Bayesian Analysis (2014) **9**, Number 1, pp. 97–128

Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms

Theodore Papamarkou [*], Antonietta Mira [†] and Mark Girolami [‡]

Abstract. Differential geometric Markov Chain Monte Carlo (MCMC) strategies exploit the geometry of the target to achieve convergence in fewer MCMC iterations at the cost of increased computing time for each of the iterations. Such computational complexity is regarded as a potential shortcoming of geometric MCMC in practice. This paper suggests that part of the additional computing required by Hamiltonian Monte Carlo and Metropolis adjusted Langevin algorithms produces elements that allow concurrent implementation of the zero variance reduction technique for MCMC estimation. Therefore, zero variance geometric MCMC emerges as an inherently unified sampling scheme, in the sense that variance reduction and geometric exploitation of the parameter space can be performed simultaneously without exceeding the computational requirements posed by the geometric MCMC scheme alone. A MATLAB package is provided, which implements a generic code framework of the combined methodology for a range of models.

Keywords: Metropolis-Hastings, Hamiltonian Monte Carlo, Metropolis adjusted Langevin algorithms, Control variates.

1 Introduction

This paper focuses on evaluating the potential of the combination of two powerful strategies recently published in the Markov Chain Monte Carlo (MCMC) literature, both exploiting information contained in the derivative of the log-target, which we assume to be available in closed form, to increased efficiency: *differential geometric MCMC* algorithms aimed at exploiting the fact that the derivative of the log-target captures the Hamiltonian dynamics on the Riemannian manifold of the parameter space, thus allowing automated and more efficient proposal transitions to be achieved; *zero variance (ZV)* techniques aimed at reducing the variance of the resulting MCMC estimators by post-processing an existing Markov path and by constructing control variates that exploit the well-known fact that the expected value of the derivative of the log-likelihood, here substituted with the log-target, is zero under mild regularity conditions related to the possibility of interchanging the order of differentiation and integration.

There is abundant statistical literature aiming at reducing the asymptotic variance of MCMC estimators. Some of the suggested variance reduction methods introduce antithetic variables in an attempt to induce negative correlation along the chain (e.g. Barone and Frigessi (1990), Craiu and Lemieux (2007)). Other variance reduction tools

[*] Department of Statistical Science, UCL, London, UK, t.papamarkou@ucl.ac.uk

[†] Swiss Finance Institute, University of Lugano, Switzerland, antonietta.mira@usi.ch

[‡] Department of Statistical Science, UCL, London, UK, m.girolami@ucl.ac.uk

DOI:10.1214/13-BA848

for MCMC include Rao-Blackwellization (Gelfand and Smith (1990), Robert and Casella (2004)), hybrid Monte Carlo (Duane et al. (1987), Fort et al. (2003)), use of Riemann sums (Philippe and Robert (2001)) or of auxiliary variables (Swendsen and Wang (1987), Mira and Tierney (2002)), exploitation of non reversible Markov chains (Diaconis et al. (2000), Geyer and Mira (2000)) and data augmentation (Dyk and Meng (2001), Solgi and Mira (2013)). Also, alternative ways of reducing the variance of MCMC estimators try to delay rejection in Metropolis-Hastings based algorithms (Tierney and Mira (1999), Green and Mira (2001)) or to avoid random walk via successive over-relaxation (Adler (1981), Barone et al. (2001)).

Another prominent variance reduction method for Monte Carlo simulation is based on *control variates*, introduced by Ripley (1987). The main challenge to the use of control variates has been their construction. Atchadé and Perron (2005) build control variates for independent Metropolis-Hastings samplers, while Hammer and Tjemeland (2008) introduce control variates for general Metropolis-Hastings samplers.

Andradóttir et al. (1993) observe that the optimum variance reduction of discrete-time finite-state Markov chains can be attained via the solution of the Poisson equation. Furthermore, Henderson (1997) notices that, for any real-valued function G defined on the state space of a Markov chain $\{X_n\}$, the mean of the function $U(x) := G(x) - E[G(X_{n+1}|X_n = x)]$ with respect to the stationary distribution of the chain is zero, therefore $U(x)$ can be utilized as a control variate. Henderson (1997) also notes that the optimal choice of G is the solution of the associated Poisson equation. The most recent development of Henderson's control variates can be found in Dellaportas and Kontoyiannis (2012), in the context of reversible MCMC samplers. A shortcoming of optimal control variates relying on the Poisson equation is that their analytic derivation or their numerical approximation is attainable in very few cases.

In a different context, Assaraf and Caffarel (1999) deploy concepts from statistical mechanics to establish the so called *zero variance* control variates, which drastically reduce the variance of MCMC estimators. Mira et al. (2012) suggest using a polynomial *trial function* for the definition of the ZV-MCMC estimators and derive the conditions for its unbiasedness and for the existence of a central limit theorem.

The trial function proposed by Mira et al. (2012) requires the calculation of the gradient of the log-target density. At the same time, this gradient is computed as part of the *Hamiltonian Monte Carlo (HMC)*, *Riemann manifold Hamiltonian Monte Carlo (RMHMC)*, *Metropolis adjusted Langevin algorithm (MALA)* and of the *manifold Metropolis adjusted Langevin algorithm (MMALA)* (see Girolami and Calderhead (2011), for details). Therefore, saving the gradient of the log-target density alongside the MCMC parameter estimates allows the derivation of the ZV-HMC, ZV-RMHMC, ZV-MALA and ZV-MMALA estimates.

This paper implements these samplers and assesses the achieved variance reduction by performing Bayesian inference on logistic regression models, probit models and on dynamical systems described by non-linear differential equations. The accompanying MATLAB package provides the geometric ZV-MCMC algorithms and it can be downloaded from `http://www.ucl.ac.uk/statistics/research/csi/software`. The package is designed to be used and possibly extended to any Bayesian model for which the log-target density, the metric tensor of the geometric MCMC and their gradients with respect to the model parameters are known in analytic form. Furthermore, routines are made available for the application of geometric ZV-MCMC to any system of differential equations in principle, which are used for modeling dynamical systems in physics, biology, engineering, economics and various other disciplines.

2 Overview of the ZV Method

Interest is in evaluating the expected value of a function $g(\boldsymbol{\theta})$ of the parameters $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ with respect to a possibly unnormalized density $\pi(\boldsymbol{\theta})$:

$$\mu[g(\boldsymbol{\theta})] := E_{\pi(\boldsymbol{\theta})}[g(\boldsymbol{\theta})] = \frac{\int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Using MCMC integration, $\mu[g(\boldsymbol{\theta})]$ is estimated by $\hat{\mu}[g(\boldsymbol{\theta})] = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}_i)$, where the samples $\boldsymbol{\theta}_i \in \mathbb{R}^{n_\theta},\ i = 1, 2, \ldots, n$, are collected along the path of an ergodic Markov chain and are asymptotically distributed according to the target $\pi(\boldsymbol{\theta})/\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta}$.

The ZV method dictates that the original function $g(\boldsymbol{\theta})$ is substituted by an auxiliary function $\tilde{g}(\boldsymbol{\theta})$ with the same mean but smaller variance. $\tilde{g}(\boldsymbol{\theta})$ is constructed by adding to $g(\boldsymbol{\theta})$ a linear combination $\mathbf{a}^T\mathbf{w}(\boldsymbol{\theta}),\ \mathbf{a} \in \mathbb{R}^{n_a}$, of control variates $\mathbf{w} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_a}$:

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \mathbf{a}^T\mathbf{w}(\boldsymbol{\theta}). \tag{1}$$

It is required that $E[\mathbf{w}(\boldsymbol{\theta})] = 0$, in order to acquire the mean equality $\mu[g(\boldsymbol{\theta})] = \mu[\tilde{g}(\boldsymbol{\theta})]$.

Assaraf and Caffarel (1999) suggest defining the auxiliary function $\tilde{g}(\boldsymbol{\theta})$ as

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \frac{H[\psi(\boldsymbol{\theta})]}{\sqrt{\pi(\boldsymbol{\theta})}}, \tag{2}$$

where $H[\psi(\boldsymbol{\theta})]$ denotes the Schrödinger Hamiltonian given by

$$H[\psi(\boldsymbol{\theta})] = -\frac{1}{2}\Delta_{\boldsymbol{\theta}}[\psi(\boldsymbol{\theta})] + \frac{\psi(\boldsymbol{\theta})}{2\sqrt{\pi(\boldsymbol{\theta})}}\Delta_{\boldsymbol{\theta}}[\sqrt{\pi(\boldsymbol{\theta})}], \tag{3}$$

$\Delta_{\boldsymbol{\theta}} := \sum_{i=1}^{n_\theta} \partial^2/\partial\theta_i^2$ is the Laplace operator and $\psi(\boldsymbol{\theta})$ is the so-called trial function, which is an arbitrary integrable function. As it is explained in Assaraf and Caffarel (1999), the condition $E[H[\psi(\boldsymbol{\theta})]/\sqrt{\pi(\boldsymbol{\theta})}] = 0$ is satisfied for the Hamiltonian defined in (3) for any integrable function $\psi(\boldsymbol{\theta})$ thus ensuring that $E\left[\mathbf{w}(\boldsymbol{\theta})\right] = 0$ by construction so that $\hat{\mu}\left[\tilde{g}(\boldsymbol{\theta})\right]$ is an asymptotically unbiased estimator of $\mu\left[g(\boldsymbol{\theta})\right]$.

2.1 Choice of $\psi(\boldsymbol{\theta})$ Based on Polynomials

Mira et al. (2012) propose selecting a polynomial for the trial function $\psi(\boldsymbol{\theta})$. Increasing the degree of the polynomial induces more control variates, yet more importantly attenuates the variance of the parameter estimators. In practice, first and second degree polynomials suffice to attain considerable variance reduction, an argument quantified later in Sections 4, 5 and 6.

Along the lines of Mira et al. (2012), define the trial function to be

$$\psi(\boldsymbol{\theta}) = P(\boldsymbol{\theta})\sqrt{\pi(\boldsymbol{\theta})}, \tag{4}$$

where $P(\boldsymbol{\theta})$ is assumed to be a polynomial. It then follows from (2) and (4) that the auxiliary function $\tilde{g}(\boldsymbol{\theta})$ and the control variates $\mathbf{z}(\boldsymbol{\theta})$ are given by

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) - \frac{1}{2}\Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] \cdot \mathbf{z}(\boldsymbol{\theta}), \tag{5}$$

$$\mathbf{z}(\boldsymbol{\theta}) := -\frac{1}{2}\nabla_{\boldsymbol{\theta}}[\ln(\pi(\boldsymbol{\theta}))], \tag{6}$$

where $\nabla_{\boldsymbol{\theta}} := \left(\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_{n_\theta}}\right)$ denotes the gradient.

