arXiv:2411.11403v3 [math.NA] 18 Dec 2025

Hadamard–Langevin dynamics for sampling the ℓ_1-prior

Ivan Cheltsov* Federico Cornalba* Clarice Poon† Tony Shardlow*

Abstract

Priors with non-smooth log-densities, such as the ℓ_1-prior, are widely used in Bayesian inverse problems for their sparsity-inducing properties. Existing Langevin-based sampling methods typically rely on proximal mappings or smooth approximations, which alter the target distribution. We propose an alternative approach based on a Hadamard product parameterization of the ℓ_1-norm, leading to a smooth but nonconvex and non–globally Lipschitz potential whose marginal law exactly recovers the desired posterior. The resulting Hadamard–Langevin dynamics (HLD) defines a diffusion process that is analytically distinct from proximal or mirror-type Langevin schemes. Our main contribution is a rigorous well-posedness theory for both the continuous and discrete HLD. We establish existence and uniqueness of strong solutions, geometric ergodicity of the continuous dynamics, and convergence of the discretized scheme as the step size tends to zero. These results provide the first theoretical foundation for sampling from nonconvex, nonsmooth posteriors through overparameterized Langevin dynamics.

1 Introduction

Sampling from sparsity-promoting distributions is a central task in high-dimensional Bayesian inference and inverse problems. A prototypical example is

$$\rho(x) = \frac{1}{Z}\exp\big(-\beta\big(\lambda\|x\|_1 + G(x)\big)\big), \qquad x \in \mathbb{R}^d, \tag{1}$$

where Z is a normalization constant, $\beta > 0$ is the inverse temperature, λ is a regularization parameter, and $G\colon \mathbb{R}^d \to \mathbb{R}$ represents the data likelihood. Due to the non-smoothness of the ℓ_1-term, existing Langevin-type methods typically rely on smoothing techniques such as the Moreau–Yosida envelope or proximal mappings.

In this work, we propose and theoretically analyse the Hadamard–Langevin dynamics (HLD) — a novel smooth overparameterized dynamics whose marginal stationary distribution

*University of Bath, ic488,fc402,t.shardlow@bath.ac.uk
†Mathematics Institute, University of Warwick, clarice.poon@warwick.ac.uk

exactly recovers ρ without smoothing and bias. The key idea is to exploit the Hadamard representation of the ℓ_1-norm,

$$\|x\|_1 = \sum_i |x_i| = \min_{u \odot v = x} \left\{ \frac{1}{2}\|u\|_2^2 + \frac{1}{2}\|v\|_2^2 \right\}, \tag{2}$$

which recasts the non-smooth prior as a smooth but nonconvex term. This representation, which has recently garnered interest in optimization to study implicit regularization and to develop efficient optimization solvers, has not yet been explored from a sampling perspective. We demonstrate that applying Langevin dynamics in the lifted space (u, v) leads to a well-defined diffusion process whose invariant measure projects exactly onto ρ.

1.1 The Hadamard-Langevin dynamics

Our starting point is the classical Gaussian-mixture identity underlying the hierarchical Bayesian formulation of the Laplace prior ([39]),

$$\exp(-a|z|) = \frac{a}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{\eta}} \exp\left(-\frac{z^2}{2\eta} - \frac{a^2\eta}{2} \right) d\eta, \qquad a \in \mathbb{R}_+,\ z \in \mathbb{R}. \tag{3}$$

Introducing latent variables $\eta_i > 0$ for each coordinate of x, one can rewrite the target distribution (1) as the marginal of a joint law over $(x, \eta) \in \mathbb{R}^d \times \mathbb{R}_+^d$:

$$\tilde{\pi}(x, \eta) \propto \frac{1}{\prod_{i=1}^d \sqrt{\eta_i}} \exp\left(-\sum_{i=1}^d \left(\frac{x_i^2}{2\eta_i} + \frac{\beta^2\lambda^2\eta_i}{2} \right) - \beta\, G(x) \right). \tag{4}$$

Since sampling from $\tilde{\pi}$ yields the correct marginal $\rho(x)$, [32] proposed a Gibbs sampler where one alternates between sampling x and η. However, such works typically restrict to the case where G is quadratic, so sampling x corresponds to sampling a Gaussian, and this in turn requires inverting a potentially high-dimensional matrix. In contrast, as we explain next, we focus on a reparameterized formulation of the above, which will allow us to derive a Langevin algorithm, under general assumptions on the data term G.

1.1.1 From hierachical models to Hadamard overparameterization

A key observation is that the Gaussian mixture representation can be viewed as the stochastic analogue of the quadratic variational form for the ℓ_1-norm. By applying a change of variables of

$$x = u \odot v, \quad \eta = u \odot u/(\lambda\beta)$$

to (4), we obtain the Hadamard parametrized joint law

$$\pi(u, v) = \frac{1}{Z_\pi} \prod_{i=1}^d u_i \exp\left(-\beta \left(\frac{1}{2}\lambda\|(u, v)\|^2 + G(u \odot v) \right) \right) \tag{5}$$

over $(u, v) \in \mathcal{X} := \mathbb{R}^d_+ \times \mathbb{R}^d$ for normalization constant Z_π and $\|(u, v)\|^2 = \|u\|^2 + \|v\|^2$. The Hadamard product $x = u \odot v$ follows the same distribution ρ defined in Eq. (1). This is formalized in the following statement.

Theorem 1.1. *If $(u, v) \sim \pi$ as defined in Eq. (5), then $x = u \odot v \sim \rho$ as defined in Eq. (1). In particular, for any smooth $\phi : \mathbb{R}^d \to \mathbb{R}$,*

$$\int_{\mathcal{X}} \phi(u \odot v)\, \pi(u, v)\, du\, dv = \int_{\mathbb{R}^d} \phi(x)\, \rho(x)\, dx.$$

Furthermore, the normalization factors Z_π and Z of π and ρ, respectively, satisfy $\frac{1}{Z} = \frac{1}{Z_\pi} \left(\sqrt{\frac{\pi}{2\beta\lambda}}\right)^d$.

Note that the log-density is smooth except for $\log(u_i)$ terms, which we show can be handled easily. It is therefore natural to derive Langevin dynamics (see (7)) to sample π, and recover samples of ρ by evaluating $u \odot v$.

1.2 Main contributions

By treating ρ using the Hadamard parameterization to obtain a differentiable log-density, we demonstrate that we can apply Langevin sampling methods to good effect. We provide a complete theoretical analysis of the Hadamard–Langevin dynamics by establishing well-posedness and stability of the continuous and discretized dynamics.

1. We establish the well-posedness on $\mathcal{X} = \mathbb{R}^d_+ \times \mathbb{R}^d$ of the Langevin system associated with π given in (5). This includes geometric convergence to the invariant measure (Section 2). We present two approaches: The first approach removes the singularity in the $\log(u)$-term via a Cartesian change of coordinates, to transform our drift into a locally bounded term. The second approach exploits connections of our Langevin system to the well-studied Cox–Ingersoll–Ross and Ornstein–Uhlenbeck processes via a Girsanov transformation. The Girsanov transformation also ensures that u_t will be positive almost surely.

2. We develop a numerical scheme to approximate the Langevin system and show the convergence of the numerical scheme to π (Section 3). We establish strong convergence of the method as well as discuss convergence of the stationary distribution of the numerical method. One of the challenges in the analysis is that the negative log-density is neither convex nor globally Lipschitz; despite this, we are able to prove geometric ergodicity of the numerical scheme by relying on two tools, namely a coupling technique for different realisations of the numerical method inside sufficiently large balls, and a rapidly-decaying bound for the escape probability from such balls.

As mentioned, the Gaussian mixture formula Eq. (3) can be seen as the equivalent of the quadratic variational form for sampling and our work can be seen as a first attempt to

deploy this trick for Langevin dynamics. Even though we restrict to sampling the ℓ_1-prior, as discussed in the conclusion, we believe that our work offers a new perspective on sampling sparse priors and future work would aim to generalize this to other structure-inducing priors.

1.3 Related background

Quadratic variational forms

It is well-known that many of the sparsity-inducing regularizers, and in particular, the ℓ_1-norm, have *quadratic variational forms* [3, 1]. In particular, if the regulariser $R(x) = \phi(x \odot x)$ for $x \in \mathbb{R}^d$ and a concave function ϕ, one can rewrite using quadratic variational forms

$$R(x) = \inf_{\eta \in \mathbb{R}^d_+} \sum_{i=1}^{d} \frac{x_i^2}{\eta_i} + \psi(\eta), \tag{6}$$

for some function ψ. This includes the ℓ_q (semi)norms for $q \in (0, 2]$ and various nonconvex regularizers such as the Lorentzian function [3]; a similar formula also holds for functionals on matrices, such as the nuclear norm [1]. In the case of $R(x) = \|x\|_1$, one can choose $\psi(\eta) = \eta/4$.

This idea has been exploited to good effect in optimization, and the computer-vision community found the new variable η can help increase the curvature of a 'flat' landscape and has led to graduated non-convexity algorithms [3]. This lifting of the ℓ_1-norm to the x and η variables is the basis of iteratively reweighted least-squares algorithms [7], where one alternates between solving least squares problems on x and exact solves on η.

The Hadamard product parameterization

More recently, there has been renewed interest in the Hadamard form of these quadratic variational formulas (6), particularly in the context of convex structure-inducing regularizers. For $\psi(\eta) = \eta/4$, by making a change of variable with $x = u \odot v$ and $\eta/2 = u \odot u$, one can equivalently write (6) as $\|x\|_1 = \inf_{x=u\odot v}\{\frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2\}$ where $\|\cdot\|$ is the Euclidean 2-norm. While the formulation in (6) retained convexity in both variables x and η, this Hadamard product-like formula is no longer convex but is now smooth. Consequently, one can make use of powerful smooth optimization solvers such as quasi-Newton methods. One of the first works that replace the ℓ_1-norm with the Hadamard product is the work of [20]; see also [35, 34] for further discussions on exploiting these forms for optimization. In the case of the ℓ_1-norm, one can show that this nonconvexity is in fact benign [34]. Furthermore, it can be shown that gradient descent with the Hadamard parameterization is implicitly linked to mirror descent with a hyperbolic entropy potential [35]. This points to a potential benefit of such parameterizations for modern optimization when working on high-dimensional and ill-conditioned problems, where [5] investigated the superior convergence behaviour of such mirror decent algorithms in the context of sparse measures. Finally, over-parameterization and its implications for implicit bias in gradient-descent training of neural networks have also been of wide interest in the machine learning community. In particular, the Hadamard

product parameterization, even without explicit regularization, causes gradient descent to bias towards sparse solutions [38, 6].

Unadjusted Langevin algorithms (ULA) and Moreau–Yosida envelopes

Langevin dynamics is a simple way to sample from the posterior distribution $\rho(x)$ [36]. Unadjusted Langevin Algorithm (ULA), in the case where $\rho(x) \propto \exp(-\beta H(x))$ for a *smooth* potential $H\colon \mathbb{R}^d \to \mathbb{R}$, is the update rule $x^{n+1} = x^n - \nabla H(x^n)\Delta t + \sqrt{2\Delta t/\beta}\,\xi^n$, for independent samples $\xi^n \sim \mathcal{N}(0, \mathrm{Id}_d)$ and time step $\Delta t > 0$. This can be interpreted as a discretization $x^n \approx x_{n\Delta t}$ of the overdamped Langevin stochastic differential equation (SDE): $dx_t = -\nabla H(x_t)\,dt + \sqrt{2/\beta}\,dW_t$, where W_t is a standard $\mathbb{R}^d$-valued Wiener process.

In the case where $H(x) = G(x) + R(x)$ and R is non-smooth (as in our case, see (1)), a canonical approach is to replace R by a smooth approximation defined by the Moreau–Yosida envelope R_γ [33]: $R_\gamma(x) := \min_{z\in\mathbb{R}^d}(R(z) + \frac{1}{2\gamma}\|x - z\|^2)$. The function R_γ is then $1/\gamma$-smooth and converges pointwise to R as $\gamma \to 0$. One can then directly apply ULA to $H_\gamma := G + R_\gamma$, leading to the Moreau–Yosida regularized Unadjusted Langevin Algorithm (MYULA). Convergence to the stationary distribution requires one to take $\gamma \to 0$ [11]. Under suitable conditions, such as convexity of G and Lipschitz smoothness, convergence results like geometric ergodicity of the stationary distribution $\rho_\gamma \propto \exp(-\beta H_\gamma)$ have been established, as well as the convergence of ρ_γ to ρ as $\gamma \to 0$. While early works replaces R by its Moreau envelope, [10] removes this approximation bias by directly incorporating the proximal operator of R. Another line of work that avoids the bias from the Moreau-envelope smoothing directly samples from H via subgradient Langevin [13] in the case where H is strongly convex. In general, smoothing by the Moreau envelope allows one to leverage all existing theory for smooth Langevin dynamics, while more recent works derive cleaner algorithms and proofs by heavily relying on techniques from convex analysis.

2 Continuous dynamics

We work with G under the following assumptions.

Assumption 1 (data term G)**.** Assume that (i) G is bounded below, that (ii) $G\colon \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, that (iii) $x^\top \nabla G(x)$ is bounded below (so $x^\top \nabla G(x) \geq -K$ for some K), and that (iv) $\|\nabla G(x)\|_\infty$ is bounded uniformly for $x \in \mathbb{R}^d$ (so $B := \sup_x \|\nabla G(x)\|_\infty < \infty$).

The data term G is bounded below so that ρ in (1) is a probability measure for some normalization constant $Z > 0$. Note that (iii) holds if G is a convex function. Important examples that satisfy (i), (ii) and (iii) include the quadratic loss $G(x) = \|Ax - y\|^2$ for an $m \times d$ matrix A and $y \in \mathbb{R}^m$; the logistic loss $G(x) = \sum_{i=1}^m \log(1 + \exp(-y_i A_i^\top x))$ for $A_i \in \mathbb{R}^d$; and smooth approximations of the unit loss, such as $G(x) = \sum_i \sigma(y_i A_i^\top x)$ for $\sigma(x) = \arctan(x)$ or $\sigma(x) = 1/(1 + \exp(x))$. Note that $G(x) = \sum_i \sigma(y_i A_i^\top x)$ satisfies condition (iv) for any function σ with bounded derivatives, which includes the logistic loss and smooth versions of the unit loss. The least-squares loss $G(x) = \|Ax - y\|^2$ does not satisfy condition (iv) on

the boundedness of ∇G, and one might consider instead the Huber loss. However, in our numerical computations, the method behaves well with the quadratic loss, so assumption (iv) is likely a limitation of our theoretical analysis rather than the method itself.

2.1 Langevin sampling

We will use Langevin sampling to produce samples from π. Write $\pi \propto \exp(-\beta H)$ for $H((u,v)) = \frac{1}{2}\lambda \|(u,v)\|^2 + G(u \odot v) - \frac{1}{\beta}\sum_i \log u_i$. The gradient is $\nabla H((u,v)) = \lambda \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} v \\ u \end{pmatrix} \odot \nabla G(u \odot v) - \begin{pmatrix} 1/\beta u \\ 0 \end{pmatrix}$, where we interpret $1/u \in \mathbb{R}^d$ as the vector with entries $1/u_i$. Therefore, the corresponding Langevin equations are

$$\begin{aligned} du_t &= -\lambda\, u_t\, dt - v_t \odot \nabla G(u_t \odot v_t)\, dt + \frac{1}{\beta\, u_t}\, dt + \sqrt{2/\beta}\, dW_t^1, \\ dv_t &= -\lambda\, v_t\, dt - \ u_t \odot \nabla G(u_t \odot v_t)\, dt + \sqrt{2/\beta}\, dW_t^2, \end{aligned} \tag{7}$$

for independent $\mathbb{R}^d$-valued Brownian motions W_t^1 and W_t^2. It is well known that π given in (5) is an invariant distribution of Eq. (7) and we explore ergodicity in Section 2.4. We will show that Eq. (7) is well posed on $\mathcal{X} = \mathbb{R}^d_+ \times \mathbb{R}^d$.

The variables (u,v) in Eq. (7) can be viewed as the cylindrical co-ordinates of an axially symmetric distribution $\rho(t,(u,v))$ with axial co-ordinate $v \in \mathbb{R}^d$ and radial co-ordinate $u \in [0,\infty)^d$. We reformulate the same equations in Cartesian coordinates $(y,z,v) \in \mathbb{R}^{3d}$, where $u = (y \odot y + z \odot z)^{1/2}$ (for the element-wise square root). For $\mathbf{x} = [y,z,v] \in \mathbb{R}^{3d}$, consider

$$d\mathbf{x} = \Big[-\lambda \mathbf{x} - \nabla G_3(\mathbf{x})\Big]\, dt + \sqrt{2/\beta}\, dW^3(t), \tag{8}$$

for $G_3(\mathbf{x}) = G((y \odot y + z \odot z)^{1/2} \odot v)$ and an $\mathbb{R}^{3d}$-valued Brownian motion $W^3(t)$. The function G_3 is locally bounded under Assumption 1, and solutions of Eq. (8) are well defined. We use these as a tool to analyze Eq. (7).

Let $P_t(X_0, B) = \mathbb{P}(\mathbf{x}_t \in B \mid \mathbf{x}_0 = X_0)$. Recall that a process is irreducible if for any Borel set B with positive Lebesgue measure and any X_0, there exists $t > 0$ so that $P_t(X_0, B) > 0$, and strong Feller if $P_t(X_0, B)$ is continuous in X_0 for all Borel sets B. We say a function $\mu(y,z,v)$ is cylindrical on $\mathbb{R}^{3d}$ if for some $\tilde{\mu}\colon \mathbb{R}^{2d} \to \mathbb{R}$ we have $\mu(y,z,v) = \tilde{\mu}(u,v)$ for all $u = ((y \odot y) + (z \odot z))^{1/2}$.

