The Annals of Applied Probability
2018, Vol. 28, No. 5, 2966–3001
https://doi.org/10.1214/18-AAP1380

ASYMPTOTIC ANALYSIS OF THE RANDOM WALK METROPOLIS ALGORITHM ON RIDGED DENSITIES

BY ALEXANDROS BESKOS[*,1], GARETH ROBERTS[†], ALEXANDRE THIERY[‡] AND NATESH PILLAI[§,2]

University College London[*], University of Warwick[†], National University of Singapore[‡] and Harvard University[§]

We study the asymptotic behaviour of the Random Walk Metropolis algorithm on "ridged" probability densities where most of the probability mass is distributed along some key directions. Such class of probability measures arise in various applied contexts including for instance Bayesian inverse problems where the posterior measure concentrates on a manifold when the noise variance goes to zero. When the target measure concentrates on a linear manifold, we derive analytically a diffusion limit for the Random Walk Metropolis Markov chain as the scale parameter goes to zero. In contrast to the existing works on scaling limits, our limiting stochastic differential equation does *not* in general have a constant diffusion coefficient. Our results show that in some cases, the usual practice of adapting the step-size to control the acceptance probability might be sub-optimal as the optimal acceptance probability is zero (in the limit).

1. Introduction. Optimal scaling of Metropolis–Hastings (MH) algorithms in high dimensions and analysis of their asymptotic behaviour has been a fruitful area of research in the last three decades. Initiated by [16], a long list of papers has been devoted to deriving the optimal scale of various local-move Markov Chain Monte-Carlo (MCMC) algorithms [2, 3, 5, 13, 17, 18]. A main result in most of these works is identifying the proper scale for the proposal distribution at which the average acceptance probability is nontrivial for increasing dimension, and obtaining an associated diffusion limit. The limiting Stochastic Differential Equation (SDE) in all of the earlier works has a constant diffusion coefficient which uniquely characterizes its "speed measure", with the latter being controlled by the step-size of the local-proposal and the average acceptance probability. For the Random-Walk Metropolis algorithm (RWM), maximizing the speed of the limiting diffusion leads to an average acceptance probability of 0.234; this was first proven

Received December 2016; revised September 2017.
[1]Supported by a Leverhulme Trust Prize and AT from a Start-Up Grant from the National University of Singapore.
[2]Supported in part by an ONR grant.
MSC2010 subject classifications. Primary 65C05, 65C40; secondary 65C30.
Key words and phrases. Manifold, random-walk metropolis, generator, diffusion limit.

in [16] for i.i.d. targets of the type $\prod_i f(x_i)$ for a one-dimensional density function $f : \mathbb{R} \to [0, \infty)$, under the assumptions of f being twice continuously differentiable, f'/f being Lipschitz continuous and $\mathbb{E}[(f'/f)^8(X)]$, $\mathbb{E}[(f''/f)^4(X))]$, with $X \sim f$, being finite. Thus the theoretical asymptotic analysis provides a simple optimality criterion that can be used as a guideline for practical implementations of these algorithms. Most of the papers cited above assume that the target measure is a product of one-dimensional densities as this facilitates explicit calculation of the diffusion limit for a single coordinate, with dynamics that are independent of the state of other components. However, the product form assumption is not essential, and there are now a number of recent generalisations to target measures that are not of a product form [4, 12, 14].

We adopt a different point of view in this paper, and study the behaviour of the RWM algorithm on a class of target distributions that have "ridges" in certain directions. Such target distributions arise in a number of examples in applied contexts. For instance, in Bayesian inverse problems, when the variance parameter in the likelihood model is small, the posterior distribution (as parametrized by the noise variance) will concentrate on a manifold or along a hyperplane as the noise variance goes to zero. In contrast to the works cited above on scaling, the parameter space in our problems needs not be high dimensional. The key issue pertaining to the mixing of MCMC algorithms in these contexts is that there is a natural separation of scale in the target distribution: a larger scale along certain key directions of interest where most of the probability mass is distributed and a smaller scale in the "orthogonal directions" where there is relatively little mass. Thus in the context of previous works cited above, the role of the dimension is played by this scale. Our contributions are summarised as follows:

(i) Motivated by inverse problems, we first describe a natural class of target distributions which have two scales [of magnitudes $\mathcal{O}(1)$ and $\varepsilon \ll 1$]. Next, we study the RWM with step-sizes of the same scale for all coordinates; indeed, in most situations, it is not possible to adapt the covariance structure of the RWM jumps in a sensible way and adopting isotropic jumps is the only viable solution. This remark will be expanded in next section when we give a more precise description of the algorithm. We focus on the case where the target distribution concentrates on a linear hyperplane as $\varepsilon \to 0$. Adapting the RWM steps in the direction of smaller scale ε, we derive a diffusion limit for the RWM for the coordinates with $\mathcal{O}(1)$ scale. In contrast with all previous results on diffusion limits of MCMC algorithms, our limiting SDE will in general have nonadditive noise, that is, the diffusion term will not have constant coefficients. We also look at the case when the step-size is allowed to vary according to the local curvature and obtain a corresponding diffusion limit. We show that diffusion limits can be useful in certain situations for providing optimality criteria in these problems. We mention that in the case of a linear manifold, the user can easily adapt the different RWM step-sizes along the various directions to obtain a very effective algorithm of cost $\mathcal{O}(1)$.

The real practical interest lies with nonlinear structures. We have focused on a linear manifold to facilitate the development of theory. We also provide a conjecture for the diffusion limit in the practical case of a family of nonlinear manifolds, but analytical derivations will require considerable future work.

(ii) When the dimension of the manifold where the target measure concentrates lags that of the entire state space only by one, we find that the usual practice of adapting the step-size to control the acceptance probability is *not* optimal. In particular, our diffusion limits imply that a more efficient chain can be obtained by keeping a large step-size and allowing the acceptance probability to drop to zero. In this case, we show that as $\varepsilon \to 0$, the optimally-scaled RWM algorithm does *not* converge weakly to a diffusion; rather it converges to a continuous-time Markov jump process.

In general, practitioners should thus be aware that, when tuning the standard RWM algorithm, the general strategy consisting in choosing the largest jump size that yields an acceptance rate bounded away from a predetermined threshold (*e.g.*, acceptance rate larger than 25%) is sometimes sub-optimal; indeed, our analysis shows that for "multiscale densities", optimal strategies can yield arbitrarily small acceptance rates.

(iii) In contrast, when the dimension of the manifold where the target measure concentrates lags that of the entire state space by at least 3, we find that the diffusion regime is optimal. Intuitively, the cost of attempting large moves (as measured by small acceptance probability) is just too large in this case. In the critical case when the dimension of the manifold where the target measure concentrates lags that of the entire state space by exactly 2, the cost of low acceptance probability is of the same order as the benefit of larger moves, so that the optimal scaling strategy can vary depending on the specific form of the target density.

(iv) As mentioned before, the most important practical benefit from deriving a diffusion limit is that it leads to an automated choice of the step-size: choose the step-size that maximizes the diffusion coefficient of the limiting SDE [16]. Doing so maximizes the speed measure of the limiting diffusion which also translates to minimizing its integrated autocorrelation time for a large class of test functions; see, for example, [17]. Since our limiting diffusions generally do not have a constant diffusion coefficient, this approach of maximizing the diffusion coefficient for choosing the step size is not valid any more. In general, there is no optimal value for the jump size of the RWM algorithm when applied to the target measures analyzed in this paper. We prove nevertheless that, if the jump size is allowed to be position dependent, optimality results can then be recovered. When the dimension of the coordinates with scale ε is large, our results yield a "local" 0.234 rule that generalize the standard global 0.234 of [16], that is, it is asymptotically optimal to tune the local jump size so that the local acceptance rate equals 0.234.

(v) Technically, our proofs for the diffusion limits are different from the usual optimal scaling literature, as the scale parameter in our target measure needs not be related to the dimension. Our main argument relies on the fact that by suitably

rescaling the coordinates corresponding to the scale ε, the RWM algorithm mixes in these directions at a much faster rate than its $\mathcal{O}(1)$ counterpart. This fast mixing on the Markov chain in the smaller scale will lead to an "asymptotic decoupling" of the two scales which then will give the diffusion limit on the larger scales. The proof technique developed here could be applicable to other contexts and is thus of independent interest.

The class of target measures we analyze in this paper should be considered as surrogates for distributions arising from real applied problems where components have various length scales that differ by orders of magnitude. We can consider the following generic model structure:

$$y = \mathcal{T}(\theta) + \xi; \qquad \mathcal{T} : \mathbb{R}^n \to \mathbb{R}^{n'}, \qquad n, n' \geq 1, \tag{1}$$

with $y \in \mathbb{R}^{n'}$ corresponding to data, $\theta \in \mathbb{R}^n$ to a vector of unknown parameters and $\xi \in \mathbb{R}^{n'}$ to the noise, such that along a number of directions the noise is small (in the context or our paper, it is of standard deviation ε with $\varepsilon \to 0$). The objective is to learn about the posterior of θ, and a main algorithmic challenge is that in the above setting the distribution of interest will concentrate on a manifold (for typical choices of $\mathcal{T}$) as $\varepsilon \to 0$. Almost nothing is known about the behaviour of MCMC algorithms in such a general context, to the best of our knowledge, and our work aims to provide some explicit analysis in this setting.

We provide MCMC results from two simple, yet illustrative, examples to motivate our investigations. Figure 1 shows RWM runs over the posterior of (θ_1, θ_2) given datum $y \sim \mathrm{N}(\theta_1^2 + \theta_2^2, \varepsilon^2)$ with $\varepsilon = 0.01$, and prior $(\theta_1, \theta_2) \sim \mathrm{N}(0, I_2)$; true parameter values were $\theta_1^\dagger = \theta_2^\dagger = 3$. The left (resp., right) panel shows the RWM scatterplot when scaling the step-size to achieve average acceptance probability of 14.8% (resp., 0.16%). Figure 2 considers the same toy-example, for several choices of ε. Here, we compared the Integrated Autocorrelation Time (IACT) for two strategies. For the first strategy, we chose a RWM covariance jump structure equal to $\varepsilon^2 I_2$. This ensures that the acceptance rate stays bounded away from zero as $\varepsilon \to 0$, as described by the theory developed in this article; we observe empirically that this strategy yields an acceptance rate of $\approx 50\%$ as $\varepsilon \to 0$. For the second strategy, we let the covariance matrix of the RWM jumps be equal to the identity, irrespectively of the value of ε. Indeed, and as described by our theory, the acceptance rate shrinks to zero as $\varepsilon \to 0$; still, as demonstrated in this text and empirically observed in Figure 2, this strategy outperforms the first one by several orders of magnitude as $\varepsilon \to 0$. At Figure 3 we consider the ODE model $x_t' = \theta_1 + \theta_2 x_t + 2\sin(\theta_3 x_t)$, $x_0 = \frac{1}{2}$, with datum $y \sim \mathrm{N}(x_{25}, 0.01^2)$ and priors as given at the figure caption; the true values were $\theta^\dagger = (2.2, 0.004, 0.5)$. The red, blue and green dots show RWM runs corresponding to average acceptance probabilities 17%, 0.57% and 0.03% respectively; the shaded surface shows the 2D-manifold $x_{25}(\theta) = x_{25}(\theta^\dagger)$. In both above cases, acceptance probabilities very

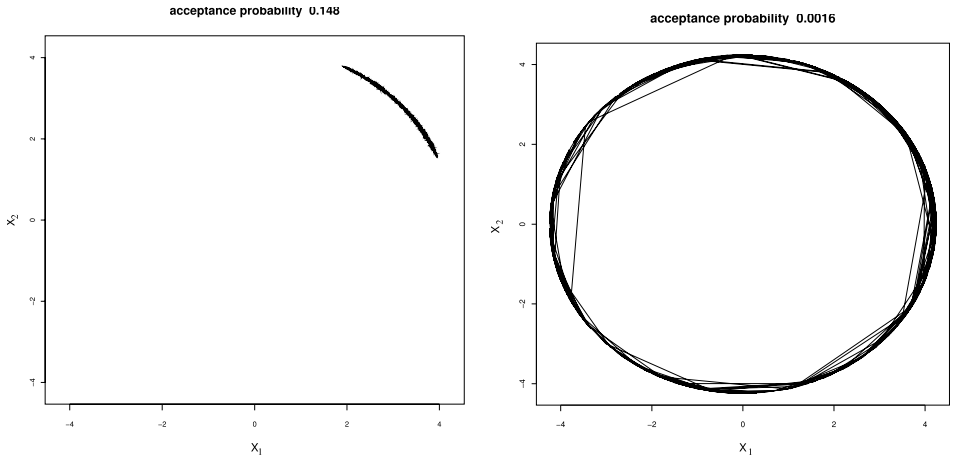

FIG. 1. *Two RWM runs (over equal amount of time) for the posterior of* $\theta = (\theta_1, \theta_2)$, *given datum* $y \sim \mathrm{N}(|\theta|^2, 0.01^2)$ *and prior* $\theta \sim \mathrm{N}(0, I_2)$. *The left (resp., right) panel is the MCMC scatterplot for step-size giving average acceptance probability* 14.8% *(resp.,* 0.16%).

close to zero seem to provide better mixing for the algorithms, fully contradicting prevailing practices for RWM.

In statistics, there are many examples that can be placed within our setting—we provide here a few, including our motivating scenario from Bayesian inverse problems.

- In inverse problems, situations where the posterior distribution concentrates in the neighbourhood of a nonlinear manifold abound; see, for example, [6, 10, 15]. A typical situation involves the posterior distribution of a high-dimensional vec-

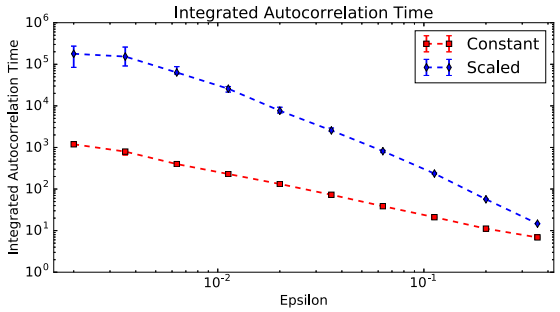

FIG. 2. *RWM runs for the posterior of* $\theta = (\theta_1, \theta_2)$, *given datum* $y \sim \mathrm{N}(|\theta|^2, \varepsilon^2)$ *and prior* $\theta \sim \mathrm{N}(0, I_2)$. *The "Scaled" (resp., "Constant") strategy consisted in choosing a RWM covariance jump structure equal to* $\varepsilon^2 I_2$ *(resp.,* I_2). *The RWM algorithms were run for* 10^6 *iterations and* 90% *confidence intervals for the IACTs were obtained through* 50 *repetitions per value of* $\varepsilon > 0$. *As will become clear from the theory developed in this article, the IACT of the "scaled" strategy is of order* $\mathcal{O}(\varepsilon^{-2})$ *while the IACT of the "constant" strategy is of order* $\mathcal{O}(\varepsilon^{-1})$.