First Degree Polynomial $P(\boldsymbol{\theta})$

Let $P(\boldsymbol{\theta})$ be a linear polynomial

$$P(\boldsymbol{\theta}) = \mathbf{a}^T\boldsymbol{\theta}, \tag{7}$$

where $\mathbf{a} \in \mathbb{R}^{n_a},\ n_a = n_\theta$. The gradient and Laplace operators for linear $P(\boldsymbol{\theta})$ equal

$$\nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] = \mathbf{a}^T,\ \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] = 0. \tag{8}$$

It follows from (5) and (8) that the auxiliary function, for a first degree polynomial $P(\boldsymbol{\theta})$, takes the form

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \mathbf{a}^T\mathbf{z}(\boldsymbol{\theta}). \tag{9}$$

The optimal polynomial coefficients $\mathbf{a}$, which minimize the variance of the auxiliary function $\tilde{g}(\boldsymbol{\theta})$, are given by

$$\mathbf{a} = -Var^{-1}[\mathbf{z}(\boldsymbol{\theta})]Cov[g(\boldsymbol{\theta}), \mathbf{z}(\boldsymbol{\theta})], \tag{10}$$

where $Var[\mathbf{z}(\boldsymbol{\theta})]$ and $Cov[g(\boldsymbol{\theta}), \mathbf{z}(\boldsymbol{\theta})]$ denote the respective variance and cross-covariance matrices

$$Var[\mathbf{z}(\boldsymbol{\theta})] := E_{\pi(\boldsymbol{\theta})}[(\mathbf{z}(\boldsymbol{\theta}) - E[\mathbf{z}(\boldsymbol{\theta})])(\mathbf{z}(\boldsymbol{\theta}) - E[\mathbf{z}(\boldsymbol{\theta})^T], \tag{11}$$

$$Cov[g(\boldsymbol{\theta}), \mathbf{z}(\boldsymbol{\theta})] := E_{\pi(\boldsymbol{\theta})}[(g(\boldsymbol{\theta}) - E[g(\boldsymbol{\theta})])(\mathbf{z}(\boldsymbol{\theta}) - E[\mathbf{z}(\boldsymbol{\theta})])]. \tag{12}$$

So, the MCMC coefficient estimators $\hat{\mathbf{a}}$ are given by

$$\hat{\mathbf{a}} = -\widehat{Var}^{-1}[\mathbf{z}(\boldsymbol{\theta})]\widehat{Cov}[g(\boldsymbol{\theta}), \mathbf{z}(\boldsymbol{\theta})], \tag{13}$$

where $\widehat{Var}[\mathbf{z}(\boldsymbol{\theta})]$ and $\widehat{Cov}[g(\boldsymbol{\theta}), \mathbf{z}(\boldsymbol{\theta})]$ are, respectively, the sample variance and sample cross-covariance matrices

$$\widehat{Var}[\mathbf{z}(\boldsymbol{\theta})] := \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}(\boldsymbol{\theta}_i) - \hat{\mu}[\mathbf{z}(\boldsymbol{\theta})])(\mathbf{z}(\boldsymbol{\theta}_i) - \hat{\mu}[\mathbf{z}(\boldsymbol{\theta})])^T, \tag{14}$$

$$\widehat{Cov}[g(\boldsymbol{\theta}), \mathbf{z}(\boldsymbol{\theta})] := \frac{1}{n-1}\sum_{i=1}^{n}(g(\boldsymbol{\theta}_i) - \hat{\mu}[g(\boldsymbol{\theta})])(\mathbf{z}(\boldsymbol{\theta}_i) - \hat{\mu}[\mathbf{z}(\boldsymbol{\theta})]), \tag{15}$$

where n is the number of post burn-in MCMC samples and the sample means are defined as $\hat{\mu}[\mathbf{z}(\boldsymbol{\theta})] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}(\boldsymbol{\theta}_i),\ \hat{\mu}[g(\boldsymbol{\theta})] = \frac{1}{n}\sum_{i=1}^{n}g(\boldsymbol{\theta}_i)$.

Note that the theoretical mean $E[z(\boldsymbol{\theta})]$ of $\mathbf{z}(\boldsymbol{\theta})$ in (11) and (12) is zero and the corresponding mean estimator $\hat{\mu}[g(\boldsymbol{\theta})]$ calculated from the MCMC output is nearly zero. Although mean-adjusting or not does not affect the ZV values nor the reduction factors, the sample variances and cross-covariances in Equations (14) and (15) are mean-adjusted to agree with the conventional statistical definition of covariance matrix.

Second Degree Polynomial $P(\boldsymbol{\theta})$

Assume now that $P(\boldsymbol{\theta})$ is a second order polynomial

$$P(\boldsymbol{\theta}) = \mathbf{c}^T\boldsymbol{\theta} + \frac{1}{2}\boldsymbol{\theta}^T B\boldsymbol{\theta}, \tag{16}$$

where $\mathbf{c}$ is a real $n_\theta \cdot 1$ vector and B is a real symmetric $n_\theta \cdot n_\theta$ matrix. The gradient and Laplace operators for quadratic $P(\boldsymbol{\theta})$ evaluate to

$$\nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] = (\mathbf{c} + B\boldsymbol{\theta})^T,\ \Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] = tr(B), \tag{17}$$

where $tr(B)$ denotes the trace of B. According to (5) and (17), the auxiliary function $\tilde{g}(\boldsymbol{\theta})$, for a second order polynomial $P(\boldsymbol{\theta})$, reduces to

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) - \frac{1}{2}tr(B) + (\mathbf{c} + B\boldsymbol{\theta})^T\mathbf{z}(\boldsymbol{\theta}), \tag{18}$$

where $\mathbf{z}(\boldsymbol{\theta})$ is given by (6). In order to conform with (1), where the auxiliary function $\tilde{g}(\boldsymbol{\theta})$ expresses as a linear combination of the original function $g(\boldsymbol{\theta})$ and of the control variates, (18) can be written as

$$\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \mathbf{a}^T\mathbf{w}(\boldsymbol{\theta}), \tag{19}$$

where the column vectors $\mathbf{a}$, $\mathbf{w}(\boldsymbol{\theta})$ have $n_\theta(n_\theta + 3)/2$ elements each, and are defined as

- $\mathbf{a} := [\mathbf{c}^T\ \mathbf{d}^T\ \mathbf{b}^T]^T$, where $\mathbf{d} := diag(B)$ is the diagonal of B and $\mathbf{b}$ is a column vector with $n_\theta(n_\theta - 1)/2$ elements, whose element in the $(2n_\theta - j)(j-1)/2 + (i-j)$ position is the lower diagonal (i, j)-th element of B.
- $\mathbf{w} := [\mathbf{z}^T\ \mathbf{u}^T\ \mathbf{v}^T]^T$, where $\mathbf{u} := \boldsymbol{\theta} \circ \mathbf{z} - \frac{1}{2}\mathbf{1}$, with $\circ$, $\mathbf{1}$ denoting the Hadamard product and the unit vector respectively, while $\mathbf{v}$ is a column vector comprising $n_\theta(n_\theta - 1)/2$ elements, whose element in the $(2n_\theta - j)(j-1)/2 + (i-j)$ position equals $\theta_i z_j + \theta_j z_i,\ j \in \{1, 2, \ldots, n_\theta\},\ i \in \{2, 3, \ldots, n_\theta\},\ j < i$.

Note that the number of polynomial coefficients for quadratic $P(\boldsymbol{\theta})$ is $n_a = n_\theta(n_\theta+3)/2$, in contrast to linear $P(\boldsymbol{\theta})$ where $n_a = n_\theta$. As for the estimation of the coefficients $\mathbf{a}$ for quadratic $P(\boldsymbol{\theta})$, formulae (10)-(15) apply by using $\mathbf{w} := [\mathbf{z}^T \ \mathbf{u}^T \ \mathbf{v}^T]^T$.

2.2 Expression of z for posterior densities

Assume that a posterior density $p(\boldsymbol{\theta}|\mathbf{x})$ is chosen in place of the unnormalized density (denoted by $\pi(\boldsymbol{\theta})$ before):

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{c}\ell(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$, $\mathbf{x}$ denote the model parameters and the data respectively, while ℓ and π represent the likelihood and the prior respectively. It is then straightforward to confirm that $\mathbf{z}(\boldsymbol{\theta})$ in (6) is half the negative sum of gradients of the log-likelihood, L, and of the log-prior:

$$\mathbf{z}(\boldsymbol{\theta}) = -\frac{1}{2}\nabla_{\boldsymbol{\theta}}[\ln(p(\boldsymbol{\theta}|\mathbf{x}))] = -\frac{1}{2}(\nabla_{\boldsymbol{\theta}}[L(\mathbf{x}|\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[\ln(\pi(\boldsymbol{\theta}))]). \tag{20}$$

2.3 Variance Reduction Factor

To quantify the variance reduction achieved by ZV, n_c independent chains are realized, each of length n. Since in the sequel we focus on estimating the posterior mean, the auxiliary function g in (1) is taken to be the identity. Therefore, the subsequent formulae for the definition of the variance reduction factor assume g to be the identity, and should be modified for posterior estimators other than the posterior mean. Let $\theta_{ijk},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n_\theta,\ k = 1, 2, \ldots, n_c$ be the i-th MCMC iteration of the j-th model parameter in the k-th chain and $\tilde{\theta}_{ijk}$ its corresponding ZV value. The asymptotic variance of the j-th parameter along the k-th chain is estimated using the *initial monotone sequence estimator*

$$\hat{\sigma}^2_{jk} := \frac{1}{n}\left(-\hat{\gamma}_0 + 2\sum_{i=0}^{m}\hat{\Gamma}_{ijk}\right),$$

$$\hat{\Gamma}_{ijk} := \hat{\gamma}_{ijk} + \hat{\gamma}_{(2i+1)jk},$$

where m denotes the largest integer such that $\hat{\Gamma}_{ijk} > 0,\ i = 0, 1, 2, \ldots, m$, and $\hat{\gamma}_{ijk}$ is the empirical estimator of the $lag(i)$ autocovariance γ_{ijk} of the j-th parameter in the k-th chain, given by

$$\hat{\gamma}_{ijk} := \frac{1}{n}\sum_{q=1}^{n-i}(\theta_{qjk} - \bar{\theta}_{.jk})(\theta_{(q+i)jk} - \bar{\theta}_{.jk}),$$

where the mean of the j-th parameter in the k-th chain calculates as $\bar{\theta}_{.jk} = \sum_{i=1}^{n}\theta_{ijk}/n$. The estimated sequence $\{\hat{\Gamma}_{ijk} : i = 0, 1, 2, \ldots, m\}$ is made monotone by reducing $\hat{\Gamma}_{ijk}$ to the minimum of the preceding sequence members, see Geyer (1992) for details. The overall variance estimator $\hat{\sigma}_j^2$ of the j-th parameter across the n_c simulated chains is obtained by taking the mean of the initial monotone sequence estimators, that is $\hat{\sigma}_j^2 = \sum_{k=1}^{n_c} \hat{\sigma}_{jk}^2 / n_c$.

Furthermore, each of the $n_\theta \cdot n_c$ coefficients $\mathbf{a}_{,jk}$ is calculated separately using the corresponding chain $\{\theta_{ijk}\}$ from which $\hat{\sigma}_{jk}^2$ is also computed. This way, the ZV values $\tilde{\theta}_{ijk}$ are obtained, whence the initial monotone estimator $\hat{\tilde{\sigma}}_{jk}^2$ for the j-th parameter in the k-th chain $\{\tilde{\theta}_{ijk}\}$ is deduced. Consequently, the variance estimator of the ZV values for the j-th parameter is analogously estimated across the n_c chains by the mean $\hat{\tilde{\sigma}}_j^2 = \sum_{k=1}^{n_c} \hat{\tilde{\sigma}}_{jk}^2 / n_c$.

Thus, the *variance reduction factor (VRF)* is naturally defined as the ratio of the asymptotic variance estimators $\hat{\sigma}_j^2$ and $\hat{\tilde{\sigma}}_j^2$:

$$r_j := \frac{\hat{\sigma}_j^2}{\hat{\tilde{\sigma}}_j^2} = \frac{\displaystyle\sum_{k=1}^{n_c} \hat{\sigma}_{jk}^2}{\displaystyle\sum_{k=1}^{n_c} \hat{\tilde{\sigma}}_{jk}^2},\ j = 1, 2, \ldots, n_\theta. \tag{21}$$

3 ZV-(RM)HMC and ZV-(M)MALA

It becomes apparent from (9), (19) and (20) that the main required component for the Bayesian application of ZV-MCMC is the gradient of the log-posterior density.

The current section outlines how the gradient of the log-target is readily available in the case of ZV-RMHMC, ZV-HMC, ZV-MMALA and ZV-MALA. Some familiarity with RMHMC and MMALA is presumed. The interested reader is referred to Girolami and Calderhead (2011) for an elaborate study of these two geometric MCMC samplers.

3.1 ZV-RMHMC and ZV-HMC

Riemann manifold Hamiltonian Monte Carlo is a Gibbs sampling scheme which defines a Hamiltonian on the Riemann manifold of probability densities as

$$H(\boldsymbol{\theta}, \mathbf{p}) := -\ln(p(\boldsymbol{\theta}|\mathbf{x})) + \frac{1}{2}\ln[(2\pi)^{n_\theta}|G(\boldsymbol{\theta})|] + \frac{1}{2}\mathbf{p}^T G^{-1}(\boldsymbol{\theta})\mathbf{p}, \tag{22}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, $\mathbf{x}$, $G(\boldsymbol{\theta})$ and $\mathbf{p}$ denote respectively the model parameters, the data, a metric tensor and the auxiliary variables $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, G(\boldsymbol{\theta}))$. The position-specific metric tensor $G(\boldsymbol{\theta})$ allows for effective transitions in RMHMC and is chosen to be the expected Fisher information. The auxiliary variables $\mathbf{p}$ and the terms $-\ln(p(\boldsymbol{\theta}|\mathbf{x})) + \frac{1}{2}\ln[(2\pi)^{n_\theta}|G(\boldsymbol{\theta})|]$ and $\frac{1}{2}\mathbf{p}^T G(\boldsymbol{\theta})^{-1}\mathbf{p}$ are interpreted as the momentum, the kinetic energy and the potential energy at a particular position $\boldsymbol{\theta}$, respectively.

