THEORETICAL PROPERTIES OF QUASI-STATIONARY MONTE CARLO METHODS

BY ANDI Q. WANG[*,1], MARTIN KOLB[†], GARETH O. ROBERTS[‡,2] AND
DAVID STEINSALTZ[*]

University of Oxford[*], *Paderborn University*[†] *and University of Warwick*[‡]

This paper gives foundational results for the application of quasi-stationarity to Monte Carlo inference problems. We prove natural sufficient conditions for the quasi-limiting distribution of a killed diffusion to coincide with a target density of interest. We also quantify the rate of convergence to quasi-stationarity by relating the killed diffusion to an appropriate Langevin diffusion. As an example, we consider in detail a killed Ornstein–Uhlenbeck process with Gaussian quasi-stationary distribution.

1. Introduction.

1.1. *Background.* Markov chain Monte Carlo (MCMC) is a staple tool for statisticians wishing to perform Bayesian inference. Suppose we wish to sample approximately from the distribution π. The celebrated Metropolis–Hastings algorithm constructs an irreducible, aperiodic Markov chain $(Y_n)_{n=1}^\infty$ that is reversible with respect to π, hence has π as its stationary distribution. General theory of Markov chains tells us that the distribution of Y_n converges to π as $n \to \infty$. The computations may, however, be intractable for large datasets and high-dimensional models, such as modern 'Big Data' applications often demand: for a dataset of size N, merely evaluating the posterior distribution, of the form

$$(1.1) \qquad \pi(x) \propto \prod_{i=1}^N f_i(x),$$

is an expensive $O(N)$ computation at each Markov chain iteration.

In [20], the authors proposed the Scalable Langevin Exact (ScaLE) algorithm as part of a new Monte Carlo framework that is provably efficient for Big-Data Bayesian inference. Starting with a diffusion $(X_t)_{t \geq 0}$ (in their case, a Brownian motion), a stopping time τ_∂, the 'killing time', is defined in such a way that the

quasi-limiting distribution (sometimes termed the *Yaglom* limit) is π. That is, we have convergence of the conditional laws

$$(1.2) \qquad \mathbb{P}_x(X_t \in \cdot \,|\, \tau_\partial > t) \to \pi(\cdot) \qquad \text{as } t \to \infty$$

in an appropriate sense from any starting point $X_0 = x \in \mathbb{R}^d$. Such a π is also *quasi-stationary*, in the sense that

$$(1.3) \qquad \mathbb{P}_\pi(X_t \in \cdot \,|\, \tau_\partial > t) = \pi(\cdot)$$

for all $t \geq 0$, where $\mathbb{P}_\pi$ denotes the law of the process conditional on $X_0 \sim \pi$. Any Monte Carlo procedure which aims to sample from a quasi-stationary distribution, for instance, using (1.2), will be termed a *quasi-stationary Monte Carlo method*.

Quasi-stationarity has long been a subject of intensive study in the probability literature, summarised recently in [7] and the bibliography of [19]. However, the ScaLE algorithm is the first application of quasi-limiting convergence to Monte Carlo sampling. Its attractiveness to the aforementioned 'Big Data' problems stems from the fact that the ScaLE algorithm can be implemented in substantially less than $O(N)$ computing time (per unit stochastic process time). In fact, the algorithm is sometimes $O(1)$ and typically no worse than $O(\log(N))$. This is because the simulation of killed diffusions can be performed perfectly through *subsampling* [usually using subsets of size $O(1)$] and, therefore, without any bias. Direct approaches based around subsampling a random subset of the N terms in (1.1) to obtain an estimate of the product have been proposed, although this results in unacceptably large errors in the target distribution unless the subset itself is $O(N)$; see, for instance, the discussions in [1].

Pollock et al. [20] gives some theory for the convergence properties of ScaLE, although this requires various regularity conditions which are difficult to check in many realistic statistical contexts. Our paper will give a much more complete picture under substantially weaker regularity conditions, and help to link quasi-stationary Monte Carlo methods with the established literature on quasi-stationarity.

Quasi-stationary convergence differs in important respects from the more familiar theory of stationary convergence. For a start, the theory of MCMC algorithms is most commonly formulated for discrete-time chains, whereas the ScaLE algorithm is fundamentally a continuous-time algorithm. There may be many probabilities π which satisfy (1.3), despite irreducibility, so we need to identify the appropriate candidate for the limit (1.2). Perhaps most significant, the conditioned laws in (1.2) are not *consistent*: they are not the marginal laws of a single Markov process at time t. This prevents us from using much of the standard probabilistic armamentarium based on conditioning and the Markov property. Instead, to prove convergence we use R. Tweedie's R-theory, [25], and to study rates of convergence we follow the approach pioneered by [14], drawing on the theory of semi-groups generated by linear differential operators.

1.2. *Summary of main results*.   We now summarise our main results, leaving
the exact mathematical setting to be explicated in Section 3. We will be assuming
throughout the following.

ASSUMPTION 0. π is a positive, smooth and integrable function on $\mathbb{R}^d$.

Consider the d-dimensional diffusion $X = (X_t)_{t \geq 0}$, defined as the (weak) solu-
tion of the stochastic differential equation (SDE)

$$(1.4) \qquad \mathrm{d}X_t = \nabla A(X_t) \, \mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = x \in \mathbb{R}^d,$$

where W is a standard d-dimensional Brownian motion and ∇ denotes the gradient
operator. We require the following.

ASSUMPTION 1. $A : \mathbb{R}^d \to \mathbb{R}$ is a smooth function such that the SDE (1.4)
has a unique nonexplosive weak solution.

Suppose we wish to sample from a distribution π on $\mathbb{R}^d$ with a Lebesgue den-
sity, which we will also denote by π—the *target density*—satisfying Assump-
tion 0. We are typically thinking of applications in which we have a statistical
model and observed data for which π is the Bayesian posterior distribution. We
would like to construct a *killing rate* $\kappa : \mathbb{R}^d \to [0, \infty)$ that makes π into the quasi-
limiting distribution of the diffusion X. That is, we define the killing time

$$(1.5) \qquad \tau_\partial := \inf\left\{ t \geq 0 : \int_0^t \kappa(X_s) \, \mathrm{d}s > \xi \right\},$$

where ξ is an exponential random variable with parameter 1 independent of X.
This killing time τ_∂, when the cumulative hazard function $t \mapsto \int_0^t \kappa(X_s) \, \mathrm{d}s$ exceeds
the (independent) threshold ξ, is equivalent to the first arrival time of a (doubly
stochastic) Poisson process with rate function $t \mapsto \kappa(X_t)$.
   We show that

$$(1.6) \qquad \begin{aligned} &\mathbb{P}_x(X_t \in E | \tau_\partial > t) \to \pi(E) \\ &\quad \text{as } t \to \infty \text{ for all } x \in \mathbb{R}^d \text{ and Borel-measurable } E \subset \mathbb{R}^d. \end{aligned}$$

To have confidence that this convergence is practically meaningful for a sam-
pling algorithm, we need in addition to have some control over the rate of the
convergence.
   Our first result gives natural conditions under which the convergence (1.6)
holds.
   To begin with, we require the following compatibility condition between the
tails of π and the underlying diffusion.

ASSUMPTION 2.

$$\int_{\mathbb{R}^d} \frac{\pi^2(y)}{\exp(2A(y))} \, dy < \infty.$$

Assumption 2 is natural from a statistical point of view. Recall that without killing, the diffusion X has invariant density proportional to $\exp(2A)$ if this quantity is integrable (and certain regularity conditions hold; see [22], Theorem 2.1). Assumption 2 can then be interpreted as requiring that the likelihood ratio $\pi(Y)/e^{2A(Y)}$ has finite variance when $Y \sim \exp(2A)$. This is what we would need to assume were we to target π by importance sampling from $\exp(2A)$.

In particular, Assumption 2 holds when the stronger 'rejection sampling' condition holds: that there exists some $M < \infty$ such that

$$(1.7) \qquad \frac{\pi(y)}{\exp(2A(y))} < M \qquad \forall y \in \mathbb{R}^d.$$

If $\exp(2A)$ is integrable, then this is precisely the condition that would allow us to sample from π using a rejection sampler with proposal density proportional to $\exp(2A)$. Informally, this demands that the asymptotic tail behavior of the diffusion be heavier than the tails of the target distribution. In particular, if the diffusion X is a Brownian motion on $\mathbb{R}^d$ [$A \equiv 0$ in (1.4)], Assumption 2 holds whenever the target density π is bounded.

We now define the appropriate killing rate κ, to be used to construct the killing time τ_∂ in (1.5). Define $\tilde{\kappa} : \mathbb{R}^d \to \mathbb{R}$ by

$$(1.8) \qquad \tilde{\kappa}(y) := \frac{1}{2} \left(\frac{\Delta \pi}{\pi} - \frac{2\nabla A \cdot \nabla \pi}{\pi} - 2\Delta A \right)(y), \qquad y \in \mathbb{R}^d$$

where Δ denotes the Laplacian operator. We require the following.

