometrika*, page asae067, 2024.

Leonhard Held and Chris C Holmes. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145–168, 2006.

PE Jacob, J O'Leary, and YF Atchadé. Unbiased Markov chain Monte Carlo with couplings (with discussion). *JR Statist. Soc. Ser. B*, 82:543–600, 2020.

James E Johndrow, Aaron Smith, Natesh Pillai, and David B Dunson. MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, 2019.

Holden Lee and Kexin Zhang. Fast mixing of data augmentation algorithms: Bayesian probit, logit, and lasso regression. *arXiv preprint arXiv:2412.07999*, 2024.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

L Mailhot. Some properties of truncated distributions connected with log-concavity of distribution functions. *Applicationes Mathematicae*, 20:531–542, 1988.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Chapman & Hall, 1989.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Andrea Pandolfi, Omiros Papaspiliopoulos, and Giacomo Zanella. Conjugate gradient methods for high-dimensional GLMMs. *arXiv preprint arXiv:2411.04729*, 2024.

Qian Qin and James P Hobert. Convergence complexity analysis of albert and chib's algorithm for bayesian probit regression. *The Annals of Statistics*, 47(4):2320–2347, 2019.

Qian Qin and James P Hobert. Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *The Annals of Applied Probability*, 32(1):124–166, 2022.

Qian Qin and Galin L Jones. Convergence rates of two-component MCMC samplers. *Bernoulli*, 28(2):859–885, 2022.

Gareth O Roberts and Jeffrey S Rosenthal. Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28(3):489–504, 2001.

Gareth O Roberts and Sujit K Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(2):291–317, 1997.

Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

Vivekananda Roy and James P Hobert. Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):607–623, 2007.

Sylvain Sardy. On the practice of rescaling covariates. *International Statistical Review*, 76 (2):285–297, 2008.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32:1–8, 2017.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

David A Van Dyk and Taeyoung Park. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *Journal of Machine Learning Research*, 23(270):1–63, 2022.

Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.

Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.

Gregor Zens, Sylvia Frühwirth-Schnatter, and Helga Wagner. Ultimate Pólya Gamma Samplers–Efficient MCMC for possibly imbalanced binary and categorical data. *Journal of the American Statistical Association*, 119(548):2548–2559, 2024.

Appendix A. Implementation and cost per iteration

We now discuss the computational cost associated to run a single iteration of P_{DA} and n iterations of P_{CG}, separating the case $n > p$ from the one $p > n$. For ease of notation, we will denote $V = (X^T X + Q_0)^{-1}$.

$n > p$. In the case of P_{DA} in (6), the main cost is associated to the computation of conditional mean and covariance matrix of β in (5). Since $n > p$, then $V = \mathrm{Var}(\beta \mid z, y)$ can be pre-computed with $\mathcal{O}(np^2)$ cost (the matrix multiplication dominates the $\mathcal{O}(p^3)$ cost of inversion). With the same cost, both VX^T and $VQ_0\mu$ can be pre-computed. Then, for every iteration, $\mathbb{E}[\beta \mid z, y] = V(Q_0\mu + X^T z)$ and $\mathbb{E}[z \mid \beta] = X\beta$ can be computed with $\mathcal{O}(np)$ cost. Thus, the overall cost is given by $\mathcal{O}(np^2)$ pre-computation and $\mathcal{O}(np)$ per iteration.

As regards P_{CG}, the conditional distribution of z_i can be written (see e.g. Held and Holmes (2006)) as

$$\pi(z_i \mid z_{-i}) \propto N\left(z_i \mid (1-h_i)^{-1} x_i^T V X^T (z - Q_0 m) - h_i (1-h_i)^{-1} z_i, (1-h_i)^{-1}\right) \mathbb{1}(y_i = g(z_i)),$$

where $h_i = x_i^T V x_i$. Thus, after pre-computing VX^T with $\mathcal{O}(np^2)$ cost, then $x_i^T V$ is given by the i-th column of XV. Then also $x_i^T V x_i$ can be pre-computed for every i, with overall cost $\mathcal{O}(np)$. After updating the i-th component, the matrix $B = VX^T$ can be updated at $\mathcal{O}(p)$ cost by noticing that

$$B = B_{\mathrm{old}} + S_i (z_i - z_{i,\mathrm{old}}),$$

where S_i is the i-th column of VX^T, while B_{old} and $z_{i,\mathrm{old}}$ refer to the values before updating z_i. Thus, the conditional mean can be obtained at $\mathcal{O}(p)$. This implies that the overall cost is given by $\mathcal{O}(np^2)$ pre-computation and $\mathcal{O}(np)$ for every n iterations. Finally, if needed, a sample of β can be obtained with an additional $\mathcal{O}(np)$ cost.

$p > n$. As regards P_{DA} we can pre-compute V using the Woodbury's identity, which reads

$$V = Q_0^{-1} - Q_0^{-1} X^T (I_n + X Q_0^{-1} X^T)^{-1} X,$$

with a $\mathcal{O}(n^2 p)$ cost (since the $\mathcal{O}(n^3)$ cost of inversion is dominated by the matrix multiplication). Similarly, it is possible to pre-compute

$$\bar{V} = VX^T = Q_0^{-1} X^T - Q_0^{-1} X^T (I_n + X Q_0^{-1} X^T)^{-1} X X^T$$

again at $\mathcal{O}(n^2 p)$ cost. Then, for every iteration it is possible to compute the conditional means $X\beta$ and $\bar{V} z$ at $\mathcal{O}(np)$ cost. This implies that the overall cost is given by $\mathcal{O}(n^2 p)$ pre-computation and $\mathcal{O}(np)$ for every iteration.

As an alternative it is possible to implement instead the Markov chain with operator $\tilde{P}_{\mathrm{DA}}$, which samples from the full conditionals of z and $\tilde{\beta} = X\beta$. The full conditionals are then given by $\tilde{\pi}(z \mid \tilde{\beta}) \propto N(z \mid \tilde{\beta}, I_n) \prod_{i=1}^n \mathbb{1}(y_i = g(z_i))$ and

$$\tilde{\pi}(\tilde{\beta} \mid z) = N\left(\beta \mid X(X^T X + Q_0)^{-1}(Q_0 m + X^T z), X(X^T X + Q_0)^{-1} X^T\right).$$

By construction $\mu P_{\mathrm{DA}}^t \in \mathcal{P}(\mathbb{R}^p \times \mathbb{R}^n)$ and $\mu \tilde{P}_{\mathrm{DA}}^t \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$, with $\int_{\mathbb{R}^p} \mu P_{\mathrm{DA}}^t(A, \mathrm{d}\beta) = \int_{\mathbb{R}^n} \nu \tilde{P}_{\mathrm{DA}}^t(A, \mathrm{d}\tilde{\beta})$ for every $A \subset \mathbb{R}^n$, $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^p)$, and $\nu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ with $\nu_2 = X \circ \mu_2$, i.e. the marginal distribution on z is the same. Moreover they are co-deinitializing in the sense of Roberts and Rosenthal (2001), which implies that the two chains enjoy the same convergence properties. This is formalized in the next lemma.

Lemma 16 *Fix $t \geq 1$ and let $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^p)$ and $\nu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^n)$ with $\nu_2 = X \circ \mu_2$. Then we have that*

$$\mathrm{KL}(\mu P_{\mathrm{DA}}^t, \pi) = \mathrm{KL}(\nu \tilde{P}_{\mathrm{DA}}^t, \tilde{\pi}).$$

Proof Consider $\tau \in \mathcal{P}(\mathbb{R}^n)$ defined as $\tau(A) = \int_{\mathbb{R}^p} \mu P_{\mathrm{DA}}^t(A, \mathrm{d}\beta) = \int_{\mathbb{R}^n} \nu \tilde{P}_{\mathrm{DA}}^t(A, \mathrm{d}\tilde{\beta})$, for every $A \subset \mathbb{R}^n$. By the chain rule for the KL divergence, we have that

$$\mathrm{KL}(\mu P_{\mathrm{DA}}^t, \pi) = \mathrm{KL}(\tau, \pi_1) + \mathbb{E}_{z\sim\tau}\left[\mathrm{KL}(\pi(\mathrm{d}\beta \mid z), \pi(\mathrm{d}\beta \mid z)\right] = \mathrm{KL}(\tau, \pi_1).$$

The same argument holds for $\tilde{P}_{\mathrm{DA}}$. ■

Using again Woodbury's identity, it is possible to compute $W = X(X^TX + Q_0)^{-1}X^T$ with $\mathcal{O}(n^2p)$ operations. Then each iteration only requires to compute Wz, at $\mathcal{O}(n^2)$ cost. This implies that the overall cost is given by $\mathcal{O}(n^2p)$ pre-computation and $\mathcal{O}(n^2)$ for every n iterations. Of course, if needed, a sample of β can be obtained with an additional $\mathcal{O}(np)$ cost.

As regards P_{CG}, the prior precision matrix $Q = (I_n + XQ_0^{-1}X^T)^{-1}$ can be pre-computed at $\mathcal{O}(n^2p)$ cost and the conditional distributions can be rewritten as

$$\pi(z_i \mid z_{-i}, y) \propto N\left(z_i \mid -Q_{ii}^{-1}Q_{i,-i}^T(z_{-i} - (Xm)_{-i}), Q_{ii}^{-1}\right) \mathbb{1}(y_i = g(z_i)),$$

where $Q_{i,-i}$ denotes the i-th row of Q without the i-th entry. Then every iteration can be performed at $\mathcal{O}(n)$ cost. This implies that the overall cost is given by $\mathcal{O}(n^2p)$ pre-computation and $\mathcal{O}(n^2)$ for every n iterations. Finally, if needed, a sample of β can be obtained with an additional $\mathcal{O}(np)$ cost.