Theorem 2.1. *Suppose that Assumption 1 holds. Subject to initial condition $(u,v)_0 = (U_0, V_0) \in \mathcal{X} = \mathbb{R}^d_+ \times \mathbb{R}^d$, there is a unique solution $(u,v)_t$ to Eq. (7) such that $u_t > 0$ almost surely and $\mathbb{E}\big[\sup_{0\le t\le T} \|(u,v)_t\|^q\big] < \infty$, for any $T > 0$, $q \ge 2$. Further, $(u,v)_t$ is strong Feller and irreducible on $\mathcal{X}$.*

The probability density function $p(t,\cdot)$ of $\mathbf{x}_t$ (see Eq. (8)) is well-defined. If the initial distribution is cylindrical, the marginal distribution of $((x_t^2 + y_t^2)^{1/2}, v_t)$ equals the law of the solution (u_t, v_t) of Eq. (7).

Proof. [37, Corollary 10.1.4] gives that $\mathbf{x}_t$ is strong Feller and has a well-defined probability density function. To apply this result, note that [37, Equation (1.7) of Theorem 10.1.3] holds as the regularised drift and diffusion (defined by multiplying by a compactly supported mollifying function) are globally Lipschitz in our case. Due to the non-degenerate and additive noise, $\mathbf{x}_t$ is irreducible by standard arguments [2, Section 6].

The generator for Eq. (7) has singular coefficients on a half space and regularity of the density does not follow directly from standard theories; we show that the u-equation in Eq. (7) arises from the solution $\mathbf{x}_t$ of Eq. (8) and that the process $(u, v)_t$ inherits properties from $\mathbf{x}_t$. Assuming that $u_t \neq 0$, we may apply the Ito formula to $u = (y \odot y + z \odot z)^{1/2}$ and obtain

$$\begin{aligned} du_t &= \frac{1}{u_t} \odot (y_t \odot (-\lambda\, y_t) + z_t \odot (-\lambda\, z_t))\, dt - \frac{1}{u_t}(y_t \odot (y_t/u_t) + z_t \odot (z_t/u_t)) \odot v_t \odot \nabla G(u_t \odot v_t)\, dt \\ &\quad + \sqrt{\frac{2}{\beta}} \frac{1}{u_t} \odot \begin{pmatrix} y_t \\ z_t \end{pmatrix} \cdot \begin{pmatrix} dW_t^{3,y} \\ dW_t^{3,z} \end{pmatrix} + \frac{1}{2}\frac{2}{\beta} \left(\frac{-1}{u^3} \odot (y_t \odot y_t + z_t \odot z_t) + \frac{2}{u_t} \right) dt \\ &= -\lambda\, u_t\, dt - v_t \odot \nabla G(u_t \odot v_t)\, dt + \frac{1}{\beta\, u_t}\, dt + \sqrt{\frac{2}{\beta}}\, dW_t, \end{aligned}$$

where the $\mathbb{R}^d$-valued process $W(t) := u_t^{-1} \odot (y_t, z_t) \cdot (dW_t^{3,y}, dW_t^{3,z})$ is again a Brownian motion due to the identity $u_t^2 = y_t^2 + z_t^2$ [23, Theorem 3.3.16]. To deal with the case $u_t = 0$, we refer the reader to [23, Proposition 3.21] (i.e., the square root function in the mapping $(y, z) \mapsto u$ is approximated by well-behaved functions and the limit is verified). This is the u-equation in Eq. (7) as required. By the uniqueness of the solutions, we see that Eq. (8) and Eq. (7) agree as long as the initial distributions agree.

The process $(u, v)_t$ inherits several properties from $\mathbf{x}_t$: namely, $(u, v)_t$ is well-posed in $\mathcal{X}$ and is strong Feller and irreducible, and the distribution $\rho(t, \cdot)$ of $(u, v)_t$ is given by $p(t, \cdot)$ if the initial distribution of $\mathbf{x}$ is cylindrical.

To derive the moment bounds, computing the Itô formula for $\|(u, v)\|^2$ for Eq. (7) gives

$$d\|u_t\|^2 = \left[-2\,\lambda\, \|u_t\|^2 - 2(u_t \odot v_t)^\top \nabla G(u_t \odot v_t) + \frac{4\,d}{\beta} \right] dt + \sqrt{\frac{2}{\beta}} 2 u_t^\top dW_t^1, \tag{9}$$

$$d\|v_t\|^2 = \left[-2\,\lambda\, \|v_t\|^2 - 2(u_t \odot v_t)^\top \nabla G(u \odot v) + \frac{2\,d}{\beta} \right] dt + \sqrt{\frac{2}{\beta}} 2 v_t^\top dW_t^2. \tag{10}$$

Under Assumption 1, $x^\top \nabla G(x)$ is bounded below and standard arguments ([27, Chapter 2, Theorem 4.4] applied to the case $G = 0$, together with the comparison argument in [22]) lead to uniform-in-time bounds $\mathbf{E}[\sup_{t\in[0,T]} \|(u, v)\|^2] \leq C(\lambda, \beta, d)$. This shows that the second moment is finite and $(u, v)_t$ does not blow-up on $[0, T]$ almost surely. Similar arguments apply to the q-moments by applying Ito's formula to $(\|(u, v)_t\|^2)^{q/2}$. □

Let $L^q(\pi)$ denote the set of integrable functions $\phi\colon \mathcal{X} \to \mathbb{R}$ such that $\int |\phi(x)|^q \pi(x)\, dx < \infty$, where π is the invariant measure Eq. (5). In the following, we show how to control expectations of $\phi((u, v)_t)$, which extends the moment bounds in Theorem 2.1 to $\phi \in L^q(\pi)$, and includes some ϕ that are singular near $u_i = 0$.

Proposition 2.2. *Let $(u,v)_t$ denote the solution of Eq. (7) with initial distribution ρ_0. Let $p \geq 1$. Assume that (p,q) are conjugate and that $h_0 = \rho_0/\pi$ belongs to $L^p(\pi)$. If $\phi \in L^q(\pi)$, then $\mathbb{E}[\phi((u,v)_t)] \leq \|\phi\|_{L^q(\pi)} \|h_0\|_{L^p(\pi)}$ for $t > 0$.*

Proof. Let $\rho(t,\cdot)$ denote the density of $(u,v)_t$. Denote $h = \rho/\pi$ and $h_0 = \rho_0/\pi$. It is well known (e.g., [28]) that h satisfies the PDE

$$\frac{\partial h}{\partial t} = \Delta h + \nabla \log \pi \cdot \nabla h, \qquad h(0,z) = h_0(z),$$

on $t > 0$ and $z \in (0,\infty)^d \times \mathbb{R}^d$. For any twice-differentiable convex function ψ, it follows [28, Eq. 9] that

$$\frac{d}{dt} \int \psi(h(t,x))\, \pi(x)\, dx = - \int \psi''(h(t,x))\, |\nabla h(t,x)|^2\, \pi(x)\, dx \leq 0.$$

Choosing $\psi(h) = h^p$ implies that $\|h(t,\cdot)\|_{L^p(\pi)}$ is a non-increasing function of time. Thus, writing

$$\mathbb{E}[\phi((u,v)_t)] = \int \phi(x)\, h(t,x)\, \pi(x)\, dx \leq \|\phi\|_{L^q(\pi)}\, \|h(t,\cdot)\|_{L^p(\pi)} \leq \|\phi\|_{L^q(\pi)}\, \|h_0\|_{L^p(\pi)}$$

completes the proof. □

As a corollary, we prove the following moment bounds on $1/u_{1,t}^r$ for $r < 2$. The density of the invariant measure π behaves like u_1 near $u_1 = 0$, and we expect $\mathbb{E}\big[1/u_{1,t}^2\big]$ to be infinite.

Corollary 2.3. *Let the assumption of Theorem 2.1 hold and the initial distribution $\rho_0((u,v))$ be such that ρ_0/π is uniformly bounded. Then, for $i = 1,\dots,d$,*

1. $\sup_{t>0} \mathbb{E}\left[\dfrac{1}{u_{i,t}^r}\right] < \infty$ *for $0 < r < 2$ (here $u_{i,t}$ denotes the ith component of u_t).*
2. *For each $T > 0$ and $\epsilon > 0$, we have $\mathbb{E}\left[\left(\int_0^T \frac{1}{u_{i,s}^{2-\epsilon}}\, ds\right)^q\right] < \infty$ for all $1 \leq q < 2$.*

Proof. Let $\phi(u,v) = u_1^{-r}$ for example. Since $r < 2$, we can choose conjugate exponents (p,q) such that $q\,r < 2$. Due to the boundedness assumption on ρ_0/π and π being a probability measure, $\rho_0/\pi \in L^p(\pi)$. Further, $\phi \in L^q(\pi)$ as

$$\int \phi^q\, \pi = \frac{1}{Z_\pi} \int \frac{1}{u_1^{q\,r-1}}\, u_2 \cdots u_d\, \exp\left(-\beta \left(\frac{1}{2}\lambda\, \|(u,v)\|^2 + G(u \odot v)\right)\right) du\, dv$$

is finite for $q\,r - 1 < 1$ as the exponential term decays rapidly at infinity due to the assumptions on G. Consequently, we may apply Theorem 2.2 to deduce $\mathbb{E}[\phi((u,v)_t)] < \infty$ as required.

2. By Ito's formula applied to $(u,v) \mapsto u_1^\epsilon$ for $\epsilon > 0$,

$$\begin{aligned} \frac{1}{\beta}\, \epsilon^2 \int_0^T u_{1,t}^{\epsilon-2}\, dt = u_{1,T}^\epsilon - u_{1,0}^\epsilon + \lambda\, \epsilon \int_0^T u_{1,t}^\epsilon\, dt \\ + \epsilon \int_0^T v_{1,t}\, [\nabla G(u_t \odot v_t)]_1\, u_{1,t}^{\epsilon-1}\, dt - \sqrt{\frac{2}{\beta}}\, \epsilon \int_0^T u_{1,t}^{\epsilon-1}\, dW_t^{1,1}. \end{aligned} \tag{11}$$

By Young's inequality with conjugate exponents $((2-\epsilon)/(1-\epsilon), (2-\epsilon))$, we have for any $\alpha > 0$ that

$$|v_1 \left[\nabla G(u \odot v)\right]_1 u_1^{\epsilon-1}| \leq \frac{1}{2-\epsilon}\alpha^{-(2-\epsilon)}|v_1 \left[\nabla G(u \odot v)\right]_1|^{(2-\epsilon)} + \frac{1-\epsilon}{2-\epsilon}\alpha^{(2-\epsilon)/(1-\epsilon)}\, u_1^{\epsilon-2}.$$

We may set $\epsilon\, \alpha^{(2-\epsilon)/(1-\epsilon)}(1-\epsilon)/(2-\epsilon) = \epsilon^2/2\beta$ by choice of α. Hence, for a constant $c(\epsilon)$,

$$\epsilon \left| \int_0^T v_{1,t} \left[\nabla G(u_t \odot v_t)\right]_1 u_{1,t}^{\epsilon-1}\, dt \right| \leq c(\epsilon) \int_0^T |v_{1,t} \left[\nabla G(u_t \odot v_t)\right]_1|^{(2-\epsilon)}\, dt + \frac{1}{2\beta}\,\epsilon^2 \int_0^T u_{1,t}^{\epsilon-2}\, dt.$$

Substituting into Eq. (11), we obtain

$$\begin{aligned}\frac{1}{2\beta}\,\epsilon^2 \int_0^T u_{1,t}^{\epsilon-2}\, dt \leq\; & u_{1,T}^{\epsilon} - u_{1,0}^{\epsilon} + \lambda\,\epsilon \int_0^T u_{1,t}^{\epsilon}\, dt \\ & + c(\epsilon) \int_0^T |v_{1,t} \left[\nabla G(u_t \odot v_t)\right]_1|^{(2-\epsilon)}\, dt + \sqrt{\frac{2}{\beta}}\,\epsilon \left| \int_0^T u_{1,t}^{\epsilon-1}\, dW_t^{1,1} \right|.\end{aligned}$$

On the right-hand side, qth moments are finite, which we now demonstrate for the last two terms. By the Burkholder–Gundy–Davis inequality [23, Theorem 3.28] and Jensen's inequality for $1 < q \leq 2$, for a constant c,

$$\mathbb{E}\left[\left| \int_0^T u_1^{\epsilon-1}\, dW_1(t) \right|^q \right] \leq c\, \mathbb{E}\left[\left[\int_0^T u_1^{2\epsilon-2}\, dt \right]^{q/2} \right] \leq c \left[\mathbb{E}\left[\int_0^T \frac{1}{u_1^r}\, dt \right] \right]^{q/2}$$

with $r = 2 - 2\epsilon$. The integrand on the right-hand side is bounded for $\epsilon > 0$ by part 1 with $r = 2 - 2\epsilon < 2$. Under the assumption on G, we have that $v_1[\nabla G(u \odot v)]_1$ is polynomially growing and has bounded qth moments by Theorem 2.1. Hence, qth moments of the integral term $c(\epsilon) \int \cdots dt$ are finite.

We conclude that q-moments of $\int_0^T u_1^{\epsilon-2}\, dt$ are finite for any $\epsilon > 0$. □

We will also require the following time-regularity bounds.

Proposition 2.4. *Let $(u, v)_t$ denote the solution of Eq. (7) with initial distribution ρ_0. Suppose that the assumptions of Theorem 2.3 hold. For $1 \leq q < 2$, $T > 0$, and $\epsilon \in (0, 1/2)$, there exists $c > 0$ such that, for $0 \leq s, t \leq T$,*

1. $\mathbb{E}\left[\|(u, v)_t - (u, v)_s\|^q\right] < c\,|t - s|^{q/2}$.
2. $\mathbb{E}\left[\frac{|u_{i,t} - u_{i,s}|}{u_{i,s}\, u_{i,t}}\right] \leq c\,|t - s|^{1/2-\epsilon}, \qquad i = 1, \ldots, d.$

Proof. 1. We treat the u_1 component only (v is handled similarly). From Eq. (7),

$$u_{1,t} = u_{1,s} + \frac{1}{\beta}\int_s^t \frac{1}{u_{1,r}}\, dr + \cdots + \sqrt{\frac{2}{\beta}} \int_s^t dW_r^1,$$

where we omit several integrals that are straightforward to estimate under the moment bound in Theorem 2.1.

By Holder's inequality with conjugate exponents $(2-\epsilon, (2-\epsilon)/(1-\epsilon))$,

$$\int_s^t \frac{1}{u_{1,r}}\,dr \le |t-s|^{(1-\epsilon)/(2-\epsilon)} \left[\int_s^t \frac{1}{|u_{1,r}|^{2-\epsilon}}\,dr\right]^{1/(2-\epsilon)}.$$

Now evaluate the qth moments and use moment bounds on Gaussian increments to find, for a constant $c_q > 0$, that

$$(\mathbb{E}|u_{1,t}-u_{1,s}|^q)^{1/q} \le |t-s|^{(1-\epsilon)/(2-\epsilon)} \frac{1}{\beta}\left(\mathbb{E}\left[\int_s^t \frac{1}{|u_{1,r}|^{2-\epsilon}}\,dr\right]^{q/(2-\epsilon)}\right)^{1/q} + \cdots + \sqrt{\frac{2}{\beta}}\, c_q\, |t-s|^{1/2}.$$

By Theorem 2.3, the $\mathbb{E}\left[\left(\int_s^t \frac{1}{|u_r|^{2-\epsilon}}\,dr\right)^q\right]$ is finite for $\epsilon > 0$ and $1 \le q < 2$. Hence, we have shown, for some constant c depending on ϵ and q, that $(\mathbb{E}\|u_t - u_s\|^q)^{1/q} \le c\,|t-s|^{1/2-\epsilon}$.

2. Let $\mathcal{R} = \{\min\{u_i(t), u_i(s)\} < \Delta\}$. Then, $\mathbb{P}(\mathcal{R}) \le \mathbb{P}(u_i(t) < \Delta) + \mathbb{P}(u_i(s) < \Delta)$. By Chebyshev's inequality, $\mathbb{P}(u_i(t) < \Delta) \le \Delta^{2-\epsilon}\,\mathbb{E}[u_i(t)^{\epsilon-2}]$ for $\epsilon > 0$. Hence, using Theorem 2.3, for a constant c, $\mathbb{P}(\mathcal{R}) \le c\,\Delta^{2-\epsilon}$. Now, split the expectation with respect to $\mathcal{R}$ as

$$\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)}\right] = \mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)} 1_{\mathcal{R}}\right] + \mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)}(1-1_{\mathcal{R}})\right].$$

We estimate each term separately.

For conjugate exponents (p, p'), Holder's inequality applied to the first term provides

$$\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)} 1_{\mathcal{R}}\right] \le \left(\mathbb{E}\left[\left|\frac{1}{u_i(t)} - \frac{1}{u_i(s)}\right|^p\right]\right)^{1/p} \mathbb{P}(\mathcal{R})^{1/p'}.$$

Choose $p' = 2+\epsilon$. Then its conjugate exponent p is less than 2, so that $\mathbb{E}\left[u_i^{-p}\right] < \infty$ as a result of Corollary 2.3. As $\mathbb{P}(\mathcal{R}) \le c\,\Delta^{2-\epsilon}$ for some c, the first term is bounded by $c\,\Delta^{(2-\epsilon)/(2+\epsilon)} = c\,\Delta^{1-2\epsilon/(2+\epsilon)}$.

