Manifold lifting: scaling Markov chain Monte Carlo to the vanishing noise regime

Khai Xiang Au
National University of Singapore, Singapore
Matthew M. Graham
University College London, UK
Alexandre H. Thiery
National University of Singapore, Singapore
E-mail: khai@u.nus.edu, m.graham@ucl.ac.uk & a.h.thiery@nus.edu.sg

Summary. Standard Markov chain Monte Carlo methods struggle to explore distributions that concentrate in the neighbourhood of low-dimensional submanifolds. This pathology naturally occurs in Bayesian inference settings when there is a high signal-to-noise ratio in the observational data but the model is inherently over-parametrized or non-identifiable. In this paper, we propose a strategy that transforms the original sampling problem into the task of exploring a distribution supported on a manifold embedded in a higher-dimensional space; in contrast to the original posterior this lifted distribution remains diffuse in the limit of vanishing observation noise. We employ a constrained Hamiltonian Monte Carlo method, which exploits the geometry of this lifted distribution, to perform efficient approximate inference. We demonstrate in numerical experiments that, contrarily to competing approaches, the sampling efficiency of our proposed methodology does not degenerate as the target distribution to be explored concentrates near low-dimensional submanifolds. Python code reproducing the results is available at https://doi.org/10.5281/zenodo.6551654.

Keywords: Hamiltonian Monte Carlo; Bayesian inverse problems; non-identifiability.

1. Introduction

Under a Bayesian framework, inference corresponds to deducing the posterior distribution over the configurations of a generative model given observed data and a prior distribution on the model configurations. While we generally expect conditioning on observed data to reduce the uncertainty about the possible model configurations, in most real-world situations we also expect to retain some uncertainty in our posterior beliefs. Accurately quantifying this uncertainty is important in downstream tasks such as using a model to make predictions or comparing how well competing models explain the data.

An obvious source of uncertainty in our posterior beliefs is measurement noise, where observations are imperfectly measured due to, for instance, instrument limitations or human error. Uncertainty also naturally arises in ill-posed or underdetermined problems, that is when the number of degrees of freedom in the unknown

variables in a model exceeds the number of degrees of freedom constrained by the observations. Even in the case of plentiful noiseless observations, uncertainty may still arise due to inherent non-identifiabilities in the model.

In this paper, we consider inference of a vector of *unknown variables* $\boldsymbol{\theta} \in \Theta = \mathbb{R}^{d_\Theta}$ given noisy *observations* $\boldsymbol{y} \in \mathcal{Y} = \mathbb{R}^{d_\mathcal{Y}}$. For ease of exposition, we first consider the special case where the model has additive, isotropic, Gaussian noise of a known variance. The observation model is then defined as

$$\boldsymbol{y} = F(\boldsymbol{\theta}) + \sigma\boldsymbol{\eta} \tag{1}$$

for a vector of *noise variables* $\boldsymbol{\eta} \sim \mathsf{Normal}(\mathbf{0}, \mathbb{I}_{d_\mathcal{Y}})$, *noise scale* $\sigma > 0$ and a *forward function* $F : \Theta \to \mathcal{Y}$, which is generally non-linear and computationally expensive to evaluate. The methodology described in this work can be applied more generally, and we will relax the assumptions of known σ and Gaussian noise in Section 3.

We adopt a Bayesian approach and specify a prior distribution on Θ with unnormalized density $\exp(-\Phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}))$ with respect to the Lebesgue measure. The negative logarithm of the posterior density $\pi^\sigma : \Theta \to \mathbb{R}_{\geq 0}$ then reads

$$-\log \pi^\sigma(\boldsymbol{\theta}) = \Phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \frac{1}{2\sigma^2} \left\|\boldsymbol{y} - F(\boldsymbol{\theta})\right\|^2 + d_\mathcal{Y} \log \sigma + \text{constant}.$$

In the case where the forward function F is linear and the prior distribution Gaussian, the posterior distribution is also Gaussian. In general, however, computations involving the posterior distribution are intractable and it is necessary to employ approximate inference methods.

There are natural situations where the dimension of the observations $d_\mathcal{Y}$ is lower, and often much lower, than the dimension of the unknown variables d_Θ. In these settings, i.e. $d_\mathcal{Y} \ll d_\Theta$, even as the noise scale σ decreases to zero, exact reconstruction of the unknown variables is typically not possible: the system is underdetermined. In *Bayesian inverse problems* (Stuart, 2010; Knapik et al., 2011; Petra et al., 2014), a particularly representative class of models where this type of scenario naturally occurs, the unknown variables $\boldsymbol{\theta}$ typically represent a spatially extended field. The forward function F describes the process of collecting a (typically small) set of measurements derived from $\boldsymbol{\theta}$. For example, in Section 6.5 we consider the problem of reconstructing a thermal conductivity field from a discrete set of noisy measurements of a temperature field; the temperature is obtained from the conductivity as the solution to an elliptic *partial differential equation* (PDE).

Conversely, even when there are ample observations compared to the number of unknowns, i.e. $d_\mathcal{Y} \gg d_\Theta$, structural non-identifiabilities in the model formulation — that multiple values of $\boldsymbol{\theta}$ give the same conditional distribution on observations $\boldsymbol{y}$ (Rothenberg, 1971) — can also mean that the posterior distribution on $\boldsymbol{\theta}$ does not collapse to a point as the observation noise vanishes ($\sigma \to 0$). As an example, in Sections 6.3 and 6.4 we consider the problem of inferring the parameters of *ordinary differential equation* (ODE) models of neuronal dynamics given a sequence of noisy observations. In some cases, such as the FitzHugh–Nagumo model in Section 6.3, while the natural parametrization of a model is such that a subset

of parameters are non-identifiable, reparametrizations can be used to remove non-identifiabilities (Dasgupta et al., 2007). However, such parametrizations may not always be possible to find, even for relatively simple models (Ballnus et al., 2017; Hines et al., 2014). Working with more complex models can exacerbate the challenge of removing non-identifiabilities, as exemplified in the Hodgkin–Huxley ODE model discussed in Section 6.4. In practice, it is also not easy to know if non-identifiabilities are present prior to performing inference (Raue et al., 2013).

Although in the above-mentioned cases it is not possible to exactly reconstruct the quantity of interest as the observation noise vanishes $\sigma \to 0$, the posterior distribution will concentrate on a subset $\mathcal{S} \subset \Theta$ of the latent space given by

$$\mathcal{S} = \{\boldsymbol{\theta} \in \Theta : F(\boldsymbol{\theta}) = \boldsymbol{y}\}.$$

Under mild regularity assumptions on the forward operator F, the subset $\mathcal{S}$ is a submanifold of dimension $d_{\mathcal{S}} = d_\Theta - d_\mathcal{Y}$ embedded in Θ and the bulk of the posterior mass is distributed in a neighbourhood of radius $\mathcal{O}(\sigma)$ around $\mathcal{S}$.

The key computation involving the posterior distribution in downstream tasks is the evaluation of posterior expectations of the form $\int_\Theta \varphi(\boldsymbol{\theta})\pi^\sigma(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ for a test function $\varphi : \Theta \to \mathbb{R}$. We focus in this text on *Markov chain Monte Carlo* (MCMC) methods for approximating such posterior expectations. We assume throughout that derivatives of the forward function F and the prior's negative log-density $\Phi_{\boldsymbol{\theta}}$ can be computed and develop MCMC methods which exploit this derivative information.

Many of the most widely-used MCMC methods for general target distributions on $\mathbb{R}^{d_\Theta}$ such as *random-walk Metropolis* (RWM) with Gaussian proposals, the *Metropolis-adjusted Langevin algorithm* (MALA) (Besag, 1994), and *Hamiltonian Monte Carlo* (HMC) (Duane et al., 1987), require the setting of a step size parameter which controls the scale of the proposed moves. The efficiency of these methods is highly dependent on an appropriate choice of this step size parameter, with overly large steps leading to a high probability of proposed moves being rejected, while overly small steps leading to slow exploration.

In the past decades, much analysis has been done to identify optimal scaling of the step size of various MCMC algorithms as the target distribution dimension becomes large, i.e. the $d_\Theta \to \infty$ asymptotic (Gupta et al., 1990; Roberts et al., 1997; Roberts and Rosenthal, 2001; Beskos et al., 2013). In contrast, comparably little attention has been devoted to the vanishing noise asymptotic, i.e. $\sigma \to 0$. As the posterior concentrates increasingly close to the manifold $\mathcal{S}$, this induces a strong anisotropy in the scaling of the posterior distribution in different directions, with large changes in posterior density in directions normal to $\mathcal{S}$ and much smaller changes in direction tangential to $\mathcal{S}$. Importantly, for non-linear manifolds $\mathcal{S}$, the tangential and normal directions vary across the manifold so that a simple global rescaling will not be sufficient to counteract the anisotropy.

In Beskos et al. (2018) the authors analyse the performance of RWM algorithms in this vanishing noise regime in the specific case where the manifold $\mathcal{S}$ is a linear subspace of Θ. They prove that the step size needs to be scaled linearly with σ to avoid the acceptance probability decreasing to zero when $\sigma \to 0$. As we will describe in the following section, this limitation also applies to gradient-based MCMC methods

such as MALA and HMC. Although Beskos et al. (2018) concentrate on the linear case, they acknowledge that the more practically relevant case is of a non-linear limiting manifold $\mathcal{S}$, and state that an important area for future work is the 'study of MCMC algorithms that better exploit the manifold structure of the support of the target distribution' where the 'manifold can be of smaller dimension than the general space'. The main contribution of this paper is precisely a practical methodology for this setting: we propose an MCMC algorithm which exploits the manifold structure of the target posterior distribution in order to remain computationally efficient in the vanishing noise regime.

The remainder of the paper is structured as follows. We begin in Section 2 by illustrating how standard MCMC methods breakdown in the limit of $\sigma \to 0$ in an illustrative two-dimensional example. In Section 3 we introduce an augmented state space formulation which lifts the target posterior distribution onto a manifold embedded in a higher-dimensional space, with this lifted target distribution not suffering the degenerate anisotropic scaling exhibited in the original space as $\sigma \to 0$. In Section 4 we describe our proposed approach of using a constrained HMC algorithm to generate samples according to this manifold-supported lifted target distribution. In Section 5 we provide some theoretical justification for the approach. Finally, in Section 6 we present the results of numerical experiments which empirically demonstrate that, in contrast to existing MCMC methods, the proposed methodology is robust to low observation noise.

2. Vanishing noise asymptotic regime

In this section we numerically illustrate in a toy example the behaviour of standard MCMC methods in the vanishing noise asymptotic $\sigma \to 0$. Consider a model of the form defined in (1) with $d_\Theta = 2$ and $d_\mathcal{Y} = 1$ and a forward function $F : \mathbb{R}^2 \to \mathbb{R}$

$$F(\boldsymbol{\theta}) = \theta_1^2 + \theta_0^2\big(\theta_0^2 - \tfrac{1}{2}\big). \tag{2}$$

We assume we observe $y = 1$ and place a standard centred Gaussian prior on the unknown parameter, i.e. $\boldsymbol{\theta} \sim \mathsf{Normal}(\mathbf{0}, \mathbb{I}_2)$. The posterior distribution is depicted in Figure 1 for three noise scales $\sigma \in \{0.5, 0.1, 0.02\}$. As $\sigma \to 0$, the posterior distribution can be seen to concentrate in the neighbourhood of the manifold $\mathcal{S}$ implicitly defined by the set of solutions to the equation $F(\boldsymbol{\theta}) = \boldsymbol{y}$.

We consider first the RWM algorithm with isotropic Gaussian proposals with standard deviation (step size) $\varepsilon > 0$. For the proposal to have a non-vanishing acceptance rate, the step size ε needs to be of order $\mathcal{O}(\sigma)$ (see top-left panel in Figure 2). As RWM chains evolve on a diffusive scale, we require $\mathcal{O}(\varepsilon^{-2}) = \mathcal{O}(\sigma^{-2})$ steps to move a $\mathcal{O}(1)$ distance in Θ (Beskos et al., 2018). To informally describe this phenomenon, we state that RWM degenerates at rate σ^{-2} as $\sigma \to 0$.

Maybe surprisingly, despite the use of gradient information, the performance of MALA chains is typically not better than RWM (see top-centre panel in Figure 2). Indeed, since in the neighbourhood of $\mathcal{S}$ the gradient of the (logarithm of the) posterior density is approximately orthogonal to the tangent plane to $\mathcal{S}$, the gradient information by itself does not help MALA chains to explore the directions in the

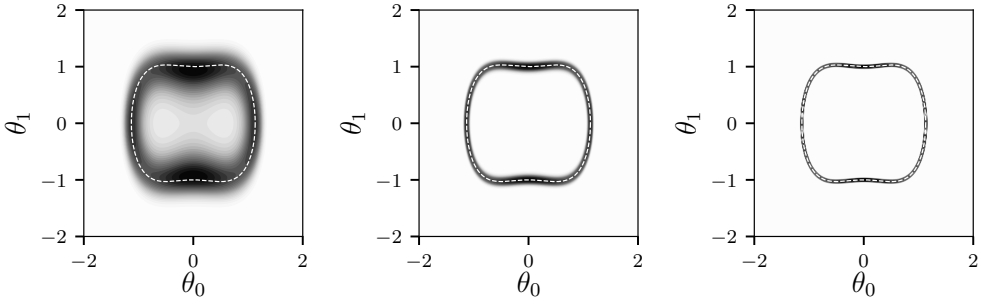

Fig. 1: Posterior distribution for noise scales $\sigma \in \{0.5, 0.1, 0.02\}$ (left, centre, right) in toy example with forward operator $F(\boldsymbol{\theta}) = \theta_1^2 + \theta_0^2\left(\theta_0^2 - \frac{1}{2}\right)$. The heatmaps show the posterior density π^σ with darker colours indicating higher density. The dashed white curves show the limiting manifold $\mathcal{S}$.

parameter space with greatest variation, that is, the directions that are tangential to $\mathcal{S}$. The use of gradient information in settings where the target distribution is highly heterogeneous can in fact often have the negative effect of making performance highly sensitive to the choice of step size (Livingstone and Zanella, 2019), with this effect evident in the example here in the sharper transition in the acceptance rate for the MALA panel in Figure 2 compared to RWM. Further, as the exploration of the parameter space remains diffusive, MALA also degenerates at rate σ^{-2}.

HMC methods with a fixed metric (mass matrix) enjoy a slightly better scaling. The step size parameter $\varepsilon > 0$ of the leapfrog integrator used to generate proposals, like the RWM step size, still needs to be chosen of order $\mathcal{O}(\sigma)$ to maintain a non-zero acceptance rate (see top-right panel in Figure 2). However for proposals generated by simulating $\mathcal{O}(\varepsilon^{-1}) = \mathcal{O}(\sigma^{-1})$ leapfrog steps, the HMC method can move a distance $\mathcal{O}(1)$ in the parameter space. Therefore, HMC degenerates at rate σ^{-1} as $\sigma \to 0$.

Riemannian-manifold Hamiltonian Monte Carlo (RM-HMC) (Girolami and Calderhead, 2011) and related position-dependent MALA (Roberts and Stramer, 2002; Xifara et al., 2014) and RWM (Livingstone, 2015) methods use a locally-varying metric to rescale the target distribution or, equivalently, locally rescale the size of the proposed moves in different directions. For an appropriate choice of metric, which differentially rescales proposed moves tangentially and orthogonally to the limiting manifold, these approaches may seem to offer a potential solution to the performance degeneration of standard RWM, MALA and HMC algorithms as $\sigma \to 0$.

In practice however these approaches still require the use of a vanishing step size as $\sigma \to 0$ in order to maintain a non-zero acceptance rate, as shown in the bottom-left and bottom-centre panels in Figure 2. These two panels show results for RM-HMC using two different metric functions: the expected Fisher information matrix plus prior covariance, as recommended by Girolami and Calderhead (2011) and a 'SoftAbs' regularization of the Hessian of the negative log posterior density as recommended by Betancourt (2013). Both metrics have the desired behaviour of rescaling the steps to make small moves normal to, and large moves tangential to,

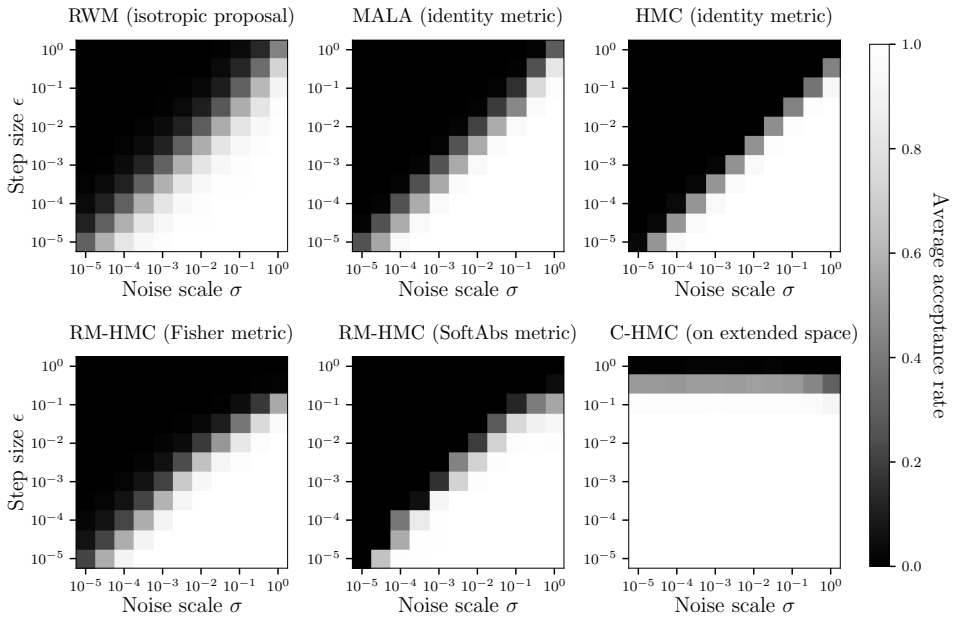

Fig. 2: Average acceptance rate for chains simulated using different MCMC methods targeting the posterior π^{σ} of the toy example for varying noise scales σ and step sizes ε. For the HMC methods $N = 10$ integrator steps were used per sample. The values displayed are means of the acceptance probabilities over four chains of 1000 samples initialised at equispaced points around the limiting manifold $\mathcal{S}$.

the limiting manifold. Despite this, the RM-HMC chains show the same pattern as HMC with a fixed metric of requiring $\varepsilon \propto \sigma$ to maintain a non-vanishing acceptance rate as $\sigma \to 0$. In Section 1 in the Supplementary Material we show that simpler MCMC schemes with position-dependent proposals also exhibit the same behaviour.

At a high-level, these phenomena can be attributed to the fact that, although the corresponding continuous time dynamics underlying RM-HMC and related approaches 'follow' the curvature of the limiting manifold, for a finite discretization step size the simulated trajectories only approximately replicate this idealised dynamic. In particular, as $\sigma \to 0$, smaller steps are needed for the dynamics to remain adequately close to the limiting manifold for the acceptance probability to remain non-zero and to ensure the iterative solvers used in the (implicit) generalized leapfrog integrator required by the RM-HMC algorithm do not diverge.

In contrast to the aforementioned standard HMC and RM-HMC approaches, the *constrained Hamiltonian Monte Carlo* (C-HMC) method proposed in this paper is able to maintain a constant acceptance rate when using a fixed integrator step size ε as $\sigma \to 0$, as shown for the toy example in the bottom-right panel of Figure 2. When combined with the coherent (non-diffusive) exploration of the state space afforded by the Hamiltonian dynamics based proposals, this means the methodology we propose allows efficient MCMC based estimates in the vanishing noise regime.

3. Lifting the posterior distribution onto a manifold

In this section, we describe our approach for constructing an MCMC method which allows us to efficiently approximate expectations with respect to the posterior distribution in situations where the noise scale σ may be small compared to the scale of the observed data $\boldsymbol{y}$. Here, we generalize our discussion from the statistical model (1) to models with σ as a function of $\boldsymbol{\theta}$,

$$\boldsymbol{y} = F(\boldsymbol{\theta}) + \sigma(\boldsymbol{\theta})\,\boldsymbol{\eta}, \tag{3}$$

for a $\mathcal{Y}$-valued noise random variable $\boldsymbol{\eta}$ with negative log-density $\Phi_{\boldsymbol{\eta}}$. Our approach relies on considering an extended state-space $\mathcal{X} = \Theta \times \mathcal{Y} = \mathbb{R}^{d_{\mathcal{X}}}$, with dimension $d_{\mathcal{X}} = d_\Theta + d_{\mathcal{Y}}$, of pairs $\boldsymbol{q} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathcal{X}$ and an extended auxiliary distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q}) = \overline{\pi}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\boldsymbol{\eta})$ that has the posterior distribution $\pi(\mathrm{d}\boldsymbol{\theta})$ as $\boldsymbol{\theta}$-marginal. We then construct a Markov chain which leaves this extended distribution invariant.

Before describing the construction of the extended distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q})$ our methodology is based upon, we recall a few standard results on conditioning of random variables. Consider a random variable $\boldsymbol{Q}$ with density $\mu : \mathcal{X} \to \mathbb{R}_{\geq 0}$ with respect to the Lebesgue measure on $\mathcal{X}$, as well as a continuously differentiable function $C : \mathcal{X} \to \mathcal{Y}$ whose Jacobian matrix $\mathbf{D}C$ has full-rank μ-almost everywhere. A consequence of the co-area formula (Rousset et al., 2010; Diaconis et al., 2013; Graham and Storkey, 2017) gives that, for any test function $\varphi : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{E}[\varphi(\boldsymbol{Q}) \,|\, C(\boldsymbol{Q}) = \mathbf{0}] = \frac{1}{Z} \int_{\mathcal{M}} \varphi(\boldsymbol{q}) \det G(\boldsymbol{q})^{-\frac{1}{2}} \mu(\boldsymbol{q})\, \mathcal{H}(\mathrm{d}\boldsymbol{q}) \tag{4}$$

where $Z = \int_{\mathcal{M}} \det G(\boldsymbol{q})^{-\frac{1}{2}} \mu(\boldsymbol{q})\, \mathcal{H}(\mathrm{d}\boldsymbol{q})$ is a normalization constant, $\mathcal{H}(\mathrm{d}\boldsymbol{q})$ is the $d_{\mathcal{M}} = d_{\mathcal{X}} - d_{\mathcal{Y}} = d_\Theta$ dimensional *Hausdorff measure* on the manifold $\mathcal{M} = C^{-1}(\mathbf{0})$, and $G(\boldsymbol{q}) = \mathbf{D}C(\boldsymbol{q})\mathbf{D}C(\boldsymbol{q})^{\mathsf{T}} \in \mathbb{R}^{\mathrm{d}_{\mathcal{Y}} \times \mathrm{d}_{\mathcal{Y}}}$ is the *Gram matrix* of C. The Gram matrix is positive-definite, and hence non-singular, for any $\sigma > 0$. Informally, this means that the posterior density is equal to the prior density μ, restricted to the manifold $\mathcal{M}$, and with a multiplicative volume term $\det G(\boldsymbol{q})^{-\frac{1}{2}}$ that accounts for the change of measure due to the non-linearity of the constraint function C.

The model (3) with a-priori $\boldsymbol{\eta} \sim \exp(-\Phi_{\boldsymbol{\eta}}(\cdot))$ and $\boldsymbol{\theta} \sim \exp(-\Phi_{\boldsymbol{\theta}}(\cdot))$ corresponds to the case with constraint function C and resulting manifold $\mathcal{M}$

$$\begin{aligned} C(\boldsymbol{q}) \equiv C(\boldsymbol{\theta}, \boldsymbol{\eta}) &= F(\boldsymbol{\theta}) + \sigma(\boldsymbol{\theta})\boldsymbol{\eta} - \boldsymbol{y} \\ \mathcal{M} &= \big\{(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathcal{X} : \boldsymbol{y} = F(\boldsymbol{\theta}) + \sigma(\boldsymbol{\theta})\boldsymbol{\eta}\big\} \end{aligned}$$

while the negative log-density of the pair $\boldsymbol{q} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ on $\mathcal{X}$ then reads

$$-\log \mu(\boldsymbol{q}) \;\equiv\; -\log \mu(\boldsymbol{\theta}, \boldsymbol{\eta}) \;=\; \Phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \Phi_{\boldsymbol{\eta}}(\boldsymbol{\eta}) + \text{constant}.$$

Consequently, Equation (4) implies that the posterior distribution on the pair $\boldsymbol{q} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ has a density $\overline{\pi}(\boldsymbol{\theta}, \boldsymbol{\eta})$ with respect to the Hausdorff measure $\mathcal{H}(\mathrm{d}\boldsymbol{q})$ on the manifold $\mathcal{M} = C^{-1}(\mathbf{0})$ given by

$$-\log \overline{\pi}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \Phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \Phi_{\boldsymbol{\eta}}(\boldsymbol{\eta}) + \tfrac{1}{2} \log \det G(\boldsymbol{\theta}, \boldsymbol{\eta}) + \text{constant}.$$

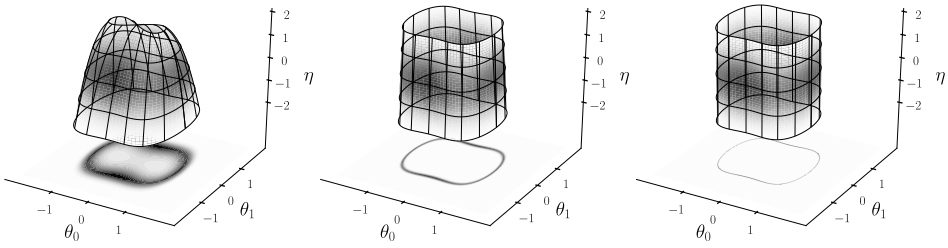

Fig. 3: Manifold $\mathcal{M}$ (wireframe) and density $\overline{\pi}$ (heatmap) in the extended space $\mathcal{X}$ for $\sigma \in \{0.5, 0.1, 0.02\}$ (left-to-right). Only subset of $\mathcal{M}$ with $|\eta| \leq 2$ is shown. The density π on Θ is shown below each manifold for comparison.