Appendix B. Model variations: tobit and cumulative probit regression

As mentioned in Remark 4, our results extend also to other models that can be expressed as partially or fully discretized Gaussian linear regression (Anceschi et al., 2023). Here we consider two examples, which differ in the regression structure and choice of discretization mechanism. The first example is the tobit model for censored data. Here the posterior, after some manipulations, can be traced back to the one of a probit model with modified data and priors, so that the mixing time bounds follow directly from Theorem 1, but with a different prior matrix. The second one is the cumulative probit model for ordinal categorical data, whose posterior does not coincide with the one of probit models, but the same proof strategy of Theorem 1 applies analogously thanks to the convexity of the discretization sets. Other examples where Theorem 1 can be applied either directly or with minor variations include the classical and sequential multinomial probit models, see respectively Sections 2.1-2.2 and Section 2.3 of Fasano and Durante (2022) for a self-contained description of such models. We do not discuss those here for brevity, since the overall message is analogous.

Tobit model The tobit model (Tobin, 1958) is a classical censored regression model where response data are observed only if they exceed a certain threshold, which is often set to 0 for convenience. The resulting Bayesian model admits a natural data augmentation representation, analogous to (3), which reads

$$\begin{aligned} &y_i = z_i \mathbb{1}(z_i > 0) && i = 1, \dots, n, \\ &z|\beta \sim N(X\beta, I_n), \quad \beta \sim N(m, Q_0^{-1}). \end{aligned} \tag{26}$$

Given the definition of y_i, it is convenient to partition the observations $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ into censored ones, which we denote as $\bar{y} = (\bar{y}_1, \dots, \bar{y}_{n_0})^T \in \mathbb{R}^{n_0}$ where $\bar{y}_i = 0$ for all $i \in \{1, \dots, n_0\}$, and non-censored ones, which we denote as $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_{n_1})^T \in \mathbb{R}^{n_1}$ where $\tilde{y}_i > 0$ for all $i \in \{1, \dots, n_1\}$. Here $n_0 + n_1 = n$. Similarly, we partition in an analogous way the latent variables $z = (z_1, \dots, z_n)^T \in \mathbb{R}^n$ into $\bar{z} = (\bar{z}_1, \dots, \bar{z}_{n_0})^T \in \mathbb{R}^{n_0}$ and $\tilde{z} = (\tilde{z}_1, \dots, \tilde{z}_{n_1})^T \in \mathbb{R}^{n_1}$, where $\bar{z}_i \leq 0$ for all $i \in \{1, \dots, n_0\}$ and $\tilde{z}_i > 0$ for all $i \in \{1, \dots, n_0\}$, and the design matrix X into $\bar{X} \in \mathbb{R}^{n_0 \times p}$ and $\tilde{X} \in \mathbb{R}^{n_1 \times p}$. Given the deterministic constraint $\tilde{z} = \tilde{y}$, one needs only to consider the joint posterior of $\bar{z}$ and β, which reads

$$\bar{\pi}(\bar{z}, \beta) \propto N(\beta \mid m, Q_0^{-1}) N(\tilde{y} \mid \tilde{X}\beta, I_{n_1}) N(\bar{z} \mid \bar{X}\beta, I_{n_0}) \prod_{i=1}^{n_0} \mathbb{1}\left(\bar{y}_i = g(\bar{z}_i)\right) \tag{27}$$

$$\propto N\left(\beta \mid m_{new}, Q_{new}^{-1}\right) N(\bar{z} \mid \bar{X}\beta, I_{n_0}) \prod_{i=1}^{n_0} \mathbb{1}\left(\bar{y}_i = g(\bar{z}_i)\right), \tag{28}$$

with $g(\bar{z}_i) = \mathbb{1}(\bar{z}_i > 0)$, $m_{new} = (\tilde{X}^T\tilde{X} + Q_0)^{-1}(Q_0 m + \tilde{X}^T\tilde{y})$ and $Q_{new} = \tilde{X}^T\tilde{X} + Q_0$. Also in the tobit context, posterior samples are often drawn using the corresponding data-augmentation algorithm (Chib, 1992), i.e. the two-block deterministic-scan Gibbs Sampler $\bar{P}_{\mathrm{DA}}$ targeting $\bar{\pi}(\bar{z}, \beta)$ by alternating the update of $\bar{z}$ from $\bar{\pi}(\bar{z} \mid \beta)$ and β from $\bar{\pi}(\beta \mid \bar{z})$. Alternatively, one could consider the collapsed kernel $\bar{P}_{\mathrm{CG}}$, targeting directly the marginal distribution $\bar{\pi}(\bar{z})$. Comparing (28) with (4) shows that $\bar{\pi}$ can be interpreted as the posterior distribution of a probit model with n_0 observations all equal to 0, and a modified prior β equal to $N(m_{new}, Q_{new}^{-1})$. It thus follows from Theorem 1 that the mixing times of $\bar{P}_{\mathrm{DA}}$ and $\bar{P}_{\mathrm{CG}}$ are upper bounded as

$$\tau_{\mathrm{mix}}(\epsilon, \mu, \bar{P}_{\mathrm{DA}}) \leq (2 + \lambda_{max}(\bar{X} Q_{new}^{-1} \bar{X}^T)) \log\left(\frac{\mathrm{KL}(\mu, \pi)}{\epsilon}\right), \tag{29}$$

$$\tau_{\mathrm{mix}}(\epsilon, \mu_1, \bar{P}_{\mathrm{CG}}^{n_0}) \leq \frac{1 + \lambda_{max}(\bar{X} Q_{new}^{-1} \bar{X}^T)}{1 + \lambda_{min}(\bar{X} Q_{new}^{-1} \bar{X}^T)} \log\left(\frac{\mathrm{KL}(\mu, \pi)}{\epsilon}\right). \tag{30}$$

Interestingly, the spectrum of $\bar{X} Q_{new}^{-1} \bar{X}^T$ behaves differently than the one of $X Q_0^{-1} X^T$. For example, under random design assumptions, the quantity $\lambda_{max}(X Q_0^{-1} X^T)$ diverges with the ratio $r \approx n/p$ of number of data points over parameters (see Corollary 6), while for $\lambda_{max}(\bar{X} Q_{new}^{-1} \bar{X}^T)$ to remain bounded it is sufficient that the ratio $\zeta \approx n_0/n_1$ of censored versus uncensored observations does so, as shown in the following corollary.

Corollary 17 *Under Assumption A we have that*

$$\tau_{\mathrm{mix}}(\epsilon, \mu, \bar{P}_{\mathrm{DA}}) \leq (2 + \zeta + \delta) \log\left(\mathrm{KL}(\mu, \pi)/\epsilon\right)$$

and

$$\tau_{\mathrm{mix}}(\epsilon, \mu_1, \bar{P}_{\mathrm{CG}}^{n_0}) \leq (2 + \zeta + \delta) \log\left(\mathrm{KL}(\mu, \pi)/\epsilon\right),$$

almost surely as $n_0, n_1 \to \infty$ with p fixed and $n_0/n_1 \to \zeta \in (0, \infty)$, for any arbitrarily small positive constant $\delta > 0$.

Proof By definition of Q_{new} and standard matrix properties, we have

$$\lambda_{max}(\bar{X}Q_{new}^{-1}\bar{X}^T) = \lambda_{max}(\bar{X}(\tilde{X}^T\tilde{X}+Q_0)^{-1}\bar{X}^T) = \lambda_{max}((\tilde{X}^T\tilde{X}+Q_0)^{-1}\bar{X}^T\bar{X})\,.$$

Since both $\tilde{X}^T\tilde{X}+Q_0$ and $\bar{X}^T\bar{X}$ have fixed dimensionality as $n_0, n_1 \to \infty$ with p fixed and $n_0/n_1 \to \zeta \in (0,\infty)$, by Assumption A, the strong law of large numbers and continuity of the matrix inverse, we have

$$(\tilde{X}^T\tilde{X}+Q_0)^{-1}\bar{X}^T\bar{X} = \frac{n_0}{n_1}\left(\frac{\tilde{X}^T\tilde{X}+Q_0}{n_1}\right)^{-1}\frac{\bar{X}^T\bar{X}}{n_0} \to \zeta I_p \tag{31}$$

almost surely. Thus, also $\lambda_{max}(\bar{X}Q_{new}^{-1}\bar{X}^T)$ converges almost surely to ζ by continuity of the operator λ_{max}. Instead, $\lambda_{min}(\bar{X}Q_{new}^{-1}\bar{X}^T) = 0$ as soon as $n_0 > p$ because $\bar{X} \in \mathbb{R}^{n_0\times p}$ and thus the rank of $\bar{X}Q_{new}^{-1}\bar{X}^T$ is at most p. The desired results then follow by combining $\lambda_{max}(\bar{X}Q_{new}^{-1}\bar{X}^T) \to \gamma$ and $\lambda_{min}(\bar{X}Q_{new}^{-1}\bar{X}^T) \to 0$ with Theorem 1. ■

