Posterior computation with the Gibbs zig-zag sampler

Sachs, Matthias; Sen, Deborshee; Lu, Jianfeng; Dunson, David

DOI:
10.1214/22-BA1319

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Sachs, M, Sen, D, Lu, J & Dunson, D 2022, 'Posterior computation with the Gibbs zig-zag sampler', *Bayesian Analysis*. https://doi.org/10.1214/22-BA1319

Bayesian Analysis (2022) **TBA**, Number TBA, pp. 1–19

Posterior Computation with the Gibbs Zig-Zag Sampler*

Matthias Sachs[†,‖,**], Deborshee Sen[‡,‖], Jianfeng Lu[§], and David Dunson[¶]

Abstract. An intriguing new class of piecewise deterministic Markov processes (PDMPs) has recently been proposed as an alternative to Markov chain Monte Carlo (MCMC). We propose a new class of PDMPs termed Gibbs zig-zag samplers, which allow parameters to be updated in blocks with a zig-zag sampler applied to certain parameters and traditional MCMC-style updates to others. We demonstrate the flexibility of this framework on posterior sampling for logistic models with shrinkage priors for high-dimensional regression and random effects, and provide conditions for geometric ergodicity and the validity of a central limit theorem.

Keywords: Gibbs sampler, Markov chain Monte Carlo, non-reversible, piecewise deterministic Markov process, sub-sampling.

1 Introduction

Despite alternative methods ranging from sequential Monte Carlo (Del Moral et al., 2006) to variational inference (Beal, 2003), Markov chain Monte Carlo (MCMC) methods remain the default approach among Bayesian statisticians and show no signs of diminishing in importance. The overwhelming majority of the literature on MCMC methods has focused on *reversible* Markov chains (that is, Markov chains which satisfy a detailed balance condition), typically constructed as instances of the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970). This includes MH samplers that obtain efficient joint proposals using gradient information, ranging from Hamiltonian Monte Carlo (HMC, Duane et al., 1987) to Metropolis-adjusted Langevin algorithms (Roberts and Tweedie, 1996). Likewise, this includes the Gibbs sampler (Geman and Geman, 1987) and generalizations that replace sampling parameters one at a time from their conditional posterior distributions with block updating using a broad class of MH steps.

Data sub-sampling has been explored as a way to speed up MCMC for large datasets (Welling and Teh, 2011; Maclaurin and Adams, 2015; Quiroz et al., 2018). Sub-samples are used to approximate transition probabilities and reduce bottlenecks in calculating

*DS and DD acknowledge support from National Science Foundation grant 1546130. MS and DS acknowledge support from grant DMS-1638521 from SAMSI. The work of JL is supported in part by the National Science Foundation via grants DMS-1454939 and CCF-1934964 (Duke TRIPODS).

[†]School of Mathematics, University of Birmingham, m.sachs@bham.ac.uk
[‡]Department of Mathematical Sciences, University of Bath, ds2469@bath.ac.uk
[§]Department of Mathematics, Duke University, jianfeng@math.duke.edu
[¶]Department of Statistical Science, Duke University, dunson@duke.edu
[‖]The two authors contributed equally to this article.
[**]Corresponding author.

 https://doi.org/10.1214/22-BA1319

likelihoods and gradients, with the current literature focusing mostly on modifications of the MH algorithm. A major drawback of these approaches is that it is typically difficult to create schemes which preserve the correct target distribution. While there has been work on quantifying the error for such approximate MCMC schemes (Pillai and Smith, 2014; Johndrow et al., 2015; Johndrow and Mattingly, 2017), it is in general difficult to do so. The pseudo-marginal approach of Andrieu and Roberts (2009) offers a potential solution, but it is generally impossible to obtain the required unbiased estimators of likelihoods using data sub-samples (Jacob and Thiery, 2015).

There is evidence to show that *non-reversible* MCMC methods can offer drastic increased sampling efficiency over reversible MCMC methods (Diaconis et al., 2000; Sun et al., 2010; Chen and Hwang, 2013; Rey-Bellet and Spiliopoulos, 2015). A recently popularized class of non-reversible stochastic processes that can be used to construct sampling algorithms (Peters, 2012; Vanetti et al., 2017; Fearnhead et al., 2018) are piecewise deterministic Markov processes (PDMPs). PDMPs follow a Markov jump process, where the process evolves deterministically according to some predefined dynamics in between jump events, with the event times being distributed according to a Poisson process. Examples of PDMPs include the bouncy particle sampler (BPS; Bouchard-Côté et al., 2018) and the zig-zag (ZZ) process (Bierkens et al., 2019a). Very interestingly, in contrast to traditional MH-based algorithms, PDMPs allow error-free sub-sampling of the data. This remarkable feature has been shown to hold for a wide range of PDMPs (Vanetti et al., 2017).

Although theoretically well-founded, PDMP approaches have not yet found widespread use in Bayesian statistics. A major reason for this is the fact that the application of these methods is in general not straightforward. The implementation of PDMPs requires the derivation of upper bounds for the gradient of the log posterior density. These upper bounds must be sufficiently tight for the sampling to remain efficient. While there have been attempts to automate the construction of such upper bounds (Pakman et al., 2017) as well as relax the need for upper bounds (Cotter et al., 2020), these lack theoretical guarantees for the exact preservation of the target measure and as such fall into a similar category as approximate MCMC schemes. An additional challenge is that the upper bounds typically deteriorate as the dimension of the parameter space increases, although this can be mitigated to a certain extent using non-uniform subsampling schemes (Sen et al., 2020).

In this article, we address the problem of increasing the versatility of PDMP-based sampling approaches by introducing a new framework which allows the inclusion of component-wise MCMC updates within a PDMP process. The main idea is to update blocks of components for which efficient upper bounds can be easily derived by a PDMP process, and update blocks of components for which such upper bounds are not easily available with a suitable MH scheme. This allows us to combine the versatility of traditional MCMC approaches with the advantages of PDMPs in sampling problems. This is particularly relevant to Bayesian hierarchical models, where is it common for certain parameters to have conditional posteriors distributions that are easy to sample from via a Gibbs step, while other parameters can be efficiently updated using a PDMP. In order to keep the presentation simple and accessible, we focus our attention on the

ZZ process in terms of PDMPs, and we refer to our framework as the Gibbs-zig-zag (GZZ) sampler/process. However, we remark that the proposed framework is generic and allows combining a wide class of PDMPs with block-wise MH updates, as detailed in Section 3 of the supplementary material (Sachs et al., 2022).

The rest of the article is organized as follows. We begin with reviewing the ZZ process in Section 2. We present the GZZ sampler in Section 3. In particular, we discuss its construction in Section 3.1, present its application to posterior sampling from Bayesian hierarchical models in Section 3.2, and summarize its main ergodic properties in Section 3.3. Section 4 contains numerical examples for two different contexts related to logistic regression. Finally, Section 5 concludes. Proofs and additional details of sampling algorithms are deferred to the supplementary material.

2 The zig-zag sampler

We review the zig-zag (ZZ) process as introduced in Bierkens et al. (2019a) in this section. Consider the problem of sampling from a probability measure

$$\pi(\mathrm{d}\zeta) = \frac{1}{Z}\exp\{-U(\zeta)\}\mathrm{d}\zeta,$$

where $U \in \mathcal{C}^2(\Omega_\zeta, \mathbb{R})$ is a smooth potential function defined on $\Omega_\zeta \subset \mathbb{R}^d$. For the remainder of this paper, we describe the ZZ sampler when $\Omega_\zeta = \mathbb{R}^d$; however, Ω_ζ can be a strict subset of $\mathbb{R}^d$ as well (Bierkens et al., 2018). The ZZ process $\{\boldsymbol{\zeta}(t), \boldsymbol{\theta}(t)\}_{t\geq 0}$ is a piecewise deterministic continuous-time Markov process which lives on an augmented phase space $\Omega_\zeta \times \{-1, 1\}^d$ and is constructed such that the process is ergodic with respect to the product measure $\widetilde{\pi}(\mathrm{d}\zeta, \theta) = \pi(\mathrm{d}\zeta)\mu(\theta)$, where μ is the uniform measure on $\{-1, 1\}^d$. The components $\boldsymbol{\zeta}(t)$ and $\boldsymbol{\theta}(t)$ are commonly referred to as the position and velocity of the process, respectively.

