Biometrika (2017), **103**, 1, *pp*. 1–14
Printed in Great Britain Advance Access publication on 31 July 2016

Large Sample Asymptotics of the Pseudo-Marginal Method

BY S. M. SCHMON, G. DELIGIANNIDIS, A. DOUCET
Department of Statistics, University of Oxford
24-29 St Giles', Oxford OX1 3LB
schmon@stats.ox.ac.uk deligiannidis@stats.ox.ac.uk doucet@stats.ox.ac.uk

AND M. K. PITT
Department of Mathematics, King's College London
Strand, London WC2R 2LS
michael.pitt@kcl.ac.uk

SUMMARY

The pseudo-marginal algorithm is a variant of the Metropolis–Hastings algorithm which samples asymptotically from a probability distribution when it is only possible to estimate unbiasedly an unnormalized version of its density. Practically, one has to trade-off the computational resources used to obtain this estimator against the asymptotic variances of the ergodic averages obtained by the pseudo-marginal algorithm. Recent works optimizing this trade-off rely on some strong assumptions which can cast doubts over their practical relevance. In particular, they all assume that the distribution of the additive error in the log-likelihood estimator is independent of the parameter value at which it is evaluated. Under weak regularity conditions we show here that, as the number of data points tends to infinity, a space-rescaled version of the pseudo-marginal chain converges weakly towards another pseudo-marginal chain for which this assumption indeed holds. A study of this limiting chain allows us to provide parameter dimension-dependent guidelines on how to optimally scale a normal random walk proposal and the number of Monte Carlo samples for the pseudo-marginal method in the large sample regime. This complements and validates currently available results.

Some key words: Asymptotic posterior normality; Intractable likelihood; Large sample theory; Metropolis–Hastings algorithm; Random measure; Weak convergence.

1. INTRODUCTION

The pseudo-marginal Metropolis–Hastings algorithm is a variant of the popular Metropolis–Hastings algorithm where an unnormalized version of the target density is replaced by a non-negative unbiased estimate. The algorithm first appeared in the physics literature (Lin et al., 2000) and has become very popular in Bayesian statistics as many intractable likelihood functions can be estimated unbiasedly using importance sampling or particle filters (Beaumont, 2003; Andrieu & Roberts, 2009; Andrieu et al., 2010).

Replacing the true likelihood in the Metropolis-Hastings algorithm with an estimate results in a trade-off: the asymptotic variance of an ergodic average of a pseudo-marginal chain typically decreases as the number of samples, N, used to obtain the likelihood estimator increases, as established by Andrieu & Vihola (2016) for importance sampling estimators; however, this comes at the cost of a higher computational burden. An important task in practice is thus to choose N

such that the computational resources required to obtain a given asymptotic variance are minimized. This problem has already been investigated by Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) where guidelines have been obtained under various assumptions either on the proposal (Pitt et al., 2012; Doucet et al., 2015) or on the proposal and target distribution (Sherlock et al., 2015). Additionally, all these contributions make the assumption that the distribution of the additive noise introduced by the log-likelihood estimator is a Gaussian of variance inversely proportional to N, its mean and variance being independent of the parameter value at which it is evaluated. A similar assumption has also been used by Nemeth et al. (2016) for the analysis of a related algorithm. This assumption can cast doubts over the practical relevance of the guidelines provided in these contributions. The normal noise assumption was motivated by Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) by the fact that the error in the log-likelihood estimator for state-space models computed using a particle filter is asymptotically normal of variance proportional to γ as $T \to \infty$ with $N = T/\gamma$ (Bérard et al., 2014) while the constant variance assumption over the parameter space was motivated in Pitt et al. (2012) and Doucet et al. (2015) by the fact that the posterior typically concentrates as T increases. However, no formal argument justifying why the pseudo-marginal chain would behave as a Markov chain for which these assumptions hold has been provided.

We carry out here an original weak convergence analysis of the pseudo-marginal algorithm in a Bayesian setting which not only justifies rigorously these assumptions but also allows us to obtain novel guidelines on how to optimally tune this algorithm as a function of the parameter dimension d. Weak convergence techniques have become very popular in the Markov chain Monte Carlo literature since their introduction in the seminal paper of Roberts et al. (1997). To the recent exception of Deligiannidis et al. (2015), all these analyses have been performed in the asymptotic regime where the parameter dimension $d \to \infty$. Results of this type typically require to make strong structural assumptions on the target distribution such as having d independent and identically distributed components as in Sherlock et al. (2015). We analyze here the pseudo-marginal scheme in the large sample asymptotic regime where the number of data points T goes to infinity while d is fixed. When the posterior distribution concentrates towards a normal, we show that a space-rescaled version of the pseudo-marginal chain converges to a pseudo-marginal chain targeting a normal distribution for which the additive error, or noise, in the log-likelihood estimator is indeed also normal of constant mean and variance. For normal random walk proposals, we provide numerical results to optimally scale the proposal and the noise variance to optimize the performance of this limiting Markov chain as a function of d. These results complement and validate the results obtained in previous contributions, bridging the gap between the guidelines proposed in Doucet et al. (2015) and Sherlock et al. (2015). All proofs can be found in the supplementary material.

2. The Pseudo-Marginal Algorithm

2·1. *Background*

Consider a Bayesian model on the Borel space $\{\Theta, \mathcal{B}(\Theta)\}$ where $\Theta \subseteq \mathbb{R}^d$. The parameter $\theta \in \Theta$ follows a prior distribution $p(\mathrm{d}\theta)$ while $\theta \mapsto p(y \mid \theta)$ denotes the likelihood function, where $y = (y_1, \ldots, y_T)$ denotes the vector of observations. When the likelihood arises from a complex latent variable model an analytic expression of $p(y \mid \theta)$ might not be available. Hence, the standard Metropolis–Hastings algorithm cannot be used to sample the posterior distribution $p(\mathrm{d}\theta \mid y) \propto p(\mathrm{d}\theta)\, p(y \mid \theta)$ as the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$ appearing in the Metropolis–Hastings acceptance probability when at parameter θ and proposing θ' cannot be computed. Assume we have access to an unbiased positive estimator $\hat{p}(y \mid \theta, U)$

of the intractable likelihood $p(y \mid \theta)$, where $U \sim m_\theta$ represents the auxiliary variables on $\{\mathcal{U}, \mathcal{B}(\mathcal{U})\}$ used to compute this estimator. We introduce the following probability measure on $\{\Theta \times \mathcal{U}, \mathcal{B}(\Theta) \times \mathcal{B}(\mathcal{U})\}$

$$\pi(\mathrm{d}\theta, \mathrm{d}u) = p(\mathrm{d}\theta \mid y) \frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)} m_\theta(\mathrm{d}u),$$

which satisfies $\pi(\mathrm{d}\theta) = p(\mathrm{d}\theta \mid y)$. The pseudo-marginal algorithm is a Metropolis–Hastings scheme targeting $\pi(\mathrm{d}\theta, \mathrm{d}u)$, hence marginally $p(\mathrm{d}\theta \mid y)$, using a proposal distribution $Q\left(\theta, u; \mathrm{d}\theta', \mathrm{d}u'\right) = q(\theta, \mathrm{d}\theta') m_{\theta'}\left(\mathrm{d}u'\right)$. This yields the acceptance probability

$$\alpha(\theta, u; \theta', u') = \min\left\{1, r(\theta, \theta') \frac{\hat{p}(y \mid \theta', u')/p(y \mid \theta')}{\hat{p}(y \mid \theta, u)/p(y \mid \theta)}\right\}, \text{ where } r(\theta, \theta') = \frac{\pi(\mathrm{d}\theta')}{\pi(\mathrm{d}\theta)} \frac{q(\theta', \mathrm{d}\theta)}{q(\theta, \mathrm{d}\theta')}.$$