Cumulative probit model The cumulative probit (or ordered probit) model (McCullagh, 1980) is a popular model for to perform regression with *ordinal* categorical data (Agresti, 2010), such as responses of questionnaires measuring levels of agreement or datasets measuring severity of diseases in a discretized scale. Similarly to (3) and (26), the cumulative probit model assumes that each ordinal observation $y_i \in \{1,\dots,K\}$, where K denotes the number of categories, arises as the discretization of some continuous latent variable z_i. Specifically, the model reads

$$\begin{aligned} &y_i = g_\alpha(z_i) = \sum_{k=1}^{K} k\mathbb{1}(\alpha_{k-1} < z_i \le \alpha_k) \quad i = 1,\dots,n, \\ &z|\beta \sim N(X\beta, I_n), \quad \beta \sim N(m, Q_0^{-1})\,, \end{aligned} \tag{32}$$

with $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_K = \infty$ being a set of real-valued thresholds. Here we assume $(\alpha_k)_{k=1}^K$ to be fixed for simplicity, and refer to Aliverti (2025) for more discussion of how these thresholds are tuned or estimated in practice. The joint posterior density of z and β coincides with (4), but with g replaced by g_α, i.e. it reads

$$\pi_\alpha(z,\beta) \propto N(\beta \mid m, Q_0^{-1})N(z \mid X\beta, I_n)\prod_{i=1}^{n}\mathbb{1}\left(y_i = g_\alpha(z_i)\right)\,. \tag{33}$$

It follows that the conditional density $\pi_\alpha(\beta \mid z)$ coincide with $\pi(\beta \mid z)$ defined in (5), while $\pi_\alpha(z \mid \beta)$ reads

$$\pi_\alpha(z \mid \beta) \propto N(z \mid X\beta, I_n)\prod_{i=1}^{n}\mathbb{1}(\alpha_{y_i-1} < z_i \le \alpha_{y_i})\,,$$

which factorizes as a product of one-dimensional Gaussians truncated on an interval, thus being easy to sample from.

Crucially, the sets $A_{y_i}^{(\alpha)} := \{z_i \in \mathbb{R} : y_i = g_\alpha(z_i)\}$ appearing in (33) are convex, like the analogous sets $A_{y_i} := \{z_i \in \mathbb{R} : \mathbb{1}(y_i = g(z_i))\}$ in (4). Thus, the proof of Theorem 1 in

Section D.1 applies analogously[2], and the bounds in (11) and (12) hold unchanged also in this context. Namely, for every $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^p)$ and $\epsilon > 0$, we have

$$\tau_{\text{mix}}(\epsilon, \mu, \bar{P}^{(\alpha)}_{\text{DA}}) \leq (2 + \lambda_{max}(XQ_0^{-1}X^T)) \log\left(\frac{\text{KL}(\mu, \pi)}{\epsilon}\right),$$
$$\tau_{\text{mix}}(\epsilon, \mu_1, (\bar{P}^{(\alpha)}_{\text{CG}})^n) \leq \frac{1 + \lambda_{max}(XQ_0^{-1}X^T)}{1 + \lambda_{min}(XQ_0^{-1}X^T)} \log\left(\frac{\text{KL}(\mu, \pi)}{\epsilon}\right),$$

where $\bar{P}^{(\alpha)}_{\text{DA}}$ is the two-block deterministic-scan Gibbs Sampler targeting $\pi_\alpha(z, \beta)$ by alternating the update of z from $\pi_\alpha(z \mid \beta)$ and β from $\pi_\alpha(\beta \mid z)$, and $\bar{P}^{(\alpha)}_{\text{CG}}$ is the n-block random-scan Gibbs Sampler defined as in (8) but with π replaced by π_α.

Appendix C. Additional useful results

C.1 Condition number of (2)

Given $\pi(\beta)$ as in (2) with $m = (0, \dots, 0)^T \in \mathbb{R}^p$, we can write

$$U(\beta) = -\log(\pi(\beta)) = \frac{\beta^T Q_0 \beta}{2} + \sum_{i=1}^{n} h\left(\text{sgn}(2y_i - 1)x_i^T\beta\right), \quad h(r) = -\log\left(\Phi(r)\right), \tag{34}$$

where sgn(u) is the sign of $u \in \mathbb{R}$. Let $\tilde{U}$ be the prior-preconditioned version of U, i.e. $\tilde{U}(\theta) = U(Q_0^{-1/2}\theta)$ for $\theta \in \mathbb{R}^p$. We define the condition number of $\tilde{U}$ as

$$\kappa(\tilde{U}) := \frac{\sup_{\theta \in \mathbb{R}^p} \lambda_{max}(\nabla^2 \tilde{U}(\theta))}{\inf_{\theta \in \mathbb{R}^p} \lambda_{min}(\nabla^2 \tilde{U}(\theta))}.$$

We have a preliminary lemma.

Lemma 18 *For every $r \in \mathbb{R}$ we have that $h''(r) \in (0, 1)$.*

Proof It is well-known that $\Phi(r)$ is a strictly log-concave function, i.e. that $h''(r) > 0$. Thus we focus on the upper bound. By simple calculations we get

$$h''(r) = \left(\frac{\phi(r)}{\Phi(r)}\right)^2 + r\frac{\phi(r)}{\Phi(r)}, \tag{35}$$

where $\phi(r) = N(r \mid 0, 1)$. Moreover, if $Z \sim N(0, 1)$, it is easy to show

$$0 \leq \text{Var}(Z \mid Z < r) = 1 - \left(\frac{\phi(r)}{\Phi(r)}\right)^2 - r\frac{\phi(r)}{\Phi(r)} = 1 - h''(r), \tag{36}$$

from which we deduce that $h''(r) < 1$. ■

The next proposition provides an upper bound on $\kappa(\tilde{U})$.

2. Specifically, the only difference relative to the proof in Section D.1 is to replace the approximating functions $U_{i,N}(z_i) = N\,\text{dist}(z_i, A_{y_i})$ with $U^{(\alpha)}_{i,N}(z_i) = N\,\text{dist}(z_i, A^{(\alpha)}_{y_i})$, where $\text{dist}(z, A) = \inf_{z' \in A}|z - z'|$ denotes the distance between a point $z \in \mathbb{R}$ and a set $A \subseteq \mathbb{R}$. In both cases, the resulting functions $U_{i,N}$ and $U^{(\alpha)}_{i,N}$ are convex and increasing in N, which allows to apply Proposition 20 as detailed in Section D.1.

Proposition 19 *It holds that* $\kappa(\tilde{U}) \leq 1 + \lambda_{max}(XQ_0^{-1}X^T)$.

Proof The Hessian of $\tilde{U}$ is

$$\nabla^2 \tilde{U}(\theta) = I_p + Q_0^{-1/2} X^T D(\theta) X Q_0^{-1/2}, \tag{37}$$

with $D(\theta)$ diagonal and $D_{ii}(\theta) = h''\left(\mathrm{sgn}(2y_i - 1)x_i^T\beta\right)$. By Lemma 18 we deduce that $I_p \succeq D(\theta) \succeq 0$, which implies

$$I_p + Q_0^{-1/2} X^T X Q_0^{-1/2} \succeq \nabla^2 \tilde{U}(\theta) \succeq I_p.$$

This implies that $\lambda_{min}\left(\nabla^2 \tilde{U}(\theta)\right) \geq 1$ and

$$\lambda_{max}\left(\nabla^2 \tilde{U}(\theta)\right) \leq 1 + \lambda_{max}\left(Q_0^{-1/2} X^T X Q_0^{-1/2}\right) = 1 + \lambda_{max}(XQ_0^{-1}X^T),$$

as desired. ■

C.2 Approximate tensorization by convex approximations

In this section we adapt the results in Ascolani et al. (2026) to accommodate for the presence of indicator functions in the target distribution of a Gibbs sampler.

Let $\mathcal{X} = \mathbb{R}^d$ and consider a partition of $\mathcal{X}$ in M blocks with length d_m, i.e. $\mathcal{X} = \bigtimes_{i=m}^M \mathcal{X}_m$ with $\mathcal{X}_m = \mathbb{R}^{d_m}$ for $m = 1, \ldots, M$ and $d = d_1 + \ldots + d_M$. For a point $x = (x_1, \ldots, x_M) \in \mathbb{R}^d$, we write $x_{-m} = (x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_M)$, which is an element of $\mathcal{X}_{-m} = \bigtimes_{i \neq m} \mathcal{X}_i$. Similarly, π_{-m} denotes the marginal distribution of $\pi \in \mathcal{P}(\mathcal{X})$ over $\mathcal{X}_{-m}$.

Define $\pi \in \mathcal{P}(\mathcal{X})$ with density

$$\pi(x) \propto \pi_0(x) e^{-\sum_{m=1}^M U_m(x_m)}, \tag{38}$$

where $\pi_0(x) = N(x \mid m, Q^{-1})$ and $U_m : \mathcal{X}_m \to \mathbb{R} \cup \{+\infty\}$. Let $L_m > 0$ be such that $Q_{mm} - L_m I_{d_m}$ is positive semi-definite, where Q_{mm} is the $d_m \times d_m$ diagonal block of Q. Define

$$\kappa^* = \frac{1}{\lambda_{min}\left(D^{-1/2} Q D^{-1/2}\right)}, \tag{39}$$

where D denotes the diagonal matrix with diagonal coefficients $L_1, L_2, \ldots, L_M$, with each L_m repeated d_m times, that is, $D_{mm} = L_m \mathrm{Id}_{d_m}$. The next proposition proves an approximate tensorization in terms of κ^*.

Proposition 20 *Let* π *be as in* (38) *and assume there exists* $\{\pi_N\}_N \subset \mathcal{P}(\mathcal{X})$ *defined as*

$$\pi_N(x) \propto \pi_0(x) e^{-\sum_{m=1}^M U_{m,N}(x_m)}$$

such that

1. $U_{m,N}$ *is convex for every* m *and* N.
2. $U_{m,N}(x_m) \to U_m(x_m)$ *as* $N \to \infty$ *for every* m *and* x_m.