For the second term, write

$$\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)}(1-1_{\mathcal{R}}) = \frac{|u_i(t)-u_i(s)|^{1-\epsilon}}{|u_i(s)\,u_i(t)|^{1-\epsilon}} \frac{|u_i(t)-u_i(s)|^{\epsilon}}{|u_i(s)\,u_i(t)|^{\epsilon}}(1-1_{\mathcal{R}}).$$

For (p, p') conjugate (different to the ones above),

$$\begin{aligned}
&\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)}(1-1_{\mathcal{R}})\right] \\
&\quad\le \left(\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|^{(1-\epsilon)p}}{|u_i(s)\,u_i(t)|^{(1-\epsilon)p}}\right]\right)^{1/p} \left(\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|^{\epsilon\,p'}}{|u_i(s)\,u_i(t)|^{\epsilon\,p'}}(1-1_{\mathcal{R}})\right]\right)^{1/p'} \\
&\quad\le \left(\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|^{(1-\epsilon)\,p}}{|u_i(s)\,u_i(t)|^{(1-\epsilon)\,p}}\right]\right)^{1/p} \Delta^{-2\epsilon} \left(\mathbb{E}\left[|u_i(t)-u_i(s)|^{\epsilon\,p'}\right]\right)^{1/p'}.
\end{aligned}$$

In the last line, we exploit $u_i(s), u_i(t) > \Delta$ outside $\mathcal{R}$. The third factor is $O(|t-s|^{\epsilon/2})$ by part 1 if $1 < \epsilon p' < 2$. Choose $p' = 3/2\epsilon$. Then $p = 1/(1-1/p') = 1/(1-(2/3)\epsilon)$.

Focus now on the first factor: for (r, r') conjugate,

$$\left(\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|^{(1-\epsilon)\,p}}{|u_i(s)\,u_i(t)|^{(1-\epsilon)\,p}}\right]\right)$$
$$\leq \left(\mathbb{E}\left[|u_i(t)-u_i(s)|^{(1-\epsilon)\,p\,r'}\right]\right)^{1/r'} \left(\mathbb{E}\left[\frac{1}{|u_i(s)\,u_i(t)|^{(1-\epsilon)\,p\,r}}\right]\right)^{1/r}.$$

This is a product of two terms. The first term is $O(|t-s|^{(1-\epsilon)\,p/2})$ by part 1 with $q = (1-\epsilon)\,p\,r'$. The second term is finite if $(1-\epsilon)\,p\,r < 1$ (by Theorem 2.3 and $2\,a\,b \leq a^2 + b^2$). We can achieve $(1-\epsilon)\,p\,r < 1$ for any $\epsilon > 0$, by choosing $1 < r < 1/(1-\epsilon)p = (1-(2/3)\epsilon)/(1-\epsilon)$. The resulting estimate is

$$\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)}(1-1_{\mathcal{R}})\right] \leq c\,|t-s|^{(1-\epsilon)/2}\,\Delta^{-\epsilon}\,|t-s|^{\epsilon/2} = c\,|t-s|^{1/2}\,\Delta^{-\epsilon}.$$

We have shown that

$$\mathbb{E}\left[\frac{|u_i(t)-u_i(s)|}{u_i(s)\,u_i(t)}\right] \leq c\,\Delta^{1-\epsilon} + c\,|t-s|^{1/2}\,\Delta^{-\epsilon}.$$

The right-hand side behaves like $\Delta^{1-2\epsilon/(2+\epsilon)} + |t-s|^{1/2}\Delta^{-2\epsilon}$. Taking $\Delta = |t-s|^{\alpha}$ and equating powers of $|t-s|$ leads to $\alpha(1-2\epsilon/(2+\epsilon)) = \frac{1}{2} - 2\alpha\epsilon$. This implies an upper bound of $c|t-s|^{\frac{1}{2}-\frac{\epsilon}{2}}$ provided that $\epsilon \leq \frac{1}{2}$. □

2.2 Links to the OU process and CIR process

We highlight a link to some well-known SDEs, namely, the *Cox–Ingersoll–Ross (CIR) process* and *Orstein–Uhlenbeck process* [12]. Girsanov's theorem allows us to transfer properties of these SDEs to our dynamics and therefore provides an alternative way of establishing well-posedness and strong Feller properties. This link will also serve as inspiration for the development of numerical schemes in Section 3.

When the data term $G \equiv 0$, our Langevin dynamics decouples into some well-known SDEs for modeling interest rates. In particular, the *Cox–Ingersoll–Ross (CIR) process* is

$$dy_t = \kappa\,(\mu - y_t)\,dt + \theta\sqrt{y_t}\,dW_t. \tag{12}$$

This is known to have unique solutions with $\mathbb{P}(y_t > 0, \forall t \geq 0) = 1$ if and only if the Feller test holds, which occurs for $2\kappa\mu \geq \theta^2$. In this case, by defining $\hat{u}_t = \sqrt{y_t}$ (this is the so-called Lamperti transformation), Itô's formula implies that

$$d\hat{u}_t = \frac{(4\kappa\mu - \theta^2)/8}{\hat{u}_t}\,dt - \frac{\kappa}{2}\,\hat{u}_t\,dt + \frac{\theta}{2}\,dW_t. \tag{13}$$

This is precisely our u process when $G \equiv 0$, with $\theta = 2\sqrt{2/\beta}$, $\kappa = 2\lambda$ and $\mu = 2/(\beta\lambda)$. Note that this corresponds to the edge case of the Feller test as $2\kappa\mu = \theta^2$ [12, 15].

The *Ornstein—Uhlenbeck process* is given by $d\hat{v}_t = -\lambda\,\hat{v}_t\,dt + \sqrt{\frac{2}{\beta}}\,dW_t^2$. This is precisely our v process with $G \equiv 0$.

Given the connections mentioned above, the general idea for establishing well-posedness of our SDE is to interpret (7) as 'perturbed' CIR and OU processes. Girsanov transformations will allow us to absorb the data term into the Brownian motion and hence inherit the required properties from CIR and OU processes. The results shown here can also be derived using the (simpler) arguments involving the cylindrical transformation Eq. (8); however, as a bonus, the arguments here additionally show that $u(t)$ is positive almost surely.

In the following, let $\hat{z}_t \coloneqq (\hat{u}, \hat{v})_t$ satisfy

$$d\hat{z}_t = \left[-\lambda z_t + \begin{pmatrix} 1/(\beta\hat{u}_t) \\ 0 \end{pmatrix}\right] dt + \sqrt{\frac{2}{\beta}}\,dW_t, \tag{14}$$

with given initial data $\hat{z}_0$. The first component $\hat{u}_t$ of $\hat{z}_t$ is a CIR process. The CIR process is positive almost surely as, following [9], each component has the form $d\hat{u}_t = ((4\kappa\lambda - \theta^2)/(8\hat{u}_t))\,dt - \frac{1}{2}\kappa\hat{u}_t\,dt + (\theta/2)\,d\hat{W}_t$ for $\theta^2 = 8/\beta$ and $2\kappa\lambda = \theta^2$, which satisfies the Feller positivity condition $2\kappa\lambda \geq \theta^2$.

In order to apply Girsanov transformations, we first establish the Novikov condition.

Lemma 2.5 (Novikov condition)**.** *Let Assumption 1 hold and* $b((u,v)) = -\binom{v\odot\nabla G(u\odot v)}{u\odot\nabla G(u\odot v)}$. *For all* $T > 0$ *and* $\gamma > 0$, *we have*

$$\mathbb{E}\left[\exp\left(\gamma\int_0^T \|b(\hat{z}_s)\|^2\,ds\right)\right] < \infty.$$

There exists a weak solution (u_t, v_t) *of Eq. (7) that is unique in law. Furthermore,* $u_t > 0$ *almost surely.*

Proof. By [23, Corollary 5.14], it is enough to find an $\epsilon > 0$ such that, for all $t \in [0,T]$, $\mathbb{E}\left[\exp\left(\int_t^{t+\epsilon}\gamma\|b(\hat{z}_s)\|^2\right)ds\right] < \infty$. Since $B = \|\nabla G\|_\infty < \infty$, we obtain

$$\mathbb{E}\left[\exp\left(\gamma\int_t^{t+\epsilon}\|b(\hat{z}_s)\|^2\,ds\right)\right] \leq \mathbb{E}\left[\exp\left(\gamma\,B^2\int_t^{t+\epsilon}\|\hat{z}_s\|^2\,ds\right)\right].$$

By Jensen's inequality,

$$\mathbb{E}\left[\exp\left(\int_t^{t+\epsilon}\gamma\,B^2\,\|\hat{z}_s\|^2\,ds\right)\right] \leq \frac{1}{\epsilon}\int_t^{t+\epsilon}\mathbb{E}\Big[\exp\big(\epsilon\,\gamma\,B^2\,\|\hat{z}_s\|^2\big)\Big]\,ds. \tag{15}$$

We now apply Itô's lemma to bound $\mathbb{E}\big[\exp(\epsilon\,\gamma\,B^2\|\hat{z}_s\|^2)\big]$. Given $\alpha > 0$ and $\hat{y}_t \coloneqq \exp\left(\frac{\alpha}{2}\|(\hat{u}_t,\hat{v}_t)\|^2\right)$, we get

$$d\hat{y}_t = \left(-\lambda\alpha + \frac{\alpha^2}{\beta}\right)\|\hat{z}_t\|^2 y_t\,dt + \left(\frac{3\alpha d}{\beta}\right)\hat{y}_t\,dt + \sqrt{\frac{2}{\beta}}\alpha\hat{y}_t\big(\hat{u}_t^\top dW_t^1 + v_t^\top dW_t^2\big).$$

Taking expectation,

$$\frac{d}{dt}\mathbb{E}\big[\hat{y}_t\big] \le \mathbb{E}\left[\left(-\lambda\alpha + \frac{\alpha^2}{\beta}\right)\|\hat{z}_t\|^2\hat{y}_t + \left(\frac{3\alpha d}{\beta}\right)\hat{y}_t\right].$$

Provided that $\alpha < \beta\lambda$, we have $\frac{d}{dt}\mathbb{E}\big[\hat{y}_t\big] \le \left(\frac{3\alpha d}{\beta}\right)\mathbb{E}\big[\hat{y}_t\big]$ and therefore $\mathbb{E}\big[y_t\big] \le \hat{y}_0 \exp\left(\left(\frac{3\alpha d}{\beta}\right)t\right)$, where $\hat{y}_0 = \exp(\alpha\|z_0\|^2/2) < \infty$. Letting $\alpha = 2\epsilon\gamma B^2$, the condition on α is satisfied if $\epsilon < \beta\lambda/(2\gamma B^2)$. Substituting for α, $\mathbb{E}\big[\hat{y}_t\big] \le y_0 \exp\left(2\epsilon\gamma B^2\left(\frac{3d}{\beta}\right)t\right)$. Returning to (15), for all $t \le T - \epsilon$, we see

$$\mathbb{E}\left[\exp\left(\int_t^{t+\epsilon} \gamma B^2\|\hat{z}_s\|^2\right)ds\right] \le \hat{y}_0 \exp\left(2\epsilon\gamma B^2\left(\frac{3d}{\beta}\right)T\right).$$

Thus, Novikov's condition is satisfied.

We can therefore apply Girsanov's theorem. Then, a solution of Eq. (7) exists and the solution is unique in law [23]. Further, for all $T > 0$, with the martingale defined by $M_T = \exp(\int_0^T b(\hat{z}_t)\cdot dW_t - \frac{1}{2}\int_0^T b(\hat{z}_t)^2 dt)$, we see $\mathbb{P}(\exists t \in [0,T],\ u_t = 0) = \mathbb{E}\Big[1_{\{\exists t\in[0,T],\ \hat{u}_t=0\}}M_T\Big]$. Since the CIR process $\hat{u}_t$ is strictly positive almost surely, $\mathbb{P}(\exists t \in [0,T],\ u_t = 0) = 0$. By the Borel–Cantelli lemma, there are only finitely many T for which the event $\{\exists t \in [0,T],\ u_t = 0\}$ occurs so $\mathbb{P}(u_t > 0, t > 0) = 1$. □

Proposition 2.6. *Let Assumption 1 hold. Then, $(u,v)_t$ is strong Feller.*

Proof. If $b(x_s) \equiv 0$, then the processes $(u,v)_t$ decouple component-wise into CIR and OU processes. Sufficient conditions are provided by [29, Theorem 2.1] to guarantee that Girsanov transformed processes are strong Feller. We check the conditions of their theorem: (i) is satisfied since each of these components is strong Feller with known densities. Condition (ii) is due to the Novikov condition proven above. Condition (iii) holds due to [29, Remark 2.3] along with Lemma 2.5. □

2.3 Link to Riemannian Langevin diffusion

When sampling distributions on non-Euclidean spaces, Riemannian Langevin dynamics [14] are typically more efficient than standard Langevin dynamics since they scale the gradient to account for the geometry of the manifold. In general, if sampling from $\pi(\mathbf{z}) \propto \exp(-H(\mathbf{z}))$ on some manifold $\mathcal{M}$ with metric $\mathfrak{g}$, the Riemannian Langevin dynamics are [40]

$$d\mathbf{z}_t = -\mathfrak{g}(\mathbf{z}_t)^{-1}\nabla H(\mathbf{z}_t)\,dt + \Gamma(\mathbf{z}_t)\,dt + \sqrt{2}\mathfrak{g}(\mathbf{z}_t)^{-1/2}dW_t, \tag{16}$$

where $\Gamma_j(\mathbf{z}) = \sum_k \partial_k[(\mathfrak{g}(\mathbf{z})^{-1})_{jk}]$. The last two terms correspond to Brownian motion on the manifold $(\mathcal{M},\mathfrak{g})$ with the correction drift term taking into account the curvature of the manifold. The special setting where the metric is the Hessian of some convex Legendre-type

function is called 'Mirror-Langevin'. For example, in the case of sampling probability densities restricted to the positive real plane $\mathbb{R}^d_+$ or the simplex, a natural choice is $\mathfrak{g}(\mathbf{z}) = \nabla^2 \mathcal{H}(z)$ where $\mathcal{H}$ is the Boltzmann–Shannon entropy $\mathcal{H}(\mathbf{z}) = \sum_i z_i \log(z_i) - z_i$ [41].

Recall from (4) for $\beta = 1$ that $\rho(x)$ is the marginal distribution to the following distribution on $\mathbf{z} = (\eta, x) \in \mathcal{X} = \mathbb{R}^d_+ \times \mathbb{R}^d$: $\tilde{\rho}(\eta, x) \propto \exp(-H(\eta, x))$ where

$$H(\eta, x) \coloneqq G(x) + \lambda \sum_{i=1}^{d} \frac{x_i^2}{8\eta_i} + 2\lambda\eta_i + \frac{1}{2}\log(\eta_i).$$

In the following proposition, we see that our Langevin dynamics (7) on (u, v) corresponds to Riemannian Langevin dynamics on $\mathbf{z}$ with metric $\mathfrak{g}(\mathbf{z}) \coloneqq M(\mathbf{z})^{-1}$, with $M(\mathbf{z})$ as given in the proposition below. The proof is a direct consequence of Ito's formula and given in Appendix B.

Proposition 2.7. *Let $(u, v)_t$ follow the SDE (7) with $\beta = 1$. Then, letting $\mathbf{z}_t = (\eta, x)_t$ with $\eta = \frac{1}{4}u^2$ and $x = u \odot v$, we have that $\mathbf{z}_t$ follows (16) with inverse metric*

$$\mathfrak{g}(\mathbf{z})^{-1} = M(\mathbf{z}) \coloneqq \begin{pmatrix} \operatorname{diag}(\eta) & \operatorname{diag}(x)/2 \\ \operatorname{diag}(x)/2 & \operatorname{diag}(x^2/(4\eta)) + 4\operatorname{diag}(\eta) \end{pmatrix}.$$

The distribution $\rho(t, \cdot)$ of $\mathbf{z}(t)$ solves the Fokker–Planck equation

$$\frac{\partial}{\partial t}\rho = \operatorname{div}\big(M(\mathbf{z})\nabla\big(H(\mathbf{z}) + \log(\rho)\big)\rho\big). \tag{17}$$

Remark. By direct computations, one can check that for $\phi(\mathbf{z}) = \frac{1}{8}\sum_{i=1}^d \frac{|x_i|^2}{\eta_i} + 2\eta_i + 4\sum_{i=1}^d \mathcal{H}(\eta_i)$, where $\mathcal{H}(\eta) = (\eta\log(\eta) - \eta)$, the corresponding metric is

$$\mathfrak{g}(\mathbf{z}) = M(\mathbf{z})^{-1} = \begin{pmatrix} \frac{1}{2}\mathrm{Id}_d & 0_{d\times d} \\ 0_{d\times d} & \mathrm{Id}_d \end{pmatrix} \nabla^2\phi(z) \begin{pmatrix} \frac{1}{2}\mathrm{Id}_d & 0_{d\times d} \\ 0_{d\times d} & \mathrm{Id}_d \end{pmatrix}.$$

2.4 Ergodicity

The $\mathcal{L}$ denotes the infinitesimal generator of Eq. (7) and let $\rho(t, \cdot)$ denote the distribution of $(u, v)_t$. Then $(e^{\mathcal{L}t}\phi)(X) = \mathbb{E}[\phi((u, v)_t) \mid (u, v)_0 = X]$ for measurable functions $\phi\colon \mathcal{X} \to \mathbb{R}$. We seek to show that $e^{\mathcal{L}t}\phi$ converges to $\pi(\phi) = \int \phi(x)\, d\pi(x)$ or, alternatively, ρ_t converges to π.

If the data term G is uniformly bounded, then one could invoke standard theories (Bakry-Emery criterion) to deduce ergodicity; see Appendix A. This is, however, a strong assumption. In the following, we prove a different ergodicity result that assumes $x^\top \nabla G(x)$ is bounded below and which holds for many unbounded examples, including $G(x) = \|Ax - y\|^2$. The proof depends on the theory of Lyapunov–Foster drift conditions [31, 26].

For the Lyapunov function $V((u, v)) = 1 + \|(u, v)\|^2$, let L^∞_V denote the Banach space of measurable functions $\phi\colon \mathcal{X} \to \mathbb{R}$ such that the norm $\|\phi\|^\infty_V \coloneqq \sup_x |\phi(x)/V(x)| < \infty$.

