MASLA: METROPOLIS-ADJUSTED SUBDIFFERENTIAL LANGEVIN SAMPLING FOR NONSMOOTH NONCONVEX POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sampling from distributions with nonsmooth nonconvex potentials challenges gradient-based Langevin methods and can complicate proximal approaches. We introduce Metropolis-adjusted subdifferential Langevin algorithm (MASLA), which combines a fixed measurable selection from a conservative field with an exact Metropolis–Hastings correction. The method accommodates locally Lipschitz potentials without requiring smoothing or a proximal mapping. We establish well-posedness, reversibility, and exact invariance of the target distribution, together with positive Harris recurrence. Our central theoretical result rigorously proves geometric ergodicity for positive-definite quadratic potentials perturbed by Lipschitz functions admitting bounded conservative fields. The result provides an explicit admissible range of step sizes and holds from every initial point, including points of nondifferentiability. Numerical experiments verify the predicted geometric convergence behavior on a convex composite target and demonstrate competitive performance against proximal and subgradient samplers. Further experiments on a nonsmooth nonconvex target show that the Metropolis correction substantially reduces sampling error relative to unadjusted subdifferential Langevin dynamics. These results establish MASLA as a theoretically grounded sampling method for a class of nonsmooth nonconvex distributions.

1 INTRODUCTION

Sampling from a density $\pi(x) \propto e^{-U(x)}$ is central to Bayesian computation, where posterior expectations and uncertainty quantification require more than a single optimizer of the potential. Langevin methods use local derivative information to guide exploration of the target distribution. When U is differentiable, the Metropolis-adjusted Langevin algorithm (MALA) uses an Euler–Maruyama step of Langevin diffusion as a proposal and applies a Metropolis–Hastings (MH) correction. This combination exploits the geometry of the target while preserving its invariant distribution. The corresponding unadjusted Langevin algorithm (ULA) can exhibit discretization bias and can even lose stability for smooth potentials (Roberts & Tweedie, 1996). Section 2 illustrates this distinction and explains why we retain the MH correction when extending Langevin sampling to nonsmooth problems.

Nonsmooth and nonconvex sampling. Nonsmooth potentials arise naturally from sparsity-inducing priors, composite regularizers, and models involving nonsmooth activations. For example, the Bayesian analogue of an ℓ_1 penalty is a Laplace prior, whose negative log-density is nondifferentiable at zero (Park & Casella, 2008; Seeger, 2008). More general model structures and penalties can also introduce nonconvexity. These features complicate classical gradient-based sampling: a gradient may not exist at every state, and tools relying on convexity may no longer apply. A useful extension should accommodate generalized derivatives, specify the update at exceptional points, and retain a rigorous connection between the implemented Markov chain and the intended target.

Related approaches. Proximal and Moreau–Yosida methods address nonsmooth sampling through implicit updates or smoothing (Pereyra, 2016; Bernton, 2018; Durmus et al., 2018; 2019); proximal MH methods also support composite posterior applications (Abry et al., 2023). Subgra-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dient methods avoid some proximal computations for convex composite potentials (Habring et al., 2024a), and Lehec (2023) gives nonasymptotic bounds for Lipschitz convex potentials without Lipschitz gradients. Lau et al. (2024) compares Langevin algorithms for nonsmooth non-log-concave targets. Inexact proximal algorithms quantify errors from approximate subproblem solves (Ehrhardt et al., 2024; Ghaderi et al., 2026); absence of a closed-form prox therefore does not make proximal MCMC inapplicable. Other approaches change the driving noise (Oechsler, 2024) or use a Hadamard parameterization of an ℓ_1 prior (Cheltsov et al., 2026). The latter establishes geometric ergodicity for continuous dynamics and convergence of a discretization as its step size vanishes. Crucinio et al. (2025) studies dimension scaling for Moreau–Yosida MH proposals; such scaling results are distinct from fixed-dimensional geometric ergodicity. Srinivasan et al. (2024) gives mixing bounds for MH-corrected mirror Langevin sampling on compact convex domains, a different geometry from our unconstrained nonsmooth potentials.

Our approach and contributions. To address nonsmooth and nonconvex targets, we introduce the Metropolis-adjusted subdifferential Langevin algorithm (MASLA), built on the framework of conservative fields. Conservative fields provide a notion of generalized derivatives that accommodates a broad class of nonsmooth nonconvex functions while retaining a chain rule (Bolte & Pauwels, 2021). MASLA uses a fixed measurable selection from a conservative field to construct proposals and an exact MH correction to preserve the target distribution, without requiring a proximal mapping or smoothing parameter. We establish its discrete-time validity directly, independently of existence results for continuous-time stochastic differential inclusions (Difonzo et al., 2022). Our contributions are threefold:

- **Exact sampling with conservative-field proposals.** We specify the MASLA kernel and establish exact invariance, reversibility, and positive Harris recurrence for every positive step size. We also characterize when the chain law is independent of subgradient choices.
- **Rigorous geometric ergodicity guarantees.** We prove geometric ergodicity for a positive-definite quadratic potential plus a Lipschitz perturbation with a bounded conservative field, with an explicit step-size range and convergence from every initial point, including points of nondifferentiability.
- **Numerical verification and sampling accuracy.** Experiments verify the predicted geometric convergence behavior on a convex composite target, compare MASLA with proximal and subgradient samplers, and demonstrate the accuracy gained from MH correction on a nonsmooth nonconvex target.

2 Motivation: Why Retain the Metropolis Correction?

For a smooth potential, Langevin diffusion and its Euler discretization take the forms

$$dX_t = -\nabla U(X_t)\,dt + \sqrt{2}\,dB_t, \qquad Y_{n+1} = Y_n - \gamma \nabla U(Y_n) + \sqrt{2\gamma}\,Z_{n+1}. \tag{2.1}$$

Here $(X_t)_{t\geq 0}$ is the continuous-time process driven by a standard d-dimensional Brownian motion $(B_t)_{t\geq 0}$, and $(Y_n)_{n\geq 0}$ is the ULA chain with fixed step size $\gamma > 0$. The noise vectors $(Z_n)_{n\geq 1}$ are independent and identically distributed (i.i.d.) as $N(0, I_d)$, independently of Y_0, where I_d is the $d \times d$-dimensional identity matrix. Even when the diffusion is well behaved, the discrete chain need not inherit its stability or invariant law (Roberts & Tweedie, 1996). We use a scaled version of the Gaussian example discussed in Sections 1.4.1 and 3 of that work (their step size h equals 2γ in our convention). For $\pi = N(0, \sigma^2)$, the potential is $U(x) = x^2/(2\sigma^2)$ and ULA becomes

$$Y_{n+1} = \left(1 - \frac{\gamma}{\sigma^2}\right) Y_n + \sqrt{2\gamma}\,Z_{n+1}. \tag{2.2}$$

This autoregression is stable for $0 < \gamma/\sigma^2 < 2$, with stationary variance $\sigma^2/(1 - \gamma/(2\sigma^2))$, and is explosive when $\gamma/\sigma^2 > 2$. We set $\sigma = 0.02$ and $\gamma = 0.001$, giving coefficient -1.5. Thus, although the absolute step is small, it exceeds the stability limit $2\sigma^2 = 0.0008$. Reducing γ below this limit restores ULA stability for this fixed target, albeit with discretization bias.

Figure 1 shows that MALA remains in the central target region and its histogram closely matches the Gaussian density, whereas all 16 ULA runs cross the numerical escape threshold within 31–41 updates. MALA has an average acceptance rate of 39.6% and a pooled sample variance of 1.0019

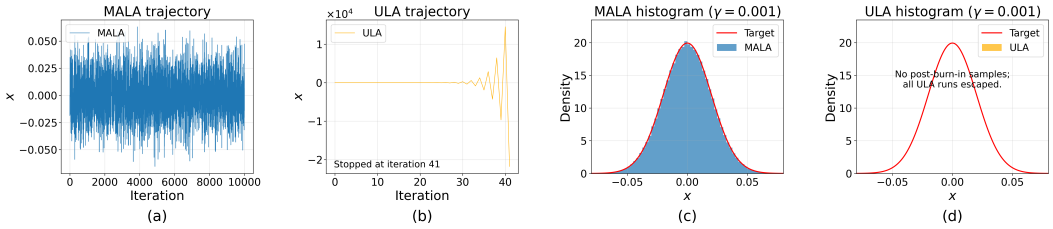

Figure 1: MALA and ULA for $\pi = N(0, 0.02^2)$ with $\gamma = 0.001$, both initialized at zero. (a)–(b) The first replicate's trajectories: MALA is shown for 10,000 updates, and ULA stops at iteration 41 upon crossing $|x|/\sigma \geq 10^6$; the axis scales differ. (c)–(d) Histograms pool the retained observations from 16 replicates, each with a nominal length of 100,000 updates and 20,000 burn-in updates; the target density is red. All ULA replicates stop before burn-in ends, so panel (d) has no histogram bars.

for X/σ, whose target variance is one. Its MH correction preserves π exactly by accounting for the target density and the forward–reverse proposal ratio. The experiment illustrates the benefit of this correction when the Euler proposal is unstable; finite-run agreement does not by itself establish a geometric convergence rate. Appendix A.1.1 gives the simulation details and the analytical basis for ULA transience.

3 Conservative Fields and the MASLA Algorithm

Throughout, the state space is $\mathbb{R}^d$, λ is Lebesgue measure, and $U : \mathbb{R}^d \to \mathbb{R}$ is finite-valued and Borel measurable with

$$0 < Z_U := \int e^{-U(x)}\, \lambda(dx) < \infty, \qquad \pi(x) = Z_U^{-1} e^{-U(x)}. \tag{3.1}$$

For the conservative-field formulation, U is locally Lipschitz and admits a conservative field $D_U : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. The double arrow assigns a set of vectors $D_U(x) \subseteq \mathbb{R}^d$ to each x, whereas $\to$ denotes a single-valued map. We fix one finite-valued Borel measurable selection $\beta : \mathbb{R}^d \to \mathbb{R}^d$ with $\beta(x) \in D_U(x)$ at every point. A locally bounded conservative field has closed graph, nonempty compact values, and the chain rule along absolutely continuous curves (Appendix B.1). Convex and locally Lipschitz semialgebraic functions are examples of potentials admitting such fields.

MASLA is formulated based on a generalized, set-valued version of the Langevin diffusion:

$$dX_t \in -D_U(X_t)dt + \sqrt{2}\, dB_t. \tag{3.2}$$

Unlike the classical MALA algorithm, where proposals are generated using a single gradient as in (2.1), MASLA draws a subgradient from the set-valued field $D_U(x)$ at each iteration:

$$\breve{Y}_{n+1} = X_n - \gamma\beta(X_n) + \sqrt{2\gamma}\, Z_{n+1}, \qquad \beta(X_n) \in D_U(X_n), \tag{3.3}$$

which is then followed by the standard MH accept-reject step. Let $\mathcal{D}_f$ denote the set of points $x \in \mathbb{R}^d$ where f is differentiable. A special case of (3.3) is the following:

$$\breve{Y}_{n+1} = X_n - \gamma\beta(X_n) + (2\gamma)^{1/2} Z_{n+1}, \tag{3.4}$$

where

$$\begin{cases} \beta(x) = \nabla U(x) & x \in \mathcal{D}_U, \\ \beta(x) \in D_U(x) & \text{otherwise.} \end{cases} \tag{3.5}$$

Remark 3.1 and Example 3.2 explain why (3.4) is special case instead of an equivalent form.

Remark 3.1 *The inclusion $\beta(x) \in D_U(x)$ holds for all $x \in \mathbb{R}^d$, because $\nabla U(x)$ is an element of $D_U(x)$ when the former exists. However, conversely, a function $\beta(x) \in D_U(x)$ does not necessarily satisfy $\beta(x) = \nabla U(x)$ if $x \in \mathcal{D}_U$. The example below shows that differentiability alone does not imply regularity.*

Example 3.2 *Let $g : \mathbb{R} \to \mathbb{R}$ be defined by*

$$g(x) := \begin{cases} x^2 \sin(1/x), & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Then g is locally Lipschitz on $\mathbb{R}$ and differentiable at $x = 0$ with

$$\nabla g(0) = 0.$$

However, its Clarke subdifferential at 0 satisfies

$$\partial g(0) = [-1, 1].$$

The proofs to Example 3.2 and Proposition 3.3 are provided in Appendix B.2.

Proposition 3.3 (Well-posedness and exact invariance) *Let $U : \mathbb{R}^d \to \mathbb{R}$ be finite-valued and Borel measurable and satisfy*

$$0 < Z_U := \int e^{-U(x)} \, \lambda(dx) < \infty, \qquad \pi(x) = Z_U^{-1} e^{-U(x)}. \tag{3.6}$$

Fix a finite-valued Borel map $\beta : \mathbb{R}^d \to \mathbb{R}^d$ (in MASLA, a selection of D_U). For every $\gamma > 0$, define

$$q_\gamma(x, y) = (4\pi\gamma)^{-d/2} \exp\left(-\frac{\|y - x + \gamma\beta(x)\|^2}{4\gamma} \right), \tag{3.7}$$

$$\alpha_\gamma(x, y) = 1 \wedge \frac{\pi(y) q_\gamma(y, x)}{\pi(x) q_\gamma(x, y)}, \qquad r_\gamma(x) = 1 - \int q_\gamma(x, y) \alpha_\gamma(x, y) \, \lambda(dy). \tag{3.8}$$

Then the Markov transition kernel

$$R_\gamma(x, A) = \int_A q_\gamma(x, y) \alpha_\gamma(x, y) \, \lambda(dy) + r_\gamma(x) \delta_x(A), \quad A \in \mathcal{B}(\mathbb{R}^d), \tag{3.9}$$

is well-defined from every initial law, reversible with respect to π, and admits π as an invariant probability distribution.

Explicit MASLA pseudocode is given in Algorithm C.1 of Appendix C; Section C.2 discusses choosing β by automatic differentiation.

4 Convergence Guarantees

4.1 Selection Independence and Recurrence

A conservative field can contain more than the classical gradient at an exceptional differentiability point. Thus, almost-everywhere agreement of fields does not justify arbitrary pointwise substitutions. The following result makes the required choice explicit.

Proposition 4.1 (Selection independence) *Suppose U is locally Lipschitz and satisfies (3.6), and D_U is a conservative field for U. Let N_U be its nondifferentiability set. Consider finite-valued Borel selections β_1, β_2 of D_U with*

$$\beta_1(x) = \beta_2(x) = \nabla U(x), \qquad x \notin N_U,$$

with possibly different values on N_U. For $i \in \{1, 2\}$ and $\gamma > 0$, let $R_{\gamma,i}$ be the MASLA kernel constructed from β_i as in (3.7)–(3.9). Then, for every $\gamma > 0$,

$$R_{\gamma,1}(x, \cdot) = R_{\gamma,2}(x, \cdot), \qquad x \notin N_U. \tag{4.1}$$

In particular, the kernels agree for λ-almost every x and for π-almost every x.

If both chains have the same initial law μ with $\mu(N_U) = 0$, then their laws on the sequence space $(\mathbb{R}^d)^{\{0,1,\dots\}}$ coincide, and each chain satisfies

$$\mathbb{P}_\mu\big(X_n \notin N_U \text{ for every } n \geq 0\big) = 1.$$

This applies, in particular, to every absolutely continuous initial law and to every deterministic initial point outside N_U. Moreover, an absolutely continuous initial law remains absolutely continuous at every iteration under either kernel.

The restriction on the initial law matters: from a nonsmooth initial point, different selections can change both the proposal and the rejection probability. Exact invariance still holds for each fixed selection. The next result does not require the initial state to avoid the nonsmooth set.

Proposition 4.2 (Irreducibility and Harris recurrence) *Under conditions of Proposition 3.3, including normalizability and a finite-valued Borel map β, the MASLA kernel R_γ is λ-irreducible and π-irreducible for every $\gamma > 0$. More precisely, $R_\gamma(x, A) > 0$ for every $x \in \mathbb{R}^d$ and every Borel set A with $\lambda(A) > 0$. Furthermore, the chain is positive Harris recurrent: for every such set A and every initial point x,*

$$\mathbb{P}_x\big(X_n \in A \text{ for infinitely many } n\big) = 1,$$

and it admits an invariant probability distribution. Its unique invariant probability distribution is π.