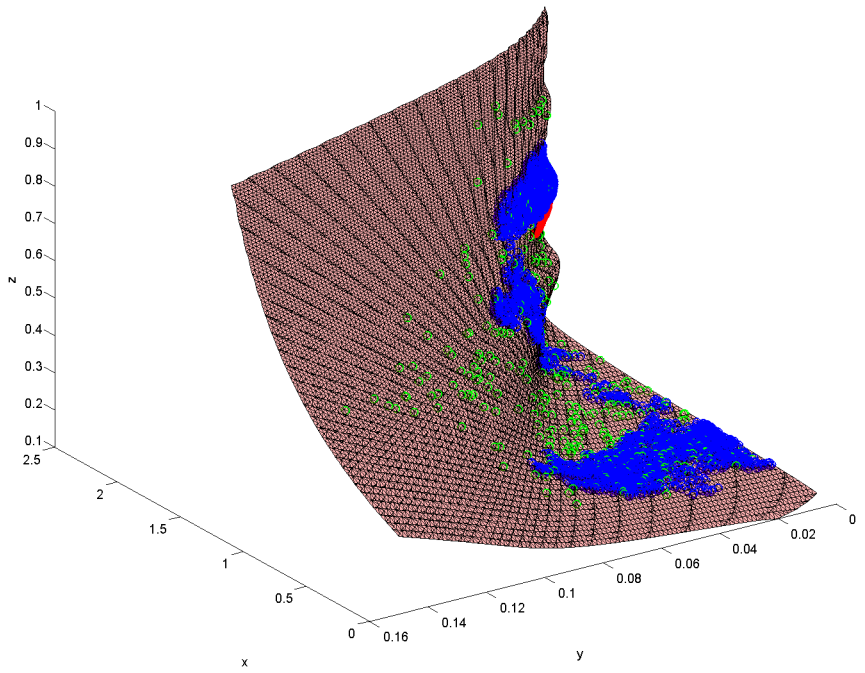

FIG. 3. *RWM runs for the posterior of* $(\theta_1, \theta_2, \theta_3)$ *given datum* $y \sim \mathrm{N}(x_{25}(\theta), 0.01^2)$ *with* $x'_t = \theta_1 + \theta_2 x_t + 2\sin(\theta_3 x_t)$, $x_0 = \frac{1}{2}$; *priors are* $\theta_1 \sim \mathrm{IG}(\mu = 2, \sigma^2 = 4)$ (*i.e., Inverse Gamma distribution of mean* 2 *and variance* 4), $\theta_2 \sim U(0, 1)$, $\theta_3 \sim \mathrm{IG}(\mu = \frac{1}{2}, \sigma^2 = 4)$, *independently over the three components. The red, blue and green dots show the runs—each produced within* 30 *mins—corresponding to average acceptance probabilities* 17%, 0.57% *and* 0.03%, *respectively*; *the shaded surface represents the 2D-manifold* $x_{25}(\theta) = x_{25}(\theta^\dagger)$.

tor $x \in \mathbb{R}^n$ observed through a low dimensional, possibly noisy, nonlinear measurement function $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ with $d \ll n$; the data collected is distributed as $y \sim \Phi(x) + (\text{noise})$. As the intensity $\varepsilon > 0$ of the noise decreases to zero, the posterior distribution $\pi(x|y)$ typically concentrates in a neighbourhood of thickness ε around the manifold $\mathcal{M} \equiv \{x \in \mathbb{R}^n : \Phi(x) = y\}$.

- Suppose that we have a sequence of posterior distributions $\{\pi_n(\theta)\}$, indexed by the data size n, where we can write $\theta = (\theta_1, \theta_2)$ with θ_1 representing the components of the posterior which are *identifiable* while θ_2 remains unidentifiable so that its standard deviation remains $\mathcal{O}(1)$ for increasing n. For example, consider the posterior $\pi_n(\sigma, \mu)$ associated to $n \geq 1$ observations $y_k = X_{k/n}$ of the diffusion process $dX = \mu\, dt + \sigma\, dW$ on $[0, 1]$; as $n \to \infty$, only the volatility coefficient is identifiable. In the so-called *regular* case, we might expect the marginal posterior of θ_1 to contract at rate $n^{-1/2}$ for instance. Therefore, it is very natural to see a scale divergence as $n \to \infty$.
- Consider the context of maximum likelihood estimation (MLE) for nonregular likelihood functions, For such problems, *super-efficiency* of MLEs is a well-

known phenomenon (see for example [20]) in which the standard deviation of the MLE shrinks to 0 faster than $n^{-1/2}$ for data size n. It may be that *super-efficiency* applies only to some of the model parameters, leading to the kind of heterogeneously scaled target distributions we consider here. For concreteness, we consider the specific problem where the data is assumed to be drawn i.i.d. from the model $X \sim \text{Exp}(\lambda)$ conditioned on $X \leq \theta$. It is well known that for this kind of example the posterior for λ contracts at the regular rate of $n^{-1/2}$ while that for θ contracts at the much more rapid rate of n^{-1}. Thus again we get scale divergence as $n \to \infty$.

The paper is structured as follows. In Section 2, we describe the RWM algorithm for a density concentrating on a hyperplane with rate $\varepsilon > 0$. In Section 3, we state the regulatory conditions and write the diffusion limits as $\varepsilon \to 0$. In Section 4, we prove the stated results. In Section 5, we prove a limiting result involving a Markov jump-process, when the step-sizes are order $\mathcal{O}(1)$, so are not adapted to the size of the smallest coordinate. In Section 6, we describe in some detail a conjecture for generalizing our diffusion limit results in the context of nonlinear manifolds. Throughout, we provide comments about the implications of the theoretical results. We finish with some conclusions and description of future work in Section 7.

2. Random-walk Metropolis on affine manifold. As explained above, we prove analytically rigorous results in the case when the manifold is flat, that is, an affine subspace of the general state space. We discuss later on in the paper extensions to more general manifold structures. We model the affine scenario as follows. We consider the target distribution $\pi_\varepsilon : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \mathbb{R}$, for integers $n_x, n_y \geq 1$ and $n = n_x + n_y$, with density with respect to the n-dimensional Lebesgue measure

$$\pi_\varepsilon(x, y) = \pi_X(x)\pi_{Y|X}(y|x) = \frac{1}{\varepsilon^{n_y}} e^{A(x)} e^{B(x,y/\varepsilon)}, \tag{2}$$

for a "small" scalar $\varepsilon > 0$. The x-marginal has density $\pi_X(x) = e^{A(x)}$ independently of ε. The distribution π_ε is a scaled version of the probability distribution

$$\pi(x, y) \equiv \pi_1(x, y) = e^{A(x)} e^{B(x,y)}. \tag{3}$$

As $\varepsilon \to 0$, the support of the sequence of distributions π_ε concentrates on the linear subspace $\mathcal{M}$ defined as

$$\mathcal{M} = \{(x, y) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} : y = 0\} \subset \mathbb{R}^{n_x + n_y}, \tag{4}$$

of dimension n_x. Integer n_x can sometimes be thought of as the dimension of the nonidentifiability and integer n_y as the part of dimension that is fully specified. For instance, in a small-noise or increasing-data context, one can consistently estimate n_y-parameters out of n. Parameter ε can be thought of as the thickness of the support of π_ε at a neighbourhood of $(x, 0) \in \mathcal{M}$.

To obtain samples from π_ε, we consider the RWM algorithm proposing moves

$$\begin{pmatrix} X'_\varepsilon \\ Y'_\varepsilon \end{pmatrix} \begin{pmatrix} X_\varepsilon \\ Y_\varepsilon \end{pmatrix} + \ell h(\varepsilon) \begin{pmatrix} Z_x \\ Z_y \end{pmatrix} \tag{5}$$

for a scaling factor $h(\varepsilon)$, tuning parameter $\ell > 0$ and noise $(Z_x, Z_y) \sim \mathrm{N}(0, I_{n_x+n_y})$. In this setting, it would be more efficient to choose different scales for the x and y coordinates. Recall, though, that we are in fact interested in the more general situation where the manifold $\mathcal{M}$ is not linear—our choice of linear manifold is only a mathematical convenience for gaining insight into the general case. For nonlinear manifolds, there is no sensible way of choosing a priori the covariance structure of the jumps of the RWM algorithm. This is the motivation for studying the algorithm (5) that employs isotropic jumps. Factor $h(\varepsilon)$ determines the scale of the jumps of the RWM algorithm and the tuning parameter ℓ allows to control the acceptance rate of the algorithm. When the context is clear, we write simply (X, Y) instead of $(X_\varepsilon, Y_\varepsilon)$. For the purpose of the analysis, we introduce the rescaled coordinate U_ε and associated proposal U'_ε,

$$U_\varepsilon = Y_\varepsilon/\varepsilon; \qquad U'_\varepsilon = Y'_\varepsilon/\varepsilon = U_\varepsilon + \ell \frac{h(\varepsilon)}{\varepsilon} Z_y.$$

If $(X_\varepsilon, Y_\varepsilon) \sim \pi_\varepsilon$, then $(X_\varepsilon, U_\varepsilon) \sim \pi$. To finish the description of the MCMC algorithm, we need to choose an accept-reject function F. To obtain detailed balance w.r.t. π_ε, one can choose any $(0, 1]$-valued function F satisfying the reversibility condition

$$e^r F(-r) = F(r), \qquad r \in \mathbb{R}. \tag{6}$$

The usual Metropolis–Hastings accept/reject correction corresponds to the choice $F(r) = F_{\mathrm{MH}}(r) = \min(1, e^r)$, so that the move $(X, Y) \mapsto (X', Y')$, or equivalently $(X, U) \mapsto (X', U')$, is accepted with probability

$$F \circ \log\left(\frac{\pi_\varepsilon(X', Y')}{\pi_\varepsilon(X, Y)}\right).$$

We allow for a general accept–reject function $F(r)$ for two main reasons. First, enforcing some smoothing upon F removes several inessential technicalities when proving diffusion limits in the sequel. Second, we would like to emphasize that the results obtained in this paper (and many others), can straightforwardly be adapted to this more general setting; for example, our setting accommodates the Barker's accept–reject function $u \mapsto 1/(1+e^u)$ [1]. For the target density (2) the acceptance probability can be written analytically as $a(X, U, h(\varepsilon)Z_x, h(\varepsilon)\varepsilon^{-1}Z_y)$ where the function $a(\cdot, \cdot, \cdot, \cdot)$ reads

$$\begin{aligned} &a(x, u, w_x, w_u) \\ &\quad = F\big(A(x + \ell w_x) - A(x) + B(x + \ell w_x, u + \ell w_u) - B(x, u)\big). \end{aligned} \tag{7}$$

The above proposal and acceptance probability give rise to the MCMC trajectory $\{(X_{\varepsilon,k}, Y_{\varepsilon,k})\}_{k\geq 0}$.

2.1. *Notation.* For $(\varepsilon, x, u) \in \mathbb{R}^+ \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$, we write $\mathbb{E}_{\varepsilon,x,u}[\cdot]$ (or simply $\mathbb{E}_{x,u}[\cdot]$) to represent conditional expectations $\mathbb{E}[\cdot|(X_{\varepsilon,0}, U_{\varepsilon,0}) = (x, u)]$. For functions $e = e(x, u)$, we write $\mathbb{E}_\pi[e(x, u)]$ to indicate that the expectation is considered under the standardised law $(x, u) \sim \pi$. We denote by $\|\cdot\|_p$ the L_p-norm of a random variable, $p \geq 1$. $|\cdot|$ denotes the standard norm on an Euclidean space $\mathbb{R}^n$ and $\langle\cdot,\cdot\rangle$ the corresponding inner product, for any $n \geq 1$. We use the notation $o_{L_1(\pi)}(1)$ to designate any sequence of functions whose expectation under π converge to zero (as $\varepsilon \to 0$). We use the symbol $(\lesssim)$ to indicate inequalities that hold under multiplication with a constant which does not depend on critical parameters implied by the context.

3. Diffusion limit. We study in this section the behaviour of RWM, as $\varepsilon \to 0$, for a scaling factor of the form

$$h(\varepsilon) = \varepsilon. \tag{8}$$

Note that it is straightforward to prove that, under the below described regularity Assumption 2 on the functions A and B, any other scaling factor $h(\varepsilon)$ such that $h(\varepsilon)/\varepsilon \to \infty$ would lead to an algorithm with an acceptance rate that decreases to zero as $\varepsilon \to 0$. This corresponds to an algorithm that proposes too large moves and such a context is investigated in Section 5. A scaling factor of the form (8) thus provides the largest choice of jump-sizes for an algorithm with acceptance rate that does not degenerate to zero in the limit $\varepsilon \to 0$. To state one of our main results, we introduce the quantity

$$a_0(x, \ell) = \int_{\mathbb{R}^{n_y}} \mathbb{E}_{x,u}\big[F\big(B(x, u + \ell Z_y) - B(x, u)\big)\big] e^{B(x,u)}\, du, \tag{9}$$

which corresponds to the limiting average acceptance probability, as $\varepsilon \to 0$, of the RWM algorithm when assuming stationarity for $U|X = x$, conditionally on a fixed position for the x-coordinate. In this article, we always assume that the functions A and B are continuous, although more stringent assumptions are required at different places. Under this continuity assumption, one can verify via the bounded convergence theorem that

$$a_0(x, \ell) = \lim_{\varepsilon \to 0} \mathbb{E}\big[a(X, U, \varepsilon Z_x, Z_y)|X = x\big], \tag{10}$$

assuming $(X, U) \sim \pi$. We prove a diffusion limit for the trajectory of the x-coordinate, after considering a proper continuous time-scale for its discrete-time trajectory $\{X_{\varepsilon,k}\}_{k\geq 0}$, where k denotes the number of MCMC iterations. In our setting, it is not interesting to study the scaling limit of y; indeed, y is of order ε and it will become clear from our analysis that only the behaviour of the slow-coordinate x dictates the complexity of the RWM algorithm in the regime $\varepsilon \to 0$. We set

$$c(\varepsilon) = \varepsilon^{-2}.$$

Our main result states that the sequence of accelerated, continuous time, càdlàg processes

$$\widetilde{X}_{\varepsilon,t} := X_{\varepsilon,\lfloor t\cdot c(\varepsilon)\rfloor} \tag{11}$$

converges weakly, as $\varepsilon \to 0$, to a nontrivial diffusion process; thus $c(\varepsilon)$ corresponds to the "diffusive time-scale".

REMARK 1. For a given positive density function $\pi : \mathbb{R}^n \to (0,\infty)$, for $n \geq 1$, and a scalar volatility function $\sigma : \mathbb{R}^n \to (0,\infty)$ we introduce the function $\mathscr{D}_{\pi,\sigma^2} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\mathscr{D}_{\pi,\sigma^2} : x \mapsto \frac{1}{2}\nabla\sigma^2(x) + \frac{1}{2}\sigma^2(x)\nabla\log\pi(x). \tag{12}$$

Consider the stochastic differential equation (SDE),

$$d\overline{X} = \mathscr{D}_{\pi,\sigma^2}(\overline{X})\,dt + \sigma(\overline{X})\,dW, \tag{13}$$

where W denotes an n-dimensional Wiener process. Under standard regularity assumptions, SDE (13) has a unique global solution and is reversible with respect to the law $\pi(dx)$. The case $\sigma \equiv$ const. corresponds to the Langevin diffusion $d\overline{X} = \frac{\sigma^2}{2}\nabla\log\pi(\overline{X})\,dt + \sigma\,dW$.

We begin with an assumption on the accept/reject function F introduced in (6).

ASSUMPTION 1. The accept–reject function F is differentiable. F and F' are globally Lipschitz and bounded over the real line.

Assumption 1 allows for straightforward use of Taylor expansions in the proofs that follow. It could probably be relaxed given substantial technical work. To work with the standard Metropolis–Hastings ratio would involve dealing with the discontinuity of the derivative at 0. We believe that this is a technical distraction that would not bring new insights into the behaviour of the algorithm and would make the proofs much more opaque. Notice that any F satisfying the reversibility condition (6) is dominated by the Metropolis–Hastings function F_{MH} in the sense that

$$F(r) \leq F_{\mathrm{MH}}(r)$$

holds for any $r \in \mathbb{R}$. To obtain our scaling limit, we assume the following regularity conditions for functions A and B involved in the specification of the density π_ε.

