Efficient Metropolis-Hastings Sampling for Nonlinear Mixed Effects Models

Belhal Karimi and Marc Lavielle

Abstract The ability to generate samples of the random effects from their conditional distributions is fundamental for inference in mixed effects models. Random walk Metropolis is widely used to conduct such sampling, but such a method can converge slowly for medium dimension problems, or when the joint structure of the distributions to sample is complex. We propose a Metropolis–Hastings (MH) algorithm based on a multidimensional Gaussian proposal that takes into account the joint conditional distribution of the random effects and does not require any tuning, in contrast with more sophisticated samplers such as the Metropolis Adjusted Langevin Algorithm or the No-U-Turn Sampler that involve costly tuning runs or intensive computation. Indeed, this distribution is automatically obtained thanks to a Laplace approximation of the original model. We show that such approximation is equivalent to linearizing the model in the case of continuous data. Numerical experiments based on real data highlight the very good performances of the proposed method for continuous data model.

Key words: Nonlinear, MCMC, Metropolis, Mixed Effects, Sampling

1 Introduction

Mixed effects models are reference models when the inter-individual variability that can exist within the same population is considered (see [9] and the references therein). Given a population of individuals, the probability distribution of the series of observations for each individual depends on a vector of individual parameters. For

Belhal Karimi
Inria, Paris, France, e-mail: belhal.karimi@inria.fr

Marc Lavielle
Inria, Paris, France, e-mail: marc.lavielle@inria.fr

complex priors on these individual parameters or models, Monte Carlo methods must be used to approximate the conditional distribution of the individual parameters given the observations. Most often, direct sampling from this conditional distribution is impossible and it is necessary to have resort to a Markov chain Monte Carlo (MCMC) procedure.

Designing a fast mixing sampler is of utmost importance for several tasks in the complex process of model building. The most common MCMC method for nonlinear mixed effects models is the *random walk Metropolis* algorithm [14, 15, 9]. Despite its simplicity, it has been successfully used in many classical examples of pharmacometry, when the number of random effects is not too large. Nevertheless, maintaining an optimal acceptance rate (advocated in [15]) most often implies very small moves and therefore a very large number of iterations in medium and high dimensions since no information of the geometry of the target distribution is used.

To make better use of this geometry and in order to explore the space faster, the Metropolis-adjusted Langevin algorithm (MALA) uses evaluations of the gradient of the target density for proposing new states which are accepted or rejected using the Metropolis-Hastings algorithm [16, 18]. The No-U-Turn Sampler (NUTS) is an extension of the Hamiltonian Monte Carlo [11] that allows an automatic and optimal selection of some of the settings required by the algorithm, [3]. Nevertheless, these methods may be difficult to use in practice, and are computationally involved, in particular when the structural model is a complex ODE based model.

The algorithm we propose is a Metropolis-Hastings algorithm, but for which the proposal is a good approximation of the the target distribution. For general data model (i.e. categorical, count or time-to-event data models or continuous data models), the Laplace approximation of the incomplete pdf $\mathbf{p}(y_i)$ leads to a Gaussian approximation of the conditional distribution $\mathbf{p}(\psi_i|y_i)$.

In the special case of continuous data, linearisation of the model leads, by definition, to a Gaussian linear model for which the conditional distribution of the individual parameter ψ_i given the data y_i is a multidimensional normal distribution that can be computed and we fall back on the results of [8].

2 Mixed Effect Models

2.1 Population Approach and Hierarchical Models

We will adopt a population approach in the sequel, where we consider N individuals and n_i observations for individual i. The set of observed data is $y = (y_i, 1 \leq i \leq N)$ where $y_i = (y_{ij}, 1 \leq j \leq n_i)$ are the observations for individual i. For the sake of clarity, we assume that each observation y_{ij} takes its values in some subset of $\mathbb{R}$. The distribution of the n_i-vector of observations y_i depends on a vector of individual parameters ψ_i that takes its values in a subset of $\mathbb{R}^p$. We assume that the pairs (y_i, ψ_i) are mutually independent and consider a parametric framework: the joint distribution of (y_i, ψ_i) is denoted by $\mathbf{p}(y_i, \psi_i; \theta)$, where θ is the vector of fixed

parameters of the model. A natural decomposition of this joint distribution writes $\mathrm{p}(y_i, \psi_i; \theta) = \mathrm{p}(y_i|\psi_i; \theta)\mathrm{p}(\psi_i; \theta)$, where $\mathrm{p}(y_i|\psi_i; \theta)$ is the conditional distribution of the observations, given the individual parameters, and where $\mathrm{p}(\psi_i; \theta)$ is the so-called population distribution used to describe the distribution of the individual parameters within the population. A particular case of this general framework consists in describing each individual parameters ψ_i as a typical value ψ_{pop}, and a vector of individual random effects η_i: $\psi_i = \psi_{\text{pop}} + \eta_i$. In the sequel, we will assume a multivariate Gaussian distribution for the random effects: $\eta_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \Omega)$. Several extensions of this model are straightforward, considering for instance transformation of the normal distribution, or adding individual covariates in the model.