Proofs to Propositions 4.1 and 4.2 are provided in Appendix B.3. When U and β are bounded on compact sets, every compact set of positive Lebesgue measure is a one-step small set and the chain is strongly aperiodic (Proposition B.17). These properties support the geometric drift argument below.

4.2 Geometric ergodicity for a nonconvex class

Exact invariance and recurrence alone do not establish a geometric convergence rate. A drift inequality must hold for the full accepted chain, including its rejection term. Theorem B.15 in the appendix states the general drift and minorization criterion. Here we give conditions directly on the potential and field that imply that criterion.

Theorem 4.3 (MASLA geometric ergodicity) *Let $H \in \mathbb{R}^{d\times d}$ be symmetric and positive definite. Write h_- and h_+ for its smallest and largest eigenvalues. Suppose that*

$$U(x) = \tfrac{1}{2}x^\top Hx + g(x), \qquad |g(x) - g(y)| \le L\|x - y\|, \qquad x, y \in \mathbb{R}^d,$$

for some $L < \infty$. Assume that g admits a conservative field D_g whose values are uniformly bounded: for some $B < \infty$,

$$\sup_{x\in\mathbb{R}^d}\sup_{v\in D_g(x)} \|v\| \le B.$$

Choose any fixed Borel measurable selection $v(x) \in D_g(x)$ and set

$$D_U(x) := Hx + D_g(x), \qquad \beta(x) := Hx + v(x).$$

For every step size

$$0 < \gamma < \frac{2}{h_+}, \tag{4.2}$$

the MASLA kernel using this selection and the exact MH acceptance probability is V-uniformly geometrically ergodic with $V(x) = 1 + \|x\|^2$. Specifically, $\pi(dx) \propto e^{-U(x)}\lambda(dx)$ is well-defined, and there exist $M_\gamma < \infty$ and $\rho_\gamma \in (0, 1)$ such that

$$\|R_\gamma^n(x, \cdot) - \pi\|_V \le M_\gamma(1 + \|x\|^2)\rho_\gamma^n, \qquad x \in \mathbb{R}^d, \quad n \ge 0.$$

This includes initial points at which U is not differentiable. More precisely, writing $a_\gamma := \|I_d - \gamma H\|_{\mathrm{op}} < 1$, the rejection probability and the drift satisfy

$$\lim_{\|x\|\to\infty} r_\gamma(x) = 0, \qquad \limsup_{\|x\|\to\infty} \frac{R_\gamma V(x)}{V(x)} \le a_\gamma^2 < 1. \tag{4.3}$$

Here $\|\xi\|_V = \sup_{|f|\le V} |\int f\, d\xi|$, with the supremum over Borel functions. The constants may depend on γ; the theorem does not assert a rate uniform over all step sizes or dimensions.

Proof outline. The quadratic lower bound makes e^{-U} integrable. Write the proposal as $Y = (I_d - \gamma H)x - \gamma v(x) + \sqrt{2\gamma}Z$. Expanding the log MH ratio gives

$$\ell_\gamma(x, y) = \frac{\gamma}{4}\big(\|Hx\|^2 - \|Hy\|^2\big) + E_\gamma(x, y), \qquad |E_\gamma(x, y)| \le K_\gamma(\|x\| + \|y\|) + \gamma B^2/4.$$

On every bounded-noise event, the positive quadratic term dominates the linear error as $\|x\| \to \infty$. Hence $r_\gamma(x) \to 0$. Retaining the rejection contribution gives

$$R_\gamma V(x) \le \mathbb{E}[V(Y)] + r_\gamma(x)V(x), \qquad \limsup_{\|x\|\to\infty} \frac{R_\gamma V(x)}{V(x)} \le a_\gamma^2 < 1.$$

A compact-set minorization completes the argument. Appendix B.3 supplies every estimate.

A nonsmooth nonconvex example. For $\tau > 0$, take $U(x) = \frac{1}{2}x^\top Hx - \tau\|x\|_1$. The perturbation has Lipschitz constant $\tau\sqrt{d}$ and admits the bounded conservative field

$$D_g(x) = \prod_{j=1}^{d} G_j(x_j), \qquad G_j(t) = \begin{cases} \{-\tau\}, & t > 0, \\ [-\tau, \tau], & t = 0, \\ \{\tau\}, & t < 0. \end{cases}$$

The potential is nonsmooth on coordinate hyperplanes and nonconvex near the origin, yet the theorem applies to every Borel selection and every $0 < \gamma < 2/h_+$. Example B.18 gives the verification. Thus the rate theorem covers a concrete nonconvex class, while the basic kernel construction applies more broadly.

5 Numerical Experiments

We evaluate MASLA on a convex benchmark against established samplers (Section 5.1), a one-dimensional nonconvex target to assess MH correction (Section 5.2), and a high-dimensional quartic target to examine numerical stability (Section 5.3).

5.1 Comparison where competing methods apply

Target and experimental protocol. To ensure a fair comparison, we adopt the exact experimental setup in Habring et al. (2024a) on sampling from a two-dimensional density with the potential

$$U(x) = F(x) + G(Kx),$$

where $F : \mathbb{R}^d \to [0, \infty)$ and $G : \mathbb{R}^{d'} \to [0, \infty)$ are both proper, convex, and lower semi-continuous, and $K : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a linear operator. They considered function F as either Lipschitz continuous with constant L_F or differentiable with a Lipschitz continuous gradient with constant $L_{\nabla F}$, while G as convex and Lipschitz continuous with constant L_G. They used the total variation (TV) functional for $G(Kx)$, defined in the two-dimensional case as $G : \mathbb{R} \to \mathbb{R}$ and $K : \mathbb{R}^2 \to \mathbb{R}$ by

$$G(p) = \lambda|p|, \qquad Kx = x_2 - x_1,$$

where $\lambda > 0$ is a scaling parameter. It is straightforward to verify that G is Lipschitz continuous with Lipschitz constant $L_G = \lambda$, and its subdifferential is given by $\partial G(p) = \{\lambda \operatorname{sign}(p)\}$ for $p \neq 0$ and $[-\lambda, \lambda]$ at zero. The operator K satisfies $\|K\|^2 \leq 2$. For $F : \mathbb{R}^2 \to \mathbb{R}$, they used the squared L^2 discrepancy, defined as

$$F(x) = \frac{1}{2\sigma^2}\|x - y\|_2^2,$$

for y being a known vector, which is m-strongly convex with $m = \sigma^{-2}$ and has a Lipschitz continuous gradient with constant $L_{\nabla F} = \sigma^{-2}$. Following Habring et al. (2024a), we set $y = (-1, 1)$, $\sigma = 1$, and $\lambda = 5$.

For each algorithm, 10,000 independent chains start at $(-1, 1)$; histograms across chains estimate the distribution at a fixed iterate. These comparisons use no burn-in removal and test step sizes $\tau \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. The first-coordinate distribution is initially concentrated near -1 and approaches the target marginal after 49,500 iterations; the initial and final density plots are given in Appendix D. These are distributions across independent chains at a fixed iteration, rather than time-averaged samples from a single trajectory.

Wasserstein comparison. Habring et al. (2024a) proposed two algorithms: the Proximal-subgradient Langevin algorithm (Prox-sub) and the Gradient-subgradient Langevin algorithm (Grad-sub). In Figure 1 of therein, the authors investigated the convergence of Grad-sub and Prox-sub in Wasserstein-2 distance $W_2(\mu, \nu) = \left(\inf_{\zeta \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, d\zeta(x, y)\right)^{1/2}$, where $\Pi(\mu, \nu)$ is the set of couplings of two probability measures μ, ν. Appendix D.1.4 details the discretized reference distribution and optimal-transport calculation of the plotted squared Wasserstein errors. Figure 2 shows that numerical convergence rates exceed theoretical expectations (explained in Appendix D.1.2) for all τ values. Experimental results indicate negligible differences in convergence rates among Prox-sub, Grad-sub, and MASLA. All algorithms exhibit exponential ergodicity to the target density.

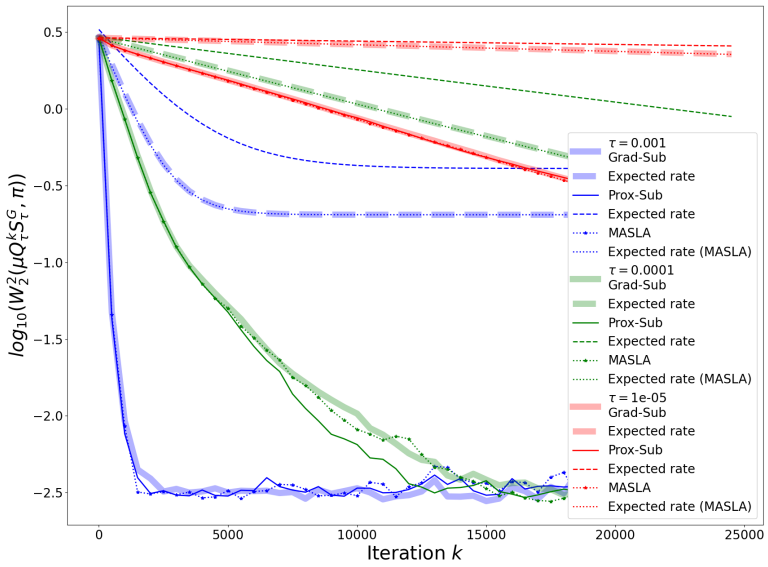

Figure 2: Comparison of MASLA, Grad-sub (Habring et al., 2024a), and Prox-sub (Habring et al., 2024a) for TV $-L^2$ sampling in Wasserstein distance. Colors denote different τ values, dashed lines represent theoretical convergence rates, and solid lines show empirical results. Thick transparent lines indicate Grad-sub, thin solid lines indicate Prox-sub, and dotted lines indicate MASLA.

Total variation comparison. In Figure 3 of Habring et al. (2024a), the authors compared Prox-sub with the Proximal Metropolis-Adjusted Langevin Algorithm (P-MALA) of Pereyra (2016) and the Moreau–Yosida Unadjusted Langevin Algorithm (MYULA) of Durmus et al. (2022) in terms of total variation (TV) distance. The TV distance between two probability measures μ and ν is defined as $\mathrm{TV}(\mu,\nu) = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|$. Appendix D.1.4 describes the histogram-based calculation and clarifies that the plotted TV norm is twice this distance. Following their comparison, we report the better-performing Prox-sub measure described in Appendix D. Figure 3 compares MASLA with Prox-sub, P-MALA, and MYULA for TV-L^2 sampling in TV distance. As shown in Figure 2, all tested algorithms exhibit empirical exponential convergence to the target distribution across the considered τ values.

We emphasize that, while exponential ergodicity is observed empirically for these methods in the these literature, we establish it rigorously for MASLA in this paper.

5.2 Evaluation beyond the competing methods' convex setting

Why this target extends the comparison. The first experiment establishes performance in a shared applicability regime. We now examine a target that violates the convexity assumptions used by those baselines and illustrates the difficulty of requiring a single-valued proximal mapping. Consider $U(x) = |x^2 - 1|$. This function is locally Lipschitz, since on a bounded interval $|U(x) - U(y)| \le |x - y|\,|x + y|$. It has the explicit Clarke subgradient

$$\partial^c U(x) = \begin{cases} \{2x\}, & |x| > 1, \\ [-2, 2], & x = \pm 1, \\ \{-2x\}, & |x| < 1. \end{cases}$$

The potential is nonconvex because $U(0) = 1 > [U(-1) + U(1)]/2 = 0$. At proximal parameter one and input zero, the minimization objective is

$$\phi(x) = |x^2 - 1| + x^2/2 = \begin{cases} 1 - x^2/2, & |x| \le 1, \\ 3x^2/2 - 1, & |x| \ge 1. \end{cases}$$

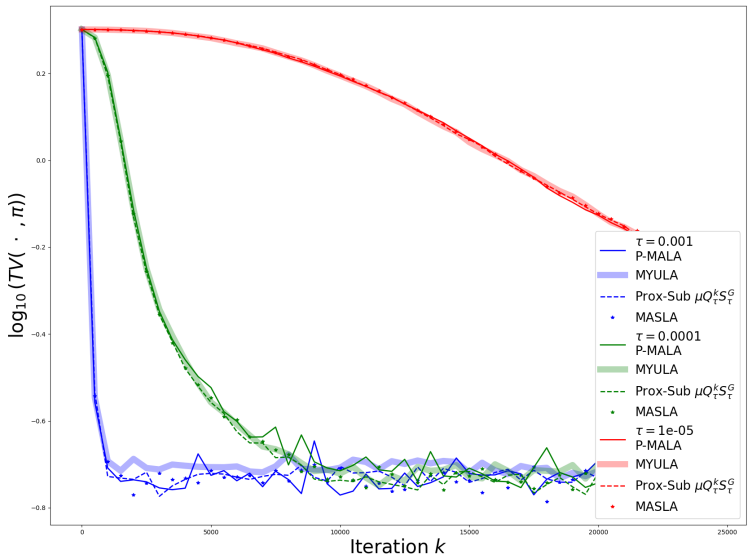

Figure 3: Comparison of MASLA, Prox-sub (Habring et al., 2024a), P-MALA (Pereyra, 2016), and MYULA (Durmus et al., 2022) for TV $-L^2$ sampling in TV distance with $\sigma = 1$, $\lambda = 5$, and step sizes $\tau \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. Colors represent different τ values. Thick transparent lines denote MYULA results, thin solid lines denote P-MALA, dashed lines denote Prox-sub, and dotted lines denote MASLA.

Both -1 and 1 minimize ϕ, yielding $\mathrm{prox}_U(0) = \{-1, 1\}$. Thus, the convex baseline guarantees do not directly apply, and a proximal construction requiring uniqueness at every parameter needs additional assumptions or a selection rule. While this does not preclude other nonconvex adaptations of proximal methods, MASLA sidesteps this difficulty entirely: by operating directly on the displayed Clarke field, it requires neither solving a proximal subproblem nor imposing conditions to ensure uniqueness of its solution, making it directly applicable in this nonconvex setting.

Isolating the Metropolis correction. The multi-valued nature of the proximal operator hinders the application of proximal-based methods, however we can still compare with the algorithmic version without the MH correction to numerically visualize the strength of the MH correction. Difonzo et al. (2022) focused on the theoretical analysis of stochastic Langevin differential inclusions of continuous-time while they ran unadjusted Langevin dynamics with the Euler-Maruyama discretization. Here, we refer it as Unadjusted Subdifferential Langevin Algorithm (USLA), which simply draws a subgradient from the set-valued field at each iteration. Both chains start at zero and run for 100,000 iterations with step size 0.1; the first 20% are discarded as burn-in. Figure 4 compares the retained samples with the true density. The reported distance errors and running times are:

Algorithm	TV error	W_2 error	Time (seconds)
USLA	0.116761	0.092183	0.9976
MASLA	0.014363	0.008199	1.5361

Timing was measured on an Apple M1 Pro laptop with 10 cores and 16 GB of memory. In this run, MASLA reduces TV error by about 8.1 times and Wasserstein error by about 11.2 times, at about 1.54 times the runtime of USLA. The comparison therefore quantifies the accuracy benefit and computational cost of correction beyond the convex benchmark.

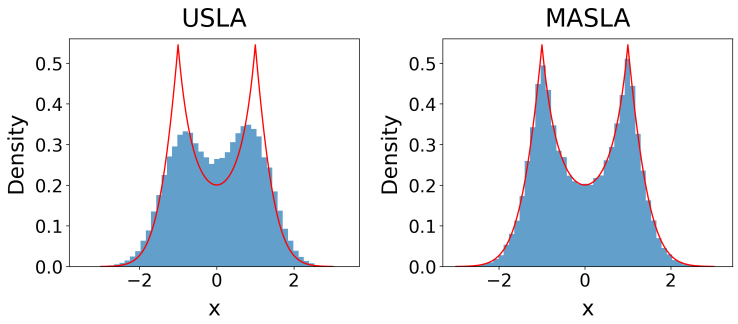

Figure 4: USLA and MASLA samples compared with the true density. Both chains start at zero and run for 100,000 iterations with step size $\tau = 0.1$; the first 20% of iterations are discarded as burn-in.