ASSUMPTION 2 (Regularity conditions on π_ε). Functions $A : \mathbb{R}^{n_x} \to \mathbb{R}$ and $B : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \to \mathbb{R}$ are differentiable and their derivatives are globally Lipschitz. Moreover, we assume that the distribution $\pi \equiv \pi_1$ in (3) possesses finite second moments so that

$$\mathbb{E}_\pi\big[|X|^2 + |U|^2\big] < \infty. \tag{14}$$

Assumption 2 is repeatedly used to control the error terms associated to the use of second-order Taylor expansions; the second moment bound allows the use of the Cauchy–Schwarz inequality. Note that these assumptions could be relaxed in several directions at the cost of increasing complexity in the proofs; for example, one could assume only a polynomial growth on the derivatives and higher moments for π. We also need the following assumption to control the behaviour of the diffusion limit identified below.

ASSUMPTION 3. The function $x \mapsto \frac{1}{2}\nabla(a_0(x,\ell)\log\pi_X(x))$ is bounded on $\mathbb{R}^{n_x}$ and there exists a $\mu \in (0,1]$ such that

$$|a_0(x,\ell) - a_0(x',\ell)| + |\nabla(a_0(x,\ell)\times\log\pi_X(x)) - \nabla(a_0(x',\ell)\times\log\pi_X(x'))| \lesssim |x-x'|^{\mu}.$$

The main theorem of this section is the following. The proof is given in Section 4.

THEOREM 1. *Let $T>0$ be a fixed time horizon. Assume that the RWM algorithm is started in stationarity, $(X_{\varepsilon,0}, Y_{\varepsilon,0}) \sim \pi_\varepsilon$, and Assumptions 1–3 hold. Then, as $\varepsilon \to 0$ the sequence of processes $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ defined via (5), (11) converges weakly in the Skorokhod space $\mathbf{D}([0,T],\mathbb{R}^{n_x})$ to the diffusion process $\{\overline{X}_t\}_{t\in[0,T]}$ specified as the solution of the SDE*

$$d\overline{X} = \mathscr{D}_{\pi_X,\sigma^2}(\overline{X})\,dt + \sigma(\overline{X})\,dW \tag{15}$$

for volatility function $\sigma^2(x) \equiv \ell^2 a_0(x,\ell)$ and initial position $\overline{X}_0 \sim \pi_X(x) = e^{A(x)}$.

$\mathbf{D}([0,T],\mathbb{R}^{n_x})$ denotes the space of $\mathbb{R}^{n_x}$-valued functions on $[0,T]$ that are right continuous with left-side limits. For a definition of the distance on $\mathbf{D}([0,T],\mathbb{R}^{n_x})$ giving rise to a metric space and to the corresponding Skorokhod topology; see [7], Chapter 3. Assumption 3 is as stated in [7], Chapter 4, Theorem 1.6, and it implies that the limiting diffusion (15) is well defined, that is, does not explode in finite time, and has a unique strong solution. Furthermore, Assumption 3 yields that for a smooth and compactly supported test function φ the function $x \mapsto \mathcal{L}\varphi(x)$, with

$$\mathcal{L} = \frac{\ell^2}{2}\nabla\{a_0(x,\ell)\log\pi_X(x)\}\cdot\nabla + \frac{\ell^2}{2}a_0(x,\ell)(\nabla\cdot\nabla) \tag{16}$$

the generator of the limiting diffusion (15), is μ-Holderian. The diffusive time scale $c(\varepsilon) = \varepsilon^{-2}$ implies that the algorithmic complexity of RWM grows as $\mathcal{O}(\varepsilon^{-2})$ as the thickness ε approaches zero; see [19] for general results on the complexity analysis of MCMC algorithms through diffusion limits. In the case where the function $(x,y) \mapsto B(x,y)$ does not depend on the x-coordinate, the limiting acceptance probability a_0 does not depend on the local position x, that is, $a_0(x,\ell) = a_0(\ell)$. In this case, the optimal value for the parameter ℓ is given by $\ell_\star = \operatorname{argmax}\ell^2 a_0(\ell)$. In the general case, however, the optimal choice of ℓ will depend on the current x-position.

3.1. *Extending Theorem 1.* We can also adopt slightly more general proposals where the variance of the proposals is allowed to depend on the current position. That is, the tuning parameter $\ell = \ell(x) > 0$ is now allowed to depend on the x-coordinate:

$$\begin{pmatrix} X'_\varepsilon \\ Y'_\varepsilon \end{pmatrix} = \begin{pmatrix} X_\varepsilon \\ Y_\varepsilon \end{pmatrix} + \ell(X_\varepsilon)\varepsilon \begin{pmatrix} Z_x \\ Z_y \end{pmatrix}. \tag{17}$$

We state the required assumptions on the function $x \mapsto \ell(x)$ that allows the relevant diffusion limit results to hold.

ASSUMPTION 4 [Regularity assumptions on $x \mapsto \ell(x)$]. Function ℓ is positive, bounded away from zero and infinity. The first two derivatives of ℓ are bounded.

We choose to work in the limited setup of Assumption 4 so that the proof of the next theorem is a straightforward adaptation of Theorem 1. The accelerated version (11) of the x-coordinate process again converges to a diffusion process.

THEOREM 2. *Let $T > 0$ be a fixed time horizon. Assume that Assumptions 1–4 hold and that the RWM algorithm is started in stationarity. As $\varepsilon \to 0$, the sequence of processes $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ defined via (17), (11) converges weakly in the Skorokhod space $\mathbf{D}([0,T],\mathbb{R}^{n_x})$ to the diffusion process $\{\overline{X}_t\}_{t\in[0,T]}$ specified as the solution of the SDE*

$$d\overline{X} = \mathscr{D}_{\pi_X,\sigma^2}(\overline{X})\,dt + \sigma(\overline{X})\,dW, \tag{18}$$

where the local volatility function is $\sigma^2(x) \equiv \ell^2(x)a_0(x,\ell(x))$. The initial distribution is $\overline{X}_0 \sim \pi_X$.

The only difference with Theorem 1 is the form of the volatility function σ. As before, the limiting distribution (18) is reversible with respect to π_X and the associated Dirichlet form reads

$$\mathcal{D}(\varphi) \equiv \frac{1}{2}\int_{\mathbb{R}^{n_x}} \big|\nabla\varphi(x)\big|^2 \ell^2(x)a_0\big(x,\ell(x)\big)\pi_X(dx). \tag{19}$$

Since the parameter $\ell = \ell(x)$ is a function of the x-coordinate, the optimal choice $\ell_\star(x)$ for the tuning parameter ℓ is

$$\ell_\star(x) \equiv \operatorname*{argmax}_{\ell>0} \ell^2 a_0(x,\ell). \tag{20}$$

As described in [18], the choice (20) maximises the expected squared jumping distance, the spectral gap of the limiting diffusion (18) and minimizes the asymptotic variance of MCMC estimators.

3.2. *High-dimensional asymptotics and local* 0.234 *rule.* In this section, we investigate the behaviour of the optimal jumping rule $x \mapsto \ell_\star(x)$ in the regime where the dimensionality of the identifiability is large, that is, $n_y \gg 1$. We adopt the setting where the dimension n_x remains fixed while the dimension n_y grows to infinity. For simplicity, we postulate a product form for the y-conditional density. That is, we investigate the sequence of densities

$$\pi_\varepsilon^{(n_y)}(x, y) = \frac{1}{\varepsilon^{n_x}} e^{A(x)+B^{(n_y)}(x,y/\varepsilon)}, \qquad x \in \mathbb{R}^{n_x}, y \in \mathbb{R}^{n_y},$$

in the regime where $n_y \to \infty$ and $\varepsilon \to 0$, with

$$B^{(n_y)}(x, u) = \sum_{j=1}^{n_y} b(x, u_j),$$

where $b : \mathbb{R}^{n_x} \times \mathbb{R} \to \mathbb{R}$ is such that for any $x \in \mathbb{R}^{n_x}$ the function $u \mapsto e^{b(x,u)}$ is a proper density function on $\mathbb{R}$. We denote by $\ell_\star^{(n_y)}(x_0)$ the optimal value of the jump size when the x-coordinate of the MCMC algorithm exploring $\pi_\varepsilon^{(n_y)}$ stands at $x_0 \in \mathbb{R}^{n_x}$. With the obvious notation, the previous sections show that $\ell_\star^{(n_y)}(x_0) = \operatorname{argmax}\{\ell^2 a_0^{(n_y)}(x_0, \ell) : \ell > 0\}$ where

$$a_0^{(n_y)}(x_0, \ell) = \mathbb{E}\left[F\left(\sum_{j=1}^{n_y} b\big(x_0, U_j^{(x_0)} + \ell Z_j\big) - b\big(x_0, U_j^{(x_0)}\big) \right) \right]$$

for an i.i.d. sequence $\{U_j^{(x_0)}\}_{j\geq 1}$ of $\mathbb{R}$-valued random variables with distribution $e^{b(x_0,u)}\,du$ and an i.i.d. sequence of standard Gaussian random variables $\{Z_j\}_{j\geq 0}$. We assume the following regularity conditions on the marginal density $u \mapsto e^{b(x_0,u)} \equiv \mu_{(x_0)}(u)$; this is the equivalent to conditions (A1) and (A2) of [16].

ASSUMPTION 5. The density $\mu_{(x_0)}(u)$ is twice differentiable, the function $u \mapsto \mu'_{(x_0)}(u)/\mu_{(x_0)}(u)$ is Lipschitz continuous and the random variable $U^{(x_0)}$ with density $\mu_{(x_0)}$ is such that

$$\mathbb{E}\left[\left(\frac{\mu'_{(x_0)}}{\mu_{(x_0)}}\big(U^{(x_0)}\big) \right)^8 \right] < \infty; \qquad \mathbb{E}\left[\left(\frac{\mu''_{(x_0)}}{\mu_{(x_0)}}\big(U^{(x_0)}\big) \right)^4 \right] < \infty.$$

A simple adaptation of Corollary 1.2 of [16] yields that, under Assumption 5, we have

$$\lim_{n_y\to\infty} a_0^{(n_y)}\big(x_0, \ell n_y^{-1/2}\big) = \mathbb{E}\big[F\big(\mathrm{N}\big(-I_{(x_0)}^2/2, I_{(x_0)}^2\big)\big)\big] \equiv \overline{a}_0(\ell);$$

$$I_{(x_0)}^2 = \mathbb{E}\left[\left(\frac{\mu'_{(x_0)}}{\mu_{(x_0)}}\big(U^{(x_0)}\big) \right)^2 \right].$$

Corollary 1.2 of [16] corresponds to the special case $F = F_{\mathrm{MH}}$ where a closed-form expression for $\overline{a}_0(\ell)$ exists; for concreteness, we will also consider the special case $F = F_{\mathrm{MH}}$ although generalization to arbitrary accept–reject functions is straightforward. The function $\ell \mapsto \ell^2 \times \overline{a}_0(\ell)$ is maximized for $\ell_\star$ such that $\overline{a}_0(\ell_\star) = 0.234$ (to three decimal places) [16]; in other words, as $n_y \to \infty$, the optimal jumping rule $\ell_\star(x)$ can be chosen such that the local acceptance rate at $x_0 \in \mathbb{R}^{n_x}$ equals approximately 0.234. Indeed, the derivation of this rule-of-thumb relies on the product form assumption of the y-marginal and is only valid in the asymptotic regime $n_y \to \infty$. Nevertheless, this type of analysis has been shown to be empirically and theoretically [2–5, 12, 14] relevant to more general distribution structures.

4. Proof of diffusion limit. This section gives rigorous proofs of Theorems 1 and 2. We first give a high-level description of the proof of Theorem 1. We introduce an intermediate time scale

$$\widetilde{c}(\varepsilon) = \varepsilon^{-\gamma} \qquad \text{for some exponent } \gamma \in \left(0, \frac{1}{2}\right),$$

and consider the sub-sampled process $\{(S_{\varepsilon,k}, V_{\varepsilon,k})\}_{k\geq 0}$ defined as

$$\begin{cases} (S_{\varepsilon,0}, S_{\varepsilon,1}, S_{\varepsilon,2}, \ldots) = (X_{\varepsilon,0}, X_{\varepsilon,\lfloor \widetilde{c}(\varepsilon)\rfloor}, X_{\varepsilon,\lfloor 2\widetilde{c}(\varepsilon)\rfloor}, \ldots), \\ (V_{\varepsilon,0}, V_{\varepsilon,1}, V_{\varepsilon,2}, \ldots) = (U_{\varepsilon,0}, U_{\varepsilon,\lfloor \widetilde{c}(\varepsilon)\rfloor}, U_{\varepsilon,\lfloor 2\widetilde{c}(\varepsilon)\rfloor}, \ldots). \end{cases}$$

On this time-scale the x-process evolves slowly (i.e., $S_{\varepsilon,k} \approx S_{\varepsilon,k+1}$) whereas the y-process is allowed enough time to mix (i.e., $V_{\varepsilon,k+1}$ is approximately independent from $V_{\varepsilon,k}$). We define the continuous-time accelerated processes

$$\widetilde{S}_{\varepsilon,t} = S_{\varepsilon,\lfloor t\cdot c(\varepsilon)/\widetilde{c}(\varepsilon)\rfloor} \equiv S_{\varepsilon,\lfloor t\cdot\varepsilon^{\gamma-2}\rfloor}; \qquad \widetilde{V}_{\varepsilon,t} \equiv V_{\varepsilon,\lfloor t\cdot\varepsilon^{\gamma-2}\rfloor}. \tag{21}$$

Note that it is natural to accelerate time by a factor $c(\varepsilon)/\widetilde{c}(\varepsilon)$ in the definition (21) since we expect that, after acceleration by a factor $c(\varepsilon)$, the original process X_ε converges to a nontrivial diffusion limit; with this choice, we thus expect the sequence $\widetilde{S}_{\varepsilon,t}$ to also converge towards the same diffusion limit. See Figure 4 for a graphical representation of all four main processes involved in our derivations. The proof is divided into two steps. First, we show that process $\widetilde{S}_\varepsilon$ converges weakly in the Skorokhod space to the diffusion (15). Then we prove that the supremum norm

$$\|\widetilde{S}_{\varepsilon,\cdot} - \widetilde{X}_{\varepsilon,\cdot}\|_{\infty,[0,T]} \equiv \sup\{|\widetilde{S}_{\varepsilon,t} - \widetilde{X}_{\varepsilon,t}| : t \in [0,T]\}$$

converges to zero in probability. By the definition of the Skorokhod metric [7], this also means that the difference $(\widetilde{S}_{\varepsilon,\cdot} - \widetilde{X}_{\varepsilon,\cdot})$ converges to zero in probability as $\varepsilon \to 0$, in the Skorokhod space. Consequently, by Slutsky's lemmas, the sequence $\widetilde{X}_\varepsilon$ itself converges weakly in the Skorohod space towards the diffusion (15).