ASSUMPTION 3. $\tilde{\kappa}$ is bounded below, and not identically zero.

We will see that the correct killing rate is

$$(1.9) \qquad \kappa = \tilde{\kappa} + K,$$

where $K := -\inf_{y \in \mathbb{R}^d} \tilde{\kappa}(y)$, chosen so that κ is nonnegative everywhere. If $\tilde{\kappa}$ is identically zero, then there is no killing and we are in the familiar realm of stationary convergence of (unkilled) Markov processes; in fact, X will be a Langevin diffusion targeting π; see [22]. To facilitate the development of intuition, some examples of κ in the case of $A \equiv 0$ are given in Section 1.5. Heuristically, this form for the killing rate makes π an eigenfunction for the generator of the killed diffusion, which corresponds to quasi-stationarity; see Section 3 for the mathematical details and further explanation.

The form of the untranslated killing rate in (1.8) also has the natural following interpretation. Writing $U := \log \pi$, which we can do since we are assuming π is

positive, and as above thinking of $\exp(2A)$ as describing the asymptotic unkilled dynamics, we can rewrite (1.8) as

$$(1.10) \qquad \tilde{\kappa}(y) = \frac{1}{2}\big(\Delta(U - 2A) + \nabla U \cdot \nabla(U - 2A)\big)(y), \qquad y \in \mathbb{R}^d.$$

Written this way, we see $\tilde{\kappa}$ is a measure of the discrepancy between the derivatives of $\log \pi$ and $2A$, and Assumption 3 states that this discrepancy cannot be arbitrarily negative.

1.3. Convergence to quasi-stationarity.

THEOREM 1. *Suppose Assumptions* 0, 1, 3 *and* 2 *and hold. Then X has quasi-limiting distribution* π. *That is, the convergence in* (1.6) *holds.*

REMARKS. 1. This significantly improves on Theorem 1 of [20]: their result only applied to killed Brownian motions, and their complicated condition on the tails of the target density has been removed. While Brownian motion—$A \equiv 0$ in (1.4)—is a natural choice of a 'proposal' diffusion, with developments in the exact simulation of diffusions, such as [3], there is potential to consider other diffusions as candidates. In Section 2, we consider an Ornstein–Uhlenbeck process targeting a Gaussian distribution.

2. We are not able to use the recent convergence results of [6]. Their approach is via minorisation-type conditions, which do not hold in our particular noncompact state space setting, and so we cannot apply their theorem on uniform exponential convergence.

3. Assumption 2 is in fact not a necessary condition. For example, in Section 4.6 of [13] the authors consider cases of low killing on $[0, \infty)$, where λ_0^κ, the bottom of the spectrum (in our case K; see Section 3.3), is not an eigenvalue in the $\mathcal{L}^2$ sense, but convergence to quasi-stationarity still occurs. Instead, the requirement is that the unkilled process be recurrent. In the context of quasi-stationary Monte Carlo methods, where we are free to choose the diffusion, Assumption 2 is a natural condition, since the excluded cases have zero spectral gap, hence inevitably poor convergence properties.

4. Theorem 1 also extends the results of [13]: there the authors considered only (one-dimensional) cases where $\lim_{y \to \infty} \kappa(y) \neq \lambda_0^\kappa$. For example, our result gives convergence of killed Brownian motions with polynomially-tailed quasi-stationary distributions: in such cases

$$\lim_{\|y\| \to \infty} \kappa(y) = \lambda_0^\kappa,$$

but the conditions of Theorem 1 still hold, so we obtain convergence to quasi-stationarity.

5. We also obtain convergence of the conditional measures $\mathbb{P}_x(X_t \in \cdot \,|\, \tau_\partial > t)$ to π in total variation distance as $t \to \infty$, as shown in the proof of Theorem 7 of [24].

1.4. *Rate of convergence.* Our second result helps us to understand the rate of convergence to quasi-stationarity. Let $Z = (Z_t)_{t \geq 0}$ be the weak solution of the related SDE

$$(1.11) \qquad \mathrm{d}Z_t = \frac{1}{2} \nabla \log \left(\frac{\pi^2}{\exp(2A)} \right)(Z_t) \, \mathrm{d}t + \mathrm{d}W_t,$$

with $Z_0 = x$. This is an example of a Langevin diffusion. Under suitable regularity conditions (see Theorem 2.1 of [22]), the law of the diffusion Z_t converges to the distribution on $\mathbb{R}^d$ with Lebesgue density proportional to $\pi^2 / \exp(2A)$ as $t \to \infty$. (Assumption 2 guarantees that this is integrable.) Let $-L^Z$ denote the infinitesimal generator of this process and let $-L^\kappa$ denote the infinitesimal generator of the process (1.4) killed at rate κ. These operators will be constructed explicitly in Section 3.3 as self-adjoint operators on the appropriate $\mathcal{L}^2$ Hilbert spaces.

Writing $\gamma := \exp(2A)$, $\Gamma(\mathrm{d}x) := \gamma(x) \, \mathrm{d}x$ for the corresponding Borel measure on $\mathbb{R}^d$, which is the reversing measure of the diffusion X, and $\varphi := \pi/\gamma$, we have the following result.

THEOREM 2. *Under the same conditions as Theorem 1, the $\mathcal{L}^2$ spectra of L^Z and L^κ agree, up to an additive constant. In particular, when L^Z has a spectral gap, the transition kernel of the killed process $p^\kappa(t, x, y)$ satisfies*

$$\left| e^{tK} p^\kappa(t, x, y) - \varphi(x)\varphi(y) \right| \leq C e^{-t(\lambda_1^Z - \lambda_0^Z)},$$

where $\lambda_1^Z > \lambda_0^Z = 0$ are the bottom two eigenvalues of the Langevin diffusion, and the constant C may depend on x and y. If the drift in (1.4) is bounded then C may be chosen independent of x and y.

If the measure Γ is such that $\Gamma(\mathbb{R}^d) < \infty$, then for an initial Γ-density $\psi \in \mathcal{L}^1(\Gamma) \cap \mathcal{L}^2(\Gamma)$,

$$\left| \mathbb{P}_\psi(X_t \in E | \tau_\partial > t) - \pi(E) \right| \leq C' e^{-t(\lambda_1^Z - \lambda_0^Z)},$$

for any measurable $E \subset \mathbb{R}^d$, where

$$C' = \frac{2(\int \psi(x)^2 \, \mathrm{d}\Gamma(x))^{1/2} \Gamma(\mathbb{R}^d)^{1/2}}{\int \psi(x)\pi(x) \, \mathrm{d}x \cdot \int \pi(x) \, \mathrm{d}x}.$$

The additive constant in Theorem 2 is K; that is, the spectrum of L^Z is the translation of the spectrum of L^κ by $+K$.

Theorem 2 tells us that the stationary convergence of the Langevin diffusion (1.11) and the quasi-stationary convergence of our killed diffusion occur at the same exponential rate, given by the equal spectral gaps. Since Langevin dynamics have been applied widely in computational statistics and the applied sciences, their rates of convergence have been studied extensively; see, for instance, the recent results of [8] and [12]. Thus for many cases of π we will be able to accurately describe the rate of convergence in (1.6).

Theorem 2 also suggests that quasi-stationary Monte Carlo methods relying on (1.6) may converge relatively slowly for densities which are *multimodal*. In the case of $A \equiv 0$ (killed Brownian motion), if π is multimodal, then π^2 will typically be even more irregular, and the Langevin diffusion targeting π^2 will converge only gradually. On the other hand, quasi-stationary Monte Carlo methods should have good success targeting densities which are unimodal, such as logconcave densities. If π is unimodal, π^2 will be even more regular and have faster tail decay, leading to faster convergence of the Langevin diffusion. Such densities appear naturally in the context of Big-Data Bayesian inference. The Bernstein–von Mises theorem ([26], Section 10.2) tells us, for instance, that for large datasets the posterior distributions are approximately Gaussian.

REMARKS.   1. A sufficient condition for the existence of a spectral gap ($\lambda_1^\kappa > \lambda_0^\kappa$) is that

$$(1.12) \qquad \liminf_{\|x\| \to \infty} \tilde{\kappa}(x) > 0.$$

See, for instance, the proof of Lemma 3.3(v) of [13], which carries over into our setting. Furthermore, if $\liminf_{\|x\| \to \infty} \tilde{\kappa}(x) = +\infty$ then this implies that the spectrum is purely discrete (the *essential spectrum* is empty). In the case of killed Brownian motion, this holds for all exponentially-tailed densities of the form $\exp(-\beta \|x\|^\alpha)$ for some $\beta > 0, \alpha \geq 1$.

2. The Langevin diffusion in (1.11) is precisely the Q-process (the diffusion conditioned never to be killed) defined by the diffusion X and the killing time τ_∂. It is defined as the limit

$$\mathbb{Q}_x(A) := \lim_{T \to \infty} \mathbb{P}_x(A | T < \tau_\partial)$$

for $A \in \sigma(X_s : s \leq t)$ for some $t \geq 0$.