5.3 High-dimensional sampling with a quartic potential

To complement the preceding one- and two-dimensional examples, we consider the 100-dimensional nonsmooth nonconvex target

$$\pi_{100}(x) \propto \exp\left[-\sum_{i=1}^{100}\left(\frac{x_i^4}{4} - x_i^2 + |x_i|\right)\right].$$

The quartic terms ensure normalizability, the absolute-value terms introduce nonsmooth coordinate hyperplanes, and the negative quadratic terms make the potential nonconvex. We use the selection $\beta_i(x) = x_i^3 - 2x_i + \mathrm{sign}(x_i)$, with $\mathrm{sign}(0) = 0$. Although the target factorizes, proposals update all 100 coordinates jointly, and MASLA applies a single MH accept–reject decision to the full vector. Appendix D.3 gives the experimental details and Figure 6 reports the results.

For each step size $\gamma \in \{0.05, 0.10, 0.20\}$, we run 16 independent replicates of each algorithm from $0.1\mathbf{1}_{100}$ for up to 10,000 iterations. We record numerical escape of USLA when $\|X_n\|_\infty \geq 10^6$ and stop that replicate after the crossing. No USLA replicate crosses the threshold at $\gamma = 0.05$ or 0.10 during the observation period; at $\gamma = 0.20$, all 16 cross it within 336 iterations. MASLA remains below the threshold in every run, with mean acceptance rates of 58.03%, 21.49%, and 1.37%, respectively. After discarding 2,000 iterations, its first-coordinate histograms agree reasonably with the target marginal at the smaller step sizes, whereas the low acceptance rate at $\gamma = 0.20$ leads to poor exploration over the available run length. This experiment illustrates both the role of the MH correction in controlling explosive excursions from a superlinear drift and the importance of step-size choice in high dimension. The observed threshold crossings provide numerical evidence of USLA instability.

6 Conclusion

We introduced MASLA for sampling from distributions with locally Lipschitz potentials that may be nonsmooth and nonconvex. Combining a fixed measurable conservative-field selection with an exact Metropolis–Hastings correction, MASLA preserves the target distribution without requiring smoothing or a proximal mapping. We established reversibility and positive Harris recurrence, characterized selection independence, and rigorously proved geometric ergodicity for positive-definite quadratic potentials with Lipschitz perturbations admitting bounded conservative fields. Numerical experiments verified the predicted geometric convergence behavior on a convex composite target and demonstrated competitive performance where existing proximal and subgradient methods apply. On a nonsmooth nonconvex target beyond their convex framework, MASLA substantially reduced sampling errors relative to unadjusted subdifferential Langevin dynamics.

AI USE STATEMENT

Generative AI tools assisted with coding and implementation, discussions of experimental design, manuscript editing, review of mathematical exposition, LaTeX preparation, and consolidation of bibliography entries.

REPRODUCIBILITY STATEMENT

The MASLA proposal, acceptance probability, and Markov kernel are specified in Section 3. Section 4 states the convergence guarantees and their assumptions, and Appendix B provides the mathematical background and full proofs. Section 5 describes the experimental targets, step sizes, initialization, chain lengths, burn-in choices, and reported error metrics; Appendix D gives additional target definitions, baseline update formulas, and experimental details. The baseline source code and the package used for Wasserstein-distance computations are cited in the experimental descriptions.

The supplementary material includes Python scripts, with a README describing their usage, dependencies, and required inputs. These scripts implement the numerical experiments reported in the paper; the experimental settings and error calculations are detailed in Appendix D.

REFERENCES

Patrice Abry, Gersende Fort, Barbara Pascal, and Nelly Pustelnik. Proximal-Langevin samplers for nonsmooth composite posteriors: Application to the estimation of COVID19 reproduction number. In *European Signal Processing Conference (EUSIPCO)*, pp. 1813–1817, 2023. URL https://eurasip.org/Proceedings/Eusipco/Eusipco2023/pdfs/0001813.pdf.

Robert Baier and Elza Farkhi. Regularity of set-valued maps and their selections through set differences. Part 1: Lipschitz continuity. *Serdica Mathematical Journal*, 39:365–390, 01 2013.

E. Bernton. Langevin Monte Carlo and JKO splitting. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1777–1798. PMLR, 2018.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188:19–51, 2021.

C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

L. Chaari, J.-Y. Tourneret, C. Chaux, and H. Batatia. A Hamiltonian Monte Carlo method for non-smooth energy sampling. *IEEE Transactions on Signal Processing*, 64(21):5585–5594, 2016.

Ivan Cheltsov, Federico Cornalba, Clarice Poon, and Tony Shardlow. Hadamard Langevin dynamics for sampling the ℓ_1-prior, 2026. URL https://arxiv.org/abs/2411.11403v4. Version 4.

Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Michel Coste. *An Introduction to o-minimal geometry*. RAAG notes. Institut de Recherche Mathématique de Rennes, 1999.

Francesca R. Crucinio, Alain Durmus, Pablo Jiménez, and Gareth O. Roberts. Optimal scaling results for Moreau–Yosida Metropolis-adjusted Langevin algorithms. *Bernoulli*, 31(3):1889–1907, 2025. doi: 10.3150/24-BEJ1790.

A. Dalalyan, E. Grappin, and Q. Paris. On the exponentially weighted aggregate with the Laplace prior. *The Annals of Statistics*, 46(5):2452–2478, 2018.

Wei Deng, Qi Feng, Liyao Gao, Faming Liang, and Guang Lin. Non-convex learning via replica exchange stochastic gradient MCMC. In *International Conference on Machine Learning*, pp. 2474–2483. PMLR, 2020.

Fabio V Difonzo, Vyacheslav Kungurtsev, and Jakub Marecek. Stochastic Langevin differential inclusions with applications to machine learning. *arXiv preprint arXiv:2206.11533*, 2022.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*, volume 4. Springer, 2018.

Alain Durmus, Éric Moulines, and Marcelo Pereyra. Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Alain Durmus, Éric Moulines, and Marcelo Pereyra. A proximal Markov chain Monte Carlo method for Bayesian inference in imaging inverse problems: When Langevin meets Moreau. *SIAM Review*, 64(4):991–1028, 2022.

Matthias J. Ehrhardt, Lorenz Kuger, and Carola-Bibiane Schönlieb. Proximal Langevin sampling with inexact proximal mapping. *SIAM Journal on Imaging Sciences*, 17(3):1729–1760, 2024. doi: 10.1137/23M1593565.

L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Chapman and Hall/CRC, London, revised edition, 2015.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, and Laetitia Chapel. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22:1–8, 2021.

Susan Ghaderi, Alireza Kabgani, Yves Moreau, and Masoud Ahookhosh. Speeding up nonsmooth Bayesian MCMC sampling via inexact proximal unadjusted Langevin algorithm, 2026. URL https://arxiv.org/abs/2605.17306.

Andreas Habring, Martin Holler, and Thomas Pock. Subgradient Langevin methods for sampling from nonsmooth potentials. *SIAM Journal on Mathematics of Data Science*, 6(4):897–925, 2024a.

Andreas Habring, Martin Holler, and Thomas Pock. Source Code to Reproduce the Results of "Subgradient Langevin Methods for Sampling from Nonsmooth Potentials". https://github.com/habring/subgradient_langevin, 2024b.

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–363, 2017.

Tim Tsz-Kit Lau, Han Liu, and Thomas Pock. Non-log-concave and nonsmooth sampling via langevin monte carlo algorithms. In Alessandro Benfenati, Federica Porta, Tatiana Alessandra Bubba, and Marco Viola (eds.), *Advanced Techniques in Optimization for Machine Learning and Imaging*, pp. 83–149, Singapore, 2024. Springer Nature Singapore.

Joseph Lehec. The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *The Annals of Applied Probability*, 33(6A):4858 – 4874, 2023. doi: 10.1214/23-AAP1935. URL https://doi.org/10.1214/23-AAP1935.

Gunther Leobacher and Michaela Szölgyenyi. A strong order $1/2$ method for multidimensional SDEs with discontinuous drift. *The Annals of Applied Probability*, 27(4):2383–2418, 2017.

Ning Ning. Convergence of Dirichlet forms for MCMC optimal scaling with dependent target distributions on large graphs. *The Annals of Applied Probability*, 35(2):898–935, 2025.

David Oechsler. Lévy Langevin Monte Carlo. *Statistics and Computing*, 34:37, 2024. doi: 10.1007/s11222-023-10345-w.

R. B. O'Hara and M. J. Sillanpää. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117, 2009.

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

M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4): 745–760, July 2016.

Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1): 255–268, 1998.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341 – 363, 1996.

Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, June 2008.

Vishwak Srinivasan, Andre Wibisono, and Ashia Wilson. Fast sampling from constrained spaces using the Metropolis-adjusted mirror Langevin algorithm. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4593–4635, 2024. URL https://proceedings.mlr.press/v247/srinivasan24a.html.

Lou Van den Dries. *Tame topology and o-minimal structures*, volume 248. Cambridge University Press, 1998.

CONTENTS

1 Introduction **1**

2 Motivation: why retain the Metropolis correction? **2**

3 Conservative fields and the MASLA algorithm **3**

4 Convergence guarantees **4**

4.1 Selection independence and recurrence 4

4.2 Geometric ergodicity for a nonconvex class 5

5 Numerical experiments **6**

5.1 Comparison where competing methods apply 6

5.2 Evaluation beyond the competing methods' convex setting 7

5.3 High-dimensional sampling with a quartic potential 9

6 Conclusion **9**

AI use statement **10**

Reproducibility statement **10**

References **11**

A Additional background and motivation **15**

A.1 Metropolis-Adjusted Langevin Algorithm 15

A.2 Metropolis-Adjusted Subdifferential Langevin Algorithm 16

B Mathematical background and full proofs **18**

B.1 Conservative fields and generalized derivatives 18

B.2 Wellposedness of MASLA 20

B.3 Convergence of MASLA 22

C MASLA implementation and automatic differentiation **31**

C.1 Explicit MASLA pseudocode 31

C.2 Choosing the selection with automatic differentiation 31

D Additional experimental details **33**

D.1 MASLA Versus State-of-the-Art Algorithms 33

D.2 MASLA Performance in Scenarios Beyond Competing Methods 37

D.3 A high-dimensional nonsmooth nonconvex quartic target 38

A ADDITIONAL BACKGROUND AND MOTIVATION

This appendix reviews the classical MALA construction and its motivation in Section A.1, followed by further motivation for the development of MASLA in Section A.2.

A.1 METROPOLIS-ADJUSTED LANGEVIN ALGORITHM

The Metropolis-Hastings (MH) algorithm is a fundamental method within the class of Markov Chain Monte Carlo (MCMC) techniques, widely employed for sampling from probability distributions, particularly when the target distribution is known only up to a normalizing constant. Given a target distribution π on $\mathbb{R}^d$ with density with respect to (w.r.t.) the Lebesgue measure, which is also denoted as π for simplicity,

$$\pi(x) = \frac{e^{-U(x)}}{\int e^{-U(y)}\,\mathrm{d}y}, \tag{A.1}$$

where $U : \mathbb{R}^d \to [0, \infty)$ is the potential function, the MH algorithm enables sampling from $\pi(x)$ even when only the relative density values, i.e., the ratio $\pi(x)/\pi(y)$, are available. Specifically, starting from an initial random variable X_0, the MH algorithm constructs a Markov chain $(X_n)_{n\geq 0}$ iteratively through the following steps. At iteration n, given the current state X_n, a candidate sample Y_{n+1} is generated from a proposal distribution with density $q(X_n, y)$ w.r.t. the Lebesgue measure. This proposed sample is then accepted with probability

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, \tag{A.2}$$

where we use the notation $a \wedge b = \min(a, b)$. If the proposal is accepted, the next state is updated as $X_{n+1} = Y_{n+1}$; otherwise, the chain remains at the current state, i.e., $X_{n+1} = X_n$.

By construction, the Markov chain $(X_n)_{n\geq 0}$ is reversible w.r.t. the target density π, ensuring that π is an invariant distribution. The efficiency of the MH algorithm heavily depends on the choice of the proposal distribution q. A common choice is the Gaussian proposal centered at $x \in \mathbb{R}^d$ as

$$q(x, y) = \frac{1}{(2\sigma^2)^{n/2}} \exp\left(-\frac{|x - y|^2}{2\sigma^2}\right).$$

For symmetric proposals satisfying $q(x, y) = q(y, x)$, the acceptance probability simplifies to

$$\alpha(x, y) = 1 \wedge \frac{\pi(y)}{\pi(x)}.$$

The Metropolis-Hastings (MH) algorithms utilizing symmetric proposal distributions are specifically referred to as Random Walk Metropolis (RWM) algorithms. To optimize performance, the asymptotic acceptance probability should be tuned to approximately 0.234, as established in Roberts et al. (1997) and supported by recent studies, such as Ning (2025).

Intuitively, if the proposal density exploits the structure of the target density, a higher acceptance probability is possible. This leads to the Metropolis-adjusted Langevin Algorithm (MALA), a notable class of algorithms derived from discrete approximations to Langevin diffusions. A Langevin diffusion for $\pi(x)$ is a natural, non-explosive diffusion process that is reversible w.r.t $\pi(x)$. By leveraging the gradient of $\pi(x)$, it tends to move more frequently in directions where $\pi(x)$ increases. As a result, MALA achieves an optimal acceptance probability of approximately 0.574 (Roberts & Rosenthal, 1998), which is significantly higher than that of the RWM algorithm. Specifically, assuming U as continuously differentiable, the MALA employs a proposal distribution induced by the following Langevin diffusion:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t, \tag{A.3}$$

where B_t is a d-dimensional Brownian motion. Under additional conditions, this diffusion has a strong solution for any initial point and admits π as its unique stationary measure (Roberts & Tweedie, 1996).

Sampling a path solution of (A.3) is challenging, so discretizations are often employed to construct a Markov chain with similar long-term behavior. The Euler-Maruyama discretization of (A.3) is given by

$$Y_{n+1} = Y_n - \gamma\nabla U(Y_n) + (2\gamma)^{1/2}Z_{n+1}, \tag{A.4}$$

for all $n \in \mathbb{N}$, where γ is the step size of the discretization, and $(Z_n)_{n\geq 1}$ is an independent and identically distributed (i.i.d.) sequence of d-dimensional standard Gaussian random variables. The iteration in (A.4) is known as the Unadjusted Langevin Algorithm (ULA). According to the classical results on page 342 of Roberts & Tweedie (1996): *"whereas the genuinely continuous-time processes often perform well on large classes of target densities, the situation is much more delicate for the approximations which would be used in practice. In particular, naive discretizations of the continuous-time models may lose not only the geometric rates of convergence but also all convergence properties, even for quite standard density π"*. To address this issue, the MALA incorporates the Markov kernel associated with the Euler-Maruyama discretization (A.4) as a proposal kernel within a MH accept-reject scheme. This adjustment ensures that the resulting Markov chain better approximates π. Specifically, the MALA generates a new Markov chain X_n according to

$$X_{n+1} = X_n + (\check{Y}_{n+1} - X_n)\mathbb{1}_{\mathbb{R}_-}(U_{n+1} - \alpha_\gamma(X_n, \check{Y}_{n+1})), \tag{A.5}$$

where $\{U_n\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. uniform random variables on $[0, 1]$,

$$\check{Y}_{n+1} = X_n - \gamma\nabla U(X_n) + (2\gamma)^{1/2} Z_{n+1}, \tag{A.6}$$

and $\alpha_\gamma : \mathbb{R}^{2d} \to [0, 1]$ represents the acceptance probability function defined in (A.2) while explicitly revealing the dependence on the step size γ.

A.1.1 Gaussian illustration of discretization instability

Figure 1 uses $U(x) = x^2/(2\sigma^2)$ with $\sigma = 0.02$ and $\gamma = 0.001$. The corresponding diffusion is an Ornstein–Uhlenbeck process with invariant law $N(0, \sigma^2)$. In contrast, writing $a = 1 - \gamma/\sigma^2 = -1.5$, the ULA recursion (2.2) gives

$$a^{-n} Y_n = Y_0 + \sqrt{2\gamma} \sum_{j=1}^{n} a^{-j} Z_j.$$

Since $|a| > 1$, the series converges almost surely to a nondegenerate Gaussian random variable, which is nonzero almost surely. Consequently, $|Y_n| \to \infty$ almost surely. This establishes transience analytically, independently of the numerical stopping threshold. The instability depends on the ratio γ/σ^2, rather than on the absolute step size alone; the example does not assert instability for arbitrarily small steps on a fixed Gaussian target.