We define some quantities needed in the sequel. Recall the definition in (16) of the generator $\mathcal{L}$ of the limiting diffusion (15). We define the appropriate space of test functions

$$\mathcal{C} = \{\varphi : \mathbb{R}^{n_x} \to \mathbb{R} : \varphi \text{ is smooth with compact support}\}.$$

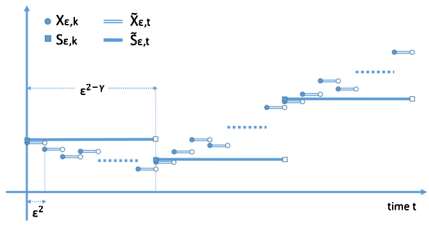

FIG. 4. *The four processes involved in the diffusion limit result. The discrete-time process* $\{S_{\varepsilon,k}\}_{k\geq 0}$ *corresponds to a thinning (of frequency* $1/\lfloor \varepsilon^{-\gamma} \rfloor$*) of* $\{X_{\varepsilon,k}\}_{k\geq 0}$*; the first process is illustrated with the filled rectangles, the latter with the filled circles. Then the continuous-time càdlàg process* $\{\widetilde{S}_{\varepsilon,t}\}_{t\geq 0}$ *is illustrated with the filled lines and the* $\{\widetilde{X}_{\varepsilon,t}\}_{t\geq 0}$ *one with the empty lines (the time instances indicated by the rectangles or the circles correspond to the jump times of the relevant processes).*

Assumption 3 implies that $\mathcal{C}$ is a core for the generator $\mathcal{L}$ in (16). For test functions $\varphi \in \mathcal{C}$ and vectors $(x,u) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$, we define the operators $\mathcal{L}_\varepsilon$, $\widetilde{\mathcal{L}}_\varepsilon$ as follows:

$$\begin{cases} \mathcal{L}_\varepsilon \varphi(x,u) = \mathbb{E}_{\varepsilon,x,u}\big[\varphi(X_{\varepsilon,1}) - \varphi(x)\big]/\varepsilon^2, \\ \widetilde{\mathcal{L}}_\varepsilon \varphi(x,u) = \mathbb{E}_{\varepsilon,x,u}\big[\varphi(S_{\varepsilon,1}) - \varphi(x)\big]/\varepsilon^{2-\gamma}. \end{cases} \tag{22}$$

We will refer to $\mathcal{L}_\varepsilon$, $\widetilde{\mathcal{L}}_\varepsilon$ as the "generators" of the càdlàg processes $\{\widetilde{X}_{\varepsilon,t}\}_{t\geq 0}$, $\{\widetilde{S}_{\varepsilon,t}\}_{t\geq 0}$, respectively, with an abuse in terminology as these processes are not Markovian with respect to their own filtration. Thus, though the domain of φ is $\mathbb{R}^{n_x}$, functions $\mathcal{L}_\varepsilon \varphi$ and $\widetilde{\mathcal{L}}_\varepsilon \varphi$ are defined on $\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$.

4.1. *Proof of weak convergence of* $\widetilde{S}_\varepsilon$ *to the limiting diffusion* (15). Our derivation of the diffusion limit proof is based on the following critical result connecting an L_1-limit between generators with weak convergence of the processes themselves. Recall that $\mathcal{L}$ is as defined in (16).

PROPOSITION 1. *Let Assumptions 1–3 hold. If the following limit holds*

$$\lim_{\varepsilon \to 0} \mathbb{E}_\pi \big| \widetilde{\mathcal{L}}_\varepsilon \varphi(x,u) - \mathcal{L}\varphi(x) \big| = 0 \tag{23}$$

for any $\varphi \in \mathcal{C}$*, then as* $\varepsilon \to 0$ *the sequence of processes* $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ *(started at time* 0 *from the distribution* π_X*) converges weakly in* $\mathbf{D}([0,T], \mathbb{R}^{n_x})$ *to the diffusion process* (15).

PROOF. The proof is given in the Appendix. We mention here that we are based on two results: (i) [7], Chapter 4, Theorem 8.2, Corollary 8.5, which give conditions under which the finite-dimensional distributions of a sequence of processes converge weakly to those of a Markov process; (ii) [7], Chapter 8, Corollary 8.6, which provides further conditions for this sequence of processes to be

relatively compact, and thus establish weak convergence of the stochastic processes themselves in $\mathbf{D}([0,T],\mathbb{R}^{n_x})$. We show in the Appendix that under Assumptions 1–3, the limit (23) implies the required conditions for the relevant results in [7]. □

It now remains to prove (23) for any $\varphi \in \mathcal{C}$. We will achieve this via the series of Lemmas 1–4 that follow.

The telescoping sum

$$\varphi(S_{\varepsilon,1}) - \varphi(S_{\varepsilon,0}) = \varphi(X_{\varepsilon,1}) - \varphi(X_{\varepsilon,0}) + \cdots + \varphi(X_{\varepsilon,\lfloor \widetilde{c}(\varepsilon)\rfloor}) - \varphi(X_{\varepsilon,\lfloor \widetilde{c}(\varepsilon)\rfloor - 1})$$

yields that

$$\widetilde{\mathcal{L}}_\varepsilon \varphi(x,u) \equiv \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma}\rfloor - 1} \mathbb{E}_{\varepsilon,x,u}\big[\mathcal{L}_\varepsilon \varphi(X_{\varepsilon,j}, U_{\varepsilon,j})\big]. \tag{24}$$

As a first step, we prove that $\mathcal{L}_\varepsilon \varphi(x,u)$ converges, in the appropriate sense, to the quantity $\mathcal{A}\varphi(x,u)$ given by

$$\begin{aligned} \mathcal{A}\varphi(x,u) &= \ell^2 \big\langle \mathbb{E}_{x,u}\big[F'(DB)\nabla_x\big\{A(x) + B(x,u+\ell Z_y)\big\}\big], \nabla\varphi(x)\big\rangle \\ &\quad + \frac{\ell^2}{2}\mathbb{E}_{x,u}\big[F(DB)\big]\Delta\varphi(x), \end{aligned} \tag{25}$$

where for notational convenience we have defined $DB = B(x,u+\ell Z_y) - B(x,u)$. Also, we have used the Laplacian notation $\Delta = \sum_{i=1}^{n_x} \partial_{x_i}^2$. In general, $\mathcal{A}\varphi(x,u)$ does not correspond to the generator of a Markov process. Note that if $\varphi \in \mathcal{C}$, under Assumptions 1–2 we have that $|\mathcal{A}\varphi(x,u)| \lesssim 1 + |u|$.

LEMMA 1. *Let Assumptions 1–2 be satisfied and $\varphi \in \mathcal{C}$ be a test function. Then we have that $|\mathcal{L}_\varepsilon\varphi(x,u) - \mathcal{A}\varphi(x,u)| \lesssim \varepsilon(1+|x|+|u|)$, thus the following limit holds*:

$$\lim_{\varepsilon\to 0} \mathbb{E}_\pi \big|\mathcal{L}_\varepsilon\varphi(x,u) - \mathcal{A}\varphi(x,u)\big| = 0. \tag{26}$$

PROOF. See Appendix A.2. □

LEMMA 2. *Let Assumptions 1–2 be satisfied and $\varphi \in \mathcal{C}$ be a test function. Then we have that the following limit holds*:

$$\lim_{\varepsilon\to 0} \mathbb{E}_\pi \left| \widetilde{\mathcal{L}}_\varepsilon\varphi(x,u) - \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma}\rfloor - 1} \mathbb{E}_{\varepsilon,x,u}\big[\mathcal{A}\varphi(x,U_{\varepsilon,j})\big] \right| = 0. \tag{27}$$

PROOF. See Appendix A.3. □

To treat the term

$$\frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma} \rfloor - 1} \mathbb{E}_{\varepsilon,x,u}\big[\mathcal{A}\varphi(x, U_{\varepsilon,j})\big]$$

in (27), we need to introduce a new Markov process *coupled* with the original one $\{(X_{\varepsilon,k}, U_{\varepsilon,k})\}_{k\geq 1}$, but with the x-coordinate pinned at its initial position. To this end, note that the Markov chain $\{X_{\varepsilon,j}, U_{\varepsilon,j}\}_{j\geq 0}$ can be described as follows. The initial position is defined as $(X_{\varepsilon,0}, U_{\varepsilon,0}) = (X, U)$ for some variables $(X, U) \sim \pi$. For a sequence $\{\xi_j\}_{j\geq 0}$ of i.i.d. random variables uniformly distributed on $(0, 1)$ and a sequence $\{(Z_{x,j}, Z_{u,j})\}_{j\geq 0}$ of i.i.d. random variables distributed as $\mathrm{N}(0, I_n)$ we then have

$$\begin{pmatrix} X_{\varepsilon,j+1} - X_{\varepsilon,j} \\ U_{\varepsilon,j+1} - U_{\varepsilon,j} \end{pmatrix} = \ell \begin{pmatrix} \varepsilon Z_{x,j} \\ Z_{y,j} \end{pmatrix} \times \mathbf{1}\big(\xi_j \leq a(X_{\varepsilon,j}, U_{\varepsilon,j}, \varepsilon Z_{x,j}, Z_{y,j})\big),$$

where the accept-reject function $a(\cdot,\cdot,\cdot,\cdot)$ is defined in (7). The new Markov chain $\{X_j^\star, U_j^\star\}_{j\geq 0}$ is defined as follows. For the *same* random variables (X, U) and $\{\xi_j\}_{j\geq 0}$ and $\{(Z_{x,j}, Z_{y,j})\}_{j\geq 0}$, we set $(X_0^\star, U_0^\star) = (X, U)$ and

$$\begin{pmatrix} X_{j+1}^\star - X_j^\star \\ U_{j+1}^\star - U_j^\star \end{pmatrix} = \ell \begin{pmatrix} 0 \\ Z_{y,j} \end{pmatrix} \times \mathbf{1}\big(\xi_j \leq a\big(X, U_j^\star, 0, Z_{y,j}\big)\big).$$

Critically, the x-coordinate of the new process $\{X_j^\star, U_j^\star\}_{j\geq 0}$ remains still and the process does not depend on the parameter ε. Also, conditionally on $X = x$, the process $\{U_j^\star\}_{j\geq 0}$ is simply a RWM Markov chain with target distribution on $\mathbb{R}^{n_y}$ proportional to $u \mapsto \exp(B(x, u))$. Thus, it readily follows from the ergodic theorem for Markov chains that for any smooth and compactly supported test function φ we have

$$\lim_{\varepsilon\to 0} \mathbb{E}_\pi \left| \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma} \rfloor - 1} \mathbb{E}_{x,u}\big[\mathcal{A}\varphi(x, U_j^\star)\big] - \int_{u\in\mathbb{R}^{n_y}} \mathcal{A}\varphi(x, u) e^{B(x,u)}\, du \right| = 0. \tag{28}$$

Furthermore, a routine calculation, whose details can be found in Section A.4, gives the following result.

LEMMA 3. *For any $x \in \mathbb{R}^{n_x}$ and any $\varphi \in \mathcal{C}$, we have*

$$\int \mathcal{A}\varphi(x, u) e^{B(x,u)}\, du = \mathcal{L}\varphi(x). \tag{29}$$

There is one result remaining to prove weak convergence of $\widetilde{S}_\varepsilon$ to the limiting diffusion.

LEMMA 4. *Let Assumptions 1–2 be satisfied and $\varphi \in \mathcal{C}$ be a test function. Then we have that the following limit holds*:

$$\lim_{\varepsilon\to 0} \mathbb{E}_\pi \left| \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma}\rfloor -1} \mathbb{E}_{\varepsilon,x,u}\big[\mathcal{A}\varphi(x, U_{\varepsilon,j})\big] - \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma}\rfloor -1} \mathbb{E}_{x,u}\big[\mathcal{A}\varphi\big(x, U_j^\star\big)\big] \right| = 0. \tag{30}$$

PROOF. It suffices to establish that, for $X_{0,\varepsilon} = X \sim \pi_X$,

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma}\rfloor -1} \big\|\mathcal{A}\varphi\big(X, U_j^\star\big) - \mathcal{A}\varphi(X, U_{\varepsilon,j})\big\|_1 = 0.$$

Under Assumptions 1–2 and due the fact that $\varphi \in \mathcal{C}$, we have that $|\mathcal{A}\varphi(x,u)| \lesssim 1 + |u|$ so that $|\mathcal{A}\varphi(X, U_j^\star) - \mathcal{A}\varphi(X, U_{\varepsilon,j})|$ is bounded by a constant multiple of $\mathbf{1}(U_j^\star \neq U_{\varepsilon,j}) \times (1 + |U_j^\star| + |U_{\varepsilon,j}|)$. Recall that both chains are started from $(X,U) \sim \pi$. Also, from stationarity, we have that U_j^* has the same law as U. By the Cauchy–Schwarz inequality, since $\mathbb{E}|U|^2 < \infty$ from Assumption 2, the conclusion follows once it is proved that

$$\lim_{\varepsilon\to 0} \frac{1}{\varepsilon^{-\gamma}} \sum_{j=0}^{\lfloor \varepsilon^{-\gamma}\rfloor -1} \mathbb{P}\big(U_j^\star \neq U_{\varepsilon,j}\big)^{1/2} = 0.$$

The definition of the coupling between $(\{X_j^\star, U_j^\star)\}_{j\geq 0}$ and $(\{X_{\varepsilon,j}, U_{\varepsilon,j})\}_{j\geq 0}$ shows that $U_j^\star = U_{\varepsilon,j}$ if, and only if, $\mathbf{1}(\xi_k \leq a(X_{\varepsilon,k}, U_{\varepsilon,k}, \varepsilon Z_{x,k}, Z_{y,k})) = \mathbf{1}(\xi_k \leq a(X, U_{\varepsilon,k}, 0, Z_{y,k}))$ for all $0 \leq k \leq j-1$. It readily follows that

$$\begin{aligned}
&\mathbb{P}\big(U_j^\star \neq U_{\varepsilon,j}\big) \\
&\quad = \mathbb{E}\left[1 - \prod_{k=0}^{j-1} \big(1 - \big|a(X_{\varepsilon,k}, U_{\varepsilon,k}, \varepsilon Z_{x,k}, Z_{y,k}) - a(X, U_{\varepsilon,k}, 0, Z_{y,k})\big|\big)\right] \\
&\quad \leq \sum_{k=0}^{j-1} \mathbb{E}\big|a(X_{\varepsilon,k}, U_{\varepsilon,k}, \varepsilon Z_{x,k}, Z_{y,k}) - a(X, U_{\varepsilon,k}, 0, Z_{y,k})\big|,
\end{aligned}$$

where we have made use of the inequality $1 - \prod_{k=0}^{j-1}(1-a_k) \leq \sum_{k=0}^{j-1} a_k$, for sequences $a_k \in [0,1]$. Under Assumptions 1–2, we have that the difference $|a(X_{\varepsilon,k}, U_{\varepsilon,k}, \varepsilon Z_{x,k}, Z_{y,k}) - a(X, U_{\varepsilon,k}, 0, Z_{y,k})|$ is less than a constant multiple of

$$\begin{aligned}
&\varepsilon\big(1 + |X| + |X_{\varepsilon_k}| + |Z_{\varepsilon,k}|\big) \times |Z_{x,k}| + \big(1 + |X| + |X_{\varepsilon_k}| + |U_{\varepsilon,k}| + |Z_{\varepsilon,k}|\big) \\
&\qquad \times |X_{\varepsilon,k} - X|.
\end{aligned}$$

Notice also that due to the RWM chain we have $|X_{\varepsilon,k} - X_{\varepsilon,0}| \lesssim \varepsilon \sum_{l=0}^{k-1} |Z_{x,l}|$. Bringing everything together, we have shown that

$$\mathbb{E}\big|a(X_{\varepsilon,k}, U_{\varepsilon,k}, \varepsilon Z_{x,k}, Z_{u,k}) - a(X, U_{\varepsilon,k}, 0, Z_{u,k})\big| \lesssim k\varepsilon.$$

Therefore, $\mathbb{P}(U_j^\star \neq U_{\varepsilon,j}) \lesssim j^2\varepsilon$. This implies that

$$\varepsilon^\gamma \sum_{j=0}^{\lfloor \varepsilon^{-\gamma} \rfloor - 1} \mathbb{P}\big(U_j^\star \neq U_{\varepsilon,j}\big)^{1/2} \lesssim \varepsilon^{1/2-\gamma}.$$

Since $\gamma \in (0, \frac{1}{2})$, the conclusion follows. $\square$

The combination of Lemmas 2 and 4, together with the ergodic result in (28) and the identity in Lemma 3 provide the proof of equation (23). Thus by Proposition 1, one can conclude that of the sequence of processes $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in the Skorokhod space $\mathbf{D}([0,T], \mathbb{R}^{n_x})$ to the diffusion process (15).