Theorem 2.8. *Under Assumption 1, the solution $(u,v)_t$ of Eq. (7) converges to the invariant π geometrically. In particular, for $V((u,v)) = 1 + \|(u,v)\|^2$ and some $C, \rho > 0$,*

$$\|e^{\mathcal{L}t}\phi - \pi(\phi)\|_V^\infty \le C\exp(-\rho t)\|\phi - \pi(\phi)\|_V^\infty, \qquad t > 0,\ \phi \in L_V^\infty.$$

Proof. In discrete time, this follows [26, Theorem 2.25] if a Lyapunov condition holds for V and a minorization condition holds for the transition density. The extension to continuous time [26, Equation 2.77] is achieved by writing $t = k\Delta t + s$ to create a Markov chain on discrete time k and correction factor depending on $s \in (0, \Delta t)$.

The process $(u,v)_t$ is strong Feller and irreducible by Theorem 2.1. Hence, the minorization conditions hold for any compact set [30, Lemma 2.3].

Under Assumption 1, $x^\top \nabla G(x) \ge -K$ for some constant $K \in \mathbb{R}$. We have, from Eq. (7),

$$\mathcal{L}V(u,v) = -2\lambda\|(u,v)\|^2 - 4(u\odot v)^\top \nabla G(u\odot v) + 2d + 4\,d/\beta \le -2\lambda\|(u,v)\|^2 + 4K + 2d + 4\,d/\beta.$$

Let c be a constant, such that $0 < c < 2\lambda$ and $b = 4K + 2d + 4d/\beta + c > 0$. We have $\mathcal{L}V(u,v) \le -cV(u,v) + b$ as required. □

3 Discretization

We consider the following time-stepping approximation: for time step Δt, we seek an approximation $(u,v)^k$ to $(u,v)_{k\Delta t}$ via

$$\begin{aligned} u^{k+1} - u^k &= -\lambda u^{k+1}\Delta t - v^k \odot \nabla G(u^k \odot v^k)\Delta t + \frac{1}{\beta u^{k+1}}\Delta t + \sqrt{\frac{2}{\beta}}\Delta W_k^1, \\ v^{k+1} - v^k &= -\lambda v^{k+1}\Delta t - u^k \odot \nabla G(u^k \odot v^k)\Delta t + \sqrt{\frac{2}{\beta}}\Delta W_k^2, \end{aligned} \tag{18}$$

where $\Delta W_k^i = W_{t_{k+1}}^i - W_{t_k}^i$. The update rule is implicit in the regularization term (involving λ) and the singularity (involving $1/\beta u$), and explicit in the data term (involving G) and the noise. For convenience, define the explicit part of the drift term as f, so that

$$f(u,v) := \begin{pmatrix} f_1(u,v) \\ f_2(u,v) \end{pmatrix} := \begin{pmatrix} v \odot \nabla G(u\odot v) \\ u \odot \nabla G(u \odot v) \end{pmatrix}. \tag{19}$$

For our analysis, we will frequently use the local Lipschitz property of f.

Lemma 3.1. *Under Assumption 2, the function f defined in (19) is locally Lipschitz on the ball of radius R with Lipschitz constant $C_R = 2LR^2 + B$: $\|f(u,v) - f(u',v')\| \le C_R\|(u-u', v-v')\|$ for all $(u,v),(u',v') \in \mathcal{X}$ such that $\|(u,v)\|, \|(u',v')\| \le R$.*

Proof. Under Assumption 2, ∇G is L-Lipschitz, $B := \sup_x\|\nabla G(x)\| < \infty$, Hence, for $v, \tilde{v}$ such that $\|v\| \le R, \|\tilde{v}\| \le R$, $\|v \odot \nabla G(u\odot v) - \tilde{v} \odot \nabla G(\tilde{u}\odot\tilde{v})\|$ is upper bounded by

$$B\|\tilde{v} - v\| + \|v\|\|\nabla G(u\odot v) - \nabla G(\tilde{u}\odot\tilde{v})\| \le \left(B + 2LR^2\right)\Big(\|v-\tilde{v}\| \vee \|u - \tilde{u}\|\Big). \qquad \square$$

Remark (Relation to CIR solvers)**.** If $G = 0$, then the SDE (7) on u_t resembles a CIR process. Our implicit-explicit scheme is inspired by the scheme introduced in [9], and our scheme also ensures that $u^k > 0$ for all k. See also [18].

Remark (Relation to proximal gradient Langevin)**.** By defining $\Psi((u,v)) \coloneqq \frac{\lambda}{2}\|(u,v)\|^2 - \beta \log(u)$, $(u,v)^{k+1} = \Pi_{\Delta t}(z^{k+1})$, one can write (18) as

$$z^{k+1} = \Pi_{\Delta t}(z^k) - \Delta t\, f(\Pi_{\Delta t}(z^k)) + \sqrt{2/\beta}\,\Delta W_k. \tag{20}$$

where $\Pi_{\Delta t}(z) = \mathrm{Prox}_{\Delta t \Psi}(z) = \mathrm{argmin}_x(\Delta t\, \Psi(x) + \frac{1}{2}\|x - z\|)$ is a non-expansive operator. This corresponds to an application of the proximal gradient Langevin algorithm introduced in [10]. Note however that their analysis does not extend to our setting since f is non-convex and lack global Lipschitz properties.

3.1 Convergence to stationary distribution

We show in Theorem 3.2 that the numerical approximation $(u,v)^k$ indeed has an invariant distribution.

We first introduce some notation. For $n > 0$, $x \in \mathcal{X}$, and Borel sets $A \subset \mathcal{X}$, let $P^n(\mu_0, A) \coloneqq \mathbb{P}(X^n \in A \mid X^0 \sim \mu_0)$ for initial distribution μ_0. For $\phi\colon \mathcal{X} \to \mathbb{R}$,

$$(P^n\phi)(x) \coloneqq \int_{\mathcal{X}} \phi(y) P^n(\delta_x, dy) = \mathbb{E}[\phi(X^n) \mid X^0 = x].$$

Given $\zeta > 0$, define $\|\phi\|_\zeta \coloneqq \sup_{x\in\mathcal{X}} \frac{|\phi(x)|}{1+\zeta\, V(x)}$ and

$$d_\zeta(x,y) \coloneqq \begin{cases} 0, & x = y, \\ 2 + \zeta\, V(x) + \zeta\, V(y), & x \neq y. \end{cases}$$

The argument is based on showing P^n is a contraction with respect to

$$d_\zeta(\mu_1,\mu_2) \coloneqq \sup_{\|\phi\|_\zeta = 1} \int_{\mathcal{X}} \phi(x)\, d(\mu_1 - \mu_2)(x), \quad \text{for measures } \mu_1, \mu_2$$

with $\|\phi\|_\zeta \coloneqq \sup_{x\neq y}|\phi(x) - \phi(y)|/d_\zeta(x,y)$. Following [17], d_ζ is equivalent to the weighted TV norm $\|\mu\|_\zeta \coloneqq \sup_{\|\phi\|_\zeta \le 1}\left|\int_{\mathcal{X}} \phi(x)\, d\mu(x)\right|$.

Theorem 3.2. *There exists $\bar\alpha \in (0,1)$, $t_0 > 0$ and $\zeta > 0$ such that, for any $\phi\colon \mathcal{X} \to \mathbb{R}$ with $\|\phi\|_\zeta \le 1$,*

$$\left|P^{\lceil t_0/\Delta t\rceil}\phi(x) - P^{\lceil t_0/\Delta t\rceil}\phi(y)\right| \le \bar\alpha\, d_\zeta(x,y), \tag{21}$$

holds for all Δt sufficiently small. As a consequence, there exists an invariant measure $\pi_{\Delta t}$ such that, for any initial distribution μ_0 and any $T > 0$,

$$\left\|P^{\lceil T/\Delta t\rceil}(\mu_0,\cdot) - \pi_{\Delta t}\right\|_\zeta \le \bar\rho^T \|\mu_0 - \pi_{\Delta t}\|_\zeta, \tag{22}$$

where $\rho = \bar\alpha^{1/t_0}$ is independent of Δt.

This is a consequence of minorisation and Lyapunov drift conditions, given in the following lemmas.

Lemma 3.3 (drift condition)**.** *Let* $V((u,v)) = 1+\|(u,v)\|^2$*. Assume that* $\Delta t < \min\{\lambda/B^2, 1/4\lambda\}$*, where* $B = \|\nabla G\|_\infty$*. Then, the numerical approximation* (u^k, v^k) *defined by Eq. (18) satisfies*

$$\mathbb{E}\big[V((u,v)^{k+1}) \mid (u,v)^k\big] \le \left(1 - \frac{1}{4}\lambda\,\Delta t\right) V((u,v)^k) + C\,\Delta t,$$

where $C \coloneqq \frac{1}{2}\lambda + 4(8 + K\beta + d)/\beta$*.*

Let $V_4((u,v)) = 1 + \|(u,v)\|^4$*. There exists* $C, c, \nu > 0$ *such that for all* $\Delta t < c\lambda$*,*

$$\mathbb{E}[V_4((u,v)^{k+1}) \mid (u,v)^k] \le (1 - \nu\,\Delta t)\, V_4((u,v)^k) + C\,\Delta t.$$

Proof. We prove the drift condition for the quadratic functional V; the calculation for the quartic V_4 is provided in the supplementary materials. We have

$$(1+\lambda\Delta t)u^{k+1} - \frac{\Delta t}{\beta u^{k+1}} = u^k - v^k \odot \nabla G(u^k \odot v^k)\,\Delta t + \sqrt{\frac{2}{\beta}}\,\Delta W_k^1, \tag{23}$$

$$(1+\lambda\Delta t)v^{k+1} = v^k - u^k \odot \nabla G(u^k \odot v^k)\,\Delta t + \sqrt{\frac{2}{\beta}}\,\Delta W_k^2. \tag{24}$$

We square the norm and expectation of the above, use the independence between the increments ΔW_k^i and the information at time t_k, and rely on the trivial fact that $\|(u^{k+1})^{-1} \odot (u^{k+1})^{-1}\| = d$. As a result, conditional on step k, we obtain

$$(1+\lambda\,\Delta t)^2\mathbb{E}\big[\|u^{k+1}\|^2 \mid (u,v)^k\big] - 2\Delta t(1+\lambda\,\Delta t)\beta^{-1}d \le \frac{2}{\beta}\,\mathbb{E}\|\Delta W_k^1\|^2 + \|u^k - \Delta t\, v^k \odot \nabla G(u^k \odot v^k)\|^2,$$

$$(1+\lambda\,\Delta t)^2\mathbb{E}\big[\|v^{k+1}\|^2 \mid (u,v)^k\big] = \frac{2}{\beta}\,\mathbb{E}\|\Delta W_k^2\|^2 + \|v^k - \Delta t\, u^k \odot \nabla G(u^k \odot v^k)\|^2.$$

Using the assumption $x^\top \nabla G(x) \ge -K$ and $B = \|\nabla G\|_\infty$, we obtain

$$\begin{aligned}\|u^k - \Delta t\, v^k \odot \nabla G(u^k \odot v^k)\|^2 &\le \|u^k\|^2 + \Delta t^2 B^2 \|v^k\|^2 - 2\,\Delta t\,(u^k \odot v^k)^\top \nabla G(u^k \odot v^k) \\ &\le \|u^k\|^2 + \Delta t^2 B^2 \|v^k\|^2 + 2\,K\,\Delta t.\end{aligned}$$

A similar statement holds for $\big\|v^k - \Delta t\, u^k \odot \nabla G(u^k \odot v^k)\big\|^2$. Combining this with $\mathbb{E}\big\|\Delta W_k^i\big\|^2 = d\,\Delta t$, we obtain

$$(1+\lambda\,\Delta t)^2\mathbb{E}\Big[\|(u,v)^{k+1}\|^2 \;\Big|\; (u,v)^k\Big] \le (1+\Delta t^2 B^2)\|(u,v)^k\|^2 + 4\frac{d\,(1+\Delta t\,\lambda) + K\,\beta + d}{\beta}\,\Delta t.$$

Note that $\frac{(1+\Delta t^2 B^2)}{(1+\lambda\Delta t)^2} \le \frac{(1+\Delta t^2 B^2)}{1+2\lambda\Delta t} < 1 - \frac{1}{2}\lambda\,\Delta t$ provided that $0 < \Delta t < \min\{\lambda/B^2, 1/4\lambda\}$. Rewrite this in terms of $V((u,v)) = 1 + \|(u,v)\|^2$:

$$\mathbb{E}\Big[V((u,v)^{k+1}) \;\Big|\; (u,v)^k\Big] \le \left(1 - \frac{1}{2}\lambda\Delta t\right)V((u,v)^k) + \left(\frac{1}{2}\lambda + 4\frac{4\,(1+\Delta t\lambda) + K\,\beta + d}{\beta(1+\lambda\Delta t)^2}\right)\Delta t.$$

Using the property $\Delta t\,\lambda \le 1$, we can replace the last term by $C\Delta t$ with $C \coloneqq \frac{1}{2}\lambda + 4(8 + K\beta + d)/\beta$. □

By first establishing Theorems 3.4 and 3.5, we prove the coupling condition in Theorem 3.6 for the numerical method.

Lemma 3.4 (escape probability)**.** *Fix $0 < r < R$. Let $k_R = \inf\{k : \|(u,v)^k\| > R\}$ and assume that $\|(u,v)^0\| \le r$. For some constants $c, C > 0$, we have, for $n\Delta t = T$,*

$$\mathbb{P}(k_R \le n) \le \exp\left(-\frac{R^2}{4T}\right)\exp\left(\frac{r^2}{2T} + \frac{C}{2}\right).$$

The constant C depend only on β and the coercivity constant of G.

Proof. The start of this proof is similar to the proof of Lemma 3.3 but we retain the Brownian increments: taking the square-norm of (23) and using $\left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^2 \le \left\|u^k\right\|^2 + \Delta t^2 B^2 \left\|v^k\right\|^2 + 2K\Delta t$, we obtain

$$(1+\lambda\Delta t)^2\|u^{k+1}\|^2 \le \|u^k\|^2 + \Delta t^2 B^2\|v^k\|^2 + 2K\Delta t + Z_k^1 + 2d\Delta t/\beta + 2d\Delta t(1+\lambda\Delta t)/\beta,$$

where $Z_k^1 \coloneqq \frac{2}{\beta}(\|\Delta W_k^1\|^2 - d\Delta t) + \sqrt{\frac{8}{\beta}}\left\langle \Delta W_k^1, u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\rangle$.

A similar computation for v^k with $Z_k^2 \coloneqq \frac{2}{\beta}(\|\Delta W_k^2\|^2 - d\Delta t) + \sqrt{\frac{8}{\beta}}\left\langle \Delta W_k^2, v^k - u^k \odot \nabla G(u^k \odot v^k)\Delta t\right\rangle$ yields

$$(1+\lambda\Delta t)^2\|v^{k+1}\|^2 \le \|v^k\|^2 + \Delta t^2 B^2\|u^k\|^2 + 2K\Delta t + Z_k^2 + 2d\Delta t/\beta.$$

We therefore have constants $c, C > 0$ such that for all $\Delta t \le c$,

$$(1+\lambda\Delta t)^2\|(u,v)^{k+1}\|^2 \le (1 + B^2\Delta t^2)\|(u,v)^k\|^2 + Z_k^1 + Z_k^2 + C\Delta t.$$

Let $\xi_k \coloneqq Z_k^1 + Z_k^2$ and let $\mathcal{F}_n$ be the σ-algebra generated by ΔW_k^j for $j = 1, 2$ and $k = 1, \dots, n-1$. Note that $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\xi_k^2 \mid \mathcal{F}_k] \le C(\Delta t^2 + (\|u^k\|^2 + \|v^k\|^2)\Delta t)$ for some $C > 0$. Let $M_n \coloneqq \sum_{k<n} \xi_k$. This is a martingale with quadratic variation $\langle M\rangle_n$ satisfying

$$\langle M\rangle_n = \sum_{k<n} \mathbb{E}\left[\xi_k^2 \mid \mathcal{F}_k\right] \le C\sum_{k<n}\left(\Delta t^2 + (\|u^k\|^2 + \|v^k\|^2)\,\Delta t\right).$$

Using this notation, we can rewrite the above inequality as

$$(1+\lambda\Delta t)^2\,\|(u,v)^{k+1}\|^2 \le \left(1 + B^2\Delta t^2\right)\|(u,v)^k\|^2 + C\Delta t + \xi_k - \frac{\alpha}{2}\mathbb{E}[\xi_k^2 \mid \mathcal{F}_k] + \frac{\alpha}{2}\mathbb{E}[\xi_k^2 \mid \mathcal{F}_k]. \quad (25)$$

Absorbing the variation term. Equation (25) can be written as

$$\|(u,v)^{k+1}\|^2 \le \|(u,v)^k\|^2 + C\Delta t + \xi_k - \frac{\alpha}{2}\,\mathbb{E}[\xi_k^2 \mid \mathcal{F}_k] + \frac{\alpha}{2}(\langle M\rangle_{k+1} - \langle M\rangle_k).$$

Summing this over the k's implies that

$$\|(u,v)^k\|^2 - \frac{\alpha}{2}\langle M\rangle_k \le \|(u,v)^0\|^2 + C\,k\,\Delta t + \sum_{j<k}(\xi_j - \frac{\alpha}{2}\,\mathbb{E}[\xi_j^2 \mid \mathcal{F}_j]).$$

It follows that, for $n\Delta t \le T$ and all $k \le n$,

$$(1-\alpha T)\sup_{k\le n}\|(u,v)^k\|^2 \le \|(u,v)^0\|^2 + C\,T + \sup_{k\le n}\Big(M_k - \frac{\alpha}{2}\,\langle M\rangle_k\Big).$$

We now apply [25, Lemma A.1]: Let $\alpha = 1/(2\,T)$. For $n\Delta t = T$, we have

$$\mathbb{P}\left(\frac{1}{2}\sup_{k\le n}\|(u,v)^{k+1}\|^2 \ge \|(u,v)^0\|^2 + CT + R\right) \le \mathbb{P}\left(\sup_{k\le n}\Big(M_k - \frac{\alpha}{2}\langle M\rangle_k\Big) \ge R\right) \le e^{-\alpha\,R} = e^{-R/(2\,T)}.$$

Estimate on escape probability. Combining our two estimates above, we have

$$\mathbb{P}\left(\frac{1}{2}\sup_{k\le n}\|(u,v)^k\|^2 \ge \|(u,v)^0\|^2 + CT + R\right) \le e^{-R/(2T)}. \tag{26}$$

So, for $n\Delta t = T$, as k_R is the first escape time from the ball of radius R,

$$\mathbb{P}(k_R \le n) = \mathbb{P}\left(\sup_{k\le n}\frac{1}{2}\|(u,v)^k\|^2 \ge \frac{R^2}{2}\right) \le \exp\left(-\frac{R^2}{4T} + \frac{\|(u,v)^0\|^2}{2T} + \frac{C}{2}\right). \qquad \square$$

The first coupling lemma considers the coupling event (i.e., $X^n = Y^n$) of two realisations X^n, Y^n of the numerical method for paths that do not leave the ball (i.e., $X^k, Y^k \in B(0,R)$ for $k = 0,\dots,n-1$).