3. $U_{m,N}(x_m)$ is increasing in N for every m and x_m.

Then for any $\mu \in \mathcal{P}(\mathbb{R}^d)$

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{KL}(\mu_{-m}, \pi_{-m}) \leq \left(1 - \frac{1}{\kappa^* M}\right) \mathrm{KL}(\mu, \pi),$$

with κ^* as in (39).

Proof By 2. we have that $\pi_N \to \pi$ weakly as $N \to \infty$. Moreover, since $U_{m,N}$ is convex, π_N satisfies Assumption B in Ascolani et al. (2026). Thus, by lower semi-continuity of the KL and Theorem 3.1 in Ascolani et al. (2026) we have

$$\frac{1}{M}\sum_{m=1}^{M} \mathrm{KL}(\mu_{-m}, \pi_{-m}) \leq \frac{1}{M}\sum_{m=1}^{M} \liminf_{N\to\infty} \mathrm{KL}(\mu_{-m}, \pi_{N,-m}) \leq \left(1 - \frac{1}{\kappa^* M}\right) \liminf_{N\to\infty} \mathrm{KL}(\mu, \pi_N).$$

Notice that

$$\begin{aligned}\mathrm{KL}(\mu, \pi_N) =& \log\left(\int_{\mathbb{R}^d} \pi_0(x) e^{-\sum_{m=1}^{M} U_{m,N}(x_m)}\,\mathrm{d}x\right) + \int_{\mathbb{R}^d} \log\left(\frac{\mu(x)}{\pi_0(x)}\right)\mu(\mathrm{d}x) \\ &+ \sum_{m=1}^{M}\int_{d_m} U_{m,N}(x_m)\mu_m(\mathrm{d}x_m).\end{aligned}$$

By dominated convergence theorem

$$\int_{\mathbb{R}^d} \pi_0(x) e^{-\sum_{m=1}^{M} U_{m,N}(x_m)}\,\mathrm{d}x \to \int_{\mathbb{R}^d} \pi_0(x) e^{-\sum_{m=1}^{M} U_m(x_m)}\,\mathrm{d}x$$

and by monotone convergence theorem (which holds by 3.)

$$\int_{\mathbb{R}^{d_m}} U_{m,N}(x_m)\mu_m(\mathrm{d}x_m) \to \int_{\mathbb{R}^{d_m}} U_m(x_m)\mu_m(\mathrm{d}x_m)$$

for every $m = 1, \ldots, M$ as $N \to \infty$. Thus we conclude that $\liminf_{N\to\infty} \mathrm{KL}(\mu, \pi_N) = \mathrm{KL}(\mu, \pi)$ from which the result follows. ■

Let now P be the transition kernel associated to the random scan Gibbs sampler on π which alternates sampling from x_m given x_{-m}.

Corollary 21 *Consider the same assumptions of Proposition 20. Then for any $\mu \in \mathcal{P}(\mathbb{R}^d)$*

$$\mathrm{KL}(\mu P, \pi) \leq \left(1 - \frac{1}{\kappa^* M}\right) \mathrm{KL}(\mu, \pi).$$

Proof The proof is analogous to Theorem 3.2 in Ascolani et al. (2026), replacing Theorem 3.1 therein with Proposition 20. ■

C.3 Mixing times bounds in χ^2

We first need a preliminary lemma.

Lemma 22 *Let $\pi, \mu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ and $t \geq 1$. Then*

$$\chi^2\left(\mu P_{\mathrm{DA}}^{t+1}, \pi\right) \leq \chi^2\left(\mu_2 P_{\mathrm{MG}}^{t}, \pi_2\right) \leq \chi^2\left(\mu P_{\mathrm{DA}}^{t}, \pi\right) .$$

In particular this implies

$$\tau_{\mathrm{mix},2}(\epsilon, \mu_2, P_{\mathrm{MG}}) \leq \tau_{\mathrm{mix},2}(\epsilon, \mu, P_{\mathrm{DA}}) \leq 1 + \tau_{\mathrm{mix},2}(\epsilon, \mu_2, P_{\mathrm{MG}}) .$$

Proof The proof is similar to the one of Proposition 11. As regards the lower bound, consider the Markov kernel from $\mathcal{X}_2$ to $\mathcal{X}$ defined as $\Pi_{2\to(1,2)}(x_2, \mathrm{d}x') = \Pi_{2\to 1}(x_2, \mathrm{d}x_1')\Pi_{1\to 2}(x_1', \mathrm{d}x_2')$, so that $\mu P_{\mathrm{DA}}^{t+1} = \mu_2 P_{\mathrm{MG}}^{t}\Pi_{2\to(1,2)}$ for all $t \geq 0$. Combining the latter with $\pi_2\Pi_{2\to(1,2)} = \pi$, and the monotonicity of χ^2 divergence we have that

$$\chi^2\left(\mu P_{\mathrm{DA}}^{t+1}, \pi\right) = \chi^2\left(\mu_2 P_{\mathrm{MG}}^{t}\Pi_{2\to(1,2)}, \pi_2\Pi_{2\to(1,2)}\right) \leq \chi^2\left(\mu_2 P_{\mathrm{MG}}^{t}, \pi_2\right),$$

as desired.

As regards the upper bound, notice that

$$\chi^2(\mu, \pi) = \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right)^2 \pi(\mathrm{d}x) - 1,$$

so it suffices to show that

$$\int_{\mathcal{X}_2} \left(\frac{\mathrm{d}\mu_2 P_{\mathrm{MG}}^{t}}{\mathrm{d}\pi_2}(x_2)\right)^2 \pi_2(\mathrm{d}x_2) \leq \int_{\mathcal{X}_2}\int_{\mathcal{X}_1} \left(\frac{\mathrm{d}\mu P_{\mathrm{DA}}^{t}}{\mathrm{d}\pi}(x)\right)^2 \pi(\mathrm{d}x) \tag{40}$$

By definition of Radon-Nykodim derivative we have that

$$\mu_2 P_{\mathrm{MG}}^{t}(A) = \int_A \frac{\mathrm{d}\mu_2 P_{\mathrm{MG}}^{t}}{\mathrm{d}\pi_2}(x_2)\pi_2(\mathrm{d}x_2), \tag{41}$$

for every $A \subset \mathcal{X}_2$. Moreover, by definition of P_{MG}, we have $\int_{\mathcal{X}_1} \mu P_{\mathrm{DA}}^{t}(\mathrm{d}x_1, A) = \mu_2 P_{\mathrm{MG}}^{t}(A)$, which means

$$\begin{aligned} \mu_2 P_{\mathrm{MG}}^{t}(A) = \mu P_{\mathrm{DA}}^{t}(\mathcal{X}_1, A) &= \int_A \int_{\mathcal{X}_1} \frac{\mathrm{d}\mu P_{\mathrm{DA}}^{t}}{\mathrm{d}\pi}(x_1, x_2)\pi(\mathrm{d}x_1, \mathrm{d}x_2) \\ &= \int_A \left[\int_{\mathcal{X}_1} \frac{\mathrm{d}\mu P_{\mathrm{DA}}^{t}}{\mathrm{d}\pi}(x_1, x_2)\pi(\mathrm{d}x_1 \mid x_2)\right] \pi_2(\mathrm{d}x_2). \end{aligned} \tag{42}$$

Combining (41) and (42) we get that

$$\frac{\mathrm{d}\mu_2 P_{\mathrm{MG}}^{t}}{\mathrm{d}\pi_2}(x_2) = \int_{\mathcal{X}_1} \frac{\mathrm{d}\mu P_{\mathrm{DA}}^{t}}{\mathrm{d}\pi}(x_1, x_2)\pi(\mathrm{d}x_1 \mid x_2),$$

which means

$$\begin{aligned}\int_{\mathcal{X}_2}\left(\frac{\mathrm{d}\mu_2 P^t_{\mathrm{MG}}}{\mathrm{d}\pi_2}(x_2)\right)^2\pi_2(\mathrm{d}x_2) &= \int_{\mathcal{X}_2}\left(\int_{\mathcal{X}_1}\frac{\mathrm{d}\mu P^t_{\mathrm{DA}}}{\mathrm{d}\pi}(x_1,x_2)\pi(\mathrm{d}x_1\mid x_2)\right)^2\pi_2(\mathrm{d}x_2)\\ &\leq \int_{\mathcal{X}_2}\int_{\mathcal{X}_1}\left(\frac{\mathrm{d}\mu P^t_{\mathrm{DA}}}{\mathrm{d}\pi}(x_1,x_2)\right)^2\pi(\mathrm{d}x_1\mid x_2)\pi_2(\mathrm{d}x_2)\\ &= \int_{\mathcal{X}_2}\int_{\mathcal{X}_1}\left(\frac{\mathrm{d}\mu P^t_{\mathrm{DA}}}{\mathrm{d}\pi}(x)\right)^2\pi(x),\end{aligned}$$

and therefore (40) is proved. ■

Denote the spectral gap of a π-reversible Markov kernel P on $\mathcal{X}$ as

$$\mathrm{Gap}(P)=\inf_f\frac{\int\int(f(y)-f(x))^2P(x,\mathrm{d}y)\pi(\mathrm{d}x)}{2\mathrm{Var}_\pi(f)} \tag{43}$$

with the infimum running on every f such that $\mathrm{Var}_\pi(f)<\infty$. It is known (Caputo, 2023, below Lemma 2.15) that

$$\mathrm{Gap}(P)\geq\frac{1-\rho_{EC}(P,\pi)}{2}, \tag{44}$$

and (Andrieu et al., 2024, equation (5)) that

$$\chi^2\left(\mu P^t,\pi\right)\leq\left(1-\mathrm{Gap}(P)\right)^t\chi^2\left(\mu,\pi\right).$$

Combining the above equations we obtain that

$$\tau_{\mathrm{mix},2}(\epsilon,\mu,P)\leq\frac{2}{1-\rho_{EC}(P,\pi)}\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right). \tag{45}$$

We can now prove the analogue of Theorem 1 for $\tau_{\mathrm{mix},2}$.