4.2. *Proof of weak convergence of $\widetilde{X}_\varepsilon$ to the limiting diffusion* (15). We have proven in Section 4.1 that the sequence of processes $\{\widetilde{S}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in $\mathbf{D}([0,T], \mathbb{R}^{n_x})$ to the limiting diffusion (15). This also implies that the sequence of processes $\{\widetilde{X}_{\varepsilon,t}\}_{t\in[0,T]}$ converges towards the same limiting diffusion if one can establish that these two processes are close to each other in the sense that

$$\lim_{\varepsilon\to 0} \mathbb{E}\big[\sup\{|\widetilde{X}_{\varepsilon,t} - \widetilde{S}_{\varepsilon,t}| : t \in [0,T]\}\big] = 0. \tag{31}$$

Since the process $\widetilde{S}_{\varepsilon,t}$ is obtained from the sequence $\{X_{\varepsilon,j}\}$ by sub-sampling at rate $\tilde{c}(\varepsilon) \equiv \varepsilon^{-\gamma}$, the triangle inequality yields that the supremum in (31) is less than a constant multiple of

$$\varepsilon \times \sup\left\{ \sum_{j=1}^{\lfloor \tilde{c}(\varepsilon) \rfloor} |Z_{i,j}| : 1 \le i \le \lfloor T c(\varepsilon)/\tilde{c}(\varepsilon) \rfloor \right\} \tag{32}$$

for independent centred and standard Gaussian random variables $\{Z_{i,j}\}_{i,j\ge 0}$ in $\mathbb{R}^{n_x}$. To show that the above quantity converges to 0 in expectation, one can for instance work as follows. We define

$$R_{i,\varepsilon} := \varepsilon \sum_{j=1}^{\lfloor \tilde{c}(\varepsilon) \rfloor} |Z_{i,j}|.$$

Then, for any $\alpha > 0$, Markov's inequality gives $\mathbb{P}(R_{i,\varepsilon}^3 > \alpha^3) \le \mathbb{E}[R_{i,\varepsilon}^3]/\alpha^3 \le C\varepsilon^{3-3\gamma}/\alpha^3$ for a constant $C > 0$. We also define $R_\varepsilon := \sup\{R_{i,\varepsilon} : 1 \le i \le$

$\lfloor T/\varepsilon^{2-\gamma}\rfloor\}$. Simple calculations give

$$\mathbb{E}[R_\varepsilon] = \int_0^\infty \mathbb{P}(R_\varepsilon > \alpha)\,d\alpha = \int_0^\infty \big[1 - \{1 - \mathbb{P}(R_{1,\varepsilon} > \alpha)\}^{\lfloor T/\varepsilon^{2-\gamma}\rfloor}\big]\,d\alpha$$

$$\leq \int_0^\infty \left[1 - \left\{1 - \frac{C}{\alpha^3}\varepsilon^{3-3\gamma}\right\}^{\lfloor T/\varepsilon^{2-\gamma}\rfloor}\right] d\alpha \tag{33}$$

$$= \int_0^\infty \left[1 - \left[\left\{1 - \frac{C}{\alpha^3}\varepsilon^{3-3\gamma}\right\}^{\varepsilon^{-3+3\gamma}}\right]^{\delta(\varepsilon)}\right] d\alpha$$

with $\delta(\varepsilon) = \varepsilon^{3-3\gamma}\cdot\lfloor T/\varepsilon^{2-\gamma}\rfloor$ vanishing in the limit since $\gamma \in (0, 1/2)$. Now, for a $\delta > 0$, we have that for big enough α

$$\left\{1 - \frac{C}{\alpha^3}\varepsilon^{3-3\gamma}\right\}^{\varepsilon^{-3+3\gamma}} \geq e^{-(C/\alpha^3)(1+\delta)}.$$

Using this bound in (33) and then calling upon the dominated convergence theorem proves that $\mathbb{E}[R_\varepsilon] \to 0$ as required.

4.3. *Proof of Theorem 2.* The proof is entirely similar to the proof of Theorem 1. We only describe the modifications necessary to deal with this more general setting. We define the quantities $S_\varepsilon, \widetilde{S}_\varepsilon, \mathcal{L}_\varepsilon\varphi, \widetilde{\mathcal{L}}_\varepsilon\varphi$ the same way as in the proof of Theorem 1. The acceptance probability of the move $(X, U) \to (X + \ell\varepsilon Z_x, U + \ell Z_y)$ reads

$$F \circ \log\left(\frac{\pi_\varepsilon(X', U')p_\varepsilon((X', U') \to (X, U))}{\pi_\varepsilon(X, U)p_\varepsilon((X, U) \to (X', U'))}\right),$$

where $p_\varepsilon[(X, U) \to (X', U')]$ is the likelihood of the move $(X, U) \to (X'U')$. Proposition 1 still holds but the limiting quantity $\mathcal{A}\varphi(x, u) = \lim_{\varepsilon\to 0}\mathcal{L}_\varepsilon\varphi(x, u)$ is now defined as

$$\mathcal{A}\varphi(x, u) = \mathbb{E}\big[F'(DB) \times \{\ell^2(x)\nabla_x\big(A(x) + B(x, u + \ell Z_y)\big) + \nabla_x\ell^2(x)\}\big]\cdot\nabla\varphi(x)$$

$$+ \frac{1}{2}\ell^2(x)\mathbb{E}\big[F(DB)\big]\Delta\varphi(x).$$

The proof uses a Taylor expansion of $\mathcal{L}_\varepsilon\varphi(x, u)$ with Assumptions 1–4 invoked to give a control on the error terms. Under boundedness assumptions on the function $x \mapsto \ell(x)$ the coupling used in the last part of the proof of Theorem 1 is still valid and the rest of the proof then follows exactly the same lines as the proof of Theorem 1.

5. Vanishing acceptance probability. We now consider the scenario where the target distribution π_ε is explored by a RWM algorithm that employs jump proposal of size $\mathcal{O}(1)$; in other words and with the notation of the previous section,

$$h(\varepsilon) = 1.$$

At an heuristic level, as $\varepsilon \to 0$, a proposal $(X, Y) \mapsto (X', Y')$ is accepted only if $|Y'|$ is of order ε, which happens with probability $\mathcal{O}(\varepsilon^{n_y})$. In order to obtain a nontrivial limiting object, one thus needs to accelerate time by a factor ε^{-n_y} and in this case the rescaled RWM trajectories converge, as $\varepsilon \to 0$, to a Markov jump process limit. In particular, we now consider the process $t \mapsto (\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})$ defined as

$$(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t}) = (X_{\varepsilon,\lfloor t\cdot\varepsilon^{-n_y}\rfloor}, U_{\varepsilon,\lfloor t\cdot\varepsilon^{-n_y}\rfloor}), \tag{34}$$

where, as in the previous section, we have used the rescaled coordinate $U_\varepsilon \equiv Y_\varepsilon/\varepsilon$. Thus, in this case we have the proposal $(X', U') = (X + \ell Z_x, U + \ell\varepsilon^{-1} Z_y)$ where (Z_x, Z_y) is a standard Gaussian random variable on $\mathbb{R}^{n_x} \times \mathbb{R}^{n_y}$.

We will again be following [7], starting from [7], Chapter 8, Theorem 3.1 which identifies the core of the generator for the limiting jump process (essentially, the space of test functions) under conditions on the jump rate and transition kernel. Then the proof of weak convergence follows similarly as in the diffusion case earlier. To apply the results in [7], we need to bound the space of u-coordinate, thus we now impose the following condition.

ASSUMPTION 6. Function $A : \mathbb{R}^{n_x} \to \mathbb{R}$ is continuous. Also, function $B = B(x, u)$ appearing in the definition of the target in (2) maps $H := \mathbb{R}^{n_x} \times M \to \mathbb{R}$, for a compact set $M \subset \mathbb{R}^{n_y}$, and is continuous on its domain H.

We redefine the acceptance probability as in (7) with an adjustment for the boundedness of u,

$$\begin{cases} a(x, u, w_x, w_u) \\ \quad = F\big(A(x + \ell w_x) - A(x) + B(x + \ell w_x, u + \ell w_u) - B(x, u)\big) \\ \qquad \text{if } u, u + \ell w_u \in M, \\ 0 \quad \text{otherwise.} \end{cases}$$

We note that smoothness assumptions are not needed in this setting, and F is allowed to be the standard Metropolis–Hastings acceptance probability.

The appropriate space of test functions is now

$$\mathcal{C}' = \{\varphi : H \to \mathbb{R} : \varphi \text{ is continuous with compact support}\}.$$

For $\varphi \in \mathcal{C}'$, the generator of the process $t \mapsto (\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})$ reads as

$$\begin{aligned} \mathcal{G}_\varepsilon \varphi(x, u) &= \varepsilon^{-n_y} \mathbb{E}_{x,u}\big[\big(\varphi\big(x + \ell Z_x, u + \ell\varepsilon^{-1} Z_y\big) - \varphi(x, u)\big) a\big(x, u, Z_x, \varepsilon^{-1} Z_y\big)\big] \\ &= \int_H \big(\varphi(\overline{x}, \overline{u}) - \varphi(x, u)\big) Q(x, u, \overline{x}, \overline{u}) \exp\big\{-\varepsilon^2 |\overline{u} - u|^2/\big(2\ell^2\big)\big\}\, d(\overline{x}, \overline{u}), \end{aligned}$$

where the function $Q(\cdot, \cdot, \cdot, \cdot)$ is given as

$$Q(x, u, \overline{x}, \overline{u}) = \frac{F(A(\overline{x}) - A(x) + B(\overline{x}, \overline{u}) - B(x, u)) \exp\{-|\overline{x} - x|^2/(2\ell^2)\}}{(2\pi\ell^2)^{n/2}}.$$

Notice that the second argument in $\varphi(x+\ell Z_x, u+\ell\varepsilon^{-1}Z_y)$ appearing in the definition of $\mathcal{G}_\varepsilon$ above can take values outside M, but the acceptance probability is then 0, so we have ignored this issue; alternatively, one can redefine $\varphi(x,u)$ for $u\in\mathbb{R}^{n_y}$ and specify it to be 0 when $u\notin M$. We will show that the sequence $t\mapsto(\widetilde{X}_{\varepsilon,t},\widetilde{U}_{\varepsilon,t})$ converges to a Markov jump process $t\mapsto(\widetilde{X}_t,\widetilde{U}_t)$ with transition kernel $K(x,u,\overline{x},\overline{u})$ and jump rate function $r(x,u)$ defined as

$$K(x,u,\overline{x},\overline{u})=Q(x,u,\overline{x},\overline{u})/r(x,u);\qquad r(x,u)=\int_H Q(x,u,\overline{x},\overline{u})\,d(\overline{x},\overline{u}).$$

The Markov process $t\mapsto(\widetilde{X}_t,\widetilde{U}_t)$ can be described as follows: when found at (x,u), the process waits an exponential time with parameter $r(x,u)$ before jumping to the new position $(\overline{x},\overline{u})$ whose density is given by $K(x,u,\overline{x},\overline{u})$. Note that under Assumption 6, $\sup_H r(x,u)<\infty$. The generator $\mathcal{G}$ of this jump process is

$$\mathcal{G}\varphi(x,u)=r(x,u)\int_H\big(\varphi(\overline{x},\overline{u})-\varphi(x,u)\big)K(x,u,\overline{x},\overline{u})\,d(\overline{x},\overline{u}). \tag{35}$$

Using an approach similar to the one of the previous section, one can prove the following result.

THEOREM 3. *Assume that the process $(\widetilde{X}_\varepsilon,\widetilde{U}_\varepsilon)$ is started at time 0 from the equilibrium distribution π. Under Assumption 6, for a fixed horizon $T>0$, the sequence of processes $\{(\widetilde{X}_{\varepsilon,t},\widetilde{U}_{\varepsilon,t}\}_{t\in[0,T]}$ converges weakly in the Skorokhod space $\mathbf{D}([0,T],H)$ to the time-homogeneous jump process $(\widetilde{X}_t,\widetilde{U}_t)$ with generator $\mathcal{G}$ in (35).*

The homogenization argument of the previous section is not necessary since the u-coordinate does not need to be averaged out, so the proof is much simpler. Since time must be accelerated by a factor ε^{-n_y} in order to observe a nontrivial limit, Theorem 3 shows that the algorithmic complexity of RWM algorithm with jump proposal of order $\mathcal{O}(1)$, when used to explore the distribution π_ε, scales as $\mathcal{O}(\varepsilon^{-n_y})$. Note nevertheless that it is not straightforward to optimize the free parameter $\ell>0$ since the limiting Markov jump processes obtained from different values of ℓ are generally not related by a simple linear change of time, as was the case for example in the original article [16].

5.1. *Proof of Theorem 3.* As in the case of the diffusion limit, we relate the convergence in the Shorokhod space with properties of the generators.

PROPOSITION 2. *Let Assumption 6 hold. If the following limit holds,*

$$\lim_{\varepsilon\to0}\mathbb{E}_\pi\big|\widetilde{\mathcal{G}}_\varepsilon\varphi(x,u)-\mathcal{G}\varphi(x,u)\big|=0 \tag{36}$$

for any $\varphi\in\mathcal{C}'$, then as $\varepsilon\to0$ the sequence of processes $\{(\widetilde{X}_{\varepsilon,t},\widetilde{U}_{\varepsilon,t})\}_{t\in[0,T]}$ (started at time 0 from the equilibrium distribution π) converges weakly in the Skorokhod space $\mathbf{D}([0,T],H)$ to the Markov jump process with generator $\mathcal{G}$ in (35).

PROOF. See the Appendix. The proof is similar to the one of Proposition 1. We briefly mention here, that under Assumption 6, [7], Chapter 8, Theorem 3.1, identifies the family $\mathcal{C}'$ as the core for the generator of the limiting jump process. Then [7], Chapter 4, Theorem 8.2, Corollary 8.5 provides the convergence of the finite-dimensional distributions and [7], Chapter 4, Corollary 8.6, the required weak convergence on the path space. $\square$

It remains to prove (36) for any $\varphi \in \mathcal{C}'$. Recall that we have the bound $F(r) \leq F_{MH}(r) = 1 \wedge e^r$, $r \in \mathbb{R}$. We have

$$\begin{aligned}
&\mathbb{E}_\pi \left| \mathcal{G}_\varepsilon \varphi(x,u) - \mathcal{G}\varphi(x,u) \right| \\
&\quad \leq \int_H |D\varphi| \big(\pi(x,u) \wedge \pi(\bar{x},\bar{u})\big) \big(1 - e^{-\frac{\varepsilon^2 |\bar{u}-u|^2}{2\ell^2}}\big) e^{-\frac{|\bar{x}-x|^2}{2\ell^2}} \frac{d(x,u,\bar{x},\bar{u})}{(2\pi\ell^2)^{n/2}} \\
&\quad \lesssim \int_{\overline{\Omega}} \min\big(\pi(x,u), \pi(\bar{x},\bar{u})\big) \big(1 - e^{-\frac{\varepsilon^2 |\bar{u}-u|^2}{2\ell^2}}\big) e^{-\frac{|\bar{x}-x|^2}{2\ell^2}} \, d(x,u,\bar{x},\bar{u}),
\end{aligned}$$

where $D\varphi \equiv \varphi(\bar{x},\bar{u}) - \varphi(x,u)$; we have used the fact that since φ has compact support, say Ω, the norm $\|D\varphi\|_\infty$ is finite and $D\varphi$ is zero outside of $\overline{\Omega} = (H \times \Omega) \cup (\Omega \times H)$. Notice that

$$\int_{\overline{\Omega}} \min\big(\pi(x,u), \pi(\bar{x},\bar{u})\big) \, d(x,u,\bar{x},\bar{u}) \leq 2 \int_\Omega dx\, du \int_H \pi(x,u)\, dx\, du < \infty,$$

so (36) follows from the dominated convergence theorem.