For a starting point $\boldsymbol{\zeta}^0$ and initial velocity $\boldsymbol{\theta}^0$, the ZZ process evolves deterministically as

$$\boldsymbol{\zeta}(t) = \boldsymbol{\zeta}^0 + \boldsymbol{\theta}^0 t, \quad \boldsymbol{\theta}(t) = \boldsymbol{\theta}^0. \tag{1}$$

At random times $(T^k)_{k\in\mathbb{N}}$, bouncing events occur which flip the sign of one component of the velocity $\boldsymbol{\theta}^{k-1}$. The process then evolves as (1) with the new velocity until the next change in velocity; that is,

$$\boldsymbol{\zeta}(T^k + s) = \boldsymbol{\zeta}^k + \boldsymbol{\theta}^k s, \quad \boldsymbol{\theta}(T^k + s) = \boldsymbol{\theta}^k, \tag{2}$$

for $s \in [0, T^{k+1} - T^k]$, where $\boldsymbol{\theta}^k = F_{I^k}(\boldsymbol{\theta}^{k-1})$, with random component index I^k as specified below and F_i denoting the operator which changes the sign of the i-th component of its argument, that is $F_i : \{-1, 1\}^d \to \{-1, 1\}^d$ with $\{F_i(\theta)\}_j = \theta_j$ if $j \neq i$ and $-\theta_j$ if $j = i$. The random event times $(T^k)_{k\in\mathbb{N}}$ correspond to arrival times of a non-homogeneous Poisson arrival process whose intensity function $m(t) = \sum_{i=1}^d m_i(t)$ depends on the current phase space value of the process, that is, $m_i(t) = \lambda_i\{\boldsymbol{\zeta}(t), \boldsymbol{\theta}(t)\}$ $(i = 1, \ldots, d)$, where $\lambda_1, \ldots, \lambda_d$ are referred to as rate functions. The k-th waiting time $\tau^k = (T^{k+1} - T^k)$ of this arrival process is $\tau^k = \tau^k_{I^k}$ with $I^k = \mathrm{argmin}_{i\in\{1,\ldots,d\}}\{\tau^k_i\}$,

where τ_i^k $(i = 1, \ldots, d)$ are random times whose densities are specified by the hazard rates $m_i^k(s) = \lambda_i\{\boldsymbol{\zeta}(T^k + s), \boldsymbol{\theta}(T^k + s)\}$.

Let $(x)^+ = \max\{0, x\}$ denote the positive part of $x \in \mathbb{R}$. If the rate functions have the form

$$\lambda_i(\zeta, \theta) = \left\{\theta_i \frac{\partial U(\zeta)}{\partial \zeta_i}\right\}^+ + \gamma_i(\zeta) \quad (i = 1, \ldots, d)$$

with $\gamma_i(\zeta) \geq 0$, this ensures that $\widetilde{\pi}$ is an invariant measure of the process (Bierkens et al., 2019a), where $\zeta = (\zeta_1, \ldots, \zeta_d)$. The γ_is are known as the refreshment rates. Slightly more restrictive conditions ensuring exponential convergence in law to the measure $\widetilde{\pi}$ and the validity of a central limit theorem can be found in Bierkens and Duncan (2017) (see also Bierkens et al., 2019b).

In general, the integrals $\int_0^s m_i^k(r)\,\mathrm{d}r$ of the rate functions $m_i^k(s)$ do not have a simple closed form, and thus the corresponding first arrival times τ_i^k cannot be sampled using a simple inverse transform. Instead, arrival times are usually sampled via a Poisson thinning step (Lewis and Shedler, 1979) as follows. Assume that we have continuous functions $M_i : \Omega_\zeta \times \{-1, 1\}^d \times \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lambda_i(\zeta + s\theta, \theta) \leq M_i(\zeta, \theta, s)$. Then

$$m_i^k(s) = \lambda_i(\boldsymbol{\zeta}^k + s\boldsymbol{\theta}^k, \boldsymbol{\theta}^k) \leq M_i(\boldsymbol{\zeta}^k, \boldsymbol{\theta}^k, s) =: M_i^k(s) \quad (i = 1, \ldots, d; \quad s \geq 0). \tag{3}$$

Let $\widetilde{\tau}_1^k, \ldots, \widetilde{\tau}_d^k$ be the first arrival times of Poisson processes with rates $M_1^k(s), \ldots, M_d^k(s)$, respectively. Let $I^k = \mathrm{argmin}_{i \in \{1,\ldots,d\}}\{\widetilde{\tau}_i^k\}$ denote the index of the smallest arrival time. Then, if

(i) $\boldsymbol{\zeta}(t)$ is evolved according to (2) for time $s = \widetilde{\tau}_{I^k}^k$, and

(ii) after time $\widetilde{\tau}_{I^k}$ the sign of θ_{I^k} is flipped with probability $m_{i_0}^k(\widetilde{\tau}_{I^k})/M_{I^k}^k(\widetilde{\tau}_{I^k})$,

the resulting process can be shown to be a ZZ process with intensities $m_i(t) = \lambda_i\{\boldsymbol{\zeta}(t), \boldsymbol{\theta}(t)\}$ $(i = 1, \ldots, d)$ (Bierkens et al., 2019a).

A particularly appealing feature of the ZZ sampler (and PDMPs in general) is that the Poisson thinning procedure can be modified in a way which allows replacing the partial derivatives of the potential function in computations of the event times of bounces by unbiased estimates without changing the invariant measure of the simulated ZZ process (Vanetti et al., 2017). The unbiased estimates can be obtained by sub-sampling of the data when observations are independent.

3 The Gibbs zig-zag sampler

3.1 Process description

In practice, derivation of tight upper bounds $M_i(t)$ as described in the previous section is often challenging. While using generalized sub-sampling schemes can help in improving the tightness of upper bounds in the setup of sub-sampling (Sen et al., 2020), the construction of upper bounds nonetheless remains a fundamental hurdle limiting the

use of PDMPs in practice. In order to simplify applications of the ZZ sampler, we introduce a novel extension which combines elements of Gibbs sampling with a PDMP framework.

Consider a decomposition of the parameter vector as

$$\zeta = (\xi, \alpha) \in \Omega_\xi \times \Omega_\alpha = \mathbb{R}^p \times \mathbb{R}^r,$$

where $d = (p + r)$, and let $\theta \in \{-1, 1\}^p =: \Omega_\theta$. The idea of the Gibbs zig-zag (GZZ) sampler is to combine updates of the component ξ via a ZZ process, which for fixed value of α preserves the conditional measure

$$\pi(\mathrm{d}\xi \mid \alpha) \propto \exp\{-U(\xi, \alpha)\}\,\mathrm{d}\xi,$$

with conventional (Markov chain) Monte Carlo updates of the second component α, which for given value of ξ preserve the conditional measure

$$\pi(\mathrm{d}\alpha \mid \xi) \propto \exp\{-U(\xi, \alpha)\}\,\mathrm{d}\alpha.$$

These updates are combined in such a way that the resulting process is a PDMP which samples the target distribution π.