In the running example (2), the original parameter space Θ is two-dimensional and, as $\sigma \to 0$, the posterior distribution concentrates in a neighbourhood of the one-dimensional manifold $\mathcal{S}$. Figure 3 shows the two-dimensional manifold $\mathcal{M}$ embedded in the three-dimensional extended space $\mathcal{X}$ and the lifted posterior density $\overline{\pi}$ on the manifold for $\sigma \in \{0.5, 0.1, 0.02\}$. It can be seen that though the geometry of the manifold does change as $\sigma \to 0$, unlike the posterior distribution $\pi(\mathrm{d}\boldsymbol{\theta})$ in the original space Θ, the lifted posterior distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q})$ does not concentrate and remains diffuse. Furthermore, note that in the limit $\sigma \to 0$ the manifold $\mathcal{M}$ tends to the product manifold $\{(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \Theta \times \mathcal{Y} : \boldsymbol{y} = F(\boldsymbol{\theta})\} = \mathcal{S} \times \mathcal{Y}$, and $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ become independent under the lifted posterior distribution. This phenomenon partly motivates why our proposed methodology does not degenerate as $\sigma \to 0$.

In summary, our methodology consists in lifting the posterior distribution $\pi(\mathrm{d}\boldsymbol{\theta})$ that is concentrated in the neighbourhood of a set $\mathcal{S} \subset \Theta$ onto a manifold $\mathcal{M}$ in a higher-dimensional space $\mathcal{X} = \Theta \times \mathcal{Y}$. The task is then to sample from the lifted distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q})$, which is supported on $\mathcal{M}$ and importantly remains diffuse in the vanishing noise limit $\sigma \to 0$. We propose to use a constrained HMC algorithm to efficiently sample from the lifted distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q})$ using first- and second-order derivative information. As will be numerically demonstrated in Section 6, the resulting sampler does not degenerate as $\sigma \to 0$.

4. Exploring the manifold-supported extended distribution

In this section, we describe the constrained HMC method we use to construct an ergodic Markov chain leaving the lifted distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q})$ invariant. The use of simulated constrained Hamiltonian dynamics within a HMC scheme has been proposed multiple times before, see for example Hartmann and Schütte (2005); Rousset et al. (2010); Brubaker et al. (2012). Our formulation most closely follows that described in Graham and Storkey (2017), which along with Lelièvre et al. (2019), applies an idea proposed in Zappa et al. (2018) to ensure the overall reversibility of the C-HMC transitions despite the use of a numerical integrator with implicit steps that may fail to converge to a unique solution.

For conceptual clarity, we proceed by first giving a very high-level description of

Algorithm 1 Unconstrained leapfrog integrator

Inputs: $\boldsymbol{p}_0$ current momentum, $\boldsymbol{q}_0$ current position, ε step size.
Outputs: $\boldsymbol{p}_1$ updated momentum, $\boldsymbol{q}_1$ updated position.

function LEAPFROGSTEP($\boldsymbol{p}_0, \boldsymbol{q}_0, \varepsilon$)
  $\boldsymbol{p}_{1/2} \leftarrow \boldsymbol{p}_0 - \frac{\varepsilon}{2}\nabla U(\boldsymbol{q}_0)$ ▷ Momentum half-step update
  $\boldsymbol{q}_1 \leftarrow \boldsymbol{q}_0 + \varepsilon \boldsymbol{p}_{1/2}$ ▷ Position full-step update
  $\boldsymbol{p}_1 \leftarrow \boldsymbol{p}_{1/2} - \frac{\varepsilon}{2}\nabla U(\boldsymbol{q}_1)$ ▷ Momentum half-step update
  return $\boldsymbol{p}_1, \boldsymbol{q}_1$

the standard HMC methodology (Duane et al., 1987; Neal, 2011) in order to draw distinctions with the C-HMC approach. In order to explore a $\mathcal{X}$-valued distribution with negative log-density $U(\boldsymbol{q})$, the HMC algorithm introduces an auxiliary, and independent, momentum variable $\boldsymbol{p} \in \mathcal{X}$ with standard Gaussian distribution. This defines an extended distribution on $\mathcal{X} \times \mathcal{X}$ with negative log-density equal, up to an irrelevant additive constant, to the Hamiltonian $H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + \frac{1}{2}\|\boldsymbol{p}\|^2$. Proposed moves in the extended state space $\mathcal{X} \times \mathcal{X}$ are computed by numerically integrating the Hamiltonian dynamics $(\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}}) = (\nabla_2, -\nabla_1)H(\boldsymbol{q}, \boldsymbol{p})$ forward in time and then accepting or rejecting the final state in a Rosenbluth–Teller–Metropolis acceptance step (Metropolis et al., 1953). To ensure ergodicity these dynamics based updates are interleaved with resampling of the momentum variable. The Hamiltonian dynamics is simulated using a reversible and volume preserving integrator, a common choice being the standard *leapfrog* or Störmer–Verlet integrator with a fixed step size $\varepsilon > 0$ (see Algorithm 1). More sophisticated integrators (Leimkuhler and Reich, 2004; Blanes et al., 2014) are also possible.

The constrained HMC algorithm allows one to instead explore a distribution supported on an embedded manifold $\mathcal{M} = C^{-1}(\mathbf{0}) \subset \mathcal{X}$ implicitly defined by a constraint function $C : \mathcal{X} \to \mathcal{Y}$.

At any point $\boldsymbol{q} \in \mathcal{M}$ we can define two orthogonal linear subspaces: the *tangent space*

$$\mathcal{T}_{\mathcal{M}}(\boldsymbol{q}) = \left\{\boldsymbol{p} \in \mathbb{R}^{d_{\mathcal{X}}} \mid \mathbf{D}C(\boldsymbol{q})\,\boldsymbol{p} = \mathbf{0}\right\},$$

which is the space of directions tangential to manifold at $\boldsymbol{q}$ and the *normal space*

$$\mathcal{N}_{\mathcal{M}}(\boldsymbol{q}) = \{\boldsymbol{n} \in \mathbb{R}^{d_{\mathcal{X}}} \mid \exists \boldsymbol{\lambda} \in \mathbb{R}^{d_{\mathcal{Y}}} : \boldsymbol{n} = \mathbf{D}C(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{\lambda}\},$$

which is the space of directions normal to the manifold at $\boldsymbol{q}$. Differentiating the constraint equation with respect to time gives us that $\mathbf{D}C(\boldsymbol{q})\,\dot{\boldsymbol{q}} = \mathbf{D}C(\boldsymbol{q})\,\boldsymbol{p} = \mathbf{0}$, i.e. the momentum $\boldsymbol{p}$ (equivalent to the velocity under an assumption of an identity metric) is in the tangent space $\mathcal{T}_{\mathcal{M}}(\boldsymbol{q})$ at $\boldsymbol{q} \in \mathcal{M}$ at all time steps.

Constrained HMC operates on the same principles as standard HMC, however additional projection steps are necessary in the integrator used to ensure that the position $\boldsymbol{q}$ remains on the manifold $\mathcal{M}$ and that the momentum $\boldsymbol{p}$ remains in the tangent space $\mathcal{T}_{\mathcal{M}}(\boldsymbol{q})$ at all time steps. The resulting *constrained leapfrog integrator* (Leimkuhler and Matthews, 2016), summarised in Algorithm 2, is a variant of the *RATTLE integrator* (Andersen, 1983) with an additional intermediate momentum

Algorithm 2 Constrained leapfrog integrator

Inputs: $\{\boldsymbol{p}_0, \boldsymbol{q}_0, \varepsilon\}$ see Algorithm 1, τ convergence tolerance and M maximum iterations for position projection.

Outputs: $\{\boldsymbol{p}_1, \boldsymbol{q}_1\}$ see Algorithm 1, FAILED flag for failed position projection.

function CONSTRAINEDSTEP($\boldsymbol{p}_0, \boldsymbol{q}_0, \varepsilon, \tau, M$)

$\tilde{\boldsymbol{p}}_{1/2} \leftarrow \boldsymbol{p}_0 - \frac{\varepsilon}{2}\nabla U(\boldsymbol{q}_0)$ ▷ Unconstrained momentum half-step update

$\bar{\boldsymbol{p}}_{1/2} \leftarrow$ PROJECTMOMENTUM($\tilde{\boldsymbol{p}}_{1/2}, \boldsymbol{q}_0$) ▷ Project $\tilde{\boldsymbol{p}}_{1/2}$ onto $\mathcal{T}_{\mathcal{M}}(\boldsymbol{q}_0)$ (see §4.1)

$\tilde{\boldsymbol{q}}_1 \leftarrow \boldsymbol{q}_0 + \varepsilon\bar{\boldsymbol{p}}_{1/2}$ ▷ Unconstrained position full-step update

$\boldsymbol{q}_1$, FAILED $\leftarrow$ PROJECTPOSITION($\tilde{\boldsymbol{q}}_1, \boldsymbol{q}_0, \tau, M$) ▷ Project $\tilde{\boldsymbol{q}}_1$ onto $\mathcal{M}$ (see §4.2)

$\boldsymbol{p}_{1/2} = \frac{1}{\varepsilon}(\boldsymbol{q}_1 - \boldsymbol{q}_0)$ ▷ Compute $\boldsymbol{p}_{1/2}$ such that $\boldsymbol{q}_1 = \boldsymbol{q}_0 + \varepsilon\boldsymbol{p}_{1/2}$

$\tilde{\boldsymbol{p}}_1 \leftarrow \boldsymbol{p}_{1/2} - \frac{\varepsilon}{2}\nabla U(\boldsymbol{q}_1)$ ▷ Unconstrained momentum half-step update

$\boldsymbol{p}_1 \leftarrow$ PROJECTMOMENTUM($\tilde{\boldsymbol{p}}_1, \boldsymbol{q}_1$) ▷ Project $\tilde{\boldsymbol{p}}_1$ onto $\mathcal{T}_{\mathcal{M}}(\boldsymbol{q}_1)$

return $\boldsymbol{p}_1, \boldsymbol{q}_1$, FAILED

projection, and, subject to an appropriate projection function, is symplectic and so volume-preserving (Leimkuhler and Skeel, 1994; Reich, 1996). As will be described in Section 4.3, additional *reversibility checks* ensure the reversibility of the method.

4.1. Momentum projection

The constrained integrator in Algorithm 2, uses a function PROJECTMOMENTUM, which orthogonally projects a momentum $\tilde{\boldsymbol{p}} \in \mathbb{R}^{d_{\mathcal{X}}}$ onto $\mathcal{T}_{\mathcal{M}}(\boldsymbol{q})$ at a position $\boldsymbol{q} \in \mathcal{M}$. The projected momentum $\boldsymbol{p}$ can be expressed as $\boldsymbol{p} = \tilde{\boldsymbol{p}} - \mathbf{D}C(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{\lambda}$ for a vector $\boldsymbol{\lambda} \in \mathbb{R}^{d_{\mathcal{Y}}}$ as the rows of $\mathbf{D}C(\boldsymbol{q})$ span $\mathcal{N}_{\mathcal{M}}(\boldsymbol{q})$. Solving $\mathbf{D}C(\boldsymbol{q})(\tilde{\boldsymbol{p}} - \mathbf{D}C(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{\lambda}) = \mathbf{0}$ for $\boldsymbol{\lambda}$ gives that

$$\boldsymbol{p} = \text{PROJECTMOMENTUM}(\tilde{\boldsymbol{p}}, \boldsymbol{q}) = \big(\mathbb{I}_{d_{\mathcal{X}}} - \mathbf{D}C(\boldsymbol{q})^{\mathsf{T}} G(\boldsymbol{q})^{-1}\mathbf{D}C(\boldsymbol{q})\big)\tilde{\boldsymbol{p}}. \tag{5}$$

As the Gram matrix $G(\boldsymbol{q})$ is non-singular, the projection in (5) is well-defined.

4.2. Position projection

For a momentum $\boldsymbol{p} \in \mathcal{T}_{\mathcal{M}}(\boldsymbol{q})$ an update to the position $\tilde{\boldsymbol{q}} \leftarrow \boldsymbol{q} + \varepsilon\boldsymbol{p}$ will move tangentially to the manifold. However, for a non-linear manifold $\mathcal{M}$ the updated position $\tilde{\boldsymbol{q}}$ in general will not lie on the manifold. To map $\tilde{\boldsymbol{q}}$ to a point on the manifold we project $\tilde{\boldsymbol{q}} \in \mathcal{X}$ along $\mathcal{N}_{\mathcal{M}}(\boldsymbol{q})$, i.e. the normal space at the previous point $\boldsymbol{q} \in \mathcal{M}$. In other words, we need to solve the non-linear equation in $\boldsymbol{\lambda} \in \mathbb{R}^{d_{\mathcal{Y}}}$

$$\chi(\boldsymbol{\lambda}) = C\big(\tilde{\boldsymbol{q}} - \mathbf{D}C(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{\lambda}\big) = \mathbf{0}. \tag{6}$$

A range of numerical methods can potentially be used for solving Equation (6). Applying Newton's method gives the iterative update,

$$\boldsymbol{\lambda}_{m+1} = \boldsymbol{\lambda}_m + \big(\mathbf{D}C(\boldsymbol{q}_m)\mathbf{D}C(\boldsymbol{q})^{\mathsf{T}}\big)^{-1}\chi(\boldsymbol{\lambda}_m), \quad \boldsymbol{q}_m = \tilde{\boldsymbol{q}} - \mathbf{D}C(\boldsymbol{q})^{\mathsf{T}}\boldsymbol{\lambda}_m. \tag{7}$$

This requires evaluating the constraint function and its Jacobian on each iteration, and solving a linear system in a $d_{\mathcal{Y}} \times d_{\mathcal{Y}}$ matrix.

An alternative quasi-Newton approach, proposed by Barth et al. (1995) and termed the *symmetric Newton method*, iterates the update

$$\boldsymbol{\lambda}_{m+1} = \boldsymbol{\lambda}_m + G(\boldsymbol{q})^{-1}\chi(\boldsymbol{\lambda}_m).$$

In contrast to the full Newton's update in (7) the matrix $G(\boldsymbol{q})$ solved for here is fixed over all iterations. This is motivated by the observation that $\mathbf{D}C(\boldsymbol{q}_m)$ often stays largely constant over the Newton iterations (Barth et al., 1995). As a single decomposition of the Gram matrix can be used for all iterations and the constraint Jacobian does not need to be recomputed on each iteration, this approach can have significantly lower per-iteration computational costs. However, if the constraint Jacobian does vary significantly, the symmetric Newton method can take more iterations to converge than Newton's method or fail to converge at all.

In practice, we stop the non-linear solvers as soon as the norm $\|\chi(\boldsymbol{\lambda}_m)\|_\infty = \|C(\boldsymbol{q}_m)\|_\infty$ falls below a tolerance threshold τ, or a maximum number M of iterations is reached. We suggest setting τ such that the projection error is comparable to the numerical error already incurred from using floating-point arithmetic. Furthermore, as mentioned in Zappa et al. (2018), we also recommend setting the maximum number of iterations M relatively small. In doing so, we avoid wasting computational effort on proposals whose position projections take abnormally long to converge. In our work, we set $\tau = 10^{-9}$ and $M = 50$, and the average number of iterations to convergence in practice was around 5 to 10 in our experiments. In the case when the numerical solver does not converge, which does happen for a non-negligible fraction of the proposals, the proposal is rejected.

We make an early remark that in most of our experiments using the full Newton's method yielded better sampling performance than the quasi-Newton approach, even after taking into account the relatively higher cost per iteration. We discuss this in more details in Section 6. As explained in the next section, the possible non-convergence of the numerical solver, as well as the possible existence of multiple solutions to the non-linear equation (6), require us to perform some additional checks to avoid compromising the correctness of the overall method.

4.3. Reversibility check

The correctness of the standard HMC method is ensured by the reversibility and volume preservation of the leapfrog integrator. Although the constrained integrator is also volume-preserving, the reversibility property needs to be checked more carefully. This is because the non-linear equation (6) that needs to be solved to implement the PROJECTPOSITION function may not in general have a unique solution, and the numerical method used to find such a solution may not converge within the prescribed number of iterations.

To enforce reversibility we follow the scheme proposed by Zappa et al. (2018) (for a constrained RWM algorithm) by manually checking if each constrained step constitutes a reversible map. For each forward step $(\boldsymbol{q}, \boldsymbol{p}) \to (\boldsymbol{q}', \boldsymbol{p}')$, we run the step in reverse (by negating the time step) $(\boldsymbol{q}', \boldsymbol{p}') \to (\boldsymbol{q}^*, \boldsymbol{p}^*)$. We proceed with the next step only if the current step is reversible, i.e., $(\boldsymbol{q}, \boldsymbol{p}) = (\boldsymbol{q}^*, \boldsymbol{p}^*)$ to within a

numerical tolerance ρ, *and* neither of the iterative solves in the position projections in the forward nor reverse steps failed to converge (indicated by the FAILED flag returned by the PROJECTPOSITION function in Algorithm 2 being TRUE); otherwise the trajectory terminates early and we reject. By construction this reversibility check guarantees the simulated trajectory defines a reversible map.

4.4. Complete algorithm

Recall that the lifted target distribution $\overline{\pi}(\mathrm{d}\boldsymbol{q})$ is supported on the manifold $\mathcal{M}$ and has a density given by

$$\frac{\mathrm{d}\overline{\pi}}{\mathrm{d}\mathcal{H}}(\boldsymbol{q}) = \frac{d\overline{\pi}}{\mathrm{d}\mathcal{H}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \propto \exp\Big(- \Phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \Phi_{\boldsymbol{\eta}}(\boldsymbol{\eta}) - \tfrac{1}{2} \log\det G(\boldsymbol{\theta}, \boldsymbol{\eta}) \Big).$$

Consequently, for C-HMC with an identity metric, the Hamiltonian reads

$$H(\boldsymbol{q}, \boldsymbol{p}) = H\big((\boldsymbol{\theta}, \boldsymbol{\eta}), \boldsymbol{p}\big) = \Phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \Phi_{\boldsymbol{\eta}}(\boldsymbol{\eta}) + \tfrac{1}{2} \log\det G(\boldsymbol{\theta}, \boldsymbol{\eta}) + \tfrac{1}{2} \|\boldsymbol{p}\|^2.$$

For initializing the C-HMC algorithm, one can choose an arbitrary initial $\theta_0 \in \Theta$ (e.g. by sampling from the prior) and map to the corresponding unique extended state $\boldsymbol{q}_0 = (\boldsymbol{\theta}_0, (\boldsymbol{y} - F(\boldsymbol{\theta}_0))/\sigma(\boldsymbol{\theta}_0)) \in \mathcal{M}$. Algorithm 3 provides a complete description of a single chain transition of the C-HMC method.

In general, the linear algebra operations involved in each constrained leapfrog integrator step will have a $\mathcal{O}(\min(d_\Theta d_\mathcal{Y}^2, d_\mathcal{Y} d_\Theta^2))$ complexity and will also require evaluation of the constraint Jacobian at $\mathcal{O}(\min(d_\Theta, d_\mathcal{Y}))$ times the cost of evaluating the forward function F (see Section 2 in the Supplementary Material for details). While these costs are significant compared to, for example, the cost of each leapfrog integrator step in the standard HMC algorithm, in the numerical experiments in Section 6 we will see that the ability to use a much larger integrator size and so take fewer integrator steps will often outweigh the increased per-step costs.

5. Lifted Hamiltonian dynamic in original latent space

To provide some theoretical insight to our approach, we now show that the lifted Hamiltonian dynamic is equivalent to a Hamiltonian dynamic on the original latent space under a particular choice of metric (mass matrix). For linear-Gaussian models the lifted dynamic corresponds to an 'optimal' choice of setting the metric to the posterior precision matrix. For general models, the lifted dynamic corresponds to a Hamiltonian dynamic with a position-dependent metric that accounts for the locally varying scale and curvature of the distribution, analogously to RM-HMC.

5.1. Linear-Gaussian models

We first consider the special case of a linear-Gaussian model with fixed $\sigma > 0$ of the form

$$\boldsymbol{y} = F\boldsymbol{\theta} + \boldsymbol{f} + \sigma\boldsymbol{\eta}, \quad \boldsymbol{\theta} \sim \mathsf{Normal}(\mathbf{0}, \mathbb{I}), \quad \boldsymbol{\eta} \sim \mathsf{Normal}(\mathbf{0}, \mathbb{I}).$$

Algorithm 3 C-HMC transition

Inputs: $\boldsymbol{q}$ current position, N number of integrator steps per transition, ρ reversibility check tolerance, $\{\varepsilon, \tau, M\}$ see Algorithm 2.

Outputs: $\boldsymbol{q}'$ updated position such that $\boldsymbol{q} \sim \overline{\pi}(\cdot) \implies \boldsymbol{q}' \sim \overline{\pi}(\cdot)$.

function ConstrainedHMC($\boldsymbol{q}, \varepsilon, \tau, \rho, N, M$)

- $\boldsymbol{q}_0 \leftarrow \boldsymbol{q}$
- $\tilde{\boldsymbol{p}}_0 \leftarrow \mathsf{Normal}(\mathbf{0}, \mathbb{I}_{d_{\mathcal{X}}})$ ▷ Sample momentum from standard Gaussian
- $\boldsymbol{p}_0 \leftarrow$ projectMomentum($\tilde{\boldsymbol{p}}_0, \boldsymbol{q}_0$) ▷ Project initial momentum onto $\mathcal{T}_{\mathcal{M}}(\boldsymbol{q}_0)$
- **for** $n \in \{1, \ldots, N\}$ **do**
  - $\boldsymbol{q}_n, \boldsymbol{p}_n,$ fwFailed $\leftarrow$ ConstrainedStep($\boldsymbol{q}_{n-1}, \boldsymbol{p}_{n-1}, \varepsilon, \tau, M$) ▷ Forward step
  - $\boldsymbol{q}_*, \boldsymbol{p}_*,$ rvFailed $\leftarrow$ ConstrainedStep($\boldsymbol{q}_n, \boldsymbol{p}_n, -\varepsilon, \tau, M$) ▷ Reverse step
  - **if** fwFailed **or** rvFailed **or** $\|\boldsymbol{q}_* - \boldsymbol{q}_{n-1}\|_\infty \geq \rho$ **then** ▷ Reversibility check
    - return $\boldsymbol{q}$ ▷ Reject proposal if non-reversible
- **if** uniform$(0,1) < \exp\big(H(\boldsymbol{q}_0, \boldsymbol{p}_0) - H(\boldsymbol{q}_n, \boldsymbol{p}_n)\big)$ **then** ▷ Acceptance step
  - return $\boldsymbol{q}_N$ ▷ Accept proposal
- **else**
  - return $\boldsymbol{q}$ ▷ Reject proposal

Remark 1. *Any linear-Gaussian model of the more general form*

$$\boldsymbol{y} = F'\boldsymbol{\theta}' + \boldsymbol{f}' + \sigma\boldsymbol{\eta}, \quad \boldsymbol{\theta}' \sim \mathsf{Normal}(\boldsymbol{r}, C), \quad \boldsymbol{\eta} \sim \mathsf{Normal}(\mathbf{0}, \mathbb{I})$$

can be reparametrized into the form above by defining $\boldsymbol{\theta} = L^{-1}(\boldsymbol{\theta}' - \boldsymbol{r})$, $F = F'L$ *and* $\boldsymbol{f} = \boldsymbol{f}' - F'\boldsymbol{r}$ *for any* L *such that* $LL^{\mathsf{T}} = C$.