We simulate 16 replicates of each method from zero, with 100,000 updates and 20,000 burn-in updates. The Python script uses seed 42 and separate NumPy random streams for MALA and ULA, obtained through `SeedSequence`; each replicate receives independent Gaussian draws and, for MALA, independent acceptance uniforms. The proposal is $y = ax + \sqrt{2\gamma}Z$, with MH acceptance probability

$$\alpha(x, y) = 1 \wedge \exp\Big(\frac{\gamma}{4\sigma^4}(x^2 - y^2)\Big).$$

Each ULA path stops at its first crossing of $|x|/\sigma \geq 10^6$, retaining the crossing value for the trajectory plot without clipping or restarting. Panels (a)–(b) display the first replicate, fixed in advance: MALA's first 10,000 updates and ULA's path through iteration 41. Panels (c)–(d) use 64 equal-width bins on $[-4\sigma, 4\sigma]$, dividing bin counts by bin width and the total planned number of retained draws, $16 \times 80,000$. Thus, stopped paths are not silently omitted from the normalization. All ULA paths stop within 31–41 updates, leaving no post-burn-in observations. The MALA histogram pools all retained draws and has standardized variance 1.0019; its mean acceptance rate across the full runs is 39.6%. These finite-run results demonstrate accurate sampling from this initialization, without implying rapid convergence from arbitrary starting points.

A.2 Metropolis-Adjusted Subdifferential Langevin Algorithm

In this paper, we introduce the Metropolis-adjusted subdifferential Langevin algorithm (MASLA), a generalization of MALA that incorporates nonsmooth, nonconvex potential functions. In the realm of non-convex and non-smooth optimization, we consider generalized derivatives in the name of conservative field, which was introduced in Bolte & Pauwels (2021) in recent years. Conservative fields coincide with classical gradients almost everywhere, having the Clarke subgradient (Clarke, 1990) as the minimal convex conservative field. If f is differentiable, ∇f is of course a conservative

field. Consider the example from Difonzo et al. (2022) with $f : \mathbb{R} \to \mathbb{R}$ and its conservative field as follows:

$$f(x) = \begin{cases} -x-1, & x < -1 \\ x+1, & -1 \le x < 0 \\ 1-x, & 0 \le x < 1 \\ x-1, & x \ge 1 \end{cases}, \quad D_f(x) = \begin{cases} -1, & x < -1 \\ [-1,1], & x = -1 \\ 1, & -1 < x < 0 \\ [-1,1], & x = 0 \\ -1, & 0 < x < 1 \\ [-1,1], & x = 1 \\ 1, & x > 1 \end{cases}.$$

A detailed explanation of a simple locally Lipschitz function, lacking a well-defined proximal operator at all points but possessing an explicit Clarke subgradient, is provided in Section D.2.

Optimization plays a fundamental role in both statistics and machine learning, as many methods in these fields are formulated as optimization problems. In statistics, optimization is essential for parameter estimation, model fitting, and inference, while in machine learning, it underpins the training of models by minimizing loss functions and improving predictive performance. To better capture learning and prediction tasks, structural constraints like sparsity or low rank are often imposed, or the objective is designed to be non-convex. This is particularly relevant for non-linear models such as tensor models and deep networks. While non-convex optimization offers immense modeling flexibility, the resulting problem is often NP-hard to solve. This challenge is common in fields like genomics and signal processing, where *data starvation* occurs. Consider a set of n covariate/response pairs $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^d$ and $y \in \mathbb{R}$, *data starvation* means that the number of data points n is much smaller than the number of model parameters p, i.e., $n \ll p$. Standard statistical methods, which require $n \ge p$, struggle in such scenarios. *Sparse recovery* addresses this issue by fitting a sparse model vector to the data:

$$\hat{w}_{\text{sp}} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (y_i - x_i^T w)^2,$$

such that $w \in \mathcal{B}_0(s) := \{x \in \mathbb{R}^d : |\operatorname{supp}(x)| \le s\}$, the set of s-sparse vectors with $\operatorname{supp}(x) := \{i : x_i \neq 0\}$. Although the objective function is convex, the constraint is non-convex. For detailed illustrations and examples, see Jain & Kar (2017). In the MCMC regime, stochastic gradient Langevin dynamics have been developed for non-convex learning; see, e.g., Deng et al. (2020).

Modern problems in AI and numerical analysis often require nonsmooth formulations, for example, deep learning models employing ReLU activation functions and square loss. In statistical M-estimation, nonsmooth regularization functions, particularly the ℓ_1-norm and its variants, have long served as foundational tools, demonstrating strong empirical and theoretical performance over decades. In the Bayesian framework, nonsmooth priors also play a central role in high-dimensional models (Seeger, 2008; O'Hara & Sillanpää, 2009). The Bayesian analogue of the ℓ_1-penalty is the Laplace prior, which has been widely studied (Park & Casella, 2008; Carvalho et al., 2008; Dalalyan et al., 2018). To address the challenges of nonsmooth potentials in Bayesian computation, Pereyra (2016) proposed applying both MALA and ULA to the Moreau–Yosida envelope of a nonsmooth potential g. Bernton (2018) analyzed the performance of the unadjusted approach in a specific setting. Durmus et al. (2018) extended these methods to composite potentials of the form $f + g$, where g is nonsmooth, by replacing g with its smooth Moreau–Yosida approximation. This smoothing approach has also been applied to Hamiltonian Monte Carlo (Chaari et al., 2016). Later Durmus et al. (2019) studied the mixing time problem for nonsmooth composite objectives using Wasserstein gradient flow. Habring et al. (2024a) further investigated the setting where the target potential takes the form $U(x) = F(x) + G(Kx)$, with $F : \mathbb{R}^d \to [0, \infty)$ and $G : \mathbb{R}^{d'} \to [0, \infty)$ both proper, convex, and lower semi-continuous, and $K : \mathbb{R}^d \to \mathbb{R}^{d'}$ a linear operator.

B MATHEMATICAL BACKGROUND AND FULL PROOFS

This appendix develops the conservative-field framework in Section B.1, establishes the well-posedness and exact invariance of MASLA in Section B.2, and provides the selection-independence and convergence proofs in Section B.3.

B.1 CONSERVATIVE FIELDS AND GENERALIZED DERIVATIVES

In this paper, we consider locally Lipschitz continuous functions in the Euclidean space $\mathbb{R}^d$, which is equipped with the canonical Euclidean inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field on $\mathbb{R}^d$. The notation A^c represents the complementary set of a set A and $\mathrm{cl}(A)$ denote the closure of A. Let λ be the Lebesgue measure on $\mathbb{R}$ and $\mathbb{R}^d$, regardless of the dimension.

Definition B.1 (Closed graph) *The graph of a set-valued map $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^q$ is given by*

$$\mathrm{Graph}(D) = \Big\{(x, z) : x \in \mathbb{R}^d,\, z \in D(x)\Big\}.$$

D is said to have closed graph or to be graph closed if $\mathrm{Graph}(D)$ is closed as a subset of $\mathbb{R}^{d+q}$.

Definition B.2 (Upper semicontinuous) *We say that a set-valued map $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^q$ is upper semi-continuous, if its graph is closed.*

Definition B.3 (Absolutely continuous curve) *An absolutely continuous curve is a continuous function $x\colon \mathbb{R} \to \mathbb{R}^d$ which admits a derivative $\dot{x}$ for Lebesgue almost all $t \in \mathbb{R}$ in which case $\dot{x}$ is Lebesgue measurable, and $x(t) - x(0)$ is the Lebesgue integral of $\dot{x}$ between 0 and t for all $t \in \mathbb{R}$.*

Definition B.4 (Conservative set-valued fields) *Let $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a set-valued map. D is a* conservative (set-valued) field *whenever it has closed graph, nonempty compact values and for any absolutely continuous loop $\gamma\colon [0,1] \to \mathbb{R}^d$, that is $\gamma(0) = \gamma(1)$, we have*

$$\int_0^1 \max_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \, \mathrm{d}t = 0,$$

which is equivalent to

$$\int_0^1 \min_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \, \mathrm{d}t = 0,$$

where the integrals are understood in the Lebesgue sense. Another equivalent characterization in term of the Aumann integral is given by

$$\int_0^1 \langle D(\gamma(t)), \dot{\gamma}(t) \rangle \, \mathrm{d}t = \{0\}. \tag{B.1}$$

Chain rule proved in Lemma 2 of Bolte & Pauwels (2021) characterizes conservativity in the following sense:

Lemma B.5 (Chain rule and conservativity, Bolte & Pauwels (2021)) *Let $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a locally bounded, graph closed set-valued map and $f\colon \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function. Then D is a conservative field for f, if and only if for any absolutely continuous curve $x\colon [0,1] \to \mathbb{R}^d$, the function $t \mapsto f(x(t))$ satisfies*

$$\frac{d}{dt} f(x(t)) = \langle v, \dot{x}(t) \rangle \qquad \forall v \in D(x(t)), \tag{B.2}$$

for almost all $t \in [0,1]$.

Definition B.6 (Potential functions of conservative fields) *Let $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a conservative field. Define a function f by any of the equivalent expressions*

$$f(x) \quad = \quad f(0) + \int_0^1 \max_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \, \mathrm{d}t$$

$$= f(0) + \int_0^1 \min_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle \, \mathrm{d}t$$
$$= f(0) + \int_0^1 \langle \dot{\gamma}(t), D(\gamma(t)) \rangle \, \mathrm{d}t,$$

where $\gamma\colon [0,1] \to \mathbb{R}^d$ is any absolutely continuous function with $\gamma(0) = 0$ and $\gamma(1) = x$. The function f is well defined and unique up to an additive constant. It is called a potential function for D. We also say that D admits f as a potential, or equivalently, that D is a conservative field for f.

Recall that the convex hull of a set S in a real vector space (e.g., $\mathbb{R}^d$) is the smallest convex set that contains S, denoted as $\mathrm{conv}(S)$.

Proposition B.7 (Bolte & Pauwels (2021), Remark 3) *The following properties hold:*

- *The potential function f defined in Definition B.6 is locally Lipschitz continuous.*
- *If D_1, D_2 are two graph closed set-valued mappings with compact nonempty values, then $D_1 \subset D_2$ and D_2 conservative implies that D_1 is conservative as well.*
- *If D is conservative, $\mathrm{conv}(D)$ is conservative as well.*

A locally Lipschitz continuous function, $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable almost everywhere by Rademacher's theorem (Evans & Gariepy, 2015).

Definition B.8 (Clarke subgradient, Clarke et al. (2008)) *For any locally Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\mathcal{D}_f$ the set of points $x \in \mathbb{R}^d$ such that f is differentiable at x, and then the Clarke subgradient of f is given for any $x \in \mathbb{R}^d$, by*

$$\partial^c f(x) = \mathrm{conv}\Big\{ y \in \mathbb{R}^d \, : \, \exists (x_n)_{n\in\mathbb{N}} \in \mathcal{D}_f^{\mathbb{N}} \text{ s.t. } (x_n, \nabla f(x_n)) \to (x, y) \Big\}. \tag{B.3}$$

The following proposition collects the results of Theorem 1, Corollary 1, and Proposition 1 of Bolte & Pauwels (2021).

Proposition B.9 (Bolte & Pauwels (2021)) *Consider a conservative field $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ for $f\colon \mathbb{R}^d \to \mathbb{R}$. Then the following hold:*

(i) $D = \{\nabla f\}$ *Lebesgue almost everywhere.*

(ii) $\partial^c f$ *is a conservative field for f, and $\partial^c f(x) \subset \mathrm{conv}(D(x))$ for all $x \in \mathbb{R}^d$.*

(iii) We say that x is D-critical for f if $D(x) \ni 0$, and the value $f(x)$ is then called a D-critical value. When $D\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ has non-empty compact values and closed graph, for $x \in \mathbb{R}^d$ being a local minimum or local maximum of f, we have $0 \in \mathrm{conv}(D(x))$.

Proposition B.10 (Bolte & Pauwels (2021), Corollary 2) *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz continuous, then the following assertions are equivalent:*

(i) f is the potential of a conservative field on $\mathbb{R}^d$.

(ii) $\partial^c f$ is a conservative field.

(iii) f has chain rule for the Clarke subgradient.

The breadth of the class of potential functions for conservative fields is highlighted by the following sufficient conditions.

Proposition B.11 (Bolte & Pauwels (2021), Proposition 2) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz continuous, the following are sufficient conditions for f to be potential of a conservative field on $\mathbb{R}^d$:*

(i) f is convex or concave.

(ii) f *or* $-f$ *is Clarke regular.*

(iii) f *or* $-f$ *is prox regular.*

(iv) f *is real semialgebraic or more generally tame.*

Tame functions are defined via o-minimal structures, which we recall in the following definition. The simplest o-minimal structure is given by the class of real semialgebraic objects. Recall that a set $A \subset \mathbb{R}^d$ is called *semialgebraic* if it is a finite union of sets of the form

$$\bigcap_{i=1}^{k} \Big\{ x \in \mathbb{R}^d \mid g_i(x) < 0,\ h_i(x) = 0 \Big\},$$

where the functions $g_i, h_i : \mathbb{R}^d \to \mathbb{R}$ are real polynomial functions and $k \geq 1$.

Definition B.12 (o-minimal structure, Coste (1999)) *An o-minimal structure on $(\mathbb{R}, +, \cdot)$ is a collection of sets $\mathcal{O} = (\mathcal{O}_d)_{d \in \mathbb{N}}$ where each $\mathcal{O}_d$ is itself a family of subsets of $\mathbb{R}^d$, such that for each $d \in \mathbb{N}$:*

(i) *$\mathcal{O}_d$ is stable by complementation, finite union, finite intersection and contains $\mathbb{R}^d$.*

(ii) *if A belongs to $\mathcal{O}_d$, then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{d+1}$;*

(iii) *if $\pi : \mathbb{R}^{d+1} \to \mathbb{R}^d$ is the canonical projection onto $\mathbb{R}^d$ then, for any $A \in \mathcal{O}_{d+1}$, the set $\pi(A)$ belongs to $\mathcal{O}_d$;*

(iv) *$\mathcal{O}_d$ contains the family of real algebraic subsets of $\mathbb{R}^d$, that is, every set of the form $\{x \in \mathbb{R}^d \mid g(x) = 0\}$ where $g : \mathbb{R}^d \to \mathbb{R}$ is a polynomial function;*

(v) *the elements of $\mathcal{O}_1$ are exactly the finite unions of intervals.*

Definition B.13 (Tame functions, Van den Dries (1998)) *A subset of $\mathbb{R}^d$ which belongs to an o-minimal structure $\mathcal{O}$ is said to be definable in $\mathcal{O}$. A set-valued function is said to be definable in $\mathcal{O}$ whenever its graph is definable in $\mathcal{O}$. The terminology tame refers to definability in an o-minimal structure.*

B.2 Wellposedness of MASLA

In this section, we demonstrate that MASLA is a well-designed algorithm with desirable theoretical properties. We adopt the term *well-posedness* from the differential equations literature, since MASLA is based on a set-valued extension of the Langevin diffusion (3.2). We begin by showing that (3.2) admits a solution and thus defines a properly posed stochastic differential inclusion (SDI). A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is said to be *piecewise Lipschitz continuous* if there exists a hypersurface Θ, referred to as the *exceptional set* of f, such that Θ has finitely many connected components $\{\theta_i\}_{i=1}^m$ and the restriction $f|_{\mathbb{R}^d \setminus \Theta}$ is intrinsically Lipschitz; see Definition 3.2 of Leobacher & Szölgyenyi (2017). Similarly, a set-valued map $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is called *piecewise Lipschitz continuous* if there exists an exceptional set Θ, such that F is Lipschitz continuous on $\mathbb{R}^d \setminus \Theta$ w.r.t. the Hausdorff metric; see Baier & Farkhi (2013). By Theorem 15 of Difonzo et al. (2022), if the conservative field D_U is piecewise Lipschitz continuous with a C^3 exceptional set Θ of positive reach, then the SDI (3.2) admits a strong solution that is unique almost everywhere. Notably, the Clarke subgradient of a continuous tame function (Definition B.13) satisfies these assumptions; we refer interested readers to Difonzo et al. (2022) for further examples. Moreover, by Corollary 1 of Bolte & Pauwels (2021), restated here as Proposition B.9, the Clarke subgradient is the minimal convex conservative field.