Theorem 23 *For every $\mu\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^p)$ and $\epsilon>0$, we have*

$$\tau_{\mathrm{mix},2}(\epsilon,\mu,P_{\mathrm{DA}})\leq(3+2\lambda_{max}(XQ_0^{-1}X^T))\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right),$$

and

$$\tau_{\mathrm{mix},2}(\epsilon,\mu_1,P^n_{\mathrm{CG}})\leq 2\,\frac{1+\lambda_{max}(XQ_0^{-1}X^T)}{1+\lambda_{min}(XQ_0^{-1}X^T)}\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right).$$

Proof By (49) below and (45) we have that

$$\tau_{\mathrm{mix},2}(\epsilon,\mu,P_{\mathrm{MG}})\leq\left(2+2\lambda_{max}\left(XQ_0^{-1}X^T\right)\right)\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right).$$

Therefore the result for P_{DA} follows by Lemma 22. The result for P_{CG} follows by combining (51) below and again (45). ■

Appendix D. Proof of the results in the main document

D.1 Proof of Theorem 1

We first prove (11).

Proof We consider a re-parametrized version of model (3), where β is replaced by $\tilde{\beta} = R\beta$ with $R = (Q_0 + X^T X)^{1/2}$ and R symmetric. Note that R is always well defined since Q_0 is positive definite. The model now reads

$$\tilde{\beta} \sim N(\tilde{\beta} \mid Rm, RQ_0^{-1}R), \quad z|\tilde{\beta} \sim N(z \mid XR^{-1}\tilde{\beta}, I_n), \quad y_i = \mathbb{1}(z_i > 0) \text{ for } i = 1, \dots, n\,, \tag{46}$$

The joint posterior of $(z, \tilde{\beta})$ given y under (46) is

$$\tilde{\pi}(z, \tilde{\beta}) \propto N\left((z, \tilde{\beta}) \mid \tilde{\mu}, \tilde{Q}^{-1}\right) \prod_{i=1}^{n} \mathbb{1}\left(y_i = g(z_i)\right), \tag{47}$$

where $\tilde{\mu} = [(Xm)^T, (Rm)^T]^T$ and, since $R^{-1}(Q_0 + X^T X)R^{-1} = I_p$,

$$\tilde{Q} = \begin{pmatrix} I_n & -XR^{-1} \\ -R^{-1}X^T & I_p \end{pmatrix}. \tag{48}$$

Since R is invertible, the two-block GS targeting π and $\tilde{\pi}$ has exactly the same mixing times in KL (see e.g. (Ascolani et al., 2026, Remark 2.3) for details).

Thus, $\pi(z, \tilde{\beta}) = \lim_{N\to\infty} \pi_N(z, \tilde{\beta})$ with

$$\pi_N(z, \tilde{\beta}) \propto N((z, \tilde{\beta}) \mid \tilde{\mu}, \tilde{Q}^{-1}) e^{-U_N(z)}$$

and $U_N(z) = \sum_{i=1}^{n} U_{i,N}(z_i)$ defined as $U_{i,N}(z_i) = N|z_i|\mathbb{1}\left(y_i \neq g(z_i)\right)$. Thus π satisfies the assumptions in Proposition 20 with $M = 2$ and then

$$\frac{\mathrm{KL}(\mu_1, \pi_1) + \mathrm{KL}(\mu_2, \pi_2)}{2} \leq \left(1 - \frac{1}{2\kappa^*}\right) \mathrm{KL}(\mu, \pi),$$

for every $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^p)$, where $\kappa^* = 1/\lambda_{min}(\tilde{Q})$ (since $\tilde{Q}$ has identity block-diagonal terms). Moreover, by standard linear algebra calculations (48) implies

$$\lambda_{min}(\tilde{Q}) = 1 - \sqrt{\lambda_{max}(XR^{-2}X^T)} = 1 - \sqrt{\lambda_{max}(X(Q_0 + X^T X)^{-1}X^T)}.$$

see e.g. Lemma 2 in Goplerud et al. (2024). Therefore, by Theorem 13 we have $\rho_{EC}(P_{\mathrm{MG}}, \pi_2) \leq \lambda_{max}(X(Q_0 + X^T X)^{-1}X^T)$. Applying Woodbury's matrix identity twice, we obtain

$$\begin{aligned} X(Q_0 + X^T X)^{-1}X^T &= X(Q_0^{-1} - Q_0^{-1}X^T(I_n + XQ_0^{-1}X^T)^{-1}XQ_0^{-1})X^T \\ &= M - M(I_p + M)^{-1}M = (I + M^{-1})^{-1} \end{aligned}$$

for $M = XQ_0^{-1}X^T$. Thus

$$\rho_{EC}(P_{\mathrm{MG}}, \pi_2) \leq \frac{1}{\lambda_{min}(I_n + M^{-1})} = \frac{1}{1 + 1/\lambda_{max}(M)} = \frac{\lambda_{max}(M)}{1 + \lambda_{max}(M)}\,. \tag{49}$$

The result then follows from (24) and $\tau_{\mathrm{mix}}(\epsilon, \mu, P_{\mathrm{DA}}) \leq 1 + \tau_{\mathrm{mix}}(\epsilon, \mu_2, P_{\mathrm{MG}})$. ■

The inequality in (12) follows instead by the next theorem.

Theorem 24 *Let π be as in (4), P_{CG} as in (8) and $\mu_1 \in \mathcal{P}(\mathbb{R}^n)$. Then*

$$\begin{aligned}\tau_{\mathrm{mix}}(\epsilon, \mu_1, P_{\mathrm{CG}}^n) &\leq \lambda_{max}(D^{1/2}(I_n + M)D^{1/2}) \log\left(\frac{\mathrm{KL}(\mu_1, \pi_1)}{\epsilon}\right) \\ &\leq \left(\frac{1 + \lambda_{max}(M)}{1 + \lambda_{min}(M)}\right) \log\left(\frac{\mathrm{KL}(\mu_1, \pi_1)}{\epsilon}\right) \\ &\leq (1 + \lambda_{max}(M)) \log\left(\frac{\mathrm{KL}(\mu_1, \pi_1)}{\epsilon}\right),\end{aligned} \tag{50}$$

with $M = XQ_0^{-1}X^T$ and D being a diagonal matrix with diagonal elements equal to the ones of $(I_n + M)^{-1}$.

Proof The marginal distribution of z under (4) is

$$\pi(z) \propto N(z \mid Xm, I_n + M) \prod_{i=1}^{n} \mathbb{1}\left(y_i = g(z_i)\right).$$

Thus, $\pi(z) = \lim_{N\to\infty} \pi_N(z)$ with $\pi_N(z) \propto N(z \mid X\mu, I_n + M)e^{-\sum_{i=1}^n U_{i,N}(z_i)}$ and $U_{i,N}(z_i) = N|z_i|\mathbb{1}\left(y_i \neq g(z_i)\right)$ as above. Thus $\pi(z)$ satisfies the assumptions in Corollary 21, implying $\rho_{EC}(P_{\mathrm{CG}}, \pi_1) \leq 1 - 1/(n\kappa^*)$ with

$$\kappa^* = 1/\lambda_{min}\left(D^{-1/2}(I_n + M)^{-1}D^{-1/2}\right) = \lambda_{max}\left(D^{1/2}(I_n + M)D^{1/2}\right).$$

Combining the latter with (24) gives the first inequality in (50). Then, by Lemma 2.4 in Ascolani et al. (2026) we have that

$$\kappa^* \leq \frac{\lambda_{max}\left((I_n + M)^{-1}\right)}{\lambda_{min}\left((I_n + M)^{-1}\right)} = \frac{1 + \lambda_{max}(M)}{1 + \lambda_{min}(M)} \leq 1 + \lambda_{max}(M),$$

and therefore

$$\rho_{EC}(P_{\mathrm{CG}}, \pi_1) \leq 1 - \frac{1}{n}\left[\frac{1 + \lambda_{min}(M)}{1 + \lambda_{max}(M)}\right], \tag{51}$$

which implies the other two inequalities in (50). ■

D.2 Proof of Proposition 8

We need some preliminary lemmas.

Lemma 25 *Let $X \sim \pi(x) \propto \exp(-U(x))$, with $U : \mathbb{R} \to \mathbb{R}$ convex and twice continuously differentiable, and $x_* = \arg\min_x U(x)$. Let also $a < x_* < b$ and $U(a) = U(b) = U(x_*) + 1$ Then:*

(i) $Var(X) \geq d'(b - a)^2$ for some universal constant $d' > 0$.

(ii) If $a \geq 0$ or $b \leq 0$, then $Var(|X|) \geq d(b - a)^2$ for some universal constant $d > 0$.

(iii) If U'' is monotone, then $b-a \geq \sqrt{2/U''(x_*)}$ and therefore $Var(X) \geq d/U''(x_*)$ for a universal constant $d>0$. If moreover $a \geq 0$ or $b \leq 0$, then $Var(|X|) \geq d'/U''(x_*)$ for a universal constant $d'>0$.