More precisely, let $\mathcal{L}_{\mathrm{ZZ}}$ denote the generator of the process which leaves the second component α constant while evolving the first component ξ in the corresponding affine subspace according to a ZZ process with rate function

$$\widetilde{m}_i(t, \alpha) = [\theta_i \partial_{\xi_i} U\{\boldsymbol{\xi}(t), \alpha\}]^+ + \gamma_i\{\boldsymbol{\xi}(t), \alpha\} \quad (i = 1, \ldots, p; \quad t \geq 0); \tag{4}$$

we have used the shorthand notation $\partial_{\xi_i} U$ to denote $(\partial/\partial\xi_i)U$. The generator $\mathcal{L}_{\mathrm{ZZ}}$ takes the form of the differential operator

$$(\mathcal{L}_{\mathrm{ZZ}} f)(\xi, \alpha, \theta) = \sum_{i=1}^{p} \theta_i \partial_{\xi_i} f(\xi, \alpha, \theta) + \lambda_i(\xi, \alpha, \theta)\,[f\{\xi, \alpha, F_i(\theta)\} - f(\xi, \alpha, \theta)]\,, \quad f \in \mathscr{S},$$

when considered as an operator on the set of smooth test functions $\mathscr{S} = \mathcal{C}^\infty(\Omega, \mathbb{R})$. Here and in the sequel, we consider $\Omega = \Omega_\xi \times \Omega_\alpha \times \Omega_\theta$ to be equipped with the product topology induced by the Euclidean norms on Ω_ξ and Ω_α, and the discrete topology on Ω_θ, so that a function $f : \Omega \to \mathbb{R}$ is continuous exactly if $f_\theta : (\xi, \alpha) \mapsto f(\xi, \alpha, \theta)$ is continuous for all $\theta \in \Omega_\theta$. Similarly, we consider the function f to be differentiable if the partial derivatives $\partial_{\xi_i} f_\theta$ $(i = 1, \ldots, p)$ and $\partial_{\alpha_i} f_\theta$ $(i = 1, \ldots, r)$ are well defined for all $\theta \in \Omega_\theta$ and measurable if f_θ is Lebesgue measurable for all $\theta \in \Omega_\theta$; we have used the shorthands $\partial_{\xi_i} f_\theta$ and $\partial_{\alpha_i} f_\theta$ to denote $(\partial/\partial_{\xi_i})f_\theta$ and $(\partial/\partial_{\alpha_i})f_\theta$, respectively.

Let $\mathcal{Q}$ be a Markov kernel which is such that for any $\xi \in \Omega_\xi$, the conditional measure $\pi(\mathrm{d}\alpha \mid \xi)$ is preserved under the action of $\mathcal{Q}$ in the sense that $\int \mathcal{Q}\{(\xi, \alpha'), A\}\pi(\mathrm{d}\alpha' \mid \xi) = \int \mathbf{1}_A(\alpha)\pi(\mathrm{d}\alpha \mid \xi)$ for any measurable set $A \subset \Omega_\alpha$, where $\mathbf{1}_A(\alpha)$ stands for the indicator function which is such that $\mathbf{1}_A(\alpha) = 1$ if $\alpha \in A$ and zero otherwise. Let $(\widetilde{T}^k)_{k\in\mathbb{N}}$ denote event times of a Poisson process with constant rate $\eta > 0$. The generator of the PDMP in Ω_α which is constant in between event times $(\widetilde{T}^k)_{k\in\mathbb{N}}$ and whose state is resampled

from the Markov kernel $\mathcal{Q}$ at event times takes the form $\eta\mathcal{L}_{\text{Gibbs}}$, where

$$(\mathcal{L}_{\text{Gibbs}} f)(\xi, \alpha, \theta) = \int_{\Omega_\alpha} \{f(\xi, \alpha', \theta) - f(\xi, \alpha, \theta)\}\, \mathcal{Q}\{(\alpha, \xi), \mathrm{d}\alpha'\}, \quad f \in \mathscr{S}. \tag{5}$$

We obtain the GZZ process by superimposing the two processes described above; that is, we construct the GZZ process as the process whose generator is

$$\mathcal{L}_{\text{GZZ}} = \mathcal{L}_{\text{ZZ}} + \eta\mathcal{L}_{\text{Gibbs}}.$$

The corresponding process $\{\boldsymbol{\xi}(t), \boldsymbol{\theta}(t), \boldsymbol{\alpha}(t)\}_{t\geq 0}$ is a PDMP whose trajectory is piecewise linear in ξ and piecewise constant in α. It follows from classical results on the simulation of non-homogeneous Poisson processes that the process can be simulated by generating skeleton points $\{(\boldsymbol{\xi}^k, \boldsymbol{\theta}^k, \boldsymbol{\alpha}^k, T^k)\}_{k\in\mathbb{N}}$ according to Algorithm 1 below, which are then linearly interpolated as

$$\boldsymbol{\xi}(t) = \boldsymbol{\xi}^k + \boldsymbol{\theta}^k(t - T^k), \quad \boldsymbol{\alpha}(t) = \boldsymbol{\alpha}^k, \quad \boldsymbol{\theta}(t) = \boldsymbol{\theta}^k, \quad \text{for } T^k \leq t < T^{k+1}. \tag{6}$$

Remark 1. *We mention two recently proposed sampling schemes that – similar to the GZZ sampler – combine ideas of PDMP with aspects of classical Gibbs sampling, but whose constructions are in fact conceptually very different from the proposed GZZ sampler.*

1. *The PDMP proposed in Wu and Robert (2020), termed the coordinate sampler, resembles aspects of a classical Gibbs sampler, but other than the fact that certain components are kept constant in between jump events it bears no resemblance to the GZZ sampler. In terms of its construction, the coordinate sampler falls into the same framework as other popular PDMP processes such as the ZZ sampler (Bierkens et al., 2019a) and the BPS (Bouchard-Côté et al., 2018), and unlike the GZZ process, it does not allow the incorporation of MH-updates.*

2. *In Zhao and Bouchard-Côté (2021), the authors propose a local BPS-within-Gibbs algorithm for the sampling of the posterior distribution of parameters of a continuous Markov chain. Unlike the GZZ sampler, which is a piecewise deterministic continuous time Markov process, the sampling scheme proposed in that work is constructed as a Markov chain where in each step parameters are block-wise updated using either a local BPS sampler or an HMC sampler. A similar construction of a Markov chain that combines a BPS sampler and a Metropolis-Hastings scheme has been also considered in Zhang et al. (2021).*

As we discuss in Section 3.3, the GZZ process is path-wise ergodic with respect to the augmented measure

$$\widetilde{\pi}(\mathrm{d}\zeta, \theta) = \pi(\mathrm{d}\zeta)\, \mu(\theta), \tag{7}$$

where μ is the uniform measure on $\{-1, 1\}^d$, under some mild conditions on the potential function U. As such, it can be used similarly to other PDMP samplers as a Monte Carlo method for the approximate computation of expectations by finite time trajectory averages, that is,

$$\mathbb{E}_{(\xi,\alpha)\sim\pi}\{f(\xi, \alpha)\} \approx \frac{1}{t}\int_0^t f\{\boldsymbol{\xi}(t), \boldsymbol{\alpha}(t)\}\, \mathrm{d}t.$$

Algorithm 1 Gibbs zig-zag (GZZ) algorithm.

Input: $(\boldsymbol{\xi}^0, \boldsymbol{\alpha}^0, \boldsymbol{\theta}^0) \in \Omega_\xi \times \Omega_\alpha \times \Omega_\theta$

for $k = 1, 2, \ldots$ **do**

Draw $\tau' \sim \text{Exponential}(\eta)$ and $\widetilde{\tau}_1, \ldots, \widetilde{\tau}_p$ such that

$$\mathbb{P}(\widetilde{\tau}_i \geq s) = \exp\left\{ -\int_0^s \widetilde{m}_i(T^k + r, \boldsymbol{\alpha}^k)\, \mathrm{d}r \right\} \ (i = 1, \ldots, p). \tag{8}$$

Let $\tau^k = \min\{\tau', \widetilde{\tau}_1, \ldots, \widetilde{\tau}_p\}$.

Set $\boldsymbol{\xi}^{k+1} = \boldsymbol{\xi}^k + \tau^k \boldsymbol{\theta}^k$ and $T^{k+1} = T^k + \tau^k$.

if $\tau = \tau'$ **then**

Draw $\boldsymbol{\alpha}^{k+1} \sim \mathcal{Q}\{(\boldsymbol{\xi}^{k+1}, \boldsymbol{\alpha}^k), \cdot\}$.