The posterior distribution in this case can be derived analytically as

$$\boldsymbol{\theta} \,|\, \boldsymbol{y} \sim \mathsf{Normal}(\boldsymbol{\mu}, \Sigma), \quad \Sigma = (\mathbb{I} + \tfrac{1}{\sigma^2} F^{\mathsf{T}} F)^{-1}, \quad \boldsymbol{\mu} = \tfrac{1}{\sigma^2}\Sigma F^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{f}).$$

We can define a Hamiltonian dynamic on Θ, the flow map of which marginally preserves the posterior, by first introducing a momentum variable $\boldsymbol{m} \in \mathbb{R}^{d_\Theta}$ which is independent of $\boldsymbol{\theta}$ and with marginal distribution $\mathsf{Normal}(\mathbf{0}, M)$ for some positive-definite matrix M. The *Hamiltonian* is then equal to the negative logarithm of the joint density on $(\boldsymbol{\theta}, \boldsymbol{m})$,

$$H(\boldsymbol{\theta}, \boldsymbol{m}) = \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}) + \tfrac{1}{2}\boldsymbol{m}^{\mathsf{T}}M^{-1}\boldsymbol{m},$$

with corresponding Hamiltonian dynamics described by the system of ODEs

$$\dot{\boldsymbol{\theta}} = M^{-1}\boldsymbol{m}, \qquad \dot{\boldsymbol{m}} = -\Sigma^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}). \tag{8}$$

Alternatively, if we lift the posterior distribution onto the manifold

$$\mathcal{M} = \{\boldsymbol{\theta} \in \mathbb{R}^{d_\Theta}, \boldsymbol{\eta} \in \mathbb{R}^{d_\mathcal{Y}} : C(\boldsymbol{\theta}, \boldsymbol{\eta}) = F\boldsymbol{\theta} + \boldsymbol{f} + \sigma\boldsymbol{\eta} - \boldsymbol{y} = 0\}$$

then lifted posterior distribution $\bar{\pi}$ has density

$$\bar{\pi}(\boldsymbol{\theta}, \boldsymbol{\eta}) \propto \exp\left(-\tfrac{1}{2}\|\boldsymbol{\theta}\|^2 - \tfrac{1}{2}\|\boldsymbol{\eta}\|^2\right) \left|FF^{\mathsf{T}} + \sigma^2\mathbb{I}\right|^{-\frac{1}{2}}$$

with respect to the d_Θ-dimensional Hausdorff measure on $\mathcal{M}$. Introducing momenta $(\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}}) \in \mathcal{T}_{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\eta})$ with density $\exp(-\frac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\theta}}\|^2 - \frac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\eta}}\|^2)$ with respect to the d_Θ-dimensional Hausdorff measure on $\mathcal{T}_{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\eta})$ and a vector of Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{d_y}$, implicitly defined by the constraint $C(\boldsymbol{\theta}, \boldsymbol{\eta}) = 0$, the Hamiltonian for the resulting constrained system is

$$\bar{H}((\boldsymbol{\theta}, \boldsymbol{\eta}), (\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}})) = \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \tfrac{1}{2}\|\boldsymbol{\eta}\|^2 + \tfrac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\theta}}\|^2 + \|\boldsymbol{p}_{\boldsymbol{\eta}}\|^2 + (F\boldsymbol{\theta} + \boldsymbol{f} + \sigma\boldsymbol{\eta} - \boldsymbol{y})^{\mathsf{T}}\boldsymbol{\lambda}$$

with corresponding constrained Hamiltonian dynamics described by the system of differential algebraic equations

$$\begin{aligned} &\dot{\boldsymbol{\theta}} = \boldsymbol{p}_{\boldsymbol{\theta}}, \quad \dot{\boldsymbol{\eta}} = \boldsymbol{p}_{\boldsymbol{\eta}}, \quad \dot{\boldsymbol{p}}_{\boldsymbol{\theta}} = -\boldsymbol{\theta} - F^{\mathsf{T}}\boldsymbol{\lambda}, \quad \dot{\boldsymbol{p}}_{\boldsymbol{\eta}} = -\boldsymbol{\eta} - \sigma\boldsymbol{\lambda} \\ &\text{subject to} \quad F\boldsymbol{\theta} + \boldsymbol{f} + \sigma\boldsymbol{\eta} - \boldsymbol{y} = \boldsymbol{0} \ \text{ and } \ F\boldsymbol{p}_{\boldsymbol{\theta}} + \sigma\boldsymbol{p}_{\boldsymbol{\eta}} = 0. \end{aligned} \tag{9}$$

The flow map corresponding to these dynamics marginally preserves the lifted distribution $\bar{\pi}$ on $(\boldsymbol{\theta}, \boldsymbol{\eta})$ and so the original posterior π on $\boldsymbol{\theta}$. We are now in a position to state our first result (proof in Appendix A).

THEOREM 1. *The continuous-time constrained Hamiltonian dynamics of $\boldsymbol{\theta}$ in (9) are equivalent to the Hamiltonian dynamics of $\boldsymbol{\theta}$ in (8) with the metric set to the posterior precision matrix of $\boldsymbol{\theta}$ under π, i.e. $M = \Sigma^{-1}$.*

This result illustrates that for linear-Gaussian models, the lifting approach results in continuous-time Hamiltonian dynamics corresponding to those produced by the 'optimal' choice of metric $M = \Sigma^{-1}$, that is resulting in dynamics corresponding to decoupled simple harmonic motions of unit angular frequency in each component of $\boldsymbol{\theta}$. Importantly this result is dependent on the model being parametrized such that $\boldsymbol{\theta}$ has a standard normal prior; for models of the more general form in Remark 1 with a non-standard normal prior, the result in Theorem 1 does not hold without reparametrization. This motivates our advice to use a standard normal prior parametrization where possible when using the manifold lifting approach. For models with Gaussian priors this can be equivalently achieved by using a non-identity metric in the constrained Hamiltonian dynamics however to avoid introducing unnecessary additional notation we assume an identity metric for the constrained dynamics here.

5.2. More general models

We now consider a more general model of the form

$$\boldsymbol{y} = F(\boldsymbol{\theta}) + \sigma(\boldsymbol{\theta})\boldsymbol{\eta}, \quad \boldsymbol{\theta} \sim \mathsf{Normal}(\boldsymbol{0}, \mathbb{I}), \quad \boldsymbol{\eta} \sim \mathsf{Normal}(\boldsymbol{0}, \mathbb{I}),$$

where F and σ are both potentially non-linear functions of latent variables $\boldsymbol{\theta}$. If we lift the posterior distribution onto the manifold

$$\mathcal{M} = \{\boldsymbol{\theta} \in \mathbb{R}^{d_\Theta}, \boldsymbol{\eta} \in \mathbb{R}^{d_y} : C(\boldsymbol{\theta}, \boldsymbol{\eta}) = F(\boldsymbol{\theta}) + \sigma(\boldsymbol{\theta})\boldsymbol{\eta} - \boldsymbol{y} = \boldsymbol{0}\}$$

then the lifted posterior distribution $\bar{\pi}$ has density

$$\bar{\pi}(\boldsymbol{\theta}, \boldsymbol{\eta}) \propto \exp\left(-\tfrac{1}{2}\|\boldsymbol{\theta}\|^2 - \tfrac{1}{2}\|\boldsymbol{\eta}\|^2\right) |G(\boldsymbol{\theta}, \boldsymbol{\eta})|^{-\frac{1}{2}}$$

with respect to the d_Θ-dimensional Hausdorff measure on $\mathcal{M}$, where

$$G(\boldsymbol{\theta}, \boldsymbol{\eta}) = (\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}))(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}))^\mathsf{T} + \sigma(\boldsymbol{\theta})^2\mathbb{I}.$$

Introducing momenta $(\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}}) \in \mathcal{T}_{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\eta})$ with density $\exp(-\frac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\theta}}\|^2 - \frac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\eta}}\|^2)$ with respect to the d_Θ-dimensional Hausdorff measure on $\mathcal{T}_{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\eta})$, the Hamiltonian function for the constrained system is then

$$\bar{H}((\boldsymbol{\theta}, \boldsymbol{\eta}), (\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}})) = \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \tfrac{1}{2}\|\boldsymbol{\eta}\|^2 + \tfrac{1}{2}\log|G(\boldsymbol{\theta}, \boldsymbol{\eta})| + \tfrac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\theta}}\|^2 + \tfrac{1}{2}\|\boldsymbol{p}_{\boldsymbol{\eta}}\|^2 + C(\boldsymbol{\theta}, \boldsymbol{\eta})^\mathsf{T}\boldsymbol{\lambda}.$$

Alternatively, working with the posterior distribution π which has Lebesgue density

$$\pi(\boldsymbol{\theta}) \propto \exp\left(-\tfrac{1}{2\sigma(\boldsymbol{\theta})^2}(\boldsymbol{y} - F(\boldsymbol{\theta}))^\mathsf{T}(\boldsymbol{y} - F(\boldsymbol{\theta})) - d_{\mathcal{Y}}\log\sigma(\boldsymbol{\theta}) - \tfrac{1}{2}\|\boldsymbol{\theta}\|^2\right),$$

if we introduce a momentum $\boldsymbol{m} \in \mathbb{R}^{d_\Theta}$ with conditional distribution $\mathsf{Normal}(0, M(\boldsymbol{\theta}))$ given $\boldsymbol{\theta}$ for a positive-definite matrix valued function M, then we can define a Hamiltonian equal to the negative logarithm of the joint density on $\boldsymbol{\theta}$ and $\boldsymbol{m}$

$$H(\boldsymbol{\theta}, \boldsymbol{m}) = \tfrac{1}{2\sigma(\boldsymbol{\theta})^2}\|\boldsymbol{y} - F(\boldsymbol{\theta})\|^2 + d_{\mathcal{Y}}\log\sigma(\boldsymbol{\theta}) + \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \tfrac{1}{2}\boldsymbol{m}M(\boldsymbol{\theta})^{-1}\boldsymbol{m} + \tfrac{1}{2}\log|M(\boldsymbol{\theta})|.$$

We then have that the constrained Hamiltonian dynamics on $\mathcal{M}$ are equivalent to the Hamiltonian dynamics on Θ with a particular *position-dependent* choice of M (proof in Appendix B).

THEOREM 2. *The continuous-time constrained Hamiltonian dynamics of $\boldsymbol{\theta}$ in the lifted system are equivalent to the Hamiltonian dynamics of $\boldsymbol{\theta}$ in the original system with the position-dependent metric (mass matrix)*

$$M(\boldsymbol{\theta}) = \mathbb{I} + \tfrac{1}{\sigma(\boldsymbol{\theta})^2}(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\theta})\mathbf{D}\sigma(\boldsymbol{\theta}))^\mathsf{T}(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\theta})\mathbf{D}\sigma(\boldsymbol{\theta})), \quad \boldsymbol{\eta}(\boldsymbol{\theta}) = \tfrac{1}{\sigma(\boldsymbol{\theta})}(\boldsymbol{y} - F(\boldsymbol{\theta})).$$

For the special case where $\sigma(\boldsymbol{\theta}) = \sigma$ is a constant function (that is the observation noise standard deviation is assumed to be known and fixed), the metric function simplifies to $M(\boldsymbol{\theta}) = \mathbb{I} + \frac{1}{\sigma^2}\mathbf{D}F(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}F(\boldsymbol{\theta})$, corresponding to the expected Fisher information matrix $\frac{1}{\sigma^2}\mathbf{D}F(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}F(\boldsymbol{\theta})$ plus the negative Hessian of the log-density of the prior $\mathbb{I}$ for this model. This specific choice of metric is recommended by Girolami and Calderhead (2011), and used in their numerical experiments.

From this equivalence of the continuous-time dynamics, the observed performance differences between our proposed C-HMC approach and RM-HMC (with a specific choice of metric) can be seen as a direct consequence of the improved performance of the constrained leapfrog integrator used in C-HMC compared to the generalized leapfrog integrator used in RM-HMC. Although both involve implicit steps, as described in Section 4.2, for each constrained leapfrog integrator step we can use Newton's method to efficiently solve the non-linear system of equations during the position projection.

In contrast in each generalized leapfrog step, two separate systems of non-linear equations must be solved, with existing implementations generally using a simple direct fixed-point iteration (Girolami and Calderhead, 2011). Empirically we observe

that these simple fixed-point iterations have a much higher tendency to diverge or converge slowly compared to the Newton iteration used for the constrained integrator. We also stress that as the fixed-point solvers used in the generalized leapfrog integrator may fail to converge, or converge to a different solution in a time-reversed step, a similar reversibility check as used for the constrained leapfrog integrator (and originally proposed in a RWM setting by Zappa et al. (2018)), is required to ensure the correctness of the RM-HMC transitions. The RM-HMC implementation we used in our experiments includes such checks.

6. Numerical experiments

In this section, we evaluate the performance of our proposed methodology on artificial and real-world inference problems. We test variants of the C-HMC algorithm using both a symmetric Newton and full Newton solver for the projection step. In each example, to provide a representative benchmark for existing MCMC methods we compare to a standard HMC algorithm with either a diagonal or dense metric. We also compare to RM-HMC on the first (toy) example, however due to its poor convergence, even with long chains, we do not include results on the other examples.

For each method (standard HMC, C-HMC, and RM-HMC), we use the updated version of the *No-U-Turn Sampler* (NUTS) algorithm (Hoffman and Gelman, 2014; Betancourt, 2017) currently underlying the probabilistic programming framework *Stan* (Carpenter et al., 2017) to automate setting the number of integrator steps in each proposal's trajectory. During an initial warm-up stage of each chain, the integrator step size $\varepsilon > 0$ was tuned with a dual-averaging algorithm to achieve an average acceptance rate of 0.9 as described in Hoffman and Gelman (2014). For the standard HMC chains, the diagonal or dense metric was additionally tuned during this warm-up stage using an equivalent scheme to that employed by *Stan*. All experiments were performed using the MCMC sampler implementations in the Python package *Mici* (Graham, 2019), which defines a high-level interface that abstracts away the details of Algorithm 3 and requires only specification of the model functions and optionally, their derivatives.

We report the minimum estimated *effective sample size* (ESS) over the quantities of interest, normalized by the total chain run time, as our measure of sampling efficiency. We compute the ESS estimates using the *ArviZ* (Kumar et al., 2019) implementation of the *bulk* ESS statistic (Vehtari et al., 2019). We also report, where informative, the maximum potential scale reduction $\hat{R}$ convergence diagnostic across the quantities of interest, using the *ArviZ* implementation of the rank-normalized split-$\hat{R}$ statistic proposed by Vehtari et al. (2019), with values greater than one indicative of non-convergence. For each of the experiments we ran three sets of four chains (twelve in total) for each method tested, with 1000 adaptive warm-up iterations and 2500 main iterations per-chain. For the standard HMC chains, the metric adaptation was shared across each set of four chains, otherwise all chains were independent and with initial states independently sampled from the prior. Only the post-warm-up samples were used in computing the ESS estimates.

All plots were produced using the Python package *Matplotlib* (Hunter, 2007) and

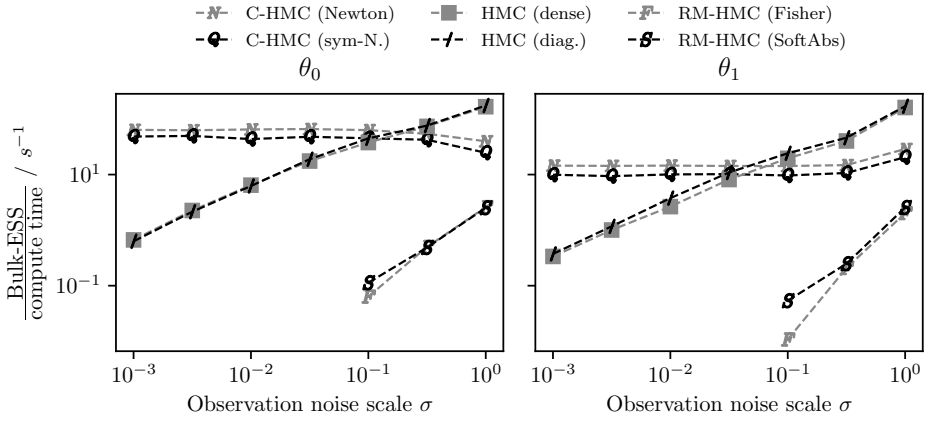

Fig. 4: Toy two-dimensional model. Sampling efficiency for varying σ.

the Python packages *NumPy* (Harris et al., 2020) and *SciPy* (Virtanen et al., 2020) were used for array and linear-algebra operations. In all cases unless otherwise mentioned, the forward models were implemented using the numerical computing framework *JAX* (Bradbury et al., 2018) to allow efficient calculation of the required derivatives using automatic differentiation.

6.1. *Toy example*

We first return to our running toy example considered in Section 2. Figure 4 shows the sampling efficiencies for standard HMC, RM-HMC and our proposed approach of C-HMC on the extended space. For RM-HMC the same Fisher information based and *SoftAbs* metrics are used as described previously. It can be immediately seen that both tested variants of HMC and RM-HMC degenerate as $\sigma \to 0$ while the efficiency of C-HMC with both projection solvers is unaffected by the vanishing noise asymptotic $\sigma \to 0$. Figure 3 in the Supplementary Material shows the $\hat{R}$ convergence statistic values and integrator step sizes ε for varying σ. For all but the largest $\sigma = 1$ value the RM-HMC chains show signs of non-convergence. To maintain a non-zero average acceptance rate, both standard HMC and RM-HMC chains require an integrator step size which decreases with σ; the C-HMC chains on the other hand maintain a stable acceptance rate with an approximately constant step size.

6.2. *State space model*

We apply our proposed methodology to a *state space model* (SSM) to illustrate that the approach can also give gains in models with high-dimensional latent spaces. A sequence of T latent states $x_{1:T}$ are generated from a linear-Gaussian dynamic

$$x_1 \sim \mathsf{Normal}(\mu, \tfrac{\gamma^2}{1-\rho^2}), \quad x_t = \mu + \rho(x_{t-1} - \mu) + \gamma\nu_t, \quad \nu_t \sim \mathsf{Normal}(0, 1) \quad \forall t \in 2{:}T$$

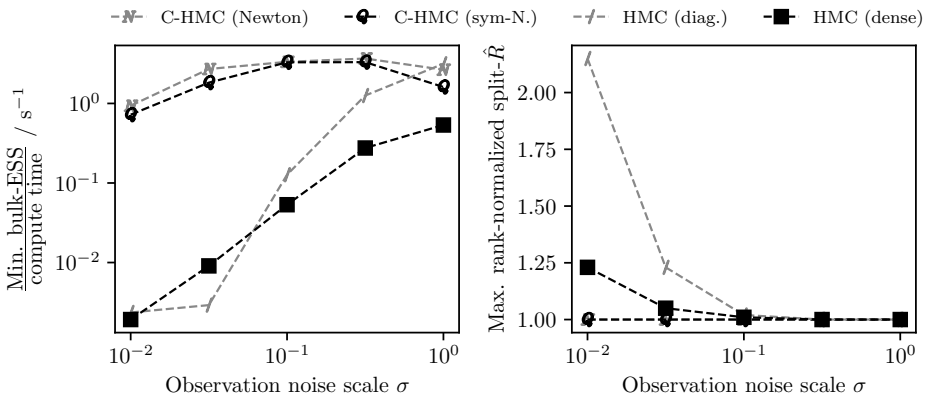

Fig. 5: Non-linear SSM model. Left: minimum $\frac{\text{ESS}}{\text{compute time}}$ for varying σ; right: maximum $\hat{R}$ for varying σ. Min./max. are taken across the model parameters $\mu, \rho, \gamma, \sigma$.

for $T = 1000$ and parameter values $\mu = -0.5$, $\gamma = 0.4$ and $\rho = 0.9$. From this latent sequence $x_{1:T}$, we generate T observations according to

$$y_t = \exp(x_t) + \sigma\eta_t, \quad \eta_t \sim \mathsf{Normal}(0, 1) \quad \forall t \in 1{:}T,$$

for each observation noise standard deviation $\sigma \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$.

Then, we jointly infer the parameters $\mu, \rho, \gamma, \sigma$ and latent state sequence $x_{1:T}$ given the observations $y_{1:T}$ for each of the true σ values.

We assign weakly informative priors on the parameters: $\mu \sim \mathsf{Normal}(0, 1)$, $\gamma \sim \mathsf{HalfNormal}(1)$, $\rho \sim \mathsf{Uniform}(-1, 1)$ and $\sigma \sim \mathsf{HalfNormal}(3)$. We use a non-centred parametrization of the model in terms of the a-priori independent noise sequence $\nu_{1:T}$ rather than the state sequence $x_{1:T}$, and express the parameters with bounded support (γ, ρ and σ) as smooth transforms of unbounded variables.

As the number of observations and latent variables both scale with T, a direct application of the constrained HMC algorithm using the lifting described in Section 3 would result in an $\mathcal{O}(T^3)$ complexity per constrained integrator step, due to the need to compute, evaluate the determinant of and solve linear systems in a $\mathcal{O}(T) \times \mathcal{O}(T)$ Gram matrix at each step. By exploiting the invertibility of the observation function here and Markovian structure of the model, the cost can be reduced to $\mathcal{O}(T)$. In brief, by using an alternative definition of the constraint function the resulting Gram matrix corresponds to the sum of a block-tridiagonal matrix and low rank product, allowing efficient $\mathcal{O}(T)$ operations. More details are given in Section 4 in the Supplementary Material.

The left panel in Figure 5 shows the measured computational efficiencies (in terms of estimated bulk-ESS per computation time) versus the true observation noise standard deviation σ for this SSM model. We see that the efficiency of our C-HMC approach (with both projection solvers) varies only minimally across all σ values, while, in contrast, the standard HMC methods both show a sustained degradation

in efficiency as σ becomes smaller. The $\hat{R}$ convergence statistic values show in the right panel in Figure 5 are also increasingly indicative of non-convergence for the HMC chains as σ becomes smaller. To give a sense of anistropy being induced in the posterior as σ decreases, Figure 4 in the Supplementary Material shows the Hessian of the negative log-density of the posterior (evaluated at a *maximum a posteriori* estimate) becoming increasingly ill-conditioned as $\sigma \to 0$.

Due to our ability to exploit additional model structure in this case, the constrained HMC method also remains competitive for even relatively large σ values despite the large latent and observation dimensions ($d_\Theta = 1004$ and $d_\mathcal{Y} = 1000$). The $\mathcal{O}(T)$ scaling with respect to the number of observation times of our approach is the same as standard HMC in this case.

6.3. FitzHugh–Nagumo ordinary differential equation model

Next, we consider parameter inference in non-linear dynamical system with noisy observations. The FitzHugh–Nagumo model (FitzHugh, 1961; Nagumo et al., 1962), a simplified description of the dynamics of action potential generation within an neuron, can be described by the following system of ODEs

$$\begin{aligned} \partial_1 x_0(t, \boldsymbol{\theta}) &= \alpha x_1(t, \boldsymbol{\theta}) - \beta x_0(t, \boldsymbol{\theta})^3 + \gamma x_1(t, \boldsymbol{\theta}), \\ \partial_1 x_1(t, \boldsymbol{\theta}) &= -\delta x_0(t, \boldsymbol{\theta}) - \varepsilon x_1(t, \boldsymbol{\theta}) + \zeta. \end{aligned} \tag{10}$$

We use a fourth-order Runge–Kutta method to discretize the system in time. The ground truth parameters used were $\alpha = 3$, $\beta = 1$, $\gamma = 3$, $\delta = 1/3$, $\varepsilon = 1/15$, and $\zeta = 1/15$, and initial state $\boldsymbol{x}(0, \cdot) = (1, -1)$. We simulated noisy observations $y_k = x_0(t_k, \boldsymbol{\theta}) + \sigma\eta_k$, with $\eta_k \sim \mathsf{Normal}(0, 1)$, of the first state coordinate at $d_\mathcal{Y} = 200$ equispaced times $0 = t_1 < \ldots < t_{d_\mathcal{Y}} = 20$.

The parametrization of the Fitzhugh–Nagumo system of ODEs in Equation (10) is non-identifiable; the dynamics remains unchanged under the transformation

$$(x_0(0, \cdot), x_1(0, \cdot), \alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \sigma) \mapsto (x_0(0, \cdot), s x_1(0, \cdot), \alpha, \beta, s^{-1}\gamma, s\delta, \varepsilon, s\zeta, \sigma)$$

for any scaling factor $s > 0$. This means that $x_2(0, \cdot)$, γ, δ and ζ are only identifiable up to a common scaling factor s. Therefore, the posterior distribution on these variables concentrates around a one-dimensional limiting manifold as $\sigma \to 0$.

We use prior distributions $\alpha \sim \mathsf{LogNormal}(0, 1)$, $\beta \sim \mathsf{LogNormal}(0, 1)$, $\gamma \sim \mathsf{LogNormal}(0, 1)$, $\delta \sim \mathsf{LogNormal}(-1, 1)$, $\varepsilon \sim \mathsf{LogNormal}(-2, 1)$, $\zeta \sim \mathsf{LogNormal}(-2, 1)$, $x_0(0) \sim \mathsf{Normal}(0, 1)$, and $x_1(0) \sim \mathsf{Normal}(0, 1)$, using a logarithmic parametrization for $(\beta, \gamma, \delta, \varepsilon, \zeta)$ to enforce non-negativity constraints.

Figure 6 shows the estimated ESS and $\hat{R}$ convergence statistic values for varying σ for this ODE model. The C-HMC chains using both the full Newton and symmetric Newton solver show a similar pattern of performance, with near constant ESS per compute time for smaller σ values and a slight drop-off in efficiency as σ becomes larger; as for the previous SSM model we speculate that this is due to a greater variation in curvature in the posterior for the more diffuse large σ settings, requiring a smaller integrator step size ε to maintain the stability of the projection solver and

in turn a drop in sampling efficiency. In support of this hypothesis, Figure 2 in the Supplementary Material shows that the adapted integrator step sizes show a small decreasing trend for large σ for the C-HMC chains. The Newton solver appears to give a small but consistent improvement in efficiency across all σ values compared to the symmetric Newton solver.

The results for the HMC chains with diagonal metric show the same decreasing sampling efficiency with σ as encountered previously, with the $\hat{R}$ values indicative of non-convergence for the smaller observation noise scales σ. The results for the HMC chains with dense metric on the other hand seem to indicate a quite different behaviour, with the sampling efficiency initially seeming to *increase* as the observation noise scale σ is *decreased* before beginning to decrease again for $\sigma < 0.1$, though at a slower rate than for the HMC chains with diagonal metric. The adapted integrator step size ε for the dense metric HMC chains (Figure 2 in the Supplementary Material) also show a similar non-monotonic relationship with σ. The split-$\hat{R}$ values also only indicate non-convergence for larger σ values for the HMC chains with dense metric, with values close to one for the smallest σ values.