3. Theorem 2 is a continuous state-space generalisation of Theorem 1 of [11]: there the authors showed that in a finite state-space, rates of convergence to quasi-stationarity in total variation distance can be bounded above and below by constant multiples of the rates of convergence to stationarity in total variation of an appropriate unkilled process.

1.5. *Examples of κ.*   In the simple and computationally important case of a killed Brownian motion [$A \equiv 0$ in (1.4)], $\tilde{\kappa}$ as defined in (1.8) simplifies down to

$$\tilde{\kappa}(y) = \frac{\Delta \pi}{2\pi}, \qquad y \in \mathbb{R}^d.$$

In the following examples it can be easily checked that the conditions of Theorem 1 are satisfied.

- *Gaussian on $\mathbb{R}^d$*. Let $\sigma^2 > 0$ and $\pi(y) \propto \exp(-\|y\|^2/(2\sigma^2))$ for $y \in \mathbb{R}^d$, where throughout $\|\cdot\|$ denotes the Euclidean norm. Then straightforward calculation gives us that $\tilde{\kappa}(y) = \frac{1}{2}(\sigma^{-4}\|y\|^2 - \sigma^{-2}d)$ for $y \in \mathbb{R}^d$ and hence

$$\kappa(y) = \frac{1}{2\sigma^4}\|y\|^2, \qquad y \in \mathbb{R}^d.$$

Since $\liminf_{\|y\|\to\infty} \tilde{\kappa}(y) > 0$ (in fact it's infinite), we expect exponential rates of convergence to quasi-stationarity, from condition (1.12). This example is considered in some detail in the case $d = 1$ in Section 2. This example also gives the independently interesting result that a Brownian motion on $\mathbb{R}^d$ killed at a quadratic rate will have a Gaussian quasi-limiting distribution.

- *Univariate exponential decay*. Consider a one-dimensional, positive, smooth target density π with tail decay $\pi \propto \exp(-\beta|y|)$ for all y outside of a compact set $E \subset \mathbb{R}$, for some $\beta > 0$. We find that for all $y \in \mathbb{R}\backslash E$, $\tilde{\kappa}(y) = \beta^2$, that is, a positive constant. The killing rate κ will then also be constant asymptotically. By condition (1.12), we expect exponential convergence to quasi-stationarity.

- *Heavy-tailed case*. Consider a univariate Cauchy target, $\pi(y) \propto 1/(1 + y^2)$ for $y \in \mathbb{R}$. Then simple calculation gives $\tilde{\kappa}(y) = \frac{3y^2-1}{(1+y^2)^2}$, for $y \in \mathbb{R}$ and then

$$\kappa(y) = \frac{3y^2 - 1}{(1 + y^2)^2} + 1, \qquad y \in \mathbb{R}^d.$$

We see here an example where $\liminf_{|y|\to\infty} \tilde{\kappa}(y) = 0$; the sufficient condition for a spectral gap (1.12) fails and we expect slower convergence.

2. Example: Ornstein–Uhlenbeck process targeting a Gaussian density.
Before turning to the mathematical technicalities, we offer a mathematically tractable example that can be readily simulated: a killed Ornstein–Uhlenbeck process targeting a Gaussian distribution. For simplicity of presentation, we discuss the univariate case $d = 1$. Analogous results hold in the multivariate case, but the notation is more cumbersome, and the calculations more involved.

Throughout this section, we write $\mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}, \sigma^2 > 0$ to denote the univariate Gaussian distribution with mean μ and variance σ^2.

In (1.4), we let $A(y) = -(\nu - y)^2/(4\tau^2)$ for each $y \in \mathbb{R}$, where $\nu \in \mathbb{R}, \tau^2 > 0$ are fixed. This defines a diffusion X as the weak solution of

$$(2.1) \qquad dX_t = \frac{1}{2\tau^2}(\nu - X_t)\,dt + dW_t, \qquad X_0 = x.$$

The Ornstein–Uhlenbeck process X has a $\mathcal{N}(\nu, \tau^2)$ stationary distribution; the corresponding density function is proportional to $\exp(2A)$.

Fix $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, and let the target density be

$$\pi(y) = \exp\left\{-\frac{1}{2\sigma^2}(y - \mu)^2\right\}$$

for each $y \in \mathbb{R}$, the (unnormalised) density of a $\mathcal{N}(\mu, \sigma^2)$ random variable. We note that the regularity conditions—Assumptions 0 and 1—hold.

The untranslated killing rate computed from (1.8) is for each $y \in \mathbb{R}$ given by

$$(2.2) \qquad \tilde{\kappa}(y) = \frac{1}{2}\left(\frac{(y-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2} + \frac{(\nu - y)(y - \mu)}{\tau^2\sigma^2} + \frac{1}{\tau^2} \right).$$

We now assume

$$(2.3) \qquad \tau^2 > \sigma^2;$$

that is, the invariant distribution of the underlying diffusion has tails that are heavier than those of the target distribution. This makes the leading coefficient in the quadratic (2.2) positive, so that $\tilde{\kappa}$ is bounded below, meaning that Assumption 3 holds. In this case, we will have a spectral gap (since the limit of the killing at infinity is $+\infty$; see (1.12), so we expect quasi-stationary convergence to occur at an exponential rate. Completing the square in (2.2) gives the minimum value

$$(2.4) \qquad K := -\inf_{y \in \mathbb{R}} \tilde{\kappa}(y) = \frac{(\mu - \nu)^2}{8\tau^2(\tau^2 - \sigma^2)} + \frac{\tau^2 - \sigma^2}{2\tau^2\sigma^2}.$$

In Section 3.3, we will identify K with λ_0^κ, the bottom of the $\mathcal{L}^2$-spectrum of the generator of the killed diffusion, and so K is also the asymptotic rate of killing (see Lemma 4.2 of [13]). We see from (2.4) that K is strictly positive, as our calculation in Section 3.2 predicts. Adding K to $\tilde{\kappa}$ and rearranging, we obtain the killing rate

$$(2.5) \qquad \kappa(y) = \frac{\tau^2 - \sigma^2}{2\tau^2\sigma^4}\left(y - \left\{ \frac{\mu + \nu}{2} + \frac{\tau^2}{\tau^2 - \sigma^2}\left(\mu - \frac{\mu + \nu}{2} \right) \right\} \right)^2$$

for $y \in \mathbb{R}$.

It remains to check Assumption 2. By direct calculation,

$$\frac{\pi^2}{\exp(2A)}(y) \propto \exp\left\{ -\frac{1}{2}\frac{2\tau^2 - \sigma^2}{\sigma^2\tau^2}\left(y - \frac{2\mu\tau^2 - \nu\sigma^2}{2\tau^2 - \sigma^2} \right)^2 \right\}.$$

Our assumption (2.3) guarantees this will be integrable, and in fact proportional to the density of the Gaussian distribution

$$(2.6) \qquad \mathcal{N}\left(\frac{2\mu\tau^2 - \nu\sigma^2}{2\tau^2 - \sigma^2}, \frac{\sigma^2\tau^2}{2\tau^2 - \sigma^2} \right).$$

So Theorem 1 allows us to conclude that π is the quasi-limiting distribution of our Ornstein–Uhlenbeck process (2.1) killed at rate (2.5), as long as (2.3) holds.

Since $\pi^2/\exp(2A)$ is the density of a Gaussian distribution, it follows that the corresponding Langevin diffusion (1.11) is another Ornstein–Uhlenbeck process, albeit with stationary distribution given by (2.6). In [15], the authors explicitly

computed the $\mathcal{L}^p$ spectra of Ornstein–Uhlenbeck operators, and by applying their Theorem 3.1 we find that the $\mathcal{L}^2$ spectrum of L^Z is given by

$$\Sigma(L^Z) = \left\{ \lambda_n^Z = \frac{n(2\tau^2 - \sigma^2)}{2\sigma^2\tau^2} : n = 0, 1, 2, \dots \right\}.$$

By Theorem 2, this coincides (up to an additive constant) with the spectrum of our killed process (2.1). In particular, the spectral gap of our killed process is

$$\lambda_1^Z - \lambda_0^Z = \frac{2\tau^2 - \sigma^2}{2\sigma^2\tau^2} = \frac{1}{\sigma^2} - \frac{1}{2\tau^2}.$$

For this example, there are two mechanisms influencing the convergence to quasi-stationarity: the drift of the underlying diffusion (2.1), along with the killing (2.5) and subsequent conditioning on survival. It is interesting to note that the spectral gap is maximised when $\tau^2 \to \infty$, in which case the drift is 0. When in addition $\mu = \nu$, we see that the killing is also maximal, as measured by, say, the asymptotic killing rate (2.4). This limit case $\tau^2 \to \infty$ corresponds to the case of killed Brownian motion [$A \equiv 0$ in (1.4)]. This suggests that the rate of convergence to quasi-stationarity is determined more by the killing mechanism than by the underlying drift. However, depending on the method of implementation, a greater rate of killing could lead to reduced computational efficiency.