Lemma 3.5 (coupling in the ball)**.** *Consider initial data $x, y \in \mathcal{X}$ with $\|x\|, \|y\| \le r$, some $r > 0$. Consider two solutions $(u,v)^n$ of the numerical method driven by different Brownian motions, denoted X^n, Y^n with initial condition $X^0 = x$, $Y^0 = y$. For $R > 0$ and $t_0 > 0$, let $E = \{X^1,\dots,X^{n-1}, Y^1,\dots,Y^{n-1} \in B(0,R)\}$ and $\tilde{E} = E \cap \{X^n \ne Y^n\}$ where $n\,\Delta t = t_0$.*

We may couple X^n, Y^n such that the following property holds: if $t_0 = 1/C_R$ and $\Delta t \in (0, t_0]$ and $r\sqrt{3\,C_R} \ge 1$, then $\mathbb{P}(\tilde{E}) \le 1 - \tilde{\delta}$ with $\tilde{\delta} = \dfrac{2}{\sqrt{2\pi}(1 + r\sqrt{3\,C_R})}\exp\left(-\dfrac{3\,r^2}{2}\,C_R\right)$.

Proof. The assumptions for [8, Theorem 4] hold: Assumption A1 holds due to the non-expansivity of $\Pi_{\Delta t}$ in Eq. (20); and Assumption A2(A) holds as the drift is Lipschitz on $B(0,R)$ with Lipschitz constant $L = C_R$ (Theorem 3.1). Hence, we may couple X^n, Y^n so that $\mathbb{P}(\tilde{E}) \le 1 - \delta$ with

$$\delta = 2\inf_{x,y\in B(0,r)}\Phi\big(-\|x-y\|/2\,\Xi_n^{1/2}(C_R)\big), \tag{27}$$

where Φ is the cumulative distribution function for the standard Gaussian, C_R is the Lipschitz constant of the drift on the set $B(0,R)$, and $\Xi_n(L) = \Delta t\sum_{k=1}^{n}(1+\Delta t\,L)^{-k}$.

For $\mu = 1/(1+\Delta t\,C_R)$, then $1-\mu = \Delta t\,C_R\,\mu$ and

$$\Xi_n(C_R) = \Delta t\,\mu\frac{1-\mu^n}{1-\mu} = \frac{1-\mu^n}{C_R} \ge \frac{1-\exp(-t_0\,C_R/(1+\Delta t\,C_R))}{C_R}, \qquad n\Delta t = t_0,$$

(as $1-1/x \le \log x$ and so $1/(1+x) \le \exp(-x/(1+x))$ for $x>0$). For $t_0 = 1/C_R$, $\Xi_n(C_R) \ge \frac{1}{C_R}(1-\exp(-1/2)) \ge \frac{1}{3C_R}$. It is well known ([16]) that $\Phi(-x) \ge \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\frac{1}{x+1/x}$ for $x>0$, and so

$$\delta = 2\,\Phi\left(-r\,\sqrt{3\,C_R}\right) \ge \frac{2}{\sqrt{2\pi}(1+r\sqrt{3\,C_R})}\exp\left(-\frac{3\,r^2}{2}\,C_R\right) =: \tilde{\delta},$$

for $r\sqrt{3C_R} \ge 1$. We conclude that $\mathbb{P}(\tilde{E}) \le 1-\tilde{\delta}$. □

Corollary 3.6 (coupling). *Under the assumptions of Theorem 3.5, for all $r>0$, there exists $t_0>0$, so that $\mathbb{P}(X^n = Y^n) \ge \alpha > 0$, uniformly for $x, y \in B(0,r)$ and $0<\Delta t \le t_0$ with $n\Delta t = t_0$.*

Proof. Let $E = \{X^1,\dots,X^{n-1},Y^1,\dots,Y^{n-1} \in B(0,R)\}$ and $\tilde{E} = E \cap \{X^n \ne Y^n\}$ where $n\,\Delta t = t_0$. Then, if C_R is sufficiently large, $t_0 = 1/C_R$ and $\Delta t \in (0,t_0]$, we have $\mathbb{P}(E) \ge 1-\delta$ and $\mathbb{P}(\tilde{E}) \le 1-\tilde{\delta}$ by Theorems 3.4 and 3.5, with

$$\delta = 2\exp\left(-\frac{C_R R^2}{4}\right)\exp\left(\frac{C_R\,r^2}{2}+\frac{C}{2}\right), \qquad \tilde{\delta} = \frac{2}{\sqrt{2\pi}(1+r\sqrt{3\,C_R})}\exp\left(-\frac{3\,r^2}{2}\,C_R\right).$$

Notice that $\tilde{E}^c$ is the union of $\{X^n = Y^n\}$ and E^c. Thus, $\mathbb{P}(X^n=Y^n) \ge \mathbb{P}(\tilde{E}^c) - \mathbb{P}(E^c) \ge \tilde{\delta}-\delta =: \alpha$. Then,

$$\alpha = \frac{2}{\sqrt{2\pi}(1+r\sqrt{3\,C_R})}\exp\left(-\frac{3\,r^2}{2}\,C_R\right) - 2\exp\left(-\frac{C_R R^2}{4}\right)\exp\left(\frac{C_R\,r^2}{2}+\frac{C}{2}\right).$$

The first term in the right-hand-side above behaves like $\exp(-3\,r^2\,C_R\,/2)$ and second like $\exp(-R^2 C_R/4)$. Thus the first term dominates in the limit $R\to\infty$ and we may choose R large enough so $\alpha>0$. □

We now show the drift and coupling conditions imply geometric convergence to the invariant measure.

Proof of Theorem 3.2. From Lemma 3.3, there exists $\nu>0$ and $C>0$ such that, for $V(x) = \|x\|^2+1$ and $V_4(x) = \|x\|^4+1$,

$$P^n(V)(x) \le (1-\nu\,\Delta t)^n V(x) + C\,n\,\Delta t, \qquad P^n(V_4)(x) \le (1-\nu\,\Delta t)^n V_4(x) + C\,n\,\Delta t.$$

The constants ν and C are independent of Δt sufficiently small. Choose $r>0$ large enough that $-\frac{1}{2}\nu + 2C/r^2 < 0$. Corresponding to this r, choose t_0 according to Theorem 3.6. Then, for $n = \lceil t_0/\Delta t\rceil$,

$$P^n(V)(x) \le \gamma\,V(x) + C_1, \qquad P^n(V_4)(x) \le \gamma\,V_4(x) + C_1, \tag{28}$$

with $\gamma = \exp(-\nu\,t_0)$ and $C_1 = C\,t_0$ and $\Delta t \le 1/2\nu$ (as $(1-\nu\,\Delta t)^n \le \exp(-\nu\,t_0)$). Later, we will exploit the inequality $\gamma + 2C_1/r^2 < 1$ (which holds as $\exp(-x) \le 1-x/2$ for $0<x\le 1$).

We closely follow the proof of [17]: the first part A), where we consider the setting of $V(x)+V(y) \geq r^2$, uses only the drift condition and is identical but repeated here verbatim only for completeness; the second part, for the case $V(x)+V(y) > r^2$, makes a small adaptation of the proof to exploit our coupling probability.

A) Using the drift condition. To show the contraction Eq. (21) when $V(x)+V(y) \geq r^2$, we simply make use of the drift condition and choose ζ accordingly:

$$P^n(\phi)(x) - P^n(\phi)(y) \leq 2 + \zeta\, P^n(V)(x) + \zeta\, P^n(V)(y) \leq 2 + \zeta\,\gamma\,(V(x)+V(y)) + 2\,\zeta\, C_1,$$

where we make use of $\|\phi\|_\zeta \leq 1$ for the first inequality and the drift condition for the second inequality. Since $V(x)+V(y) \geq r^2$, by choosing $\gamma_0 = \gamma + 2C_1/r^2 < 1$, we have

$$2 + \zeta\gamma(V(x)+V(y)) + 2\zeta C_1 \leq 2 + \zeta\gamma_0(V(x)+V(y)) \leq 2\gamma_1 + \zeta\gamma_1(V(x)+V(y)),$$

where the last inequality is due to $2(1-\gamma_1) \leq \zeta(\gamma_1-\gamma_0)r^2 \leq \zeta(\gamma_1-\gamma_0)(V(x)+V(y))$ by choice of $\gamma_1 \coloneqq \frac{2+\zeta\gamma_0 r^2}{(2+\zeta r^2)} < 1$.

B) Using the coupling probability. We now consider the setting where $V(x)+V(y) \leq r^2$, so $x, y \in B(0,r)$ and Theorem 3.6 yields $\mathbb{P}(X^n \neq Y^n) \leq 1-\alpha$. Given a test function ϕ such that $\|\phi\|_\zeta \leq 1$,

$$\begin{aligned}\mathbb{E}[(\phi(X^n)-\phi(Y^n))1_{\{X^n\neq Y^n\}}] &\leq 2\,\mathbb{P}(X^n \neq Y^n) + \zeta\,\mathbb{E}\big[(V(X^n)+V(Y^n))1_{\{X^n\neq Y^n\}}\big] \\ &\leq 2(1-\alpha) + \zeta\sqrt{(1-\alpha)}\sqrt{\mathbb{E}[V(X^n)^2 + V(Y^n)^2]}.\end{aligned}$$

Using the drift condition on V_4, we deduce that

$$\begin{aligned}\mathbb{E}[\phi(X^n)-\phi(Y^n)] &\leq 2(1-\alpha) + \zeta\sqrt{(1-\alpha)}\sqrt{\gamma V(x)^2 + \gamma V(y)^2 + 2C_1} \\ &\leq 2(1-\alpha) + \zeta\sqrt{(1-\alpha)\gamma}\Big(V(x)+V(y)+\sqrt{2C_1/\gamma}\Big).\end{aligned}$$

Choose ζ such that $2(1-\alpha) + \zeta\sqrt{2C_1(1-\alpha)} < 2(1-\tilde{\alpha})$ for some $0 < \tilde{\alpha} < \alpha$; that is, $\zeta < 2(\alpha-\tilde{\alpha})/\sqrt{2C_1(1-\alpha)}$. With $\gamma_2 \coloneqq \max((1-\tilde{\alpha}), \sqrt{(1-\alpha)\gamma})$, we now have

$$|P^n(\phi)(x) - P^n(\phi)(y)| \leq 2(1-\tilde{\alpha}) + \zeta\sqrt{(1-\alpha)\gamma}(V(x)+V(y)) \leq \gamma_2 d_\zeta(x,y).$$

Putting A) and B) together, we see $d_\zeta(P^n(\mu_1,\cdot), P^n(\mu_2,\cdot)) \leq \bar{\alpha}\, d_\zeta(\mu_1,\mu_2)$, and P^n is a contraction with respect to $\|\cdot\|_\zeta$ with contraction constant $\bar{\alpha} \coloneqq \min(\gamma_1,\gamma_2)$. This implies geometric convergence to the invariant measure in the equivalent weighted TV norm as stated in Eq. (22). □

3.2 Finite-time convergence of discretized scheme

We establish convergence of our discretized scheme on a fixed time interval $[0,T]$ under the following assumptions:

Assumption 2. (i) G satisfies Assumption 1 and ∇G is L-Lipschitz.

(ii) (u^0, v^0) and $(u(0), v(0))$ follow an initial distribution $\rho_0 \ll \pi$ such that $\frac{\rho_0}{\pi}$ is uniformly bounded.

For $t_k = k\,\Delta t$, let $e_1(t_k) \coloneqq u_{t_k} - u^k$ and $e_2(t_k) \coloneqq v_{t_k} - v^k$, and $e(t_k) \coloneqq (e_1(t_k), e_2(t_k))$. Our main convergence result is as follows.

Theorem 3.7. *Suppose that Assumption 2 holds. Then, for any $R > 0$,*

$$\mathbb{E}\left[\sup_{0\le t_k\le T}\|e(t_k)\|\right] \le c_R\, d\, \exp(2\,T\,C_R)\Delta t^{\frac{1}{2}-\epsilon} + C\,\sqrt{T}\,e^{-C\,(R^2-T)},$$

where c_R and C_R depends on R. We can choose R and Δt to make the right-hand side arbitrarily small, and hence $\mathbb{E}\big[\sup_{0\le t_k\le T}\|e(t_k)\|\big] \to 0$ as $\Delta t \to 0$.

The rest of this subsection is dedicated to proving Theorem 3.7. Before proceeding with the proof, we first point out the two main technical challenges that we address.

Limited inverse moments. As mentioned, our implicit-explicit scheme with $G \equiv 0$ is the same as the scheme proposed in [9], who studied the discretization scheme for the Lamperti-transformed version of the CIR process (12) under the parameter choice $2\kappa\mu > \theta^2$. Their main result is the strong error bound $\mathbb{E}\left[\sup_{0\le t_k\le T}\left\|u_{t_k} - u^k\right\|\right] = \mathcal{O}(\Delta t^{\frac{1}{2}})$. Their analysis required inverse-moment bounds of the form $\mathbb{E}[u_t^{-2}] < \infty$: this is *false* when $2\kappa\mu = \theta^2$, which is precisely the case for our u process when $G \equiv 0$.

To overcome this, we derive inverse-moment bounds of order up to but excluding 2 (we will make use of Theorem 2.3 and Theorem 2.4 in our proof). As a result, in the case where $G \equiv 0$, our proof shows that the discretized CIR process with parameters $2\kappa\mu = \theta^2$ satisfies for any $\epsilon > 0$, the strong error bound $\mathbb{E}\left[\sup_{0\le t_k\le T}\left\|u_{t_k} - u^k\right\|\right] = \mathcal{O}(\Delta t^{\frac{1}{2}-\epsilon})$. We remark that similar ideas and error bounds were applied in [21].

Lack of global Lipschitzness. The drift in Eq. (18) comprises the explicit part f and implicit part $S_\lambda(u) \coloneqq \frac{1}{\beta\,u} - \lambda u$. The second difficulty is that neither S_λ nor f are globally Lipschitz. In particular, f is at best only locally Lipschitz. To handle this, we will follow the analysis of [19] to stop our process whenever $\|(u, v)\| \ge R$ and upper bound $\mathbb{P}(\|(u, v)_t\| \ge R)$. Intuitively, this can be done since our stationary distribution has exponentially decaying tails.

The proof of Theorem 3.7 is a consequence of Theorem 3.9 below, which shows convergence for a stopped process confined to a ball of radius R. While we manage to show $\Delta t^{1/2-\epsilon}$ rate for a stopped process (restricted to the ball of radius R), our main convergence result, Theorem 3.7, does not provide a rate because our error bound has an exponential dependence on C_R and hence R.

Define the following stopping times

$$k_R \coloneqq \inf\{k \ge 0 :\ \|(u, v)^k\| \ge R\}, \quad \gamma_R \coloneqq \inf\{t \ge 0 :\ \|(u, v)_t\| \ge R\} \tag{29}$$

and $\theta_R \coloneqq k_R\Delta t \wedge \gamma_R$. We have estimates on the tail behaviour of k_R in Theorem 3.4; similar estimates holds for γ_R.

Lemma 3.8. *For some c_1, c_2, c_3 independent of T,R and r, the following holds for $x \in B(0,r)$,*

$$\mathbb{P}(\gamma_R < T) \le \exp(-c_1 R^2)\, \exp(-c_2(r^2 + c_3 T)).$$

Proof. From Eqs. (9) and (10) that, for some $C > 0$ $dV(x_t) \le -2\lambda V(x_t)\, dt + C\, dt + dM_t$, where $M_t = 2\sqrt{2/\beta}\int_0^t \langle x^t, dW_t\rangle$. This is a martingale with quadratic variation $\langle M\rangle_t = \frac{8}{\beta}\int_0^t \|x_s\|^2\, ds$. For $\alpha = \lambda\beta/4$,

$$\begin{aligned} V(x_t) &\le V(x_0) - 2\lambda \int_0^t V(x_s)\, ds + \frac{1}{2}\alpha\frac{8}{\beta}\int_0^t \|x_s\|^2\, ds + C\, t + M_t - \frac{1}{2}\alpha\langle M\rangle_t \\ &\le V(x_0) + C\, t + M_t - \frac{1}{2}\alpha\langle M\rangle_t. \end{aligned}$$

Then, by Doob's martingale inequality,

$$\mathbb{P}\left(\sup_{0\le t\le T}(V(x_t) - V(x_0) - C\, t) \ge R\right) \le \mathbb{P}\left(\sup_{0\le t\le T}\left(M_t - \frac{1}{2}\alpha\langle M\rangle_t\right) \ge R\right) \le \exp(-R\,\alpha).$$

As γ_R is the time of first escape from $B(0,R)$, $V(x_{\gamma_R}) = R^2 + 1$ and

$$\begin{aligned} \mathbb{P}(\gamma_R < T) &\le \mathbb{P}\left(\sup_{0\le t\le T}(V(x_t) - V(x_0) - C\, t) \ge R^2 - \|x_0\|^2 - c\,T\right) \le \exp(-(R^2 - r^2 - cT)\lambda\,\beta/4) \\ &\le \exp(-R^2\,\lambda\,\beta/4)\, \exp(-(r^2 + C\,T)\lambda\,\beta/4). \end{aligned}$$

□

Up to time θ_R, we have the following error.