It follows from (22) that Hamilton's equation for RMHMC takes the form

$$\frac{\partial H(\boldsymbol{\theta}, \mathbf{p})}{\partial \theta_i} = 2z_i(\boldsymbol{\theta}) + \frac{1}{2} tr \left[G^{-1}(\boldsymbol{\theta}) \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_i} \right] - \frac{1}{2} \mathbf{p}^T G^{-1}(\boldsymbol{\theta}) \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_i} G^{-1}(\boldsymbol{\theta}) \mathbf{p}, \qquad (23)$$

for $i \in \{1, 2, \ldots, n_\theta\}$, where $z_i(\boldsymbol{\theta})$ is the i-th element of $\mathbf{z}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}}(\ln(p(\boldsymbol{\theta}|\mathbf{x})))/2$. Thus, $\mathbf{z}(\boldsymbol{\theta})$ is computed at each iteration of RMHMC as an intermediate result. So ZV-RMHMC, for the linear and quadratic polynomials proposed by Mira et al. (2012), requires only to save $\mathbf{z}(\boldsymbol{\theta})$ at each iteration of the Gibbs sampler.

Hamiltonian Monte Carlo (HMC) can be viewed as a simplified version of RMHMC, where the metric tensor $G(\boldsymbol{\theta})$ is substituted by the constant mass matrix M. This means that the momentum has the simpler covariance matrix M, as per $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, M)$. Then $\mathbf{z}(\boldsymbol{\theta})$ is given directly by Hamilton's equation according to $2\mathbf{z}(\boldsymbol{\theta}) = \partial H(\boldsymbol{\theta}, \mathbf{p})/\partial \boldsymbol{\theta}$.

3.2 ZV-MMALA, ZV-SMMALA and ZV-MALA

MMALA defines a Langevin diffusion with invariant distribution $p(\boldsymbol{\theta}|\mathbf{x})$, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, on the Riemann manifold of probability densities with metric tensor $G(\boldsymbol{\theta})$. A first order Euler integrator is used for solving the diffusion. This Euler approximation induces some discretization error, so a Metropolis-Hastings step is employed to account for the associated bias. The proposal density and the standard acceptance probability are defined to be

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k) = \mathcal{N}(\boldsymbol{\theta}^*|\boldsymbol{\mu}(\boldsymbol{\theta}^k, \epsilon), \epsilon^2 G^{-1}(\boldsymbol{\theta}^k)),$$
$$\min\{1, p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*)/p(\boldsymbol{\theta}^k)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^k)\},$$

where ϵ is the integration step size and the ℓ-th coordinate $\boldsymbol{\mu}_\ell(\boldsymbol{\theta}^k, \epsilon)$ of the mean is

$$\boldsymbol{\mu}_\ell(\boldsymbol{\theta}^k, \epsilon) = \boldsymbol{\theta}_\ell - \epsilon^2 (G^{-1}(\boldsymbol{\theta})\mathbf{z}(\boldsymbol{\theta}))_\ell - \epsilon^2 \sum_{i=1}^{n_\theta} \sum_{j=1}^{n_\theta} G_{ij}^{-1}(\boldsymbol{\theta}) \boldsymbol{\Gamma}_{ij}^\ell. \qquad (24)$$

$\boldsymbol{\Gamma}_{ij}^\ell$ are the Christoffel symbols of the second kind in local coordinates and $\mathbf{z}(\boldsymbol{\theta})$ is minus half the gradient of the log-target, that is $\mathbf{z}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}}(\ln(p(\boldsymbol{\theta}|\mathbf{x})))/2$. It thus becomes apparent that ZV-MMALA requires only to store $\mathbf{z}(\boldsymbol{\theta})$ at each iteration of the Metropolis-Hastings sampler.

Simplified MMALA (SMMALA) assumes a manifold of constant curvature, which means that the Christoffel symbols are zero, thereby the last term in (24) vanishes. A further simplification would be to select a constant metric tensor $G(\boldsymbol{\theta}) = M$, in which case MMALA coincides with the *Metropolis adjusted Langevin algorithm (MALA)* with preconditioning matrix M (see Roberts and Stramer (2002), for details).

4 ZV-MCMC for Bayesian Logistic Regression

Consider a Bayesian logistic regression model (see for example Gelman et al. (2004)). Let X be the $n_d \cdot n_\theta$ design matrix of the model consisting of n_d samples, each with

n_θ covariates, $y \in \{0,1\}^{n_d}$ the binary response variable and $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ the regression coefficients. The log-likelihood of the model is given by

$$L(\mathbf{y}|X,\boldsymbol{\theta}) = (X\boldsymbol{\theta})^T\mathbf{y} - \sum_{i=1}^{n_d} \ln\left[1+\exp(\boldsymbol{\theta}^T\mathbf{x}_{\mathbf{i},})\right], \tag{25}$$

where $\mathbf{x}_{\mathbf{i},}$ denotes the i-th row of the design matrix X. A Normal prior is assumed for the model parameters $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, aI)$, where a is a positive real number. $a = 100$ is chosen in the example of Section 4.2, which implies a rather flat prior.

4.1 ZV Estimates and Metric Tensor

Differentiating (25) yields the gradient of the log-likelihood

$$\nabla_{\boldsymbol{\theta}}(L(\mathbf{y}, X|\boldsymbol{\theta})) = X^T\left[\mathbf{y} - \frac{1}{1+\exp(-X\boldsymbol{\theta})}\right]. \tag{26}$$

The gradient of the log-prior evaluates to

$$\nabla_{\boldsymbol{\theta}}(\ln(\pi(\boldsymbol{\theta}))) = -\frac{1}{a}\boldsymbol{\theta}. \tag{27}$$

(20), (26) and (27) give the gradient of the log-posterior and consequently $\mathbf{z}(\boldsymbol{\theta})$:

$$\mathbf{z}(\boldsymbol{\theta}) = -\frac{1}{2}\nabla_{\boldsymbol{\theta}}(\ln(p(\boldsymbol{\theta}|\mathbf{x}))) = -\frac{1}{2}\left\{X^T\left[\mathbf{y} - \frac{1}{1+\exp(-X\boldsymbol{\theta})}\right] - \frac{1}{a}\boldsymbol{\theta}\right\}. \tag{28}$$

To frame the parameter estimation in a Monte Carlo context, consider n post burn-in MCMC samples for each of the n_θ regression coefficients. This implies introducing $n\cdot n_\theta$ original functions of interest $g_{ij}(\boldsymbol{\theta}) := \theta_{ij},\ i = 1,2,\ldots,n,\ j = 1,2,\ldots,n_\theta$, and the corresponding auxiliary functions $\tilde{g}_{ij}(\boldsymbol{\theta}) := \tilde{\theta}_{ij}$, where θ_{ij} is the i-th MCMC iteration of the j-th coefficient and $\tilde{\theta}_{ij}$ denotes the ZV counterpart of θ_{ij}. Hence, for linear polynomial P, the $n \cdot n_\theta$ realizations of (9) express as

$$\tilde{\theta}_{ij} = \theta_{ij} + \mathbf{a}_{,j}^T \cdot \mathbf{z}_{i,},\ i = 1,2,\ldots,n,\ j = 1,2,\ldots,n_\theta, \tag{29}$$

where the n_θ-dimensional row vector $\mathbf{z}_{i,}$ for the i-th MCMC iteration is given by (28) and the n_θ-dimensional column vector $\mathbf{a}_{,j}$ holds the coefficients of the linear polynomial of Equation (7) for the j-th regression coefficient. Using matrix notation, (29) can be written more concisely as

$$\tilde{\Theta} = \Theta + ZA, \tag{30}$$

where Θ, $\tilde{\Theta}$ are the respective $n \cdot n_\theta$ matrices of MCMC and ZV-MCMC estimates, Z is the $n \cdot n_\theta$ matrix whose i-th row is the vector $\mathbf{z}_{i,}$ appearing in (29) and A is the $n_\theta \cdot n_\theta$ matrix of polynomial coefficients whose j-th column is the vector $\mathbf{a}_{,j}$ of (29).

The estimates $\hat{\mathbf{a}}_{,j},\ j = 1, 2, \ldots, n_\theta$, of the coefficients of the linear polynomial for the j-th regression coefficient are deduced from (13):

$$\hat{\mathbf{a}}_{,j}^T = -\widehat{Var}(\mathbf{z})^{-1}\widehat{Cov}(\theta_j, \mathbf{z}), \tag{31}$$

$$\widehat{Var}(\mathbf{z}) := \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_{i,} - \bar{\mathbf{z}}_{.,})(\mathbf{z}_{i,} - \bar{\mathbf{z}}_{.,})^T, \tag{32}$$

$$\widehat{Cov}(\theta_j, \mathbf{z}) := \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})(\mathbf{z}_{i,} - \bar{\mathbf{z}}_{.,}), \tag{33}$$

where $\bar{\mathbf{z}}_{.,} := \sum_{i=1}^{n}\mathbf{z}_{i,}/n$ and $\bar{\theta}_{.j} := \sum_{i=1}^{n}\theta_{ij}/n$.

In a similar way, (16) takes the following matrix form for the derivation of the ZV estimates for quadratic polynomials:

$$\tilde{\Theta} = \Theta + WA, \tag{34}$$

where Θ, $\tilde{\Theta}$ denote the respective $n \cdot n_\theta$ matrices of MCMC and ZV-MCMC estimates, W is the $n \cdot n_a$ matrix, $n_a = n_\theta(n_\theta + 3)/2$, whose i-th row is the n_a-dimensional vector $\mathbf{w}_{i,}$ defined in Section 2.1 and A is the $n_a \cdot n_\theta$ matrix of polynomial coefficients whose j-th column is the n_a-dimensional vector $\mathbf{a}_{,j}$ of Section 2.1 holding the coefficients of the quadratic polynomial (16) for the j-th regression coefficient. The matrix of polynomial coefficients A for quadratic polynomials is estimated analogously to (31)-(33), using $\mathbf{w}_{i,} - \bar{\mathbf{w}}_{.,}$ in place of $\mathbf{z}_{i,} - \bar{\mathbf{z}}_{.,}$.

To implement RMHMC and MMALA, the metric tensor $G(\boldsymbol{\theta})$ and its derivatives $dG(\boldsymbol{\theta})/d\theta_j,\ j = 1, 2, \ldots, n_\theta$ are also required. For Bayesian inference $G(\boldsymbol{\theta})$ is taken to be the expected Fisher information minus the Hessian of the log-prior, so as to incorporate any prior information when exploiting the local curvature of the manifold. It can be shown that for the Bayesian logistic regression model with a Normal prior $\mathcal{N}(\mathbf{0}, aI)$ the metric tensor and its derivatives are given by

$$G(\boldsymbol{\theta}) = X^T\Lambda(\boldsymbol{\theta})X + \frac{1}{a}I,\ \frac{\partial G(\boldsymbol{\theta})}{\partial\theta_j} = X^T M^j(\boldsymbol{\theta})X, \tag{35}$$

where the (i, i)-th elements of the $n_d \cdot n_d$ diagonal matrices $\Lambda(\boldsymbol{\theta})$ and $M^j(\boldsymbol{\theta})$ are

$$\Lambda(\boldsymbol{\theta}) := diag[p_i(1 - p_i)],\ M^j(\boldsymbol{\theta}) := \frac{\partial\Lambda(\boldsymbol{\theta})}{\partial\theta_j} = diag[p_i(1 - p_i)(1 - 2p_i)x_{ij}], \tag{36}$$

where $p_i := P(y_i = 1) = \exp(\boldsymbol{\theta}^T\mathbf{x}_{\mathbf{i},})/(1 + \exp(\boldsymbol{\theta}^T\mathbf{x}_{\mathbf{i},})),\ i = 1, 2, \ldots, n_d$, and $x_{ij},\ \mathbf{x}_{\mathbf{i},}$ denote the (i, j)-th element and the i-th row of the design matrix X, respectively.