This simple example is amenable to simulation, as shown in Figure 1. The figure shows the conditional distributions $\mathbb{P}_x(X_T \in \cdot \mid \tau_\partial > T)$ for $T = 1, 5, 10, 20$ for the choices $\nu = 2, \tau^2 = 4, \mu = -1, \sigma^2 = 2$, and initial value $X_0 = x = 3$.
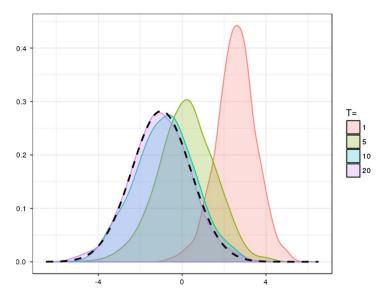

FIG. 1.   *Estimates of the conditioned laws $\mathbb{P}_x(X_T \in \cdot \mid \tau_\partial > T)$ for the process (2.1) with parameters $\nu = 2$, $\tau^2 = 4$ started at $x = 3$ for various T. The dashed black line shows the quasi-stationary density, a Gaussian density with mean and variance $\mu = -1, \sigma^2 = 2$, respectively.*

3. Mathematical preliminaries.

3.1. *Definitions.* Fix $d \in \mathbb{N}$. Let ∇ denote the gradient operator; the d-dimensional vector with components $\nabla_i = \partial/\partial x_i$, $i = 1, \ldots, d$. We will denote the Laplacian operator by $\Delta := \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$. We are given functions $A : \mathbb{R}^d \to \mathbb{R}$ and $\pi : \mathbb{R}^d \to [0, \infty)$ that satisfy Assumptions 0, 1 and 2. For brevity, we write

$$\gamma := \exp(2A).$$

π is our target density, which need not be normalised. In a slight abuse of notation, we will also write π for the Borel probability measure on $\mathbb{R}^d$ with Lebesgue density proportional to π.

Let $C \equiv C([0, \infty), \mathbb{R}^d)$ denote the space of continuous functions mapping $[0, \infty) \to \mathbb{R}^d$, and let ω be a typical element. For each $t \geq 0$, let $X_t : C \to \mathbb{R}^d$ be the coordinate mapping $X_t(\omega) = \omega(t)$, and let $\mathcal{C} := \sigma(\{X_t : t \geq 0\})$ be the cylinder σ-algebra. For any $x \in \mathbb{R}^d$, let $\tilde{\mathbb{P}}_x$ be the measure on $(C, \mathcal{C})$ such that under $\tilde{\mathbb{P}}_x$, $X = (X_t)_{t \geq 0}$ is the weak solution to (1.4).

Define $\kappa : \mathbb{R}^d \to [0, \infty)$ by

$$\kappa(y) := \tilde{\kappa}(y) + K \qquad \forall y \in \mathbb{R}^d,$$

where $\tilde{\kappa}$, defined in (1.8), is required to satisfy Assumption 3, so that $K := -\inf_{y \in \mathbb{R}^d} \tilde{\kappa}(y)$ is finite. We augment our probability space to include an independent unit exponential random variable ξ, and define killing at rate κ as in (1.5), denoting this augmented space by $(\Omega, \mathcal{F}, \mathbb{P}_x)$.

We define $\mathcal{L}^2(\Gamma) \equiv \mathcal{L}^2(\mathbb{R}^d, \Gamma)$ to be the Hilbert space of (equivalence classes of) Borel-measurable square-integrable functions $f, g : \mathbb{R}^d \to \mathbb{R}$ with respect to the inner product

$$\langle f, g \rangle_{\mathcal{L}^2(\Gamma)} = \int_{\mathbb{R}^d} f(y) g(y) \, d\Gamma(y),$$

where the measure Γ is given by $d\Gamma(y) = \gamma(y) \, dy$, with dy denoting Lebesgue measure on $\mathbb{R}^d$. We denote the corresponding norm by $\| \cdot \|_{\mathcal{L}^2(\Gamma)}$.

Define $\varphi : \mathbb{R}^d \to \mathbb{R}$ by

$$\varphi := \frac{\pi}{\exp(2A)},$$

which is smooth and positive. By construction, we have that φ is integrable with respect to Γ:

$$\int_{\mathbb{R}^d} \varphi(y) \, d\Gamma(y) = \int_{\mathbb{R}^d} \pi(y) \, dy < \infty.$$

We will generally be working in the function space $\mathcal{L}^2(\Gamma)$, as this is the space on which the generator of the killed diffusion can be realised as a self-adjoint operator, which we will do explicitly in Section 3.3. As such, we will want consider densities

with respect to Γ—rather than Lebesgue measure—and hence we will work with φ, rather than directly with π. Of course in the case of killed Brownian motion, $A \equiv 0$, π and φ coincide.

Following this line of thought, Assumption 2 states that indeed $\varphi \in \mathcal{L}^2(\Gamma)$:

$$\|\varphi\|_{\mathcal{L}^2(\Gamma)}^2 = \int_{\mathbb{R}^d} \frac{\pi^2(y)}{\exp(2A(y))} \, dy < \infty.$$

Without loss of generality, we can rescale π so that this quantity is 1.

3.2. *The killed Markov semi-group.* Our results depend on the spectral theory of self-adjoint linear operators on Hilbert spaces. The proof of Theorem 1 avoids the heavy machinery of this theory by drawing on R. Tweedie's *R-theory*, which provides some of the results of operator theory most relevant to asymptotics of stochastic processes in a somewhat probabilistic package. We review the essentials of operator theory in Section 3.3, but it will be required only for the proof of Theorem 2.

The diffusion X killed at rate κ has a formal infinitesimal generator $-\tilde{L}^\kappa$ described by

$$(3.1) \qquad \tilde{L}^\kappa = -\frac{1}{2\exp(2A)} \nabla \cdot \exp(2A)\nabla + \kappa = -\frac{1}{2}\Delta - \nabla A \cdot \nabla + \kappa.$$

Under Assumption 3, this formal differential operator can be realized as a positive self-adjoint operator L^κ on an $\mathcal{L}^2$ Hilbert space. It is this theory that we defer to Section 3.3.

Straightforward calculation shows that

$$(3.2) \qquad\qquad\qquad \tilde{L}^\kappa \varphi = K\varphi.$$

So φ is an eigenfunction of the formal differential operator $\tilde{L}^\kappa$ with eigenvalue K. Since we have assumed that φ is in $\mathcal{L}^2(\Gamma)$,

$$
\begin{aligned}
K \int \varphi(y)^2 \, d\Gamma(y) &= \int \varphi(y)\tilde{L}^\kappa \varphi(y) \, d\Gamma(y) \\
&= -\frac{1}{2} \int \varphi(y)(\Delta\varphi(y) + 2\nabla A \cdot \nabla\varphi(y)) \, d\Gamma(y) \\
&\quad + \int \kappa(y)\varphi(y)^2 \, d\Gamma(y) \\
&\geq \int \kappa(y)\varphi(y)^2 \, d\Gamma(y) \\
&> 0.
\end{aligned}
$$

The first inequality follows since φ is smooth, so an application of Green's identity shows that the integral term is nonnegative. The final strict inequality follows since

φ and γ are strictly positive and κ is not identically 0, by Assumption 3. Thus we conclude $K > 0$.

Recall from [24] that a finite nonnegative measurable function f with $\int f(x)\,dx > 0$ is said to be λ-*invariant* for a continuous-time semi-group $(P_t)_{t \geq 0}$ if for all $t > 0$,

$$f(x) = e^{\lambda t} P_t f(x) \qquad \text{for almost every } x$$

and a σ-finite nontrivial measure ν is λ-*invariant* for continuous-time $(P_t)_{t \geq 0}$ if for all $t > 0$,

$$\nu(A) = e^{\lambda t} \nu P_t(A) \qquad \text{for every measurable } A.$$

Analogous notions of R-invariance of functions and measures are similarly defined for discrete-time processes as well; the requirement $t > 0$ is replaced with $t \in \mathbb{N}$, and $e^{\lambda t}$ is replaced is replaced by R^t.

All we need for present purposes is the following lemma.