Proposition 3.9. *Let the assumptions of Theorem 3.7 hold and $\Delta t \le \frac{1}{2C_R}$. For all $\epsilon > 0$,*

$$\mathbb{E}\left[\sup_{0\le t_k\le T}\|e(t_k \wedge \theta_R)\|\right] \le c\, \exp(2\,T\,C_R)\, \Delta t^{\frac{1}{2}-\epsilon}, \tag{30}$$

where c that depends on ϵ, C_R, d, and λ.

Proof. In the following, assume that $t_k \le \theta_R$.

We first subtract the discretized equations with the continuous counterparts,

$$\begin{aligned} u_{t_{k+1}} &= u_{t_k} + \int_{t_k}^{t_{k+1}} \left(-\lambda u_t + v_t \odot \nabla G(u_t \odot v_t) + \frac{1}{u_t}\right) dt + \sqrt{2/\beta}\int_{t_k}^{t_{k+1}} dW_t^1, \\ v_{t_{k+1}} &= v_{t_k} + \int_{t_k}^{t_{k+1}} \left(-\lambda v_t + u_t \odot G(u_t \odot v_t)\right) dt + \sqrt{2/\beta}\int_{t_k}^{t_{k+1}} dW_t^2, \end{aligned}$$

to obtain

$$e_1(t_{k+1}) = e_1(t_k) - \Delta t\left(f_1((u,v)_{t_k}) - f_1((u,v)^k)\right) + \Delta t\left(S_\lambda(u_{t_{k+1}}) - S_\lambda(u^{k+1})\right) + r_{1,k}, \tag{31}$$

$$e_2(t_{k+1}) = e_2(t_k) - \lambda\,\Delta t\, e_2(t_{k+1}) - \Delta t\left(f_2((u,v)_{t_k}) - f_2((u,v)^k)\right) + r_{2,k}, \tag{32}$$

where

$$r_{1,k} \coloneqq \int_{t_k}^{t_{k+1}} (f_1((u,v)_{t_k}) - f_1((u,v)_t))dt + \int_{t_k}^{t_{k+1}} \left(S_\lambda(u_t) - S_\lambda(u_{t_{k+1}})\right)\, dt,$$
$$r_{2,k} \coloneqq \int_{t_k}^{t_{k+1}} (f_2((u,v)_{t_k}) - f_2((u,v)_t))\, dt - \lambda \int_{t_k}^{t_{k+1}} \left(v_t - v_{t_{k+1}}\right)\, dt.$$

Taking inner product of (31) with $e_1(t_{k+1})$ and (32) with $e_2(t_{k+1})$, we have

$$\left(\frac{1}{2} + \lambda\,\Delta t - \frac{C_R\,\Delta t}{2}\right)\|e_1(t_{k+1})\|^2 \le \frac{1}{2}\|e_1(t_k)\|^2 + \frac{C_R\,\Delta t}{2}\|(e_1(t_k), e_2(t_k))\|^2 + \|r_{1,k}\|\,\|e_1(t_{k+1})\|,$$
$$\left(\frac{1}{2} + \lambda\,\Delta t - \frac{C_R\,\Delta t}{2}\right)\|e_2(t_{k+1})\|^2 \le \frac{1}{2}\|e_2(t_k)\|^2 + \frac{C_R\,\Delta t}{2}\|(e_1(t_k), e_2(t_k))\|^2 + \|r_{2,k}\|\,\|e_2(t_{k+1})\|,$$

where we used that $a^\top b \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$, that f_i is C_R-Lipschitz, and that

$$\left(S_\lambda(u_{t_{k+1}}) - S_\lambda(u^{k+1})\right)^\top e_1(t_{k+1}) \le -\lambda\|e_1(t_{k+1})\|^2$$

due to monotonicity arguments. Denote $E_k = \|(e_1(t_k), e_2(t_k))\|$ and $R_k = \|(r_{1,k}, r_{2,k})\|$ and sum the two equations to find

$$(1 + 2\,\lambda\,\Delta t - C_R\,\Delta t)E_{k+1}^2 \le (1 + 2\,C_R\,\Delta t)E_k^2 + 2\,E_{k+1}R_k.$$

Define $c_1 = (1 + 2\,C_R\,\Delta t)/(1 + 2\lambda\,\Delta t - \,C_R\,\Delta t)$ and $c_2 = 2/(1 + 2\,\lambda\,\Delta t - C_R\,\Delta t)$. Then,

$$E_{k+1}^2 \le c_1 E_k^2 + c_2 E_{k+1}R_k \le c_1^{k+1}E_0^2 + c_2\sum_{j=0}^{k} c_1^j E_{k+1-j}R_{k-j} \le c_2 \sup_{k\le M\wedge\theta_R} E_k \sum_{j=0}^{k} c_1^j R_{k-j}.$$

We therefore have $\mathbb{E}\left[\sup_{k\le M\wedge\theta_R} E_k\right] \le c_2\sum_{j=0}^{M-1} c_1^j\,\mathbb{E}[R_{k-j}]$. Note that for $\Delta t \le 1/(2C_R)$, we have $c_1 \le (1 + 2\,\Delta t\,C_R)^2$: this is a consequence of the simple inequality $(1-a)^{-1} \le 1 + 2a$ applied to $a = \Delta t\,C_R \le 1/2$. Also using the geometric sum formula and the inequality $e^z \ge 1 + z$ for $z \ge 0$, we get

$$\sum_{j=0}^{M-1} c_1^j \le \sum_{j=0}^{M-1}(1 + 2\,\Delta t\,C_R)^{2j} \le \frac{\exp(2\,C_R\,M\,\Delta t)}{4\,C_R\,\Delta t + 4\,C_R^2\,\Delta t^2}.$$

On the other hand, $\mathbb{E}[R_k] \le \mathbb{E}[\|r_{1,k}\|] + \mathbb{E}[\|r_{2,k}\|]$. Observe that by Theorem 2.4 and the definition S_λ in (19),

$$\begin{aligned}\mathbb{E}\|r_{1,k}\| &\le \int_{t_k}^{t_{k+1}} \mathbb{E}\|f_1((u,v)_{t_k}) - f_1((u,v)_t)\| + \lambda\mathbb{E}\left\|u_t - u_{t_{k+1}}\right\|\, dt + \frac{1}{\beta}\sum_{i=1}^{d}\int_{t_k}^{t_{k+1}} \mathbb{E}\left[\frac{\left|u_{i,t} - u_{i,t_{k+1}}\right|}{u_{i,t}u_{i,t_{k+1}}}\right] dt \\ &\le c\Big((C_R + \lambda)\,\Delta t^{3/2} + d\,\Delta t^{\frac{3}{2}-\epsilon}\Big).\end{aligned}$$

Similarly, $\mathbb{E}[\|r_{2,k}\|] \le c\,(C_R+\lambda)\,\Delta t^{3/2}$. It follows that $\mathbb{E}\left[\sup_{0\le t_k\le T\wedge\theta_R} E_k\right] \le c_3\,\exp(2\,C_R\,T)\,\Delta t^{\frac{1}{2}-\epsilon}$, for a constant c_3 that depends on λ, C_R, and d, as required. □

We are now ready to prove the main result.

Proof of Theorem 3.7. Define $\tau_R \coloneqq k_R \Delta t$ (see Eq. (29)). First note that $\mathbb{E}\big[\sup_{0\le t_k\le T}\|e(t_k)\|\big] \le A+B$ where $A \coloneqq \mathbb{E}\big[\sup_{0\le t_k\le T}\|e(t_k)\|1_{\{\tau_R>T,\gamma_R>T\}}\big]$ and $B \coloneqq \mathbb{E}\big[\sup_{0\le t_k\le T}\|e(t_k)\|1_{\{\tau_R<T \text{ or } \gamma_R<T\}}\big]$. By Theorem 3.9, $A \le c\,\exp(2\,T\,C_R)\,d\,\Delta t^{\frac{1}{2}-\epsilon}$. Using the Cauchy–Schwarz inequality, we bound B as $B \le \sqrt{\mathbb{E}\big[\sup_{0\le t_k\le T}\|e(t_k)\|^2\big]}\sqrt{\mathbb{P}(\gamma_R<T)+\mathbb{P}(\tau_R<T)}$. From (26) with $\alpha = 1/T$,

$$\mathbb{E}\left[\sup_{0\le t_k\le T}\|(u,v)^k\|^2\right] \le \|(u,v)^0\|^2 + C\,T + \int_0^\infty e^{-\alpha\,r}\,dr,$$

which is uniformly bounded with respect to Δt. Finally, applying Theorems 3.4 and 3.8, $B \le C\sqrt{T}e^{-C(R^2-T)}$, which completes the proof. □

3.3 Invariant measure

To conclude this section, we combine Theorem 3.2 and Eq. (7) to show that the stationary distribution $\pi_{\Delta t}$ of $(u,v)^k$ converges to π. Here W_1 denotes the usual Wasserstein 1-norm.

Theorem 3.10 (convergence to invariant measure)**.** *Let Assumption 2 hold. Then,* $W_1(\pi_{\Delta t},\pi) \to 0$ *as* $\Delta t \to 0$.

Proof. By geometric ergodicity of the numerical approximation in Theorem 3.2 with $t_k = T$, we have $|\int \phi(x)\,P^k(\pi,dx) - \int \phi(x)\pi_{\Delta t}(dx)| \le \bar\rho^T(\|\pi\|_\zeta + \|\pi_{\Delta t}\|_\zeta))$ for any 1-Lipschitz function ϕ. As $\pi(V)$ and $\pi_{\Delta t}(V)$ are bounded uniformly in Δt, we may choose c_3 independent of Δt such that $\|\pi\|_\zeta + \|\pi_{\Delta t}\|_\zeta \le c_3$ and

$$W_1(P^k(\pi,\cdot),\pi_{\Delta t}) \le c_3\bar\rho^T. \tag{33}$$

Denote by P^t the transition density of the solution x^t of Eq. (7). The initial distribution $\rho_0 = \pi$ satisfies Assumption 2(ii) and Theorem 3.7 gives

$$\sup_{0\le t_k\le T} W_1(P^k(\pi,\cdot),P^{t_k}(\pi,\cdot)) \to 0, \qquad \text{as } \Delta t \to 0. \tag{34}$$

Divide the error into finite-time numerical error and distance to the long-time average:

$$W_1(\pi,\pi_{\Delta t}) = W_1(P^k(\pi,\cdot),P^{t_k}(\pi,\cdot)) + W_1(P^k(\pi,\cdot),\pi_{\Delta t}).$$

We now choose $T = |\log(\delta/2\,c_3)|/|\log\bar\rho|$, in Eq. (33) with $t_k = T$ to ensure $W_1(P^k(\pi,\cdot),\pi_{\Delta t}) \le \frac{1}{2}\delta$; and the timestep Δt in Eq. (34) so that $W_1(P^k(\pi,\cdot),P^{t_k}(\pi,\cdot)) \le \frac{1}{2}\delta$. We see now that $W_1(\pi,\pi_{\Delta t}) \le \delta$. □

4 Conclusion and outlook

We proposed a new approach for sampling distributions with the Laplace prior via a Hadamard parameterization. Our resulting Hadamard–Langevin Dynamics can be seen as an implicit form of Riemannian Langevin. The analysis in this work covered well-posedness of the continuous-time dynamics, strong convergence of the discretized dynamics, as well as geometric ergodicity results in both continuous and discrete time. These results provide a theoretical foundation for developing sampling schemes based on overparameterized Langevin dynamics. In the following, we detail the main remaining missing piece in our analysis of the Hadamard parametrization, and also the challenges in extending this to other regularizers.

Theoretical limitations. On a theoretical level, for our numerical scheme, we established a strong convergence result (i.e., convergence as stepsize diminishes over a fixed time interval), and convergence of the invariant distribution $\pi_{\Delta t}$ for the discretized process. However, we are not able to establish quantitative rates of convergence. As mentioned in [26, Section 3], such convergence rates are typically obtained by understanding the Taylor expansion of the semigroup operator $P_{\Delta t}\phi = \mathbb{E}[\phi(x_{t_{k+1}}) \mid x_{t_k} = \cdot]$. The difficulty of a theoretical proof is that these Taylor expansions would include singular $1/u$ terms and this lack of differentiability prevent the direct use of results such as [26, Theorem 3.3]. One important direction of future work is to quantify the bias of $\pi_{\Delta t}$ and to provide quantitative strong convergence rates.

One-dimensional experiment. In the absence of a theoretical characterization, we provide some numerical evidence on a simple one-dimensional problem in Figure 1 below that this bias is $\mathcal{O}(\Delta t)$.

We consider the case where $d = 1$, and choose $A = 1$, $y = 3$, and $\lambda = 2.7$. We let $G(x) = \|x - y\|^2/2$. We carry out two experiments with $\phi(x) = x^2$. To estimate the impact of Δt for the stationary distribution, for different stepsizes, we run our Hadamard–Langevin (HLD) algorithm for $T = 800$ before recording 10^6 samples. Figure 1(a) shows the error between the mean of $\phi(u \odot v)$ evaluated via these samples compared with the integral of the true density ρ (i.e., (1)) against ϕ (we compute this numerically using scipy.integrate). This procedure is also repeated for Proximal Langevin, where the Moreau-envelope regularization choice for Prox-l1 are taken to be $\gamma = 1/(KL)$ for some $K \geq 1$ and $\Delta t = \gamma/(5(\gamma L + 1)$, where $L = \|A\|^2$ is the Lipschitz constant of the gradient of G. This is consistent with the parameter choices recommended in [11]. We observe $\mathcal{O}(\Delta t)$ errors for Proximal Langevin and just slightly slower than $\mathcal{O}(\Delta t)$ for HLD, although the error is typically lower for moderate stepsizes.

To assess the convergence to the stationary distribution with fixed stepsize, fix $\Delta t = 5 \times 10^{-4}$, and generate 5×10^4 paths of HLD and record the means $\phi(u \odot v)$ over these paths for 10^4 iterations. Figure 1(b) shows the error between the mean of $\phi(u \odot v)$ over the first 3000 iterations against the mean of $\phi(u \odot v)$ over the last 1000 iterations (which we expect to estimate the integral of ϕ against the stationary distribution). The same procedure is repeated for Proximal Langevin. In both cases, we observe exponential decay of the error, which corroborates the exponential ergodicity result of Theorem 3.2.

The code for reproducing this numerical experiment and also further higher-dimensional

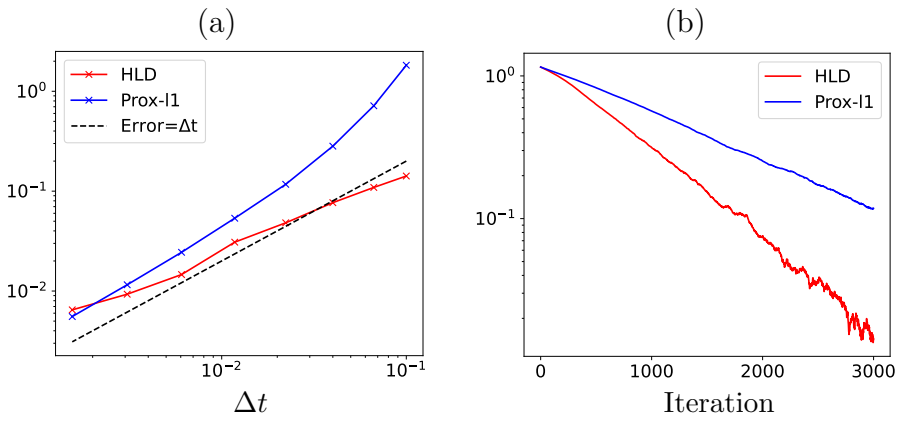


Figure 1: Estimation of the convergence rate for a 1D problem. (a) Error of the 'stationary distribution' for different stepsizes. (b) Error decay with increasing iterations for fixed stepsize.

implementations of our method can be found online[1].

Other sparse priors. Although this study is specific to the Laplace prior, one central idea is the development of Langevin dynamics based on Gaussian mixture representation of this prior. We first remark that the extension to the group-ℓ_1 prior is trivial. The group-ℓ_1 prior is $\|x\|_{1,2} \coloneqq \sum_{j=1}^{K} \left\|x_{\mathbf{b}_j}\right\|^2$, where $\cup_{j=1}^{K} \mathbf{b}_j$ is a partition of $\{1, \dots, d\}$. In this case, one can consider for $(u, v) \in \mathcal{X} = \mathbb{R}_+^K \times \mathbb{R}^d$

$$\pi(u, v) \propto \prod_{i=1}^{K} u_i \exp\left(-\beta\left(\frac{\lambda}{2}\|(u, v)\|^2 + G(u \odot v)\right)\right),$$

where $u \odot v \coloneqq (u_i\, v_{\mathbf{b}_i})_{i=1}^K$. Samples of $\pi(x) \propto \exp\Big(-\beta\|x\|_{1,2} - \beta G(x)\Big)$ can be retrieved by sampling from π and forming the product $u \odot v$ as before. For this setting, $u \in \mathbb{R}_+^K$ while $v \in \mathbb{R}^d$ with $K \leq d$, and the only difference here is the different definition of $\odot$. Likewise, the Langevin dynamics are identical to the ℓ_1 setting, except $\odot$ and the different dimensions of u and v.