Proof *Part (i).* Assuming $U(x_*)=0$ without loss of generality (w.l.o.g.), we have

$$Z = \int_{\mathbb{R}} \exp(-U(x))dx \leq (b-a) + \int_{-\infty}^{a} \exp(-U(x))dx + \int_{b}^{\infty} \exp(-U(x))dx\,.$$

Using $U(b)=1\geq 0$ and $0=U(x_*)\geq U(b)+U'(b)(x_*-b)$, we deduce $U(x)\geq U'(b)(x-b)\geq \frac{U(b)}{b-x_*}(x-b)=\frac{x-b}{b-x_*}$ for $x\geq b$ and thus $\int_b^\infty \exp(-U(x))dx \leq b-x_*$. By similar arguments $\int_{-\infty}^{a}\exp(-U(x))dx \leq x_*-a$. We thus obtain $Z\leq 2(b-a)$ and

$$\pi(x) \geq \frac{e^{-1}}{2(b-a)} \qquad\qquad x\in(a,b)\,. \tag{52}$$

Then, given $A=(a,b)\subseteq\mathbb{R}$ and $B=\{x\in\mathbb{R}\,:\,|x-\mu|\geq (b-a)/4\}\subseteq\mathbb{R}$ with $\mu\in\mathbb{R}$, we have

$$|A\cap B| = |A| - |A\cap B^c| \geq |A| - |B^c| = (b-a)-(b-a)/2 = (b-a)/2\,. \tag{53}$$

Taking $\mu=\mathbb{E}[X]$ and combining (52) and (53) we obtain

$$\mathbb{P}\left(|X-\mathbb{E}[X]|\geq \frac{b-a}{4}\right) \geq \frac{e^{-1}}{2(b-a)}\frac{b-a}{2} \geq \frac{e^{-1}}{4}$$

which implies

$$\mathrm{Var}(X) = \mathbb{E}[|X-\mathbb{E}[X]|^2] \geq \frac{e^{-1}}{4}\frac{(b-a)^2}{4^2}\,,$$

as desired.

Part (ii). Assume without loss of generality that $a\geq 0$. Then, given $A=(a,b)\subseteq\mathbb{R}$ and $C=\{x\in\mathbb{R}\,:\,||x|-\mu|\geq (b-a)/8\}\subseteq\mathbb{R}$ with $\mu\in\mathbb{R}$, reasoning as in the previous point we have that

$$|A\cap C| \geq |A| - |C^c| = (b-a)-(b-a)/2 = (b-a)/2\,. \tag{54}$$

Notice that the density of $|X|$, denoted by ν, is such that $\nu(x)\geq\pi(x)$ for every $x\geq 0$ and in particular for $X\in A$ by assumption. Taking $\mu=\mathbb{E}[|X|]$ and combining (52) and (54) we obtain

$$\mathbb{P}\left(\left||X|-\mathbb{E}[|X|]\right|\geq\frac{b-a}{8}\right) \geq \int_{A\cap C}\pi(x)\mathrm{d}x \geq \frac{e^{-1}}{2(b-a)}\frac{b-a}{2}\geq\frac{e^{-1}}{4}$$

which implies

$$\mathrm{Var}(X) = \mathbb{E}\left[(|X|-\mathbb{E}[|X|])^2\right] \geq \frac{e^{-1}}{4}\frac{(b-a)^2}{8^2}\,,$$

as desired.

Part (iii). Let $a<x_*<b$ be such that $U(a)=U(b)=U(x_*)+1$. Note that a and b exist and are unique by convexity of U and integrability of $\exp(-U)$. Assuming U'' non-increasing w.l.o.g., we have $U''(x)\leq U''(x_*)$ for $x\geq x_*$ which, together with $U'(x_*)=0$,

implies $U(x) \leq U(x_*) + U''(x_*)(x - x_*)^2/2$ for $x \geq x_*$ and thus $b \geq x_* + \sqrt{2/U''(x_*)}$. Thus $(b - a) \geq (b - x_*) \geq \sqrt{2/U''(x_*)}$ and the two conclusions follow by parts (i) and (ii). ■

Lemma 26 *Let $U(x) = x^2/(2c) + nh(x)$ with $x \in \mathbb{R}$ and $h = -\log \Phi$ as in (34). Then U'' is non-increasing and*

(i) If $cn \geq 3$, then $1/c \leq U''(x_*) \leq 5\log(cn)/c$.

(ii) Let $a < x_*$ such that $U(a) = U(x_*) + 1$. If $cn \geq 8$, then $a \geq 0$.

Proof We know that h'', and thus U'', is decreasing by the representation in (36) and (Mailhot, 1988, Corollary 4). By (35) and $h' = -\phi/\Phi$, we have $h''(x) = h'(x)^2 - xh'(x)$ for all x.

Part(i). Take first $c = 1$. Since $U'(x_*) = x_* + nh'(x_*) = 0$, we deduce $h'(x_*) = -x_*/n$ and

$$1 \leq U''(x_*) = 1 + nh''(x_*) = 1 + n\left(\frac{x_*^2}{n^2} + \frac{x_*^2}{n}\right) = 1 + x_*^2\left(1 + \frac{1}{n}\right) \leq 1 + 2x_*^2.$$

Let $\tilde{x} = \sqrt{2\log(n)}$. Since $\tilde{x} > 0$, we have $h'(\tilde{x}) \geq -2\phi(\tilde{x}) = -\sqrt{2/\pi}\exp(-\tilde{x}^2/2) = -\sqrt{2/\pi}n^{-1} \geq -n^{-1}$. Since $n \geq 3$ we also have $\tilde{x} > 1$, and thus $U'(\tilde{x}) = \tilde{x} + nh'(\tilde{x}) \geq \tilde{x} - 1 \geq 0 = U'(x_*)$. Since U' is increasing we deduce $x_* \leq \tilde{x} = \sqrt{2\log(n)}$ and thus $U''(x_*) \leq 1 + 2x_*^2 \leq 1 + 4\log(n) \leq 5\log(n)$.

For general $c > 0$, write $cU(x) = x^2/2 + cnh(x)$ and apply the result with $c = 1$ and n replaced by nc.

Part(ii). It suffices to prove the result for $c = 1$, reasoning as in point (i). Let $x \leq x_*$. Since U'' is non-increasing we have that

$$U(x) \geq U(x_*) + U''(x_*)\frac{(x - x_*)^2}{2},$$

which implies

$$x_* - a \leq \sqrt{\frac{2}{U''(x_*)}} \leq \sqrt{2}. \tag{55}$$

Moreover let $\tilde{x} = \sqrt{\log n}$. Since $n \geq 8$, we have that $\tilde{x} > 0$ and therefore

$$U'(\tilde{x}) \leq \sqrt{\log n} - \sqrt{\frac{n}{2\pi}} < U'(x_*) = 0.$$

Since $U'(x)$ is increasing we deduce $x_* \geq \tilde{x} = \sqrt{\log n}$ Combining this with (55) we have that

$$a \geq x_* - \sqrt{2} \geq \sqrt{\log n} - \sqrt{2} \geq 0,$$

since $n \geq 8$. ■

Lemma 27 *Consider model* (2) *with* $p = 1$, $m = 0$, $Q_0^{-1} = c > 0$ *and* $x_i = 1$ *for every* i. *Then, if* $y_i = 1$ *for every* i *or* $y_i = 0$ *for every* i, *we have that*

$$Var_\pi(\beta_1) \geq dc/\left(\log(cn)\right) \quad \text{and} \quad Var_\pi(|\beta_1|) \geq d'c/\left(\log(cn)\right)$$

for every $n \geq 8/c$ *and some universal constants* $d > 0$ *and* $d' > 0$.

Proof Assume without loss of generality that $y_i = 1$ for every i. Then $\pi(\beta_1) \propto \exp(-U(\beta_1))$ with U as in Lemma 26 and the result follows combining Lemma 26 with Lemma 25(iii). ■

We are finally ready to prove Proposition 8.

Proof Define $P_{\mathrm{MG}} = \Pi_{z\to\beta}\Pi_{\beta\to z}$, using a notation analogous to Proposition 11. It is well-known (Roberts and Rosenthal, 2001, Section 3.3) that P_{MG} is $\pi(\beta)$-reversible.

Part (i). Choosing $f(\beta) = \beta_1$ in (43) we obtain

$$\mathrm{Gap}(P_{\mathrm{MG}}) \leq \frac{\mathbb{E}\left[\mathrm{Var}_\pi(\beta_1 \mid z)\right]}{\mathrm{Var}_\pi(\beta_1)}\,.$$

By (5) we have $\mathrm{Var}_\pi(\beta_1 \mid z) = 1/(1/c + n)$ for every z and y. We thus obtain

$$\mathrm{Gap}(P_{\mathrm{MG}}) \leq \frac{1}{(1/c+n)\mathrm{Var}_\pi(\beta_1)} \leq \frac{\log(cn)}{dc(1/c+n)}, \tag{56}$$

where the last inequality follows from the lower bound on $\mathrm{Var}_\pi(\beta_1)$ in Lemma 27. Therefore, combining Proposition 11 with (44) we have that

$$\rho_{EC}(P_{\mathrm{DA}}, \pi) \geq \rho_{EC}(P_{\mathrm{MG}}, \pi_2) \geq 1 - \frac{2\log(cn)}{dc(1/c+n)}\,,$$

as desired.

Part (ii). The upper bound follows immediately from Theorem 23 and $XQ_0^{-1}X^T = cn$.