Set $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k$.

else

Set $\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k$.

Bounce: $\boldsymbol{\theta}^{k+1} = F_{i_0}(\boldsymbol{\theta}^k)$, with $i_0 = \text{argmin}_{i \in \{1,\ldots,p\}} \widetilde{\tau}_i$.

end if

end for

Output: Skeleton points $\{(\boldsymbol{\xi}^k, \boldsymbol{\alpha}^k, \boldsymbol{\theta}^k, T^k)\}_{k \in \mathbb{N}}$.

Practically, ZZ updates of the ξ component can be performed using Poisson thinning. In this case, an upper bound $\widetilde{M}_i : \Omega_\xi \times \Omega_\alpha \times \{-1, 1\}^d \times \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $\lambda_i(\xi + s\theta, \alpha, \theta) \leq \widetilde{M}_i(\xi, \alpha, \theta, s)$ for all $s \geq 0$ is required.

The approach is particularly useful if the restriction of the ZZ process onto the component ξ simplifies construction of upper bounds, and efficient MCMC updates for the remaining component α are available. Such a decomposition is often naturally available in the context of Bayesian posterior distributions with hierarchical priors. We describe the application of the GZZ sampler to such models in Section 3.2. In addition, in the context of Bayesian posterior computation, the GZZ sampler can be modified to allow for sub-sampling of data while exactly preserving the measure $\widetilde{\pi}$. This is explored numerically in Section 4.

3.2 Bayesian posterior sampling with hierarchical priors

Hierarchical Bayesian models can often be specified as

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x \mid \xi), \quad \xi \mid \alpha \sim h(\xi \mid \alpha), \quad \alpha \sim p_0(\alpha),$$

where α are hyperparameters with hyper-prior $p_0(\alpha)$, h denotes the conditional distribution of the parameters given the hyperparameters, and f denotes the likelihood of observations given parameters. Bayesian posterior sampling typically proceeds by

sampling from

$$p(\xi, \alpha \mid X_1, \ldots, X_n) \propto \prod_{j=1}^{n} f(X_j \mid \xi) \times h(\xi \mid \alpha)\, p_0(\alpha),$$

whose marginal distribution for ξ is the posterior distribution of ξ given $X_1, \ldots, X_n$, which is $p(\xi \mid X_1, \ldots, X_n) \propto \prod_{j=1}^{n} f(X_j \mid \xi) \int_{\Omega_\alpha} h(\xi \mid \alpha)\, p_0(\alpha)\, \mathrm{d}\alpha$. Letting $\zeta = (\xi, \alpha)$, this corresponds to sampling the Gibbs measure $\pi(\mathrm{d}\zeta) = Z^{-1} \exp\{-U(\zeta)\}\, \mathrm{d}\zeta$ with potential function

$$U(\zeta) = U^0(\xi, \alpha) + \sum_{j=1}^{n} U^j(\xi), \tag{9}$$

where $U^0(\xi, \alpha) = -\log h(\xi \mid \alpha) - \log p_0(\alpha)$ and $U^j(\xi) = -\log f(X_j \mid \xi)$. The GZZ sampler can readily be applied in this context, with $\mathcal{Q}$ corresponding to either an exact update for the hyperparameters (which is the case when using conditionally conjugate priors), or using a suitable Metropolis-Hastings (MH) scheme such as random walk MH or Hamiltonian Monte Carlo (HMC), when such an exact update is not possible. We consider numerical examples of this in Section 4.

3.3 Ergodic properties

We provide a high level overview of theoretical results on the ergodic properties of the GZZ sampler in this section. Detailed conditions and theorems as well as proofs are deferred to Section 1 of the supplementary material. These results pertain to the long term properties of trajectory averages

$$\widehat{\varphi}_t = \frac{1}{t} \int_0^t \varphi\{\boldsymbol{\xi}(s), \boldsymbol{\theta}(s), \boldsymbol{\alpha}(s)\}\, \mathrm{d}s$$

of $\widetilde{\pi}$-integrable functions φ, and as such are intended to justify the usage of the GZZ sampler as a Monte Carlo method. First, Theorem 1 states that under relatively mild conditions (summarized in Assumption 1) on the Markov transition kernel $\mathcal{Q}$ and the refreshment rate functions λ_i $(i = 1, \ldots, p)$, the trajectory averages converge almost surely to corresponding expectations, that is,

$$\lim_{t \to \infty} \widehat{\varphi}_t = \mathbb{E}_{(\xi,\alpha,\theta) \sim \widetilde{\pi}}\{\varphi(\xi, \alpha, \theta)\} \quad \text{almost surely.}$$

Secondly, we show in Theorem 3 that under additional conditions on the potential function U (see Assumption 2), the GZZ process is geometrically ergodic. This implies a central limit theorem (Corollary 2) of the form

$$\sqrt{t}\left[\widehat{\varphi}_t - \mathbb{E}_{(\xi,\alpha,\theta) \sim \widetilde{\pi}}\{\varphi(\xi, \alpha, \theta)\}\right] \xrightarrow[t \to \infty]{\text{law}} \mathcal{N}(0, \sigma_\varphi^2) \quad \text{for some } \sigma_\varphi^2 > 0$$

which holds for a wide class of real-valued functions φ.

Assumption 1 explicitly requires the refreshment rate function λ_i to be strictly positive. This drastically simplifies the proof of irreducibility of the process. By doing so,

we circumvent difficulties (as described and tackled in Bierkens et al., 2019b in the case of the standard ZZ process) in the proof of the irreducibility of the process. From a theoretical perspective, this makes the presented results less interesting. However, we expect that the GZZ process will typically be used in combination with a sub-sampling scheme so that vanishing refreshment rates are unlikely in practice. Likewise, conditions on the potential function U are rather restrictive in terms of the tail properties of the corresponding density and the coupling between the parameters ξ and the hyperparameters α. We acknowledge that some of these conditions might not hold in practice. Instead, we demonstrate in numerical experiments (see Section 4) that the central limit theorem remains valid in settings beyond what it is covered by our theoretical results.

4 Numerical examples

Consider the following generic logistic regression model,

$$Y_j \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\psi_j}}\right) \quad (j = 1, \ldots, n), \tag{10}$$

where $Y_1, \ldots, Y_n \in \{0, 1\}$ denote observations and $\psi_1, \ldots, \psi_n$ are linear predictors that are further assigned a model in a context-specific manner. This is a highly flexible model for which various complexities can be induced by considering different forms for the predictors. PDMP methods for logistic regression with simple (non-hierarchical) priors tend to be efficient (Bouchard-Côté et al., 2018; Bierkens et al., 2019a), but it is not straightforward to modify these samplers to account for hierarchical structure.

We run the GZZ sampler for 10^7 iterations in all our examples. That is, we run Algorithm 1 for $k = 1, \ldots, 10^7$ attempts. We consider sub-sampling, with a sub-sample of size $n_1 < n$ meaning that n/n_1 iterations of the GZZ sampler corresponds to approximately one epoch of data evaluation,[1] and therefore 10^7 iterations of GZZ corresponds to $n_1/n \times 10^7$ epochs of data evaluation. We compare the GZZ sampler to HMC-within-Gibbs. We run the Gibbs sampler for a total of 10^4 iterations for each setting, which means that we make a total of 10^4 HMC steps as well. For HMC, we consider the leapfrog integrator as described in, for example, Section 2.3 of Neal (2011). Therefore, for L leapfrog steps, we have $L \times 10^4$ epochs of data evaluation for HMC-within-Gibbs.