LEMMA 3. *The sub-Markovian semi-group $(P_t^\kappa)_{t \geq 0}$ of the killed process X has a unique self-adjoint generator that is an extension of $-\tilde{L}^\kappa$ on smooth compactly supported functions. π is a λ-invariant measure for this semi-group, and φ a λ-invariant function, for $\lambda = K$.*

Except for some technical complications, which we will describe in the context of presenting the operator-theory framework in Section 3.3, this should be reasonably intuitive. We have already pointed out in (3.2) that φ is an eigenfunction of the generator with eigenvalue $-K$. Direct calculation shows that $\tilde{L}^\kappa$ is symmetric with respect to the measure Γ; that is, for $f, g \in \mathcal{L}^2(\Gamma)$ in the domain of $\tilde{L}^\kappa$ we have that

$$\langle \tilde{L}^\kappa f, g \rangle_{\mathcal{L}^2(\Gamma)} = \langle f, \tilde{L}^\kappa g \rangle_{\mathcal{L}^2(\Gamma)}.$$

Heuristically, since our assumptions ensure that the generator of the killed diffusion L^κ is symmetric, using (3.2) we obtain the following manipulations, for any nonnegative test function $f \in \mathcal{L}^2(\Gamma)$:

$$\begin{aligned}
\mathbb{E}_\pi \big[L^\kappa f(Y) \big] &= \int \pi(y) L^\kappa f(y)\,dy = \int \varphi(y) L^\kappa f(y)\,d\Gamma(y) \\
&= \int L^\kappa \varphi(y) f(y)\,d\Gamma(y) = K \int \varphi(y) f(y)\,d\Gamma(y) \\
&= K \mathbb{E}_\pi \big[f(Y) \big].
\end{aligned}$$

Bearing in mind that L^κ is *minus* the generator of the killed diffusion, this shows that started in π the process will remain in π, except with a *mass loss at rate K*. That is to say, π is quasi-stationary. For an unkilled diffusion, if π were stationary,

we would expect a similar expression to hold for any appropriate f, except with the right-hand side being exactly zero, reflecting the fact that the mass is preserved.

If we think of the adjoint operator—acting on measures—as acting on densities with respect to Γ, we have $(P_t^\kappa)^* g = (P_t^\kappa) g$. On the other hand, if g is a density with respect to Lebesgue measure the action is

$$(3.3) \qquad\qquad P_t^\kappa g = \gamma P_t^\kappa (g/\gamma).$$

3.3. *Operator theory.* This section gives the mathematical background necessary for the proof of Theorem 2 in Section 5. Readers interested in the proof of Theorem 1 can move straight to Section 4.

Our operator $\tilde{L}^\kappa$ on $C_c^\infty(\mathbb{R}^d)$, smooth compactly supported functions, is a symmetric semi-bounded operator and, therefore, has a self-adjoint extension, for instance the Friedrichs extension; see [9], Section 4.4. As a matter of fact, our operator is essentially self-adjoint—proven in Section 5.1— and thus has a unique self-adjoint extension L^κ, so its completions are self-adjoint.

Recall that (3.1) describes the formal infinitesimal generator of our killed process. This gives rise to a closable densely defined positive quadratic form $\tilde{q}^\kappa$ on $\mathcal{L}^2(\Gamma)$ given by

$$\tilde{q}^\kappa(f) = \frac{1}{2} \int_{\mathbb{R}^d} \nabla f \cdot \nabla f(y) \gamma(y) \,\mathrm{d}y + \int_{\mathbb{R}^d} \kappa(y) |f(y)|^2 \gamma(y) \,\mathrm{d}y$$

for $f \in \mathcal{D}_\kappa$, where

$$\mathcal{D}_\kappa := \left\{ f \in \mathcal{L}^2(\Gamma) : f \text{ continuously differentiable, } \tilde{q}^\kappa(f) < \infty \right\}.$$

We note that Assumption 3 is essential here. From a probabilistic point of view, we need $\tilde{\kappa}$ to be bounded below since a sensible killing rate must be nonnegative (which amounts to putting a bound on the Radon–Nikodým derivative; see [20], Appendix B). From a functional-analytic point of view, we also need $\tilde{\kappa}$ to be bounded below since we require $\tilde{q}^\kappa$ to be closable. The semi-boundedness assumption on $\tilde{\kappa}$ implies that for all compactly supported, twice differentiable $f \in C_c^2(\mathbb{R}^d)$, $\tilde{q}^\kappa(f)$ is a nonnegative quadratic form associated to the symmetric operator $\tilde{L}^\kappa$. By Lemma 1.29, Assertion 2 of [16], we therefore conclude that the quadratic form $\tilde{q}^\kappa$ is closable.

Now let us denote the closure of $\tilde{q}^\kappa$ by q^κ. To this quadratic form, there is associated a unique positive self-adjoint operator L^κ, with dense domain $\mathcal{D}(L^\kappa) \subset \mathcal{L}^2(\Gamma)$; see [16], Section 1.2.3. For smooth functions the action of L^κ is identical to that of $\tilde{L}^\kappa$.

Let $\Sigma(L^\kappa)$ denote the $\mathcal{L}^2(\Gamma)$-spectrum of L^κ. Since L^κ is self-adjoint and positive, we have that $\Sigma(L^\kappa) \subset [0, \infty)$. We have seen in (3.2) that $K \in \Sigma(L^\kappa)$; in particular $\Sigma(L^\kappa)$ is nonempty, so let us write λ_0^κ for the bottom of the spectrum. In fact, we have that $K = \lambda_0^\kappa$. This follows from general operator theory since φ is positive everywhere. We also have that λ_0^κ is a simple eigenvalue, with φ being its

unique eigenfunction up to constant multiples. A reference for these assertions is
[21], Section XIII.12.

We now make use of the spectral calculus for self-adjoint operators using
projection-valued measures, as discussed in [9], Section 2.5. This gives us the ex-
istence of a family of spectral projections $(E_\lambda^\kappa)_{\lambda \in [\lambda_0^\kappa, \infty)}$ and allows us to define
$\phi(L^\kappa)$ for Borel-measurable $\phi : \mathbb{R} \to \mathbb{R}$, via

$$\phi(L^\kappa)f = \int_{\Sigma(L^\kappa)} \phi(\lambda) \, dE_\lambda^\kappa f,$$

$$\mathcal{D}(\phi(L^\kappa)) = \left\{ f \in \mathcal{L}^2(\Gamma) : \int_{\Sigma(L^\kappa)} |\phi(\lambda)|^2 \, d\langle E_\lambda^\kappa f, f \rangle_{\mathcal{L}^2(\Gamma)} < \infty \right\},$$

$$\|\phi(L^\kappa)f\|_{\mathcal{L}^2(\Gamma)}^2 = \int_{\Sigma(L^\kappa)} |\phi(\lambda)|^2 \, d\langle E_\lambda^\kappa f, f \rangle_{\mathcal{L}^2(\Gamma)}.$$

Now the Feynman–Kac representation states that for each $t > 0$

$$(e^{-tL^\kappa} f)(x) = \mathbb{E}_x [f(X_t) 1_{\{\tau_\partial > t\}}]$$

for $f \in \mathcal{L}^2(\Gamma)$. Furthermore, for each $t > 0$ the operator e^{-tL^κ} is a contraction on
$\mathcal{L}^2(\Gamma)$ (*cf.* the derivation in [10]).

The spectral theorem allows us to write the diffusion semi-group as

$$P_t^\kappa f(x) = \mathbb{E}_x [f(X_t) 1_{\{\tau_\partial > t\}}] = e^{-tL^\kappa} f(x) = \int_{\Sigma(L^\kappa)} e^{-t\lambda} \, dE_\lambda^\kappa f(x)$$

for $f \in \mathcal{L}^2(\Gamma)$. The $(E_\lambda^\kappa)_{\lambda \in [\lambda_0^\kappa, \infty)}$ are orthogonal projections; in particular, $E_{\lambda_0^\kappa}^\kappa$
projects onto the span of φ. We can write

$$e^{-tL^\kappa} f = e^{-t\lambda_0^\kappa} \varphi \langle f, \varphi \rangle_{\mathcal{L}^2(\Gamma)} + \int_{\Sigma(L^\kappa) \setminus \{\lambda_0^\kappa\}} e^{-t\lambda} \, dE_\lambda^\kappa f.$$

Thus

$$(3.4) \qquad e^{t\lambda_0^\kappa} e^{-tL^\kappa} f = \varphi \langle f, \varphi \rangle_{\mathcal{L}^2(\Gamma)} + \int_{\Sigma(L^\kappa) \setminus \{\lambda_0^\kappa\}} e^{-t(\lambda - \lambda_0^\kappa)} \, dE_\lambda^\kappa f.$$

For a given $f \in \mathcal{L}^2$, we are interested in the convergence to 0 of the integral term
in (3.4). We note here that the convergence in this discussion is convergence in
$\mathcal{L}^2(\Gamma)$. Ultimately, we will be interested in convergence in $\mathcal{L}^1(\Gamma)$; we will return
to this issue later.

4. Proof of Theorem 1. We wish to apply the results of [24]. In order to do
this, we first need to check that $(P_t^\kappa)_{t \geq 0}$ is 'simultaneously ϕ-irreducible', that is,
the resolvent kernel is strictly positive for discrete versions of the process discre-
tised with respect to arbitrary time-steps. Ordinary ϕ-irreducibility holds for dif-
fusions with smooth drift and locally bounded by the Stroock–Varadhan support

theorem, [17], Section 2.6. Simultaneous ϕ-irreducibility follows then immediately from Theorem 1 of [24] since our process has a jointly continuous transition density with respect to the reversing measure; see Remark 1 after this proof.