We believe these results are due a pathological behaviour of the adaptation of the HMC chains with dense metric for small σ in this model, causing the chains to become effectively non-ergodic. It appears that for small σ the step size adaptation typically produces step sizes which give a vanishingly low probability of the HMC trajectories being able to enter the 'narrowest' regions of the posterior in the tails. Figure 6 in the Supplementary Material shows pairwise scatter plots for samples of the non-identifiable parameters $(x_1(0), \gamma, \delta, \zeta)$ for both HMC (with dense metric) and C-HMC (with Newton solver) chains for $\sigma = 0.01$ and $\sigma = 0.1$. It is immediately evident that the HMC chains have explored only a subset of the posterior mass covered by the C-HMC chain samples, with the HMC chains seeming to have particularly failed to explore the narrow regions of the posterior in the tails along the directions spanned by the variables in which the limiting manifold is non-linear.

The generated HMC trajectories rarely approach these narrow posterior regions and when they do, as the integrator step size is too large for the local curvature, they typically diverge leading to a rejection; however, as the proportion of the posterior mass in these regions is small (for $\sigma = 0.01$ just over two percent of the chain transitions of HMC with dense metric were rejected due to integrator divergence), these rejections do not significantly affect the average acceptance rate and so fail to cause the step size to be adapted to a more appropriate smaller value. Eventually for long enough runs the chains will enter the narrower regions and will typically then get 'stuck' and reject for many successive iterations due to the inappropriately large step size for the local curvature - trace plots for a HMC (with dense metric) chain showing such a sticking event are shown in Figure 7 in the Supplementary Material. However, for finite-length chains, such sticking events are rare and do not appear to have adversely affected the $\hat{R}$ convergence diagnostic.

This issue seems to particularly affect the HMC chains with dense metric in the following ways. In the local region of the posterior where the limiting manifold $\mathcal{S}$ is close to linear, the equivalent linear transformation effected by the adapted dense metric leads to a beneficial rescaling and decorrelation of the variables. Conversely,

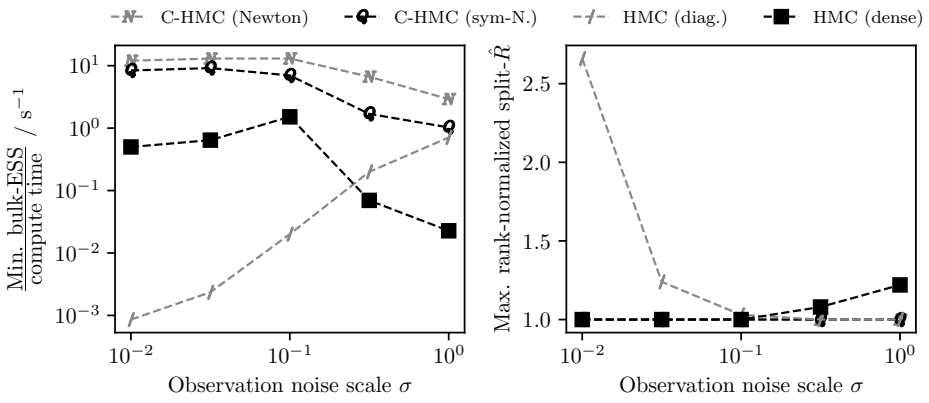

Fig. 6: FitzHugh–Nagumo ODE model. Left: minimum $\frac{\text{ESS}}{\text{compute time}}$ for varying σ; right: maximum $\hat{R}$ for varying σ. Min./max. are taken across the 9 model parameters.

this linear transformation has a detrimental effect in the tail regions of the posterior as an even smaller step size is needed for stability in these regions. This pathology and the difficulty in diagnosing it using standard heuristics provide a clear illustration of the benefits of using an MCMC method, such as our proposed C-HMC based approach, that is robust to a range of posterior scales.

6.4. Hodgkin–Huxley ordinary differential equation model

While the parameter non-identifiability in the previous example (Section 6.3) could be removed by reparametrization, we now consider a non-linear dynamical model with structural non-identifiabilities (Daly et al., 2015) that, to the best of the authors' knowledge, does not admit such a reparametrization.

Conductance-based or *Hodgkin–Huxley* models (Hodgkin and Huxley, 1952), are a family of non-linear dynamical system that describe the time-varying membrane voltage of neurons in terms of the ionic currents associated with voltage-gated ion channels. This class of mechanistic models is widely used in computational neuroscience, and inference in the vanishing noise regime is of particular relevance as modern intracellular recording technologies allow high-fidelity recording of cell membrane voltages with high signal-to-noise ratios (Weckstrom, 2010).

Here, we specifically consider a conductance-based model using three channel types, with the model behaviour determined by a 'minimal' set of 6 free parameters $(\bar{g}_{\text{Na}}, \bar{g}_{\text{K}}, \bar{g}_{\text{M}}, g_{\text{L}}, V_T, k_{\tau_p})$ which nonetheless allow good fits to experimental data from a broad range of cell types with differing characteristic behaviours (Pospischil et al., 2008). We consider the case of a cell subject to a square-wave stimulus current and with the cell membrane voltage noisily observed at $d_{\mathcal{Y}} = 1000$ equispaced time points. We infer the $d_\Theta = 7$ model parameters $\boldsymbol{\theta} = (\bar{g}_{\text{Na}}, \bar{g}_{\text{K}}, \bar{g}_{\text{M}}, g_{\text{L}}, V_T, k_{\tau_p}, \sigma)$ for observations generated for each of the true observation noise standard deviations $\sigma \in \{10^{-2}, 10^{-1.5}, 10^{-1}, 10^{0}\}$. A full description of the model used including the

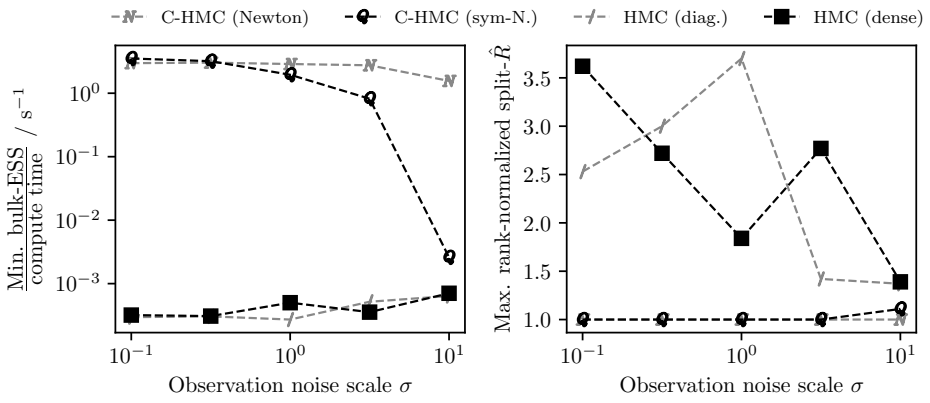

Fig. 7: Hodgkin–Huxley ODE model. Left: minimum $\frac{\text{ESS}}{\text{compute time}}$ for varying σ; right: maximum $\hat{R}$ for varying σ. Min./max. are taken across the 7 model parameters.

priors and time discretization is given in Section 5 in the Supplementary Material. Figure 7 shows the estimated sampling efficiencies and convergence diagnostics for varying true observation noise scales σ. HMC with both diagonal and dense metrics fails to converge for all σ values with $\hat{R}$ values much greater than one. For small σ values C-HMC with both projection solvers show similarly good efficiency and $\hat{R}$ values indicative of convergence, however for the larger σ values tested there is a noticeable drop-off in efficiency for the C-HMC chains with symmetric-Newton projection solver (and some indication of non-convergence for $\sigma = 1$). This appears to be due to the larger variation in curvatures in this more diffuse setting causing the symmetric-Newton solver to become unstable in some regions of the state space, potentially due to the constant Jacobian approximation becoming increasingly poor in such regions. The rejections due to non-convergence of the projection solver cause the step size to be adapted to a smaller value to maintain the acceptance rate, this in turn leads to a reduction in sampling efficiency. As for C-HMC with the Newton solver, there is only a relatively small drop-off in efficiency, suggesting that the Newton solver is able to remain stable for larger step sizes.

6.5. Poisson partial differential equation model

To illustrate the applicability of our model to inference in PDE constrained Bayesian inverse problems, we consider a two-dimensional domain $\Omega \subset \mathbb{R}^2$ and the problem of inferring a log-conductivity field $\phi : \Omega \to \mathbb{R}$ from a discrete set of noisy observations of a temperature field $u : \Omega \to \mathbb{R}$ and a known source term $f : \Omega \to \mathbb{R}$. These quantities satisfy the following Poisson PDE on the spatial domain Ω,

$$-\nabla \cdot \big(\exp(\phi)\nabla u\big) = f, \tag{11}$$

with vanishing Dirichlet boundary conditions $u_{|\partial\Omega} = 0$. Noisy observations of the temperature field $y_k = u(x_k) + \sigma\eta_k \quad \forall k \in \{1, \ldots, d_{\mathcal{Y}}\}$, are available at $d_{\mathcal{Y}} = 10$

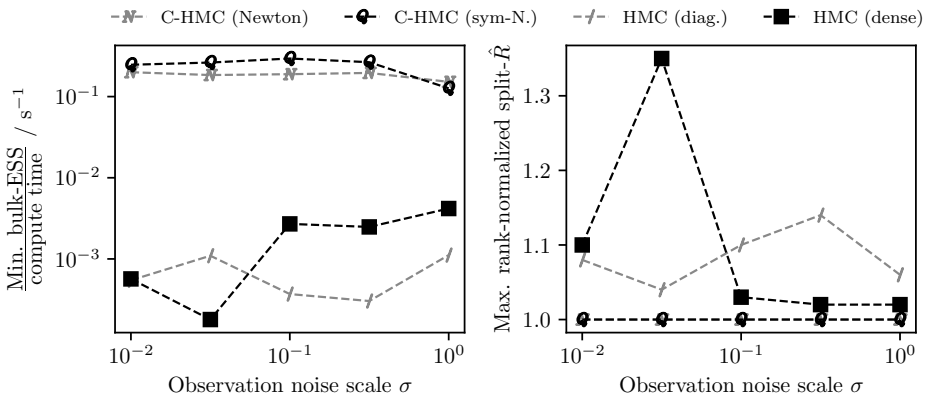

Fig. 8: Poisson PDE model. Left: minimum $\frac{\text{ESS}}{\text{compute time}}$ for varying σ; right: maximum $\hat{R}$ for varying σ. Min./max. are taken across the estimates of mean($\boldsymbol{z}$), std($\boldsymbol{z}$), and the parameter σ.

distinct locations $x_k \in \Omega$ (right panel of Figure 5 in the Supplementary Material). The library *FEniCS* (Alnæs et al., 2015) was used to numerically solved the PDE (11) with a first order *finite element method* (FEM) on a triangular mesh discretization of Ω with dimension $K = 469$ (left panel of Figure 5 in the Supplementary Material).

Specifically the discretized log-conductivity field can be expressed as $\phi(x) = \sum_{k=1}^{K} z_k e_k(x)$ for piecewise linear basis functions $e_k : \Omega \to \mathbb{R}$ and coefficients $\boldsymbol{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$.

We use a Gaussian random field with a Matérn covariance (Lindgren et al., 2011) as a prior model for ϕ. This can be realized as a FEM discretization of the Poisson equation $(\kappa^2 - \Delta)\phi = w$, with scaling parameter $\kappa = 0.1$, for a spatial Gaussian white noise $w : \Omega \to \mathbb{R}$ with unit variance. This induces a zero-mean Gaussian prior on the coefficients $\boldsymbol{z}$ with covariance matrix $\Gamma \in \mathbb{R}^{K \times K}$.

Our objective is then to approximate the posterior on the $d_\Theta = K + 1$ vector $\boldsymbol{\theta} = (z_1, \ldots, z_K, \sigma)$ of coefficients and observation noise scale, given the observations $\boldsymbol{y} = (y_1, \ldots, y_{d_y})$. Note that each evaluation of the posterior density involves solving the Poisson equation in (11). The automatic differentiation functionality of the *unified form language* (Alnæs et al., 2014) library in *FEniCS* was used to calculate the required derivatives of the forward operator implicitly defined by solving (11).

Figure 8 shows estimates of how the sampling efficiency and convergence to stationarity varies with σ for the different methods tested. As for the Hodgkin–Huxley ODE model in the previous section, the standard HMC chains with both dense and diagonal metrics, appear to have not converged for all σ values tested, with $\hat{R}$ values greater than one. While the non-convergence mean the ESS estimates for the HMC chains are unreliable, the adapted integrator step sizes ε, shown in Figure 2 in the Supplementary Material, suggest a similar requirement of ε scaling proportionally to σ for the HMC chains to maintain a stable acceptance rate, which would be expected to lead to a similar drop in efficiency with σ. The C-HMC chains

with both projection solvers do not show any signs of non-convergence across all σ values and the ESS estimates are close to constant aside from a small drop-off in efficiency when using the symmetric Newton solver for the largest σ values. For this model the symmetric Newton solver however in general appears to be slightly more performant than the Newton solver, with the large latent dimension d_Θ compared to observation dimension $d_\mathcal{Y}$ here meaning there is a greater gain in reducing the number of constraint Jacobian evaluations and associated linear algebra operations.

6.6. Ordinary differential equation models fitted to real data

So far, the data in our experiments have been simulated using the underlying generative models. This gave direct control over the observation noise scale σ (and so the signal-to-noise ratio in the model) which allowed us to show the robustness of the performance of our proposed methodology to differing signal-to-noise ratios.

In this section, we demonstrate the competitive performance of our proposed methodology in approximating the posteriors of models fitted to real data:

(a) A Lotka–Volterra predator-prey model (Lotka, 1925; Volterra, 1926) fitted to a dataset of hare and lynx populations measured by the Hudson Bay Company in the years 1900–1920 (Hewitt, 1921; Howard, 2009).
(b) A two-pool model for carbon flow within soil fitted to data from two soil incubation experiments taken from the *R* package *SoilR* (Sierra et al., 2014). The two datasets, *HN-T35* and *AK-T25* use soil samples from two different sites and incubated at different temperatures.
(c) A conductance-based model for action potential generation in the giant squid axon proposed by Hodgkin and Huxley (Hodgkin and Huxley, 1952) fitted to a digitization of the original voltage clamp experimental data from Daly et al. (2015). Posteriors are separately fitted to the sodium (Na) and potassium (K) ionic conductance data. Compared to the conductance-based model fitted in Section 6.4, here the membrane voltage is assumed to be clamped resulting in a set of decoupled linear ODEs for each of the channel gating variables.

Further details of all three models are given in Section 6 in the Supplementary Material. Approximate samples from each posterior were generated using: (i) our proposed approach of a constrained HMC algorithm targeting the lifted posterior (using a Newton projection solver), (ii) a standard HMC algorithm with diagonal metric, (iii) a standard HMC algorithm with dense metric.

Table 1 shows the parameter dimension d_Θ and observation dimension $d_\mathcal{Y}$ for each of the posteriors as well as an estimate of the effective sample size per compute time in seconds, with the values shown corresponding to the median (across the three independent chain sets) of the minimum per-parameter time normalized effective sample size (across the d_Θ model parameters). The constrained HMC algorithm outperforms standard HMC with a diagonal metric in all cases, and standard HMC with a dense metric in all but one case.

In general, while the cost of each constrained integrator step in the constrained HMC algorithm is higher compared to the leapfrog integrator step cost in the stan-

Posterior	d_Θ	$d_\mathcal{Y}$	Min. bulk-ESS per compute time / s^{-1} C-HMC	HMC (diag.)	HMC (dense)
Lotka–Volterra	8	42	35	15	**300**
Soil carbon (HN-T35)	7	25	**6.6**	2.1	3.4
Soil carbon (AK-T25)	7	25	**17**	4.4	2.8
Hodgkin-Huxley (K)	7	136	**52**	23	29
Hodgkin-Huxley (Na)	11	142	**3.9**	1.6	3.0

Table 1: Problem dimensions and measured minimum bulk-ESS per compute time for real-data examples. Bold values indicate best performing method for each posterior.

dard HMC algorithm, typically the adaptive tuning of the integrator step size results in much higher step sizes for a given target acceptance rate for the constrained HMC algorithm, resulting in fewer integrator steps per sample and so often an overall reduced computational cost per sample.

When the posterior is close to Gaussian, however, standard HMC with an appropriately tuned metric can also support larger integrator step sizes. In this case, due to its lower per-step computational cost standard HMC will typically outperform constrained HMC. This is seen in the Lotka-Volterra results where the posterior in the unconstrained space is well approximated by a Gaussian unlike the other posteriors (see pair plots in Section 7 in the Supplementary Material).

7. Conclusion

While a large literature has been devoted to designing MCMC samplers with good mixing properties when used in high-dimensional settings, comparatively little attention has been devoted to the problem of exploring distributions concentrating near low-dimensional manifolds or exhibiting multi-scale structures (Beskos et al., 2018; Livingstone and Zanella, 2019; Schillings et al., 2019). To the best of our knowledge, the sampling efficiency of all currently existing general-purpose MCMC samplers degenerates in the vanishing noise $\sigma \to 0$ asymptotic studied in this text. By reformulating the original problem into the task of exploring a distribution supported on a manifold embedded in a higher-dimensional space, our proposed methodology is able to maintain high sampling efficiency in the $\sigma \to 0$ regime. Our work opens up a number of avenues for future work in this area. While we concentrated in this work on additive observation noise, there are natural possible extensions to more general observation processes. Furthermore, an important open question is whether it is possible to define variants of the proposed approach that improve scalability in the regime where both latent and observation dimensions, d_Θ and $d_\mathcal{Y}$, are high, with the current scheme having a $\mathcal{O}(\min(d_\Theta d_\mathcal{Y}^2, d_\mathcal{Y} d_\Theta^2))$ computational complexity for general problems. The approach used to achieve a linear scaling in the number of observations for the SSM in Section 6.2, illustrates some initial progress along this line by exploiting structure in specific model classes.

Acknowledgement: The authors acknowledge support from a National University of Singapore (NUS) Young Investigator Award Grant (R-155-000-180-133) and a Singapore Ministry of Education Academic Research Funds Tier 1 (R-155-000-228-114). K. X. Au is supported by the NUS Integrative Sciences and Engineering Programme (ISEP) Scholarship.

References

Alnæs, M. S., Blechta, J., Hake, J., Johansson, A., Kehlet, B., Logg, A., Richardson, C., Ring, J., Rognes, M. E. and Wells, G. N. (2015) The FEniCS project version 1.5. *Archive of Numerical Software*, **3**.

Alnæs, M. S., Logg, A., Ølgaard, K. B., Rognes, M. E. and Wells, G. N. (2014) Unified form language: A domain-specific language for weak formulations of partial differential equations. *ACM Transactions on Mathematical Software (TOMS)*, **40**, 1–37.

Andersen, H. C. (1983) RATTLE: A 'velocity' version of the SHAKE algorithm for molecular dynamics calculations. *Journal of Computational Physics*, **52**, 24–34.

Ballnus, B., Hug, S., Hatz, K., Görlitz, L., Hasenauer, J. and Theis, F. J. (2017) Comprehensive benchmarking of markov chain monte carlo methods for dynamical systems. *BMC systems biology*, **11**, 1–18.

Barth, E., Kuczera, K., Leimkuhler, B. and Skeel, R. D. (1995) Algorithms for constrained molecular dynamics. *Journal of Computational Chemistry*, **16**, 1192–1209.

Besag, J. (1994) Comments on 'Representations of knowledge in complex systems' by Grenander and Miller. *Journal of the Royal Statistical Society, Series B*, **56**, 591–592.

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

Beskos, A., Roberts, G., Thiery, A. and Pillai, N. (2018) Asymptotic analysis of the random walk Metropolis algorithm on ridged densities. *The Annals of Applied Probability*, **28**, 2966–3001.

Betancourt, M. (2013) A general metric for riemannian manifold hamiltonian monte carlo. In *International Conference on Geometric Science of Information*, 327–334. Springer.

Betancourt, M. (2017) A conceptual introduction to Hamiltonian Monte Carlo. *arXiv e-prints*.

Blanes, S., Casas, F. and Sanz-Serna, J. (2014) Numerical integrators for the Hybrid Monte Carlo method. *SIAM Journal on Scientific Computing*, **36**, A1556–A1580.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D. and Wanderman-Milne, S. (2018) JAX: composable transformations of Python+NumPy programs. URL: http://github.com/google/jax.

Brubaker, M., Salzmann, M. and Urtasun, R. (2012) A family of MCMC methods on implicitly defined manifolds. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 161–172. Proceedings of Machine Learning Research.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**, 1.

Daly, A. C., Gavaghan, D. J., Holmes, C. and Cooper, J. (2015) Hodgkin–Huxley revisited: reparametrization and identifiability analysis of the classic action potential model with approximate Bayesian methods. *Royal Society open science*, **2**, 150499.

Dasgupta, A., Self, S. G. and Gupta, S. D. (2007) Non-identifiable parametric probability models and reparametrization. *Journal of Statistical Planning and Inference*, **137**, 3380–3393.

Diaconis, P., Holmes, S. and Shahshahani, M. (2013) *Sampling from a Manifold*, vol. Volume 10 of *Collections*, 102–125. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

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

FitzHugh, R. (1961) Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, **1**, 445–466.

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**, 123–214.

Graham, M. M. (2019) Mici: manifold MCMC methods in Python. URL: https://git.io/mici.py.

Graham, M. M. and Storkey, A. J. (2017) Asymptotically exact inference in differentiable generative models. *Electronic Journal of Statistics*, **11**, 5105–5164.

Gupta, S., Irbäck, A., Karsch, F. and Petersson, B. (1990) The acceptance probability in the hybrid Monte Carlo method. *Physics Letters, B*, **242**.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. and Oliphant, T. E. (2020) Array programming with NumPy. *Nature*, **585**, 357–362. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hartmann, C. and Schütte, C. (2005) A constrained hybrid Monte-Carlo algorithm and the problem of calculating the free energy in several variables. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics*, **85**, 700–710.

Hewitt, C. (1921) *The Conservation of thc Wild Life of Canada*. Charles Scribner's Sons.

Hines, K. E., Middendorf, T. R. and Aldrich, R. W. (2014) Determination of parameter identifiability in nonlinear biophysical models: A bayesian approach. *Journal of General Physiology*, **143**, 401–416.

Hodgkin, A. L. and Huxley, A. F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, **117**, 500.

Hoffman, M. D. and Gelman, A. (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Howard, P. (2009) Modeling basics, lecture notes for Math 442. Texas A&M University. URL: http://www.math.tamu.edu/~phoward/m442/modbasics.pdf.

Hunter, J. D. (2007) Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, **9**, 90–95.

Knapik, B. T., van der Vaart, A. W. and van Zanten, J. H. (2011) Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, **39**, 2626–2657.

Kumar, R., Carroll, C., Hartikainen, A. and Martin, O. A. (2019) ArviZ a unified library for exploratory analysis of Bayesian models in Python. *The Journal of Open Source Software*.

Leimkuhler, B. and Matthews, C. (2016) Efficient molecular dynamics using geodesic integration and solvent–solute splitting. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **472**, 20160138.

Leimkuhler, B. and Reich, S. (2004) *Simulating Hamiltonian dynamics*, vol. 14. Cambridge University Press.

Leimkuhler, B. J. and Skeel, R. D. (1994) Symplectic numerical integrators in constrained Hamiltonian systems. *Journal of Computational Physics*, **112**, 117–125.

Lelièvre, T., Rousset, M. and Stoltz, G. (2019) Hybrid Monte Carlo methods for sampling probability measures on submanifolds. *Numerische Mathematik*.

Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Livingstone, S. (2015) Geometric ergodicity of the random walk Metropolis with position-dependent proposal covariance. *arXiv preprints*.

Livingstone, S. and Zanella, G. (2019) The Barker proposal: combining robustness and efficiency in gradient-based MCMC. *arXiv preprint arXiv:1908.11812*.

Lotka, A. (1925) *Elements of physical biology*. Williams and Wilkins.

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**, 1087–1092.

Nagumo, J., Arimoto, S. and Yoshizawa, S. (1962) An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, **50**, 2061–2070.

Neal, R. M. (2011) *MCMC Using Hamiltonian Dynamics*, chap. Chapter 5. CRC Press. URL: https://www.routledgehandbooks.com/doi/10.1201/b10905-7.

Petra, N., Martin, J., Stadler, G. and Ghattas, O. (2014) A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM Journal on Scientific Computing*, **36**, A1525–A1555.

Pospischil, M., Toledo-Rodriguez, M., Monier, C., Piwkowska, Z., Bal, T., Frégnac, Y., Markram, H. and Destexhe, A. (2008) Minimal hodgkin–huxley type models for different classes of cortical and thalamic neurons. *Biological cybernetics*, **99**, 427–441.

Raue, A., Kreutz, C., Theis, F. J. and Timmer, J. (2013) Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**, 20110544. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2011.0544.

Reich, S. (1996) Symplectic integration of constrained Hamiltonian systems by composition methods. *SIAM Journal on Numerical Analysis*, **33**, 475–491.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, **7**, 110–120.

Roberts, G. O. and Rosenthal, J. S. (2001) Optimal scaling for various Metropolis–Hastings algorithms. *Statistical science*, **16**, 351–367.

Roberts, G. O. and Stramer, O. (2002) Langevin diffusions and Metropolis–Hastings algorithms. *Methodology and computing in applied probability*, **4**, 337–357.

Rothenberg, T. J. (1971) Identification in parametric models. *Econometrica: Journal of the Econometric Society*, 577–591.