As in previous contributions (Andrieu & Roberts, 2009; Pitt et al., 2012; Andrieu & Vihola, 2015; Doucet et al., 2015; Sherlock et al., 2015), we analyze the pseudo-marginal algorithm using additive noise in the log-likelihood estimator, writing $Z(\theta) = \log \hat{p}(y \mid \theta, U) - \log p(y \mid \theta)$. This parameterization allows us to write the target distribution as a measure on $\{\Theta \times \mathbb{R}, \mathcal{B}(\Theta) \times \mathcal{B}(\mathbb{R})\}$ with

$$\pi(\mathrm{d}\theta, \mathrm{d}z) = p(\mathrm{d}\theta \mid y) \exp(z)\, g(\mathrm{d}z \mid \theta),$$

where $Z(\theta) \sim g(\cdot \mid \theta)$ when $U \sim m_\theta$ and the pseudo-marginal kernel is

$$P\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q(\theta, \mathrm{d}\theta') g(\mathrm{d}z' \mid \theta') \alpha\left(\theta, z; \theta', z'\right) + \rho(\theta, z) \delta_{(\theta, z)}(\mathrm{d}\theta', \mathrm{d}z'),$$

with acceptance probability

$$\alpha\left(\theta, z; \theta', z'\right) = \min\left\{1, r(\theta, \theta') \exp\left(z' - z\right)\right\},$$

and corresponding rejection probability $\rho(\theta, z)$.

2·2. *Literature Review*

We briefly review here recent research motivating this work. To this end, we first need to introduce a few additional notation. Let μ be a probability measure on $\{\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\}$ and $\Pi\colon \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ a Markov transition kernel. For any measurable function f and measurable set A, we write $\mu(f) = \int f(x) \mu(\mathrm{d}x)$, $\mu(A) = \mu\left\{\mathbb{I}_A(\cdot)\right\}$ and $\Pi f(x) = \int \Pi(x, \mathrm{d}y) f(y)$. We consider the Hilbert space $L^2(\mu)$ with inner product $\langle f, g \rangle_\mu = \int f(x) g(x) \mu(\mathrm{d}x)$. For a function $f \in L^2(\mu)$, the asymptotic variance of averages of a stationary Markov chain $(X_k)_{k \geqslant 1}$ of μ-invariant transition kernel Π is defined as

$$\mathrm{var}(f, \Pi) = \lim_{M \to \infty} \frac{1}{M} E\left\{\sum_{k=1}^{M} f(X_k) - \mu(f)\right\}^2,$$

and $\mathrm{var}(f, \Pi) = \mathrm{var}_\mu(f)\, \tau(f, \Pi)$ whenever the integrated autocorrelation time, $\tau(f, \Pi)$, defined by

$$\tau(f, \Pi) = 1 + 2 \sum_{k=1}^{\infty} \frac{\mathrm{cov}\left\{f(X_0), f(X_k)\right\}}{\mathrm{var}\left\{f(X_0)\right\}},$$

is finite. We denote by $\varphi(x; m, \Lambda)$ the normal density of argument x, mean m and covariance Λ.

In order to obtain guidelines to balance computational cost and accuracy of the likelihood estimator Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) make the simplifying assumption that $g(\mathrm{d}z \mid \theta) = \varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)$, that $\sigma^2 \propto 1/N$, and focus on functions

$f \in L^2(\pi)$ such that $f(\theta, z) = f(\theta, z')$ for any z, z'. Under these simplifying assumptions, it was first proposed by Pitt et al. (2012) to minimize

$$\textsc{ct}(f, P_\sigma) = \frac{\tau(f, P_\sigma)}{\sigma^2}, \tag{1}$$

with respect to σ where

$$P_\sigma(\theta, z; \mathrm{d}\theta', \mathrm{d}z') = q(\theta, \mathrm{d}\theta')\varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)\alpha(\theta, z; \theta', z') + \rho_\sigma(\theta, z)\delta_{(\theta,z)}(\mathrm{d}\theta', \mathrm{d}z'), \tag{2}$$

$\rho_\sigma(\theta, z)$ being the corresponding rejection probability. The criterion (1) arises from the fact that the computational time required to evaluate the likelihood is typically proportional to N. Under the additional assumption that $q(\theta, \mathrm{d}\theta') = \pi(\mathrm{d}\theta')$, the minimizer of $\textsc{ct}(f, P_\sigma)$ is $\sigma = 0{\cdot}92$ (Pitt et al., 2012). For more general proposal distributions Doucet et al. (2015) minimize upper bounds on $\textsc{ct}(f, P_\sigma)$. This results in guidelines stating that one should select indeed σ around $1{\cdot}0$ when the Metropolis–Hastings algorithm using the exact likelihood would provide an estimator having a small integrated autocorrelation time and around $1{\cdot}7$ when this integrated autocorrelation time is very large (Doucet et al., 2015). In practical scenarios, the integrated autocorrelation time of the Metropolis–Hastings algorithm using the exact likelihood is unknown as the algorithm cannot be implemented and the results in Doucet et al. (2015) suggest to select σ around $1{\cdot}2$ as a robust default choice. A slightly different approach is taken by Sherlock et al. (2015). In addition to similar noise assumptions, it is assumed that the posterior factorizes into d independent and identically distributed components and that one uses an isotropic normal random walk proposal of jump size proportional to ℓ. In this context, one maximizes with respect to (σ, ℓ) the expected squared jump distance associated to the pseudo-marginal sequence of the first parameter component $(\vartheta_{1,k})_{k\geqslant 0}$ at stationarity divided by the noise variance as $d \to \infty$. In this asymptotic regime, a time-rescaled version of $(\vartheta_{1,k})_{k\geqslant 0}$ converges weakly to a diffusion process and the adequately rescaled expected squared jumping distance converges to the squared diffusion coefficient of this process. In this context, however, maximizing the diffusion coefficient (which also appears in the drift) speeds up the diffusion, decreasing the variance of any Monte Carlo estimate (see, e.g. Roberts & Rosenthal, 2014). Thus, maximizing the scaled expected squared jump distance is asymptotically equivalent to minimizing $\textsc{ct}(f, P_\sigma)$ irrespective of f and its maximizing arguments are $\sigma = 1{\cdot}8$ and $\ell = 2{\cdot}56$ (Sherlock et al., 2015, Corollary 1).