We now show that $(P_t^\kappa)_{t\geq 0}$ is λ-positive, with $\lambda = K$, and that the K-invariant measure is precisely the target density π. This will then imply convergence to quasi-stationarity by an application of Theorem 7 of [24], which states that λ-positive processes, when $\lambda > 0$, exhibit quasi-limiting convergence as in (1.6), where the quasi-limiting distribution is the (unique) λ-invariant measure.

By Theorem 4(ii) of [24], showing $(P_t^\kappa)_{t\geq 0}$ is λ-positive is equivalent to showing that each (discrete-time) skeleton chain generated by P_h^κ, for any $h > 0$, is $e^{\lambda h}$-positive in the discrete-time sense, as defined in [25]. This involves showing that each skeleton chain is R-recurrent with $R = e^{\lambda h}$ and that the corresponding integral of the $e^{\lambda h}$-invariant function against the $e^{\lambda h}$-invariant measure is finite. So let us fix $h > 0$.

It follows from (3.2) and the Kolmogorov equations that

$$e^{hK} P_h^\kappa \varphi = \varphi.$$

This is exactly the definition of φ being e^{hK}-invariant for the discrete-time semi-group generated by P_h^κ. By (3.3), the measure π with Lebesgue density $\gamma\varphi$ is similarly e^{hK}-invariant for the discrete-time chain. (Definitions of λ-invariance are included in Section 3.2 for convenience.)

By Assumption 2,

$$\int_{\mathbb{R}^d} \varphi(y)\pi(\mathrm{d}y) = \int_{\mathbb{R}^d} \frac{\pi^2(y)}{\gamma(y)}\,\mathrm{d}y < \infty.$$

Thus by Proposition 3.1 and Proposition 4.3 of [25] the skeleton chain defined by operator P_h^κ, $(X_{nh})_{n=1}^\infty$, is R-recurrent, with $R = e^{hK}$. Theorem 7 of [25] then tells us that this skeleton chain is e^{hK}-positive. Since $h > 0$ was arbitrary, we obtain that $(P_t^\kappa)_{t\geq 0}$ is λ-positive, with $\lambda = K$. Theorem 4(iii) of [24] also tells us that φ and π are the unique K-invariant function and measure for (P_t^κ), respectively.

We are now in a position to utilise Theorem 7 of [24]. Since $K > 0$, killing happens almost surely, hence the key assumption (B) of Theorem 7 of [24] requires simply that $\int \pi(y)\,\mathrm{d}y < \infty$, which is certainly true. The conclusion of the theorem implies convergence to quasi-stationarity (1.6); that is, for any measurable $E \subset \mathbb{R}^d$ there is a set of starting points x of full Lebesgue measure such that

$$\lim_{t\to\infty} \mathbb{P}_x(X_t \in E | \tau_\partial > t) = \frac{\int_E \pi(y)\,\mathrm{d}y}{\int_{\mathbb{R}^d} \pi(y)\,\mathrm{d}y}.$$

In fact, this convergence holds for every starting point x. Since we have a continuous transition density $p^\kappa(t, x, y)$ (see Remark 1 after this proof), we have for any measurable set $E \subset \mathbb{R}^d$

$$\mathbb{P}_x(X_{t+1} \in E) = \int_{\mathbb{R}^d} p^\kappa(1, x, y)\mathbb{P}_y(X_t \in E)\,\mathrm{d}\Gamma(y).$$

Since we have convergence for y in some set of full measure, we obtain convergence for all $x \in \mathbb{R}^d$, which completes the desired result.

REMARKS.   1. Assumption 2 can be interpreted in terms of spectral theory. It tells us that $\varphi \in \mathcal{L}^2(\mathbb{R}^d, \Gamma)$, so φ is also an eigenfunction of L^κ in the sense of $\mathcal{L}^2$ spectral theory. It is then possible to prove Theorem 1 analogously to Lemma 4.4 of [13]. Following the derivation of [10], it follows that we have a continuous integral kernel $p^\kappa(t, x, y)$ with $\mathbb{E}_x[f(X_t)1_{\{\tau_\partial > t\}}] = \int p^\kappa(t, x, y) f(y) \, d\Gamma(y)$. We can then apply [23] to see that $e^{t\lambda_0^\kappa} p^\kappa(t, x, y) \to c\varphi(x)\varphi(y)$ as $t \to \infty$, where $c = \|\varphi\|_{\mathcal{L}^2(\Gamma)}^{-2}$ and the proof of Theorem 1 can proceed analogously.

2. Our argument here relies fundamentally on self-adjointness of the operators and subsequent properties such as (3.2), so there is no way we can circumvent the assumption of a gradient-form drift in (1.4). In one dimension, this always holds, since we can simply take the integral of the drift function.

5. Rates of convergence.

5. Rates of convergence.   Practitioners hoping to implement quasi-stationary Monte Carlo methods, such as the ScaLE Algorithm of [20], having been reassured that the procedure indeed converges to the correct distribution, will naturally inquire about the rate of convergence. Our result in this section draws heavily on the spectral theory for self-adjoint (unbounded) operators that we have outlined in Section 3.3.

When there is a *spectral gap*, that is, when $\lambda_1^\kappa > \lambda_0^\kappa$, the integral term will vanish at an exponential rate. Thus, it suffices to understand the spectrum $\Sigma(L^\kappa)$. To do this, we will adapt an idea of [18], to transform our operator into one whose spectrum is already understood. Here, it will be the infinitesimal generator of a certain Langevin diffusion.

5.1. *Proof of Theorem 2.*   Consider the formal differential operator

$$\tilde{L}^{\tilde{\kappa}} = -\frac{1}{2\gamma}\nabla \cdot \gamma \nabla + \tilde{\kappa},$$

where $\tilde{\kappa}$ is defined in (1.8), acting on $C_c^\infty(\mathbb{R}^d)$, the set of smooth compactly-supported functions. This is very similar to the formal differential operator we began with in (3.1), differing only by an additive constant K, which will have the effect of merely translating the spectrum accordingly. $\tilde{L}^{\tilde{\kappa}}$ can be realised as a non-negative, self-adjoint operator $L^{\tilde{\kappa}}$ on $\mathcal{L}^2(\Gamma)$, by taking the Friedrichs extension of the appropriate quadratic form as before.

Now let $\mathcal{L}^2(\pi^2/\gamma) \equiv \mathcal{L}^2(\mathbb{R}^d, \pi^2/\gamma)$ denote the Hilbert space of (equivalence classes of) measurable functions $u, v : \mathbb{R}^d \to \mathbb{R}$ which are square-integrable with respect to the inner product

$$\langle u, v \rangle_{\mathcal{L}^2(\pi^2/\gamma)} = \int_{\mathbb{R}^d} u(y)v(y)\frac{\pi^2(y)}{\gamma(y)} \, dy.$$

The multiplication operator

$$Uf = \frac{\gamma}{\pi} f$$

is a bounded unitary transformation $U : \mathcal{L}^2(\Gamma) \to \mathcal{L}^2(\pi^2/\gamma)$, with inverse given by $U^{-1}u = \frac{\pi}{\gamma}u$.

We now define a second formal differential operator

$$\tilde{L}^Z = -\frac{1}{2}\Delta - \frac{1}{2}\nabla \log\left(\frac{\pi^2}{\gamma}\right) \cdot \nabla,$$

which is minus the generator of the Langevin diffusion given in (1.11), targeting the density π^2/γ. $\tilde{L}^Z$ can similarly be realized as a positive, self-adjoint operator L^Z on $\mathcal{L}^2(\pi^2/\gamma)$. Our two formal operators are related through

$$\tilde{L}^{\tilde{\kappa}} = U^{-1}\tilde{L}^Z U.$$

We can also conjugate L^Z to obtain $U^{-1}L^Z U$, a self-adjoint operator on $\mathcal{L}^2(\Gamma)$.

Theorem 2 will be an immediate consequence if we show that in fact $U^{-1}L^Z U = L^{\tilde{\kappa}}$. This is the same as showing that the following diagram commutes:

$$
\begin{array}{ccc}
\tilde{L}^Z & \xrightarrow{\text{Friedrichs ext.}} & L^Z \\
\Big\downarrow{\scriptstyle\text{Conjugate with } U} & & \Big\downarrow{\scriptstyle\text{Conjugate with } U} \\
\tilde{L}^{\tilde{\kappa}} & \xrightarrow{\text{Friedrichs ext.}} & L^{\tilde{\kappa}}
\end{array}
$$

An operator is said to be *essentially self-adjoint* if it has a unique self-adjoint extension, which is given by the closure. From the background in Section 3.3, we see that the diagram commutes, and so Theorem 2 will follow, if we can show that $\tilde{L}^{\tilde{\kappa}}$ acting on $C_c^\infty(\mathbb{R}^d)$ is essentially self-adjoint. After all, the conjugate $U^{-1}\tilde{L}^Z U$ is a self-adjoint extension of $\tilde{L}^{\tilde{\kappa}}$; if the extension is unique it must be the same as $L^{\tilde{\kappa}}$.