5.2. *Jump process versus diffusion limit.* The general heuristic forming the basis of most diffusion (or jump process, through the work in this paper) limit approaches is the following. Consider a sequence of MCMC algorithms $\{M_k^{(\varepsilon)}\}_{k \geq 0}$ indexed by a parameter ε; if, as $\varepsilon \to 0$, the time-rescaled sequence of Markov chains obtained by accelerating time by a factor $\varepsilon^{-\kappa}$ converges towards a nontrivial limiting process (in this article, either a nontrivial ergodic diffusion or Markov jump process), it can be argued that, in the setting $\varepsilon \to 0$, the Markov chain M^ε requires $\mathcal{O}(\varepsilon^{-\kappa})$ steps to mix. We refer the reader to [19] for rigorous results that form the basis of this approach.

When using the RWM algorithm to explore π_ε, the scaling limit Theorems 2 and 3 reveal that in the case where the dimension of identifiability equals one, that is, $n_y = 1$, it is asymptotically more efficient (as $\varepsilon \to 0$) to apply jump proposals of size $\mathcal{O}(1)$, and thus use an algorithm with vanishing acceptance probability that behaves like a Markov jump process. This is more effective than adopting jump proposals of size $\mathcal{O}(\varepsilon)$ which leads to an acceptance rate bounded away from zero and one; indeed, we have proven that in this case, jumps sizes of order $\mathcal{O}(1)$ lead to an algorithm whose complexity scales as $\mathcal{O}(\varepsilon^{-1})$ whereas jump sizes of order $\mathcal{O}(\varepsilon)$ yield to a complexity that scales as $\mathcal{O}(\varepsilon^{-2})$. The standard rule-of-thumb that advocates tuning the mean acceptance probability of the random walk algorithm to

a given optimal value $a_\star \in (0, 1)$, for example, $a_\star = 0.234$ in the high-dimensional setting of [16] does not generally apply in the setting investigated in this article.

In the case where $n_y = 2$, a similar argument shows that the approaches consisting in setting a jump size of $\mathcal{O}(1)$ or $\mathcal{O}(\varepsilon)$ both yield to algorithms whose complexity scales as $\mathcal{O}(\varepsilon^{-2})$. In the case where the dimensionality of the weak identifiability is large, that is, $n_y \geq 3$, it is preferable to adopt jump proposal sizes of order $\mathcal{O}(\varepsilon)$; the resulting algorithm scales as $\mathcal{O}(\varepsilon^{-2})$ whereas the choice of jump sizes $\mathcal{O}(1)$ scales as $\mathcal{O}(\varepsilon^{-n_y})$.

6. Towards a diffusion limit for a larger class of manifolds. We give here a conjecture for a diffusion limit for the case of a nonlinear manifold. An analytic proof is left for future work. The presentation will sidestep technicalities and will also serve to highlight the main building blocks of the earlier proof of Theorem 1.

We define an n_x-dimensional manifold by assuming existence of an invertible global coordinate chart $r : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x+n_y}$, so we have

$$\mathcal{M} = \{v \in \mathbb{R}^{n_x+n_y} : v = r(x), x \in \mathbb{R}^{n_x}\}.$$

We denote by $T_v\mathcal{M}$ the tangent space of $\mathcal{M}$ at $v = r(x)$. The plane $T_v\mathcal{M}$ is n_x-dimensional with a canonical basis comprised of the linearly independent vectors $\{(\partial r/\partial x_i)(x)\}_{i=1}^{n_x}$. The mapping r gives rise to the metric tensor:

$$G(x) = \big(\langle (\partial r/\partial x_i)(x), (\partial r/\partial x_j)(x)\rangle\big)_{i,j=1}^{n_x} \in \mathbb{R}^{n_x \times n_x}.$$

We will use the standard decomposition in terms of the tangent and its perpendicular normal space $N_v\mathcal{M}$ defined via the linear system:

$$N_v\mathcal{M} = \{w \in \mathbb{R}^{n_x+n_y} : \langle w, (\partial r/\partial x_i)(x)\rangle = 0, 1 \leq i \leq n_x\}. \tag{37}$$

That is, we have $\mathbb{R}^{n_x+n_y} = T_v\mathcal{M} \oplus N_v\mathcal{M}$. Let $(q_i(v))_{i=1}^{n_y}$ denote an orthonormal basis for $N_v\mathcal{M}$. This could for instance be generated after applying Gram–Schmidt iteration on the solutions of the linear system in (37). Some care is needed in the basis construction to ensure smoothness of $v \mapsto q_i(v)$, $1 \leq i \leq n_y$. For $w \in N_v\mathcal{M}$, we denote by $Q_v w$ the ordered coordinates of w w.r.t. the basis $(q_i(v))_{i=1}^{n_y}$. We assume well posedness of the projection $\mathrm{proj}_{\mathcal{M}} : \mathbb{R}^{n_x+n_y} \to \mathcal{M}$, mapping each element of $\mathbb{R}^{n_x+n_y}$ to its closest on $\mathcal{M}$ defined as

$$\mathrm{proj}_{\mathcal{M}}(w) = r \circ \Big\{\arg\min_{x \in \mathbb{R}^{n_x}} |r(x) - w|^2\Big\}.$$

We will need the derivatives for the projection mapping. For a mapping $H : \mathbb{R}^k \to \mathbb{R}^l$, we denote $DH(x) = (\partial H_i/\partial x_j)_{1 \leq i \leq l, 1 \leq j \leq k}$. We have

$$D\big(r^{-1}(\mathrm{proj}_{\mathcal{M}})\big)(v) = G^{-1}(x)\{Dr(x)\}^\top, \tag{38}$$

$$D\mathrm{proj}_{\mathcal{M}}(v) = Dr(x)G^{-1}(x)\{Dr(x)\}^\top. \tag{39}$$

The above can be found by standard Taylor expansion of the distance metric $|r(x) - w|^2$ around its maximiser, akin to the procedure used for proving a CLT for the MLE; see, for instance, Chapter 18 of [8]. For $w \in \mathbb{R}^{n_x+n_y}$, a natural decomposition to be used in this set-up is

$$w = (x, y) \equiv \big(r^{-1}(\mathrm{proj}_{\mathcal{M}}(w)), Q_v(w - \mathrm{proj}_{\mathcal{M}}(w))\big); \qquad x \in \mathbb{R}^{n_x}, y \in \mathbb{R}^{n_y}.$$

The target distribution is assumed to be of the form

$$\pi_\varepsilon(dw) = \pi_\varepsilon(dx, dy) = \frac{1}{\varepsilon^{n_y}} e^{A(x)+B(x,y/\varepsilon)}\, dx\, dy. \tag{40}$$

To standardize, we set $u = y/\varepsilon$. As in (5), we consider the standard RWM on $\mathbb{R}^{n_x+n_y}$ with target $\pi_\varepsilon(x, y)$ and proposal

$$w' = w + \ell\varepsilon Z, \qquad Z \sim \mathrm{N}(0, I_{n_x+n_y}).$$

These dynamics give rise to the RWM trajectory $\{(X_{\varepsilon,k}, Y_{\varepsilon,k})\}_{k\geq 0}$, and the standardised trajectory $U_{\varepsilon,k} = Y_{\varepsilon,k}/\varepsilon$. We have that $w' = (x', y')$ with $x' = r^{-1}(\mathrm{proj}_{\mathcal{M}}(w'))$. A straightforward Taylor expansion using (38)–(39) will give

$$\begin{aligned} x' &= x + \ell\varepsilon J(x)Z + \mathcal{O}(\varepsilon^2); \\ u' &= u + \ell K(x)Z + \mathcal{O}(\varepsilon), \end{aligned}$$

where we have defined

$$\begin{aligned} J(x) &= G^{-1}(x)\big\{Dr(x)\big\}^\top, \\ K(x) &= Q(v)^\top\big(I - Dr(x)G^{-1}(x)\big\{Dr(x)\big\}^\top\big) \end{aligned}$$

for $Q(v) = [q_1(v), \ldots, q_{n_y}(v)]$. We thus get that

$$\begin{aligned} \varphi(x') - \varphi(x) = \big(\ell\varepsilon J(x)Z + \mathcal{O}(\varepsilon^2)\big)\big(\nabla\varphi(x)\big)^\top \\ + \frac{1}{2}\ell^2\varepsilon^2\big\langle J(x)Z, \nabla^2\varphi(x)J(x)Z\big\rangle + \mathcal{O}(\varepsilon^3). \end{aligned}$$

We now turn our attention to the acceptance probability term, and we have

$$\begin{aligned} &F\big(A(x') - A(x) + B(x', u') - B(x, u)\big) \\ &\quad = F\big(B\big(x, u + \ell K(x)Z\big) - B(x, u)\big) + \mathcal{O}(\varepsilon). \end{aligned}$$

Similar to (9), we define the limiting average acceptance probability at position x:

$$a_0(x, \ell) = \int_{\mathbb{R}^{n_y}} \mathbb{E}\big[F\big(B\big(x, u + \ell K(x)Z\big) - B(x, u)\big)\big] e^{B(x,u)}\, du. \tag{41}$$

Following the crux of the analytical proof for the case of affine manifold, we start by looking at the one-step generator:

$$
\begin{aligned}
&\mathcal{L}_\varepsilon \varphi(x,u)\\
&\quad= \mathbb{E}\left[\frac{\varphi(X_{\varepsilon,1})-\varphi(X_{\varepsilon,0})}{\varepsilon^2}\Big| X_{\varepsilon,0}=x, U_{\varepsilon,0}=u\right]\\
&\quad= \mathbb{E}_{x,u}\left[\frac{\varphi(x')-\varphi(x)}{\varepsilon^2}\cdot F\big(A(x')-A(x)+B(x',u')-B(x,u)\big)\right]\\
&\quad= \mathbb{E}_{x,u}\big[\big(\ell\varepsilon^{-1}J(x)Z+\mathcal{O}(1)\big)F\big(B\big(x,u+\ell K(x)Z\big)-B(x,u)\big)\big]\cdot\big(\nabla\varphi(x)\big)^\top\\
&\qquad+\mathbb{E}_{x,u}\left[\frac{1}{2}\ell^2\big\langle J(x)Z,\nabla^2\varphi(x)J(x)Z\big\rangle\cdot F\big(B\big(x,u+\ell K(x)Z\big)-B(x,u)\big)\right]\\
&\qquad+\mathcal{O}(\varepsilon).
\end{aligned}
\tag{42}
$$

Notice that due to the orthogonality of $T_v\mathcal{M}$, $N_v\mathcal{M}$, we have that

$$K(x)^\top J(x)=0,$$

which implies the independence $J(x)Z\perp K(x)Z$. Thus, continuing from (42) we have that

$$
\begin{aligned}
\mathcal{L}_\varepsilon\varphi(x,u)=\big\langle\mathcal{O}(1),\nabla\varphi(x)\big\rangle&\\
+\mathbb{E}_x\left[\frac{1}{2}\ell^2\big\langle J(x)Z,\nabla^2\varphi(x)J(x)Z\big\rangle\right]&\\
\times\mathbb{E}_{x,u}\big[F\big(B\big(x,u+\ell K(x)Z\big)-B(x,u)\big)\big]&\\
+\mathcal{O}(\varepsilon).&
\end{aligned}
\tag{43}
$$

Following closely the affine case, we consider the sped-up process $\widetilde{S}_{\varepsilon,t}=S_{\varepsilon,\lfloor t\cdot c(\varepsilon)/\widetilde{c}(\varepsilon)\rfloor}\equiv S_{\varepsilon,\lfloor t\cdot\varepsilon^{\gamma-2}\rfloor}$ for the sub-sampled trajectory $S_{\varepsilon,k}$, $V_{\varepsilon,k}$. Recall that the idea here is the u-trajectory will have enough time to mix during the sub-sampled times, whereas the x-trajectory will still make local moves and provide a diffusion limit. Thus, we make the following conjecture for the generator $\widetilde{\mathcal{L}}_\varepsilon$ of the process $\widetilde{S}_\varepsilon$:

$$
\begin{aligned}
\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u)&=\frac{1}{\varepsilon^{-\gamma}}\mathbb{E}\left[\sum_{j=0}^{\lfloor\varepsilon^{-\gamma}\rfloor-1}\mathcal{L}_\varepsilon\varphi(X_{\varepsilon,j},U_{\varepsilon,j})|X_{\varepsilon,0}=x,U_{\varepsilon,0}=u\right]\\
&=\mathbb{E}_x\big[\mathcal{L}_\varepsilon\varphi(x,u)\big]+o(1)\\
&=\big\langle\mathcal{O}(1),\nabla\varphi(x)\big\rangle+\mathbb{E}_x\left[\frac{1}{2}\ell^2\big\langle J(x)Z,\nabla^2\varphi(x)J(x)Z\big\rangle\right]\cdot a_0(x,\ell)+o(1).
\end{aligned}
$$

It is easy to check that $\{Dr(x)\}^\top Dr(x)=G(x)$. Thus, the above expression, and in particular the quantity involving $\nabla^2\varphi(x)$, suggests a diffusion limit with diffusion

coefficient σ such that

$$(\sigma\sigma^\top)(x) = G(x)^{-1} a_0(x,\ell)\ell^2,$$

with a corresponding expression for the limiting SDE

$$d\overline{X} = \text{drift}(\pi_X, \sigma\sigma^\top)(\overline{X})\,dt + \sigma(\overline{X})\,dW \tag{44}$$

for $D_0 \sim \exp\{A(x)\}$ and $\text{drift}(\pi_X, \sigma\sigma^\top) = \frac{1}{2}\sigma\sigma^\top\nabla + \frac{1}{2}(\sigma\sigma^\top)\nabla \log \pi_X$, with $\pi_X(x) = \exp\{A(x)\}$. The expression we obtained for the diffusion coefficient is the same as the one for the Langevin SDE on a manifold obtained in [9, 11] with the addition of the average acceptance probability term $a_0(x,\ell)\ell^2$.

7. Conclusions/future work. As far as we are aware, ours is the first attempt towards analytically studying the behaviour and complexity of MCMC algorithms when applied to target densities with a multi-scale structure. We acknowledge here that the practical advice stemming out of our results are probably not as strong as in the case of diffusion limits in high dimensions. Still, we believe that our analysis provides inroads for the investigation of MCMC algorithmic performance in a different direction from the one followed so far in the literature. Our work opens up a number of avenues for future work in this area. We highlight a few of these below:

- In many practical problems, the limiting manifold will be nonlinear and the directions of small size can vary in different parts of the state space, and thus one cannot predetermine narrow directions and adjust the step-sizes. The conjectures about diffusions limits on manifolds in Section 6 thus have immediate impact in applications but might require substantial amount of work to be proved in full generality.
- In a wider perspective, we believe that the results in the paper open new directions also for the study of MCMC algorithms that better exploit the manifold structure of the support of the target distribution; this direction also connects with recent advances in the development of Riemannian MALA and HMC methods as in, for example, [9, 11]. The set-up in our paper is a bit more involved as the manifold can be of smaller dimension that the general space (in the above works it is of the same dimension). To be more explicit, following the notation of Section 6, it would be of interest, for instance, to study RWM with location-specific step-sizes, say of the form

$$\begin{aligned} w' &= w + \ell(\partial r/\partial x_1, \partial r/\partial x_2, \ldots, \partial r/\partial x_{n_x} | q_1, q_2, \ldots, q_{n_y}) \begin{pmatrix} h(\varepsilon) Z_x \\ \varepsilon Z_y \end{pmatrix} \\ &= w + \ell h(\varepsilon) Dr(x) Z_x + \ell\varepsilon Q(v) Z_y, \end{aligned}$$

  so that the method moves along the tangent space $T_v\mathcal{M}$ with a step of size $h(\varepsilon)$ and along the perpendicular normal space with step of size ε (as this is the

size of the probability mass around $\mathcal{M}$). Of interest here would be the optimal selection for $h(\varepsilon)$ and the specification of the computational cost for the method. For example, consider the case when the manifold $\mathcal{M}$ corresponds to a circle of radius 1 in $\mathbb{R}^2$. Then it appears that controlling the acceptance probability would require $h(\varepsilon) = \sqrt{\varepsilon}$, as this is the order of the size of the string of the circle which is perpendicular to a radius at position of distance ε from the circle surface. Such questions could be investigated for general manifold structures. The above of course corresponds to an idealized algorithm, when the method uses explicit information about the tangent and normal space. In practice, it will be of interest to investigate also practical recent algorithms using for instance information about the curvature of the log-target distribution or the Hessian to scale the step-sizes in the different directions, and contrast their effect with the idealised scenario.