A natural question is how to handle priors corresponding to ℓ_q semi-norms with $q \in (0, 1)$. These priors also have a Gaussian mixture representation [39]

$$e^{-|z|^q} \propto \int_0^\infty \frac{1}{\sqrt{2\pi\eta}}\, e^{-z^2/(2\eta)}\, \frac{1}{\eta^{3/2}}\, g_{q/2}\left(\frac{1}{2\eta}\right) d\eta,$$

where $g_{q/2}$ is the density of a positive stable random variable with index $q/2$. Unfortunately, this has no closed-form expression in general so it is unclear how to perform Langevin sampling for such priors. Nonetheless, given the developments of our present work, it is perhaps conceivable that one could construct new sparsity-inducing priors and efficient Langevin sampling schemes via Gaussian-mixture representations, and we believe that this

[1] https://hadamardlangevin.readthedocs.io/

is an exciting direction for future work. In particular, recent works such as [4] derived generalizations of the hierarchical samplers for the Bayesian lasso to new sparsity inducing Gaussian mixture models. Our article can be seen as a first attempt at deriving Langevin dynamics for such priors.

A Ergodicity under bounded assumptions

Consider two measures π_1, π_2 such that π_1 is absolutely continuous with respect to π_2 and denote the Radon–Nikodym derivative by $d\pi_1/d\pi_2$. Define the entropy of π_1 with respect to π_2 (also called the Kullback–Leibler or KL divergence) by $\mathcal{H}(\pi_1 \mid \pi_2) = \mathbb{E}_{\pi_1}(\log(d\pi_1/d\pi_2))$ and the Fisher information $\mathcal{I}(\pi_1 \mid \pi_2) = \mathbb{E}_{\pi_1}\big[\|\nabla \log(d\pi_1/d\pi_2)\|^2\big]$. We say π_2 satisfies a log-Sobolev inequality LSI(R) if, for a constant $R > 0$,

$$\mathcal{H}(\pi_1 \mid \pi_2) \le \frac{1}{2\,R}\mathcal{I}(\pi_1 \mid \pi_2), \qquad \forall \pi_1 \ll \pi_2.$$

The theory of log Sobolev inequalities (e.g., [26]) gives the following geometric convergence of $\rho(t,\cdot)$ to the invariant measure π.

Theorem A.1. *Assume that G is bounded. The solution $(u,v)_t$ of Eq. (7) with initial distribution ρ_0 converges to the invariant measure π geometrically. In particular,*

$$\mathcal{H}(\rho(t,\cdot) \mid \pi) \le \exp(-2\,R\,t/\beta)\,\mathcal{H}(\rho_0 \mid \pi), \quad t > 0, \tag{SM35}$$

for $R = \beta\,\lambda\,e^{\inf G - \sup G}$.

The assumption that G is bounded holds for $G(x) = \sum \sigma(y_i a_i^\top x)$ when σ is bounded such as the smooth approximations of the unit loss that are mentioned in the introduction.

Proof. In terms of the density $p(t,\cdot)$ for Eq. (8) and the Cartesian co-ordinates $\mathbf{x}_t \in \mathbb{R}^{3d}$, standard theory (Bakry–Emery criterion, Holley–Stroock theorem) applies to the potential $\frac{1}{2}\lambda\|\mathbf{x}\|^2 + G_3(\mathbf{x})$, as it is a bounded perturbation (given by G) of a convex potential (given by $\frac{1}{2}\lambda\|\mathbf{x}\|^2$). See [26, Proposition 2.9]. This gives $\mathcal{H}(p(t,\cdot) \mid \pi) \le \exp(-2\,R\,t/\beta)$ for R defined above. Taking a cylindrical initial distribution, $p(t,\cdot) = \rho(t,\cdot)$, we have Eq. (SM35). □

B Link to Riemannian Langevin

Proof of Proposition 2.7. The proof is a direct application of Itô's formula. By Itô's formula,

$$d\mathbf{z}_t = \underbrace{\begin{bmatrix} (-2\lambda\eta_t - \frac{1}{2}x_t \odot \nabla G(x_t) + \frac{1}{2}) + \frac{1}{2} \\ -(4\eta_t + \frac{x_t^2}{4\eta_t}) \odot \nabla G(x_t) - 2\lambda x_t + \frac{x_t}{4\eta_t} \end{bmatrix}}_{\mu(\mathbf{z}_t)} dt + \sqrt{2}\underbrace{\begin{bmatrix} \frac{1}{2}\mathrm{diag}(u_t) & 0 \\ \mathrm{diag}(v_t) & \mathrm{diag}(u_t) \end{bmatrix}}_{S(\mathbf{z}_t)} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix}.$$

Note that for $\mathbf{g}(\mathbf{z}) = M(z)^{-1}$,

$$-M(\mathbf{z})\nabla H(\mathbf{z}) = \begin{bmatrix} -2\lambda\eta_t - \frac{1}{2}x_t \odot \nabla G(x_t) - \frac{1}{2} \\ -\left(4\eta_t + \frac{x_t}{4\eta_t}\right) \odot \nabla G(x_t) - 2\lambda x_t - \frac{x_t}{4\eta_t} \end{bmatrix} \quad \text{and} \quad \Gamma(\mathbf{z}_t) = \frac{1}{2}\begin{bmatrix} 3 \cdot 1_d \\ x_t/\eta_t \end{bmatrix}.$$

So, $\mu(z) = -\mathbf{g}(\mathbf{z})^{-1}\nabla H(\mathbf{z}) + \Gamma(\mathbf{z}_t)$. Note also that $M(\mathbf{z}) = S(\mathbf{z})S(\mathbf{z})^\top$.

Letting $\rho(t, \cdot)$ denote the distribution of $\mathbf{z}_t$, we have

$$\frac{\partial}{\partial t}\rho = -\mathrm{div}\big(\mu^\top(\mathbf{z})\rho\big) + \sum_i \sum_j \frac{\partial^2}{\partial_{\mathbf{z}_i}\partial_{\mathbf{z}_j}}(M_{i,j}(\mathbf{z}, t)\rho).$$

Note that

$$\begin{aligned} \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\partial^2}{\partial_{\mathbf{z}_i}\partial_{\mathbf{z}_j}}(M_{i,j}(\mathbf{z}, t)\rho) &= \mathrm{div}(M\nabla\rho) + \mathrm{div}\left(\left(\sum_j \frac{\partial}{\partial \mathbf{z}_j} M_{i,j}(\mathbf{z}, t)\rho\right)_i\right) \\ &= \mathrm{div}(M\nabla\rho) + \mathrm{div}\left(\begin{pmatrix} 3/2 \\ \frac{x}{2\eta} \end{pmatrix}\rho\right). \end{aligned}$$

So,

$$\begin{aligned} \frac{\partial}{\partial t}\rho &= \mathrm{div}\left(\begin{bmatrix} (2\lambda\eta + \frac{1}{2}x \odot \nabla G(x) - \frac{1}{2}) - \frac{1}{2} + \frac{3}{2} \\ (4\eta + \frac{x^2}{4\eta})\nabla G(x) + 2\lambda x - \frac{1}{2}\frac{x}{2\eta} + \frac{1}{2}\frac{x}{\eta} \end{bmatrix}\right) + \mathrm{div}(M\nabla\rho) \\ &= \mathrm{div}\left(\begin{pmatrix} \eta & \frac{1}{2}x \\ \frac{1}{2}x & \frac{x^2}{4\eta} + 4\gamma \end{pmatrix}\left(\begin{pmatrix} 2\lambda - \lambda\frac{x^2}{8\eta^2} + \frac{1}{2\eta} \\ \nabla G(x) + \lambda\frac{x}{4\eta} \end{pmatrix} + \nabla \log\rho\right)\rho\right) \\ &= \mathrm{div}(M(\mathbf{z})\nabla(H(\mathbf{z}) + \log(\rho)\rho). \end{aligned}$$

□

C Additional proofs for the discretized scheme

We first remark that our numerical scheme (18) can be written explicitly as

$$\begin{aligned} \begin{pmatrix} u^{k+\frac{1}{2}} \\ v^{k+\frac{1}{2}} \end{pmatrix} &= \begin{pmatrix} u^k \\ v^k \end{pmatrix} - \Delta t \begin{pmatrix} v^k \odot \nabla G(u^k \odot v^k) \\ u^k \odot \nabla G(u^k \odot v^k) \end{pmatrix} + \sqrt{\frac{2}{\beta}}\begin{pmatrix} \Delta W_k^1 \\ \Delta W_k^2 \end{pmatrix}, \\ u^{k+1} &= \frac{u^{k+\frac{1}{2}} + \sqrt{u^{k+\frac{1}{2}} \odot u^{k+\frac{1}{2}} + 4(\Delta t/\beta)(1 + \Delta t\lambda)}}{2(1 + \Delta t\lambda)}, \\ v^{k+1} &= \frac{1}{1 + \Delta t\lambda} v^{k+\frac{1}{2}}. \end{aligned} \tag{SM36}$$

where the square root is interpreted element-wise and we have used the fact that, for $u > 0$,

$$u = \Delta t \frac{1}{\beta u} - \Delta t \lambda u + w \iff (1 + \Delta t\lambda)u^2 - w u - \Delta t/\beta = 0$$
$$\iff u = \frac{w + \sqrt{w^2 + 4(\Delta t/\beta)(1 + \Delta t\lambda)}}{2(1 + \Delta t\lambda)}.$$

It is straightforward to see from this that positivity of u^k is preserved.

C.1 Explicit form of the transition density

Lemma C.1. *Up to constants, the log transition kernel P for moving from $x = (u, v)^k$ to $\hat{x} = (u, v)^{k+1}$ under Eq. (18) is*

$$\log(P)(x, \hat{x}) = -\frac{\beta}{4\Delta t}\left\|(1 + \Delta t\,\lambda)u^{k+1} - \frac{\Delta t}{\beta u^{k+1}} - u^k + \Delta t\, v^k \odot \nabla G(u^k \odot v^k))\right\|^2$$
$$- \frac{\beta}{4\Delta t}\left\|(1 + \Delta t\,\lambda)v^{k+1} - v^{k+1} + \Delta t\, u^k \odot \nabla G(u^k \odot v^k)))\right\|^2$$
$$+ \sum_{i=1}^{d} \log\left(1 + \Delta t\,\lambda + \frac{\Delta t}{\beta (u_i^k)^2}\right).$$

Proof. Let $x = (u, v)$ and $g(x) = (v \odot \nabla G(u \odot v), u \odot \nabla G(u \odot v))$. From Eq. (18), we have $P(x) = P_1(P_0(x))$ for

$$P_0(x) = x - \Delta t\, g(x) + \sqrt{\frac{2\Delta t}{\beta}}\mathcal{N}(0, \mathrm{Id}_{2d})$$

and $P_1(x) = (P_{1,u}(x), P_{1,v}(x))$ with

$$P_{1,u}(x) = \frac{u + \sqrt{u \odot u + \frac{4}{\beta}\Delta t(1 + \Delta t\lambda))}}{2(1 + \Delta t\lambda)}, \qquad P_{1,v}(x) = \frac{v}{(1 + \Delta t\lambda)}.$$

Note that $P_0(x) \sim \mathcal{N}(x - \Delta t\, g(x), (2\,\Delta t/\beta)\,\mathrm{Id}_{2d})$ is Gaussian with transition density

$$f_0(z) \propto \exp\left(-\beta \frac{(z - (x - \Delta t\, g(x)))^2}{4\,\Delta t}\right)$$

To work out the density of $u = P_{1,u}(z)$ with $z = P_0(x)$, write the cumulative distribution function $F_{1,u}(t)$ as

$$F_{1,u}(t) = \mathbb{P}(u \le t) = \mathbb{P}\left(\frac{z + \sqrt{z \odot z + \frac{4}{\beta}\Delta t\,(1 + \Delta t\,\lambda))}}{2(1 + \Delta t\,\lambda)} \le t\right)$$
$$= \mathbb{P}\left(z \le (1 + \Delta t\,\lambda)t - \frac{\Delta t}{t\,\beta}\right).$$

It follows that the density of Y_u is

$$f_{1,u}(t) = \frac{d}{dt}F_{1,u}(t) \propto f_0\bigg((1+\Delta t\,\lambda)\,t - \frac{\Delta t}{\beta\,t}\bigg)\bigg((1+\Delta t\,\lambda) + \frac{\Delta t}{\beta\,t^2}\bigg).$$

The density of $P_{1,v}(v)$ is $f_{1,v}(t) \propto f_0((1+\Delta t\,\lambda)\,t)$. □

C.2 Proof of drift condition of order 4

Lemma C.2. *Let $V_4((u,v)) = 1 + \|(u,v)\|^4$. There exists $C, c, \nu > 0$ such that for all $\Delta t < c\lambda$,*

$$\mathbb{E}[V_4((u,v)^{k+1}) \mid (u,v)^k] \le (1-\nu\,\Delta t)\,V_4((u,v)^k) + C\,\Delta t.$$

Proof of Lemma C.2. Since

$$(1+\lambda\,\Delta t)u^{k+1} - \frac{\Delta t}{\beta\,u^{k+1}} = u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t + \sqrt{\frac{2}{\beta}}\Delta W_k^1, \tag{SM37}$$

taking square-norm yields

$$(1+\lambda\Delta t)^2\|u^{k+1}\|^2 - \frac{2\,d\,\Delta t\,(1+\lambda\,\Delta t)}{\beta} \le \left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^2 + \frac{2}{\beta}\,\|\Delta W_k^1\|^2 \tag{SM38}$$

$$+ \sqrt{\frac{4}{\beta}}\left\langle \Delta W_k^1, u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\rangle. \tag{SM39}$$

Squaring this again and taking expectation conditional on u^k, v^k gives

$$\begin{aligned}
&(1+\lambda\Delta t)^4\mathbb{E}[\|u^{k+1}\|^4 \mid (u,v)^k] - \frac{4d\Delta t(1+\lambda\Delta t)^3}{\beta}\mathbb{E}[\|u^{k+1}\|^2 \mid (u,v)^k] \\
&\le \left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^4 + \frac{4}{\beta^2}\mathbb{E}\|\Delta W_k^1\|^4 + \frac{4}{\beta}\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\|^2\mathbb{E}\|\Delta W_k^1\|^2 \\
&\quad + \frac{4}{\beta}\mathbb{E}\|\Delta W_k^1\|^2\left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^2.
\end{aligned}$$

We now use the fact that $\mathbb{E}_k\|\Delta W_k^1\|^2 = d\,\Delta t$ and $\mathbb{E}_k\|\Delta W_k^1\|^4 = (d^2+2d)\,\Delta t^2$ to obtain

$$\begin{aligned}
&(1+\lambda\Delta t)^4\mathbb{E}[\|u^{k+1}\|^4 \mid (u,v)^k] - \frac{4d\Delta t(1+\lambda\Delta t)^3}{\beta}\mathbb{E}[\|u^{k+1}\|^2 \mid (u,v)^k] \\
&\le \left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^4 + \frac{4(d^2+2d)\Delta t^2}{\beta^2} + \frac{8d\Delta t}{\beta}\left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^2.
\end{aligned}$$

Applying Young's inequality for products gives that, for any $\alpha_1, \alpha_2 > 0$,

$$\begin{aligned}
&\big((1+\lambda\Delta t)^4 - \alpha_1\Delta t\big)\mathbb{E}[\|u^{k+1}\|^4 \mid (u,v)^k] \\
&\le (1+\alpha_2\Delta t)\left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^4 + \frac{4(d^2+2d)\Delta t^2}{\beta^2} + \frac{64d^2\Delta t}{\beta^2\alpha_2} + \frac{4d^2\Delta t(1+\lambda\Delta t)^6}{\alpha_1\beta^2}.
\end{aligned}$$

As in the proof of Theorem 3.3, we have

$$\left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^2 \leq \|u^k\|^2 + \Delta t^2 B^2 \|v^k\|^2 + 2K\Delta t$$

and squaring this gives

$$\begin{aligned}
&\left\|u^k - v^k \odot \nabla G(u^k \odot v^k)\Delta t\right\|^4 \\
&\leq \|u^k\|^4 + \Delta t^4 B^4 \|v^k\|^4 + 2\|u^k\|^2 \Delta t^2 B^2 \|v^k\|^2 + 4K\Delta t\|u^k\|^2 + 4K\Delta t^3 B^2 \|v^k\|^2 + 4K^2\Delta t^2 \\
&\leq (1 + \Delta t^2 B^2)\|u^k\|^4 + (B^2\Delta t^2 + \Delta t^4 B^4)\|v^k\|^4 \\
&\qquad + \alpha_3 \Delta t\|u^k\|^4 + \frac{4K^2\Delta t}{\alpha_3} + \alpha^3 \Delta t^2 \|v^k\|^4 + \frac{4K^2\Delta t^4 B^4}{\alpha_3} + 4K^2\Delta t^2 \\
&= (1 + \alpha_3\Delta t + \Delta t^2 B^2)\|u^k\|^4 + \Delta t^2(\alpha_3 + B^2 + \Delta t^2 B^4)\|v^k\|^4 + 4K^2\Delta t(\alpha_3^{-1} + \Delta t^3 B^4 \alpha_3^{-1} + \Delta t).
\end{aligned}$$