In practice, the standard deviation of the log-likelihood estimator varies over the parameter space and one selects N such that this standard deviation is approximately equal to the desired σ for a parameter value around the mode of the posterior density obtained through a preliminary run.

3. Large Sample Asymptotics of the Pseudo-Marginal Algorithm

3·1. *Notation and Assumptions*

Our analysis of the pseudo-marginal algorithm relies on the assumption that the posterior concentrates (Assumption 1) which is most commonly formulated using convergence in probability with respect to the data distribution, denoted $\mathbb{P}^Y$. For our result to hold under this weak assumption we take into account the randomness induced by the data, resulting in a random Markov chain. This induces some technical difficulties dealing with weak convergence of random probability measures. To make this more precise we introduce the following notation.

The observations $(Y_t)_{t\geqslant 1}$ are regarded as random variables defined on a probability space $\{\mathsf{Y}^{\mathbb{N}}, \mathcal{B}(\mathsf{Y})^{\mathbb{N}}, \mathbb{P}^Y\}$, where $\mathcal{B}(\mathsf{Y})^{\mathbb{N}}$ denotes the Borel σ-algebra and we write $\Omega = \mathsf{Y}^{\mathbb{N}}$ for brevity. For $T \geqslant 1$ we can define the random variables $Y_{1:T} = (Y_1, \ldots, Y_T)$ as the coordinate projections to Y^T. Then, given $\omega = (y_t)_{t\geqslant 1} \in \Omega$, $\pi_T^\omega(\mathrm{d}\theta) = p(\mathrm{d}\theta \mid y_{1:T})$ denotes a regular version of the target posterior distribution and, for any $\theta \in \Theta$, $g_T^\omega\,(\mathrm{d}z \mid \theta)$ the distribution of the error in the log-likelihood estimator for observations $y_{1:T}$. The measures π_T^ω and g_T^ω can be interpreted as random measures. Relevant results for random measures are briefly discussed in Section 4 and in more detail in the supplementary material. In the following we will use a superscript ω to highlight that a certain quantity depends on the data. All probability densities considered hereafter are with respect to the Lebesgue measure and we use the same symbols for distributions and densities, e.g., $\mu\,(\mathrm{d}\theta) = \mu\,(\theta)\,\mathrm{d}\theta$.

In this context, the target distribution of the pseudo-marginal algorithm is

$$\pi_T^\omega(\mathrm{d}\theta, \mathrm{d}z) = \pi_T^\omega(\mathrm{d}\theta)\exp\,(z)\; g_T^\omega\,(\mathrm{d}z \mid \theta)\,,$$

and its transition kernel is

$$P_T^\omega\left(\theta, z; \mathrm{d}\theta', \mathrm{d}z'\right) = q_T(\theta, \mathrm{d}\theta')g_T^\omega(\mathrm{d}z' \mid \theta')\alpha_T^\omega\big(\theta, z; \theta', z'\big) + \rho_T^\omega(\theta, z)\delta_{(\theta,z)}\left(\mathrm{d}\theta', \mathrm{d}z'\right),$$

where

$$\alpha_T^\omega\big(\theta, z; \theta', z'\big) = \min\left\{1, \frac{\pi_T^\omega(\mathrm{d}\theta')}{\pi_T^\omega(\mathrm{d}\theta)}\frac{q_T(\theta', \mathrm{d}\theta)}{q_T(\theta, \mathrm{d}\theta')}\exp\left(z' - z\right)\right\}$$

and $\rho_T^\omega(\theta, z)$ is the corresponding rejection probability.

Our first assumption is that the posterior distributions concentrate towards a normal. We denote by $\mathcal{Y}_T$ the σ-algebra spanned by $Y_{1:T}$.

Assumption 1. The posterior distributions $\{\pi_T^\omega\,(\mathrm{d}\theta)\}_{T\geqslant 1}$ admit densities and there exists a $d \times d$ positive definite matrix Σ, a parameter value $\bar\theta \in \Theta$ and a sequence $(\hat\theta_T^\omega)_{T\geqslant 1}$ of $\mathcal{Y}_T$-adapted random variables such that as $T \to \infty$

$$\int \left|\pi_T^\omega(\theta) - \varphi\big(\theta; \hat\theta_T^\omega, \Sigma/T\big)\right| \mathrm{d}\theta \to 0, \qquad \hat\theta_T^\omega \to, \bar\theta \tag{3}$$

both convergence being in $\mathbb{P}^Y$-probability.

In particular, Assumption 1 is satisfied if a Bernstein-von Mises theorem holds; see, e.g., Van der Vaart (2000, Theorem 10.1) for the classical version or Kleijn & Van der Vaart (2012) for the misspecified case. Under this assumption, the posterior concentrates at rate $1/\sqrt{T}$. Our second assumption is that we use random walk proposal distributions whose increments are appropriately scaled.

Assumption 2. The proposal distributions $\{q_T(\theta, \mathrm{d}\theta')\}_{T\geqslant 1}$ admit densities of the form

$$q_T(\theta, \theta') = T^{1/2}\nu\left\{T^{1/2}(\theta' - \theta)\right\},$$

where ν is a continuous probability density on $\mathbb{R}^d$ with $\int \|\theta\|\,\nu(\mathrm{d}\theta) < \infty$ for the Euclidean norm $\|\cdot\|$.

Finally, we assume that the error in the log-likelihood estimator satisfies a central limit theorem conditional upon $\mathcal{Y}_T$ and that this convergence holds uniformly in a neighbourhood of $\bar\theta$.

Assumption 3. There exists an ε-ball $B(\bar\theta)$ around $\bar\theta$ such that the distributions of the error in the log-likelihood estimator $\left\{g_T^\omega(\mathrm{d}z \mid \theta)\right\}_{T\geqslant 1}$ satisfy as $T \to \infty$

$$\sup_{\theta\in B(\bar\theta)} d_{\mathrm{BL}}\left[g_T^\omega\,(\,\cdot \mid \theta)\,, \varphi\left\{\,\cdot\,; -\sigma^2(\theta)/2, \sigma^2(\theta)\right\}\right] \to 0, \tag{4}$$

in $\mathbb{P}^Y$-probability, where d_{BL} denotes the bounded Lipschitz metric, $\sigma : \Theta \to [0, \infty)$ is continuous at $\bar\theta$ with $0 < \sigma(\bar\theta) < \infty$. An analogous result holds for $\bar g_T^\omega(z \mid \theta) = \exp(z) g_T^\omega(z \mid \theta)$, the distribution of this error at equilibrium, that is as $T \to \infty$

$$\sup_{\theta \in B(\bar\theta)} d_{\mathrm{BL}} \left[\bar g_T^\omega(\,\cdot \mid \theta)\,, \varphi\left\{\cdot; \sigma^2(\theta)/2, \sigma^2(\theta)\right\}\right] \to 0 \tag{5}$$

in $\mathbb{P}^Y$-probability.