Independently of these SDI existence results, Proposition 3.3 establishes that the MASLA kernel is well-defined, admits π as an invariant distribution, and is reversible w.r.t. π for every step size $\gamma > 0$, using only a fixed Borel measurable selection and integrability of the target density.

Example 3.2 (restated). *Let $g : \mathbb{R} \to \mathbb{R}$ be defined by*

$$g(x) := \begin{cases} x^2 \sin(1/x), & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Then g is locally Lipschitz on $\mathbb{R}$ and differentiable at $x = 0$ with

$$\nabla g(0) = 0.$$

However, its Clarke subdifferential at 0 satisfies

$$\partial g(0) = [-1, 1].$$

Proof of Example 3.2. Since g is continuously differentiable on $\mathbb{R} \setminus \{0\}$ and bounded by $|g(x)| \leq |x|^2$, it is locally Lipschitz on $\mathbb{R}$. Differentiability at 0 follows from

$$\nabla g(0) = \lim_{h \to 0} \frac{g(h) - g(0)}{h} = \lim_{h \to 0} h \sin(1/h) = 0.$$

For $x \neq 0$, the derivative of g is

$$g'(x) = 2x \sin(1/x) - \cos(1/x).$$

As $x \to 0$, the term $2x \sin(1/x)$ converges to 0, while $\cos(1/x)$ oscillates between -1 and 1. Hence, for any $a \in [-1, 1]$, there exists a sequence $(x_n)_{n \geq 1}$ converging to 0 such that

$$g'(x_n) \to a.$$

By definition of the Clarke subdifferential (B.3),

$$\partial g(0) = \operatorname{conv}\Big\{ \lim_{n \to \infty} g'(x_n) : \ x_n \to 0, \ g \text{ differentiable at } x_n \Big\} = [-1, 1].$$

■

Proposition 3.3 (Well-posedness and exact invariance) (restated). *Let $U : \mathbb{R}^d \to \mathbb{R}$ be finite-valued and Borel measurable and satisfy*

$$0 < Z_U := \int e^{-U(x)} \, \lambda(dx) < \infty, \qquad \pi(x) = Z_U^{-1} e^{-U(x)}.$$

Fix a finite-valued Borel map $\beta : \mathbb{R}^d \to \mathbb{R}^d$ (in MASLA, a selection of D_U). For every $\gamma > 0$, define

$$q_\gamma(x, y) = (4\pi\gamma)^{-d/2} \exp\left(-\frac{\|y - x + \gamma\beta(x)\|^2}{4\gamma} \right),$$

$$\alpha_\gamma(x, y) = 1 \wedge \frac{\pi(y) q_\gamma(y, x)}{\pi(x) q_\gamma(x, y)}, \qquad r_\gamma(x) = 1 - \int q_\gamma(x, y) \alpha_\gamma(x, y) \, \lambda(dy).$$

Then the Markov transition kernel

$$R_\gamma(x, A) = \int_A q_\gamma(x, y) \alpha_\gamma(x, y) \, \lambda(dy) + r_\gamma(x) \delta_x(A), \quad A \in \mathcal{B}(\mathbb{R}^d),$$

is well-defined from every initial law, reversible with respect to π, and admits π as an invariant probability distribution.

Proof of Proposition 3.3. Fix $\gamma > 0$. Conditional on $X_n = x$, the proposal is Gaussian with mean $x - \gamma\beta(x)$ and covariance $2\gamma I_d$, which gives (3.7). Since β is Borel measurable and finite-valued, q_γ is jointly Borel measurable, strictly positive, and finite, and $\int q_\gamma(x, y) \, \lambda(dy) = 1$ for every x. Likewise, π is Borel measurable, strictly positive, and finite. Consequently, the ratio in (3.8) is well-defined for every pair (x, y), and α_γ is Borel measurable with $0 < \alpha_\gamma(x, y) \leq 1$.

Measurability of parameter integrals shows that r_γ is Borel measurable, with $0 \leq r_\gamma(x) \leq 1$. For every Borel set A, $x \mapsto R_\gamma(x, A)$ is thus Borel measurable. For each x, $R_\gamma(x, \cdot)$ is a nonnegative measure of total mass one by (3.8). This proves that R_γ is a Markov kernel and defines the MASLA chain for any initial law.

To prove reversibility, observe that for every $x, y \in \mathbb{R}^d$,

$$\begin{aligned} \pi(x) q_\gamma(x, y) \alpha_\gamma(x, y) &= \min\{\pi(x) q_\gamma(x, y), \pi(y) q_\gamma(y, x)\} \\ &= \pi(y) q_\gamma(y, x) \alpha_\gamma(y, x). \end{aligned}$$

The rejection term is also symmetric: for any Borel sets A, B,

$$\int_A \pi(x) r_\gamma(x) \delta_x(B)\, \lambda(dx) = \int_{A\cap B} \pi(x) r_\gamma(x)\, \lambda(dx).$$

More explicitly, write $\pi(dx) = \pi(x)\lambda(dx)$ and expand the kernel in (3.9). For any Borel sets A, B, we obtain

$$\begin{aligned}
&\int_A \pi(dx) R_\gamma(x, B) \\
&= \int_A \int_B \pi(x) q_\gamma(x,y) \alpha_\gamma(x,y)\, \lambda(dy)\lambda(dx) + \int_{A\cap B} \pi(x) r_\gamma(x)\, \lambda(dx) \\
&= \int_A \int_B \pi(y) q_\gamma(y,x) \alpha_\gamma(y,x)\, \lambda(dy)\lambda(dx) + \int_{A\cap B} \pi(y) r_\gamma(y)\, \lambda(dy) \\
&= \int_B \pi(y) \left[\int_A q_\gamma(y,x) \alpha_\gamma(y,x)\, \lambda(dx) \right] \lambda(dy) + \int_B \pi(y) r_\gamma(y) \delta_y(A)\, \lambda(dy) \\
&= \int_B \pi(dy) R_\gamma(y, A).
\end{aligned}$$

The second equality uses the pointwise symmetry of the accepted-move term proved above and renames the integration variable in the rejection term. The third equality follows from Tonelli's theorem, since the double-integral integrand is nonnegative, and from $\delta_y(A) = \mathbb{1}_A(y)$. The last equality is precisely the definition of $R_\gamma(y, A)$. This establishes detailed balance. Taking $A = \mathbb{R}^d$ yields $\int \pi(dx) R_\gamma(x, B) = \pi(B)$, proving invariance. Since $\gamma > 0$ was arbitrary, all conclusions hold for every positive step size. ■

B.3 Convergence of MASLA

Proposition 4.1 (Selection independence) (restated). *Suppose U is locally Lipschitz and satisfies (3.6), and D_U is a conservative field for U. Let N_U be its nondifferentiability set. Consider finite-valued Borel selections β_1, β_2 of D_U with*

$$\beta_1(x) = \beta_2(x) = \nabla U(x), \qquad x \notin N_U,$$

with possibly different values on N_U. For $i \in \{1, 2\}$ and $\gamma > 0$, let $R_{\gamma,i}$ be the MASLA kernel constructed from β_i as in (3.7)–(3.9). Then, for every $\gamma > 0$,

$$R_{\gamma,1}(x, \cdot) = R_{\gamma,2}(x, \cdot), \qquad x \notin N_U.$$

In particular, the kernels agree for λ-almost every x and for π-almost every x.

If both chains have the same initial law μ with $\mu(N_U) = 0$, then their laws on the sequence space $(\mathbb{R}^d)^{\{0,1,\dots\}}$ coincide, and each chain satisfies

$$\mathbb{P}_\mu\big(X_n \notin N_U \text{ for every } n \geq 0\big) = 1.$$

This applies, in particular, to every absolutely continuous initial law and to every deterministic initial point outside N_U. Moreover, an absolutely continuous initial law remains absolutely continuous at every iteration under either kernel.

Proof of Proposition 4.1. The nondifferentiability set N_U is Borel, and Rademacher's theorem gives $\lambda(N_U) = 0$. Since π has a Lebesgue density, also $\pi(N_U) = 0$. Fix $\gamma > 0$, and write $q_{\gamma,i}$, $\alpha_{\gamma,i}$, and $r_{\gamma,i}$ for the proposal density, acceptance probability, and rejection probability associated with β_i. Set

$$p_{\gamma,i}(x,y) := q_{\gamma,i}(x,y)\alpha_{\gamma,i}(x,y) = \min\left\{ q_{\gamma,i}(x,y), \frac{\pi(y)}{\pi(x)} q_{\gamma,i}(y,x) \right\}.$$

For a fixed $x \notin N_U$, equality of the drifts at x implies $q_{\gamma,1}(x,y) = q_{\gamma,2}(x,y)$ for every y. For $y \notin N_U$, equality of the drifts at y also implies $q_{\gamma,1}(y,x) = q_{\gamma,2}(y,x)$. Thus

$$p_{\gamma,1}(x,y) = p_{\gamma,2}(x,y), \qquad x, y \notin N_U.$$

The reverse proposal density can differ when $y \in N_U$, but this set is Lebesgue null. Consequently, for every Borel set A and every $x \notin N_U$,

$$\int_A p_{\gamma,1}(x,y)\,\lambda(dy) = \int_A p_{\gamma,2}(x,y)\,\lambda(dy).$$

Taking $A = \mathbb{R}^d$ also gives $r_{\gamma,1}(x) = r_{\gamma,2}(x)$. Substituting these identities into (3.9) proves (4.1), including equality of the rejection atoms.

Next, fix either i and a current state $x \notin N_U$. Applying (3.9) to the Borel set N_U gives

$$R_{\gamma,i}(x, N_U) = \int_{N_U} p_{\gamma,i}(x,y)\,\lambda(dy) + r_{\gamma,i}(x)\mathbb{1}_{N_U}(x).$$

Both terms vanish separately. For the accepted-move term, the bound $0 \le p_{\gamma,i}(x,y) \le q_{\gamma,i}(x,y)$ and $\lambda(N_U) = 0$ imply

$$0 \le \int_{N_U} p_{\gamma,i}(x,y)\,\lambda(dy) \le \int_{N_U} q_{\gamma,i}(x,y)\,\lambda(dy) = 0.$$

Indeed, the Gaussian proposal has a Lebesgue density, so it assigns zero probability to the Lebesgue-null set N_U; restricting to accepted proposals cannot increase this probability. For the rejection term, $\delta_x(N_U) = \mathbb{1}_{N_U}(x) = 0$ because $x \notin N_U$: rejection leaves the chain at x, which is still outside N_U. Thus $R_{\gamma,i}(x, N_U) = 0$ for every $x \notin N_U$.

To make the induction explicit, let $\mu_{n,i}$ denote the law of the chain at iteration n under $R_{\gamma,i}$. If $\mu_{n,i}(N_U) = 0$, then the Markov property gives

$$\begin{aligned}\mu_{n+1,i}(N_U) &= \int_{N_U^c} R_{\gamma,i}(x, N_U)\,\mu_{n,i}(dx) + \int_{N_U} R_{\gamma,i}(x, N_U)\,\mu_{n,i}(dx)\\ &= 0.\end{aligned}$$

The first integral is zero by the pointwise identity just proved, and the second is zero because $\mu_{n,i}(N_U) = 0$ and the kernel takes values in $[0,1]$. Starting from $\mu_{0,i} = \mu$ with $\mu(N_U) = 0$, induction therefore gives $\mathbb{P}_\mu(X_n \in N_U) = 0$ for each n. A countable union of these null events shows that the chain avoids N_U at all iterations almost surely. Since the two kernels coincide at every point of N_U^c, iteration of the kernels from the same initial law gives identical finite-dimensional distributions. Cylinder sets generate the product sigma-algebra, so the two laws on sequence space coincide.

Finally, let $\eta \ll \lambda$ and let A be any Lebesgue-null Borel set. For either kernel,

$$\begin{aligned}(\eta R_{\gamma,i})(A) &= \int_{\mathbb{R}^d} \eta(dx) \int_A p_{\gamma,i}(x,y)\,\lambda(dy) + \int_A r_{\gamma,i}(x)\,\eta(dx)\\ &= 0.\end{aligned}$$

To justify the vanishing of the first term explicitly, the Gaussian proposal density satisfies

$$0 \le p_{\gamma,i}(x,y) = q_{\gamma,i}(x,y)\alpha_{\gamma,i}(x,y) \le q_{\gamma,i}(x,y) \le (4\pi\gamma)^{-d/2}$$

for every x, y, since $\alpha_{\gamma,i}(x,y) \le 1$ and the exponential factor in the Gaussian density is at most one. Therefore, for each fixed x,

$$0 \le \int_A p_{\gamma,i}(x,y)\,\lambda(dy) \le (4\pi\gamma)^{-d/2}\lambda(A) = 0.$$

The outer integral with respect to $\eta(dx)$ thus integrates the identically zero function. For the second term, absolute continuity $\eta \ll \lambda$ implies $\eta(A) = 0$, and the rejection probability satisfies $0 \le r_{\gamma,i}(x) \le 1$. Hence

$$0 \le \int_A r_{\gamma,i}(x)\,\eta(dx) \le \int_A 1\,\eta(dx) = \eta(A) = 0.$$

These bounds show explicitly that neither density weighting in the first term nor rejection-probability weighting in the second can give positive mass to a set that is null for the respective integrating measure. Hence $\eta R_{\gamma,i} \ll \lambda$, and induction proves preservation of absolute continuity. ■

Remark B.14 *The requirement that both selections equal ∇U at every differentiability point is explicit here; membership in $D_U(x)$ alone does not guarantee this pointwise equality, as discussed in Remark 3.1. Also, the proposition does not assert selection independence for an initial point in N_U: the drift at such a point can affect the proposal, acceptance, and rejection probabilities. For initial laws that give N_U zero mass, however, arbitrary changes of the selection on N_U have no effect on the law of the MASLA chain.*

Proposition 4.2 (Irreducibility and Harris recurrence) (restated). *Under conditions of Proposition 3.3, including normalizability and a finite-valued Borel map β, the MASLA kernel R_γ is λ-irreducible and π-irreducible for every $\gamma > 0$. More precisely, $R_\gamma(x, A) > 0$ for every $x \in \mathbb{R}^d$ and every Borel set A with $\lambda(A) > 0$. Furthermore, the chain is positive Harris recurrent: for every such set A and every initial point x,*

$$\mathbb{P}_x\big(X_n \in A \text{ for infinitely many } n\big) = 1,$$

and it admits an invariant probability distribution. Its unique invariant probability distribution is π.

Proof of Proposition 4.2. Fix $\gamma > 0$. Since U and β are finite-valued, both $\pi(x)$ and $q_\gamma(x, y)$ are strictly positive and finite for all x, y. It follows that the accepted-move density satisfies

$$p_\gamma(x, y) = q_\gamma(x, y)\alpha_\gamma(x, y) = \min\left\{q_\gamma(x, y), \frac{\pi(y)}{\pi(x)} q_\gamma(y, x)\right\} > 0.$$

Consequently, if $\lambda(A) > 0$, then

$$R_\gamma(x, A) \geq \int_A p_\gamma(x, y)\, \lambda(dy) > 0 \qquad \text{for every } x. \tag{B.4}$$

For completeness, strict positivity of the integral follows by writing

$$A = \bigcup_{k\geq 1} \{y \in A : p_\gamma(x, y) \geq 1/k\}.$$

At least one of these sets has positive Lebesgue measure, and its contribution to the integral is strictly positive. Thus every set of positive Lebesgue measure can be reached in one step from every state, which proves λ-irreducibility. Also, π and λ have the same null sets because π has a strictly positive density, so the chain is π-irreducible as well. Taking $A = \mathbb{R}^d$ in the integral above shows that

$$1 - r_\gamma(x) = \int_{\mathbb{R}^d} p_\gamma(x, y)\, \lambda(dy) > 0 \qquad \text{for every } x. \tag{B.5}$$

We now prove Harris recurrence directly. Fix a Borel set A with $\lambda(A) > 0$, and define its hitting time and avoidance probability by

$$\tau_A := \inf\{n \geq 0 : X_n \in A\}, \qquad h(x) := \mathbb{P}_x(\tau_A = \infty),$$

where $\inf \varnothing = \infty$. The function h is measurable, $0 \leq h \leq 1$, and $h = 0$ on A. By conditioning on the first step,

$$h(x) = (R_\gamma h)(x) := \int h(y) R_\gamma(x, dy), \qquad x \in A^c. \tag{B.6}$$

Proposition 3.3 gives invariance of π. Hence

$$\begin{aligned}\int_{\mathbb{R}^d} (R_\gamma h)(x)\, \pi(dx) &= \int_{\mathbb{R}^d} h(x)\, \pi(dx) \\ &= \int_{A^c} h(x)\, \pi(dx) = \int_{A^c} (R_\gamma h)(x)\, \pi(dx).\end{aligned}$$

All these integrals are finite since $h \leq 1$. Subtracting the A^c integral gives

$$\int_A (R_\gamma h)(x)\, \pi(dx) = 0.$$

Since $R_\gamma h \geq 0$ and $\pi(A) > 0$, there exists $x_* \in A$ with $(R_\gamma h)(x_*) = 0$. Expanding the kernel and using $h(x_*) = 0$ yields

$$0 = (R_\gamma h)(x_*) = \int_{\mathbb{R}^d} p_\gamma(x_*, y) h(y)\, \lambda(dy).$$

The integrand is nonnegative, so it is zero for λ-almost every y. Since $p_\gamma(x_*, y) > 0$ everywhere, this implies $h(y) = 0$ for λ-almost every y.