Rousset, M., Stoltz, G. and Lelièvre, T. (2010) *Free Energy Computations: A Mathematical Perspective*. Imperial College Press.

Schillings, C., Sprungk, B. and Wacker, P. (2019) On the convergence of the Laplace approximation and noise level robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *arXiv preprints*.

Sierra, C., Müller, M. and Trumbore, S. E. (2014) Modeling radiocarbon dynamics in soils: SoilR version 1.1. *Geoscientific Model Development*, **7**, 1919–1931.

Stuart, A. M. (2010) Inverse problems: a Bayesian perspective. *Acta numerica*, **19**, 451–559.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.-C. (2019) Rank-normalization, folding, and localization: An improved r for assessing convergence of MCMC. *arXiv e-prints*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. and SciPy 1.0 Contributors (2020) SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**, 261–272.

Volterra, V. (1926) Fluctuations in the abundance of a species considered mathematically. *Nature*, **118**, 558–560.

Weckstrom, M. (2010) Intracellular recording. *Scholarpedia*, **5**, 2224. Revision #121415.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S. and Girolami, M. (2014) Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, **91**, 14–19.

Zappa, E., Holmes-Cerfon, M. and Goodman, J. (2018) Monte Carlo on manifolds: sampling densities and integrating functions. *Communications on Pure and Applied Mathematics*, **71**, 2609–2647.

A. Proof of Theorem 1

From the momentum condition in the constrained Hamiltonian dynamics and differentiating with respect to time we have that

$$\boldsymbol{p}_{\boldsymbol{\eta}} = -\tfrac{1}{\sigma} F \boldsymbol{p}_{\boldsymbol{\theta}} \implies \dot{\boldsymbol{p}}_{\boldsymbol{\eta}} = -\tfrac{1}{\sigma} F \dot{\boldsymbol{p}}_{\boldsymbol{\theta}}.$$

Substituting $\dot{\boldsymbol{p}}_{\boldsymbol{\theta}} = -\boldsymbol{\theta} + F^{\mathsf{T}}\boldsymbol{\lambda}$ and $\dot{\boldsymbol{p}}_{\boldsymbol{\eta}} = -\boldsymbol{\eta} + \sigma\boldsymbol{\lambda}$ and solving for $\boldsymbol{\lambda}$ gives

$$\boldsymbol{\lambda} = -(\sigma^2 \mathbb{I} + FF^{\mathsf{T}})^{-1}(F\boldsymbol{\theta} + \sigma\boldsymbol{\eta}) = -(\sigma^2 \mathbb{I} + FF^{\mathsf{T}})^{-1}(\boldsymbol{y} - \boldsymbol{f}),$$

with the last equality coming from the constraint equation $F\boldsymbol{\theta} + \boldsymbol{f} + \sigma\boldsymbol{\eta} - \boldsymbol{y} = \mathbf{0}$.

This gives that

$$\begin{aligned}\dot{\boldsymbol{p}}_{\boldsymbol{\theta}} &= -\boldsymbol{\theta} + F(\sigma^2\mathbb{I} + FF^{\mathsf{T}})^{-1}(\boldsymbol{y} - \boldsymbol{f}) \\ &= -\boldsymbol{\theta} + \tfrac{1}{\sigma^2}(\mathbb{I} + \tfrac{1}{\sigma^2}F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{f}) \\ &= -(\boldsymbol{\theta} - \boldsymbol{\mu})\end{aligned}$$

with the equality between the first and second lines arising from the push-through identity, and between the second and third lines from the definition of $\boldsymbol{\mu}$ above.

Differentiating $\dot{\boldsymbol{\theta}} = \boldsymbol{p}_{\boldsymbol{\theta}}$ with respect to time and substituting the above gives that

$$\ddot{\boldsymbol{\theta}} = -(\boldsymbol{\theta} - \boldsymbol{\mu})$$

which is equivalent to the second-order dynamics of $\boldsymbol{\theta}$ in the original Hamiltonian system for the choice of metric $M = \Sigma^{-1}$.

B. Proof of Theorem 2

Our proof uses the following standard result from classical mechanics

LEMMA 3. *If $\gamma : \Theta \to \mathcal{M}$ is a diffeomorphism, then the canonical transformation $\Gamma : \mathcal{T}^*\Theta \to \mathcal{T}^*\mathcal{M}$; $(\boldsymbol{\theta}, \boldsymbol{m}) \mapsto (\boldsymbol{q}, \boldsymbol{p})$ implicitly defined by*

$$\boldsymbol{q} = \gamma(\boldsymbol{\theta}), \quad \mathbf{D}\gamma(\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{p} = \boldsymbol{m}$$

preserves the form of Hamiltonian's equations that is

$$(\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}}) = (\nabla_2, -\nabla_1)\bar{H}(\boldsymbol{q}, \boldsymbol{p}) \implies (\dot{\boldsymbol{\theta}}, \dot{\boldsymbol{m}}) = (\nabla_2, -\nabla_1)H(\boldsymbol{\theta}, \boldsymbol{m})$$

for a Hamiltonian function $\bar{H} : \mathcal{T}^*\mathcal{M} \to \mathbb{R}$ and corresponding Hamiltonian function $H : \mathcal{T}^*\Theta \to \mathbb{R}$ defined by $H(\boldsymbol{\theta}, \boldsymbol{m}) = \bar{H} \circ \Gamma(\boldsymbol{\theta}, \boldsymbol{m})$.

If we define

$$\gamma(\boldsymbol{\theta}) = (\boldsymbol{\theta}, (\boldsymbol{y} - F(\boldsymbol{\theta}))/\sigma(\boldsymbol{\theta}))$$

then $\mathcal{M}$ is the image of Θ under γ, i.e. $\gamma(\Theta) = \mathcal{M}$. If we assume F and σ are smooth then the corestriction of γ onto $\mathcal{M}$ is a diffeomorphism from Θ to $\mathcal{M}$ and we have that $\mathcal{M}$ is diffeomorphic to Θ. Differentiating γ we have that

$$\mathbf{D}\gamma(\boldsymbol{\theta}) = \left(\mathbb{I}, -\tfrac{1}{\sigma(\boldsymbol{\theta})}\left(\mathbf{D}F(\boldsymbol{\theta}) + \tfrac{1}{\sigma(\boldsymbol{\theta})}(\boldsymbol{y} - F(\boldsymbol{\theta}))\mathbf{D}\sigma(\boldsymbol{\theta})\right)\right)$$

As $(\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}}) \in \mathcal{T}^*_{\mathcal{M}}(\boldsymbol{\theta}, \boldsymbol{\eta}) \implies \mathbf{D}c(\boldsymbol{\theta}, \boldsymbol{\eta})(\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}}) = 0$ and $\boldsymbol{\eta} = \frac{1}{\sigma(\boldsymbol{\theta})}(\boldsymbol{y} - F(\boldsymbol{\theta}))$ we have

$$\boldsymbol{p}_{\boldsymbol{\eta}} = -\tfrac{1}{\sigma(\boldsymbol{\theta})}(\mathbf{D}F(\boldsymbol{\theta}) + \tfrac{1}{\sigma(\boldsymbol{\theta})}(\boldsymbol{y} - F(\boldsymbol{\theta}))\mathbf{D}\sigma(\boldsymbol{\theta}))\boldsymbol{v}_\theta$$

and so $(\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}}) = \mathbf{D}\gamma(\boldsymbol{\theta})\boldsymbol{p}_{\boldsymbol{\theta}}$. For a canonical transformation $\Gamma : \mathcal{T}^*\Theta \to \mathcal{T}^*\mathcal{M}$; $(\boldsymbol{\theta}, \boldsymbol{m}) \mapsto (\boldsymbol{q}, \boldsymbol{v})$ implicitly defined by

$$\boldsymbol{q} = (\boldsymbol{\theta}, \boldsymbol{\eta}) = \gamma(\boldsymbol{\theta}), \quad \mathbf{D}\gamma(\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{v} = \mathbf{D}\gamma(\boldsymbol{\theta})^{\mathsf{T}}(\boldsymbol{p}_{\boldsymbol{\theta}}, \boldsymbol{p}_{\boldsymbol{\eta}}) = \boldsymbol{m}$$

we therefore have that

$$\boldsymbol{m} = \mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}(\boldsymbol{p_\theta}, \boldsymbol{p_\eta}) = \mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta})\boldsymbol{p_\theta} \implies \boldsymbol{p_\theta} = (\mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta}))^{-1}\boldsymbol{m}$$

and so the following explicit definition for Γ

$$((\boldsymbol{\theta}, \boldsymbol{\eta}), (\boldsymbol{p_\theta}, \boldsymbol{p_\eta})) = \Gamma(\boldsymbol{\theta}, \boldsymbol{m}) = (\gamma(\boldsymbol{\theta}), \mathbf{D}\gamma(\boldsymbol{\theta})(\mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta}))^{-1}\boldsymbol{m}).$$

By Lemma 3 the constrained Hamiltonian dynamics described by

$$(\dot{\boldsymbol{q}}, \dot{\boldsymbol{p}}) = (\nabla_2, -\nabla_1)\bar{H}(\boldsymbol{q}, \boldsymbol{p})$$

on $\mathcal{T}^*\mathcal{M}$ are equivalent to the Hamiltonian dynamics $(\dot{\boldsymbol{\theta}}, \dot{\boldsymbol{m}}) = (\nabla_2, -\nabla_1)H(\boldsymbol{\theta}, \boldsymbol{m})$ on $\mathcal{T}^*\Theta$ with Hamiltonian function H defined by

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{m}) &= \bar{H}(\Gamma(\boldsymbol{\theta}, \boldsymbol{m})) \\ &= \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 + \tfrac{1}{2\sigma(\boldsymbol{\theta})^2}(\boldsymbol{y} - F(\boldsymbol{\theta}))^\mathsf{T}(\boldsymbol{y} - F(\boldsymbol{\theta})) + \tfrac{1}{2}\log G(\gamma(\boldsymbol{\theta})) \\ &\quad + \tfrac{1}{2}\boldsymbol{m}^\mathsf{T}(\mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta}))^{-1}\mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta})(\mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta}))^{-1}\boldsymbol{m} \end{aligned}$$

By the matrix determinant lemma we have that

$$\begin{aligned} |G(\boldsymbol{\theta}, \boldsymbol{\eta})| &= \left|(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}))(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}))^\mathsf{T} + \sigma(\boldsymbol{\theta})^2\mathbb{I}\right| \\ &= \left|\tfrac{1}{\sigma(\boldsymbol{\theta})^2}(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}))^\mathsf{T}(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta})) + \mathbb{I}\right| |\sigma(\boldsymbol{\theta})^2\mathbb{I}| \end{aligned}$$

therefore we have that

$$\begin{aligned} H(\boldsymbol{\theta}, \boldsymbol{m}) = \tfrac{1}{2\sigma(\boldsymbol{\theta})^2}(\boldsymbol{y} - F(\boldsymbol{\theta}))^\mathsf{T}(\boldsymbol{y} - F(\boldsymbol{\theta})) + d_\mathcal{Y}\log\sigma(\boldsymbol{\theta}) + \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 \\ + \tfrac{1}{2}\boldsymbol{m}M(\boldsymbol{\theta})^{-1}\boldsymbol{m} + \tfrac{1}{2}\log|M(\boldsymbol{\theta})| \end{aligned}$$

with M defined for $\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{1}{\sigma(\boldsymbol{\theta})}(\boldsymbol{y} - F(\boldsymbol{\theta}))$ by

$$\begin{aligned} M(\boldsymbol{\theta}) &= \mathbf{D}\gamma(\boldsymbol{\theta})^\mathsf{T}\mathbf{D}\gamma(\boldsymbol{\theta}) \\ &= \mathbb{I} + \tfrac{1}{\sigma(\boldsymbol{\theta})^2}(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\theta})\mathbf{D}\sigma(\boldsymbol{\theta}))^\mathsf{T}(\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}(\boldsymbol{\theta})\mathbf{D}\sigma(\boldsymbol{\theta})). \end{aligned}$$

Web-based supporting materials for Manifold lifting: scaling Markov chain Monte Carlo to the vanishing noise regime

Khai Xiang Au
National University of Singapore, Singapore
Matthew M. Graham
University College London, UK
Alexandre H. Thiery
National University of Singapore, Singapore
E-mail: khai@u.nus.edu, m.graham@ucl.ac.uk & a.h.thiery@nus.edu.sg

1. Position-dependent MCMC methods in the vanishing noise regime

In Section 2 in the main text, we compared the performance of various existing MCMC methods in terms of how the average acceptance rate varies as a function of the observation noise scale σ and integrator step size ε in our toy two-dimensional model. There we compared the performance of both 'simple' schemes such as *random-walk Metropolis* (RWM), *Metropolis-adjusted Langevin algorithm* (MALA) and *Hamiltonian Monte Carlo* (HMC) with a fixed (isotropic) proposal covariance or metric, and the more complex *Riemannian-manifold Hamiltonian Monte Carlo* (RM-HMC), which accounts for the locally varying geometry of the posterior by simulating Hamiltonian dynamics on a Riemannian manifold with position dependent metric, with this requiring using an integrator with implicit steps involving solving a non-linear system of equations. For all these methods we found that the step size ε needed to be proportional to σ for the acceptance rate to not vanish as $\sigma \to 0$.

In the case of RM-HMC we suggested the requirement of reducing ε with σ to maintain a non-zero acceptance rate was due to the need to maintain the stability of the fixed-point solver used in the implicit integrator. As there are MCMC schemes which use position-dependent preconditioners to account for locally varying geometry in the target distribution but do not require solving a non-linear system of equations on each step as in RM-HMC, an obvious question is whether such simpler schemes suffer from the same degradation in performance as $\sigma \to 0$. If not, it would seem that such methods would provide a less implementationally complex approach to get comparable behaviour to the *constrained Hamiltonian Monte Carlo* (C-HMC) approach described in the paper.

As an initial concrete example, Girolami and Calderhead (2011) suggests a simple position-dependent MALA variant, which ignores terms involving derivatives of the metric function M (accounting for manifolds with varying curvature), with Gaussian

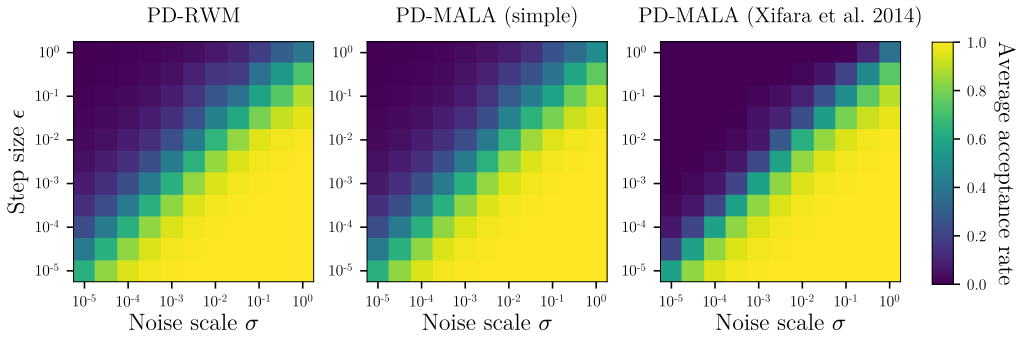

Fig. 1: Average acceptance rate for chains simulated using different position-dependent MCMC methods (using a Fisher information based metric) targetting the posterior π^σ of toy example for varying noise scales σ and (integrator) step sizes ε. The values displayed are means of the acceptance probabilities over four chains of 1000 samples initialised at equispaced points around the limiting manifold $\mathcal{S}$.

proposals generated according to

$$\boldsymbol{\theta}' \,|\, \boldsymbol{\theta} \sim \mathsf{Normal}\big(\boldsymbol{\theta} - \tfrac{\varepsilon}{2} M(\boldsymbol{\theta})^{-1} \nabla \Phi(\boldsymbol{\theta}), \varepsilon^2 M^{-1}(\boldsymbol{\theta})\big), \tag{1}$$

and accepted or rejected in a *Rosenbluth–Teller-Metropolis–Hastings* acceptance step (Metropolis et al., 1953; Hastings, 1970).

In Xifara et al. (2014), the authors note that the diffusion from which (1) is derived (using an Euler–Maruyama discretization) does not have the target distribution $\pi(\mathrm{d}\boldsymbol{\theta}) \propto \exp(-\Phi(\boldsymbol{\theta}))\,\mathrm{d}\boldsymbol{\theta}$ as its stationary distribution. They show that by adding a correction to the drift term, a diffusion with the correct invariant distribution can be derived, at the cost of requiring evaluation of derivatives of the metric function M. Again using an Euler–Maruyama discretization, the resulting scheme generates Gaussian proposals from

$$\boldsymbol{\theta}' \,|\, \boldsymbol{\theta} \sim \mathsf{Normal}\Big(\boldsymbol{\theta} - \tfrac{\varepsilon}{2} M(\boldsymbol{\theta})^{-1} \big(\nabla \Phi(\boldsymbol{\theta}) + \Xi(M)(\boldsymbol{\theta})\big), \varepsilon^2 M^{-1}(\boldsymbol{\theta})\Big),$$

with $\Xi_i(M)(\boldsymbol{\theta}) = \sum_{j,k=1}^{d_\Theta} \partial_k M_{ij}(\boldsymbol{\theta}) M^{-1}_{jk}(\boldsymbol{\theta}),\ i \in 1{:}d_\Theta$.

An alternative approach analysed by Livingstone (2015) is to omit the drift term altogether, and instead use M as a position-dependent proposal covariance in a RWM scheme, that is generate Gaussian proposals from

$$\boldsymbol{\theta}' \,|\, \boldsymbol{\theta} \sim \mathsf{Normal}\big(\boldsymbol{\theta}, \varepsilon^2 M^{-1}(\boldsymbol{\theta})\big).$$

In all three cases, as for the metric in RM-HMC, the pre-conditioner or proposal covariance function M, can be chosen as any positive-definite matrix valued function. One possible choice is the expected Fisher information plus the negative Hessian of the log prior density, which for a model with fixed observation noise scale σ and standard normal prior corresponds to

$$M(\boldsymbol{\theta}) = \mathbb{I} + \tfrac{1}{\sigma^2} \mathbf{D}F(\boldsymbol{\theta})^{\mathsf{T}} \mathbf{D}F(\boldsymbol{\theta}). \tag{2}$$

For large σ the identity term will dominate and the position-dependent RWM and MALA schemes will behave increasingly similar to RWM and MALA schemes with a fixed identity metric (proposal covariance), while for small σ the Fisher information term will dominate, resulting in proposals which adjust for the differing scales in the tangential and normal directions to the limiting manifold $\{\boldsymbol{\theta} : F(\boldsymbol{\theta}) = \boldsymbol{y}\}$.

In Figure 1 we show comparable heatmaps of the average acceptance rate as a function of the observation noise scale σ and step size ε to those in Figure 2 in the main paper, for these three MCMC scheme with position-dependent proposals when applied to our toy two-dimensional model, using the metric function defined in (2). We see a similar pattern of performance as observed for the RWM and MALA schemes with fixed identity metric and RM-HMC with this same choice of metric, with all three schemes requiring for the step size ε to be proportional to σ to maintain a non-zero acceptance rate (though the constants of proportionality are larger than for the schemes with fixed metric).

We therefore see that these simpler position-dependent schemes do not provide a remedy to the issue of poor scaling in the small noise regime. We also note that, while these position-dependent schemes are implementationally simpler than RM-HMC and C-HMC due to the lack of any requirement to solve non-linear systems of equations, for the choice of metric in (2), they retain the same overall per-step $\min(d_{\mathcal{Y}} d_{\Theta}^2, d_{\mathcal{Y}}^2 d_{\Theta})$ complexity due to the requirement to evaluate the determinant of (and solve linear systems in) the metric function M.

2. Computational cost of constrained leapfrog integrator

Typically the operations dominating the computational cost of each constrained leapfrog integrator step (Algorithm 2) will be: evaluation of the constraint Jacobian $\mathbf{D}C$, solving linear systems of the form $\mathbf{D}C(\boldsymbol{q})\mathbf{D}C(\boldsymbol{q}')^{\mathsf{T}}\boldsymbol{x} = \boldsymbol{b}$ and $G(\boldsymbol{q})\boldsymbol{x} = \boldsymbol{b}$ and evaluating the Gram matrix log-determinant $\log\det G(\boldsymbol{q})$ and its derivative.

The exact number of operations performed will vary on each integrator step depending on the number of iterations taken for the projection solver to reach convergence. We found in practice however both the Newton and symmetric Newton solvers, when they did converge, tended to do so within a small number ($\sim 5 - 10$) of iterations and that this was reflected across all of the different models used in the experiments. This suggests that the average number of iterations for convergence is not strongly affected by either the dimensions of the unknowns d_{Θ} or observations $d_{\mathcal{Y}}$ and so can be assumed to be roughly constant with respect to these variables.

For a model of the form in (3), the constraint function $C(\boldsymbol{\theta}, \boldsymbol{\eta}) = F(\boldsymbol{\theta}) + \sigma(\boldsymbol{\theta})\boldsymbol{\eta} - \boldsymbol{y}$ has a Jacobian of the form

$$\mathbf{D}C(\boldsymbol{\theta}, \boldsymbol{\eta}) = (\mathbf{D}F(\boldsymbol{\theta}) + \boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}), \sigma(\boldsymbol{\theta})\mathbb{I}_{d_{\mathcal{Y}}}).$$

By a standard result from algorithmic differentiation (Griewank, 1993), the Jacobian of a function $\mathbb{R}^m \to \mathbb{R}^n$ can always be evaluated at a cost proportional to $\min(m, n)$ times the cost of evaluating the function itself. Evaluation of $\mathbf{D}F$ will therefore cost $\mathcal{O}(\min(d_{\Theta}, d_{\mathcal{Y}}))$ times the cost of evaluating the forward function F while evaluation of $\mathbf{D}\sigma$ will be comparable to evaluating (the scalar valued) σ itself.

The projection steps require solving linear systems with matrix terms of the form

$$\mathbf{D}C(\boldsymbol{\theta},\boldsymbol{\eta})\mathbf{D}C(\boldsymbol{\theta}',\boldsymbol{\eta}')^{\mathsf{T}} = (\mathbf{D}F(\boldsymbol{\theta})+\boldsymbol{\eta}\mathbf{D}\sigma(\boldsymbol{\theta}))(\mathbf{D}F(\boldsymbol{\theta}')+\boldsymbol{\eta}'\mathbf{D}\sigma(\boldsymbol{\theta}'))^{\mathsf{T}}+\sigma(\boldsymbol{\theta})\sigma(\boldsymbol{\theta}')\mathbb{I}_{d_{\mathcal{Y}}}.$$

corresponding to the Gram matrix when $(\boldsymbol{\theta},\boldsymbol{\eta}) = (\boldsymbol{\theta}',\boldsymbol{\eta}')$. For *underdetermined models* where $d_\Theta > d_\mathcal{Y}$ direct computation of the matrix product will have a $\mathcal{O}(d_\Theta d_\mathcal{Y}^2)$ cost and solving a linear system in the resulting $d_\mathcal{Y} \times d_\mathcal{Y}$ matrix will have a $\mathcal{O}(d_\mathcal{Y}^3)$. For *overdetermined models* with $d_\Theta < d_\mathcal{Y}$, by recognising that the above corresponds to a rank d_Θ correction to a diagonal matrix, we can exploit the Woodbury matrix identity to solve the linear systems at $\mathcal{O}(d_\Theta^2 d_\mathcal{Y})$ cost.

By a similar argument, evaluating the determinant of the Gram matrix $\det G(\boldsymbol{q})$, can in the undetermined setting be performed directly at $\mathcal{O}(d_\mathcal{Y}^3)$ cost plus a $\mathcal{O}(d_\Theta d_\mathcal{Y}^2)$ to compute the Gram matrix itself, and in the overdetermined setting, by exploiting the matrix determinant lemma at a $\mathcal{O}(d_\Theta^2 d_\mathcal{Y})$ cost. As the log-determinant is a scalar valued function computing the derivative of $\log\det G(\boldsymbol{q})$ with respect to $\boldsymbol{q}$ by reverse-model algorithmic differentiation will cost no more than a constant multiple of the cost of evaluating $\log\det G(\boldsymbol{q})$ itself.

Overall we therefore have that the linear algebra operations involved in each constrained integrator step have a $\mathcal{O}(\min(d_\Theta d_\mathcal{Y}^2, d_\mathcal{Y} d_\Theta^2))$ complexity, with each constraint Jacobian evaluation additionally having a cost of $\mathcal{O}(\min(d_\Theta, d_\mathcal{Y}))$ times the (model dependent) cost of evaluating the forward function F (assuming evaluation of the noise scale σ is negligible in comparison to the cost of evaluating F).