4.1 Random effects model

Random effects models are routinely applied in a wide variety of disciplines. We consider the following model as illustration,

$$\begin{aligned} Y_{ij} \mid \beta_j &\sim \text{Bernoulli}\left(\frac{1}{1+e^{-\psi_{ij}}}\right), \\ \psi_{ij} &= m + \beta_j + X_{ij}^\top \upsilon, \\ \beta_j &\overset{\text{iid}}{\sim} \text{Normal}(0, \phi^{-1}), \end{aligned} \tag{11}$$

[1] In reality, this is actually less than one epoch as updating the hyperparameter α does not involve the data.

where $j = 1, \ldots, K$ index K groups and $i = 1, \ldots, n$ index n subjects per group,[2] and β_j is the random effect for the jth group. For the ith observation from the jth group, $Y_{ij} \in \{0, 1\}$ denotes the response variable and $X_{ij} = (X_{ij1}, \ldots, X_{ijp}) \in \mathbb{R}^p$ denote covariates. In addition, m denotes an overall intercept, $\upsilon = (\upsilon_1, \ldots, \upsilon_p)$ denotes the fixed effect coefficients, and $X_{ij}^\top \upsilon = \sum_{l=1}^p X_{ijl}\upsilon_l$. We consider the following priors:

$$m \sim \text{Normal}(0, \phi^{-1}), \quad \upsilon_l \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2) \quad (l = 1, \ldots, p),$$
$$\phi \sim \text{Ga}(a_\phi, b_\phi), \quad \sigma^2 \sim \text{IG}(a_\sigma, b_\sigma),$$

where Ga denotes a gamma distribution and IG denotes an inverse-gamma distribution. For this problem, we can use a ZZ process with sub-sampling to update $(\upsilon, m, \beta_1, \ldots, \beta_K)$ conditionally on the hyperparameters (ϕ, σ^2), while the conditional distributions for the hyperparameters can be exactly sampled from; details are provided in Section 4.1 of the supplementary material. In the notation of Section 3.1, we have $\xi = (\upsilon_1, \ldots, \upsilon_p, m, \beta_1, \ldots, \beta_K)$ and $\alpha = (\phi, \sigma^2)$.

We consider synthetic data generated from model (11) with true $(m, \delta, \xi) = (m_{\text{true}}, \delta_{\text{true}}, \xi_{\text{true}}) \in \mathbb{R}^{1+K+p}$. The covariates X_{ijl}s are sampled from the mixture distribution $\varrho_\epsilon(\mathrm{d}x) = \epsilon\delta_0(\mathrm{d}x) + (1 - \epsilon)\rho(\mathrm{d}x)$, where $\delta_0(\mathrm{d}x)$ is a point mass at zero, ρ is a standard normal density, and $\epsilon \in (0, 1]$ denotes the level of sparsity among the covariates.

In a first experiment, we study the effect of the switching rate η on the mixing of process; recall that this is given in equation (5). To this end, we consider a simple setup with $n = 10$, $K = 2$, and $p = 2$, and we also choose $\epsilon = 0.5$. We run the GZZ sampler for various values of the switching rate for mini-batch size ten. We plot the integrated auto-correlation time of the slowest component of ξ in the left panel of Figure 1. The mixing improves to a certain point as the switching rate increases, beyond which the improvement tapers off. In particular, for $\eta \leq 10^{-1}$ the integrated auto-correlation time of the slowest component is approximately proportional to η^{-1}.

In another experiment, we compare the mixing of the process to the size of the mini-batch used. This is shown in the right panel of Figure 1. When the switching rate is low, increasing the mini-batch size does not have a noticeable effect on the mixing of the process. However, when the switching rate is in the "flat" part of the left panel of Figure 1 (that is, $\eta = 6.47$), increasing the mini-batch size has a clear effect on the mixing of the process.

Remark 2. *We obtain the integrated auto-correlation times as follows. We first extract equally-spaced samples from the continuous-time process obtained by running the GZZ sampler. We plot each component of ξ's auto-correlation function and calculate its integrated auto-correlation time by observing when the auto-correlation function converges to zero and then summing the auto-correlation function up to that time. We do not display these plots here, but these can be found in the notebooks accompanying our code.*

Next, we compare the GZZ sampler to HMC-within-Gibbs. We choose $\epsilon = 5 \times 10^{-2}$, which means that the covariates are 95% sparse. For HMC-within-Gibbs, we replace the

[2] We assume that each group is of the same size for simplicity; however, this can easily be extended.

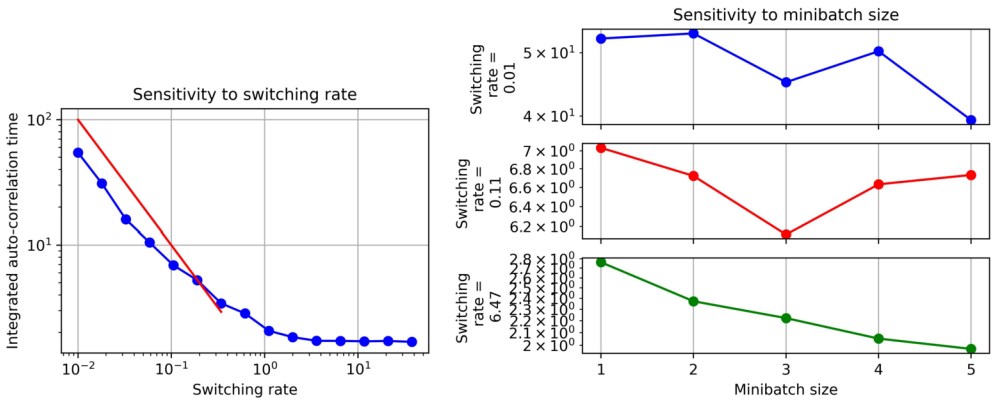

Figure 1: Sensitivity to switching rate η and mini-batch size for the random effects model. The red line in the left plot shows the graph of η^{-1}.

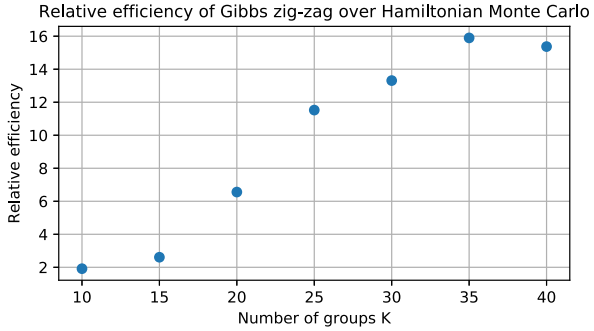

Figure 2: Comparison of effective sample size per epoch of data evaluation for Gibbs zig-zag with sub-sampling and Hamiltonian Monte Carlo for the random effects model.

ZZ updating by HMC. In this case, we choose $n = 100$ and $p = 5$, and vary the number of groups K. As K increases, both the dimension of the sampling problem $(1+K+p)$ as well as the total number of observations $K \times n$ increases. We tune HMC by choosing a range of different leapfrog steps and stepsizes, and looking at cases where the acceptance rate is close to the optimal acceptance rate of 0.651 (Beskos et al., 2013). Among them, we choose the combination of step-size and number of leapfrog steps which gives the highest effective sample size per epoch of data evaluation. We plot the relative effective sample size per epoch of data evaluation for GZZ with sub-sampling divided by the same for HMC in Figure 2, where we observe that the relative performance of using GZZ over HMC increases as the number of groups increases.

Remark 3. *The total number of observations Kn increases as the number of groups K increases, which makes each iteration of HMC slower. However, since GZZ uses sub-sampling, the time per iteration of GZZ remains the same. The number of random*

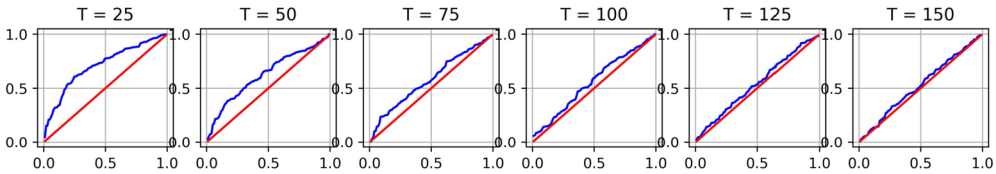

Figure 3: Quantile-Quantile (QQ) plots for the mixed effects model providing empirical evidence for a central limit theorem.

effects increases as the number of groups increases, which means that the dimension of the parameter space increases (if we treat the random effects as "parameters" whose posterior is to be sampled from). While it is true that GZZ performs worse as the dimension increases, this is offset by the slower run time of HMC.