We will refer to convergence in probability with respect to the bounded Lipschitz metric as weak convergence in probability. In Section 5, we provide sufficient conditions under which this assumption is satisfied by likelihood estimators obtained through importance sampling for random effects models.

3·2. *Weak Convergence in the Large Sample Regime*

Denote $(\vartheta_{T,k}^\omega, Z_{T,k}^\omega)_{k\geqslant 0}$ the stationary Markov chain defined by the pseudo-marginal kernel, i.e. $(\vartheta_{T,0}^\omega, Z_{T,0}^\omega) \sim \pi_T^\omega$ and $(\vartheta_{T,k}^\omega, Z_{T,k}^\omega) \sim P_T^\omega(\vartheta_{T,k-1}^\omega, Z_{T,k-1}^\omega; \cdot)$ for $k \geqslant 1$. Let $\chi_T^\omega = (\tilde\vartheta_{T,k}^\omega, Z_{T,k}^\omega)_{k\geqslant 0}$ where $\tilde\vartheta_{T,k}^\omega = T^{1/2}(\vartheta_{T,k}^\omega - \hat\theta_T^\omega)$ is the Markov chain arising from rescaling the parameter component of the pseudo-marginal chain. Its transition kernel is thus

$$\tilde P_T^\omega(\tilde\theta, z; \mathrm{d}\tilde\theta', \mathrm{d}z') = \tilde q_T(\tilde\theta, \mathrm{d}\tilde\theta')\tilde g_T^\omega(\mathrm{d}z'|\tilde\theta')\tilde\alpha_T^\omega\big(\tilde\theta, z; \tilde\theta', z'\big) + \tilde\rho_T^\omega(\tilde\theta, z)\delta_{(\tilde\theta,z)}(\mathrm{d}\tilde\theta', \mathrm{d}z'),$$

where

$$\tilde\alpha_T^\omega\big(\tilde\theta, z; \tilde\theta', z'\big) = \min\left\{1, \frac{\tilde\pi_T^\omega(\mathrm{d}\tilde\theta')}{\tilde\pi_T^\omega(\mathrm{d}\tilde\theta)}\frac{\tilde q_T(\tilde\theta', \mathrm{d}\tilde\theta)}{\tilde q_T(\tilde\theta, \mathrm{d}\tilde\theta')}\exp\left(z' - z\right)\right\},$$

$\tilde\rho_T^\omega(\theta, z)$ is the corresponding rejection probability, $\tilde\pi_T^\omega(\tilde\theta) = \pi_T^\omega(\hat\theta_T^\omega + \tilde\theta/T^{1/2})/T^{1/2}$, $\tilde q_T(\tilde\theta, \tilde\theta') = q_T(\hat\theta_T^\omega + \tilde\theta/T^{1/2}, \hat\theta_T^\omega + \tilde\theta'/T^{1/2})/T^{1/2}$ and $\tilde g_T^\omega(z \mid \tilde\theta) = g_T^\omega(z \mid \hat\theta_T^\omega + \tilde\theta/T^{1/2})$. Under Assumption 2, we have

$$\tilde q_T(\tilde\theta, \tilde\theta') = \nu(\tilde\theta' - \tilde\theta) = \tilde q(\tilde\theta, \tilde\theta').$$

We now state the main result of this paper.

THEOREM 1. *Under Assumptions* 1, 2 *and* 3, *the sequence of stationary Markov chains* $(\chi_T^\omega)_{T\geqslant 1}$ *converges weakly in* $\mathbb{P}^Y$*-probability as* $T \to \infty$ *to the law of a stationary Markov chain of initial distribution*

$$\tilde\pi(\mathrm{d}\tilde\theta, \mathrm{d}z) = \varphi(\mathrm{d}\tilde\theta; 0, \Sigma)\varphi\left(\mathrm{d}z; \sigma^2/2, \sigma^2\right)$$

and transition kernel

$$\tilde P(\tilde\theta, z; \mathrm{d}\tilde\theta', \mathrm{d}z') = \tilde q(\tilde\theta, \mathrm{d}\tilde\theta')\varphi\left(\mathrm{d}z'; -\sigma^2/2, \sigma^2\right)\tilde\alpha(\tilde\theta, z; \tilde\theta', z') + \tilde\rho(\tilde\theta, z)\delta_{(\tilde\theta,z)}(\mathrm{d}\tilde\theta', \mathrm{d}z') \tag{6}$$

where $\sigma = \sigma(\bar\theta)$,

$$\tilde\alpha(\tilde\theta, z; \tilde\theta', z') = \min\left\{1, \frac{\varphi(\tilde\theta'; 0, \Sigma)}{\varphi(\tilde\theta; 0, \Sigma)}\frac{\tilde q(\tilde\theta', \tilde\theta)}{\tilde q(\tilde\theta, \tilde\theta')}\exp\left(z' - z\right)\right\},$$

and $\tilde\rho(\theta, z)$ *is the corresponding rejection probability.*

Under this asymptotic regime, the limiting transition kernel (6) is thus a pseudo-marginal transition kernel where the noise distribution is independent of the parameter and given by

$\varphi\left(\mathrm{d}z; -\sigma^2/2, \sigma^2\right)$ as assumed in previous analyses (Pitt et al., 2012; Doucet et al., 2015; Sherlock et al., 2015). For large T, this suggests that some characteristics of the pseudo-marginal kernel can indeed be captured by those of the kernel (2) which can be obtained from (6) by using the change of variables $\theta = \hat{\theta}_T^{\omega} + \tilde{\theta}/T^{1/2}$ and substituting the true target for its normal approximation $\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)$, hence removing a level of approximation.

4. Outline of the Proof of the Main Result

4·1. *Random Markov Chains*

The proof of Theorem 1 follows from a slightly more general result on weak convergence of random Markov chains on Polish spaces given in Theorem 2. We introduce here some notation and recall some definitions concerning random probability measures that we need in order to define random Markov chains. For more details we refer the readers to the supplementary material or Crauel (2003).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and S a Polish space endowed with its Borel σ-algebra $\mathcal{B}(S)$. We equip the product space $\Omega \times S$ with the product σ-algebra $\mathcal{F} \otimes \mathcal{B}(S)$. We denote by $\mathcal{P}(S)$ the space of Borel probability measures which is itself endowed with the Borel σ-algebra $\mathcal{B}\{\mathcal{P}(S)\}$ generated by the weak topology. Finally, $C_b(S)$, respectively $\mathrm{BL}(S)$, denote the sets of continuous bounded functions, respectively the set of bounded Lipschitz functions.