We apply Theorem 2.13 of [5]. The smooth boundaryless manifold we are working in is simply $\mathbb{R}^d$, with smooth positive measure measure Γ. In their notation, we take D to be $\frac{1}{\sqrt{2}}\nabla$, which is elliptic. The formal adjoint D^* is given by $-\frac{1}{\sqrt{2}}(\nabla \cdot + 2\nabla A \cdot)$. We set $V = \tilde{\kappa}$, and the resulting operator H_V is precisely $\tilde{L}^{\tilde{\kappa}}$.

The result follows immediately if V satisfies their Assumptions A and B, which ask for a decomposition of V into well-behaved nonnegative parts and a mild technical condition. Assumption A is immediate by writing

$$V = \underbrace{\tilde{\kappa} + K}_{V_+} + \underbrace{(-K)}_{V_-},$$

where clearly $V_+ \geq 0$ and $V_- \leq 0$. V_- trivially satisfies (ii) of Assumption A since it is constant.

Assumption B follows from their Theorem 2.3(ii), since our operator acts on scalar functions. The final condition of Theorem 2.13 is completeness of the metric g^{TM}, which is satisfied since it is equivalent to geodesic completeness of the manifold, which is true for $\mathbb{R}^d$.

Since unitary transformations leave spectra invariant it follows that the $\mathcal{L}^2(\Gamma)$ spectrum of $L^{\tilde{\kappa}}$ coincides with the $\mathcal{L}^2(\pi^2/\gamma)$ spectrum of L^Z, and hence the $\mathcal{L}^2$ spectra of $L^{\tilde{\kappa}}$ and L^Z coincide after translation by K.

We now would like to extend our proof of $\mathcal{L}^2$ convergence to $\mathcal{L}^1$ convergence in the case when there is a spectral gap. Let $\psi \in \mathcal{L}^1(\Gamma) \cap \mathcal{L}^2(\Gamma)$ be any initial density (with respect to the measure Γ). For the rest of this section, all norms and inner products will be with respect to $\mathcal{L}^2(\Gamma)$. Writing $\lambda_1^\kappa := \inf\{\Sigma(L^\kappa) \setminus \{\lambda_0^\kappa\}\}$, from our earlier results we have that

$$
\begin{aligned}
\left\| e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi - \langle \psi, \varphi \rangle \varphi \right\|^2 &= \left\| \int_{\lambda_1^\kappa}^\infty e^{-t(\lambda - \lambda_0^\kappa)} \, dE_\lambda^\kappa \psi \right\|^2 \\
&= \int_{\lambda_1^\kappa}^\infty e^{-t \cdot 2(\lambda - \lambda_0^\kappa)} \, d\langle E_\lambda^\kappa \psi, \psi \rangle \\
&\leq \|\psi\|^2 \cdot e^{-t \cdot 2(\lambda_1^\kappa - \lambda_0^\kappa)}.
\end{aligned}
$$
(5.1)

We now link this to $\mathcal{L}^1$ convergence. Let $H \subset \mathbb{R}^d$ be a compact set. From the Cauchy–Schwarz inequality, we know that

$$
\begin{aligned}
\int_H &\left| e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi(y) - \langle \psi, \varphi \rangle \varphi(y) \right| d\Gamma(y) \\
&\leq \left\| e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi - \langle \psi, \varphi \rangle \varphi \right\| \cdot \Gamma(H)^{1/2} \\
&\leq \|\psi\| \cdot \Gamma(H)^{1/2} \cdot e^{-t(\lambda_1^\kappa - \lambda_0^\kappa)}.
\end{aligned}
$$

So we have the appropriate convergence in $\mathcal{L}^1(\Gamma)$ on compact sets. We could similarly obtain convergence for test functions $f \in \mathcal{L}^2(\Gamma)$, that is,

$$
\left| \langle e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi, f \rangle - \langle \psi, \varphi \rangle \langle \varphi, f \rangle \right| \leq \|\psi\| \cdot \|f\| \cdot e^{-t(\lambda_1^\kappa - \lambda_0^\kappa)}.
$$
(5.2)

We see that when Γ is a finite measure, we will obtain $\mathcal{L}^1$ convergence at this rate on all measurable sets, not just compact ones. This is the case when the (unkilled) diffusion has a strong inward drift.

Now assume that $\Gamma(\mathbb{R}^d) < \infty$ and fix some $E \subset \mathbb{R}^d$. Writing $\mathbb{P}_\psi$ for the law of the killed process starting from ψ, we have [recalling that $\int_{\mathbb{R}^d} \pi(x) \, dx = \langle \varphi, 1 \rangle$],

$$
\begin{aligned}
&\left| \mathbb{P}_\psi(X \in E | \tau_\partial > t) - \pi(E) \right| \\
&= \left| \int_E \left(\frac{e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi(y)}{e^{t\lambda_0^\kappa} \mathbb{P}_\psi(\tau_\partial > t)} - \frac{\varphi(y)}{\int_{\mathbb{R}^d} \pi(x) \, dx} \right) d\Gamma(y) \right|
\end{aligned}
$$

$$= \frac{1}{\langle \psi, \varphi \rangle \int \pi(x)\, dx} \left| \int_E \left(e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi(y) - \langle \psi, \varphi \rangle \varphi(y) \right) d\Gamma(y) \right.$$

$$\left. - \left(\int_E e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi(y)\, d\Gamma(y) \right) \left(\frac{e^{t\lambda_0^\kappa} \mathbb{P}_\psi(\tau_\partial > t) - \langle \psi, \varphi \rangle \int \pi(x)\, dx}{e^{t\lambda_0^\kappa} \mathbb{P}_\psi(\tau_\partial > t)} \right) \right|.$$

Note that

$$\frac{\int_E e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi(y)\, d\Gamma(y)}{e^{t\lambda_0^\kappa} \mathbb{P}_\psi(\tau_\partial > t)} = \frac{\langle e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi, \mathbf{1}_E \rangle}{\langle e^{t\lambda_0^\kappa} e^{-tL^\kappa} \psi, \mathbf{1} \rangle} \le 1,$$

so

$$(5.3) \qquad \left| \mathbb{P}_\psi(X \in E | \tau_\partial > t) - \pi(E) \right| \le \frac{2\|\psi\| \Gamma(\mathbb{R}^d)^{1/2}}{\langle \psi, \varphi \rangle \int \pi(x)\, dx} e^{-t(\lambda_1^\kappa - \lambda_0^\kappa)}.$$

It remains to derive the rate of pointwise convergence for

$$e^{t\lambda_0^\kappa} p^\kappa(t, x, y) \to \varphi(x)\varphi(y) \qquad \text{as } t \to \infty.$$

This argument does not require us to assume $\Gamma(\mathbb{R}^d) < \infty$. Following the approach of [23], for $x, y \in \mathbb{R}^d$ let us write $g_x(y) := e^{\lambda_0^\kappa} p^\kappa(1, x, y)$. First, note that $g_x \in \mathcal{L}^2(\Gamma)$:

$$\|g_x\|_{\mathcal{L}^2(\Gamma)}^2 = e^{2\lambda_0^\kappa} \int p^\kappa(1, x, y) p^\kappa(1, x, y)\, d\Gamma(y)$$

$$= e^{2\lambda_0^\kappa} \int p^\kappa(1, x, y) p^\kappa(1, y, x)\, d\Gamma(y)$$

$$= e^{2\lambda_0^\kappa} p^\kappa(2, x, x)$$

$$< \infty,$$

using symmetry and the semi-group property. By the invariance of φ,

$$\langle g_x, \varphi \rangle = e^{\lambda_0^\kappa} \int p^\kappa(1, x, z)\varphi(z)\, d\Gamma(z) = \varphi(x).$$

Now for $t > 2$, $x, y \in \mathbb{R}^d$,

$$e^{t\lambda_0^\kappa} p^\kappa(t, x, y) = \int \int g_x(z) e^{\lambda_0^\kappa(t-2)} p^\kappa(t - 2, z, w) g_y(w)\, d\Gamma(z)\, d\Gamma(w)$$

$$= \langle e^{-(t-2)L^\kappa} e^{(t-2)\lambda_0^\kappa} g_x, g_y \rangle.$$

By (5.2), this converges to

$$\langle g_x, \varphi \rangle \langle g_y, \varphi \rangle = \varphi(x)\varphi(y),$$

with rate given by

$$(5.4) \quad \left| e^{t\lambda_0^\kappa} p^\kappa(t, x, y) - \varphi(x)\varphi(y) \right| \le e^{2\lambda_0^\kappa} \left(p^\kappa(2, x, x) p^\kappa(2, y, y) \right)^{1/2} e^{-t(\lambda_1^\kappa - \lambda_0^\kappa)}.$$

If the drift is bounded then the transition density is bounded as well, so this is bounded by $Ce^{-t(\lambda_1^\kappa - \lambda_0^\kappa)}$ for a universal constant C.

6. Discussion. In this paper, we have proven natural sufficient conditions for the quasi-limiting distribution of a diffusion of the form (1.4) killed at an appropriate state-dependent rate to coincide with a target density π. We have also quantified the rate of convergence to quasi-stationarity by relating the rate of this convergence to the rate of convergence to stationarity of a related unkilled process.