We acknowledge that the assumptions made in order to prove the central limit theorem (Corollary 2), namely Assumptions 1 and 2, do not hold for this example. We nevertheless demonstrate numerically that a central limit theorem does appear to hold for this example. To this end, we choose a simple setting with $K = 5$ groups and $n = 50$ observations per group, and we choose $p = 10$ covariates. We let the test function φ be simply the identity function and estimate $\varphi_{\text{true}} := \mathbb{E}_{(\xi,\alpha,\theta)\sim\tilde{\pi}}\{\varphi(\xi,\alpha,\theta)\}$ by running Algorithm 1 for a very long time. We then choose several different time horizons $T_1 < \cdots < T_M$ and run the GZZ sampler up to each T_m; this can be achieved by running Algorithm 1 till the total time T^k reaches T_m. We run the GZZ sampler independently $R = 2 \times 10^2$ times for each T_m and obtain estimates $\widehat{\varphi}_{T_m,r}$ for $m = 1, \ldots, M$ and $r = 1, \ldots, R$. Corollary 2 implies that $\{T_m^{1/2}(\widehat{\varphi}_{T_m,r} - \varphi_{\text{true}})\}_{r=1}^R$ should converge to samples from a Gaussian distribution with zero mean as m increases. We demonstrate this by noting that if $X \sim \text{Normal}_p(\mu, \Sigma)$, then $(X-\mu)^\top \Sigma^{-1}(X-\mu) \sim \chi_p^2$. This means that $\{[T_m^{1/2}(\widehat{\varphi}_{T_m,r} - \varphi_{\text{true}})]^\top \widehat{\Sigma}_m^{-1}[T_m^{1/2}(\widehat{\varphi}_{T_m,r} - \varphi_{\text{true}})]\}_{r=1}^R$ should converge to samples from a χ_p^2 distribution as m increases, where $\widehat{\Sigma}_m$ is the empirical covariance matrix of $\{T_m^{1/2}[\widehat{\varphi}_{T_m,r} - \varphi_{\text{true}}]\}_{r=1}^R$. We display QQ plots for this in Figure 3, where we observe that a central limit theorem appears to hold in this setting.

4.2 Shrinkage prior

Consider the case where we have p covariates and let $X_j = (X_{j1}, \ldots, X_{jp})$ be the covariates for the jth observation Y_j. Equation (10) then corresponds to a typical logistic regression model with $\psi_j = \upsilon_0 + \sum_{i=1}^p X_{ji}\upsilon_j$, where $\upsilon = (\upsilon_1, \ldots, \upsilon_p)$ are coefficients for the covariates and υ_0 is an intercept term. Even when p is relatively small compared to n, the posterior for υ is not concentrated around a reference point if the covariates are sparse and the prior is isotropic Gaussian. We instead use the GZZ sampler to employ a shrinkage prior for υ. A popular shrinkage prior is the spike-and-slab prior (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005), which is a mixture of a spike at zero and a higher variance component. We consider the following specification of the spike-and-slab

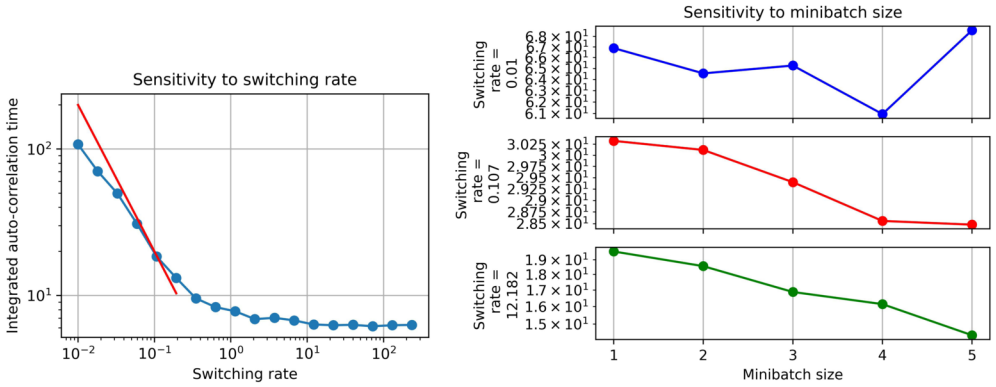

Figure 4: Sensitivity to switching rate η and mini-batch size for logistic regression with spike-and-slab prior. The red line in the left plot shows the graph of η^{-1}.

prior:

$$\begin{aligned}
\upsilon_i &\overset{\text{ind}}{\sim} \gamma_i \,\text{Normal}(0, \nu\tau_i^2) + (1-\gamma_i)\,\text{Normal}(0, \tau_i^2), \\
\gamma_i &\overset{\text{iid}}{\sim} \text{Bernoulli}(\pi) \quad (i = 1, \ldots, p), \\
\nu &\sim \text{IG}(a_\nu, b_\nu), \quad \pi \sim \text{Beta}(a_\pi, b_\pi),
\end{aligned}$$

where $\gamma_i \in \{0, 1\}$ $(i = 1, \ldots, p)$, and we choose $\upsilon_0 \sim \text{Normal}(0, \sigma_0^2)$ for the intercept. Finally, as recommended by Polson and Scott (2012), we choose i.i.d. half-Cauchy priors for the τ_is as $p_0(\tau_i) \propto (1 + \tau_i^2/d_\tau)^{-(d_\tau+1)/2}$ $(i = 1, \ldots, p)$. In terms of the notation of Section 3.1, a ZZ process with sub-sampling can be used to update $\xi = (\upsilon_0, \ldots, \upsilon_p)$ conditionally on the hyperparameters $\alpha = (\gamma_1, \ldots, \gamma_p, \tau^2, \pi, \nu)$, while the conditional distributions for the hyperparameters can be sampled from using MCMC update steps. Details are provided in Section 4.2 of the supplementary material. In contrast to the random effects model of Section 4.1, the dimension of the hyperparameter α is more than twice that of the parameter ξ in this case. We consider synthetic data with the covariates being generated in the same way in Section 4.1. The responses Y_i are sampled from model (10) with "true" $\upsilon = \upsilon_{\text{true}} \in \mathbb{R}^{p+1}$.

In a first experiment, we study the effect of varying mini-batch sizes and varying switching rates η on the efficiency of the GZZ sampler. We choose a simple example with $n = 50$ and $p = 20$, and $\epsilon = 0.4$, and we make the "true" $(\upsilon_1, \ldots, \upsilon_p)$ sparse by setting only 20% of its components to be non-zero. We run the GZZ sampler for various values of the switching rate for mini-batch size ten and plot the integrated auto-correlation time of the slowest component of ξ in the left panel of Figure 4, and look at the sensitivity to the mini-batch size in the right panel of Figure 4. The results are similar to those in Section 4.1. In particular, the mixing improves to a certain point with increasing switching rate, beyond which it tapers off, and for $\eta \leq 10^{-1}$ the integrated auto-correlation time of the slowest component is again approximately proportional to η^{-1}. Increasing the mini-batch size does not have a noticeable effect on the mixing for low switching rates and has a clear effect when the switching rate is sufficiently high.