It follows that

$$\big((1 + \lambda\Delta t)^4 - \alpha_1\Delta t\big)\mathbb{E}\big[\|u^{k+1}\|^4 \mid (u,v)^k\big] \leq c_1\|u^k\|^4 + c_2\Delta t^2\|v^k\|^4 + c_3\Delta t, \tag{SM40}$$

where

$$c_1 = (1 + \alpha_2\Delta t)(1 + \alpha_3\Delta t + \Delta t^2 B^2), \qquad c_2 = (1 + \alpha_2\Delta t)(\alpha_3 + B^2 + \Delta t^2 B^4),$$

and

$$c_3 = (1 + \alpha_2\Delta t)4K^2(\alpha_3^{-1} + \Delta t^3 B^4 \alpha_3^{-1} + \Delta t) + \frac{64d^2}{\beta^2\alpha_2} + \frac{4d^2(1 + \lambda\Delta t)^6}{\alpha_1\beta^2}.$$

One can carry out a similar computation for v^{k+1} to obtain

$$(1 + \lambda\Delta t)^4\mathbb{E}[\|v^{k+1}\|^4 \mid (u,v)^k] \leq c_1\|v^k\|^4 + c_2\Delta t^2\|u^k\|^4 + \tilde{c}_3\Delta t,$$

with

$$\tilde{c}_3 = (1 + \alpha_2\Delta t)4K^2(\alpha_3^{-1} + \Delta t^3 B^4 \alpha_3^{-1} + \Delta t) + \frac{64d^2}{\beta^2\alpha_2}.$$

It follows that

$$\left((1 + \lambda\Delta t)^4 - \alpha_1\Delta t\right)\mathbb{E}[\|(u,v)^{k+1}\|^4 \mid (u,v)^k] \leq (c_1 + c_2\Delta t^2)\|(u,v)^k\|^4 + 2c_3\Delta t. \tag{SM41}$$

Finally, define

$$\alpha \coloneqq \frac{c_1 + c_2\Delta t^2}{(1 + \lambda\Delta t)^4 - \alpha_1\Delta t}. \tag{SM42}$$

Then, $\alpha < 1$ provided that

$$\alpha_1\Delta t + (1 + \alpha_2\Delta t)(1 + \alpha_3\Delta t + \Delta t^2 B^2) + \Delta t^2(1 + \alpha_2\Delta t)(\alpha_3 + B^2 + \Delta t^2 B^4) < (1 + \lambda\Delta t)^4.$$

The leading exponent to Δt^0 on both sides is 1, and the leading exponent to Δt is 4λ and $\alpha_1 + \alpha_2 + \alpha_3$ on the right and left and side respectively. By choosing $\alpha_1, \alpha_2, \alpha_3 \lesssim \lambda$ and $\Delta t \lesssim \lambda$, we have $\alpha < 1$. In other words, we may choose $\nu > 0$ and write $\alpha \leq 1 - \nu\Delta t$ for Δt sufficiently small. □

D Numerical experiments

The code for reproducing our numerical experiments can be found online[2]. Throughout, we consider the setting where $G(x) = \|Ax - y\|^2/2$. The stepsize and Moreau-envelope regularization choice for Prox-l1 are taken to be $\gamma = 1/(KL)$ for some $K \geq 1$ and $\Delta t = \gamma/(5(\gamma L+1)$, where $L = \|A\|^2$ is the Lipschitz constant of the gradient of G. This is consistent with the parameter choices recommended in [11].

D.1 Synthetic experiment

We consider the case of $d = 20$, and $A \in \mathbb{R}^{40\times 20}$ is a random Gaussian matrix of mean zero and variance $\frac{1}{16m}$. We let $\lambda = \frac{1}{2}\left\|A^\top y\right\|_\infty$ and $y = Ax_0$ where x_0 is a 2-sparse vector. For Proximal and Hadamard Langevin, we run 10^4 burn-in iterations before recording 10^5 samples, while for Gibbs (due to its slower running time and faster convergence), we record 10^4 samples after 10 burn-in samples. To assess the convergence of each scheme, we evaluate the effective sample size (ESS) which computed using the package arviz [24]. The sample means for each of the 20 dimensions are shown in Figure 2. Figure 2 also shows the ESS in each dimension. Note that Gibbs has the highest ESS, followed by Hadamard. In general, for these low-dimensional settings, the Gibbs sampler displays fast mean convergence behavior, and the Hadamard sampler also displays a similar pattern (although less sharp than the Gibbs sampler). Additional imaging experiments can be found in Section D.

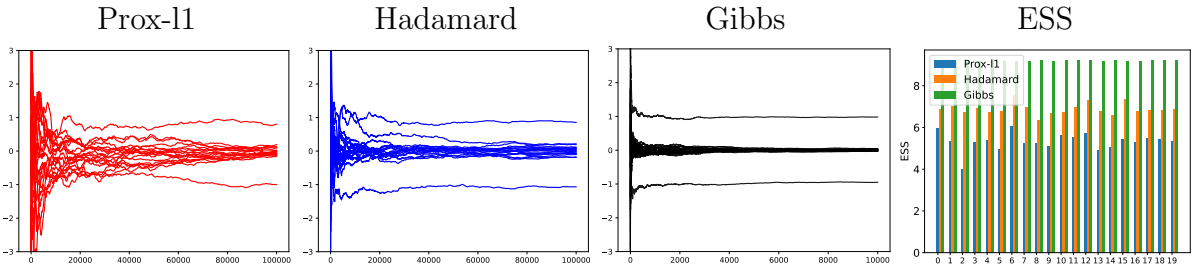


Figure 2: Plots of sample means against iteration. The right figure shows the ESS in log-scale at of the 20 dimensions. The minimum ESS (in log scale) for Prox-l1 is 4.0, for Hadamard is 6.4 and for Gibbs is 9.2.

D.2 Deconvolution with Haar wavelets

We consider Gaussian deconvolution of a piecewise-constant signal (length $d = 1024$). The matrix $A = UW^\top$ where U is a convolution operator and W is an orthogonal Haar transform. The regularization parameter is $\lambda = 1$. We run Proximal-Langevin and Hadamard at stepsize $\tau = 0.01$, with 10^2 burn-in iterations and record 10^5 samples. The sample means and ESS scores are shown in Figure 3. The ground truth is displayed as x0. In log scale, we also

[2] https://hadamardlangevin.readthedocs.io/

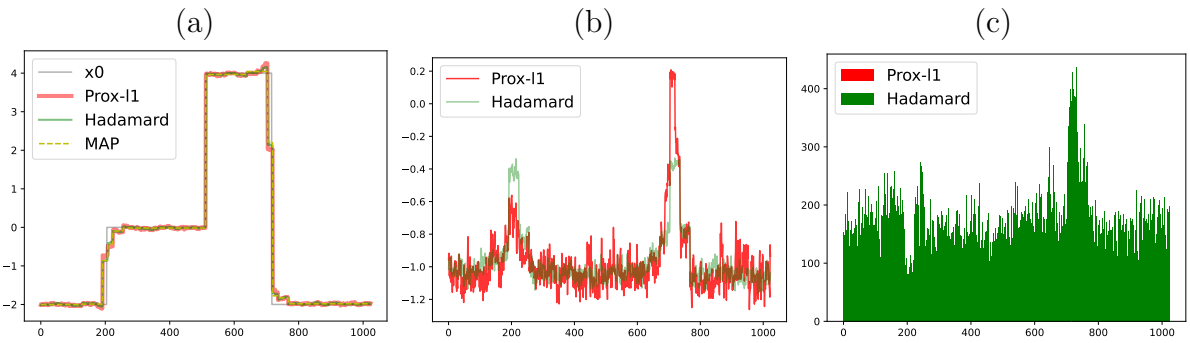


Figure 3: 1d Haar wavelet deconvolution. (a) the mode and mean signals. (b) the difference between the 95 quantile and the 5 quantile. (c) the ESS at each pixel. The minimum ESS across all pixels for Hadamard is 286,433.7 while the minimum ESS for Prox-l1 is 4.1.

plot the difference between the top 95 quantile and the bottom 5 quantile. Notice that both methods have a higher difference around the jumps, which naturally aligns with the fact that reconstruction of the precise jump locations is more unstable.

D.3 Image inpainting

We perform inpainting: $A = UW^\top$ where U is a masking operator and W is the wavelet transform with the Haar-wavelet, with $\beta = 5, 100, 500$. For both Proximal and Hadamard Langevin, we run 10,000 burn-in iterations, then record every 20 samples out of 6000 samples. On the 300 recorded samples, we compute the mean images and quantile differences. The corrupted input as well as the mode image are shown in Figure 4 in the case of $\beta = 5$. The mean images and difference in 95 and 5 quantiles are shown in figure 5. For small β, the largest variations in the quantiles for both methods occur where the obscured pixels are. By increasing the inverse temperature β, the distribution concentrates more around the mode. In this case, we observe that the largest variation occurs around the image edges, which suggests that the uncertainty as β increases is more around the location of the image edges. The observations of this latter experiment are in line with the observations of [11].

References

[1] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[2] L. R. Bellet. Ergodic Properties of Markov Processes. In S. Attal, A. Joye, and C.-A. Pillet, editors, *Open Quantum Systems II: The Markovian Approach*, pages 1–39. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[3] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection,

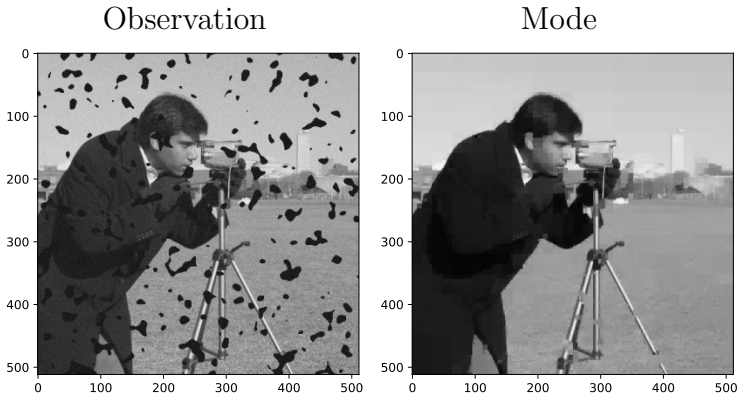


Figure 4: Observation and mode image.

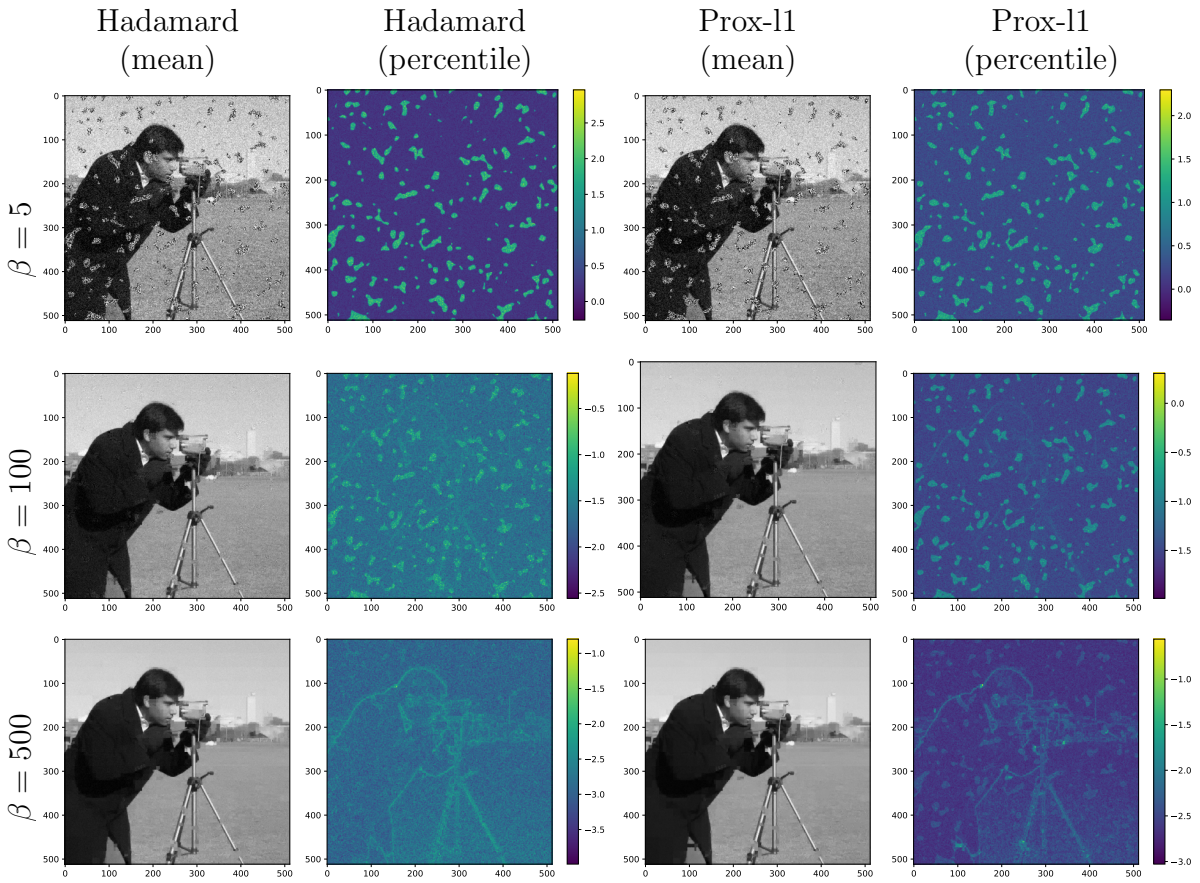


Figure 5: Mean images and log differences between 95 and 5 quantiles for different choices of β.

and robust statistics with applications in early vision. *Int. J. Comput. Vis.*, 19(1):57–91, 1996.

[4] D. Calvetti and E. Somersalo. Computationally efficient sampling methods for sparsity promoting hierarchical Bayesian models. *SIAM/ASA Journal on Uncertainty Quantification*, 12(2):524–548, 2024.

[5] L. Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *Open Journal of Mathematical Optimization*, 3:1–19, 2022.

[6] H.-H. Chou, J. Maly, and H. Rauhut. More is less: inducing sparsity via overparameterization. *Information and Inference*, 12(3):1437–1460, 2023.

[7] I. Daubechies, R. DeVore, M. Fornasier, and S. Gunturk. Iteratively re-weighted least squares minimization: Proof of faster than linear rate for sparse recovery. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 26–29. IEEE, 2008.

[8] V. De Bortoli and A. Durmus. Convergence of diffusions and their discretizations: from continuous to discrete processes and back, 2020.

[9] S. Dereich, A. Neuenkirch, and L. Szpruch. An Euler-type method for the strong approximation of the Cox–Ingersoll–Ross process. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 468(2140):1105–1115, Dec. 2011.

[10] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[11] A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau. *SIAM J. Imaging Sci.*, 11(1):473–506, Jan. 2018.

[12] W. Feller. Two singular diffusion problems. *Annals of Mathematics*, 54(1):173–182, 1951.

[13] L. Fruehwirth and A. Habring. Ergodicity of Langevin dynamics and its discretizations for non-smooth potentials. *arXiv*, 2024.

[14] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Series B Stat. Methodol.*, 73(2):123–214, Mar. 2011.

[15] P. Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer, 2004.

[16] R. D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

[17] M. Hairer and J. C. Mattingly. Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, volume 63 of *Progress in Probability*, pages 109–117. Birkhäuser, Basel, 2011.

[18] M. Hefter and A. Herzwurm. Strong convergence rates for Cox-–Ingersoll-–Ross processes — Full parameter range. *J. Math. Anal. Appl.*, 459(2):1079–1101, Mar. 2018.

[19] D. J. Higham, X. Mao, and A. M. Stuart. Strong convergence of Euler-type methods for nonlinear stochastic differential equations. *SIAM J. Numer. Anal.*, 40(3):1041–1063, Jan. 2002.

[20] P. D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

[21] M. Hutzenthaler, A. Jentzen, and M. Noll. Strong convergence rates and temporal regularity for Cox–Ingersoll–Ross processes and Bessel processes with accessible boundaries. *arXiv [math.NA]*, Mar. 2014.

[22] N. Ikeda and S. Watanabe. A comparison theorem for solutions of stochastic differential equations and its applications. *Osaka Journal of Mathematics*, 14(3):619–633, Dec. 1977.

[23] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer Science & Business Media, 1991.

[24] R. Kumar, C. Carroll, A. Hartikainen, and O. Martin. Arviz a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software*, 4(33):1143, 2019.

[25] H. Lamba, J. C. Mattingly, and A. M. Stuart. An adaptive Euler–Maruyama scheme for SDEs: convergence and stability. *IMA Journal of Numerical Analysis*, 27(3):479–506, July 2007.

[26] T. Lelièvre and G. Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer.*, 25:681–880, May 2016.

[27] X. Mao. *Stochastic Differential Equations and Applications*. Chichester: Horwood, 2 edition, 2008.

[28] P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker–Planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

[29] B. Maslowski and J. Seidler. Probabilistic approach to the strong Feller property. *Probab. Theory Relat. Fields*, 118(2):187–210, Oct. 2000.

[30] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stoch. Process. Their Appl.*, 101(2):185–232, Oct. 2002.

[31] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, Dec. 2012.

[32] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[33] M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Stat. Comput.*, 26(4):745–760, July 2016.

[34] C. Poon and G. Peyré. Smooth bilevel programming for sparse regularization. *Advances in Neural Information Processing Systems*, 34:1543–1555, 2021.

[35] C. Poon and G. Peyré. Smooth over-parameterized solvers for non-smooth structured optimization. *Math. Program.*, 201(1-2):897–952, Sept. 2023.

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

[37] D. W. Stroock and S. R. S. Varadhan. *Multidimensional Diffusion Processes*. Classics in Mathematics. Springer, Berlin, Germany, July 2013.

[38] T. Vaskevicius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

[40] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.

[41] K. S. Zhang, G. Peyré, J. Fadili, and M. Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Conference on Learning Theory*, pages 3814–3841. PMLR, 2020.