4.2 Swiss Banknotes Example

The Bayesian logistic regression model, as outlined in the current section, is employed to run ZV-MALA, ZV-SMMALA, ZV-MMALA, ZV-HMC and ZV-RMHMC on the bank dataset taken from Flury and Riedwyl (1988). *ZV-Metropolis-Hastings (ZV-MH)* is also run as a standard benchmark. The dataset holds the measurements of four covariates on 200 Swiss banknotes, of which 100 are genuine and 100 counterfeit, representing the length of the bill, the width of the left and the right edge, and the bottom margin width. Therefore, the design matrix X has dimensions $n_d \cdot n_\theta = 200 \cdot 4$ and the $n_\theta = 4$ regression coefficients constitute the model parameters to be estimated. The binary response variable is the type of banknote, 0 being genuine and 1 counterfeit.

$55,000$ iterations are run for the realization of each chain, of which the first $5,000$ burn-in are discarded, so $n = 50,000$ MCMC samples are retained from each chain. Figure R.1 serves as a first visual demonstration of the effectiveness of ZV. The traces of ordinary RMHMC, θ, are overlaid with the ZV-RMHMC traces, $\tilde{\theta}$, for first and second order polynomials P on the left and right columns of Figure R.1, respectively. As it can be seen the ZV traces exhibit negligible variability, especially the ones based on quadratic P.

Figure R.2 (see Appendix) displays a graphical comparison between the original and the ZV estimates for linear and quadratic polynomials for each parameter and each MCMC scheme. The mean of the boxplots of Figure R.2 remains unaltered between the MCMC and the corresponding ZV-MCMC samples, agreeing with the theory of control variates. At the same time, the variance diminishes substantially among the ZV estimators.

For a more systematic assessment of ZV, Table R.1 (see Appendix) provides the variance reduction factors (VRFs), defined by (21). All subsequent boxplots as well as tables with the variance reduction factors are available in the Appendix. A total of $n_c = 100$ chains are run for each of the six MCMC algorithms to obtain Monte Carlo estimates of the asymptotic variances and consequently of the variance reduction factors according to the procedure outlined in Section 2.3 (see Equation (21)).

Table R.1 shows that the VRF between the ZV estimates for linear polynomials and the ordinary estimates range roughly in the region of $10 - 50$. As for the ZV estimates for quadratic polynomials and the original estimates, they differ by three orders of magnitude, that is they differ by a variance reduction factor of about $1000 - 8000$. It is also apparent in Table R.1 that the more effective the MCMC algorithm the smaller the VRF. To quantify this argument, let the effectiveness of the MCMC algorithm be measured by its *effective sample size (ESS)*, with larger ESS obviously indicating a more effective sampler. So, larger ESS results in smaller asymptotic variance and consequently the ZV strategy induces smaller variance in the already effective sampler. For example, RMHMC usually has the largest ESS among the six compared MCMC schemes (see Girolami and Calderhead (2011)) and therefore the smallest asymptotic variance, which is confirmed by Table R.1 for all four parameters of the logistic regression model. Although the corresponding asymptotic variance of the ZV estimator for linear and quadratic polynomials is the smallest in the case of RMHMC, the algorithm is more effective than the remaining five, thus the attained reduction factor is the smallest. Furthermore, Table R.1 shows that the least effective Metropolis-Hastings (MH) algorithm has the highest asymptotic variance, followed by MALA, SMMALA, MMALA, HMC and RMHMC. Along these lines, Table R.1 provides circumstantial

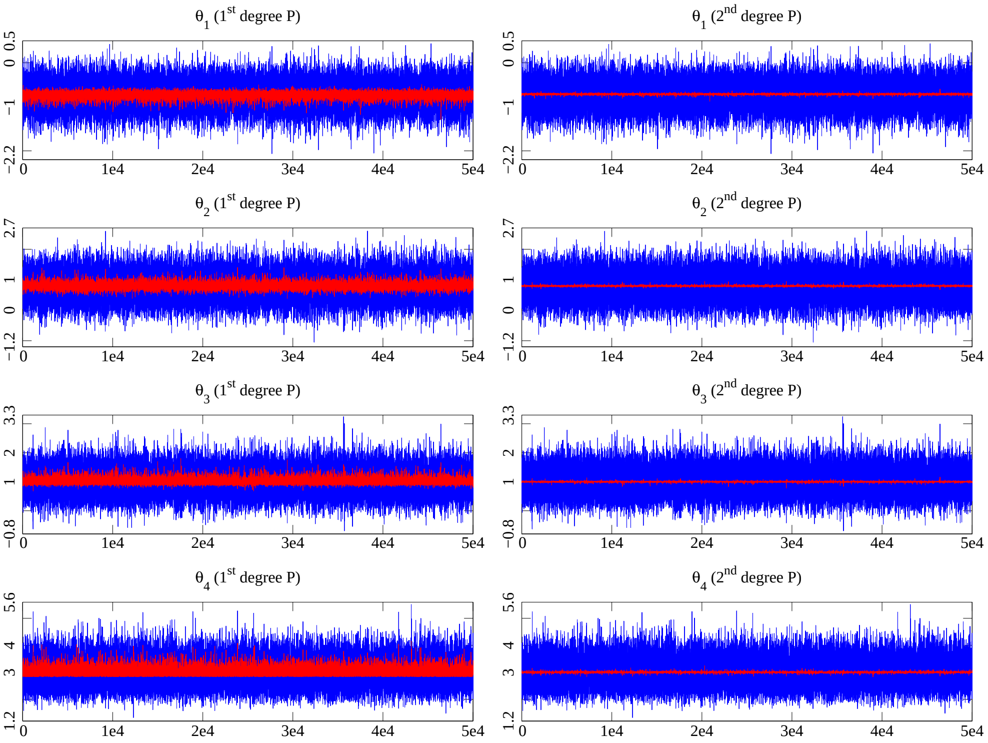

Figure R.1: Traces of RMHMC (blue line) and ZV-RMHMC (red line) for different parameters (rows) of the Bayesian logistic regression model and for different degree polynomials (columns). The blue-coloured RMHMC trace plots in the left and right columns are identical for each parameter (across each row) and represent the parameters $\theta_i,\ i = 1, 2, 3, 4$. On the other hand, the red-coloured trace plots in the left and right columns represent the ZV-RMHMC values $\tilde{\theta}_i$ for linear and quadratic polynomials, calculated respectively by (9) and (18).

evidence that in most cases MH, being the least effective sampler, exhibits the highest VRF, then MALA, SMMALA and MMALA cluster having similar VRFs and finally HMC and RMHMC possess the smallest VRF as the two most effective samplers. In fact, RMHMC has consistently the smallest VRF across all four parameters, agreeing with the previously reported findings on its ESS superiority.

As a demonstration of the gain in variance reduction independently of the selected model, Table R.2 shows the VRFs for the Bayesian logit and probit models implemented on the Swiss banknotes data. The same covariates and consequently the same model parameters are involved, while both the logit and probit link functions are used. As it can be seen, reduction in variance is observed irrespectively of the choice of link function. It appears that the probit model achieves greater reduction in variance than the logit but the reduction "pattern" is preserved.

5 ZV-MCMC for Bayesian Probit Regression

A Bayesian probit regression model is considered in its ordinary form, along the lines of Albert and Chib (1993):

$$\Phi^{-1}\left(p_i\right) = \sum_{j=1}^{n_\theta} \theta_j x_{ij} = \boldsymbol{\theta}^T \mathbf{x}_{i,},\ i = 1, 2, \ldots, n_d,$$
$$p_i := P(y_i = 1|\boldsymbol{\theta}, \mathbf{x}_{i,}) = p(y_i = 1|\boldsymbol{\theta}, \mathbf{x}_{i,}),$$

where $\mathbf{x}_{i,}$ is the i-th row of the $n_d \cdot n_\theta$ design matrix X, $y \in \{0,1\}^{n_d}$ the binary response variable, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ the regression coefficients and Φ^{-1} the link function of the probit model, which is the inverse cumulative distribution of $\mathcal{N}(0,1)$. The n_d latent variables $X\boldsymbol{\theta} + \boldsymbol{\epsilon}$ are assumed, where $\epsilon \ \sim \mathcal{N}(\mathbf{0}, I)$ as suggested by Albert and Chib (1993). Note that the assumption $\epsilon \ \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ is not preferred because it can cause problems of identifiability, as explained in the commentary of Girolami and Calderhead (2011) by Stathopoulos and Filippone. Under the notation $\rho_i := \boldsymbol{\theta}^T \mathbf{x}_{i,}$ and the assumption $y_i \sim Bernoulli(1, p_i)$, the probit model expresses as

$$p(y_i = 1|\boldsymbol{\theta}, \mathbf{x}_{i,}) = \Phi(\rho_i) = p_i^{y_i}(1 - p_i)^{1-y_i},\ i = 1, 2, \ldots, n_d, \tag{37}$$

whence the log-likelihood of the model is derived:

$$L(\mathbf{y}|\boldsymbol{\theta}, X) = \sum_{i=1}^{n_d} \ln\left[p(y_i = 1|\boldsymbol{\theta}, \mathbf{x}_{i,})\right] = \sum_{i=1}^{n_d} \left[y_i \ln(\Phi(\rho_i)) + (1 - y_i)\ln(\Phi(-\rho_i))\right]. \tag{38}$$

A Normal prior is assumed for the parameters $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, aI)$. A flat Normal prior is chosen in the example of Section 5.2 by setting $a = 100$.

5.1 ZV Estimates and Metric Tensor

In the framework outlined above, (20), (27) and (38) give the gradient of the log-posterior and the control variate is:

$$\mathbf{z}(\boldsymbol{\theta}) = -\frac{1}{2}\nabla_{\boldsymbol{\theta}}(\ln(p(\boldsymbol{\theta}|\mathbf{x}))) = -\frac{1}{2}\left(X^T s - \frac{1}{a}\boldsymbol{\theta}\right), \tag{39}$$

where the elements of the n_d-dimensional column vector $\mathbf{s}$ are

$$s_i := y_i \xi(\rho_i) - (1 - y_i)\xi(-\rho_i),\ i = 1, 2, \ldots, n_d, \tag{40}$$

and $\xi(\rho_i) := \mathcal{N}(\rho_i)/\Phi(\rho_i)$, with $\mathcal{N}$ and Φ denoting the density and cumulative distribution of a standard Gaussian random variable. Equations (30)-(34) give the ZV estimators for linear and quadratic polynomials for the Bayesian probit regression model, where the matrices Z and W are calculated using $\mathbf{z}(\boldsymbol{\theta})$ as specified by (39).

The metric tensor and its derivatives for the probit model, which are required for running the RMHMC and MMALA samplers, are given by (35), where the (i,i)-th elements of the $n_d \cdot n_d$ diagonal matrices $\Lambda(\boldsymbol{\theta})$ and $M^j(\boldsymbol{\theta})$ calculate as

$$\Lambda(\boldsymbol{\theta}) := diag\left[\xi(\rho_i)\xi(-\rho_i)\right] = diag\left[\frac{\mathcal{N}^2(\rho_i)}{\Phi(\rho_i)\Phi(-\rho_i)}\right], \tag{41}$$

$$M^j(\boldsymbol{\theta}) := \frac{\partial \Lambda(\boldsymbol{\theta})}{\partial \theta_j} = diag\left\{\frac{\xi^2(\rho_i)}{\Phi(-\rho_i)}\{\xi(-\rho_i) - 2\left[\mathcal{N}(\rho_i) + \rho_i \Phi(\rho_i)\right]\}\right\}x_{ij}. \tag{42}$$

5.2 Vaso Constriction Example

To exemplify simulation of ZV-RMHMC and ZV-MMALA for the Bayesian probit regression model, the vaso constriction data from Finney (1947) are used. The data come from an experiment conducted on human physiology to study the effect of taking a single deep breath on the occurrence of a reflex vaso constriction in the skin of the digits. 39 samples from three individuals are available, each of them contributing 9, 8 and 22 samples. Although the data represent repeated measurements, Pregibon (1981) claims that the observations can be assumed to be independent, therefore the Bayesian probit model can be applied. Two explanatory variables are included in the study, namely the rate of inhalation and the inhaled volume of air per individual. An intercept is also added, so $n_\theta = 3$ regression coefficients comprise the parameters of the model and the design matrix X has dimensions $n_d \cdot n_\theta = 39 \cdot 3$. The occurrence or non-occurrence of vaso constriction in the skin of the digits of each subject, corresponding to 1 and 0, plays the role of the binary response.