Definition 1 (*Random probability measure*). *A* random probability measure *is a map* $\mu \colon \Omega \times \mathcal{B}(S) \to [0, 1]$, $(\omega, B) \mapsto \mu(\omega, B) = \mu^{\omega}(B)$, *such that for every* $B \in \mathcal{B}(S)$ *the map* $\omega \mapsto \mu(\omega, B)$ *is measurable while* $\mu^{\omega} \in \mathcal{P}(S)$ $\mathbb{P}$*—almost surely.*

For all bounded and measurable functions $g \colon \Omega \times S \to \mathbb{R}$, $\omega \mapsto \int_S g(\omega, x)\mu^{\omega}(\mathrm{d}x)$ is measurable (Crauel, 2003, Proposition 3.3) and thus the map $\omega \mapsto \mu^{\omega}(f)$ is a random variable for bounded measurable functions $f \colon S \to \mathbb{R}$. Consequently, $\mu^{\omega} \colon \Omega \to \mathcal{P}(S)$ is a Borel measurable map. Conversely, it can be shown that any random element of $\{\mathcal{P}(S), \mathcal{B}(\mathcal{P}(S)\}$ fulfils the conditions set out in Definition 1; see Crauel (2003, Remark 3.20 (i)) or Kallenberg (2006, Lemma 1.37) for details.

Definition 2 (*Random Markov kernel*). *A* random Markov kernel *is a map* $K \colon \Omega \times S \times \mathcal{B}(S) \to [0, 1]$, $(\omega, x, B) \mapsto K(\omega, x, B) = K^{\omega}(x, B)$, *such that*

(i) $(\omega, x) \mapsto K^{\omega}(x, B)$ is $\mathcal{F} \otimes \mathcal{B}(S)$-measurable for every $B \in \mathcal{B}(S)$,
(ii) $K^{\omega}(x, \cdot) \in \mathcal{P}(S)$ $\mathbb{P}$—almost surely for every $x \in S$.

Lemma 1 (*Random Markov chain*). Given a random probability measure μ^{ω} and random Markov kernel K^{ω}, there exists a (almost surely) unique random probability measure $\mu^{\mathbb{N},\omega}$ on $S^{\mathbb{N}}$ such that

$$\mu^{\mathbb{N},\omega}(A_1 \times \ldots \times A_k \times E_{k+1}) = \int_{A_1} \mu^{\omega}(\mathrm{d}x_1) \int_{A_2} K^{\omega}(x_1, \mathrm{d}x_2) \ldots \int_{A_k} K^{\omega}(x_{k-1}, \mathrm{d}x_k)$$

for any $A_i \in \mathcal{B}(S)$ $(i = 1, \ldots, k)$, $k \in \mathbb{N}$ *and* $E_{k+1} = \times_{i=k+1}^{\infty} S$.

4·2. *Convergence of Random Markov Chains*

For a sequence of random probability measures $(\mu_n^{\omega})_{n \geqslant 1}$, respectively a sequence of random Markov kernels $(K_n^{\omega})_{n \geqslant 1}$, converging in a suitable sense towards a probability measure μ, respectively a Markov kernel K, we show here that the distributions of the associated Markov chains $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ defined in Lemma 1 converge weakly in probability to the distribution $\mu^{\mathbb{N}}$ of the homogeneous Markov chain of initial distribution μ and Markov kernel K.

THEOREM 2 (WEAK CONVERGENCE OF RANDOM MARKOV CHAINS). *If the following assumptions hold,*

(T.1) the random probability measures $\left(\mu_n^{\omega}\right)_{n\geqslant 1}$ converge weakly in probability to a probability measure μ as $n \to \infty$,

(T.2) the random Markov transition kernels $\left(K_n^{\omega}\right)_{n\geqslant 1}$ satisfy

$$\int \left|K_n^{\omega} f(x) - Kf(x)\right| \mu_n^{\omega}(\mathrm{d}x) \to 0$$

in probability as $n \to \infty$ for all $f \in \mathrm{BL}(S)$ where K is a Markov transition kernel ,

(T.3) the transition kernel K is such that $x \mapsto Kf(x)$ is continuous for any $f \in C_b(S)$,

then, as $n \to \infty$, the measures $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ on $S^{\mathbb{N}}$ converge weakly in probability to the measure $\mu^{\mathbb{N}}$ induced by the Markov chain with initial distribution μ and transition kernel K.

4·3. *Application to the Pseudo-Marginal Algorithm*

Theorem 1 follows from Theorem 2 by noting that under Assumptions 1, 2 and 3 all conditions set out in Theorem 2 are fulfilled. Firstly, as we increase the number of data points, the stationary distribution of the Markov chain will converge weakly to the limiting stationary distribution of Theorem 2.

PROPOSITION 1. *Under Assumptions* 1 *and* 3*, we have*

$$\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) \to \tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z),$$

weakly in $\mathbb{P}^Y$-probability as $T \to \infty$ where $\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) = \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})\exp(z)\, \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})$.

This follows as the marginal $\pi_T^{\omega}(\mathrm{d}\theta)$ concentrates around the limiting parameter value $\bar{\theta}$ while the noise uniformly converges towards a normal distribution in a neighbourhood around $\bar{\theta}$. The next proposition ensures the stability of the transition and can be proven using similar arguments.

PROPOSITION 2. *Under Assumptions* 1, 2 *and* 3*, we have for any $f \in \mathrm{BL}(\mathbb{R}^{d+1})$*

$$\int |\tilde{P}_T^{\omega} f(\theta, z) - \tilde{P} f(\theta, z)| \tilde{\pi}_T^{\omega}(\mathrm{d}\theta, \mathrm{d}z) \to 0$$

in $\mathbb{P}^Y$-probability as $T \to \infty$.

A further requirement to ensure the stability of the transition is that the the application of the transition operator conserves continuity.

PROPOSITION 3. *Under Assumption* 2*, the map $(\theta, z) \mapsto \tilde{P} f(\theta, z)$ is continuous for every $f \in C_b(\mathbb{R}^{d+1})$.*

Theorem 1 now follows from a direct application of Theorem 2 as the assumptions *(T.1)*, *(T.2)* and *(T.3)* hold by Proposition 1, 2 and 3, respectively.

5. Random effects models

5·1. *Statistical Model and Likelihood Estimator*

We provide here sufficient conditions under which weak convergence of the pseudo-marginal algorithm is verified for an important class of latent variable models. Consider the model

$$X_t \sim f(\cdot \mid \theta), \qquad Y_t \mid X_t \sim g(\cdot \mid X_t, \theta), \tag{7}$$

where $(X_t)_{t\geqslant 1}$ are independent $\mathbb{R}^k$-valued latent variables, $f(x \mid \theta)$ is a density with respect to Lebesgue measure and $(Y_t)_{t\geqslant 1}$ are Y-valued observations distributed according to a conditional density $g(y \mid x, \theta)$ with respect to a dominating measure, Y being a topological space. For observations $Y_{1:T} = y_{1:T}$ the likelihood is

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} p(y_t \mid \theta) = \prod_{t=1}^{T} \int g(y_t \mid x_t, \theta) f(x_t \mid \theta) \mathrm{d}x_t.$$