- It is of interest to provide connections with locally adaptive methods currently looked at in the literature. Further exploration of such advanced methods in a similar manifold setting may provide analytical results that should be contrasted with the ones here and illustrate their superiority.
- Finally, all of our results assume that the Markov chain is at stationarity. There is a parallel literature on the scaling and the behaviour of MCMC algorithms in the transient phase. The limiting process in the transient phase is usually an ordinary differential equation instead of an SDE. It is natural next step to obtain analyze the transient phase of our algorithm.

APPENDIX: PROOFS

Notation. Recall that we use the expression "$e = e(\varepsilon, x, u) = o_{L_1(\pi)}(1)$" to indicate that for given function $e(\varepsilon, x, u)$ we have $\lim_{\varepsilon \to 0} \mathbb{E}_\pi |e(\varepsilon, x, u)| = 0$.

A.1. Proof of Proposition 1. For clarity, the proof is divided into two main steps. First, one proves that the finite dimensional distributions of $\widetilde{S}_\varepsilon$ converge to those of the limiting diffusion. Second, one proves that the sequence $\widetilde{S}_\varepsilon$ is relatively weakly compact in the appropriate topology.

Convergence of the finite dimensional distributions of $\widetilde{S}_\varepsilon$.

We follow closely [7], Chapter 4, Theorem 8.2 and Corollary 8.5, and apply those results for the set of functions $E = \{(\varphi, \mathcal{L}\varphi) : \varphi \in \mathcal{C}\}$. Under Assumption 3, [7], Chapter 8, Theorem 1.6, gives that the closure of E generates a Feller semigroup $\{T(t)\}$ (corresponding to the solution X of the SDE) on the Banach space L of continuous functions vanishing at infinity. Thus, all conditions at the statement of [7], Chapter 4, Theorem 8.2, are satisfied; it remains to prove part (e) of [7], Chapter 4, Corollary 8.5. Given an arbitrary test function $\varphi \in \mathcal{C}$, we set $f_\varepsilon(x, u) = \varphi(x)$ and $g_\varepsilon(x, u) = \widetilde{\mathcal{L}}_\varepsilon \varphi(x, u)$. We need to prove equations (8.8)–(8.9)

and (8.11) of [7], Chapter 4, Theorem 8.2, [equation (8.10) is trivially satisfied]; that is, one must show that

$$\sup_{\varepsilon>0}\sup_{t\le T}\mathbb{E}\big|\varphi(\widetilde{S}_{\varepsilon,t})\big|<\infty; \tag{45}$$

$$\sup_{\varepsilon>0}\sup_{t\le T}\mathbb{E}\big|\widetilde{\mathcal{L}}_{\varepsilon}\varphi(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})\big|<\infty; \tag{46}$$

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\big(\widetilde{\mathcal{L}}_{\varepsilon}\varphi(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})-\mathcal{L}\varphi(\widetilde{S}_{\varepsilon,t})\big)\prod_{i=1}^{k}h_i(\widetilde{S}_{\varepsilon,t_i})\right]=0, \tag{47}$$

for any $k\ge 1$, any times $0\le t_1<\cdots<t_k\le t\le T$, and any functions h_i that can be assumed to be bounded. Indeed, [7], Chapter 8, Theorem 1.6, shows that the above equations involving the generators and expectations at infinitesimally small increments from instance t can imply convergence over finite times. The first requirement follows from the boundedness of φ; the second requirement is implied by the stationarity of $(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})$, equation (23) and the fact that $\mathbb{E}_{\pi}|\mathcal{L}\varphi(x)|<\infty$ [we have that $\sup_{x\in\mathbb{R}^{n_x}}|\mathcal{L}\varphi(x)|<\infty$ from the boundedness of the gradient of the drift function of the limiting diffusion on compact domains, since it is continuous from Assumption 3]. The third requirement is also implied from (23). We have now proven the required convergence of the finite dimensional distributions of $\widetilde{S}_{\varepsilon,t}$ to those of the solution of the limiting SDE (15).

Relative weak compactness of $\widetilde{S}_{\varepsilon}$. We follow [7], Chapter 4, Corollary 8.6. First, we remark that the process $\widetilde{S}_{\varepsilon}$ is started at stationarity and the space $\mathcal{C}\subset L$ of smooth functions with compact support is an algebra that strongly separates points. As noted in the proof of [7], Chapter 4, Corollary 8.5, the pair $(f_{\varepsilon}(\widetilde{S}_{\varepsilon,t}),g_{\varepsilon}(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t}))$, with $f_{\varepsilon},g_{\varepsilon}$ as defined above, in general does not belong to the approximate generator defined in equation (8.6) of [7], Chapter 4, and one needs to consider instead the pair

$$\big(f_{\varepsilon}(\widetilde{S}_{\varepsilon,t})+\big(t-\lfloor\varepsilon^{\gamma-2}t\rfloor/\varepsilon^{\gamma-2}\big)g_{\varepsilon}(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t}),g_{\varepsilon}(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})\big)$$

to account for the fact that the process $(X_{\varepsilon},Y_{\varepsilon})$ is a discrete time Markov chain (note here the typo in equation (8.28) of [7], Chapter 4; the correct term involves the quantity $t-\lfloor\alpha_n t\rfloor/\alpha_n$). Since $(t-\lfloor\varepsilon^{\gamma-2}t\rfloor/\varepsilon^{\gamma-2})<\varepsilon^{2-\gamma}$, to prove equations (8.33) and (8.34) of [7], Chapter 4, Corollary 8.6, we must show that, for some exponent $p>1$, some $\varepsilon_0>0$ and for $(X,U)\sim\pi$, we have

$$\begin{aligned}&\sup_{\varepsilon\in(0,\varepsilon_0)}\big\|\widetilde{\mathcal{L}}_{\varepsilon}\varphi(X,U)\big\|_p<\infty;\\ &\lim_{\varepsilon\to 0}\varepsilon^{2-\gamma}\cdot\mathbb{E}\Big[\sup_{t\in[0,T]}\big|\widetilde{\mathcal{L}}_{\varepsilon}\varphi(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})\big|\Big]=0.\end{aligned} \tag{48}$$

These will imply that the sequence $\widetilde{S}_{\varepsilon}$ is relatively weakly compact in the Skorokhod topology. Note that we have exploited the fact that the algorithm is started

at stationarity. For the first result in (48), we choose $1 < p < 2$. Then the telescoping expansion (24) yields that

$$\begin{aligned}\big\|\widetilde{\mathcal{L}}_\varepsilon\varphi(X,U)\big\|_p &\le \frac{1}{\varepsilon^{-\gamma}}\sum_{j=0}^{\lfloor\varepsilon^{-\gamma}\rfloor-1}\big\|\mathcal{L}_\varepsilon\varphi(X_{\varepsilon,j},U_{\varepsilon,j})-\mathcal{A}\varphi(X_{\varepsilon,j},U_{\varepsilon,j})\big\|_p\\ &\quad+\frac{1}{\varepsilon^{-\gamma}}\sum_{j=0}^{\lfloor\varepsilon^{-\gamma}\rfloor-1}\big\|\mathcal{A}\varphi(X_{\varepsilon,j},U_{\varepsilon,j})\big\|_p\\ &\lesssim \big\|\mathcal{L}_\varepsilon\varphi(X,U)-\mathcal{A}\varphi(X,U)\big\|_p+\big\|\mathcal{A}\varphi(X,U)\big\|_p.\end{aligned}$$

In the second line, we have exploited the fact that the RWM chain is started at stationarity. The required result follows immediately from the given upper bound in Proposition 1 and Assumption 2. For the second result in (48), since the process $t \mapsto \widetilde{\mathcal{L}}_\varepsilon\varphi(\widetilde{S}_{\varepsilon,t},\widetilde{V}_{\varepsilon,t})$ makes at most $\lfloor T/\varepsilon^{2-\gamma}\rfloor$ jumps on the interval $t\in[0,T]$, it suffices to show that each jump is $o(1)$ in L_1-norm; equivalently, due to the stationarity assumption, one needs to prove that the expectation $\|\widetilde{\mathcal{L}}_\varepsilon\varphi(S_{\varepsilon,1},V_{\varepsilon,1})-\widetilde{\mathcal{L}}_\varepsilon\varphi(S_{\varepsilon,0},V_{\varepsilon,0})\|_1$ converges to zero as $\varepsilon\to 0$. Adding and subtracting $\mathcal{L}\varphi(S_{\varepsilon,1})-\mathcal{L}\varphi(S_{\varepsilon,0})$, equation (23) and the stationarity assumption yield that

$$\big\|\widetilde{\mathcal{L}}_\varepsilon\varphi(S_{\varepsilon,1},V_{\varepsilon,1})-\widetilde{\mathcal{L}}_\varepsilon\varphi(S_{\varepsilon,0},V_{\varepsilon,0})\big\|_1\le o(1)+\big\|\mathcal{L}\varphi(S_{\varepsilon,1})-\mathcal{L}\varphi(S_{\varepsilon,0})\big\|_1.$$

Under Assumption 3, for a smooth and compactly test function φ, the function $x\mapsto\mathcal{L}\varphi(x)$ is μ-Holderian so that it suffices to show that $\|S_{\varepsilon,1}-S_{\varepsilon,0}\|_1$ converges to zero as $\varepsilon\to 0$; this is immediate since

$$\|S_{\varepsilon,1}-S_{\varepsilon,0}\|_1\lesssim\varepsilon\sum_{j=1}^{\lfloor\varepsilon^{-\gamma}\rfloor}\|Z_{x,j}\|_1\lesssim\varepsilon^{1-\gamma}$$

and γ was chosen inside the interval $(0,\frac{1}{2})$.

A.2. Proof of Lemma 1. Recall the definition of the one-step generator $\mathcal{L}_\varepsilon$ in (22). Using the notation $\mathsf{v}=\nabla\varphi(x)\in\mathbb{R}^{n_x}$, $\mathsf{S}=\nabla^2\varphi(x)\in\mathbb{R}^{n_x\times n_x}$, a second-order Taylor expansion yields

$$\begin{aligned}\mathcal{L}_\varepsilon\varphi(x)&=\varepsilon^{-2}\mathbb{E}_{\varepsilon,x,u}\big[\big(\varphi(x+\ell\varepsilon Z_x)-\varphi(x)\big)\times a(x,u,\varepsilon Z_x,Z_y)\big]\\ &=\ell\varepsilon^{-1}\mathbb{E}_{\varepsilon,x,u}\big[\langle\mathsf{v},Z_x\rangle\times a(x,u,\varepsilon Z_x,Z_y)\big]\\ &\quad+\frac{1}{2}\ell^2\mathbb{E}_{\varepsilon,x,u}\big[\langle Z_x,\mathsf{S}Z_x\rangle\times a(x,u,\varepsilon Z_x,Z_y)\big]+o_{L_1(\pi)}(1),\end{aligned}$$

where the remainder term has been identified as $o_{L_1(\pi)}(1)$ for $\varepsilon\to 0$ as its absolute value is upper bounded by $C\varepsilon\mathbb{E}|Z_x|^3$, for a constant $C>0$, due to φ being smooth

and of compact support. Thus, to prove the stated limiting result, it suffices to prove that the following two identities hold:

$$\begin{aligned}&\varepsilon^{-1}\mathbb{E}_{\varepsilon,x,u}\big[\langle \mathsf{v}, Z_x\rangle \times a(x,u,\varepsilon Z_x, Z_y)\big]\\&\quad=\ell\mathbb{E}_{x,u}\big[F'(DB)\big\langle \mathsf{v}, \nabla A(x)+\nabla_x B(x,u+\ell Z_y)\big\rangle\big]\\&\qquad+o_{L_1(\pi)}(1),\end{aligned}\tag{49}$$

$$\begin{aligned}&\mathbb{E}_{\varepsilon,x,u}\big[\langle Z_x, \mathsf{S}Z_x\rangle \times a(x,u,\varepsilon Z_x, Z_y)\big]\\&\quad=\operatorname{Tr}(\mathsf{S})\times\mathbb{E}_{x,u}\big[F(DB)\big]+o_{L_1(\pi)}(1).\end{aligned}\tag{50}$$

Recall the shorthand notation $DB = B(x,u+\ell Z_y) - B(x,u)$. Expression (7) for the acceptance probability function $a(\cdot,\cdot,\cdot,\cdot)$ together with regularity Assumption 2 on functions A and B yield

$$\begin{aligned}a(x,u,\varepsilon Z_x, Z_y) = F(DB) + \ell\varepsilon F'(DB)\big\langle&\nabla A(x)\\&+\nabla_x B(x,u+\ell Z_y), Z_x\big\rangle+\varepsilon\times o_{L_1(\pi)}(1).\end{aligned}\tag{51}$$

The remainder term has been identified as $\varepsilon\times o_{L_1(\pi)}(1)$ as it is upper bounded in absolute value by $C\varepsilon^2(1+|x|+|u|+|Z_x|+|Z_y|)\times|Z_x|$, for a constant $C>0$, due to F' being bounded and Lipschitz, and $\nabla A, \nabla B$ being Lipschitz; also, π has finite absolute first moments. Using this expression gives that the quantity $\varepsilon^{-1}\mathbb{E}_{\varepsilon,x,u}[\langle \mathsf{v}, Z_x\rangle \times a(x,u,\varepsilon Z_x, Z_y)]$ equals

$$\ell\mathbb{E}_{x,u}\big[\langle \mathsf{v}, Z_x\rangle \times F'(DB)\times\big\langle\nabla A(x)+\nabla_x B(x,u+\ell Z_y), Z_x\big\rangle\big]+o_{L_1(\pi)}(1),$$

which gives immediately (49) after taking the expectation over Z_x. Also, (50) follows immediately from (51). Finally, the stated upper bound in the proposition follows immediately from the explicit upper bounds given above for the remainder terms.

A.3. Proof of Lemma 2. Under Assumptions 1–2 (we require that F, F' are bounded and Lipschitz, $\nabla A, \nabla B$ are Lipschitz), for a smooth and compactly supported test function φ, one can verify that

$$\big|\mathcal{A}\varphi(x,u)-\mathcal{A}\varphi(\overline{x},u)\big|\lesssim\big(1+|x|+|\overline{x}|+|u|\big)|x-\overline{x}|\tag{52}$$

for any $\overline{x}, x\in\mathbb{R}^{n_x}$ and $u\in\mathbb{R}^{n_y}$. We now make use of the telescoping equation (24) to get that

$$\begin{aligned}&\mathbb{E}_\pi\left|\widetilde{\mathcal{L}}_\varepsilon\varphi(x,u)-\frac{1}{\varepsilon^{-\gamma}}\sum_{j=0}^{\lfloor\varepsilon^{-\gamma}\rfloor-1}\mathbb{E}_{\varepsilon,x,u}\big[\mathcal{A}\varphi(x,U_{\varepsilon,j})\big]\right|\\&\quad\le\mathbb{E}_\pi\big|\mathcal{L}_\varepsilon\varphi(x,u)-\mathcal{A}\varphi(x,u)\big|\\&\qquad+\frac{1}{\varepsilon^{-\gamma}}\sum_{j=0}^{\lfloor\varepsilon^{-\gamma}\rfloor-1}\mathbb{E}\big|\mathcal{A}\varphi(X_{\varepsilon,j},U_{\varepsilon,j})-\mathcal{A}\varphi(X_{\varepsilon,0},U_{\varepsilon,j})\big|.\end{aligned}\tag{53}$$

We consider the last term. Using (52) together with Cauchy–Schwarz, and the RWM upper bound $|X_{\varepsilon,j} - X_{\varepsilon,0}| \lesssim \varepsilon \sum_{k=1}^{j} |Z_{x,k}|$, we obtain

$$\mathbb{E}\big|\mathcal{A}\varphi(X_{\varepsilon,j}, U_{\varepsilon,j}) - \mathcal{A}\varphi(X_{\varepsilon,0}, U_{\varepsilon,j})\big| \lesssim \varepsilon \sum_{k=1}^{j} \|Z_{x,k}\|_2 = \mathcal{O}(j\varepsilon).$$

Thus, the last term in (53) is $\mathcal{O}(\varepsilon^{1-\gamma})$ and vanishes in the limit since $\gamma < 1$. The proof is complete.