$50,000$ MCMC samples are obtained per chain after a burn-in stage of $5,000$ samples. Figure R.3 shows, for a single chain, via boxplots, the considerable variance reduction achieved by the ZV estimators, especially for quadratic polynomials when the variance nearly vanishes.

Table R.3 reports the asymptotic variances and associated VRFs for $n_c = 100$ simulated chains for each of the six samplers. The tabulated results for the probit model fully agree with the ones reported in Table R.1 for the logit model. More specifically, Table R.3 demonstrates that MH has the highest asymptotic variance, thus the smallest ESS, followed by MALA, SMMALA, MMALA, HMC and RMHMC. This order in the levels of asymptotic variance holds both for the original chains and for their ZV counterparts. Moreover, this order translates into MH and RMHMC being the algorithms with the highest and lowest VRFs respectively. In fact, the succession of the intermediate samplers MALA, SMMALA, MMALA and HMC ranging from higher to lower VRFs is violated in a minority of cases. Finally, HMC is the second most effective sampler, having consistently the second smallest asymptotic variance and VRF.

By examining Tables R.1 and R.3 concurrently, it is deduced from the examples on the logit and probit models that in terms of minimal asymptotic variance Hamiltonian Monte Carlo methods are always preferable, followed by the Metropolis adjusted

Langevin family of algorithms and lastly by the Metropolis-Hastings sampler. Furthermore, ZV-(RM)HMC reduces the asymptotic variance by one to three magnitudes of order even in the case of the most effective Hamiltonian Monte Carlo algorithms.

Furthermore, Table R.4 shows the VRFs for the Bayesian logit and probit models implemented on the Vaso constriction data. Both Tables R.2 and R.4 show that the ZV estimators reduce the variance irrespectively of the selected link function and that the achieved reduction in variance is greater under the probit model for both datasets (Swiss banknotes and vaso constriction data).

Table R.5 reports the runtime of MCMC in comparison to ZV-MCMC for the Bayesian logit and probit models, each as applied on the Swiss banknotes and on the vaso constriction data. The difference in the runtime between MCMC and ZV-MCMC is due to the calculation of the polynomial coefficients of the trial function, and it is less than 0.3% for both linear and quadratic polynomials for each of the four tabulated examples.

6 ZV-MCMC for Ordinary Differential Equations

Consider a dynamical system modeled by the n_x ordinary differential equations (ODEs) $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}, t)$, where t denotes time, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ the model parameters, $\mathbf{x}(t) \in \mathbb{R}^{n_x}$ the state of the system at time t and $\dot{\mathbf{x}}(t)$ its time derivative. It is assumed that the state $\mathbf{x}(t)$ is degraded by noise $\boldsymbol{\epsilon}(t)$ so that $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$ is observed. Observations are made at n_t distinct time points, so the observed state of the system is described by $Y = X + E$, where the matrices Y, X and E have dimensions $n_t \cdot n_x$. Given the parameters $\boldsymbol{\theta}$ and some initial conditions $\mathbf{x}_0 \in \mathbb{R}^{n_x}$, the initial value problem is solved for X leading to the solution $X(\boldsymbol{\theta}, \mathbf{x}_0)$ of the ODEs at the n_t specified time points.

Gaussian noise $\boldsymbol{\epsilon}_{,j} \sim \mathcal{N}(\mathbf{0}, \sigma_j^2 I_{n_t}),\ j = 1, 2, \ldots, n_x$, is assumed, where $\boldsymbol{\epsilon}_{,j}$ is the j-th column of the error matrix E and I_{n_t} is the $n_t \cdot n_t$ identity matrix. Then it follows that $\mathbf{y}_{,j} | (\mathbf{x}_{,j}(\boldsymbol{\theta}, \mathbf{x}_0), \sigma_j^2) \sim \mathcal{N}(\mathbf{x}_{,j}, \sigma_j^2 I_{n_t})$, hence the log-likelihood of the model takes the form

$$L(Y|\boldsymbol{\theta}, X(\boldsymbol{\theta}, \mathbf{x}_0)) = -\frac{1}{2} \sum_{j=1}^{n_x} \left[\frac{1}{\sigma_j^2} \|\mathbf{y}_{,j} - \mathbf{x}_{,j}\| + n_t \ln(2\pi\sigma_j^2) \right], \tag{43}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\mathbf{y}_{,j}$ is the j-th column of matrix Y and $\mathbf{x}_{,j}$ is the j-th column of the ODE solution $X(\boldsymbol{\theta}, \mathbf{x}_0)$.

6.1 Sensitivities and Expected Fisher Information

The first order sensitivities, defined as the $n_x \cdot n_\theta$ Jacobian matrix $S := \partial\mathbf{x}/\partial\boldsymbol{\theta}$ of partial derivatives of the states over the parameters, and the second order sensitivities $\partial S/\partial\theta_i,\ i = 1, 2, \ldots, n_\theta$, are required for calculating the gradient of the log-target as well as the metric tensor and its derivatives to run ZV-RMHMC and ZV-MMALA. Successive differentiation of the system's ODEs with respect to the parameters, using the chain rule, generates a new set of ODEs which entail the first and second order sensitivities:

$$\dot{S} = \mathbf{f}^{\mathbf{x}}\, S + \mathbf{f}^{\boldsymbol{\theta}}, \tag{44}$$

$$\dot{\mathbf{s}}^{ij}_{k,} = \mathbf{f}^{\mathbf{x}}\, \mathbf{s}^{ij}_{k,} + \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}}{\partial \theta_i}\mathbf{s}^{j}_{k,} + \sum_{\ell=1}^{n_x} s^{i}_{k\ell}\frac{\partial \mathbf{f}^{\mathbf{x}}}{\partial x_\ell}\mathbf{s}^{j}_{k,} + \sum_{\ell=1}^{n_x} s^{i}_{k\ell}\frac{\partial \mathbf{f}^{\boldsymbol{\theta}}_{,j}}{\partial x_\ell} + \frac{\partial \mathbf{f}^{\boldsymbol{\theta}}_{,j}}{\partial \theta_i}, \tag{45}$$

where $\mathbf{f}^{\mathbf{x}} := \partial \mathbf{f}/\partial \mathbf{x}$ is the $n_x \cdot n_x$ Jacobian matrix of $\mathbf{f}$ over the states, $\mathbf{f}^{\boldsymbol{\theta}} := \partial \mathbf{f}/\partial \boldsymbol{\theta}$ the $n_x \cdot n_\theta$ Jacobian of $\mathbf{f}$ over the parameters, $\partial \mathbf{f}^{\boldsymbol{\theta}}_{,j}/\partial x_\ell$ the partial derivative of the j-th column of $\mathbf{f}^{\boldsymbol{\theta}}$ with respect to x_ℓ and $\partial \mathbf{f}^{\boldsymbol{\theta}}_{,j}/\partial \theta_i$ the partial derivative of the j-th column of $\mathbf{f}^{\boldsymbol{\theta}}$ with respect to θ_i. As for the sensitivities, $\mathbf{s}^{ij}_{k,}$ denotes the n_x-dimensional column vector of second order sensitivities with elements $s^{ij}_{k\ell} := \partial\,^2 x_{k\ell}/\partial\theta_i\partial\theta_j,\ \ell = 1, 2, \ldots, n_x$, and $\dot{\mathbf{s}}^{ij}_{k,}$ its time derivatives, while $\mathbf{s}^{j}_{k,}$ and $s^{i}_{k\ell} := \partial x_{k\ell}/\partial\theta_i$ are the j-th column and the (ℓ, i)-th element of the matrix S of first order sensitivities, respectively. The differential equation (45) refers to a single time point k, therefore (45) describes n_t ODEs for $k = 1, 2, \ldots, n_t$. Note that due to the symmetry $s^{ij}_{k\ell} = s^{ji}_{k\ell}$ of second order sensitivities, (45) needs to be calculated only for $i \leq j$.

The ODEs of the system are augmented by the sensitivity equations (44) and (45) to compute numerically the solution $X(\boldsymbol{\theta}, \mathbf{x}_0)$ of the system together with the first and order sensitivities $S,\ \partial S/\partial\theta_i$. Then the sensitivities are used for deriving the gradient of the log-likelihood, the expected Fisher information $F := -E_{\mathbf{y}|\boldsymbol{\theta},X}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}L(\mathbf{y}|\boldsymbol{\theta}, X)\right]$ and its derivatives,

$$\frac{\partial L(Y|\boldsymbol{\theta}, X(\boldsymbol{\theta}, \mathbf{x}_0))}{\partial \theta_q} = \sum_{\ell=1}^{n_x}\left[\frac{1}{\sigma^2_\ell}(\mathbf{y}_{,\ell} - \mathbf{x}_{,\ell}) \cdot s^{q}_{,\ell}\right],\ q = 1, 2, \ldots, n_\theta, \tag{46}$$

$$F_{ij} = \sum_{\ell=1}^{n_x}\frac{1}{\sigma^2_\ell}\mathbf{s}^{i}_{,\ell} \cdot \mathbf{s}^{j}_{,l},\ \ i, j \in \{1, 2, \ldots, n_\theta\}, \tag{47}$$

$$\frac{\partial F_{ij}}{\partial \theta_q} = \sum_{\ell=1}^{n_x}\frac{1}{\sigma^2_\ell}(\mathbf{s}^{qi}_{,\ell} \cdot \mathbf{s}^{j}_{,l} + \mathbf{s}^{i}_{,\ell} \cdot \mathbf{s}^{qj}_{,l}),\ q = 1, 2, \ldots, n_\theta, \tag{48}$$

where the dot between vectors denotes the inner product and F_{ij} is the (i, j)-th element of F. It thus becomes possible to evaluate the metric tensor and its derivatives once a prior $\pi(\boldsymbol{\theta})$ is selected and the Hessian of the log-prior $\partial^2 \ln(\pi(\boldsymbol{\theta}))/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ is calculated.

6.2 Prior for Positive Model Parameters

In some areas of research the model parameters are constrained to the positive real line $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}_{+}$. On the other hand, (ZV-)RMHMC and (ZV-)MMALA operate on an unbounded parameter space, therefore these samplers are run on some reparameterization $\boldsymbol{\phi} = \mathbf{h}(\boldsymbol{\theta}) \in \mathbb{R}^{n_\theta}$ and the obtained estimates $\hat{\boldsymbol{\phi}}$ are transformed to obtain the estimates of the original model parameters $\hat{\boldsymbol{\theta}} = \mathbf{h}^{-1}(\hat{\boldsymbol{\phi}})$. In such cases, a log-Gamma or Normal prior $\pi(\boldsymbol{\phi})$ is usually assumed to run the (ZV-)MCMC algorithm, implying a respective Gamma or log-Normal prior $\pi(\boldsymbol{\theta})$ for the model parameters.

Log-Gamma Prior

Assign a Gamma prior to the positive model parameters, $\boldsymbol{\theta} \sim G(\boldsymbol{\kappa}, \boldsymbol{\lambda})$, which introduces $2n_\theta$ hyperparameters $\boldsymbol{\kappa}, \boldsymbol{\lambda}$. Perform the transformation $\boldsymbol{\phi} = \ln(\boldsymbol{\theta}/\boldsymbol{\lambda})$, where the involved vector division is Hadamard element-wise. Then the transformed parameters follow a log-Gamma prior $\boldsymbol{\phi} \sim logG(\boldsymbol{\kappa})$, see for example Chan (1993) for details.

So the log-prior of $\boldsymbol{\phi}$ is

$$\ln\left(\pi(\boldsymbol{\phi})\right) = \sum_{i=1}^{n_\phi} \left[k_i \phi_i - \exp\left(\phi_i\right) - \ln(\Gamma(k_i))\right], \tag{49}$$

whence its gradient is found to be $\nabla_{\boldsymbol{\phi}}(\ln\left(\pi(\boldsymbol{\phi})\right)) = \boldsymbol{\kappa} - \exp\left(\boldsymbol{\phi}\right)$. Thus, the gradient of the log-posterior and subsequently $\mathbf{z}(\boldsymbol{\phi})$ for the ODE model are given by

$$z_q(\boldsymbol{\phi}) = -\frac{1}{2}\frac{\partial \ln(p(\boldsymbol{\phi}|Y, X(\boldsymbol{\phi}, \mathbf{x}_0)))}{\partial \phi_q} = -\frac{1}{2}\left[\frac{\partial L}{\partial \phi_q} - \kappa_q + \exp\left(\phi_q\right)\right], \tag{50}$$

where $z_q,\ q = 1, 2, \ldots, n_\theta$, denotes the q-th element of $\mathbf{z}(\boldsymbol{\phi})$. The derivative of the log-likelihood $\partial L/\partial \phi_q$ is available in (46) by replacing θ_q and $s^q_{k\ell} = \partial x_{k\ell}/\partial \theta_q$ by ϕ_q and $s^q_{k\ell} = \partial x_{k\ell}/\partial \phi_q$, respectively.