2.2 Continuous Data Models

A regression model is used to express the link between continuous observations and individual parameters:

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \;, \tag{1}$$

where y_{ij} is the j-th observation for individual i measured at time t_{ij}, ε_{ij} is the residual error, f is the structural model assumed to be a twice differentiable function of ψ_i. We start by assuming that the residual errors are independent and normally distributed with zero-mean and a constant variance σ^2. Let $t_i = (t_{ij}, 1 \leq n_i)$ be the vector of observation times for individual i. Then, the model for the observations rewrites $y_i|\psi_i \sim \mathcal{N}(f_i(\psi_i), \sigma^2 \mathtt{Id}_{n_i \times n_i})$, where $f_i(\psi_i) = (f(t_{i,1}, \psi_i), \ldots, f(t_{i,n_i}, \psi_i))$. If we assume that $\psi_i \sim_{\text{i.i.d.}} \mathcal{N}(\psi_{\text{pop}}, \Omega)$, then the parameters of the model are $\theta = (\psi_{\text{pop}}, \Omega, \sigma^2)$.

3 Sampling from Conditional Distributions

The conditional distribution $\mathrm{p}(\psi_i|y_i; \theta)$ plays a crucial role in most methods used for inference in nonlinear mixed effects models.

One of the main task to perform is to compute the maximum likelihood (ML) estimate of θ, $\hat{\theta}_{\text{ML}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, y)$, where $\mathcal{L}(\theta, y) \triangleq \log \mathrm{p}(y; \theta)$. The stochastic approximation version of EM [7] is an iterative procedure for ML estimation that requires to generate one or several realisations of this conditional distribution at each iteration of the algorithm.

Metropolis-Hasting algorithm is a powerful MCMC procedure widely used for sampling from a complex distribution [4]. To simplify the notations, we remove the dependency on θ. For a given individual i, the MH algorithm, to sample from the conditional distribution $\mathrm{p}(\psi_i|y_i)$, is described as:

Algorithm 1 Metropolis-Hastings algorithm

Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution ξ_i .
Iteration k: given the current state of the chain $\psi_i^{(k-1)}$:

1. Sample a candidate ψ_i^c from a proposal distribution $q_i(\,\cdot\,|\psi_i^{(k-1)})$.
2. Compute the MH ratio:

$$\alpha(\psi_i^{(k-1)}, \psi_i^c) = \frac{\mathrm{p}(\psi_i^c\,|y_i)}{\mathrm{p}(\psi_i^{(k-1)}|y_i)} \frac{q_i(\psi_i^{(k-1)}|\psi_i^c)}{q_i(\psi_i^c\,|\psi_i^{(k-1)})} \,. \tag{2}$$

3. Set $\psi_i^{(k)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(k-1)})$ (otherwise, keep $\psi_i^{(k)} = \psi_i^{(k-1)}$).

Current implementations of the MCMC algorithm, to which we will compare our new method, in Monolix [5], saemix (R package) [6], nlmefitsa (Matlab) and NONMEM [2] mainly use the same combination of proposals. The first proposal is an independent Metropolis-Hasting algorithm which consists in sampling the candidate state directly from the marginal distribution of the individual parameter ψ_i. The other proposals are component-wise and block-wise random walk procedures [10] that updates different components of ψ_i using univariate and multivariate Gaussian proposal distributions. Nevertheless, those proposals fail to take into account the nonlinear dependence structure of the individual parameters. A way to alleviate these problems is to use a proposal distribution derived from a discretised Langevin diffusion whose drift term is the gradient of the logarithm of the target density leading to the Metropolis Adjusted Langevin Algorithm (MALA) [16, 18]. The MALA proposal is given by:

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma \nabla_{\psi_i} \log \mathrm{p}(\psi_i^{(k)}|y_i), 2\gamma) \,, \tag{3}$$

where γ is a positive stepsize. These methods still do not take into consideration the multidimensional structure of the individual parameters. Recent works include efforts in that direction, such as the Anisotropic MALA for which the covariance matrix of the proposal depends on the gradient of the target measure [1]. The MALA algorithm is a special instance of the Hybrid Monte Carlo (HMC), introduced in [11]; see [4] and the references therein, and consists in augmenting the state space with an auxiliary variable p, known as the velocity in Hamiltonian dynamics.