As regards the lower bound, choose $\mu \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R})$ such that $\mu_2 \in \mathcal{P}(\mathbb{R})$ has density $\mu_2(\beta_1) \propto |\beta_1|\pi_2(\beta_1)$. Then

$$f_2(\beta_1) = \frac{\mu_2(\beta_1)}{\pi_2(\beta_1)} = \frac{|\beta_1|}{\int_{\mathbb{R}} |\beta|\pi_2(\beta)\,\mathrm{d}\beta}$$

and

$$\begin{aligned} \frac{\int\int (f_2(\beta') - f_2(\beta))^2 P_{\mathrm{MG}}(\beta, \mathrm{d}\beta')\pi(\mathrm{d}\beta)}{2\mathrm{Var}_\pi(f_2)} &= \frac{\mathbb{E}[\mathrm{Var}_\pi\left(|\beta_1| \mid z\right)]}{\mathrm{Var}_\pi\left(|\beta_1|\right)} \\ &\leq \frac{\mathbb{E}[\mathrm{Var}_\pi\left(\beta_1 \mid z\right)]}{\mathrm{Var}_\pi\left(|\beta_1|\right)} \leq \frac{\log(cn)}{d'c(1/c+n)}, \end{aligned} \tag{57}$$

where the last inequality follows from the lower bound on $\mathrm{Var}_\pi(|\beta_1|)$ in Lemma 27 and $\mathrm{Var}_\pi(\beta_1 \mid z) = 1/(1/c + n)$.

Combining (57) with Corollary 7 in Wu et al. (2022) we have that

$$\tau_{\mathrm{mix},2}(\epsilon, \mu, P_{\mathrm{MG}}) \geq \frac{d'}{2}\left(\frac{1+cn}{\log(cn)}\right)\log\left(\frac{\chi^2(\mu,\pi)}{\epsilon}\right),$$

and the result follows by Lemma 22. ■

D.3 Proof of Proposition 15

Proof By definition of μ and the chain rule we have $\mathrm{KL}(\mu, \pi) = \mathrm{KL}(\mu_2, \pi_2)$. We thus study the latter. By definition of μ_2 and π, and Bayes Theorem, we have $\frac{\mu_2(\beta)}{\pi_2(\beta)} \leq \frac{1}{m(y)}$ for every $\beta \in \mathbb{R}^p$, where

$$m(y) = \int_{\mathbb{R}^p} \mathbb{P}(y_1 = g(z_1), \dots, y_n = g(z_n) \mid \beta) p(\mathrm{d}\beta) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} \Phi(\mathrm{sgn}(2y_i - 1)\eta_i) p_\eta(\mathrm{d}\eta)$$

is the marginal likelihood of y, $p(\beta) = N(\beta \mid (0, \dots, 0)^T, Q_0^{-1})$ is the prior of β and $p_\eta(\eta) = N(\eta \mid (0, \dots, 0)^T, M)$ with $M = XQ_0^{-1}X^T$ and $\eta = (\eta_1, \dots, \eta_n) = X\beta$. Thus $\mathrm{KL}(\mu_2, \pi_2) \leq -\log(m(y))$. Also,

$$m(y) \geq \int_K \prod_{i=1}^{n} \Phi(\mathrm{sgn}(2y_i - 1)\eta_i) p_\eta(\mathrm{d}\eta) \geq \Phi(-1)^n p_\eta(K),$$

with $K = \{\eta \in \mathbb{R}^n \mid \|\eta\|^2 \leq 1\} \subset \{\eta \in \mathbb{R}^n \mid |\eta_i| \leq 1 \text{ for all } i\}$. By $p_\eta = N((0, \dots, 0)^T, M)$ we have $p_\eta(K) \geq \gamma(K)$ with $\gamma = N((0, \dots, 0)^T, \lambda_{max}(M) I_n)$. Then, using $[-n^{-1/2}, n^{-1/2}]^n \subset K$ we obtain

$$\gamma(K) \geq \gamma([-n^{-1/2}, n^{-1/2}]^n) = (N([-n^{-1/2}, n^{-1/2}] \mid 0, \lambda_{max}(M)))^n = a(\, 1/\sqrt{n\lambda_{max}(M)}\,)^n,$$

with $a(\epsilon) := \Phi(\epsilon) - \Phi(-\epsilon) = 2(\Phi(\epsilon) - 0.5)$ for $\epsilon > 0$. Using $\Phi(\epsilon) - \Phi(0) \geq (\Phi(1) - \Phi(0))\epsilon$ for $\epsilon \in (0, 1)$ we deduce $\Phi(\epsilon) - 0.5 \geq \min\{(\Phi(1) - 0.5)\epsilon, \Phi(1) - 0.5\}$ for $\epsilon \in (0, \infty)$ and thus

$$\Phi(\epsilon) - 0.5 \geq \min\{(\Phi(1) - 0.5)\epsilon, \Phi(1) - 0.5\} \geq \frac{1}{4}\min\{1, \epsilon\} = \frac{1}{4\max\{1, \epsilon^{-1}\}} \geq \frac{1}{4(1 + \epsilon^{-2})}, \tag{58}$$

where we used $\Phi(1) - 0.5 \geq 1/4$. The above implies $1/a(\epsilon) \leq 2(1 + \epsilon^{-2})$ and thus

$$-\log(m(y)) \leq n\log(\Phi(-1)^{-1}) - n\log(a(\, 1/\sqrt{n\lambda_{max}(M)}\,)) \leq 2n + n\log(2(1 + n\lambda_{max}(M))),$$

where we used $\log(\Phi(-1)^{-1}) \leq 2$. ■

Appendix E. Empirical comparison between P_{DA} and P_{CG}

In Figure 2 the upper bounds on $\tau^{TV}_{\mathrm{mix}}(\epsilon, \mu, P_{\mathrm{DA}})$ and $\tau^{TV}_{\mathrm{mix}}(\epsilon, \mu, P_{\mathrm{CG}})$, with $\epsilon = 0.1$ and X generated according to Assumption A, are depicted as a function of the ratio n/p. Coherently with the results of Corollary 6, since $c = 10$ is relatively large, when n/p is small P_{CG} yields faster convergence than P_{DA} and viceversa when n/p is large, even if the differences are overall moderate. Interestingly, the behaviour as n grows differ depending on the data generating mechanism. The left plot of Figure 2, regarding data generated from the model, exhibits an increase of the mixing times; both chains seem instead to converge faster in the right plot, where $y_i = 1$ for every i. As mentioned in Section 6, it would be interesting to complement the worst case results of Theorem 1 with upper bounds that depend on the observed y.

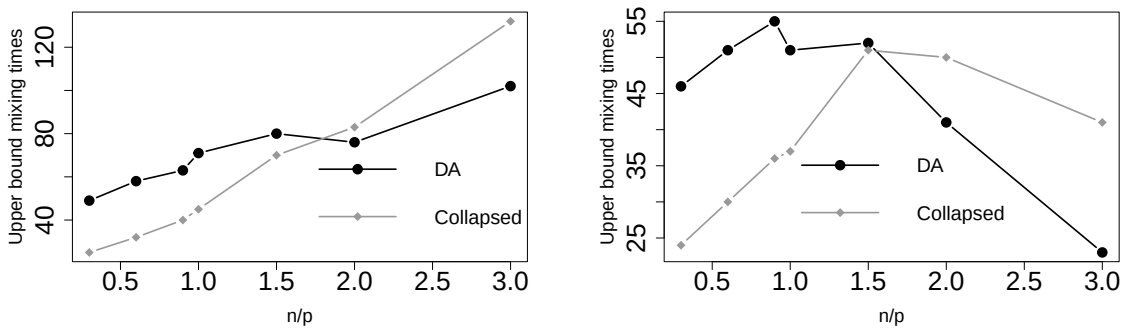

Figure 2: Upper bounds on $\tau^{TV}_{\text{mix}}(\epsilon, \mu, P_{\text{DA}})$ and $\tau^{TV}_{\text{mix}}(\epsilon, \mu, P_{\text{CG}})$ as a function n/p, with $p = 200$, $\epsilon = 0.1$, $Q_0^{-1} = 10 I_p$, $\mu(\mathrm{d}z, \mathrm{d}\beta) = N(\mathrm{d}\beta \mid 0, Q_0^{-1})\pi(\mathrm{d}z \mid \beta)$ and X generated according to Assumption A with $F = N(0, 1)$. Bounds are obtained from (16), taking $L = 200$ and estimating $\bar{d}(t)$ with $N = 500$ independent simulations of $\tau^{(L)}$. Observations are generated according to model (3) (left) and with $y_i = 1$ (right).

Appendix F. Couplings

F.1 Auxiliary couplings

We list below the auxiliary couplings for the algorithms used in Section 5. Algorithms 5 and 6 contain the pseudocode for maximal couplings, i.e. that maximize the probability of the random variables being exactly equal after one step: Algorithm 6 applies only to normal distributions with the same covariance matrix. Algorithms 7 and 8 instead refer to optimal contractive couplings, minimizing the expected squared distance between the random variables after one step: Algorithm 7 only applies to univariate distributions, while Algorithm 8 to Gaussian distributions. Finally, Algorithm 9 provides the pseudocode for a coupling between kernels of Random walk Metropolis operators starting from different points. See Sections A and B of the Supplementary Material of Ceriani et al. (2026) for more details.