It is further deduced from (49) that the Hessian of the log-prior is the $n_\phi \cdot n_\phi$ diagonal matrix $\Xi(\boldsymbol{\phi}) := \partial^2 \ln\left(\pi(\boldsymbol{\phi})\right)/\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T = -diag\left[\exp(\boldsymbol{\phi})\right]$, where $diag\left[\exp(\boldsymbol{\phi})\right]$ is the diagonal matrix whose (i,i)-th element is $\exp(\phi_i)$, and the derivatives of the Hessian are $\partial\,\Xi(\boldsymbol{\phi})/\partial\phi_i = \emptyset\left[\exp(\phi_i)\right],\ i = 1, 2, \ldots, n_\phi$, where $\emptyset\left[\exp(\phi_i)\right]$ denotes the $n_\phi \cdot n_\phi$ matrix whose only non-zero element $\exp(\phi_i)$ is the one in the (i,i)-th position.

RMHMC and MMALA are run on the log-Gamma distributed parameters $\boldsymbol{\phi}$. The expected Fisher information, F, and its derivatives, $\partial F/\partial \phi_i$, are available in (47) and (48) by using $\boldsymbol{\phi}$ in place of $\boldsymbol{\theta}$ and by using the first and second order sensitivities with respect to $\boldsymbol{\phi}$. F combined with the Hessian of the log-prior Ξ and its derivatives gives the metric tensor $G(\boldsymbol{\phi}) = F(\boldsymbol{\phi}) - \Xi(\boldsymbol{\phi})$ and its derivatives $\partial G(\boldsymbol{\phi})/\partial \phi_i$.

Once RMHMC and MMALA are run, the obtained chains for $\boldsymbol{\phi}$ are transformed to the original parameters $\boldsymbol{\theta} = \boldsymbol{\lambda} \circ \exp(\boldsymbol{\phi})$. Furthermore, the gradient of the log-target $p_{\boldsymbol{\phi}}$ is converted to the gradient of the log-posterior $p_{\boldsymbol{\theta}}$ using the inverse transform theorem:

$$\nabla_{\boldsymbol{\theta}}(p_{\boldsymbol{\theta}}(\boldsymbol{\theta}|Y, X(\boldsymbol{\theta}, \mathbf{x}_0))) = \left[\nabla_{\boldsymbol{\phi}}(p_{\boldsymbol{\phi}}(\boldsymbol{\phi}|Y, X(\boldsymbol{\phi}, \mathbf{x}_0))) - 1\right]/\boldsymbol{\theta}, \tag{51}$$

where the vector division is Hadamard element-wise. This consequently means that $\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = (\mathbf{z}_{\boldsymbol{\phi}}(\boldsymbol{\phi}) + 1/2)/\boldsymbol{\theta}$. To compute the ZV estimates $\tilde{\boldsymbol{\theta}}$ for linear and quadratic polynomials under the ODE model with the Gamma prior $\boldsymbol{\theta} \sim G(\boldsymbol{\kappa}, \boldsymbol{\lambda})$, the estimates $\boldsymbol{\theta}$ and $\mathbf{z}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ are then embedded in Equations (30)-(34).

Normal Prior

Alternatively to the Gamma prior, a log-Normal prior $\boldsymbol{\theta} \sim log\mathcal{N}(\mathbf{0}, aI)$, can be assumed for the original parameters of the ODEs. Then the transformation $\boldsymbol{\phi} = \ln\left(\boldsymbol{\theta}\right)$ allows to use the Normally distributed parameters $\boldsymbol{\phi} \sim \pi(\boldsymbol{\phi}) = \mathcal{N}(\mathbf{0}, aI)$ for running ZV-RMHMC and ZV-MMALA. The only required component of the log-prior $\ln(\pi(\boldsymbol{\phi}))$ is its gradients $\nabla_{\boldsymbol{\phi}}(\ln(\pi(\boldsymbol{\phi}))) = -\boldsymbol{\phi}/a$, since the Hessian of $\ln(\pi(\boldsymbol{\phi}))$ vanishes.

After the MCMC run, the estimates $\boldsymbol{\theta} = \exp(\boldsymbol{\phi})$ are uncovered from the simulated $\boldsymbol{\phi}$, while $\mathbf{z}_{\boldsymbol{\theta}}$ is derived from $\mathbf{z}_{\boldsymbol{\phi}}$ using (51), which holds under the current scheme of Normal prior too. The ZV estimates $\tilde{\boldsymbol{\theta}}$ are then obtained with the help of Equations (30)-(34).

6.3 Example: Fitzhugh-Nagumo ODEs

As an example of ZV-RMHMC and ZV-MMALA under an ODE model, consider the Fitzhugh-Nagumo ODEs (see for example Ramsay et al. (2007))

$$\dot{V} = c\left(V - \frac{V^3}{3} + R\right), \quad \dot{R} = -\frac{1}{c}\left(V - a + bR\right). \tag{52}$$

$\mathbf{x}^T = (V, R)$ are the states and $\boldsymbol{\theta}^T = (a, b, c)$ are the parameters of the model. Following the example of Girolami and Calderhead (2011), data are generated from the Fitzhugh-Nagumo differential equations at $n_t = 200$ distinct time points between $t = 0$ and $t = 20$ using the initial condition $\mathbf{x}_0^T = (V(0), R(0)) = (-1, 1)$, the parameter values $(a, b, c) = (0.2, 0.2, 3)$ and Gaussian noise with variance equal to $\sigma_j^2 = 0.25, \ j = 1, 2$.

Attention must be paid to MCMC simulations of non-linear ODEs, as the chains may converge to the wrong mode (see Calderhead and Girolami (2009), for details). A population MCMC scheme can be employed to ensure convergence to the target distribution as suggested for example in Calderhead et al. (2009). Nevertheless, it is enough to sample a single chain initialized on the true mode for the purpose of comparing the original to the ZV estimates.

Figure R.4 displays the boxplots of such a single chain realization of length $n = 10,000$ after a burn-in phase of $10,000$ points. A log-Gamma prior $logG(\boldsymbol{\kappa})$, $\boldsymbol{\kappa}^T = (2, 2, 2)$, is chosen for the simulation of $\boldsymbol{\phi}$ via ZV-RMHMC, ZV-MMALA, ZV-SMMALA and ZV-MH. The simulated chains are transformed to $\boldsymbol{\theta}$, corresponding to samples drawn with the help of a Gamma prior $G(\boldsymbol{\kappa}, \boldsymbol{\lambda})$, $\boldsymbol{\lambda}^T = (2, 2, 2)$. The boxplots demonstrate drastic reduction in variance in all six MCMC schemes.

Table R.6 quantifies the variance reduction by means of the VRFs after running $n_c = 100$ chains for each of the six MCMC schemes. The patterns of asymptotic variance across the six samplers, as previously noted in Tables R.1 and R.3, do not fully manifest themselves in Table R.6. This is to be expected given the complexity of the ODE system comparatively to the simpler logit and probit models. More specifically, the non-linear dynamics of the ODE system pose challenges in terms of convergence and fine-tuning of the MCMC scheme, which are not conducive to a full assessment of the ESS and of the MCMC mean estimator of the ODE model. As previously stated, alternative MCMC schemes such as population MCMC can potentially alleviate the lack of convergence. Nevertheless, Table R.6 serves the purpose of demonstrating the reduction in variance achieved by ZV-MCMC when the ZV method is implemented in a complex dynamical system.

7 Discussion

This paper provides a simulation-based assessment of the zero variance principle on Hamiltonian Monte Carlo and Metropolis adjusted Langevin algorithms. The examples of Bayesian logit and probit regression and of the non-linear Fitzhugh-Nagumo ODEs are conclusive in that the variance is attenuated drastically by using the ZV strategy with linear polynomials and vanishes nearly completely in ZV estimators with quadratic polynomials. The examples of logit and probit models further assess the asymptotic variances of the considered MCMC samplers, confirming that the Hamiltonian Monte Carlo methods are more effective exhibiting smaller variance and that RMHMC entails the smallest variance.

Given that the variance reduction is achieved at no extra computational cost, it becomes instructive to embody the ZV method in any application of Metropolis-Hastings type algorithms in particular those where, in order to run the sampler, one needs to compute the gradient of the log-target. Such algorithms include MALA, MMALA, HMC and RMHMC.

The accompanying MATLAB package facilitates the incentive of integrating the two methods by providing a general and extensible computational framework for ZV-(M)MALA and ZV-(RM)HMC. The user input required for running the ZV-MCMC routines of the package is a single file, in the form of a MATLAB class, which defines the log-target and its gradient as well as the metric tensor and its derivatives with respect to the model parameters. Care has been taken in the implementation of the ODE framework in the package, which uses the ODE solvers of the Systems Biology Pharmacodynamic (SBPD) MATLAB package. This implies that the ODE model is compiled to C code and that the efficient SUNDIALS C solvers are invoked. This way, highly performing MMALA and RMHMC samplers with negligible variance are at the user's disposal for a broad set of user-defined ODE models.

Finally, a computational framework is currently being set-up, where the ideas presented and discussed in this paper can be applied even if the derivative of the log-target is not available in closed form. This is achievable by employing numerical or symbolic computation techniques for automatic differentiation in place of the unknown functional expressions of the log-target, metric tensor and their associated derivatives (see for example Gay (2006), Smith (1995), Siskind and Pearlmutter (2008) and Naumann (2008)).

References

Adler, S. L. (1981). "Over-Relaxation Method for the Monte Carlo Evaluation of the Partition Function for Multiquadratic Actions." *Physical Review D*, 23: 2901–2904. 98

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous

Response Data." *Journal of the American Statistical Association*, 88(422): 669–679. 109

Andradóttir, S., Heyman, D. P., and Ott, T. J. (1993). "Variance Reduction Through Smoothing and Control Variates for Markov Chain Simulations." *ACM Transactions on Modeling and Computer Simulation*, 3(3): 167–189. 98

Assaraf, R. and Caffarel, M. (1999). "Zero-Variance Principle for Monte Carlo Algorithms." *Physical Review Letters*, 83: 4682–4685. 98, 99

Atchadé, Y. F. and Perron, F. (2005). "Improving on the Independent Metropolis-Hastings Algorithm." *Statistica Sinica*, 15(1): 3–18. 98

Barone, P. and Frigessi, A. (1990). "Improving Stochastic Relaxation for Gaussian Random Fields." *Probability in the Engineering and Informational Sciences*, 4(03): 369–389. 97

Barone, P., Sebastiani, G., and Stander, J. (2001). "General Over-Relaxation Markov Chain Monte Carlo Algorithms for Gaussian Densities." *Statistics and Probability Letters*, 52(2): 115 – 124. 98

Calderhead, B. and Girolami, M. (2009). "Estimating Bayes factors via thermodynamic integration and population MCMC." *Computational Statistics and Data Analysis*, 53(12): 4028 – 4045. 114

Calderhead, B., Girolami, M., and Lawrence, N. (2009). "Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes." *Advances in Neural Information Processing Systems, MIT Press*, 21. 114

Chan, P. S. (1993). "A Statistical Study of Log-Gamma Distribution." Ph.D. thesis, McMaster University. 113

Craiu, R. V. and Lemieux, C. (2007). "Acceleration of the Multiple-Try Metropolis Algorithm Using Antithetic and Stratified Sampling." *Statistics and Computing*, 17(2): 109–120. 97

Dellaportas, P. and Kontoyiannis, I. (2012). "Control Variates for Estimation Based on Reversible Markov Chain Monte Carlo Samplers." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1): 133–161. 98

Diaconis, P., Holmes, S., and Neal, R. M. (2000). "Analysis of a Nonreversible Markov Chain Sampler." *The Annals of Applied Probability*, 10(3): pp. 726–752. 98