All those methods aim at finding the proposal q that accelerates the convergence of the chain. Unfortunately they are computationally involved and can be difficult to implement (stepsizes and numerical derivatives need to be tuned and implemented).

We see in the next section how to define a multivariate Gaussian proposal for both continuous and noncontinuous data models, that is easy to implement and that takes into account the multidimensional structure of the individual parameters in order to accelerate the MCMC procedure.

4 A Multivariate Gaussian Proposal

For a given parameter value θ, the MAP estimate, for individual i, of ψ_i is the one that maximises the conditional distribution $\mathrm{p}(\psi_i|y_i,\theta)$:

$$\hat{\psi}_i = \arg\max_{\psi_i} \mathrm{p}(\psi_i|y_i,\theta) = \arg\max_{\psi_i} \mathrm{p}(y_i|\psi_i,\theta)\mathrm{p}(\psi_i,\theta)$$

4.1 General Data Models

For both continuous and noncontinuous data models, the goal is to find a simple proposal, a multivariate Gaussian distribution in our case, that approximates the target distribution $\mathrm{p}(\psi_i|y_i)$. In our context, we can write the marginal pdf $\mathrm{p}(y_i)$ that we aim to approximate as $\mathrm{p}(y_i) = \int e^{\log \mathrm{p}(y_i,\psi_i)} \mathrm{d}\psi_i$. Then, the Taylor expansion of $\log(\mathrm{p}(y_i,\psi_i)$ around the MAP $\hat{\psi}_i$ (that verifies by definition $\nabla \log \mathrm{p}(y_i,\hat{\psi}_i) = 0$) yields the Laplace approximation of $-2\log(\mathrm{p}(y_i))$ as follows:

$$-2\log \mathrm{p}(y_i) \approx -p\log 2\pi - 2\log \mathrm{p}(y_i,\hat{\psi}_i) + \log\left(\left|-\nabla^2 \log \mathrm{p}(y_i,\hat{\psi}_i)\right|\right) \ .$$

We thus obtain the following approximation of $\log \mathrm{p}(\hat{\psi}_i|y_i)$:

$$\log \mathrm{p}(\hat{\psi}_i|y_i) \approx -\frac{p}{2}\log 2\pi - \frac{1}{2}\log\left(\left|-\nabla^2 \log \mathrm{p}(y_i,\hat{\psi}_i)\right|\right) \ ,$$

which is precisely the log-pdf of a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance $-\nabla^2 \log \mathrm{p}(y_i,\hat{\psi}_i)^{-1}$, evaluated at $\hat{\psi}_i$.

Proposition 1 *The Laplace approximation of the conditional distribution $\psi_i|y_i$ is a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance*

$$\Gamma_i = -\nabla^2 \log p(y_i,\hat{\psi}_i)^{-1} = \left(-\nabla^2 \log p(y_i|\hat{\psi}_i) + \Omega^{-1}\right)^{-1} \ .$$

We shall now see another method to derive a Gaussian proposal distribution in the specific case of continuous data models.

4.2 Nonlinear Continuous Data Models

When the model is described by (1), the approximation of the target distribution can be done twofold: either by using the Laplace approximation, as explained above, or by linearizing the structural model f_i for any individual i of the population. Once the MAP estimate $\hat{\psi}_i$ has been computed, using an optimisation procedure, the method is based on the linearisation of the structural model f around $\hat{\psi}_i$:

$$f_i(\psi_i) \approx f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}(\psi_i - \hat{\psi}_i)\,, \tag{4}$$

where $\mathrm{J}_{f_i(\hat{\psi}_i)}$ is the Jacobian matrix of the vector $f_i(\hat{\psi}_i)$. Defining $z_i \triangleq y_i - f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}\hat{\psi}_i$ yields a linear model $z_i = \mathrm{J}_{f_i(\hat{\psi}_i)}\psi_i + \epsilon_i$ which tractable conditional distribution can be used for approximating $\mathrm{p}(\psi_i|y_i, \theta)$:

Proposition 2 *Under this linear model, the conditional distribution $\psi_i|y_i$ is a Gaussian distribution with mean μ_i and variance-covariance Γ_i where*

$$\mu_i = \hat{\psi}_i \quad \textit{and} \quad \Gamma_i = \left(\frac{J'_{f_i(\hat{\psi}_i)} J_{f_i(\hat{\psi}_i)}}{\sigma^2} + \Omega^{-1}\right)^{-1}\,. \tag{5}$$

We can note that linearizing the structural model is equivalent to using the Laplace approximation with the expected information matrix. Indeed:

$$\mathbb{E}_{y_i|\hat{\psi}_i}\left(-\nabla^2 \log \mathrm{p}(y_i|\hat{\psi}_i)\right) = \frac{\mathrm{J}'_{f_i(\hat{\psi}_i)}\mathrm{J}_{f_i(\hat{\psi}_i)}}{\sigma^2}\,. \tag{6}$$

We then use this normal distribution as a proposal in algorithm 1 for model (1).