It remains to obtain the conclusion at every initial point, rather than only almost everywhere. For any $x \in A^c$, expanding (B.6) now gives

$$h(x) = \int_{\mathbb{R}^d} p_\gamma(x, y) h(y)\, \lambda(dy) + r_\gamma(x) h(x) = r_\gamma(x) h(x).$$

The integral vanishes because $h = 0$ Lebesgue almost everywhere. By (B.5), $1 - r_\gamma(x) > 0$, so $h(x) = 0$. Together with $h = 0$ on A, this proves $\mathbb{P}_x(\tau_A < \infty) = 1$ for every x.

There are in fact infinitely many visits. For each integer $m \geq 0$, the Markov property at time m and $h \equiv 0$ give

$$\mathbb{P}_x(X_n \notin A \text{ for every } n \geq m) = \mathbb{E}_x[h(X_m)] = 0.$$

The event of only finitely many visits to A is the countable union of these events over m, and therefore has probability zero. This establishes Harris recurrence. Since π is an invariant probability distribution (Proposition 3.3), the chain is positive Harris recurrent. The standard uniqueness theorem for invariant measures of Harris recurrent chains then implies that π is the unique invariant probability distribution; see Theorem 11.2.5 of Douc et al. (2018). ■

Theorem B.15 (A geometric drift criterion) *Fix $\gamma > 0$ and assume the conditions of Proposition 3.3.*

Suppose there exist a finite-valued Borel function $V : \mathbb{R}^d \to [1, \infty)$, a Borel set C, constants $a \in (0, 1)$ and $b \in [0, \infty)$, a probability measure ν, and $\epsilon \in (0, 1]$ such that

$$R_\gamma V(x) := \int_{\mathbb{R}^d} V(y) R_\gamma(x, dy) \leq a V(x) + b \mathbb{1}_C(x), \qquad x \in \mathbb{R}^d, \tag{B.7}$$

$$R_\gamma(x, A) \geq \epsilon \nu(A), \qquad x \in C, \quad A \in \mathcal{B}(\mathbb{R}^d), \tag{B.8}$$

and $\nu(C) > 0$. Then $\pi(V) := \int V\, d\pi < \infty$, and R_γ is V-uniformly geometrically ergodic: there exist constants $M_\gamma < \infty$ and $\rho_\gamma \in (0, 1)$ such that

$$\|R_\gamma^n(x, \cdot) - \pi\|_V \leq M_\gamma V(x) \rho_\gamma^n, \qquad x \in \mathbb{R}^d, \quad n \geq 0. \tag{B.9}$$

Here, for a signed measure ξ with $\int V\, d|\xi| < \infty$, the weighted total variation norm is

$$\|\xi\|_V := \sup_{f \text{ Borel}, |f| \leq V} \left| \int f\, d\xi \right|.$$

In particular, convergence holds geometrically in total variation from every initial point. More generally, for any initial probability measure μ with $\mu(V) < \infty$,

$$\|\mu R_\gamma^n - \pi\|_V \leq M_\gamma \mu(V) \rho_\gamma^n, \qquad n \geq 0. \tag{B.10}$$

Proof. By Proposition 4.2, R_γ is λ-irreducible and positive Harris recurrent, with unique invariant probability distribution π. The minorization (B.8) makes C a one-step small set with minorizing measure $\epsilon\nu$, and every small set is petite. We verify aperiodicity directly using the everywhere-positive accepted-move density: every Borel set of positive λ-measure is reached in one step, by (B.4). In particular, no cyclic decomposition of period $k \geq 2$ is possible. Such a decomposition would have disjoint sets $D_0, \ldots, D_{k-1}$ of positive λ-measure with transitions from D_j to D_{j+1} (indices modulo k) having probability one. But for $x \in D_j$, (B.4) gives $R_\gamma(x, D_j) > 0$, contradicting that requirement. Thus R_γ is aperiodic.

We next verify integrability of V without presupposing $\pi(V) < \infty$. Dropping the indicator from (B.7) and iterating gives

$$R_\gamma^n V(x) \leq a^n V(x) + b \sum_{j=0}^{n-1} a^j \leq a^n V(x) + \frac{b}{1 - a}.$$

For $L > 0$, invariance of π and the inequalities $R_\gamma^n(V \wedge L) \le R_\gamma^n V$ and $R_\gamma^n(V \wedge L) \le L$ imply

$$\pi(V \wedge L) = \pi\big(R_\gamma^n(V \wedge L)\big) \le \int \left[\left(a^n V(x) + \frac{b}{1-a}\right) \wedge L\right] \pi(dx).$$

Since $V(x) < \infty$ for every x and the integrand is bounded by L, dominated convergence as $n \to \infty$ gives $\pi(V \wedge L) \le [b/(1-a)] \wedge L$. Monotone convergence as $L \to \infty$ then yields $\pi(V) \le b/(1-a) < \infty$. Integrating the original drift inequality is now legitimate and gives the sharper bound

$$\pi(V) = \pi(R_\gamma V) \le a\pi(V) + b\pi(C), \qquad \pi(V) \le \frac{b\pi(C)}{1-a}.$$

The geometric drift theorem, Theorem 15.2.4(a) of Douc et al. (2018), applies to an irreducible, aperiodic Markov kernel satisfying a geometric drift inequality toward a petite set. All of its hypotheses have now been verified for R_γ, V, and C. It gives constants $M_\gamma < \infty$ and $\rho_\gamma \in (0,1)$ satisfying (B.9).

For a probability measure μ with $\mu(V) < \infty$, the iterated drift bound gives $(\mu R_\gamma^n)(V) < \infty$. For every Borel function f with $|f| \le V$, we may therefore write

$$\begin{aligned} |\mu R_\gamma^n f - \pi(f)| &= \left| \int \big(R_\gamma^n f(x) - \pi(f)\big)\, \mu(dx) \right| \\ &\le \int M_\gamma V(x) \rho_\gamma^n \, \mu(dx) = M_\gamma \mu(V) \rho_\gamma^n. \end{aligned}$$

Taking the supremum over these functions proves (B.10). Finally, $V \ge 1$ implies $\|\xi\|_{\mathrm{TV}} \le \|\xi\|_V$, where $\|\xi\|_{\mathrm{TV}} := \sup_{|f| \le 1} |\int f \, d\xi|$, so the same bounds imply geometric convergence in total variation. ■

Remark B.16 (Verifying the drift and minorization) *The geometric ergodicity conclusion requires (B.7); exact invariance and Harris recurrence alone do not establish a geometric rate. The drift must be checked for the full MASLA kernel, including rejection. In terms of the accepted-move density, its exact expression is*

$$\begin{aligned} R_\gamma V(x) &= \int V(y) p_\gamma(x,y)\, \lambda(dy) + r_\gamma(x) V(x) \\ &= V(x) + \int \big(V(y) - V(x)\big) p_\gamma(x,y)\, \lambda(dy). \end{aligned}$$

Thus, for a given potential and selection, it suffices to find $V \ge 1$ and $a < 1$ such that the last integral is at most $-(1-a)V(x)$ for $x \notin C$, and to bound $\sup_{x \in C} R_\gamma V(x) < \infty$. A drift estimate for the unadjusted Gaussian proposal by itself does not establish this inequality for the accepted chain.

If C is compact with $\lambda(C) > 0$, and U and β are bounded on compact sets, Proposition B.17 below provides (B.8) with $\nu = \lambda(\cdot \cap C)/\lambda(C)$, so that $\nu(C) = 1$. The remaining task is then to verify the drift inequality for the particular target and step size. The function V, the drift and minorization constants, and the resulting M_γ, ρ_γ may depend on γ; no uniform rate over all positive step sizes is asserted.

Proposition B.17 (Small sets and strong aperiodicity) *Fix $\gamma > 0$ and assume the conditions of Proposition 3.3. Suppose also that U and β are bounded on compact sets. Then every compact set $C \subset \mathbb{R}^d$ with $\lambda(C) > 0$ is an accessible one-step small set for R_γ. More precisely, define*

$$\nu_C(A) := \frac{\lambda(A \cap C)}{\lambda(C)}, \qquad A \in \mathcal{B}(\mathbb{R}^d),$$

and let

$$L_C := \sup_{x \in C} \|\beta(x)\|, \qquad \Delta_C := \sup_{x \in C} U(x) - \inf_{x \in C} U(x),$$
$$D_C := \sup_{x,y \in C} \|x - y\|.$$

Then the constant

$$\epsilon_{\gamma,C} := \lambda(C)(4\pi\gamma)^{-d/2} \exp\left(-\frac{(D_C + \gamma L_C)^2}{4\gamma} - \Delta_C\right) \tag{B.11}$$

belongs to $(0, 1]$ *and satisfies*

$$R_\gamma(x, A) \geq \epsilon_{\gamma,C}\nu_C(A), \qquad x \in C, \quad A \in \mathcal{B}(\mathbb{R}^d). \tag{B.12}$$

In particular, R_γ *is strongly aperiodic, as well as* λ*-irreducible and* π*-irreducible.*

Proof. Fix a compact set C with $\lambda(C) > 0$. Compactness gives $\lambda(C) < \infty$ and $D_C < \infty$, while the boundedness assumptions give $L_C < \infty$ and $0 \leq \Delta_C < \infty$. Thus ν_C is a probability measure with $\nu_C(C) = 1$.

For $x, y \in C$, the triangle inequality yields

$$\|y - x + \gamma\beta(x)\| \leq D_C + \gamma L_C, \qquad \|x - y + \gamma\beta(y)\| \leq D_C + \gamma L_C.$$

By (3.7), both $q_\gamma(x, y)$ and $q_\gamma(y, x)$ are therefore bounded below by

$$m_{\gamma,C} := (4\pi\gamma)^{-d/2} \exp\left(-\frac{(D_C + \gamma L_C)^2}{4\gamma}\right) > 0.$$

Also, for $x, y \in C$,

$$\frac{\pi(y)}{\pi(x)} = e^{U(x)-U(y)} \geq e^{-\Delta_C}.$$

It follows that the accepted-move density satisfies

$$\begin{aligned} p_\gamma(x, y) &:= q_\gamma(x, y)\alpha_\gamma(x, y) \\ &= \min\left\{q_\gamma(x, y), \frac{\pi(y)}{\pi(x)}q_\gamma(y, x)\right\} \\ &\geq \min\{m_{\gamma,C}, m_{\gamma,C}e^{-\Delta_C}\} = m_{\gamma,C}e^{-\Delta_C}, \qquad x, y \in C. \end{aligned}$$

Here the last equality uses $\Delta_C \geq 0$.

For every $x \in C$ and every Borel set A, the kernel formula (3.9) and nonnegativity of the rejection term now give

$$\begin{aligned} R_\gamma(x, A) &= \int_A p_\gamma(x, y)\,\lambda(dy) + r_\gamma(x)\mathbb{1}_A(x) \\ &\geq \int_{A\cap C} p_\gamma(x, y)\,\lambda(dy) \\ &\geq m_{\gamma,C}e^{-\Delta_C}\lambda(A \cap C) = \epsilon_{\gamma,C}\nu_C(A). \end{aligned}$$

The constant $\epsilon_{\gamma,C}$ is strictly positive because all its factors are strictly positive. Taking $A = C$ in this inequality also shows that $\epsilon_{\gamma,C} \leq R_\gamma(x, C) \leq 1$. This proves the asserted one-step minorization and hence the small-set property (Definition 9.1.1 of Douc et al. (2018)).

By Proposition 4.2, the kernel is λ-irreducible and π-irreducible. Moreover, its one-step positivity property (B.4) gives

$$R_\gamma(x, C) > 0, \qquad x \in \mathbb{R}^d,$$

since $\lambda(C) > 0$. Thus C is accessible from every state. Finally, the minorizing measure $\epsilon_{\gamma,C}\nu_C$ assigns mass $\epsilon_{\gamma,C}\nu_C(C) = \epsilon_{\gamma,C} > 0$ to C. The existence of this accessible one-step small set proves strong aperiodicity, by Definitions 9.1.2 and 9.3.5 of Douc et al. (2018). ■

Theorem 4.3 verifies the drift condition for a class of nonsmooth potentials, allowing nonconvex perturbations and discontinuous selections. It establishes the hypotheses of Theorem B.15 from conditions directly on the potential and its conservative field.

Theorem 4.3 (MASLA geometric ergodicity) (restated). *Let* $H \in \mathbb{R}^{d\times d}$ *be symmetric and positive definite. Write* h_- *and* h_+ *for its smallest and largest eigenvalues. Suppose that*

$$U(x) = \tfrac{1}{2}x^\top Hx + g(x), \qquad |g(x) - g(y)| \leq L\|x - y\|, \qquad x, y \in \mathbb{R}^d,$$

for some $L < \infty$. Assume that g admits a conservative field D_g whose values are uniformly bounded: for some $B < \infty$,

$$\sup_{x\in\mathbb{R}^d} \sup_{v\in D_g(x)} \|v\| \le B.$$

Choose any fixed Borel measurable selection $v(x) \in D_g(x)$ and set

$$D_U(x) := Hx + D_g(x), \qquad \beta(x) := Hx + v(x).$$

For every step size

$$0 < \gamma < \frac{2}{h_+},$$

the MASLA kernel using this selection and the exact MH acceptance probability is V-uniformly geometrically ergodic with $V(x) = 1 + \|x\|^2$. Specifically, $\pi(dx) \propto e^{-U(x)}\lambda(dx)$ is well-defined, and there exist $M_\gamma < \infty$ and $\rho_\gamma \in (0, 1)$ such that

$$\|R_\gamma^n(x, \cdot) - \pi\|_V \le M_\gamma(1 + \|x\|^2)\rho_\gamma^n, \qquad x \in \mathbb{R}^d, \quad n \ge 0.$$

This includes initial points at which U is not differentiable. More precisely, writing $a_\gamma := \|I_d - \gamma H\|_{\mathrm{op}} < 1$, the rejection probability and the drift satisfy

$$\lim_{\|x\|\to\infty} r_\gamma(x) = 0, \qquad \limsup_{\|x\|\to\infty} \frac{R_\gamma V(x)}{V(x)} \le a_\gamma^2 < 1.$$

Proof. The Lipschitz bound on g gives

$$U(x) \ge \tfrac{1}{2}h_-\|x\|^2 - L\|x\| + g(0).$$

Consequently, e^{-U} is integrable and its normalizing constant is strictly positive. The chain rule for D_g, together with the usual chain rule for the quadratic term, shows that D_U is a conservative field for U. The selection β is Borel measurable and bounded on compact sets, as is U. Proposition 3.3 therefore applies. Fix a step size satisfying (4.2) and abbreviate $a = a_\gamma$. The eigenvalues of $I_d - \gamma H$ lie in $(-1, 1)$, so $a < 1$.