As mentioned in the Introduction, this framework is foundational for the recently-developed class of quasi-stationary Monte Carlo algorithms to sample from Bayesian posterior distributions, introduced in [20]. This framework promises improvement over more traditional MCMC approaches particularly for Bayesian inference on large datasets, since the killed diffusion framework enables the use of *subsampling* techniques. As detailed in [20], Section 4, these allow the construction of estimators which scale exceptionally well as the size of the underlying dataset grows.

Quasi-stationary Monte Carlo methods are likely to be particularly effective compared to established Monte Carlo methods for Bayesian inference for *tall data*; that is, where parameter spaces have moderate dimension (allowing diffusion simulation to be feasible) but where data sizes are high. This includes the 'Big-Data' context where data size is so large it cannot even be stored locally on computers implementing the algorithm. This is because subsampling can take place 'offline' with only the subsets being stored locally. This adds significantly to the potential applicability of quasi-stationary Monte Carlo methods.

Our approach in this present work is also slightly more general than that of [20] in that we allow for a nonzero drift term in our diffusion (1.4). This raises the question of how to select among several possible drift functions the one that results in the most practical computational outcomes. While a detailed answer to this question is beyond the scope of this present work, we suggest the following guidelines. There is a critical trade-off between overall killing and the essential rate of convergence described in (1.6). As mentioned in the example of Section 2, higher rates of killing will tend to *increase* the essential rate of convergence, while increasing the computational burden imposed by simulating killing events. Depending on the details of the implementation, this trade-off could go either way in terms of optimality. When scalable estimators for the killing events are available, such as in [20], it would be sensible to choose a drift that makes the killing rate high, for instance choosing a Brownian motion, so $A \equiv 0$. Of course, any Gaussian process allows straightforward simulation of the unkilled dynamics, and the choice of Brownian motion also simplifies Assumptions 2 and 3. Formally answering this question of the choice of drift would be an interesting avenue for future exploration.

We comment briefly now on some of our assumptions. Assumption 2 is generally straightforward to verify, especially in light of the stronger 'rejection sampling' formulation in (1.7). For instance, if A is uniformly bounded below then Assumption 2 holds if π is a bounded density function.

Assumption 3 is generally the most challenging. When $A \equiv 0$, this is mostly straightforward to verify, especially since densities on $\mathbb{R}^d$ are often convex in the tails. Verifying Assumption 3 in general can be done using the equivalent expression for $\tilde{\kappa}$ in (1.10), by comparing the decay of derivatives $2A$ with those of $\log \pi$. Indeed, ensuring a practically useful form of $\tilde{\kappa}$—so that verification of Assumption 3 is straightforward—could influence the choice of A in the first place.

In practice, Assumption 3 also involves computing a lower bound for $\tilde{\kappa}$. It is actually not necessary to compute the precise value of $\inf_{y \in \mathbb{R}^d} \tilde{\kappa}(y)$; our results still hold if K in (1.9) is replaced by any constant such that the resulting κ is nonnegative everywhere. Intuitively, taking a larger constant K amounts to merely adding additional killing events according to a homogeneous, independent Poisson process.

Depending on the choices of π and A, $\tilde{\kappa}$ can be a convex function in the tails, even in cases of nonzero A, as in our example of Section 2. A precise recipe for computing K in general is currently unavailable; readers interested in these more implementational details are encouraged to look at [20].

We conclude this discussion by indicating some potential future directions. As mentioned above, there are important questions of how to choose the underlying diffusion to optimize the computation for a given target density. One could also consider extensions of this work to entirely different underlying processes, such as jump diffusions or Lévy processes. Finally, another potential question is the exploration of alternative approaches to that described in [20] for the simulation of quasi-stationary distributions, such as the stochastic approximation approaches as discussed in [4] and [2].

REFERENCES

[1] BARDENET, R., DOUCET, A. and HOLMES, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res*. **18** Paper No. 47, 43. MR3670492

[2] BENAIM, M., CLOEZ, B. and PANLOUP, F. (2018). Stochastic approximation of quasi-stationary distributions on compact spaces and applications. *Ann. Appl. Probab*. **28** 2370–2416. MR3843832

[3] BESKOS, A., PAPASPILIOPOULOS, O. and ROBERTS, G. O. (2006). Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli* **12** 1077–1098. MR2274855

[4] BLANCHET, J., GLYNN, P. and ZHENG, S. (2016). Analysis of a stochastic approximation algorithm for computing quasi-stationary distributions. *Adv. in Appl. Probab*. **48** 792–811. MR3568892

[5] BRAVERMAN, M., MILATOVICH, O. and SHUBIN, M. (2002). Essential selfadjointness of Schrödinger-type operators on manifolds. *Uspekhi Mat. Nauk* **57** 3–58. MR1942115

[6] CHAMPAGNAT, N. and VILLEMONAIS, D. (2016). Exponential convergence to quasi-stationary distribution and Q-process. *Probab. Theory Related Fields* **164** 243–283. MR3449390

[7] COLLET, P., MARTÍNEZ, S. and SAN MARTÍN, J. (2013). *Quasi-Stationary Distributions*: *Markov Chains*, *Diffusions and Dynamical Systems*. Springer, Heidelberg. MR2986807

[8] DALALYAN, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **79** 651–676. MR3641401

[9] DAVIES, E. B. (1995). *Spectral Theory and Differential Operators*. *Cambridge Studies in Advanced Mathematics* **42**. Cambridge Univ. Press, Cambridge. MR1349825

[10] DEMUTH, M. and VAN CASTEREN, J. A. (2000). *Stochastic Spectral Theory for Selfadjoint Feller Operators*: *A Functional Integration Approach*. Birkhäuser, Basel. MR1772266

[11] DIACONIS, P. and MICLO, L. (2015). On quantitative convergence to quasi-stationarity. *Ann. Fac. Sci. Toulouse Math.* (6) **24** 973–1016. MR3434264

[12] DURMUS, A. and MOULINES, É. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.* **27** 1551–1587. MR3678479

[13] KOLB, M. and STEINSALTZ, D. (2012). Quasilimiting behavior for one-dimensional diffusions with killing. *Ann. Probab.* **40** 162–212. MR2917771

[14] MANDL, P. (1961). Spectral theory of semi-groups connected with diffusion processes and its application. *Czechoslovak Math. J.* **11 (86)** 558–569. MR0137143

[15] METAFUNE, G., PALLARA, D. and PRIOLA, E. (2002). Spectrum of Ornstein–Uhlenbeck operators in L^p spaces with respect to invariant measures. *J. Funct. Anal.* **196** 40–60. MR1941990

[16] OUHABAZ, E. M. (2005). *Analysis of Heat Equations on Domains*. *London Mathematical Society Monographs Series* **31**. Princeton Univ. Press, Princeton, NJ. MR2124040

[17] PINSKY, R. G. (1995). *Positive Harmonic Functions and Diffusion*. *Cambridge Studies in Advanced Mathematics* **45**. Cambridge Univ. Press, Cambridge. MR1326606

[18] PINSKY, R. G. (2009). Explicit and almost explicit spectral calculations for diffusion operators. *J. Funct. Anal.* **256** 3279–3312. MR2504526

[19] POLLETT, P. K. (2015). Quasi-stationary distributions: A bibliography. http://www.maths.uq.edu.au/~pkp/papers/qsds/qsds.html.

[20] POLLOCK, M., FEARNHEAD, P., JOHANSEN, A. M. and ROBERTS, G. O. (2016). The scalable Langevin exact algorithm: Bayesian inference for big data. Preprint. Available at arXiv:1609.03436.

[21] REED, M. and SIMON, B. (1978). *Methods of Modern Mathematical Physics*. *IV*. *Analysis of Operators*. Academic Press, New York. MR0493421

[22] ROBERTS, G. O. and TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273

[23] SIMON, B. (1993). Large time behavior of the heat kernel: On a theorem of Chavel and Karp. *Proc. Amer. Math. Soc.* **118** 513–514. MR1139473

[24] TUOMINEN, P. and TWEEDIE, R. L. (1979). Exponential decay and ergodicity of general Markov processes and their discrete skeletons. *Adv. in Appl. Probab.* **11** 784–803. MR0544195

[25] TWEEDIE, R. L. (1974). *R*-theory for Markov chains on a general state space. I. Solidarity properties and *R*-recurrent chains. *Ann. Probab.* **2** 840–864. MR0368151

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

A. Q. WANG
D. STEINSALTZ
DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
24-29 ST GILES
OXFORD OX1 3LB
UNITED KINGDOM
E-MAIL: a.wang@stats.ox.ac.uk
          steinsal@stats.ox.ac.uk

M. KOLB
DEPARTMENT OF MATHEMATICS
PADERBORN UNIVERSITY
WARBURGER STR. 100
33098 PADERBORN
GERMANY
E-MAIL: kolb@math.uni-paderborn.de

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