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). "Hybrid Monte Carlo." *Physics Letters B*, 195(2): 216 – 222. 98

Dyk, D. A. v. and Meng, X.-L. (2001). "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics*, 10(1): pp. 1–50. 98

Finney, D. J. (1947). "The Estimation from Individual Records of the Relationship Between Dose and Quantal Response." *Biometrika*, 34(3-4): 320–334. 110

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics*. Chapman and Hall. 107

Fort, G., Moulines, E., Roberts, G. O., and Rosenthal, J. S. (2003). "On the Geometric Ergodicity of Hybrid Samplers." *Journal of Applied Probability*, 40(1): pp. 123–146. 98

Gay, D. M. (2006). *Semiautomatic Differentiation for Efficient Gradient Computations*, volume 50, chapter 13, 147–158. Berlin/Heidelberg: Springer-Verlag. 115

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Quarterly of applied mathematics*, 85(410): 398–409. 98

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall. 104

Geyer, C. J. (1992). "Practical Markov Chain Monte Carlo." *Statistical Science*, 7(4): 473–483. 102

Geyer, C. J. and Mira, A. (2000). "On Non-Reversible Markov chains." In *Institute Communications, Volume 26: Monte Carlo Methods*, 93–108. American Mathematical Society. 98

Girolami, M. and Calderhead, B. (2011). "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214. 98, 103, 107, 109, 114

Green, P. J. and Mira, A. (2001). "Delayed Rejection in Reversible Jump Metropolis-Hastings." *Biometrika*, 88(4): pp. 1035–1053. 98

Hammer, H. and Tjemeland, H. (2008). "Control Variates for the Metropolis-Hastings Algorithm." *Scandinavian Journal of Statistics*, 35(3): 400–414. 98

Henderson, S. G. (1997). "Variance Reduction via an Approximating Markov Process." Ph.D. thesis, Stanford University. 98

Mira, A., Solgi, R., and Imparato, D. (2012). "Zero Variance Markov Chain Monte Carlo for Bayesian Estimators." *Statistics and Computing*, 1–10. 98, 99, 100, 104

Mira, A. and Tierney, L. (2002). "Efficiency and Convergence Properties of Slice Samplers." *Scandinavian Journal of Statistics*, 29(1): pp. 1–12. 98

Naumann, U. (2008). "Optimal Jacobian accumulation is NP-complete." *Mathematical Programming*, 112(2): 427–441. 115

Philippe, A. and Robert, C. P. (2001). "Riemann Sums for MCMC Estimation and Convergence Monitoring." *Statistics and Computing*, 11(2): 103–115. 98

Pregibon, D. (1981). "Logistic Regression Diagnostics." *The Annals of Statistics*, 9(4): pp. 705–724. 110

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). "Parameter Estimation for Differential Equations: a Generalized Smoothing Approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5): 741–796. 114

Ripley, B. (1987). *Stochastic Simulation*. John WIley & Sons. 98

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics, 2nd edition. 98

Roberts, G. O. and Stramer, O. (2002). "Langevin Diffusions and Metropolis-Hastings Algorithms." *Methodology and Computing in Applied Probability*, 4: 337–357. 104

Siskind, J. M. and Pearlmutter, B. A. (2008). "Nesting Forward-Mode AD in a Functional Framework." *Higher Order and Symbolic Computation*, 21(4): 361 – 376. 115

Smith, S. P. (1995). "Differentiation of the Cholesky Algorithm." *Journal of Computational and Graphical Statistics*, 4(2): 134 – 147. 115

Solgi, R. and Mira, A. (2013). "A Bayesian Semiparametric Multiplicative Error Model with an Application to Realized Volatility." *Journal of Computational and Graphical Statistics*, 22(3): 558–583. 98

Swendsen, R. H. and Wang, J.-S. (1987). "Nonuniversal Critical Dynamics in Monte Carlo Simulations." *Physical Review Letters*, 58: 86–88. 98

Tierney, L. and Mira, A. (1999). "Some Adaptive Monte Carlo Methods for Bayesian Inference." *Statistics in Medicine*, 18: 2507–2515. 98

Acknowledgments

The authors are grateful to the SysMO consortium and the SilicoTryp project for funding Theodore Papamarkou during his research on this paper. Antonietta Mira acknowledges Swiss National Foundation research grant 200021-132766 on Zero-Variance Markov chain Monte Carlo. Mark Girolami is supported by an EPSRC Established Career Research Fellowship, EP/J016934/1 and a Royal Society Wolfson Research Merit Award. The authors are also thankful to Reza Solgi and Daniele Imparato for useful and constructive comments on the ZV principle and some initial code for the implementation of ZV. Finally, the authors would like to thank Ben Calderhead for helpful discussions on Geometric MCMC methods.

Appendix: Boxplots and Variance Reduction Factors

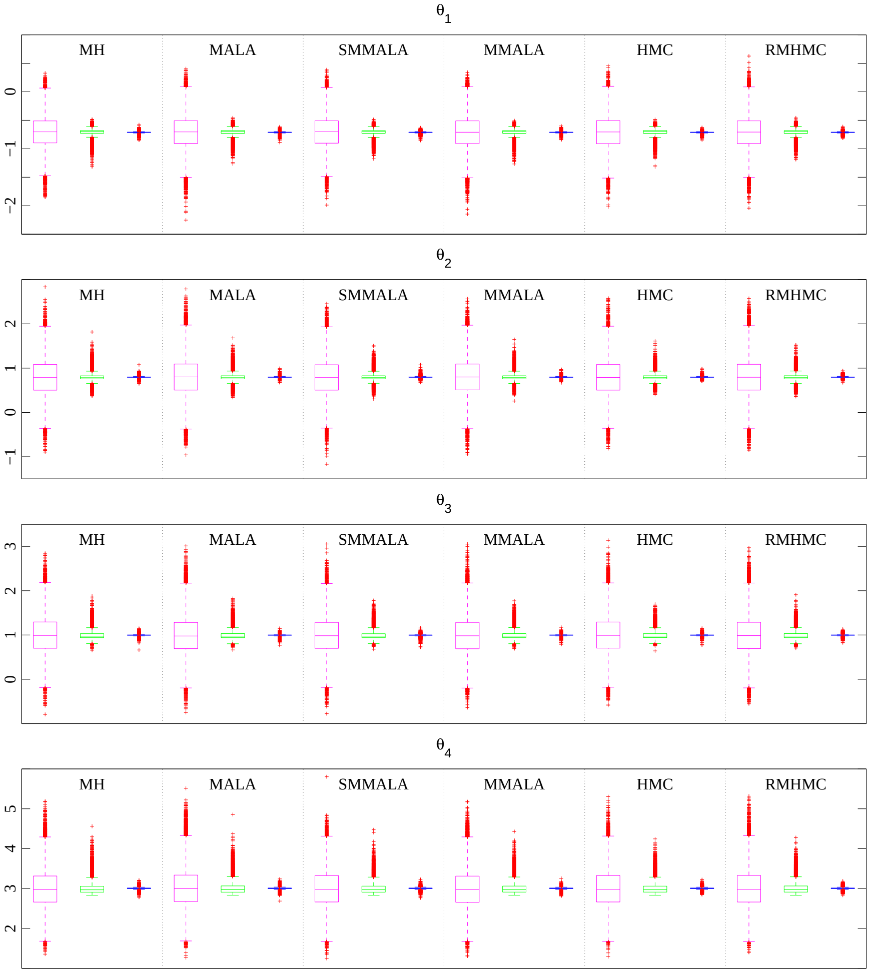

Figure R.2: Boxplots of ordinary MCMC (magenta), ZV-MCMC with linear polynomials (green) and quadratic polynomials (blue) for each parameter of the Bayesian logistic regression model as applied on the Swiss banknotes.

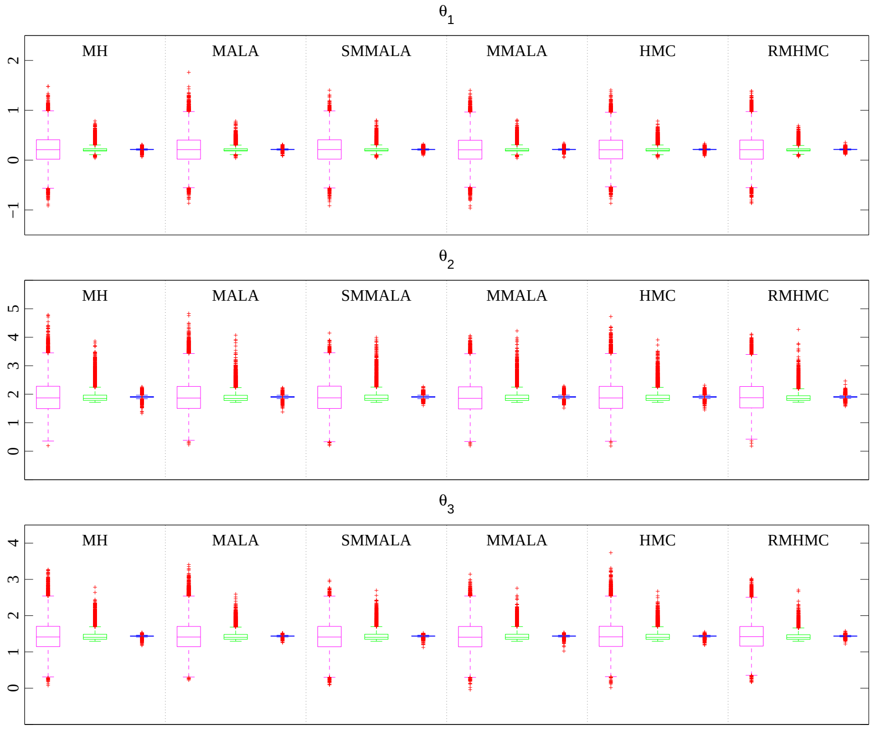

Figure R.3: Boxplots of ordinary MCMC (magenta), ZV-MCMC with linear polynomials (green) and quadratic polynomials (blue) for each parameter of the Bayesian probit regression model as applied on the vaso constriction data.

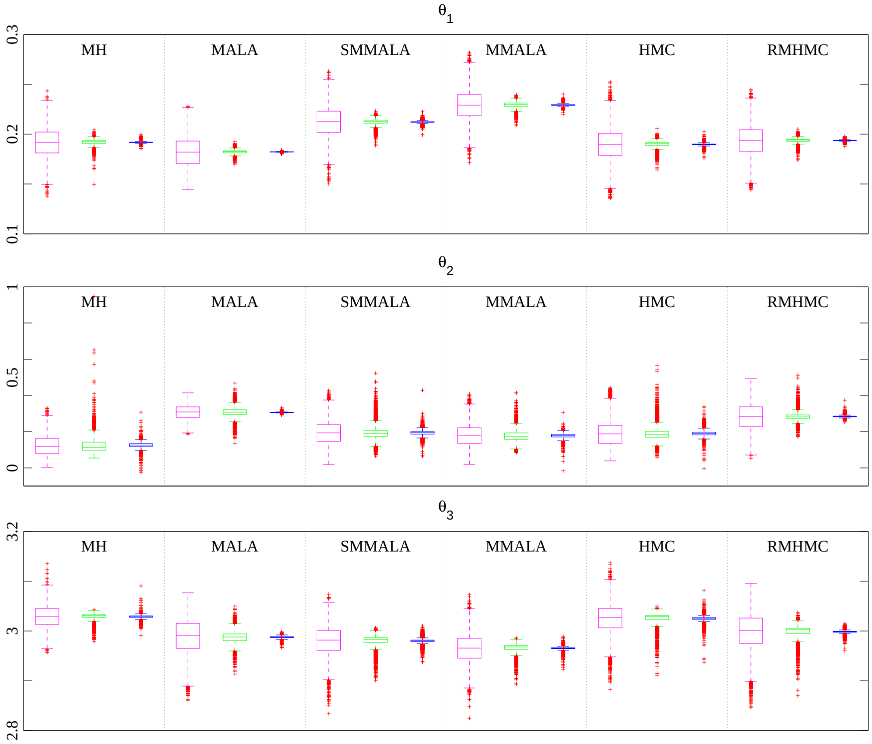

Figure R.4: Boxplots of MCMC (magenta), ZV-MCMC with linear polynomials (green) and quadratic polynomials (blue) for each parameter of the Fitzhugh-Nagumo ODEs.