A.4. Proof of Lemma 3. To keep the exposition as simple as possible, we suppose that $\ell = 1$ and $n_x = n_y = 1$. The multi-dimensional case is entirely similar. The proof of (29) consists in verifying that for all $x \in \mathbb{R}$ the following identity holds:

$$\int_{u\in\mathbb{R}} \mathcal{A}\varphi(x,u) e^{B(x,u)}\, du = \mathcal{L}\varphi(x), \tag{54}$$

where $\mathcal{L}$ is the generator of the limiting diffusion (15), $\varphi \in \mathcal{C}$ is a test function in the core of $\mathcal{L}$ and $\mathcal{A}\varphi(x,u)$ reads

$$\begin{aligned}\mathcal{A}\varphi(x,u) &= \mathbb{E}\big[F'(DB)\big(A'(x) + \partial_x B(x, u + Z_y)\big)\big]\varphi'(x) \\ &\quad + \frac{1}{2}\mathbb{E}\big[F(DB)\big]\varphi''(x),\end{aligned} \tag{55}$$

where $DB = B(x, u + Z_y) - B(x,u)$ and $Z_y \sim \mathrm{N}(0,1)$. The proof is a routine calculation that is based on the symmetry of the Gaussian distribution and the fact that the accept–reject function F verifies the reversibility condition (6). More specifically, the derivative of equation (6) also reads

$$F(r) = F'(r) + e^r F'(-r). \tag{56}$$

This identity also holds for the standard MH function $F_{\mathrm{MH}}(r) = \min(1, e^r)$ but has to be interpreted in the sense of distributions. In the scalar case $n_x = 1$ with $\ell = 1$, the generator of (15) reads

$$\mathcal{L}\varphi(x) = \frac{1}{2}\big(a_0(x)A'(x) + a_0'(x)\big)\varphi'(x) + \frac{1}{2}a_0(x)\varphi''(x),$$

where $a_0(x) \equiv a_0(x,1)$ is the mean acceptance probability $a_0(x) = \int_{u\in\mathbb{R}} \mathbb{E}[F(DB)]e^{B(x,u)}\, du$. To prove (54), it suffices to verify that

$$\mathbb{E}\big[F'(DB)\partial_x B(x, u+Z)\big] = \frac{1}{2}a_0'(x); \qquad \mathbb{E}\big[F'(DB)\big] = \frac{1}{2}a_0(x). \tag{57}$$

Let us prove that the first identity holds. Assumption 2 justify the following derivation under the integral sign:

$$\begin{aligned}\partial_x a_0(x) = \int \mathbb{E}\big[&F'(DB)\big(\partial_x B(x, u+Z_y) - \partial_x B(x,u)\big) \\ &+ F(DB)\partial_x B(x,u)\big]e^{B(x,u)}\, du.\end{aligned}$$

Equation (56) shows that $F(DB) = F'(DB) + F(-DB)e^{DB}$; since $e^{DB}e^{B(x,u)} = e^{B(x,u+Z_y)}$, we have

$$\partial_x a_0(x) = \int \mathbb{E}\big[F'(DB)\partial_x B(x, u + Z_y)e^{B(x,u)}\big]\,du + \int \mathbb{E}\big[F'(-DB)\partial_x B(x, u)e^{B(x,u+Z_y)}\big]\,du.$$

The symmetry of the Gaussian distribution $Z_y \sim \mathrm{N}(0, 1)$ then shows that

$$\int \mathbb{E}\big[F'(DB)\partial_x B(x, u + Z_y)e^{B(x,u)}\big]\,du = \int \mathbb{E}\big[F'(-DB)\partial_x B(x, u)e^{B(x,u+Z_y)}\big]. \tag{58}$$

This concludes the proof of the first identity of (57). The proof of the second identity is similar and even simpler, and thus omitted.

A.5. Proof of Proposition 2. The proof is very similar to the one of Proposition 1. Thus, as a first step, we prove that the finite dimensional distributions of $(\widetilde{X}_\varepsilon, \widetilde{U}_\varepsilon)$ converge to those of $(\widetilde{X}, \widetilde{U})$. The proof is then concluded by proving that the sequence $(\widetilde{X}_\varepsilon, \widetilde{U}_\varepsilon)$ is weakly compact in the appropriate topology. Before the above, we work with [7], Chapter 8, Theorem 3.1, identifying $\mathcal{C}'$ as the core of the generator of the limiting jump process. This requires us to prove equations (3.3)–(3.5) at the statement of Theorem 3.1. Avoiding too many details, we proceed as follows: (i) equation (3.3) follows trivially due to the boundedness of the domain of the u-coordinate; equations (3.4), (3.5) also follow immediately when someone specifies the γ and η functions appearing in the statement as $\gamma(x, u) = \eta(x, u) = \exp\{|x|^2/(2c)\}$ with $c > \ell$ (this choice allows for analytical evaluation of the corresponding integrals once upped-bounding the acceptance probability by 1).

Convergence of the finite dimensional distributions of $(\widetilde{X}_\varepsilon, \widetilde{U}_\varepsilon)$. [7], Chapter 8, Theorem 3.1, implies that all conditions at the statement of [7], Chapter 4, Theorem 8.2, are satisfied. We need only prove part (e) of [7], Chapter 4, Corollary 8.5. Given a test function $\varphi \in \mathcal{C}'$, we set $f_\varepsilon(x, u) = \varphi(x)$ and $g_\varepsilon(x, u) = \mathcal{G}_\varepsilon\varphi(x, u)$. We must now prove equations (8.8)–(8.9) and (8.11) of [7], Chapter 4, Theorem 8.2, which are now expressed as

$$\sup_{\varepsilon>0}\sup_{t\le T}\mathbb{E}\big|\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})\big| < \infty;$$

$$\sup_{\varepsilon>0}\sup_{t\le T}\mathbb{E}\big|\mathcal{G}_\varepsilon\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})\big| < \infty;$$

$$\lim_{\varepsilon\to 0}\mathbb{E}\left[\big(\mathcal{G}_\varepsilon\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t}) - \mathcal{G}\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})\big)\prod_{i=1}^{k} h_i(\widetilde{X}_{\varepsilon,t_i}, \widetilde{U}_{\varepsilon,t_i})\right] = 0,$$

for any $\varphi \in \mathcal{C}'$, any $k \geq 1$, times $0 \leq t_1 < \cdots < t_k \leq t \leq T$, and bounded functions h_i. These are very similar to (45)–(47) in the proof of Proposition 1. The first equation follows immediately due to stationarity of process; for the second we add and subtract $\mathcal{G}\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})$ inside the absolute value, so the result follows from (23) and the fact that $\mathbb{E}_\pi|\mathcal{G}\varphi(x,u)| < \infty$; the last equation follows immediately from (23). (In both of the last equations, we also used the stationarity of the processes.)

Relative weak compactness of $(\widetilde{X}_\varepsilon, \widetilde{U}_\varepsilon)$. As in the proof of Proposition 1 and (48), we follow [7], Chapter 4, Corollary 8.6, and have to prove that for some $\varepsilon_0 > 0$, $p > 1$,

$$\begin{aligned} &\sup_{\varepsilon\in(0,\varepsilon_0)} \big\|\mathcal{G}_\varepsilon\varphi(X,U)\big\|_p < \infty; \\ &\lim_{\varepsilon\to 0} \varepsilon^{n_y} \cdot \mathbb{E}\Big[\sup_{t\in[0,T]} |\mathcal{G}_\varepsilon\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})|\Big] = 0. \end{aligned} \tag{59}$$

Following the definitions in the main text, we have that Ω is a compact set outside which φ is zero and $\overline{\Omega} = (H \times \Omega) \cup (\Omega \times H)$. For the first equation in (59), we have that, since the density $\pi(x,y) = e^{A(x)+B(x,y)}$ is strictly positive and continuous,

$$\begin{aligned} &\big\|\mathcal{G}_\varepsilon\varphi(X,U)\big\|_p^p \\ &\quad\leq \frac{1}{(2\pi\ell^2)^{np/2}} \\ &\qquad\times \int_H \bigg\{\int_H |D\varphi| \min\bigg(\frac{1}{\pi(\overline{x},\overline{u})}, \frac{1}{\pi(x,u)}\bigg)\pi(\overline{x},\overline{u})\, d(\overline{x},\overline{u})\bigg\}^p \pi(x,u)\, d(x,u) \\ &\quad\leq \frac{1}{(2\pi\ell^2)^{np/2}} 2^p \|\varphi\|_\infty^p \\ &\qquad\times \int_{\overline{\Omega}} \min\bigg(\frac{1}{\pi(\overline{x},\overline{u})}, \frac{1}{\pi(x,u)}\bigg)^p \pi(\overline{x},\overline{u})\pi(x,u)\, d(x,u,\overline{x},\overline{u}) \\ &\quad\leq \frac{1}{(2\pi\ell^2)^{np/2}} 2^p \|\varphi\|_\infty^p \sup\bigg\{\min\bigg(\frac{1}{\pi(\overline{x},\overline{u})}, \frac{1}{\pi(x,u)}\bigg)^p : (x,u,\overline{x},\overline{u}) \in \overline{\Omega}\bigg\} \\ &\quad= \frac{1}{(2\pi\ell^2)^{np/2}} 2^p \|\varphi\|_\infty^p \sup\big\{\pi(x,u)^{-p} : (x,u) \in \Omega\big\} < \infty, \end{aligned}$$

which proves the required statement. For the second equation in (59), recall that $(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})$ is piece-wise constant on intervals of length ε^{n_y}, and attempts $\lfloor T/\varepsilon^{n_y} \rfloor$ transitions on $[0,T]$. We can thus bound, due to stationarity,

$$\mathbb{E}\Big[\sup_{t\in[0,T]} |\mathcal{G}_\varepsilon\varphi(\widetilde{X}_{\varepsilon,t}, \widetilde{U}_{\varepsilon,t})|\Big] \lesssim \varepsilon^{-n_y} \cdot \big\|\mathcal{G}_\varepsilon\varphi(X_{\varepsilon,1}, U_{\varepsilon,1}) - \mathcal{G}_\varepsilon\varphi(X_{\varepsilon,0}, U_{\varepsilon,0})\big\|_1.$$

Thus, due to equation (36), it suffices to show that $\|\mathcal{G}\varphi(X_{\varepsilon,1}, U_{\varepsilon,1}) - \mathcal{G}\varphi(X_{\varepsilon,0}, U_{\varepsilon,0})\|_1$ converges to zero as $\varepsilon \to 0$. The definition of $\mathcal{G}$ yields that for

any bounded test function φ the function $(x,u) \mapsto \mathcal{G}\varphi(x,u)$ is also bounded. Consequently,

$$\left|\mathcal{G}\varphi(X_{\varepsilon,1},U_{\varepsilon,1}) - \mathcal{G}\varphi(X_{\varepsilon,0},U_{\varepsilon,0})\right| \lesssim \mathbf{1}\big((X_{\varepsilon,1},U_{\varepsilon,1}) \neq (X_{\varepsilon,0},U_{\varepsilon,0})\big).$$

Since, as $\varepsilon \to 0$, the probability that $(X_{\varepsilon,1},U_{\varepsilon,1})$ is different from $(X_{\varepsilon,0},U_{\varepsilon,0})$ converges to zero, the conclusion follows.

Acknowledgements. We thank Sam Livingstone at UCL for several useful conservations on the topic. We thank the referees and the Associate Editor for several suggestions that have improved the content of the paper.

REFERENCES

[1] BARKER, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics* **18** 119–134.

[2] BÉDARD, M. (2007). Weak convergence of Metropolis algorithms for non-i.i.d. target distributions. *Ann. Appl. Probab.* **17** 1222–1244. MR2344305

[3] BÉDARD, M. and ROSENTHAL, J. S. (2008). Optimal scaling of Metropolis algorithms: Heading toward general target distributions. *Canad. J. Statist.* **36** 483–503. MR2532248

[4] BESKOS, A., ROBERTS, G. and STUART, A. (2009). Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions. *Ann. Appl. Probab.* **19** 863–898. MR2537193

[5] BREYER, L. A., PICCIONI, M. and SCARLATTI, S. (2004). Optimal scaling of MaLa for nonlinear regression. *Ann. Appl. Probab.* **14** 1479–1505. MR2071431

[6] COTTER, S. L., DASHTI, M., ROBINSON, J. C. and STUART, A. M. (2009). Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Probl.* **25** 115008, 43. MR2558668

[7] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. Wiley, New York. MR0838085

[8] FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London. MR1699953

[9] GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492

[10] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881

[11] LIVINGSTONE, S. and GIROLAMI, M. (2014). Information-geometric Markov chain Monte Carlo methods using diffusions. *Entropy* **16** 3074–3102. MR3234224

[12] MATTINGLY, J. C., PILLAI, N. S. and STUART, A. M. (2012). Diffusion limits of the random walk Metropolis algorithm in high dimensions. *Ann. Appl. Probab.* **22** 881–930. MR2977981

[13] PASARICA, C. and GELMAN, A. (2010). Adaptively scaling the Metropolis algorithm using expected squared jumped distance. *Statist. Sinica* **20** 343–364. MR2640698

[14] PILLAI, N. S., STUART, A. M. and THIÉRY, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.* **22** 2320–2356. MR3024970

[15] RADDE, N. E. and OFFTERMATT, J. (2014). Convergence of posteriors for structurally non-identified problems using results from the theory of inverse problems. *J. Inverse Ill-Posed Probl.* **22** 251–276. MR3187940

[16] ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751

[17] ROBERTS, G. O. and ROSENTHAL, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statist. Sci.* **16** 351–367. MR1888450

[18] ROBERTS, G. O. and ROSENTHAL, J. S. (2014). Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *Ann. Appl. Probab.* **24** 131–149. MR3161644

[19] ROBERTS, G. O. and ROSENTHAL, J. S. (2016). Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits. *J. Appl. Probab.* **53** 410–420. MR3514287

[20] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

A. BESKOS
DEPARTMENT OF STATISTICAL SCIENCE
UNIVERSITY COLLEGE LONDON
1-19 TORRINGTON PLACE
LONDON, WC1E 7HB
UNITED KINGDOM
E-MAIL: a.beskos@ucl.ac.uk

G. ROBERTS
DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY, CV4 7AL
UNITED KINGDOM
E-MAIL: gareth.o.roberts@warwick.ac.uk

A. THIERY
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
6 SCIENCE DRIVE 2
SINGAPORE, 117546
SINGAPORE
E-MAIL: a.h.thiery@nus.edu.sg

N. PILLAI
DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
1 OXFORD STREET
BOSTON, MASSACHUSETTS 02138-2901
USA
E-MAIL: pillai@fas.harvard.edu