F.2 Couplings used in the illustrations

Algorithms 10, 11 and 12 show the pseudocode for the couplings used in Section 5 to upper bound distance from stationarity through (16). We also use the notation $\theta^{(t)} = (z^{(t)}, \beta^{(t)})$ and $d(x, y) = \sqrt{\sum_{j=1}^{J}(x_j - y_j)^2}$ for every $x, y \in \mathbb{R}^J$. In the pseudocode we implicitly assume that the couplings are relative to the corresponding steps in Algorithms 1, 2 and 3 respectively. Moreover we explored a range of values for ϵ and we obtained the tighter bounds with $\epsilon = 1/10$ (Algorithms 10 and 12) and $\epsilon = 1/1000$ (Algorithm 11).

Algorithm 5 Maximal rejection coupling of $p, q \in \mathcal{P}(\mathbb{R}^d)$

Sample $\mathbf{X} \sim p$
Sample $W \sim U(0,1)$
if $Wp(\mathbf{X}) \leq q(\mathbf{X})$ **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 $W^* \sim U(0,1)$
  end while
  Set $\mathbf{Y} = \mathbf{Y}^*$
end if

Algorithm 6 Maximal reflection coupling of $N(\boldsymbol{\xi}, \Sigma)$ and $N(\boldsymbol{\nu}, \Sigma)$

Set $\mathbf{z} = \Sigma^{-1/2}(\boldsymbol{\xi} - \boldsymbol{\nu})$, $\mathbf{e} = \mathbf{z}/||\mathbf{z}||$
Sample $\dot{\mathbf{X}} \sim N(\mathbf{0}, I)$, $W \sim U(0,1)$
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
  Set $\dot{\mathbf{Y}} = \dot{\mathbf{X}} - \mathbf{2}(\mathbf{e}^\top \dot{\mathbf{X}})\mathbf{e}$
end if
Set $\mathbf{X} = \Sigma^{1/2}\dot{\mathbf{X}} + \boldsymbol{\xi}$
Set $\mathbf{Y} = \Sigma^{1/2}\dot{\mathbf{Y}} + \boldsymbol{\nu}$

Algorithm 7 Monotone map for $p, q \in \mathcal{P}(\mathbb{R})$ with distribution functions F_p and F_q

Sample $U \sim U(0,1)$
Set $X = F_p^{-1}(U)$
Set $Y = F_q^{-1}(U)$

Algorithm 8 Common random number coupling of $N(\boldsymbol{\xi}, \Sigma)$ and $N(\boldsymbol{\nu}, \Sigma)$

Set $F = \Sigma^{1/2}$.
Sample $\mathbf{Z} \sim N(\mathbf{0}, I)$
Set $\mathbf{X} = \boldsymbol{\xi} + F\mathbf{Z}$
Set $\mathbf{Y} = \boldsymbol{\nu} + F\mathbf{Z}$

Algorithm 9 Coupling of Random walk Metropolis kernels $P(x, \cdot)$ and $P(y, \cdot)$ with proposal variance σ^2, which are invariant with respect to $p, q \in \mathcal{P}(\mathbb{R})$

Sample $U \sim U(0,1)$
Sample (X', Y') with Algorithm 6 such that $X' \sim N(x, \sigma^2)$ and $Y' \sim N(y, \sigma^2)$.
if $U \leq p(X')/p(x)$ **then**
  Set $X = X'$.
else
  Set $X = x$.
end if
if $U \leq q(Y')/q(y)$ **then**
  Set $Y = Y'$.
else
  Set $Y = y$.
end if

Algorithm 10 (Sampling meeting times $\tau^{(L)}$ for P_{DA})

Set $\epsilon = 1/10$. Initialize $\theta_2^{(0)} \sim \mu$ and $\theta_1^{(0)} \sim \mu$, $\theta_1^{(t)} \mid \theta_1^{(t-1)} \sim P_{\mathrm{DA}}(\theta_1^{(t-1)}, \cdot)$ for $t = 1, \ldots, L$.
for $t > L$ **do**
  if $d\left(\theta_1^{(t-1)}, \theta_2^{(t-L-1)}\right) > \epsilon$ **then**
    Sample $\left(z_1^{(t)}, z_2^{(t-L)}\right)$ by coupling $\pi\left(z_{1i} \mid \beta_1^{(t-1)}\right)$ and $\pi\left(z_{2i} \mid \beta_2^{(t-L-1)}\right)$ for every $i = 1, \ldots, n$ with Algorithm 7.
    Sample $\left(\beta_1^{(t)}, \beta_2^{(t-L)}\right)$ by coupling $\pi\left(\beta_1 \mid z_1^{(t)}\right)$ and $\pi\left(\beta_2 \mid z_2^{(t-L)}\right)$ with Algorithm 8.
  else
    Sample $\left(z_1^{(t)}, z_2^{(t-L)}\right)$ by coupling $\pi\left(z_{1i} \mid \beta_1^{(t-1)}\right)$ and $\pi\left(z_{2i} \mid \beta_2^{(t-L-1)}\right)$ for every $i = 1, \ldots, n$ with Algorithm 5.
    Sample $\left(\beta_1^{(t)}, \beta_2^{(t-L)}\right)$ by coupling $\pi\left(\beta_1 \mid z_1^{(t)}\right)$ and $\pi\left(\beta_2 \mid z_2^{(t-L)}\right)$ with Algorithm 6.
  end if
  If $\theta_1^{(t)} = \theta_2^{(t-L)}$, then return $\tau^{(L)} = t$.
end for

Algorithm 11 (Sampling meeting times $\tau^{(L)}$ for P^n_{CG})

Set $\epsilon = 1/1000$. Initialize $z_2^{(0)} \sim \mu$ and $z_1^{(0)} \sim \mu$, $z_1^{(t)} \mid z_1^{(t-1)} \sim P^n_{\mathrm{CG}}(z_1^{(t-1)}, \cdot)$ for $t = 1, \dots, L$.
for $t > L$ **do**
  Sample $(I_1, \dots, I_n) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{1, \dots, n\})$.
  if $d\left(z^{(t-1)}, z^{(t-L-1)}\right) > \epsilon$ **then**
    for $i = 1, \dots, n$ **do**
      Sample $\left(z_{1I_i}^{(t)}, z_{2I_i}^{(t-L)}\right)$ by coupling $\pi\left(z_{1I_i} \mid z_{1,-I_i}\right)$ and $\pi\left(z_{2i} \mid z_{2,-I_i}\right)$ with Algorithm 7.
    end for
  else
    for $i = 1, \dots, n$ **do**
      Sample $\left(z_{1I_i}^{(t)}, z_{2I_i}^{(t-L)}\right)$ by coupling $\pi\left(z_{1I_i} \mid z_{1,-I_i}\right)$ and $\pi\left(z_{2i} \mid z_{2,-I_i}\right)$ with Algorithm 5.
    end for
  end if
  If $z_1^{(t)} = z_2^{(t-L)}$, then return $\tau^{(L)} = t$.
end for

Algorithm 12 (Sampling meeting times $\tau^{(L)}$ for $P_{\mathrm{DA,mod}}$)

Set $\epsilon = 1/10$. Initialize $\theta_2^{(0)} \sim \mu$ and $\theta_1^{(0)} \sim \mu$, $\theta_1^{(t)} \mid \theta_1^{(t-1)} \sim P_{\mathrm{DA}}(\theta_1^{(t-1)}, \cdot)$ for $t = 1, \dots, L$.
for $t > L$ **do**
  if $d\left(\theta_1^{(t-1)}, \theta_2^{(t-L-1)}\right) > \epsilon$ **then**
    Sample $\left(\tilde{\beta}_{11}, \tilde{\beta}_{21}\right)$ according to Algorithm 9 with $p = \pi\left(\beta_{11} \mid \beta_{1,-1}^{(t-1)}\right)$ and $q = \pi\left(\beta_{21} \mid \beta_{2,-1}^{(t-1)}\right)$.
    Sample $\left(z_1^{(t)}, z_2^{(t-L)}\right)$ by coupling $\pi\left(z_{1i} \mid \tilde{\beta}_{11}, \beta_{1,-1}^{(t-1)}\right)$ and $\pi\left(z_{2i} \mid \tilde{\beta}_{21}, \beta_{2,-1}^{(t-1)}\right)$ for every $i = 1, \dots, n$ with Algorithm 7.
    Sample $\left(\beta_1^{(t)}, \beta_2^{(t-L)}\right)$ by coupling $\pi\left(\beta_1 \mid z_1^{(t)}\right)$ and $\pi\left(\beta_2 \mid z_2^{(t-L)}\right)$ with Algorithm 8.
  else
    Sample $\left(\tilde{\beta}_{11}, \tilde{\beta}_{21}\right)$ according to Algorithm 9 with $p = \pi\left(\beta_{11} \mid \beta_{1,-1}^{(t-1)}\right)$ and $q = \pi\left(\beta_{21} \mid \beta_{2,-1}^{(t-1)}\right)$.
    Sample $\left(z_1^{(t)}, z_2^{(t-L)}\right)$ by coupling $\pi\left(z_{1i} \mid \tilde{\beta}_{11}, \beta_{1,-1}^{(t-1)}\right)$ and $\pi\left(z_{2i} \mid \tilde{\beta}_{21}, \beta_{2,-1}^{(t-1)}\right)$ for every $i = 1, \dots, n$ with Algorithm 5.
    Sample $\left(\beta_1^{(t)}, \beta_2^{(t-L)}\right)$ by coupling $\pi\left(\beta_1 \mid z_1^{(t)}\right)$ and $\pi\left(\beta_2 \mid z_2^{(t-L)}\right)$ with Algorithm 6.
  end if
  If $\theta_1^{(t)} = \theta_2^{(t-L)}$, then return $\tau^{(L)} = t$.
end for