3. Additional figures for simulated data experiments

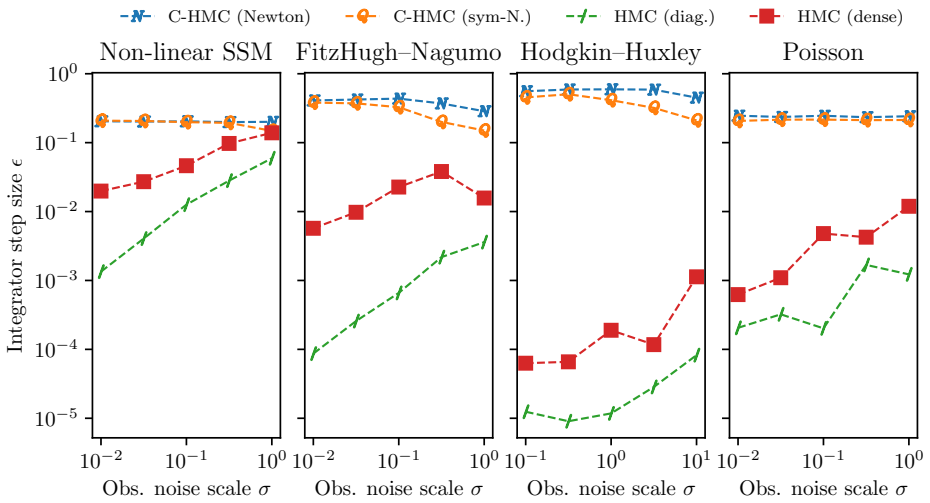

Fig. 2: Adapted integrator step size ε for varying σ for non-linear SSM, FitzHugh–Nagumo and Hodgkin–Huxley ODE and Poisson PDE simulated data models.

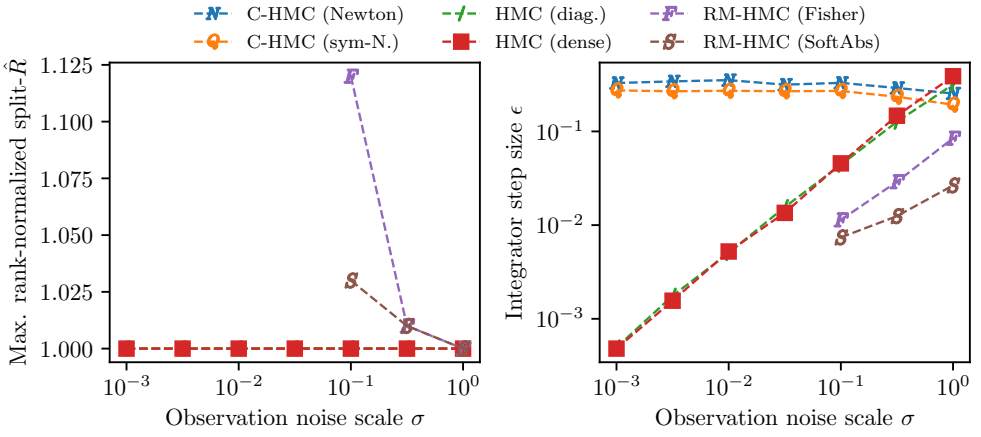

Fig. 3: Toy two-dimensional model. Maximum (across θ_0 and θ_1) rank-normalized split-$\hat{R}$ convergence diagnostic and integrator step size ε for varying σ.

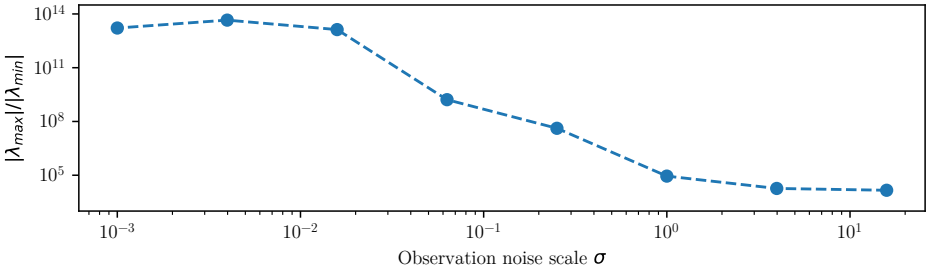

Fig. 4: Nonlinear SSM model. Condition number of the Hessian of the negative log-density evaluated at a MAP estimate for varying σ. $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalues (by magnitude) respectively.

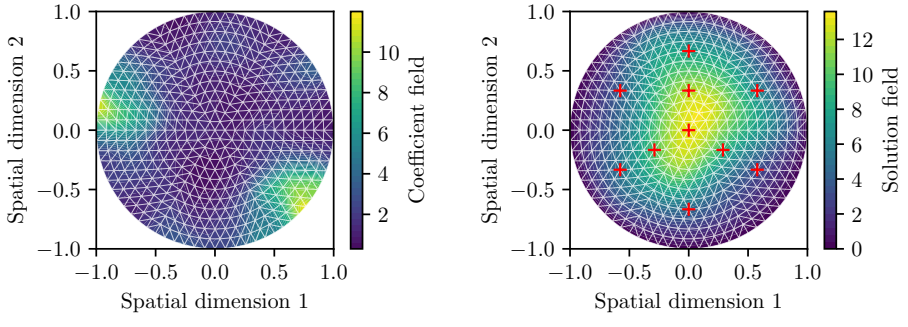

Fig. 5: Poisson PDE model. Left: True coefficient field; right: solution field with observed locations shown by red crosses and mesh by white triangulation.

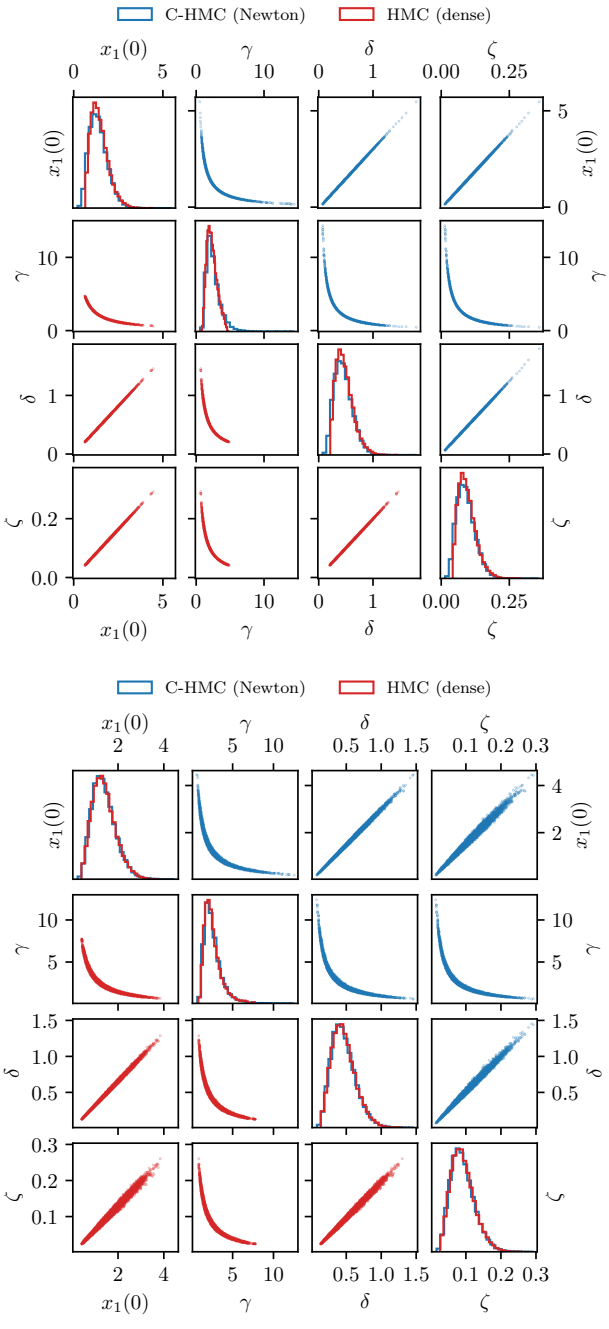

Fig. 6: FitzHugh–Nagumo ODE model. Pairwise scatter plots and histograms of non-identifiable parameters $x_1(0), \gamma, \delta, \zeta$ for true observation scale $\sigma = 0.01$ (top) and $\sigma = 0.1$ (bottom) for HMC (with dense metric, red) and C-HMC (with Newton projection solver, blue) chains. The HMC chains have failed to explore the narrower regions of the posterior for the pairs of variables with a non-linear relationship.

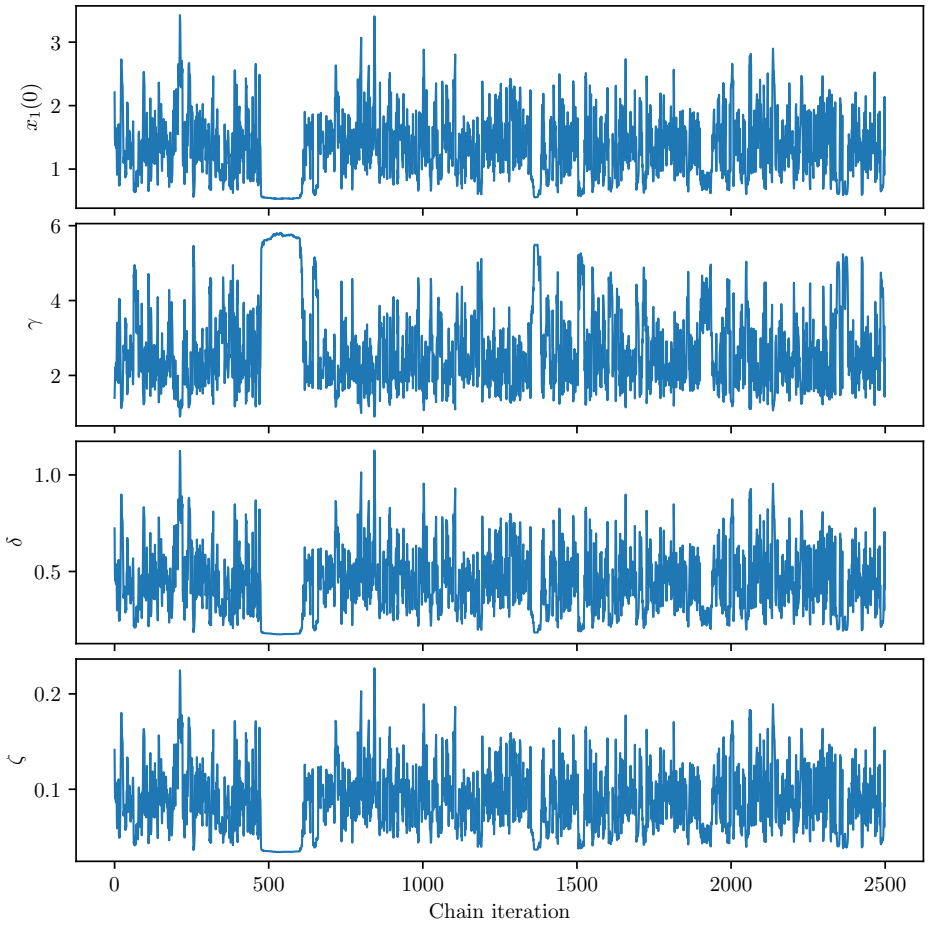

Fig. 7: FitzHugh–Nagumo ODE model. Example trace plots of non-identifiable parameters $x_1(0), \gamma, \delta, \zeta$ for true observation scale $\sigma = 10^{-1.5}$ for HMC chain (with dense metric) illustrating 'sticking' behaviour.

4. Exploiting Markovian structure in state space models

A SSM of the form

$$\begin{aligned}
\boldsymbol{x}_1 &= G_1(\boldsymbol{\phi}, \boldsymbol{\nu}_1), \\
\boldsymbol{x}_t &= G_t(\boldsymbol{\phi}, \boldsymbol{\nu}_t)(\boldsymbol{x}_{t-1}) \quad t \in 2{:}T, \\
\boldsymbol{y}_t &= H_t(\boldsymbol{\phi})(\boldsymbol{x}_t) + \sigma_t(\boldsymbol{\phi})\boldsymbol{\eta}_t \quad t \in 1{:}T,
\end{aligned}$$

can be considered as a special case of the general model class described by (3), with $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\nu})$, $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_T)$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)$ and $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$. If $\boldsymbol{\phi} \in \mathbb{R}^{d_\Phi}$, $\boldsymbol{\nu}_t \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$ for all $t \in 1{:}T$ then $d_\Theta = d_\Phi + T d_x$ and $d_\mathcal{Y} = T d_y$. Using

the lifting described in Section 3 in the main paper, with constraint function

$$\begin{aligned} C(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) &= (C_1, \ldots, C_T)(\boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\eta}), \\ C_t(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) &= H_t(\boldsymbol{\phi}) \circ \big(\bigcirc_{s=1}^{t} G_s(\boldsymbol{\phi}, \boldsymbol{\nu}_t) \big) + \sigma_t(\boldsymbol{\phi}) \boldsymbol{\eta}_t - \boldsymbol{y}_t \quad t \in 1{:}T, \end{aligned}$$

evaluating the $Td_y \times (d_\Phi + Td_x + Td_y)$ Jacobian of the constraint function will have an $\mathcal{O}(T^2)$ cost (as evaluating each C_t and its Jacobian $\mathbf{D}C_t$ via reverse-mode automatic differentiation will have an $\mathcal{O}(t)$ cost) and evaluating the determinant of the $Td_y \times Td_y$ Gram matrix or solving a linear system in it will have an $\mathcal{O}(T^3)$ cost. This would suggest that our proposed approach will be of limited applicability in SSMs unless the number of observation times T is small.

In some cases however we can exploit the additional structure in SSMs to significantly reduce the computational costs. Specifically if the observation functions $(H_t(\boldsymbol{\phi}))_{t=1}^T$ are invertible for all values of the parameters $\boldsymbol{\phi}$ then can reduce the cost of evaluating the constraint Jacobian and Gram matrix to $\mathcal{O}(T)$. In this case, we have that the system states $\boldsymbol{x}_{1:T}$ at each time step can be recovered from the observations $\boldsymbol{y}_{1:T}$, parameters $\boldsymbol{\phi}$ and observation noise vectors $\boldsymbol{\eta}_{1:T}$ using

$$\boldsymbol{x}_t = \bar{\boldsymbol{x}}_t(\boldsymbol{\phi}, \boldsymbol{\eta}_t) = (H_t(\boldsymbol{\phi}))^{-1}(\boldsymbol{y}_t - \sigma_t(\boldsymbol{\phi})\boldsymbol{\eta}_t) \quad t \in 1{:}T,$$

with we explicitly differentiating between the values of the states $\boldsymbol{x}_t$ and functions of the parameters and noise vectors mapping to those values $\bar{\boldsymbol{x}}_t$ here for clarity.

We can then define an alternative constraint function

$$\begin{aligned} \bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) &= (\bar{C}_1, \ldots, \bar{C}_T)(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}), \\ \bar{C}_1(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) &= G_1(\boldsymbol{\phi}, \boldsymbol{\nu}_1) - \bar{\boldsymbol{x}}_1(\boldsymbol{\phi}, \boldsymbol{\eta}_1), \\ \bar{C}_t(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) &= G_t(\boldsymbol{\phi}, \boldsymbol{\nu}_t)(\bar{\boldsymbol{x}}_{t-1}(\boldsymbol{\phi}, \boldsymbol{\eta}_{t-1})) - \bar{\boldsymbol{x}}_t(\boldsymbol{\phi}, \boldsymbol{\eta}_t) \quad t \in 2{:}T, \end{aligned}$$

with the property that the zero-level sets of $\bar{C}$ and C coincide; that is the constraint equations $C(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) = \mathbf{0}$ and $\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) = \mathbf{0}$ describe the same manifold $\mathcal{M}$.

Importantly however we can evaluate the Jacobian of the constraint function $\bar{C}$ and perform computations with the associated Gram matrix at much lower cost than for the corresponding computations for the original constraint function C. Firstly, as each $\bar{C}_t$ can be evaluated at cost that is independent of t (compared to the $\mathcal{O}(t)$ evaluation cost for separately evaluating each C_t), $\mathbf{D}\bar{C}_t(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})$ can be evaluated at $\mathcal{O}(1)$ cost (with respect to t) using reverse-mode automatic differentiation and the overall constraint Jacobian at $\mathcal{O}(T)$ cost.

Secondly, we have that $\bar{C}_t$ depends only on $\boldsymbol{\phi}$, $\boldsymbol{\nu}_t$, $\boldsymbol{\eta}_t$ and $\boldsymbol{\eta}_{t-1}$ for all $t \in 2{:}T$ and $\bar{C}_1$ depends only on $\boldsymbol{\phi}$, $\boldsymbol{\nu}_1$ and $\boldsymbol{\eta}_1$. Therefore, $\mathbf{D}_2\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})$ is block diagonal with T blocks of size $d_y \times d_x$ (and zeros elsewhere) and $\mathbf{D}_3\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})$ is block tridiagonal, with T blocks of size $d_y \times d_y$ along the main diagonal and a further $T-1$ blocks of size $d_y \times d_y$ in the first lower diagonal (and zeros elsewhere). The corresponding Gram matrix

$$\bar{G}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta}) = \mathbf{D}\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})\mathbf{D}\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})^\mathsf{T} = \sum_{i \in 1:3} \mathbf{D}_i\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})\mathbf{D}_i\bar{C}(\boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{\eta})^\mathsf{T}$$

can therefore be decomposed as the sum of a rank-d_Φ matrix

$$\mathbf{D}_1\bar{C}(\boldsymbol{\phi},\boldsymbol{\nu},\boldsymbol{\eta})\mathbf{D}_1\bar{C}(\boldsymbol{\phi},\boldsymbol{\nu},\boldsymbol{\eta})^\mathsf{T},$$

and the block tri-diagonal matrix

$$\mathbf{D}_2\bar{C}(\boldsymbol{\phi},\boldsymbol{\nu},\boldsymbol{\eta})\mathbf{D}_2\bar{C}(\boldsymbol{\phi},\boldsymbol{\nu},\boldsymbol{\eta})^\mathsf{T} + \mathbf{D}_3\bar{C}(\boldsymbol{\phi},\boldsymbol{\nu},\boldsymbol{\eta})\mathbf{D}_3\bar{C}(\boldsymbol{\phi},\boldsymbol{\nu},\boldsymbol{\eta})^\mathsf{T},$$

with T blocks of size $d_y \times d_y$ along the main diagonal, and $T-1$ blocks of size $d_y \times d_y$ on the lower and upper diagonals. Linear systems in the block tridiagonal matrix and its determinant can both be evaluated at $\mathcal{O}(T)$ cost, therefore by further using the matrix determinant lemma and Woodbury identity to account for the rank-d_Φ term, we can then compute the determinant of and solve linear systems in the Gram matrix at an $\mathcal{O}(T)$ cost.

5. Hodgkin–Huxley current stimulus model details

For the numerical experiments in Section 6.4 we use a *Hodgkin–Huxley* (Hodgkin and Huxley, 1952) conductance-based model of neuronal dynamics subject to injected current stimulus, in particular following the particular model formulation described in Pospischil et al. (2008). The dynamics are described by the following system of ODEs

$$\begin{aligned}
\partial_1 V(t,\boldsymbol{\theta}) &= \tfrac{1}{C_m}\big(I_\mathrm{S}(t) - \textstyle\sum_{\mathrm{ch}\in\{\mathrm{Na,K,M,L}\}} I_\mathrm{ch}(V(t,\boldsymbol{\theta}),(m,n,h,p)(t,\boldsymbol{\theta}),\boldsymbol{\theta})\big),\\
\partial_1 m(t,\boldsymbol{\theta}) &= \alpha_m(V(t,\boldsymbol{\theta}),\boldsymbol{\theta})(1-m(t,\boldsymbol{\theta})) - \beta_m(V(t,\boldsymbol{\theta}),\boldsymbol{\theta})m(t,\boldsymbol{\theta}),\\
\partial_1 n(t,\boldsymbol{\theta}) &= \alpha_n(V(t,\boldsymbol{\theta}),\boldsymbol{\theta})(1-n(t,\boldsymbol{\theta})) - \beta_n(V(t,\boldsymbol{\theta}),\boldsymbol{\theta})n(t,\boldsymbol{\theta}),\\
\partial_1 h(t,\boldsymbol{\theta}) &= \alpha_h(V(t,\boldsymbol{\theta}),\boldsymbol{\theta})(1-h(t,V,\boldsymbol{\theta})) - \beta_h(V(t,\boldsymbol{\theta}),\boldsymbol{\theta})h(t,\boldsymbol{\theta}),\\
\partial_1 p(t,\boldsymbol{\theta}) &= \big(p_\infty(V(t,\boldsymbol{\theta}),\boldsymbol{\theta}) - p(t,\boldsymbol{\theta})\big)/\tau_p(V(t,\boldsymbol{\theta}),\boldsymbol{\theta}),
\end{aligned}$$

where V is the real-valued *membrane voltage* and m, n, h and p are ion channel gating variables, which takes values in $[0,1]$.

The ODE for the membrane voltage V includes terms for the *stimulus* current I_S, here defined as a square wave

$$I_\mathrm{S}(t) = \begin{cases} 3.248, & \text{if } 10 \leq t \leq 110 \\ 0, & \text{otherwise} \end{cases}$$

and various ionic channel current components, namely the sodium ionic current I_Na, the potassium ionic current I_K, the 'delayed-rectifier' potassium ionic current I_M (Traub and Miles, 1991) and a catch-all leakage current I_L, with these defined by

$$\begin{aligned}
I_\mathrm{Na}(V,(m,n,h,p),\boldsymbol{\theta}) &= \bar{g}_\mathrm{Na}m^3h(V-53), & I_\mathrm{K}(V,(m,n,h,p),\boldsymbol{\theta}) &= \bar{g}_\mathrm{K}n^4(V+107),\\
I_\mathrm{M}(V,(m,n,h,p),\boldsymbol{\theta}) &= \bar{g}_\mathrm{M}p(V+107), & I_\mathrm{L}(V,(m,n,h,p),\boldsymbol{\theta}) &= g_\mathrm{L}(V+75).
\end{aligned}$$

The dynamics of the gating variables $\{m, n, h\}$ are governed by rate functions

$$\alpha_n(V, \boldsymbol{\theta}) = \frac{0.032\,(V - V_T - 15)}{\exp(-(V - V_T - 15)/5) - 1}, \qquad \beta_n(V, \boldsymbol{\theta}) = 0.5\exp(-(V - V_T - 10)/40),$$
$$\alpha_m(V, \boldsymbol{\theta}) = \frac{0.32(V - V_T - 13)}{\exp(-(V - V_T - 13)/4) - 1}, \qquad \beta_m(V, \boldsymbol{\theta}) = \frac{0.28(V - V_T - 40)}{\exp(-(V - V_T - 40)/5) - 1},$$
$$\alpha_h(V, \boldsymbol{\theta}) = 0.128\exp(-(V - V_T - 17)/18), \quad \beta_h(V, \boldsymbol{\theta}) = \frac{4}{\exp(-(V - V_T - 40)/5) + 1},$$

which can be used to define *steady state* and *time constant* functions

$$x_\infty(V, \boldsymbol{\theta}) = \frac{\alpha_x(V, \boldsymbol{\theta})}{\alpha_x(V, \boldsymbol{\theta}) + \beta_x(V, \boldsymbol{\theta})}, \quad \tau_x(V, \boldsymbol{\theta}) = \frac{1}{\alpha_x(V, \boldsymbol{\theta}) + \beta_x(V, \boldsymbol{\theta})}, \quad x \in \{m, n, h\},$$

corresponding respectively to the steady state and time constant of the solutions to governing ODEs for fixed V, for which the ODEs are linear. In the case of p the steady-state and time-constant functions are defined directly as

$$p_\infty(V, \boldsymbol{\theta}) = \big(1 + \exp\big(-(V + 35)/10\big)\big)^{-1},$$
$$\tau_p(V, \boldsymbol{\theta}) = k_{\tau_p}\big(3.3\exp((V + 35)/20) + \exp(-(V + 35)/20)\big)^{-1}.$$

The initial state of the ODEs system is then defined as

$$V(0, \boldsymbol{\theta}) = -75, \quad x(0, \boldsymbol{\theta}) = x_\infty(-75, \boldsymbol{\theta}) \quad \forall x \in \{m, n, h, p\}.$$

The ODE system is numerically solved using a Lie–Trotter splitting based approach that exploits the property that the system is *conditionally linear* (Chen et al., 2020): the ODEs for the gating variables $\{m, n, h, p\}$ for a fixed membrane voltage V are linear as is the ODE for the membrane voltage V for fixed values of the gating variables $\{m, n, h, p\}$. A integrator timestep of $\Delta = 0.1$ is used. The $d_{\mathcal{Y}} = 1000$ observations are modelled as being generated according to

$$y_i = V(10 + i\Delta, \boldsymbol{\theta}) + \sigma\eta_i, \quad \eta_i \sim \mathsf{Normal}(0, 1), \quad \forall i \in 1{:}1000.$$

that is, observations of the membrane voltage only, at discrete times and subject to independent additive Gaussian noise.