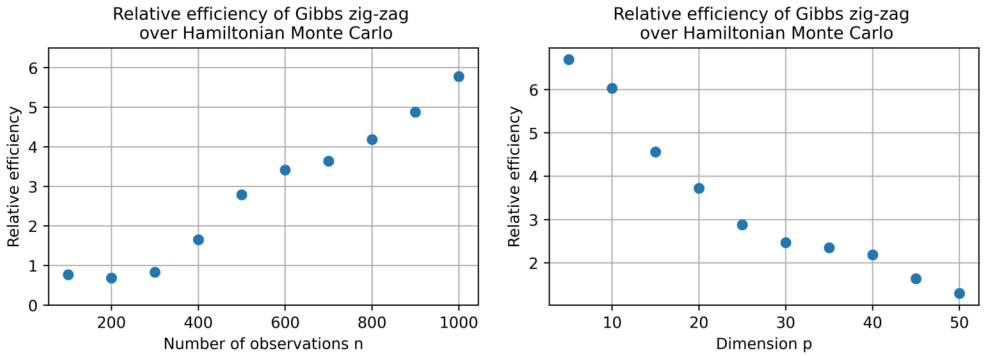

Figure 5: Comparison of effective sample size per epoch of data evaluation for Gibbs zig-zag with sub-sampling and Hamiltonian Monte Carlo while using a shrinkage prior; the left plot is for $p = 10^2$, and the right plot is for $n = 10^2$.

We compare the GZZ sampler to HMC-within-Gibbs. We consider $p = 10^2$ and varying values of n. We also choose the "true" $(\upsilon_1, \ldots, \upsilon_p)$ to be sparse with only 10% of its components being non-zero. For each value of n, we choose ϵ such that $\epsilon \times n$ is fixed at 50. We tune HMC in the same way as in Section 4.1 and compare the effective sample size per epoch of data evaluation of the GZZ sampler with sub-sampling and HMC in the left plot of Figure 5. We observe that as n increases, the GZZ sampler improves upon HMC; this is due to GZZ being faster due to sub-sampling.

We also perform experiments where we fix $n = 10^2$ and consider increasing values of p. The "true" $(\upsilon_1, \ldots, \upsilon_p)$ is again chosen to be sparse with only 10% of its components being non-zero for each value of p. The right plot of Figure 5 displays the effective sample size per epoch of data evaluation of the GZZ sampler with sub-sampling as compared to HMC. We observe that HMC becomes more efficient as compared to GZZ as the dimension p increases.

Finally, in a last set of experiments, we follow the same procedure as in Section 4.1 to emperically demonstrate the validity of a centeral limit theorem; see Figure 6.

4.3 Choice of rate parameter η

In the above reported numerical experiments, we observed that the integrated autocorrelation time of the slowest mixing component is monotonically decreasing, approximately proportional to η^{-1} for sufficiently small η, and approximately constant for large values of η. This observation is consistent with the large-deviation results for similar systems (see Lu and Vanden-Eijnden, 2019) and can be used to derive the following heuristic for the parametrization of η.

Let C_{Gibbs} and C_{ZZ} denote the total computational time incurred for updates of the hyperparameters α and the parameters ξ during a simulation of the GZZ process, respectively. We suggest choosing η such that C_{Gibbs} and C_{ZZ} are of comparable magni-

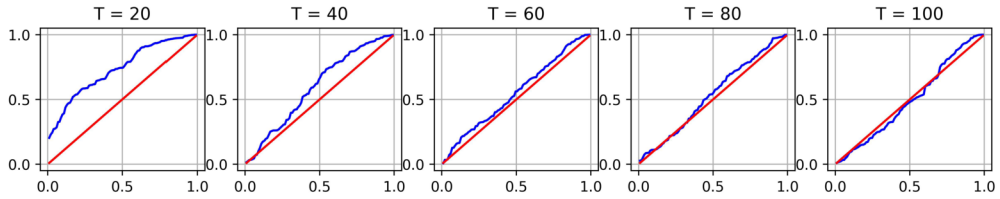

Figure 6: QQ plots for the spike-and-slab prior.

tude (for example, such that $r_{\text{Gibbs}} = C_{\text{Gibbs}}(C_{\text{Gibbs}} + C_{\text{ZZ}})^{-1} \approx 0.2$). This can be easily achieved in practice since $C_{\text{Gibbs}} \propto \eta T$, and thus the computational time associated with Gibbs updates can be easily controlled by varying the value of η.

The motivation behind the heuristic is as follows. Let $\eta_{\text{opt}} > 0$ denote a value of η that results in maximal sampling efficiency (measured in terms of effective sample size of the slowest mixing component per computational cost). If the determined value of η is larger than η_{opt}, then it follows that the loss in sampling efficiency relative to an optimal choice of η is bounded from above by r_{Gibbs}. Otherwise, if the determined value of η is smaller than η_{opt}, then, since the decrease of the integrated auto-correlation time is at most proportional to η^{-1}, a further increase of η would not result in significant increase of sampling efficiency.

5 Discussion

Piecewise deterministic Markov process (PDMP) methods present a promising alternative to traditional (reversible) MCMC algorithms for sampling from posteriors in Bayesian inference. In this paper, we have combined one of the popular PDMPs, the zig-zag process, with Gibbs-like updates. Other variants of the framework that incorporate different PDMPs can be straightforwardly implemented as well (see Section 3 of the supplementary material). PMDP-based sampling schemes have found limited applications in past years, mainly due to the fact that for many sampling problems the construction of suitable tight upper bound is cumbersome, if not impossible. This includes the type of posterior sampling problems with hierarchical priors considered in this article. Thus, the proposed framework contributes to extending the applicability of PDMP-based sampling.

In terms of performance, our approach inherits both positive and negative features of classical PDMP-based approaches. Exact sub-sampling allows for computationally very efficient and asymptotically exact sampling in the presence of large data. However, as also demonstrated for other PDMP sampling methods (see, for example, Quiroz et al., 2021), sampling efficiency in comparison to HMC-within-Gibbs tends to deteriorate as the dimensionality of the sampling problem increases relative to the number of observations.

There are many interesting follow-up directions. While we have focused on PDMP schemes that preserve the exact target distribution, it could be useful to combine Gibbs-

like updates with PDMP schemes that only approximately preserve the target distribution like those in Pakman (2017); Cotter et al. (2020). Theoretically, it would be interesting to study high-dimensional scaling limits of the GZZ process along the lines of Bierkens et al. (2022); Deligiannidis et al. (2021). Moreover, the derivation of η-dependent spectral estimates for the generator of the GZZ process using the Hypocoercivity framework by Dolbeault et al. (2015) (see Andrieu et al. (2021) for an adoption of that framework to PDMPs) would be of interest in order to gain a better understanding of the effect of parameter choices for η on sampling efficiency of the GZZ sampler.

Supplementary Material

Posterior computation with the Gibbs zig-zag sampler. Supplementary material (DOI: 10.1214/22-BA1319SUPP; .pdf).

References

Andrieu, C., Durmus, A., Nüsken, N., and Roussel, J. (2021). "Hypocoercivity of piecewise deterministic Markov process-Monte Carlo." *The Annals of Applied Probability*, 31(5): 2478–2517. MR4332703. doi: https://doi.org/10.1214/20-aap1653. 16

Andrieu, C. and Roberts, G. O. (2009). "The pseudo-marginal approach for efficient Monte Carlo computations." *The Annals of Statistics*, 37(2): 697–725. MR2502648. doi: https://doi.org/10.1214/07-AOS574. 2

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London, London. 1

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). "Optimal tuning of the hybrid Monte Carlo algorithm." *Bernoulli*, 19(5A): 1501–1534. MR3129023. doi: https://doi.org/10.3150/12-BEJ414. 11

Bierkens, J., Bouchard-Côté, A., Doucet, A., Duncan, A. B., Fearnhead, P., Lienart, T., Roberts, G., and Vollmer, S. J. (2018). "Piecewise deterministic Markov processes for scalable Monte Carlo on restricted domains." *Statistics & Probability Letters*, 136: 148–154. MR3806858. doi: https://doi.org/10.1016/j.spl.2018.02.021. 3

Bierkens, J. and Duncan, A. (2017). "Limit theorems for the zig-zag process." *Advances in Applied Probability*, 49(3): 791–825. MR3694318. doi: https://doi.org/10.1017/apr.2017.22. 4

Bierkens, J., Fearnhead, P., and Roberts, G. (2019a). "The zig-zag process and super-efficient sampling for Bayesian analysis of big data." *The Annals of Statistics*, 47(3): 1288–1320. MR3911113. doi: https://doi.org/10.1214/18-AOS1715. 2, 3, 4, 6, 9