In many practical scenarios, the likelihood is not available analytically. If one wants to perform Bayesian inference about the parameter θ in this context, we can use the pseudo-marginal algorithm as it is possible to obtain an unbiased non-negative estimator of $p(y_{1:T} \mid \theta)$ using importance sampling. Indeed, we can consider $\hat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \hat{p}(y_t \mid \theta, U_t)$ where $U = (U_1, ..., U_T)$, $U_t = \left(U_{t,1}, ..., U_{t,N}\right)$, $U_{t,i}$ is $\mathbb{R}^k$-valued, N denotes the number of Monte Carlo samples and $\hat{p}(y_t \mid \theta, U_t)$ is an importance sampling estimator of $p(y_t \mid \theta)$ given by

$$\hat{p}(y_t \mid \theta, U_t) = \frac{1}{N} \sum_{i=1}^{N} \varpi(y_t, U_{t,i}, \theta), \qquad \varpi(y_t, U_{t,i}, \theta) = \frac{g(y_t \mid U_{t,i}, \theta) f(U_{t,i} \mid \theta)}{h(U_{t,i} \mid y_t, \theta)},$$

where $U_{t,i} \sim h(\cdot \mid y_t, \theta)$, $h(u \mid y_t, \theta)$ being a probability density on $\mathbb{R}^k$ with respect to Lebesgue measure. In this case the joint density of all the auxiliary variates used to obtain the likelihood estimator is

$$m_{T,\theta}(u) = \prod_{t=1}^{T} \prod_{i=1}^{N} h(u_{t,i} \mid y_t, \theta).$$

We will assume subsequently that the true observations are independent and identically distributed samples taken from some unspecified probability measure μ. The joint data distribution is then just the product measure $\mathbb{P}^Y(\mathrm{d}\omega) = \prod_{t=1}^{\infty} \mu(\mathrm{d}y_t)$.

5·2. *Verifying the assumptions*

The Bernstein–von Mises theorem holds under weak regularity assumptions; see, e.g., Van der Vaart (2000, Theorem 10.1). This ensures Assumption 1 is satisfied while Assumption 2 is easy to satisfy, e.g., select a multivariate normal proposal of covariance scaling as $1/\sqrt{T}$. Assumption 3 is more complicated as it requires to establish uniform conditional central limit theorems for $\hat{p}(Y_{1:T} \mid \theta, U)$ in scenarios where $U \sim m_{T,\theta}$ arise from the proposal, so $Z \sim g_T^{\omega}(\cdot \mid \theta)$, or at stationarity where $U \sim \pi_T^{\omega}(\cdot \mid \theta)$ with

$$\pi_T^{\omega}(u \mid \theta) = \frac{\hat{p}(y_{1:T} \mid \theta, u)}{p(y_{1:T} \mid \theta)} m_{T,\theta}(u),$$

implying that $Z \sim \bar{g}_T^{\omega}(\cdot \mid \theta)$. We denote

$$\sigma^2(y, \theta) = E\left\{\epsilon_T(y, \theta)^2\right\} = \mathrm{var}\left\{\overline{w}(y, U_{1,1}, \theta)\right\}, \quad \sigma^2(\theta) = E\left\{\sigma^2(Y_1, \theta)\right\},$$

with $U_{1,1} \sim h(\cdot \mid y, \theta)$, $Y_1 \sim \mu$ and

$$\overline{w}(Y_t, U_{t,i}, \theta) = \frac{w(Y_t, U_{t,i}, \theta)}{p(Y_t \mid \theta)}. \tag{8}$$

We make the following assumption.

Assumption 4. There exists a closed ε-ball $B(\bar{\theta})$ around $\bar{\theta}$ and a function g such that the normalized weight $\overline{w}(y, U_{1,1}, \theta)$ defined in (8) satisfies

$$\sup_{\theta \in B(\bar{\theta})} E\left\{\left|\overline{w}(y, U_{1,1}, \theta)\right|^7\right\} \leq g(y), \tag{9}$$

where $U_{1,1} \sim h(\cdot \mid y, \theta)$ and $\mu(g) < \infty$. Additionally $\theta \mapsto \sigma^2(y, \theta)$ is continuous in θ on $B(\bar{\theta})$ for all $y \in \mathsf{Y}$. Further, there exist real constants $\delta, C, a > 0$ such that for all $t \in [0, a)$

$$\sup_{\theta \in B(\bar{\theta}), y \in \mathsf{Y}} \mathbb{P}\{\overline{w}(y, U_{1,1}, \theta) \leq t\} \leq C t^{1+\delta}.$$

Under these conditions, we obtain the following uniform version of the central limit theorem for the error in the log-likelihood estimator.

THEOREM 3 (UNIFORM CENTRAL LIMIT THEOREM). *Under Assumption* 4, *Assumption* 3 *is satisfied.*

Theorem 3 strengthens earlier results of Deligiannidis et al. (2015, Theorem 1) which obtain standard central limit theorems for the error in the log-likelihood estimator.

6. OPTIMIZATION OF THE PSEUDO-MARGINAL RANDOM WALK ALGORITHM

6·1. *Optimization Problem*

We propose to optimize the performance of the limiting pseudo-marginal chain identified in Theorem 1 as a proxy for the optimization of the original pseudo-marginal chain. We assume that the limiting covariance matrix Σ in (3) is the identity matrix I_d with d denoting the parameter dimension. For more general covariance matrices, we can use a Cholesky decomposition and a change of variables as in Sherlock et al. (2015); Nemeth et al. (2016). We denote by $\tilde{P}_{\ell,\sigma}$ the transition kernel (6) using the proposal density

$$q(\theta, \theta') = \varphi\left(\theta'; \theta, \ell^2 I_d/d\right).$$

As Pitt et al. (2012) and Doucet et al. (2015), we propose to minimize

$$\text{CT}(f, \tilde{P}_{\ell,\sigma}) = \frac{\tau(f, \tilde{P}_{\ell,\sigma})}{\sigma^2}$$

with respect to the noise standard deviation σ but, contrary to Pitt et al. (2012) and Doucet et al. (2015), also with the scale parameter ℓ. We restrict attention here to the case where $f(\theta, z) = \theta_1$, the first component of θ, and write $\text{CT}(f, \tilde{P}_{\ell,\sigma}) = \text{CT}(\ell, \sigma)$ in this case. As this criterion is not available in closed-form, we simulate the limiting Markov chain initialized in its stationary regime with different noise levels σ and different values of ℓ on a fine grid to obtain empirical estimates of $\text{CT}(f, \tilde{P}_{\ell,\sigma})$ computed using the overlapping batch mean estimator. Other estimators did not provide significantly different results. This simulation is straightforward as the target and noise distributions in this limiting case are both Gaussian. We then find the approximate minimizer $(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$ of $\text{CT}(f, \tilde{P}_{\ell,\sigma})$ over this grid. This set-up is applied for selected scenarios with parameter dimension d ranging from 1 to 50.