The $d_\Theta = 7$ parameters, $\boldsymbol{\theta} = (\bar{g}_{\mathrm{Na}}, \bar{g}_{\mathrm{K}}, \bar{g}_{\mathrm{M}}, g_{\mathrm{L}}, V_T, k_{\tau_p}, \sigma)$, are given priors

$$\begin{aligned} \bar{g}_{\mathrm{Na}} &\sim \mathsf{LogNormal}(4, 1), & V_T &\sim \mathsf{Normal}(-60, 10), \\ \bar{g}_{\mathrm{K}} &\sim \mathsf{LogNormal}(2, 1), & k_{\tau_p} &\sim \mathsf{LogNormal}(8, 1), \\ \bar{g}_{\mathrm{M}} &\sim \mathsf{LogNormal}(-3, 1), & \sigma &\sim \mathsf{LogNormal}(0, 1). \\ g_{\mathrm{L}} &\sim \mathsf{LogNormal}(-3, 1), \end{aligned}$$

6. Details for ordinary differential equation models fitted to real data

6.1. Lotka–Volterra model

A classic model of the dynamics of an ecosystem in which two species interact, one as predator and one as prey, due to Lotka (1925) and Volterra (1926). The system

dynamics are described by the ODEs

$$\partial_1 x_1(t, \boldsymbol{\theta}) = (\alpha - \beta x_2(t, \boldsymbol{\theta}))x_1(t, \boldsymbol{\theta}), \qquad \partial_1 x_2(t, \boldsymbol{\theta}) = (\delta - \gamma x_1(t, \boldsymbol{\theta}))x_2(t, \boldsymbol{\theta}),$$

with initial conditions

$$x_1(t_0, \boldsymbol{\theta}) = x_{1,0}, \qquad x_2(t_0, \boldsymbol{\theta}) = x_{2,0},$$

where x_1 describes the prey population and x_2 the predator population. The model was fitted to a dataset of annual measurements from 1900 to 1920 of snowshoe hare and Canadian Lynx populations, based on pelt numbers collected by the Hudson's Bay Company (Hewitt, 1921; Howard, 2009). Following the approach of Carpenter (2018), a log-normal observation model is assumed with the $d_{\mathcal{Y}} = 42$ observations $y_{1:42}$ of the prey x_1 and predator x_2 populations for the years $t_j = 1900+j$, $\forall j \in 0{:}20$ assumed to be generated according to

$$\log y_{2j+i} = \log x_i(t_j, \boldsymbol{\theta}) + \sigma_i \eta_{2j+i}, \quad \eta_{2j+i} \sim \mathsf{Normal}(0, 1) \quad \forall i \in \{1, 2\}, j \in 0{:}20.$$

The ODE system is numerically solved using a fourth-order Runge–Kutta method with fixed time step $\Delta = 0.1$ days. Also following Carpenter (2018), the $d_\Theta = 8$ parameters, $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \delta, x_{1,0}, x_{2,0}, \sigma_1, \sigma_2)$, are given priors

$$\begin{aligned}
&\alpha \sim \mathsf{TruncatedNormal}(1, 0.5, 0, \infty), && \delta \sim \mathsf{TruncatedNormal}(0.05, 0.05, 0, \infty),\\
&\beta \sim \mathsf{TruncatedNormal}(0.05, 0.05, 0, \infty), && x_{i,0} \sim \mathsf{LogNormal}(\log(10), 1) \quad \forall i \in \{1, 2\},\\
&\gamma \sim \mathsf{TruncatedNormal}(1, 0.5, 0, \infty), && \text{and } \sigma_i \sim \mathsf{LogNormal}(-1, 1) \quad \forall i \in \{1, 2\}.
\end{aligned}$$

6.2. Soil carbon two-pool model

A two-pool model for carbon flow within soil, based on a *Stan* case study (Carpenter, 2014). The system is described by the linear ODEs

$$\partial_1 x_1(t, \boldsymbol{\theta}) = -k_1 x_1(t, \boldsymbol{\theta}) + \alpha_{12} k_2 x_2(t, \boldsymbol{\theta}), \quad \partial_1 x_2(t, \boldsymbol{\theta}) - k_2 x_2(t, \boldsymbol{\theta}) + \alpha_{21} k_1 x_1(t, \boldsymbol{\theta}),$$

with initial conditions

$$x_1(t_0, \boldsymbol{\theta}) = \gamma\, C_0, \qquad x_2(t_0, \boldsymbol{\theta}) = (1 - \gamma)\, C_0.$$

The observed data correspond to measurements of the carbon cumulatively evolved by the system over time, with an assumption of independent normal noise in the observations of unknown standard deviation, with the observation model for a set of measurements at $d_{\mathcal{Y}}$ times $t_{1:d_{\mathcal{Y}}}$ then defined by

$$y_i = (C_0 - x_1(t_i, \boldsymbol{\theta}) - x_2(t_i, \boldsymbol{\theta})) + \sigma \eta_i \quad \forall i \in 1{:}d_{\mathcal{Y}}.$$

Data from two soil incubation experiments are taken from the *R* package *SoilR* (Sierra et al., 2014). The two datasets, *HN-T35* and *AK-T25* use soil samples from two different sites and incubated at different temperatures, and both have

measurements at $d_{\mathcal{Y}} = 25$ times. The ODE system can be exactly solved in this case. The $d_\Theta = 7$ parameters, $\boldsymbol{\theta} = (k_1, k_2, \alpha_{12}, \alpha_{21}, \gamma, C_0, \sigma)$ are given priors

$$\begin{aligned} k_1 &\sim \mathsf{HalfNormal}(1), & \gamma &\sim \mathsf{Uniform}(0,1), \\ k_2 &\sim \mathsf{TruncatedNormal}(0,1,0,k_1), & C_0 &\sim \mathsf{LogNormal}(1,2), \\ \alpha_{12} &\sim \mathsf{Uniform}(0,1), & \text{and } \sigma &\sim \mathsf{HalfNormal}(1). \\ \alpha_{21} &\sim \mathsf{Uniform}(0, 1-\alpha_{12}), \end{aligned}$$

6.3. Hodgkin–Huxley voltage clamp model

A conductance-based model of action potential generation in neurons, as proposed in Hodgkin and Huxley (1952). Compared to the model described in Section 5, which corresponded to the free dynamics of a neuron subject to a current stimulus, in accordance with the experimental protocol of Hodgkin and Huxley (1952) the model here is used to simulate the dynamics of a giant squid axon under a *voltage clamp*, with the membrane voltage held at a specified level using a feedback circuit. In addition to clamping the voltage, in the experiments of Hodgkin and Huxley (1952), the dynamics of specific ionic channels were isolated by manipulating the ions present in the extracellular fluid during experiments, with this allowing individual measurements to be made of the potassium and sodium channel conductances.

The conductance of a particular set of ion channels is modelled as being determined by one or more *gating variables* with temporally varying activations in $[0,1]$, with the overall conductance then a polynomial expression in the gating variable values and an overall maximum conductance coefficient. Each of the gating variables is assumed to have dynamics governed by the linear (for fixed parameters $\boldsymbol{\theta}$ and membrane voltage V) ODE

$$\partial_1 x(t, V, \boldsymbol{\theta}) = \alpha_x(V, \boldsymbol{\theta})\big(1 - x(t, V, \boldsymbol{\theta})\big) - \beta_x(V, \boldsymbol{\theta}) x(t, V, \boldsymbol{\theta})$$

where α_x and β_x are channel-specific parametrized and voltage-dependent rate functions. For a fixed membrane voltage V this ODE has the exact solution

$$x(t, V, \boldsymbol{\theta}) = x(0, V, \boldsymbol{\theta}) + \big(x_\infty(V, \boldsymbol{\theta}) - x(0, V, \boldsymbol{\theta})\big)\big(1 - \exp(-t/\tau_x(V, \boldsymbol{\theta}))\big)$$

where x_∞ and τ_x, respectively the *steady state* and *time constant*, are defined

$$x_\infty(V, \boldsymbol{\theta}) = \frac{\alpha_x(V, \boldsymbol{\theta})}{\alpha_x(V, \boldsymbol{\theta}) + \beta_x(V, \boldsymbol{\theta})}, \quad \tau_x(V, \boldsymbol{\theta}) = \frac{1}{\alpha_x(V, \boldsymbol{\theta}) + \beta_x(V, \boldsymbol{\theta})}.$$

We assume in all cases the initial condition for a gating variable is the steady state value for a membrane voltage $V = 0$, corresponding to the resting potential, i.e.

$$x(0, V, \boldsymbol{\theta}) = x_\infty(0, \boldsymbol{\theta}).$$

We fit the models to a digitization of the original voltage clamp experimental data from Daly et al. (2015), specifically sequences of sodium and potassium channel conductances for varying depolarizations (fixed values of the membrane voltage).

6.3.1. Potassium channel conductances

For the potassium channel conductances, the model uses a single gating variable n with corresponding parametrized rate functions

$$\alpha_n(V, \boldsymbol{\theta}) = \frac{k_{\alpha_n 1}\,(V + k_{\alpha_n 2})}{\exp\left((V + k_{\alpha_n 2})/k_{\alpha_n 3}\right) - 1}, \qquad \beta_n(V, \boldsymbol{\theta}) = k_{\beta_n 1}\,\exp(V/k_{\beta_n 2}).$$

The potassium conductances are then assumed to be defined as $\bar{g}_{\mathrm{K}} n^4$, with the observed noisy conductance data therefore assumed to be generated according to

$$y_i = \bar{g}_{\mathrm{K}} n(t_i, V_i, \boldsymbol{\theta})^4 + \sigma\eta_i, \quad \eta_i \sim \mathsf{Normal}(0, 1) \quad \forall i \in 1{:}d_{\mathcal{Y}}$$

where $(t_i, V_i)_{i=1}^{d_{\mathcal{Y}}}$ are the set of $d_{\mathcal{Y}} = 136$ measurement times and applied membrane voltages (depolarizations) used in the experiments.

The $d_\Theta = 7$ parameters $\boldsymbol{\theta} = (k_{\alpha_n 1}, k_{\alpha_n 2}, k_{\alpha_n 3}, k_{\beta_n 1}, k_{\beta_n 2}, \bar{g}_{\mathrm{K}}, \sigma)$ are given priors

$$\begin{aligned}
k_{\alpha_n 1} &\sim \mathsf{LogNormal}(-3, 1), & k_{\beta_n 2} &\sim \mathsf{LogNormal}(2, 1),\\
k_{\alpha_n 2} &\sim \mathsf{LogNormal}(2, 1), & \bar{g}_{\mathrm{K}} &\sim \mathsf{LogNormal}(2, 1),\\
k_{\alpha_n 3} &\sim \mathsf{LogNormal}(2, 1), & \text{and } \sigma &\sim \mathsf{LogNormal}(0, 1).\\
k_{\beta_n 1} &\sim \mathsf{LogNormal}(-3, 1),
\end{aligned}$$

6.3.2. Sodium channel conductances

For the sodium channel conductances, the model uses two gating variables m and h with corresponding parametrized rate functions

$$\alpha_m(V, \boldsymbol{\theta}) = \frac{k_{\alpha_m 1}\,(V + k_{\alpha_m 2})}{\exp((V + k_{\alpha_m 2})/k_{\alpha_m 3}) - 1}, \quad \beta_m(V, \boldsymbol{\theta}) = k_{\beta_m 1}\,\exp(V/k_{\beta_m 2}),$$

$$\alpha_h(V, \boldsymbol{\theta}) = k_{\alpha_h 1}\,\exp(V/k_{\alpha_h 2}), \quad \beta_h(V, \boldsymbol{\theta}) = \frac{1}{\exp((V + k_{\beta_h 1})/k_{\beta_h 2}) - 1}.$$

The sodium conductances are then assumed to be defined as $\bar{g}_{\mathrm{Na}} m^3 h$, with the observed noisy conductance data therefore assumed to be generated according to

$$y_i = \bar{g}_{\mathrm{Na}} m(t_i, V_i, \boldsymbol{\theta})^3 h(t_i, V_i, \boldsymbol{\theta}) + \sigma\eta_i, \quad \eta_i \sim \mathsf{Normal}(0, 1) \quad \forall i \in 1{:}d_{\mathcal{Y}}$$

where $(t_i, V_i)_{i=1}^{d_{\mathcal{Y}}}$ are the set of $d_{\mathcal{Y}} = 142$ measurement times and applied membrane voltages (depolarizations) used in the experiments.

The $d_\Theta = 11$ parameters

$$\boldsymbol{\theta} = (k_{\alpha_m 1}, k_{\alpha_m 2}, k_{\alpha_m 3}, k_{\beta_m 1}, k_{\beta_m 2}, k_{\alpha_h 1}, k_{\alpha_h 2}, k_{\beta_h 1}, k_{\beta_h 2}, \bar{g}_{\mathrm{Na}}, \sigma)$$

are given priors

$$\begin{aligned}
k_{\alpha_m 1} &\sim \mathsf{LogNormal}(-3, 1), & k_{\beta_m 2} &\sim \mathsf{LogNormal}(2, 1), & k_{\beta_h 2} &\sim \mathsf{LogNormal}(1, 1),\\
k_{\alpha_m 2} &\sim \mathsf{LogNormal}(2, 1), & k_{\alpha_h 1} &\sim \mathsf{LogNormal}(-3, 1), & \bar{g}_{\mathrm{Na}} &\sim \mathsf{LogNormal}(2, 1),\\
k_{\alpha_m 3} &\sim \mathsf{LogNormal}(2, 1), & k_{\alpha_h 2} &\sim \mathsf{LogNormal}(2, 1), & \text{and } \sigma &\sim \mathsf{LogNormal}(0, 1).\\
k_{\beta_m 1} &\sim \mathsf{LogNormal}(0, 1), & k_{\beta_h 1} &\sim \mathsf{LogNormal}(2, 1),
\end{aligned}$$