We first control the rejection probability without differentiating g. Define the log MH ratio by

$$\ell_\gamma(x, y) := \log \frac{\pi(y)q_\gamma(y, x)}{\pi(x)q_\gamma(x, y)}.$$

Substituting $U(x) = x^\top Hx/2 + g(x)$ and $\beta(x) = Hx + v(x)$ into the Gaussian proposal formula and expanding the squared norms yields

$$\ell_\gamma(x, y) = \frac{\gamma}{4}\big(\|Hx\|^2 - \|Hy\|^2\big) + E_\gamma(x, y), \tag{B.13}$$

where

$$\begin{aligned} E_\gamma(x, y) = g(x) - g(y) + \tfrac{1}{2}\langle y - x, v(x) + v(y)\rangle - \tfrac{\gamma}{2}\big(\langle Hy, v(y)\rangle - \langle Hx, v(x)\rangle\big) \\ - \tfrac{\gamma}{4}\big(\|v(y)\|^2 - \|v(x)\|^2\big). \end{aligned}$$

In obtaining (B.13), the terms $\frac{1}{2}(x^\top Hx - y^\top Hy)$ cancel with the corresponding cross terms from the two proposal densities. Since $\|v(x)\|, \|v(y)\| \le B$ and $\|H\|_{\mathrm{op}} = h_+$,

$$\begin{aligned} |E_\gamma(x, y)| &\le (L + B)\|x - y\| + \tfrac{\gamma h_+ B}{2}(\|x\| + \|y\|) + \tfrac{\gamma B^2}{4} \\ &\le K_\gamma(\|x\| + \|y\|) + \tfrac{\gamma B^2}{4}, \end{aligned}$$

where $K_\gamma := L + B + \frac{\gamma h_+ B}{2}$. These bounds hold at all points, including nonsmooth points, and use only the boundedness of the selected field values at both x and y.

Write a proposal from x as

$$Y = (I_d - \gamma H)x + w_x(Z), \qquad w_x(Z) := -\gamma v(x) + \sqrt{2\gamma}\, Z, \qquad Z \sim N(0, I_d).$$

Fix $T > 0$. On the event $\|Z\| \le T$, we have $\|w_x(Z)\| \le t_T := \gamma B + \sqrt{2\gamma}\, T$ and $\|Y\| \le a\|x\| + t_T$. Since H commutes with $I_d - \gamma H$,

$$\|H(I_d - \gamma H)x\| \le a\|Hx\|.$$

Expanding $\|HY\|^2$ and applying the triangle and Cauchy–Schwarz inequalities gives, on the same event,

$$\begin{aligned}\|Hx\|^2 - \|HY\|^2 &\geq (1-a^2)\|Hx\|^2 - 2a\|Hx\|h_+t_T - h_+^2t_T^2 \\ &\geq (1-a^2)h_-^2\|x\|^2 - 2ah_+^2t_T\|x\| - h_+^2t_T^2.\end{aligned}$$

Combining this bound with (B.13) yields

$$\begin{aligned}\ell_\gamma(x,Y) \geq \frac{\gamma}{4}(1-a^2)h_-^2\|x\|^2 - \left(\frac{\gamma}{2}ah_+^2t_T + K_\gamma(1+a)\right)\|x\| \\ - \frac{\gamma}{4}h_+^2t_T^2 - K_\gamma t_T - \frac{\gamma B^2}{4}.\end{aligned}$$

The coefficient of $\|x\|^2$ is strictly positive. Hence there is $R_T < \infty$ such that $\ell_\gamma(x,Y) \geq 0$ whenever $\|x\| \geq R_T$ and $\|Z\| \leq T$. Such proposals are accepted with probability one. Therefore,

$$0 \leq r_\gamma(x) = \mathbb{E}\big[1 - \alpha_\gamma(x,Y)\big] \leq \mathbb{P}(\|Z\| > T), \qquad \|x\| \geq R_T.$$

First taking the limit superior as $\|x\| \to \infty$, and then letting $T \to \infty$, proves $r_\gamma(x) \to 0$.

We now derive the drift inequality for the accepted chain. For $V(x) = 1 + \|x\|^2$, the Gaussian second-moment formula gives

$$\begin{aligned}\mathbb{E}[V(Y)] &= 1 + \|(I_d - \gamma H)x - \gamma v(x)\|^2 + 2\gamma d \\ &\leq 1 + a^2\|x\|^2 + 2a\gamma B\|x\| + \gamma^2B^2 + 2\gamma d.\end{aligned}$$

Keeping the rejection contribution explicitly and using $V \geq 0$,

$$\begin{aligned}R_\gamma V(x) &= \mathbb{E}\big[\alpha_\gamma(x,Y)V(Y) + (1-\alpha_\gamma(x,Y))V(x)\big] \\ &\leq \mathbb{E}[V(Y)] + r_\gamma(x)V(x).\end{aligned}$$

Dividing by $V(x)$ and using $r_\gamma(x) \to 0$ proves the second assertion of (4.3).

Choose any $\theta \in (a^2, 1)$. There exists $R \geq 1$ such that $R_\gamma V(x) \leq \theta V(x)$ for $\|x\| > R$. For the compact ball $C = \{x : \|x\| \leq R\}$, the preceding moment bound and $r_\gamma \leq 1$ give

$$\sup_{x \in C} R_\gamma V(x) \leq 2 + (1+a^2)R^2 + 2a\gamma BR + \gamma^2B^2 + 2\gamma d =: b_C < \infty.$$

Consequently, the full MASLA kernel satisfies the derived drift inequality

$$R_\gamma V(x) \leq \theta V(x) + b_C\,\mathbb{1}_C(x), \qquad x \in \mathbb{R}^d.$$

Since U and β are bounded on compact sets, Proposition B.17 supplies the minorization on C with $\nu_C(C) = 1$. Theorem B.15 now gives the stated V-uniform geometric convergence. Every estimate above holds for the chosen selection at every point; no differentiability of g or continuity of v was required. ■

Example B.18 (A nonsmooth nonconvex target) *Fix $\tau > 0$ and a symmetric positive definite matrix H. Set*

$$U(x) = \tfrac{1}{2}x^\top Hx - \tau\|x\|_1.$$

An additive constant may be included if a nonnegative potential is desired; it does not change the target distribution or the algorithm. The perturbation $g(x) = -\tau\|x\|_1$ is Lipschitz with constant $\tau\sqrt{d}$. A conservative field for g is

$$D_g(x) = \prod_{j=1}^d G_j(x_j), \qquad G_j(t) = \begin{cases}\{-\tau\}, & t > 0, \\ [-\tau,\tau], & t = 0, \\ \{\tau\}, & t < 0.\end{cases}$$

Indeed, along an absolutely continuous curve each coordinate has derivative zero almost everywhere on its zero set, while the usual derivative of $-\tau|t|$ applies off that set. Thus every element of D_g satisfies the chain rule, and the field has closed graph and nonempty compact values. Moreover, $\|v\| \leq \tau\sqrt{d}$ for every $v \in D_g(x)$.

The potential is nonsmooth on the coordinate hyperplanes and is nonconvex: for a coordinate vector e_j and sufficiently small $s > 0$,

$$\frac{U(se_j) + U(-se_j)}{2} = \tfrac{1}{2}H_{jj}s^2 - \tau s < U(0) = 0.$$

Nevertheless, Theorem 4.3 proves geometric ergodicity for every $0 < \gamma < 2/h_+$ and every Borel selection from D_g, including arbitrary choices in $[-\tau, \tau]$ when a coordinate is zero. This verifies geometric ergodicity for a concrete nonsmooth, nonconvex MASLA target without assuming a kernel drift inequality.

C MASLA IMPLEMENTATION AND AUTOMATIC DIFFERENTIATION

C.1 EXPLICIT MASLA PSEUDOCODE

Algorithm C.1 implements the kernel in Proposition 3.3. The input potential is evaluated up to an additive constant, so its normalizing constant is not required. The selection β is fixed before simulation and used at both the current state and the proposal.

Algorithm C.1 (MASLA with a fixed measurable selection) Input: A finite-valued potential U satisfying (3.6); a finite-valued Borel selection $\beta(x) \in D_U(x)$; a step size $\gamma > 0$; an initial point $x_0 \in \mathbb{R}^d$; a number of updates N; and a burn-in length $0 \le B < N$.

1. Initialize $x \leftarrow x_0$, $u \leftarrow U(x)$, $b \leftarrow \beta(x)$, and the acceptance count $A \leftarrow 0$.
2. For $n = 0, \dots, N-1$:
   (a) Draw $z \sim N(0, I_d)$ and $a \sim \mathrm{Uniform}(0,1)$, independently of each other and of previous draws.
   (b) Propose $y \leftarrow x - \gamma b + \sqrt{2\gamma}\, z$.
   (c) Evaluate $u_y \leftarrow U(y)$ and $b_y \leftarrow \beta(y)$ using the same selection rule as at x.
   (d) Compute the log MH ratio
   $$\ell \leftarrow u - u_y - \frac{\|x - y + \gamma b_y\|^2}{4\gamma} + \frac{\|y - x + \gamma b\|^2}{4\gamma}.$$
   (e) If $\log a < \min\{0, \ell\}$, set $(x, u, b) \leftarrow (y, u_y, b_y)$ and $A \leftarrow A + 1$. Otherwise, retain (x, u, b).
   (f) Record $X_{n+1} \leftarrow x$, including repeated states following rejection.
3. Return $(X_{B+1}, \dots, X_N)$ and the acceptance fraction A/N.

The two squared-norm terms are respectively the reverse and forward Gaussian log proposal densities, with their common constants canceled. In particular, the reverse density requires $\beta(y)$ even when the proposal is subsequently rejected. Caching $U(x)$ and $\beta(x)$ means that each update requires only one new potential evaluation and one new selection evaluation, at y. The logarithmic comparison avoids exponentiating a large log ratio. A burn-in choice alone does not certify convergence.

C.2 CHOOSING THE SELECTION WITH AUTOMATIC DIFFERENTIATION

Suppose U is represented by a fixed finite computational graph of locally Lipschitz primitives, each equipped with a locally bounded conservative field. Reverse-mode automatic differentiation, using a fixed Borel selection from each primitive field, then produces a selection β_{AD} from a conservative field for U; see Bolte & Pauwels (2021, Section 5, Theorem 8). Smooth primitives use their usual derivatives. For absolute value, choose $\mathrm{sign}(t)$ for $t \neq 0$ and a fixed value in $[-1, 1]$ at zero; for ReLU, choose 0 below zero, 1 above zero, and a fixed value in $[0, 1]$ at zero. Fixing these conventions and the computational graph makes the selection reproducible. This conclusion requires valid primitive rules: it is not a blanket guarantee for arbitrary custom backward operations or branching programs.

The resulting conservative field can be larger than the Clarke subdifferential of the composite potential. Thus, one should not automatically identify $\beta_{\mathrm{AD}}(x)$ with an element of $\partial U(x)$ at every point. For the conservative-field formulation of MASLA, use a field containing the computed selection and verify any additional assumptions of the convergence result being invoked.

For the quartic target in Appendix D.3, differentiating the expression $\sum_i (x_i^4/4 - x_i^2 + |x_i|)$ with the absolute-value derivative set to zero at zero gives exactly

$$\beta_{\mathrm{AD},i}(x) = x_i^3 - 2x_i + \mathrm{sign}(x_i), \qquad \mathrm{sign}(0) = 0,$$

which belongs to the explicit Clarke field in (D.5). This provides a direct check of the autodifferentiation output, including on the nonsmooth hyperplanes. In practice, inspect the library's primitive rules and specify a custom rule when needed to enforce the chosen convention.

For Algorithm C.1, all evaluations must use the same deterministic map β: independently resampling a subgradient or a minibatch at each evaluation does not implement its stated Gaussian proposal kernel. The MH ratio must also use the full potential U. Proposition 3.3 gives exact invariance for any fixed finite-valued Borel drift when its actual proposal density is used, but interpreting the drift as a conservative selection and applying the later convergence results require their respective hypotheses. In particular, automatic differentiation by itself does not establish the pointwise gradient agreement required by Proposition 4.1.

D ADDITIONAL EXPERIMENTAL DETAILS

This appendix details the convex composite benchmark, competing algorithms, and distance comparisons in Section D.1, then examines the nonsmooth nonconvex example, its proximal nonuniqueness and explicit Clarke subgradient, and the effect of MH correction in Section D.2. Section D.3 adds a 100-dimensional nonsmooth nonconvex quartic experiment examining numerical escape of USLA and the effect of the MH correction.

D.1 MASLA VERSUS STATE-OF-THE-ART ALGORITHMS

D.1.1 TARGET DISTRIBUTION DESCRIPTION

To ensure a fair comparison, we adopt the exact experimental setup in Habring et al. (2024a) on sampling from a two-dimensional density with the potential

$$U(x) = F(x) + G(Kx),$$

where $F : \mathbb{R}^d \to [0, \infty)$ and $G : \mathbb{R}^{d'} \to [0, \infty)$ are both proper, convex, and lower semi-continuous, and $K : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a linear operator. They considered function F as either Lipschitz continuous with constant L_F or differentiable with a Lipschitz continuous gradient with constant $L_{\nabla F}$, while G as convex and Lipschitz continuous with constant L_G. They used the total variation (TV) functional for $G(Kx)$, defined in the two-dimensional case as $G : \mathbb{R} \to \mathbb{R}$ and $K : \mathbb{R}^2 \to \mathbb{R}$ by

$$G(p) = \lambda|p|, \qquad Kx = x_2 - x_1,$$

where $\lambda > 0$ is a scaling parameter. It is straightforward to verify that G is Lipschitz continuous with Lipschitz constant $L_G = \lambda$, and its subdifferential is given by

$$\partial G(p) = \begin{cases} [-\lambda, \lambda] & \text{if } p = 0, \\ \{\text{sign}(p)\lambda\} & \text{otherwise.} \end{cases}$$

The operator K satisfies $\|K\|^2 \leq 2$. For $F : \mathbb{R}^2 \to \mathbb{R}$, they used the squared L^2 discrepancy, defined as

$$F(x) = \frac{1}{2\sigma^2}\|x - y\|_2^2,$$

for y being a known vector, which is m-strongly convex with $m = \sigma^{-2}$ and has a Lipschitz continuous gradient with constant $L_{\nabla F} = \sigma^{-2}$. Following Habring et al. (2024a), we set $y = (-1, 1)$, $\sigma = 1$, and $\lambda = 5$.

To approximate the distributions of iterates of the algorithms, we need i.i.d. samples of each iterate. Thus, for each algorithm in this section we compute $10,000$ independent Markov chains and then estimate the distribution of the kth iterate as a histogram based on these samples. The initial values of the chains are taken as $(-1, 1)$. Figure 5 reports the comparison of the true density and MASLA-sampled distribution after the first step, which is the initial sample distribution, followed by the comparison of the true density and MASLA-sampled distribution after 49,500 iterations. All experiments in this section has no burn-in discartion We can clearly see that the distribution of the first dimension is clustered around value -1 at the beginning, which is expected. After running the MASLA chain, the samples follow the true distribution.

D.1.2 COMPARISON IN WASSERSTEIN DISTANCE

Habring et al. (2024a) proposed two algorithms: the Proximal-subgradient Langevin algorithm (Prox-sub) and the Gradient-subgradient Langevin algorithm (Grad-sub). These methods, building on Durmus et al. (2019), are designed for non-smooth potentials while achieving improved convergence for smooth cases. In Figure 1 of Habring et al. (2024a), the authors investigated the convergence of Grad-sub and Prox-sub in Wasserstein-2 distance, often shortened to Wasserstein distance. It is a metric on the space of probability measures with finite second moments. Rooted in optimal transport theory, it captures the minimal cost of transporting mass to transform one distribution into another, where cost is measured by squared Euclidean distance. Formally, let μ and ν be two probability measures on $\mathbb{R}^d$ with finite second moments, i.e.,

$$\int_{\mathbb{R}^d} \|x\|^2 \, d\mu(x) < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} \|y\|^2 \, d\nu(y) < \infty.$$

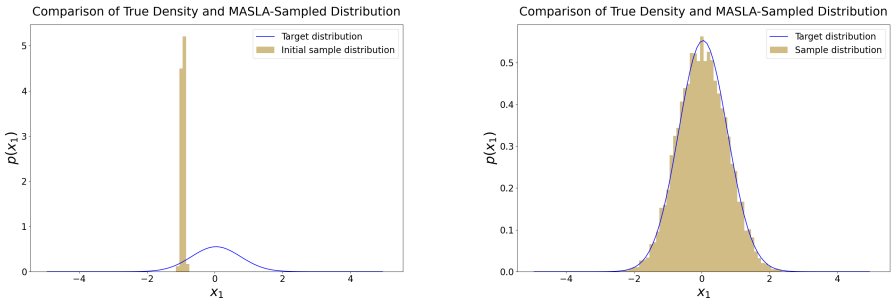

Figure 5: Comparison of MASLA-sampled initial and final distributions with the true density. The top plot shows the initial sample distribution of the first dimension after the first step, clustered around -1 as expected. The bottom plot displays the distribution after 49,500 iterations, demonstrating convergence to the true density.