Table 1. *Optimal values for scaling ℓ and noise σ and associated value of computing time and average acceptance probability. All simulations with* 10 *repetitions. We report the mean of the minimizers as well as the standard deviation over the 10 repetitions.*

Dimension d	$\hat{\ell}_{\mathrm{opt}}$	$\hat{\sigma}_{\mathrm{opt}}$	$\mathrm{CT}(\hat{\sigma}_{\mathrm{opt}}, \hat{\ell}_{\mathrm{opt}})$	$\mathrm{pr}_{\mathrm{acc}}(\hat{\sigma}_{\mathrm{opt}}, \hat{\ell}_{\mathrm{opt}})$
$d = 1$	2·05 (0·25)	1·16 (0·07)	8·47	25·73%
$d = 2$	1·97 (0·14)	1·21 (0·06)	12·71	22·92%
$d = 3$	2·11 (0·07)	1·24 (0·05)	16·79	19·97%
$d = 5$	2·17 (0·12)	1·30 (0·05)	23·18	17·35%
$d = 10$	2·20 (0·08)	1·44 (0·05)	37·93	14·27%
$d = 15$	2·33 (0·08)	1·50 (0·00)	53·43	12·07%
$d = 20$	2·34 (0·10)	1·54 (0·05)	65·62	11·44%
$d = 30$	2·36 (0·11)	1·61 (0·03)	90·46	10·41%
$d = 50$	2·41 (0·10)	1·74 (0·05)	136·38	8·66%

Table 2. *Comparison of the computing time for different noise levels. $\hat{\sigma}_{\mathrm{opt}}$ denotes the minimizer of the estimated integrated autocorrelation time, as shown in Table 1.*

Dimension d	$\mathrm{CT}(\hat{\sigma}_{\mathrm{opt}}, \ell_{\infty})$	$\mathrm{CT}(\sigma = 1.2, \ell_{\infty})$	$\mathrm{CT}(\sigma_{\infty}, \ell_{\infty})$
$d = 1$	9·04 (0·25)	9·05 (0·21)	17·10 (1·34)
$d = 2$	13·48 (0·32)	13·37 (0·28)	22·45 (0·81)
$d = 3$	17·63 (0·28)	17·43 (0·26)	26·71 (0·64)
$d = 5$	24·38 (0·44)	24·72 (0·31)	34·14 (0·88)
$d = 10$	40·17 (0·71)	41·60 (0·24)	47·08 (1·03)
$d = 15$	53·69 (0·72)	58·01 (0·50)	59·08 (0·79)
$d = 20$	67·15 (0·53)	74·34 (0·36)	71·41 (1·48)
$d = 30$	91·36 (0·95)	106·08 (0·34)	93·73 (1·08)
$d = 50$	136·49 (1·18)	167·83 (0·75)	135·92 (1·27)

6·2. *Numerical Results*

The simulation results are collected in Table 1. In addition to $(\hat{\ell}_{\mathrm{opt}}, \hat{\sigma}_{\mathrm{opt}})$, we give also the computing time at these values as well as the average acceptance probability of the proposal under $\tilde{P}_{\ell,\sigma}$ at stationary using a chain length of K equal to 5 million. The results we obtain are consistent with those in Doucet et al. (2015) and Sherlock et al. (2015). For low dimensions, $1 \leq d \leq 5$, the ideal Metropolis–Hastings algorithm mixes well and $\hat{\sigma}_{\mathrm{opt}}$ is around 1·1-1·3 as suggested by Doucet et al. (2015) and it increases slowly as d increases to the values $(\sigma_{\infty}, \ell_{\infty}) = (1{\cdot}81, 2{\cdot}56)$ obtained by the diffusion limit (Sherlock et al., 2015). For example, for $d = 50$, we obtain $(\hat{\sigma}_{\mathrm{opt}}, \hat{\ell}_{\mathrm{opt}}) = (1{\cdot}74, 2{\cdot}41)$ and the resulting optimal computing time $\mathrm{CT}(\hat{\sigma}_{\mathrm{opt}}, \hat{\ell}_{\mathrm{opt}})$ is close to $\mathrm{CT}(\sigma_{\infty}, \ell_{\infty})$. For lower dimensions, however, the performance in terms of computing time can be increased by reducing the noise of the estimator and the proposed jumping distance in comparison to ℓ_{∞} and σ_{∞}; see Table 2. We also observed empirically that the cost function $\ell \mapsto \mathrm{CT}(\ell, \sigma)$ is fairly flat as noticed in the limiting case by Sherlock et al. (2015).

7. Simulation study: Random Effects Model

We now illustrate how the guidelines derived from the limiting pseudo-marginal chain compare to a practical implementation of pseudo-marginal Metropolis-Hastings. We consider a Bayesian logistic mixed effects model applied to a real data set. Mixed models are popular in econometrics, survey analysis and medical statistics amongst others and are often used to describe heterogeneity between groups. Here we consider a subset of a cohort study of Indonesian preschool children. This data was previously analyzed using Bayesian mixed models by Zeger & Karim (1991). Overall, the dataset contains 1200 observations of 275 children. We model the probability of a respiratory infection based on the following covariates: age, sex, height, an indicator for presence of vitamin deficiency, an indicator for subnormal height and two seasonal components. Including the intercept we have an overall of 8 covariates. Cluster effects due to repeated measurements of the same children are modelled with individual random intercepts. In this case the linear predictor of a regression model based on covariates $c_{t,j}$ $(t = 1, \dots, T, j = 1, \dots J)$ reads

$$\eta_{t,j} = c_{t,j}^{\mathrm{T}} \beta + X_t, \quad X_t \sim \mathcal{N}(0, \tau),$$

where X_t denotes the random intercept for children $t = 1, \dots, T$ and β the regression parameters. For every child we have an observation vector $y_t = \left(y_{t,1} \dots, y_{t,J}\right) \in \{0, 1\}^J$. The observations are assumed conditionally independent given the random effects and are modelled through

$$p(y_{1:T} \mid \beta, \tau, x_{1:T}) = \prod_{t=1}^{T} \prod_{j=1}^{J} \frac{\exp(y_{t,j} \eta_{t,j})}{1 + \exp(\eta_{t,j})}.$$

Inference in mixed effects models often aims at finding the population effects and thus one is interested in integrating out the random effects. Since the marginal likelihood contains intractable integrals, this model lends itself to the pseudo-marginal approach. We obtain an unbiased estimator of the marginal likelihood by estimating the integrals using the prior distribution of the random effects as importance distribution. For the covariate parameters we assume a diffuse Gaussian prior and the variance of the random effects are assigned an inverse gamma prior. The unknown parameter is $\theta = (\beta, \tau) \in \mathbb{R}^d$ where $d = 9$. We run a pseudo-marginal algorithm with a Gaussian random walk proposal for one million iterations. The covariance of the proposal is set equal to the covariance matrix of the parameters estimated in a preliminary run and scaled by $\ell / \surd d = 2{\cdot}2 / \surd 9$. We compare the integrated autocorrelation time and the acceptance rate with that of the limiting chain using the same $\ell = 2{\cdot}2$ and $\sigma = \hat{\sigma}$. Here, $\hat{\sigma}$ is the standard deviation of the log-likelihood estimator obtained using 10000 samples of the marginal likelihood evaluated at $\bar{\theta} = (\bar{\beta}, \bar{\tau})$ also estimated in a preliminary run. The results are summarized in Table 3. For a given number of particles N we report the associated estimate of the noise in the log-likelihood estimator, the integrated autocorrelation time averaged over the 9 dimensions and the average acceptance rate.