5 A Pharmacokinetic Example

5.1 Data and Model

32 healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis [12], for who we measure warfarin plasmatic concentration at different times. We will consider a one-compartment pharmacokinetics (PK) model for oral administration, assuming first-order absorption and linear elimination processes:

$$f(t, ka, V, k) = \frac{D\,ka}{V(ka - k)}(e^{-ka\,t} - e^{-k\,t})\,, \tag{7}$$

where ka is the absorption rate constant, V the volume of distribution , k the elimination rate constant, and D the dose administered. Here, ka, V and k are PK parameters that can change from one individual to another. Then, let $\psi_i = (ka_i, V_i, k_i)$ be the vector of individual PK parameters for individual i lognormally distributed. We will assume in this example that the residual errors are independent and normally distributed with mean 0 and variance σ^2. We can use the proposal given by Proposition 2 and based on a linearisation of the structural model f proposed in (7). For the method to be easily extended to any structural model, the gradient is calculated by automatic differentiation using the R package 'Madness' [13].

5.2 MCMC Convergence Diagnostic

We will consider one of the 32 individuals for this study and fix θ to some arbitrary value, close to the Maximum Likelihood (ML) estimate obtained with SAEM (saemix R package [6]): $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.01$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. First, we compare our our nlme-IMH, which is a MH sampler using our new proposal, with the RWM, the MALA, which proposal, at iteration k, is defined by $\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma_k \nabla \log \pi(\psi_i^{(k)}), 2\gamma_k)$. The stepsize ($\gamma = 10^{-2}$) is constant and is tuned such that the optimal acceptance rate of 0.57 is reached [15]. 20 000 iterations are run for each algorithm. Figure 1 highlights quantiles stabilisation using the MALA similar to our method for all orders and dimensions. The NUTS, implemented in rstan (R Package [17]), is fast and steady and presents similar, or even better convergence behaviors for some quantiles and dimension, than our method (see Figure 1).

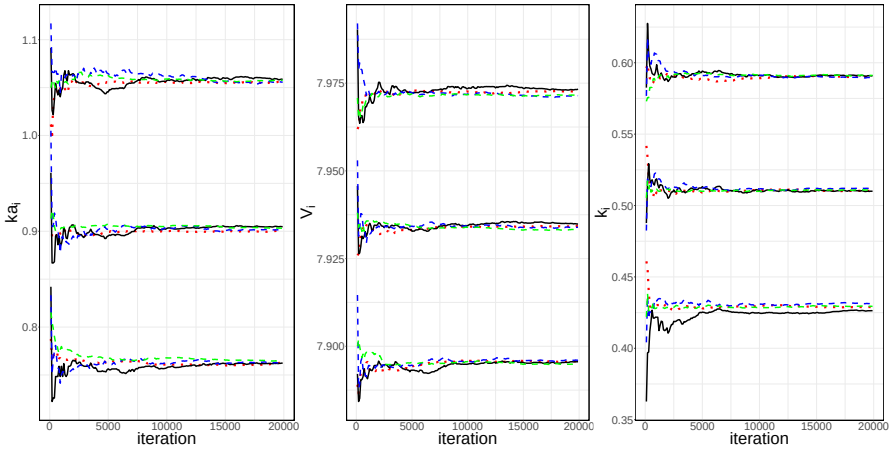

Fig. 1 Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\text{p}(\psi_i | y_i; \theta)$ for a single individual. Our new MH algorithm is in red and dotted, the RWM is in black and solid, the MALA is in blue and dashed and the NUTS is in green and dashed.

Then, we produce 100 independent runs of the RWM, the IMH using our proposal distribution (called the nlme-IMH algorithm), the MALA and the NUTS for 500 iterations. The boxplots of the samples drawn at a given iteration threshold are presented Figure 2 against the ground truth (calculated running the NUTS for 100 000 iterations) for the parameter **ka**.