	MCMC	L-ZV-MCMC		Q-ZV-MCMC	
	Variance	Variance	VRF	Variance	VRF
θ_1					
MH	0.7085	0.0143	49.53	0.0002	3903.24
MALA	0.4206	0.0140	30.15	0.0002	2472.55
SMMALA	0.4817	0.0155	31.08	0.0003	1732.64
MMALA	0.3878	0.0102	37.87	0.0002	2353.82
HMC	0.2187	0.0127	17.23	0.0002	1202.89
RMHMC	**0.0860**	**0.0079**	**10.90**	**0.0001**	**989.36**
θ_2					
MH	2.5078	0.0306	81.86	0.0003	7316.08
MALA	2.5608	0.0467	54.84	0.0003	7447.15
SMMALA	1.0381	0.0306	33.90	0.0004	2814.97
MMALA	0.8259	0.0209	39.56	0.0002	3366.95
HMC	0.5160	0.0316	16.32	0.0003	1737.94
RMHMC	**0.1809**	**0.0168**	**10.78**	**0.0001**	**1420.74**
θ_3					
MH	2.1869	0.0413	52.92	0.0004	6164.20
MALA	2.4584	0.0467	52.66	0.0004	6671.19
SMMALA	1.0875	0.0348	31.27	0.0004	2941.56
MMALA	0.8697	0.0243	35.73	0.0003	3410.03
HMC	0.5532	0.0404	13.70	0.0003	1786.99
RMHMC	**0.1887**	**0.0205**	**9.20**	**0.0001**	**1415.10**
θ_4					
MH	1.3925	0.1215	11.46	0.0008	1736.54
MALA	3.2107	0.1471	21.83	0.0010	3369.44
SMMALA	1.6293	0.1013	16.08	0.0010	1678.99
MMALA	1.2678	0.0682	18.59	0.0006	1993.76
HMC	0.8620	0.1161	7.42	0.0007	1223.61
RMHMC	**0.2563**	**0.0513**	**4.99**	**0.0003**	**941.38**

Table R.1: Estimated asymptotic variances (rescaled by multiplying by the simulation length $n = 50,000$ of each chain) and variance reduction factors for each parameter of the Bayesian logit model applied on the Swiss banknotes data after $n_c = 100$ realized chains. The estimated asymptotic variances of both MCMC and ZV-MCMC are smaller for more effective samplers. As a consequence, the VRFs of more effective algorithms are smaller. The variances and VRFs have been rounded to the fourth and second decimal places respectively, while the cells in bold represent the smallest variances and reduction factors across the six samplers demonstrating that RMHMC is the most effective sampler.

	L-ZV-MCMC		Q-ZV-MCMC	
	Logit	Probit	Logit	Probit
		θ_1		
MH	49.53	224.76	3903.24	36691.02
MALA	30.15	108.13	2472.55	16872.81
SMMALA	31.08	124.27	1732.64	23792.15
MMALA	37.87	150.12	2353.82	30983.78
HMC	17.23	71.09	1202.89	15302.85
RMHMC	10.90	35.01	989.36	10889.57
		θ_2		
MH	81.86	211.32	7316.08	48569.46
MALA	54.84	139.47	7447.15	41132.25
SMMALA	33.90	76.58	2814.97	21681.99
MMALA	39.56	91.62	3366.95	26443.08
HMC	16.32	54.86	1737.94	11711.14
RMHMC	10.78	19.78	1420.74	7091.47
		θ_3		
MH	52.92	158.12	6164.20	39463.93
MALA	52.66	130.14	6671.19	31738.70
SMMALA	31.27	80.33	2941.56	18124.88
MMALA	35.73	92.19	3410.03	21372.32
HMC	13.70	47.72	1786.99	9884.59
RMHMC	9.20	20.18	1415.10	6054.57
		θ_4		
MH	11.46	27.02	1736.54	27297.00
MALA	21.83	40.49	3369.44	29738.73
SMMALA	16.08	36.89	1678.99	28860.57
MMALA	18.59	42.93	1993.76	31953.08
HMC	7.42	15.42	1223.61	13729.68
RMHMC	4.99	10.19	941.38	11939.83

Table R.2: Estimated variance reduction factors (VRFs) for each parameter of the Bayesian logit and probit models applied on the Swiss banknotes data after $n_c = 100$ realized chains, each of post burn-in length $n = 50,000$. The same model is simulated under the logit and probit link functions, and therefore the parameter estimators are directly comparable between the logit and the probit implementation. VRFs are obtained using both linear and quadratic polynomials, corresponding to the columns labelled as L-ZV-MCMC and Q-ZV-MCMC.

	MCMC	L-ZV-MCMC		Q-ZV-MCMC	
	Variance	Variance	VRF	Variance	VRF
θ_1					
MH	0.5619	0.0219	25.65	0.0002	2429.47
MALA	0.4128	0.0242	17.03	0.0004	1078.57
SMMALA	0.5701	0.0277	20.57	0.0006	943.24
MMALA	0.4835	0.0170	28.46	0.0003	1503.32
HMC	0.1711	0.0120	14.30	0.0002	854.04
RMHMC	**0.0727**	**0.0057**	**12.81**	**0.0001**	**714.91**
θ_2					
MH	4.5668	0.3055	14.95	0.0029	1552.23
MALA	5.5498	0.3640	15.25	0.0048	1151.53
SMMALA	3.3729	0.3213	10.50	0.0065	515.48
MMALA	2.3447	0.1930	12.15	0.0037	631.06
HMC	0.9252	0.1825	**5.07**	0.0022	424.87
RMHMC	**0.3807**	**0.0702**	5.42	**0.0010**	**368.14**
θ_3					
MH	2.3059	0.1461	15.79	0.0006	3960.80
MALA	2.2995	0.1607	14.31	0.0011	2185.78
SMMALA	1.8359	0.1709	10.74	0.0012	1471.48
MMALA	1.2912	0.0943	13.70	0.0008	1639.25
HMC	0.4752	0.0819	5.80	0.0005	966.86
RMHMC	**0.2009**	**0.0348**	**5.78**	**0.0003**	**750.29**

Table R.3: Estimated asymptotic variances (rescaled by multiplying by the simulation length $n = 50,000$ of each chain) and variance reduction factors for each parameter of the Bayesian probit model applied on the vaso constriction data after $n_c = 100$ realized chains. The estimated asymptotic variances of both MCMC and ZV-MCMC are smaller for more effective samplers. As a consequence, the VRFs of more effective algorithms are smaller. The variances and VRFs have been rounded to the fourth and second decimal places respectively, while the cells in bold represent the smallest variances and reduction factors across the six samplers demonstrating that RMHMC is the most effective sampler.

	L-ZV-MCMC		Q-ZV-MCMC	
	Logit	Probit	Logit	Probit
		θ_1		
MH	17.59	25.65	865.19	2429.47
MALA	10.90	17.03	465.33	1078.57
SMMALA	18.71	20.57	613.84	943.24
MMALA	24.13	28.46	790.76	1503.32
HMC	13.11	14.30	464.73	854.04
RMHMC	9.22	12.81	369.02	714.91
		θ_2		
MH	6.52	14.95	1613.51	1552.23
MALA	6.86	15.25	1394.05	1151.53
SMMALA	5.70	10.50	625.98	515.48
MMALA	6.47	12.15	724.15	631.06
HMC	2.71	5.07	424.42	424.87
RMHMC	2.56	5.42	293.50	368.14
		θ_3		
MH	6.67	15.79	1088.45	3960.80
MALA	6.46	14.31	792.99	2185.78
SMMALA	5.83	10.74	536.11	1471.48
MMALA	6.79	13.70	642.47	1639.25
HMC	2.87	5.80	352.39	966.86
RMHMC	2.65	5.78	242.09	750.29

Table R.4: Estimated variance reduction factors (VRFs) for each parameter of the Bayesian logit and probit models applied on the vaso constriction data after $n_c = 100$ realized chains, each of post burn-in length $n = 50,000$. The same model is simulated under the logit and probit link functions, and therefore the parameter estimators are directly comparable between the logit and the probit implementation. VRFs are obtained using both linear and quadratic polynomials, corresponding to the columns labelled as L-ZV-MCMC and Q-ZV-MCMC.

	MCMC	ZV-MCMC	$\mathbf{a}$ for L-ZV		$\mathbf{a}$ for Q-ZV	
	Time	Time	Time	%	Time	%
Logit on Swiss banknotes						
MALA	1437.55	1441.90	0.67	0.0465	3.68	0.2552
SMMALA	2880.91	2887.63	1.21	0.0418	5.51	0.1910
MMALA	4426.95	4432.03	0.78	0.0177	4.29	0.0969
HMC	3043.35	3048.40	0.78	0.0256	4.27	0.1402
RMHMC	28479.06	28484.63	0.86	0.0030	4.70	0.0165
Probit on Swiss banknotes						
MALA	6627.71	6634.68	1.28	0.0193	5.69	0.0858
SMMALA	6442.69	6448.35	0.85	0.0132	4.80	0.0745
MMALA	16104.12	16109.87	0.87	0.0054	4.89	0.0303
HMC	14168.16	14175.66	1.38	0.0097	6.12	0.0432
RMHMC	93272.78	93278.47	0.96	0.0010	4.72	0.0051
Logit on vaso constriction data						
MALA	2022.26	2026.92	1.17	0.0577	3.50	0.1725
SMMALA	2384.82	2387.57	0.44	0.0183	2.31	0.0968
MMALA	5429.94	5434.12	0.90	0.0165	3.28	0.0603
HMC	3758.81	3763.25	1.00	0.0265	3.45	0.0916
RMHMC	23119.01	23121.60	0.47	0.0020	2.12	0.0092
Probit on vaso constriction data						
MALA	5052.16	5055.41	0.65	0.0128	2.59	0.0513
SMMALA	5841.20	5844.39	0.61	0.0104	2.58	0.0441
MMALA	12179.25	12182.31	0.52	0.0043	2.55	0.0209
HMC	11772.57	11776.24	0.77	0.0065	2.90	0.0247
RMHMC	72031.10	72033.78	0.53	0.0007	2.16	0.0030

Table R.5: Runtime in seconds of each geometric ZV-MCMC simulation using the Bayesian logit and probit models on the Swiss banknotes and vaso constriction data. The time (in seconds) required for the calculation of the coefficients of linear and quadratic polynomials is, respectively, less than 0.05% and 0.26% of the total ZV-MCMC runtime for each of the five ZV-MCMC algorithms in all tabulated examples.

	MCMC	L-ZV-MCMC		Q-ZV-MCMC	
	Variance	Variance	VRF	Variance	VRF
θ_1					
MH	0.0080	0.0004	21.34	0.0000	380.59
MALA	0.1789	0.0029	62.61	0.0000	4549.38
SMMALA	0.0095	0.0004	25.59	**0.0000**	743.08
MMALA	**0.0065**	**0.0002**	41.65	0.0000	424.64
HMC	0.0079	0.0007	**10.57**	0.0001	**109.26**
RMHMC	0.0393	0.0004	90.84	0.0000	1528.78
θ_2					
MH	**0.2101**	0.0461	4.56	0.0038	54.70
MALA	6.2353	0.9301	6.70	0.0031	2012.30
SMMALA	0.3032	0.0925	**3.28**	**0.0011**	270.58
MMALA	0.2260	0.0510	4.43	0.0043	**52.11**
HMC	2.9346	0.3626	8.09	0.0558	52.63
RMHMC	0.6390	**0.0298**	21.46	0.0020	316.37
θ_3					
MH	0.0393	0.0068	5.77	**0.0002**	163.49
MALA	0.9877	0.1418	6.97	0.0022	444.03
SMMALA	0.0884	0.0237	3.74	0.0011	83.19
MMALA	0.0305	**0.0066**	4.64	0.0007	43.45
HMC	**0.0228**	0.0110	**2.07**	0.0010	**22.13**
RMHMC	0.2059	0.0070	29.31	0.0004	577.33

Table R.6: Estimated asymptotic variances (rescaled by multiplying by the simulation length $n = 10,000$ of each chain) and variance reduction factors for each parameter of the Fitzhugh-Nagumo ODE model after $n_c = 100$ realized chains. Despite the simplification in the implementation, and more specifically without developing a population MCMC framework for the ODE model, the VRFs show that ZV reduces the variance across the simulated chains.