We find that the integrated autocorrelation time and the acceptance rate are very close to the respective values of the limiting algorithm. This is visualized in Figure 1 where we plot acceptance rate and integrated autocorrelation time of both algorithms against each other. The computing time of the pseudo-marginal algorithm targeting the exact posterior $\mathrm{CT}(\theta_1, P_T^{\omega}) = N \tau(\theta_1, P_T^{\omega})$ and the computing time of the limiting algorithm $\mathrm{CT}(\theta_1, \tilde{P}_{\ell,\sigma})$ are both optimized for $N = 45$ particles or $\hat{\sigma} = 1{\cdot}42$, respectively, as we would expect from Table 1. This demonstrates that the limiting kernel captures well the behaviour of the pseudo-marginal kernel for large data sets. In these scenarios, Table 1 thus provides useful dimension dependent guidelines on how to tune the pseudo-marginal kernel. We further illustrate the relevance of these guidelines for another example in the supplementary material.

Table 3. *For N particles: standard deviation $\hat{\sigma}$ of the log-likelihood estimator at the mean, average integrated autocorrelation time $\hat{\tau}$ and average acceptance probability* $\hat{\mathrm{pr}}_{\mathrm{acc}}$ *for pseudo-marginal kernel with* $\ell = 2{\cdot}2$ *and limiting kernel* $\tilde{P}_{\ell=2\cdot2,\hat{\sigma}}$. *The row associated with the minimum values for the computing time is highlighted.*

Particles N	$\hat{\sigma}$	$\hat{\tau}$	$\hat{\mathrm{pr}}_{\mathrm{acc}}$	$\hat{\tau}\left(\tilde{P}_{\ell=2\cdot2,\sigma=\hat{\sigma}}\right)$	$\hat{\mathrm{pr}}_{\mathrm{acc}}\left(\tilde{P}_{\ell=2\cdot2,\sigma=\hat{\sigma}}\right)$
25	1·90	141·32	9·09%	151·18	8·68%
30	1·73	117·51	10·68%	105·43	10·65%
35	1·60	92·57	11·89%	91·05	12·18%
40	1·52	81·07	13·20%	83·82	13·01%
45	**1·42**	**71·95**	**14·29%**	**69·50**	**14·59%**
50	1·35	67·93	15·08%	64·48	15·45%
55	1·30	62·72	15·85%	58·84	16·33%
60	1·24	56·91	16·66%	55·98	16·33%
65	1·19	55·41	17·32%	50·71	17·94%

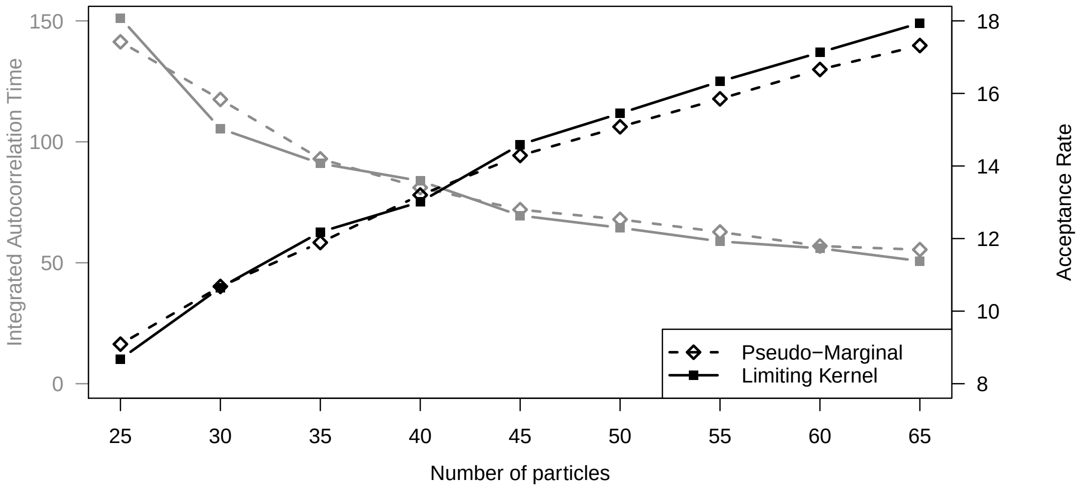

Fig. 1. For N particles: integrated autocorrelation time and acceptance rate for the pseudo-marginal algorithm with $\ell = 2{\cdot}2$ and limiting kernel $\tilde{P}_{\ell=2\cdot2,\hat{\sigma}}$

Acknowledgement

Sebastian M. Schmon's research is supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/K503113/1 and Arnaud Doucet's research is partially supported by the EPSRC grant EP/K000276/1.

Supplementary material

Supplementary material available at *Biometrika* online includes the proofs to all propositions and theorems as well as a short review of weak convergence of random measures and some further simulation studies, including a 3-dimensional Lotka-Volterra model.

References

Andrieu, C., Doucet, A. & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with Discussion). *J. R. Statist. Soc.* B **72**, 269–342.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.

Andrieu, C. & Vihola, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.* **25**, 1030–1077.

Andrieu, C. & Vihola, M. (2016). Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Probab.* **26**, 2661–2696.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–1160.

Bérard, J., Del Moral, P. & Doucet, A. (2014). A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.* **19**, 1–28.

Crauel, H. (2003). *Random Probability Measures on Polish Spaces*. CRC press.

Deligiannidis, G., Doucet, A. & Pitt, M. K. (2015). The correlated pseudo-marginal method. *J. R. Statist. Soc.* B, *to appear - arXiv preprint arXiv:1511.04992* .

Doucet, A., Pitt, M. K., Deligiannidis, G. & Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer-Verlag: New York.

Kleijn, B. J. K. & Van der Vaart, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Statist.* **6**, 354–381.

Lin, L., Liu, K. & Sloan, J. (2000). A noisy Monte Carlo algorithm. *Phys. Rev.* D **61**, 074505.

Nemeth, C., Sherlock, C. & Fearnhead, P. (2016). Particle Metropolis-adjusted Langevin algorithms. *Biometrika* **103**, 701–717.

Pitt, M. K., dos Santos Silva, R., Giordani, P. & Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics* **171**, 134–151.

Roberts, G., Gelman, A. & Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Roberts, G. O. & Rosenthal, J. S. (2014). Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *Ann. Appl. Probab.* **24**, 131**–**149.

Sherlock, C., Thiery, A. H., Roberts, G. O., Rosenthal, J. S. et al. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–275.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Zeger, S. L. & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. R. Statist. Soc.* B .

[*Received* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]