7. Posterior pair plots for models fitted to real data

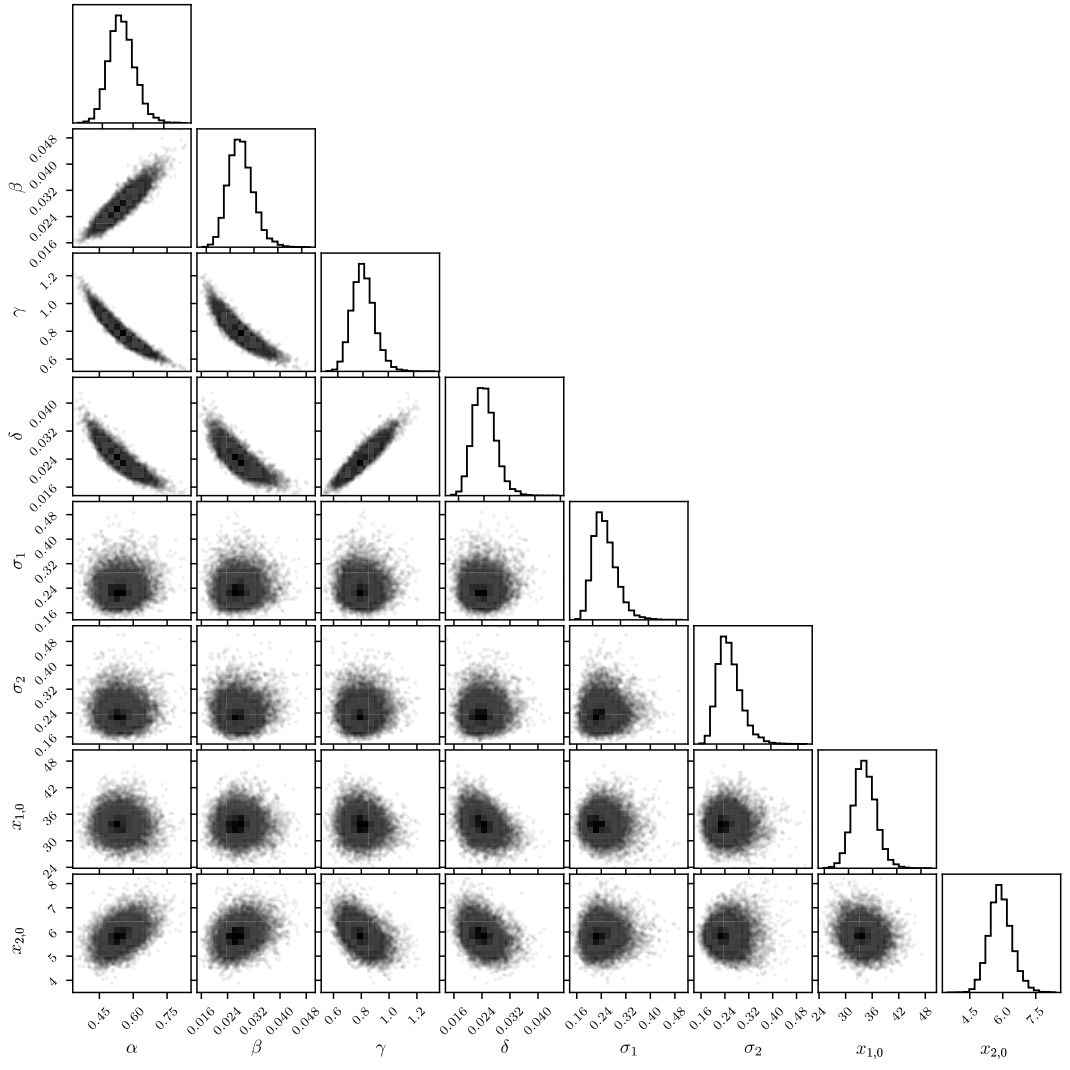

Fig. 8: *Lotka–Volterra*: Parameter space

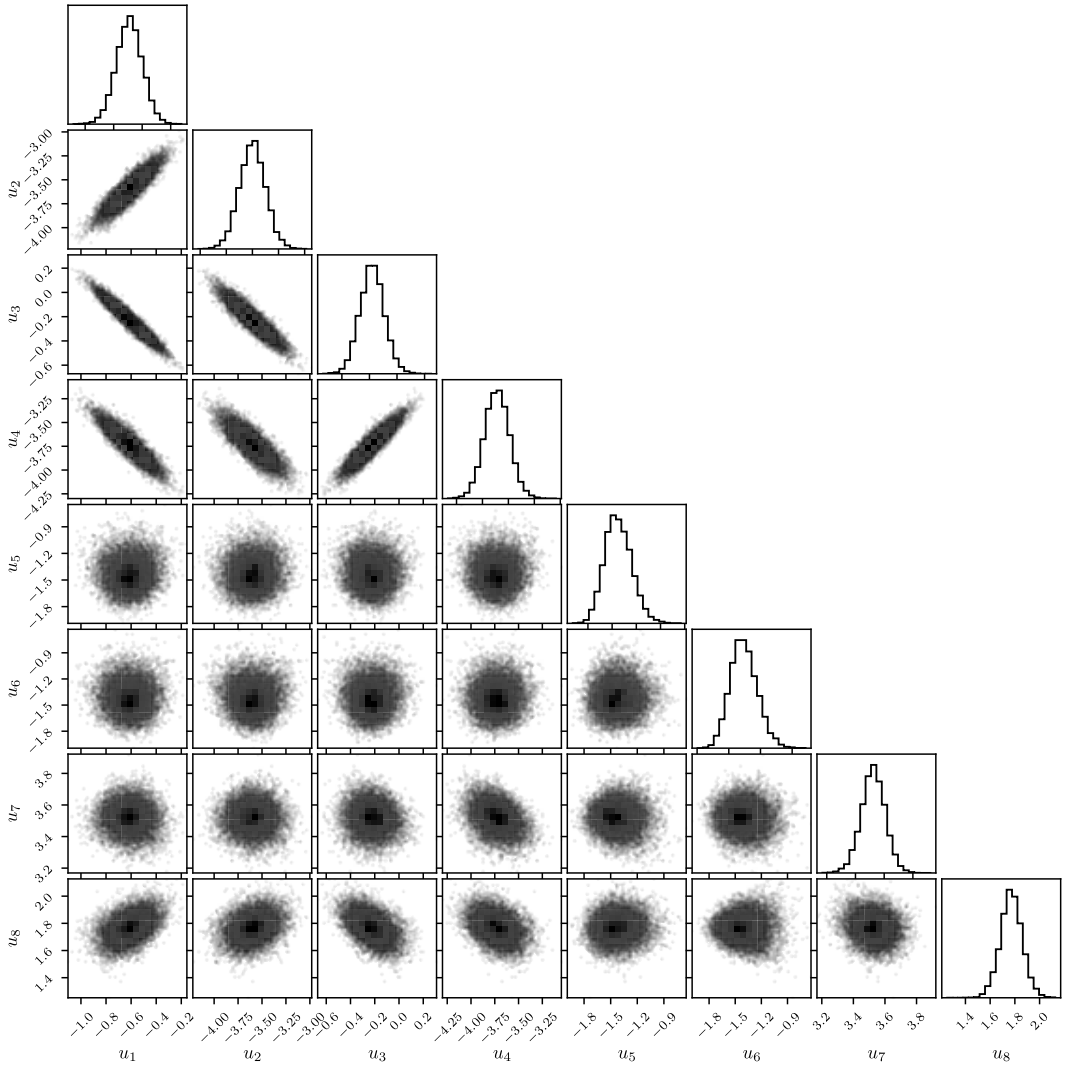

Fig. 9: *Lotka–Volterra*: Unconstrained space

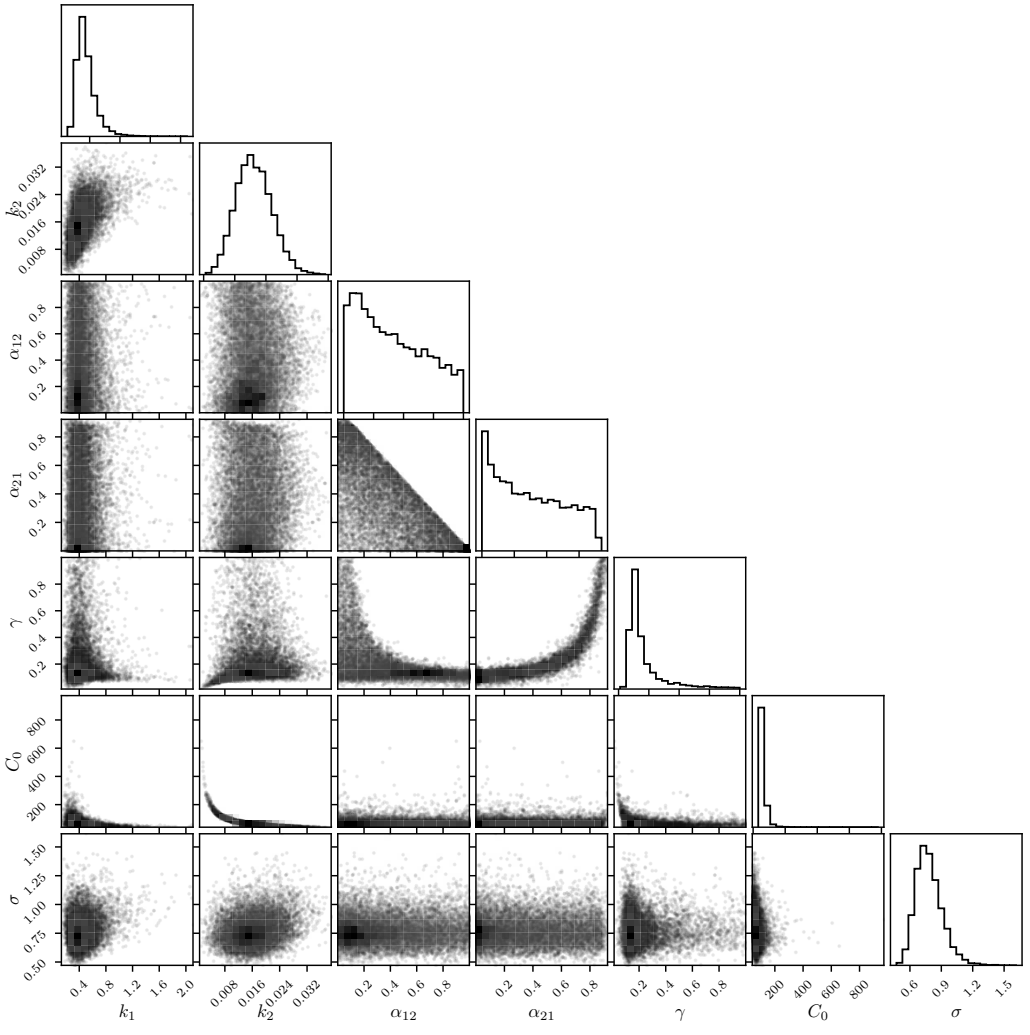

Fig. 10: *Soil incubation (AK-T25)*: Parameter space

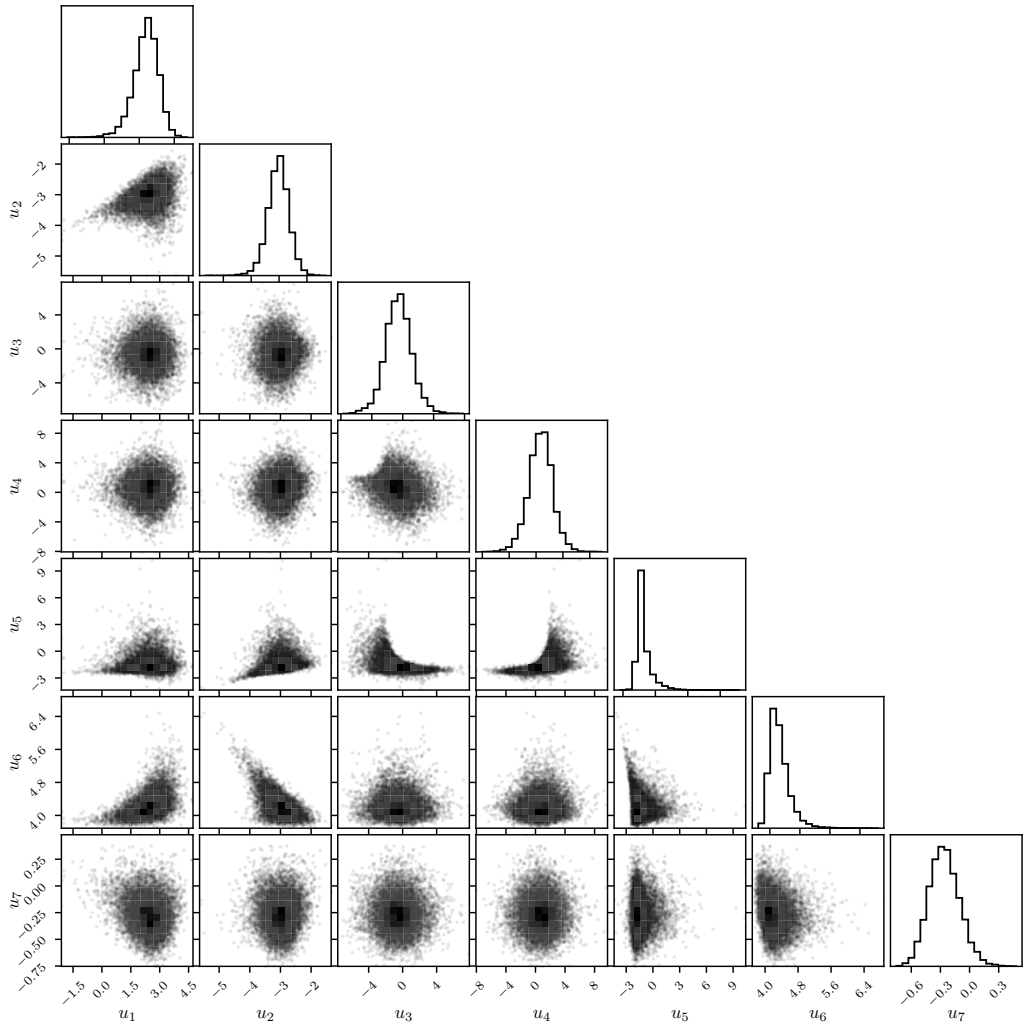

Fig. 11: *Soil incubation (AK-T25)*: Unconstrained space

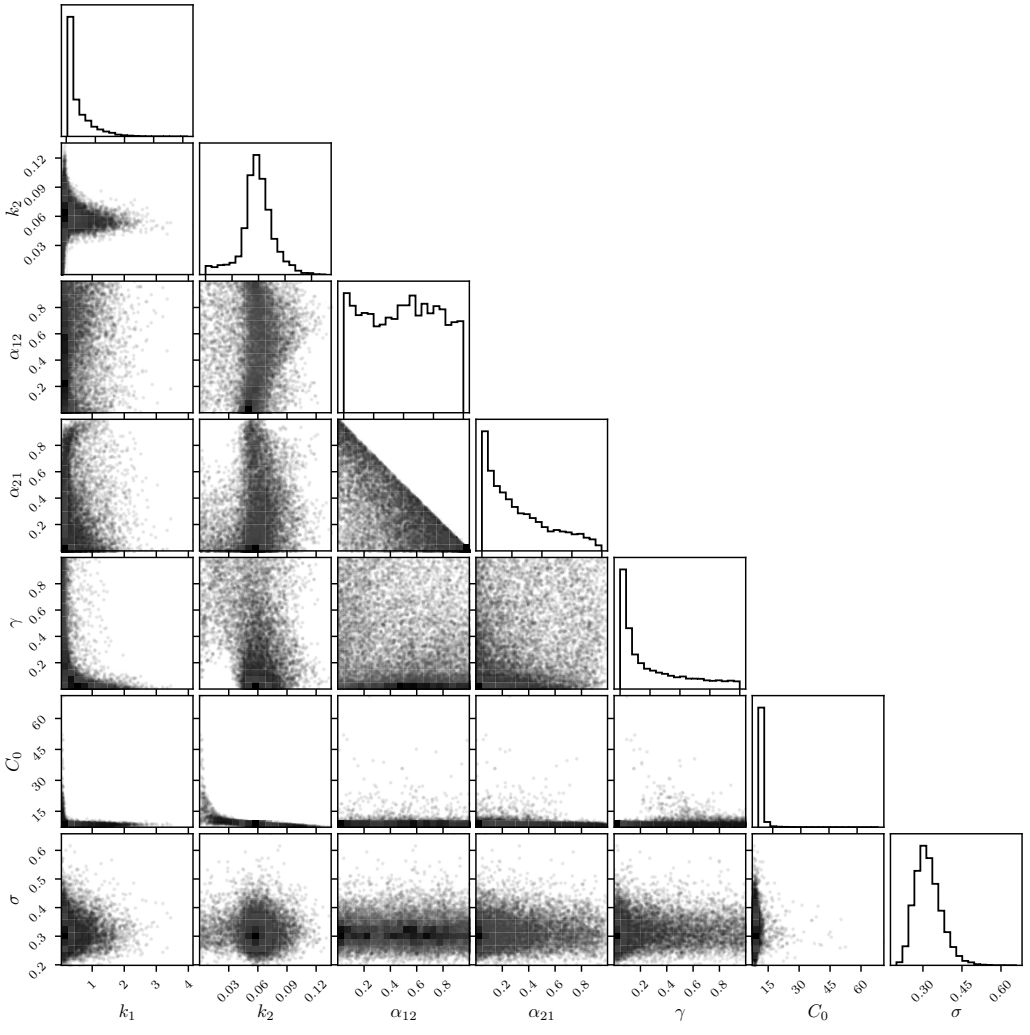

Fig. 12: *Soil incubation (HN-T35)*: Parameter space

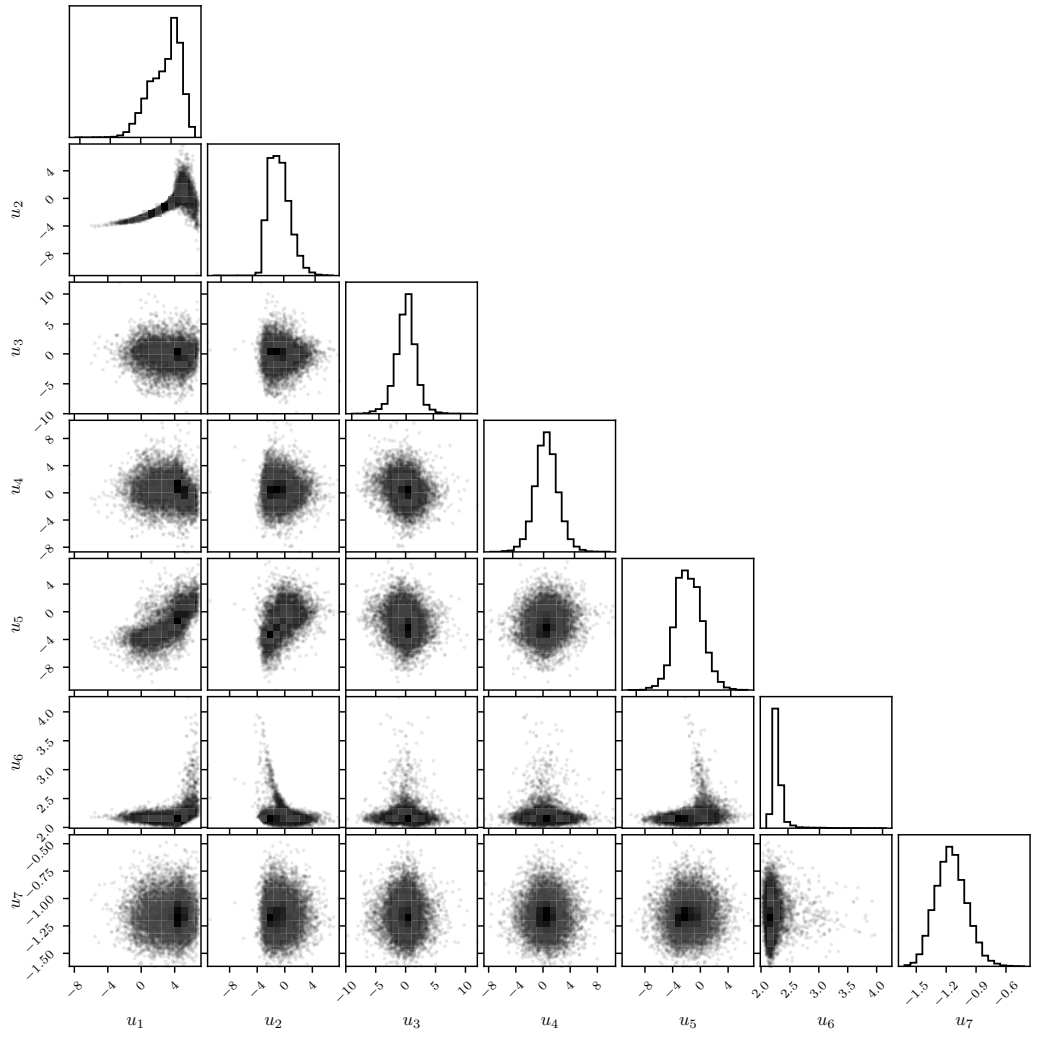

Fig. 13: *Soil incubation (HN-T35)*: Unconstrained space

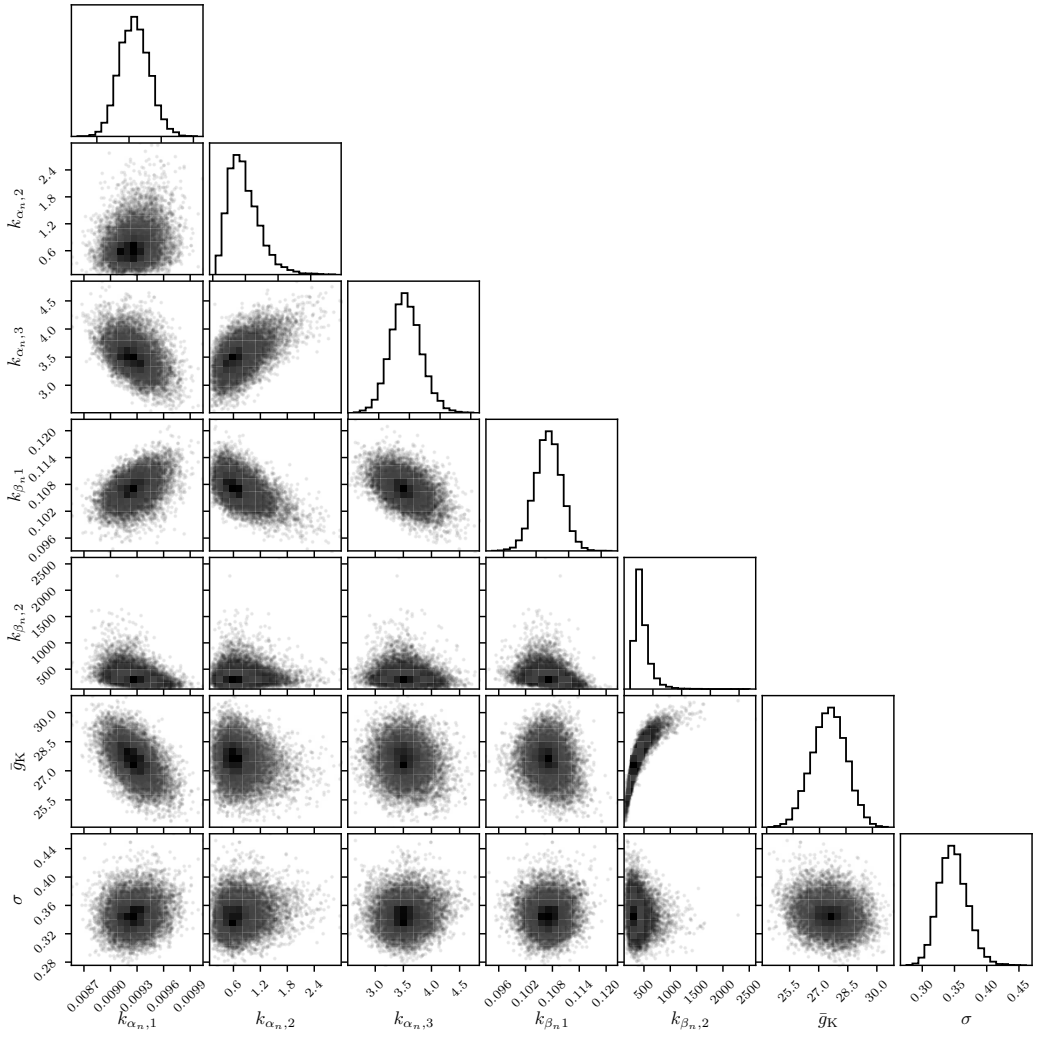

Fig. 14: *Hodgkin–Huxley voltage clamp (potassium)*: Parameter space

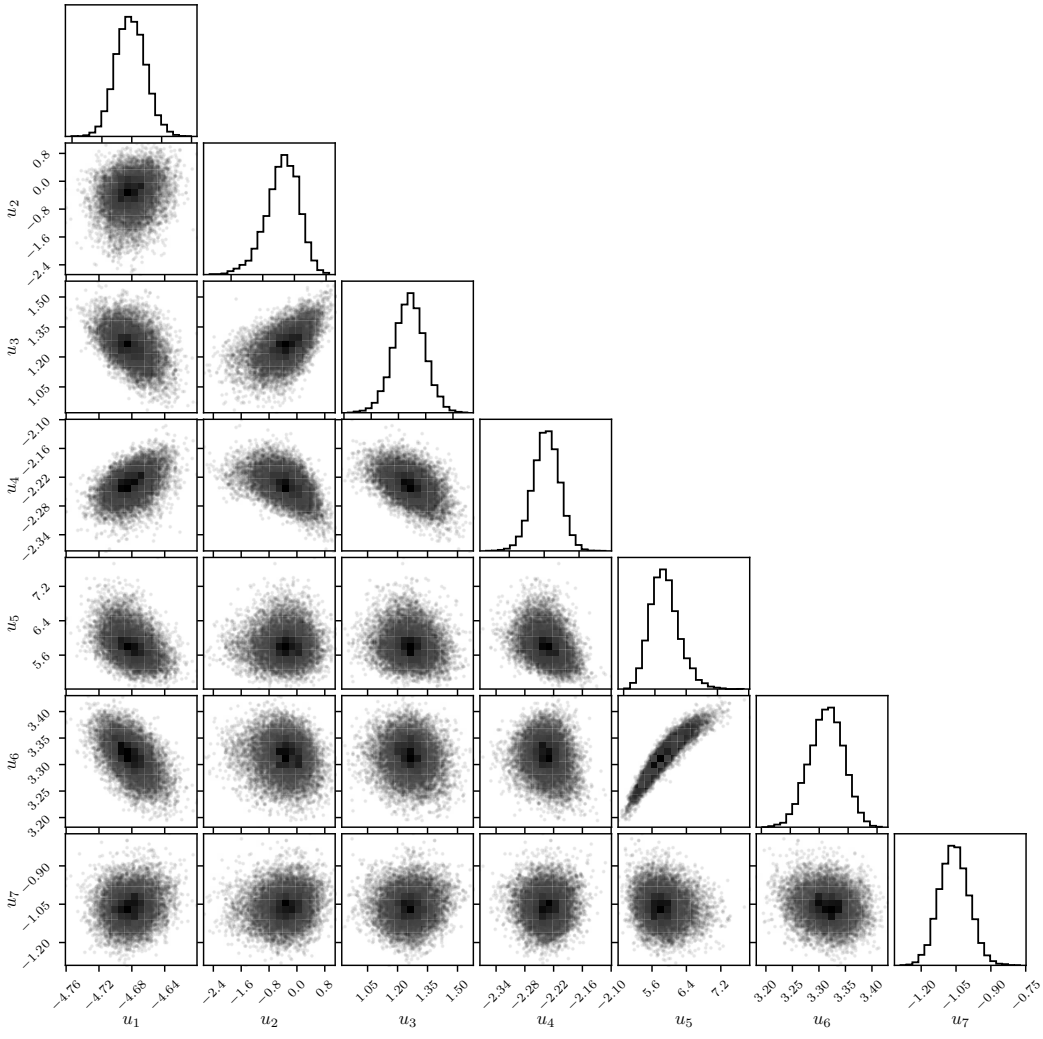

Fig. 15: *Hodgkin–Huxley voltage clamp (potassium)*: Unconstrained space

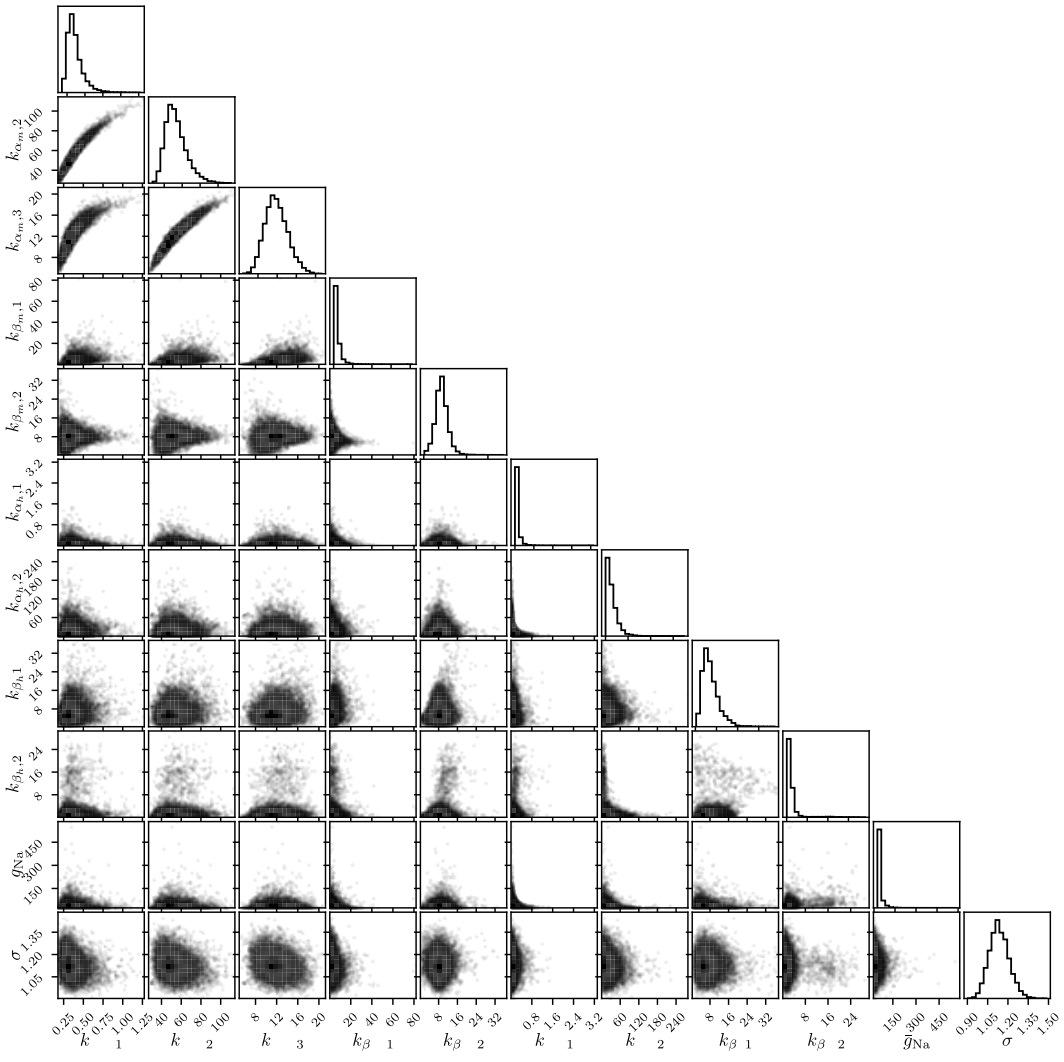

Fig. 16: *Hodgkin–Huxley voltage clamp (sodium)*: Parameter space

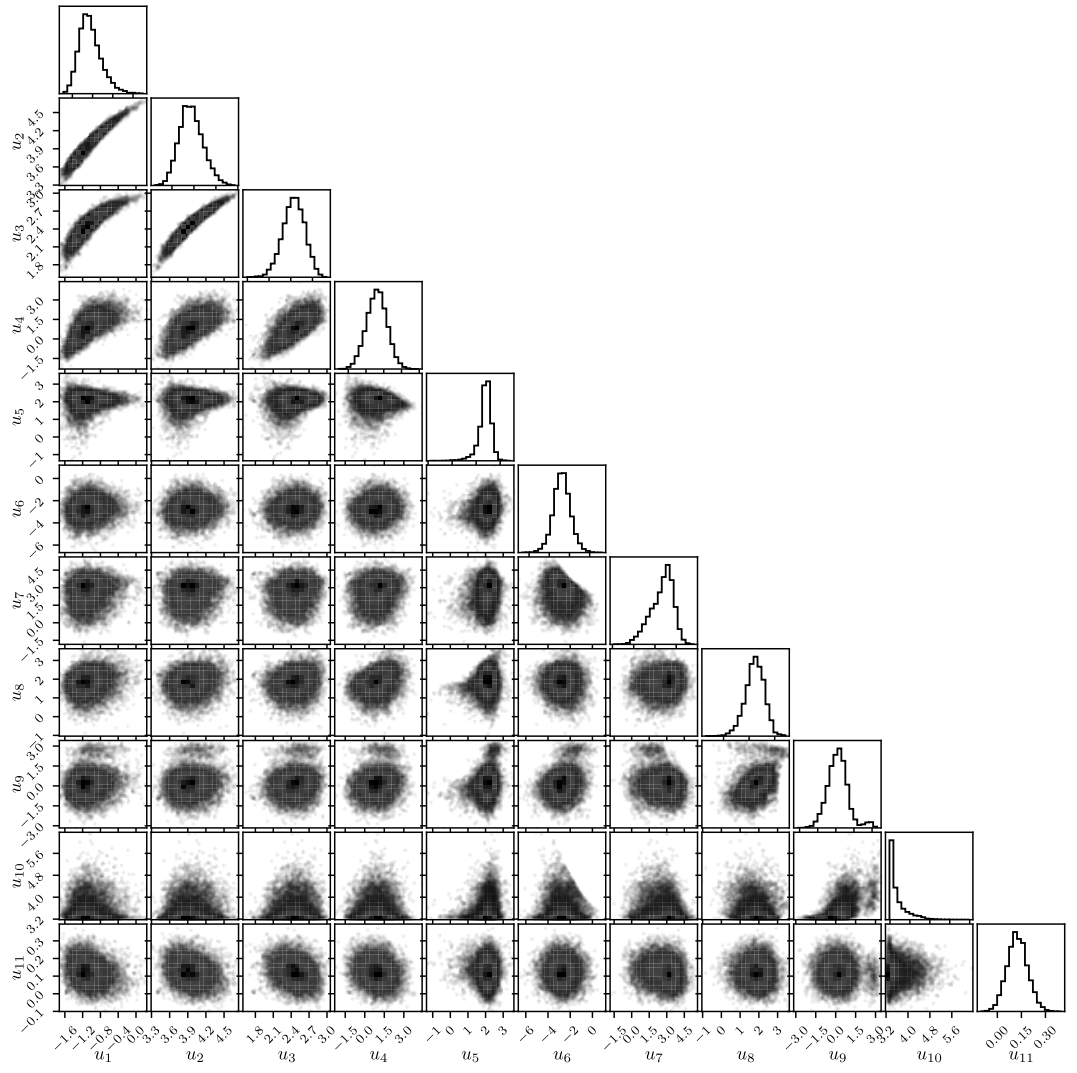

Fig. 17: *Hodgkin–Huxley voltage clamp (sodium)*: Unconstrained space

References

Carpenter, B. (2014) Soil carbon modeling with RStan. *Stan case studies: Volume 1*. URL: https://mc-stan.org/users/documentation/case-studies/soil-knit.html.

— (2018) Predator-prey population dynamics: the Lotka-Volterra model in Stan. *Stan case studies: Volume 5*. URL: https://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html.

Chen, Z., Raman, B. and Stern, A. (2020) Structure-preserving numerical integrators for Hodgkin–Huxley-type systems. *SIAM Journal on Scientific Computing*, **42**, B273–B298.

Daly, A. C., Gavaghan, D. J., Holmes, C. and Cooper, J. (2015) Hodgkin–Huxley revisited: reparametrization and identifiability analysis of the classic action potential model with approximate Bayesian methods. *Royal Society open science*, **2**, 150499.

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**, 123–214.

Griewank, A. (1993) Some bounds on the complexity of gradients, Jacobians, and Hessians. In *Complexity in numerical optimization*, 128–162. World Scientific.

Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 97–109.

Hewitt, C. (1921) *The Conservation of thc Wild Life of Canada*. Charles Scribner's Sons.

Hodgkin, A. L. and Huxley, A. F. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, **117**, 500.

Howard, P. (2009) Modeling basics, lecture notes for Math 442. Texas A&M University. URL: http://www.math.tamu.edu/~phoward/m442/modbasics.pdf.

Livingstone, S. (2015) Geometric ergodicity of the random walk Metropolis with position-dependent proposal covariance. *arXiv preprints*.

Lotka, A. (1925) *Elements of physical biology*. Williams and Wilkins.

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**, 1087–1092.

Pospischil, M., Toledo-Rodriguez, M., Monier, C., Piwkowska, Z., Bal, T., Frégnac, Y., Markram, H. and Destexhe, A. (2008) Minimal hodgkin–huxley type models for different classes of cortical and thalamic neurons. *Biological cybernetics*, **99**, 427–441.

Sierra, C., Müller, M. and Trumbore, S. E. (2014) Modeling radiocarbon dynamics in soils: SoilR version 1.1. *Geoscientific Model Development*, **7**, 1919–1931.

Traub, R. D. and Miles, R. (1991) *Physiology of single neurons: voltage- and ligand-gated ionic channels*, 34–45. Cambridge University Press.

Volterra, V. (1926) Fluctuations in the abundance of a species considered mathematically. *Nature*, **118**, 558–560.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S. and Girolami, M. (2014) Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, **91**, 14–19.