Bierkens, J., Roberts, G. O., and Zitt, P.-A. (2019b). "Ergodicity of the zigzag process." *The Annals of Applied Probability*, 29(4): 2266–2301. MR3983339. doi: https://doi.org/10.1214/18-AAP1453. 4, 9

Bierkens, J., Kamatani, K., and Roberts, G. O. (2022). "High-dimensional scaling limits

of piecewise deterministic sampling algorithms." *The Annals of Applied Probability (to appear)*. 16

Bouchard-Côté, A., Vollmer, S. J., and Doucet, A. (2018). "The bouncy particle sampler: A nonreversible rejection-free Markov chain Monte Carlo method." *Journal of the American Statistical Association*, 1–13. MR3832232. doi: https://doi.org/10.1080/01621459.2017.1294075. 2, 6, 9

Chen, T.-L. and Hwang, C.-R. (2013). "Accelerating reversible Markov chains." *Statistics & Probability Letters*, 83(9): 1956–1962. MR3079029. doi: https://doi.org/10.1016/j.spl.2013.05.002. 2

Cotter, S., House, T., and Pagani, F. (2020). "The NuZZ: Numerical ZigZag Sampling for General Models." *arXiv preprint* arXiv:2003.03636. 2, 16

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3): 411–436. MR2278333. doi: https://doi.org/10.1111/j.1467-9868.2006.00553.x. 1

Deligiannidis, G., Paulin, D., Bouchard-Côté, A., and Doucet, A. (2021). "Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates." *The Annals of Applied Probability*, 31(6): 2612–2662. MR4350970. doi: https://doi.org/10.1214/20-aap1659. 16

Diaconis, P., Holmes, S., and Neal, R. M. (2000). "Analysis of a nonreversible Markov chain sampler." *Annals of Applied Probability*, 726–752. MR1789978. doi: https://doi.org/10.1214/aoap/1019487508. 2

Dolbeault, J., Mouhot, C., and Schmeiser, C. (2015). "Hypocoercivity for linear kinetic equations conserving mass." *Transactions of the American Mathematical Society*, 367(6): 3807–3828. MR3324910. doi: https://doi.org/10.1090/S0002-9947-2015-06012-7. 16

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physics Letters B*, 195(2): 216–222. MR3960671. doi: https://doi.org/10.1016/0370-2693(87)91197-x. 1

Fearnhead, P., Bierkens, J., Pollock, M., and Roberts, G. O. (2018). "Piecewise deterministic Markov processes for continuous-time Monte Carlo." *Statistical Science*, 33(3): 386–412. MR3843382. doi: https://doi.org/10.1214/18-STS648. 2

Geman, S. and Geman, D. (1987). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." In *Readings in Computer Vision*, 564–584. Elsevier. 1

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57(1): 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 1

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist

and Bayesian strategies." *The Annals of Statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 12

Jacob, P. E. and Thiery, A. H. (2015). "On nonnegative unbiased estimators." *The Annals of Statistics*, 43(2): 769–784. MR3319143. doi: https://doi.org/10.1214/15-AOS1311. 2

Johndrow, J. E. and Mattingly, J. C. (2017). "Error bounds for Approximations of Markov chains." *arXiv preprint* arXiv:1711.05382. 2

Johndrow, J. E., Mattingly, J. C., Mukherjee, S., and Dunson, D. (2015). "Approximations of Markov chains and high-dimensional Bayesian inference." *arXiv preprint* arXiv:1508.03387. 2

Lewis, P. W. and Shedler, G. S. (1979). "Simulation of nonhomogeneous Poisson processes by thinning." *Naval Research Logistics Quarterly*, 26(3): 403–413. MR0546120. doi: https://doi.org/10.1002/nav.3800260304. 4

Lu, J. and Vanden-Eijnden, E. (2019). "Methodological and computational aspects of parallel tempering methods in the infinite swapping limit." *Journal of Statistical Physics*, 174(3): 715–733. MR3911783. doi: https://doi.org/10.1007/s10955-018-2210-y. 14

Maclaurin, D. and Adams, R. P. (2015). "Firefly Monte Carlo: exact MCMC with subsets of data." In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 4289–4295. 1

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *The Journal of Chemical Physics*, 21(6): 1087–1092. 1

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 12

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics." *Handbook of Markov chain Monte Carlo*, 2(11): 2. MR2858447. 9

Pakman, A. (2017). "Binary Bouncy Particle Sampler." *arXiv preprint* arXiv:1711.00922. 16

Pakman, A., Gilboa, D., Carlson, D., and Paninski, L. (2017). "Stochastic Bouncy Particle Sampler." In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2741–2750. PMLR. 2

Peters, E. A. (2012). "Rejection-free Monte Carlo sampling for general potentials." *Physical Review E*, 85(2): 026703. 2

Pillai, N. S. and Smith, A. (2014). "Ergodicity of approximate MCMC chains with applications to large data sets." *arXiv preprint* arXiv:1405.0182. 2

Polson, N. G. and Scott, J. G. (2012). "On the half-Cauchy prior for a global scale

parameter." *Bayesian Analysis*, 7(4): 887–902. MR3000018. doi: https://doi.org/10.1214/12-BA730. 13

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2018). "Speeding up MCMC by efficient data subsampling." *Journal of the American Statistical Association*, 1–13. MR3963184. doi: https://doi.org/10.1080/01621459.2018.1448827. 1

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2021). "The block-Poisson estimator for optimally tuned exact subsampling MCMC." *Journal of Computational and Graphical Statistics*, 30(4): 877–888. MR4356592. doi: https://doi.org/10.1080/10618600.2021.1917420. 15

Rey-Bellet, L. and Spiliopoulos, K. (2015). "Irreversible Langevin samplers and variance reduction: a large deviations approach." *Nonlinearity*, 28(7): 2081. MR3366637. doi: https://doi.org/10.1088/0951-7715/28/7/2081. 2

Roberts, G. O. and Tweedie, R. L. (1996). "Exponential convergence of Langevin distributions and their discrete approximations." *Bernoulli*, 2(4): 341–363. MR1440273. doi: https://doi.org/10.2307/3318418. 1

Sachs, M., Sen, D., Lu, J., and Dunson, D. (2022). "Posterior computation with the Gibbs zig-zag sampler. Supplementary material." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1319SUPP. 3

Sen, D., Sachs, M., Lu, J., and Dunson, D. B. (2020). "Efficient posterior sampling for high-dimensional imbalanced logistic regression." *Biometrika*, 107(4): 1005–1012. MR4186502. doi: https://doi.org/10.1093/biomet/asaa035. 2, 4

Sun, Y., Schmidhuber, J., and Gomez, F. J. (2010). "Improving the asymptotic performance of Markov chain Monte-Carlo by inserting vortices." In *Advances in Neural Information Processing Systems*, 2235–2243. 2

Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. (2017). "Piecewise Deterministic Markov Chain Monte Carlo." *arXiv preprint* arXiv:1707.05296. 2, 4

Welling, M. and Teh, Y. W. (2011). "Bayesian learning via stochastic gradient Langevin dynamics." In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 681–688. 1

Wu, C. and Robert, C. P. (2020). "Coordinate sampler: a non-reversible Gibbs-like MCMC sampler." *Statistics and Computing*, 30(3): 721–730. MR4065228. doi: https://doi.org/10.1007/s11222-019-09913-w. 6

Zhang, Z., Nishimura, A., Bastide, P., Ji, X., Payne, R. P., Goulder, P., Lemey, P., and Suchard, M. A. (2021). "Large-scale inference of correlation among mixed-type biological traits with phylogenetic multivariate probit models." *The Annals of Applied Statistics*, 15(1): 230–251. MR4255276. doi: https://doi.org/10.1214/20-aoas1394. 6

Zhao, T. and Bouchard-Côté, A. (2021). "Analysis of high-dimensional continuous time Markov chains using the local bouncy particle sampler." *Journal of Machine Learning Research*, 22(91): 1–41. MR4279742. 6