For the three numbers of iteration considered in Figure 2, the median of the nlme-IMH and NUTS samples are closer to the ground truth. Figure 2 also highlights that all those methods succeed in sampling from the whole distribution after 500 iterations. Similar comments can be made for the other parameters.

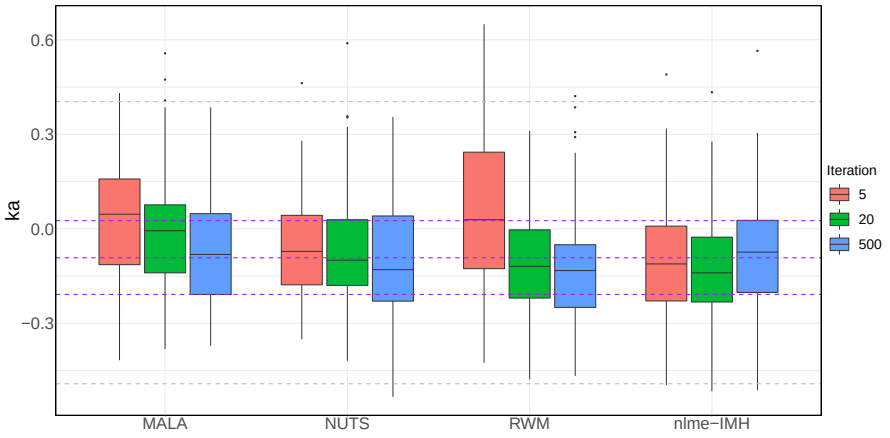

Fig. 2 Modelling of the warfarin PK data: Boxplots, over 100 parallel runs, for the RWM, the nlme-IMH, the MALA and the NUTS algorithm. The ground truth median, 0.25 and 0.75 percentiles are plotted as a dashed purple line and its maximum and minimum as a dashed grey line.

We decided to conduct a comparison between those sampling methods just in terms of number of iterations (one iteration is one transition of the Markov Chain). We acknowledge that the transition cost is not the same for each of those algorithms, though, our nmle-IMH algorithm, except the initialisation step that requires a MAP and a Jacobian computation, has the same iteration cost as RWM. The call to the structural model f being very costly in real applications (when the model is the solution of a complex ODE for instance), the MALA and the NUTS, computing its first order derivatives at each transition, are thus far computationally involved.

Since computational costs per transition are hard to accurately define for each sampling algorithm and since runtime depends on the actual implementation of those methods, comparisons are based on the number of iterations of the chain here.

6 Conclusion and Discussion

We presented in this article an independent Metropolis-Hastings procedure for sampling random effects from their conditional distributions in nonlinear mixed effects models. The numerical experiments that we have conducted show that the proposed sampler converges to the target distribution as fast as state-of-the-art samplers. This good practical behaviour is partly explained by the fact that the conditional mode of the random effects in the linearised model coincides with the conditional mode of the random effects in the original model. Initial experiments embedding this fast and easy-to-implement IMH algorithm within the SAEM algorithm [7], for Maximum Likelihood Estimation, indicate a faster convergence behavior.

References

[1] Allassonniere, S. and Kuhn, E. (2013). Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation. *arXiv preprint arXiv:1207.5938*.

[2] Beal, S. and Sheiner, L. (1980). The NONMEM system. *The American Statistician*, 34(2):118–119.

[3] Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

[4] Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

[5] Chan, P. L. S., Jacqmin, P., Lavielle, M., McFadyen, L., and Weatherley, B. (2011). The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. *Journal of Pharmacokinetics and Pharmacodynamics*, 38(1):41–61.

[6] Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software*, 80(3):1–42.

[7] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128.

[8] Karimi, B., Lavielle, M., and Moulines, E. (2017). Non linear mixed effects models: bridging the gap between independent metropolis-hastings and variational inference. In *ICML 2017 Implicit Models Workshop*.

[9] Lavielle, M. (2014). *Mixed Effects Models for The Population Approach: Models, Tasks, Methods and Tools*. CRC press.

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

[11] Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

[12] O'Reilly, R. A. and Aggeler, P. M. (1968). Studies on Coumarin anticoagulant drugs initiation of Warfarin therapy without a lading dose. *Circulation*, 38(1):169–177.

[13] Pav, S. E. (2016). Madness: a package for multivariate automatic differentiation.

[14] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics.

[15] Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120.

[16] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

[17] Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3.

[18] Stramer, O. and Tweedie, R. L. (1999). Langevin-type models i: diffusions with given stationary distributions and their discretizations. *Methodology And Computing In Applied Probability*, 1(3):283–306.