The Wasserstein distance between μ and ν is defined as

$$W_2(\mu,\nu) := \left(\inf_{\zeta\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \, d\zeta(x,y)\right)^{1/2},$$

where $\Pi(\mu,\nu)$ denotes the set of all couplings of μ and ν, i.e., all probability measures ζ on $\mathbb{R}^d\times\mathbb{R}^d$ such that

$$\zeta(A\times\mathbb{R}^d) = \mu(A) \quad \text{and} \quad \zeta(\mathbb{R}^d\times B) = \nu(B)$$

for all Borel sets $A, B \subset \mathbb{R}^d$.

Prox-sub employs a proximal step for the smooth component F and a subgradient step for the nonsmooth $G(Kx)$, ensuring effective handling of nondifferentiability. Let $\partial G(KX_n)$ denote a subgradient of G at KX_n, and the update step is given by plugging $Y_{n+1} \in \partial G(KX_n)$ into

$$X_{n+1} = \operatorname*{prox}_{\tau F}(X_n - \tau K^* Y_{n+1}) + \sqrt{2\tau} Z_{n+1}, \tag{Prox-sub}$$

where $\operatorname{prox}_{\tau F}$ is the proximal mapping of F with step size τ and K^* is the adjoint of K. Grad-sub employs an explicit gradient step for the smooth component F and a subgradient step for the nonsmooth $G(Kx)$. The update step is given by plugging $Y_{n+1} \in \partial G(KX_n)$ into:

$$\begin{aligned} X_{k+1/2} &= X_n - \tau K^* Y_{n+1} \qquad \text{(Grad-sub)} \\ X_{n+1} &= X_{k+1/2} - \tau \nabla F(X_n) + \sqrt{2\tau} Z_{n+1}, \end{aligned}$$

where $\nabla F(X_n)$ is the gradient of F at X_n. The formulation of both algorithms ensures nonasymptotic convergence in the Wasserstein-2 distance under appropriate regularity conditions.

Comparison of MASLA, Grad-sub, and Prox-sub for TV $-L^2$ Sampling in Wasserstein Distance is reported in Figure 2. The Wasserstein distance between approximate distributions is computed using Python's POT package (Flamary et al., 2021). Experiments were conducted with step sizes $\tau \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. The source code for Grad-sub and Prox-sub was taken from Habring et al. (2024b). Figure 2 shows that numerical convergence rates exceed theoretical expectations for all τ values. Experimental results indicate negligible differences in convergence rates among Prox-sub, Grad-sub, and MASLA, consistent with Figure 1 of Habring et al. (2024a). All algorithms exhibit exponential ergodicity to the target density.

Calculation of the expected-rate curves. The dashed "expected rate" curves in Figure 1 of Habring et al. (2024a) evaluate upper bounds on the squared Wasserstein error, rather than fitted empirical rates. Their Theorems 4.13 and 5.2 give, for constant step size τ,

$$\begin{aligned} B_{\text{Prox}}(k) &= (1 - m\tau/2)^k E_0 + 2C\tau/m, \\ B_{\text{Grad}}(k) &= (1 - m\tau)^k E_0 + C\tau/m, \end{aligned}$$

where $E_0 = W_2^2(\mu S_\tau^G, \pi)$ is the initial squared Wasserstein error after the subgradient step, S_τ^G denotes the deterministic update $x \mapsto x - \tau K^* s(Kx)$ with the chosen $s(Kx) \in \partial G(Kx)$, and $C = 2dL_{\nabla F} + L_G^2\|K\|^2$. The plotted ordinates are $\log_{10} B(k)$. Each bound combines a geometrically decreasing initial-error term with a discretization-error bound proportional to τ, explaining the eventual plateau of the dashed curves. In the authors' plotting notebook (Habring et al., 2024b), the initial error is approximated numerically by optimal transport between the stored initial samples and a discretized target. The notebook uses $d = 2$, $m = L_{\nabla F} = 1$, $L_G = 5\sqrt{2}$, and an operator-norm bound of 2, giving $C = 204$. Thus its curves are $(1-\tau)^k \widehat{E}_0 + 204\tau$ for Grad-sub and $(1-\tau/2)^k \widehat{E}_0 + 408\tau$ for Prox-sub. At $\tau = 10^{-3}$, their limiting logarithmic levels are approximately -0.690 and -0.389, respectively. These constants are conservative: the stated two-dimensional model permits $L_G = 5$ and $\|K\| = \sqrt{2}$, giving the tighter value $C = 54$. Empirical errors below these bounds are therefore consistent with the theory. The bounds concern Grad-sub and Prox-sub; they do not provide a MASLA convergence rate or establish exact convergence to π at a fixed positive step size for the unadjusted methods.

D.1.3 Comparison in Total Variation Distance

In Figure 3 of Habring et al. (2024a), the authors compared the results of Prox-sub with those obtained using the Proximal Metropolis-Adjusted Langevin Algorithm (P-MALA) by Pereyra (2016) and the Moreau-Yosida Unadjusted Langevin Algorithm (MYULA) by Durmus et al. (2022) in TV distance. The TV distance is a fundamental measure of dissimilarity between two probability distributions. Intuitively, it quantifies the maximum difference in probability that the two measures assign to the same event. Let μ and ν be two probability measures on a measurable space $(\Omega, \mathcal{F})$. The TV distance between μ and ν is defined as

$$\mathrm{TV}(\mu, \nu) = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

If μ and ν are both absolutely continuous w.r.t. a common measure λ (e.g., Lebesgue measure), with densities $f = \frac{d\mu}{d\lambda}$ and $g = \frac{d\nu}{d\lambda}$ respectively, then the TV distance can also be written as

$$\mathrm{TV}(\mu, \nu) = \frac{1}{2} \int_\Omega |f(x) - g(x)| \, d\lambda(x).$$

The TV distance takes values in $[0, 1]$, where 0 indicates the distributions are identical, and 1 indicates they are mutually singular.

For P-MALA, the proposed update is

$$X^* = \operatorname*{prox}_{\tau(F+G\circ K)}(X_n) + \sqrt{2\tau} Z_{n+1}, \tag{P-MALA}$$

accepted as X_{n+1} with probability $\min\left\{1, \frac{\pi(X^*)q(X_n|X^*)}{\pi(X_n)q(X^*|X_n)}\right\}$, where

$$q(x|y) = \mathcal{N}\Big(\operatorname*{prox}_{\tau(F+G\circ K)}(y),\ 2\tau I_d\Big),$$

and $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate Gaussian density with mean μ and covariance Σ; otherwise, $X_{n+1} = X_n$. For MYULA, the update is

$$X_{n+1} = \left(1 - \frac{\tau}{\theta}\right) X_n - \tau \nabla F(X_n) + \frac{\tau}{\theta} \operatorname*{prox}_{\theta G\circ K}(X_n) + \sqrt{2\tau} Z_{n+1}, \tag{MYULA}$$

where $\theta > 0$ is the Moreau-Yosida parameter and the step size satisfies $\tau \leq \frac{\theta}{\theta L_{\nabla F} + 1}$. MYULA approximates the non-smooth term $G \circ K$ with its differentiable Moreau envelope

$$(G \circ K)^\theta(x) = \min_z \frac{1}{2\theta}\|z - x\|^2 + G(Kz), \tag{D.1}$$

which converges to $G \circ K$ as $\theta \to 0$. Applying the ULA to the potential $U^\theta = F + (G \circ K)^\theta$ yields the MYULA iteration, approximating the target density in total variation as $\theta \to 0$. We follow Habring et al. (2024a) in setting $\theta = 0.01$, which achieved favorable convergence.

In Figure 3 of Habring et al. (2024a), the authors evaluated the TV distance for Prox-sub using both the iterates $(\nu_n^N)_{n\in\mathbb{N}}$ with $N = 0$ and the measure $\mu Q_n^\tau S\tau^G$, with the latter demonstrating superior performance across all step sizes τ. Here, we compare only the optimal results of Prox-sub. Figure 3 illustrates the comparison of MASLA with Prox-sub (Habring et al., 2024a), P-MALA (Pereyra, 2016), and MYULA (Durmus et al., 2022) for TV $-L^2$ sampling in TV distance. Experiments were conducted with tested step sizes $\tau \in \{10^{-3}, 10^{-4}, 10^{-5}\}$. As shown in Figure 2, all algorithms exhibit exponential ergodicity to the target density for all tested τ values.

D.1.4 Numerical calculation of Wasserstein and TV errors

For the two-dimensional convex benchmark, the reference-density construction and error evaluation follow Habring et al. (2024a, Section 6.1.1) and their evaluation notebook (Habring et al., 2024b). The paper describes numerical integration of the target and comparison with sampled distributions; the implementation details below specify the discrete quantities actually evaluated. These calculations concern distributions across independent chains at a fixed iteration, rather than the time-averaged samples of the one-dimensional experiment in Section D.2.

Discretized reference distribution. The notebook uses the square $\Omega = [-5, 5]^2$, partitioned into 100×100 rectangular bins B_j of side length 0.1. For $U = F + G \circ K$, it numerically integrates

$$Z_\Omega = \int_\Omega e^{-U(x)}\, dx, \qquad b_j = Z_\Omega^{-1} \int_{B_j} e^{-U(x)}\, dx,$$

and renormalizes the computed bin masses to sum to one. Thus the reference is obtained by quadrature, without drawing target samples. Write c_j for the center of bin B_j and $J = 10{,}000$ for the number of bins. For transport computations the target is represented by $\pi_h = \sum_{j=1}^J b_j \delta_{c_j}$; for density comparisons it is represented by the piecewise constant density $p_h(x) = b_j/|B_j|$ on B_j.

Wasserstein error. Let $x_k^{(1)}, \dots, x_k^{(n)}$ be the states of independent chains at iteration k, with empirical measure $\widehat{\mu}_k = n^{-1} \sum_{r=1}^n \delta_{x_k^{(r)}}$. The notebook evaluates the squared transport cost

$$\widehat{W}_{2,k}^2 = \min_{P\geq 0} \sum_{r=1}^n \sum_{j=1}^J P_{rj} \|x_k^{(r)} - c_j\|^2, \qquad \sum_j P_{rj} = \frac{1}{n}, \quad \sum_r P_{rj} = b_j. \tag{D.2}$$

This is computed using the Python Optimal Transport package (Flamary et al., 2021): `ot.dist` constructs the squared Euclidean cost matrix and `ot.emd2` returns its minimum transport cost. In particular, despite its name, the latter function returns W_2^2 for this cost, so a square root is needed to report W_2. The authors' Wasserstein notebook uses the first 5,000 stored chains and evaluates every 500 iterations; its convergence figure plots $\log_{10} \widehat{W}_{2,k}^2$. The sample locations themselves form the source support in (D.2); only the target is replaced by bin centers.

TV error and normalization convention. For TV, both distributions are compared on the same bins. Let $N_{k,j}$ count the sampled states in B_j and set $a_{k,j} = N_{k,j}/\sum_\ell N_{k,\ell}$. The density histogram returned by `numpy.histogram2d` with `density=True` is multiplied by the bin areas to obtain these probability masses. Under the convention $\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$ used above, the histogram-based estimate is

$$\widehat{\mathrm{TV}}_k = \frac{1}{2} \sum_{j=1}^J |a_{k,j} - b_j| = \frac{1}{2} \int_\Omega |\widehat{p}_{k,h}(x) - p_h(x)|\, dx,$$

where $\widehat{p}_{k,h}(x) = a_{k,j}/|B_j|$ on B_j. However, the authors' notebook computes $\sum_j |a_{k,j} - b_j|$ without the factor $1/2$, using the total variation norm of the signed difference, with values in $[0, 2]$. The two-dimensional comparison code used here follows that convention for Figure 3. Its plotted TV errors are therefore twice the distance defined above; converting the logarithmic curves to the $[0, 1]$ convention subtracts $\log_{10} 2$ from every ordinate and does not change their relative ordering or slopes.

D.2 MASLA Performance in Scenarios Beyond Competing Methods

So far we have shown that MASLA matches the superior performance when those leading algorithms are applicable. Now, we are going to reveal MASLA's strengths in cases where these methods are not applicable. We provide a simple example of a function, $f(x) = |x^2 - 1|$, defined for all $x \in \mathbb{R}$, which is locally Lipschitz continuous, lacks a single-valued proximal operator at some points, and has an explicit Clarke subgradient.

D.2.1 Locally Lipschitz Property

A function f is locally Lipschitz continuous at $x_0 \in \mathbb{R}$, if there exists a neighborhood of x_0 and a constant $L > 0$ such that $|f(x) - f(y)| \leq L|x - y|$ for all x, y in that neighborhood. We verify this for $f(x) = |x^2 - 1|$ as below:

- *Differentiable points*: The function is differentiable except at $x = \pm 1$, where $x^2 = 1$. For $x^2 \neq 1$:
  - If $x^2 > 1$ ($x > 1$ or $x < -1$), then $f(x) = x^2 - 1$, so $f'(x) = 2x$.
  - If $x^2 < 1$ ($-1 < x < 1$), then $f(x) = 1 - x^2$, so $f'(x) = -2x$.

  The derivative $f'(x) = \pm 2x$ is continuous in regions where f is differentiable. In a compact interval $[a, b]$, $|f'(x)| = |2x| \leq 2\max(|a|, |b|)$. By the mean value theorem, for x, y in a differentiable region, there exists ξ such that:
  $$|f(x) - f(y)| = |f'(\xi)||x - y| \leq 2\max(|a|, |b|)|x - y|.$$
- *Non-differentiable points*: At $x = 1$, for small h,
  $$f(1 + h) = |(1 + h)^2 - 1| = |2h + h^2| \approx 2|h|,$$
  and $f(1) = 0$. Thus:
  $$|f(1 + h) - f(1)| \approx 2|h|,$$
  indicating local Lipschitz continuity with constant approximately 2. Similarly at $x = -1$.

Hence, f is locally Lipschitz on $\mathbb{R}$, with the Lipschitz constant depending on the neighborhood size due to the linear growth of $f'(x)$.

D.2.2 Proximal Operator Not Well-Defined Everywhere

The proximal operator of f at $y \in \mathbb{R}$ with parameter $\lambda > 0$ is given by
$$\text{prox}_{\lambda f}(y) = \arg\min_{x \in \mathbb{R}} \left\{ f(x) + \frac{1}{2\lambda}(x - y)^2 \right\}.$$
For $\text{prox}_{\lambda f}(y)$ to be well-defined, the minimizer must be unique. Set $y = 0$, $\lambda = 1$ and define
$$\phi(x) := |x^2 - 1| + \frac{1}{2}x^2.$$
We present our discussion below:

- For $x^2 < 1$, we have $|x^2 - 1| = 1 - x^2$, and then
  $$\phi(x) = 1 - x^2 + \frac{1}{2}x^2 = 1 - \frac{1}{2}x^2.$$
- For $x^2 > 1$, we have $|x^2 - 1| = x^2 - 1$, and then
  $$\phi(x) = x^2 - 1 + \frac{1}{2}x^2 = \frac{3}{2}x^2 - 1.$$
- For $x = \pm 1$,
  $$\phi(1) = \frac{1}{2} \quad \text{and} \quad \phi(-1) = \frac{1}{2}.$$

The function $\phi(x)$ is maximized at $x = 0$ ($\phi(0) = 1$) for $|x| < 1$, decreases to $\phi(\pm 1) = \frac{1}{2}$, and increases for $|x| > 1$. The global minima occur at $x = \pm 1$, yielding:
$$\text{prox}_f(0) = \{-1, 1\}.$$
Since the minimizer is not unique, the proximal operator is not single-valued at $y = 0$, and thus not well-defined everywhere as a single-valued mapping, due to the non-convexity of f.

D.2.3 Explicit Clarke Subgradient

By the Clarke subgradient definition provided in Definition B.8, since f is differentiable except at $x = \pm 1$,

- For $x > 1$ or $x < -1$, we have $f'(x) = 2x$, and then $\partial f(x) = \{2x\}$.
- For $-1 < x < 1$, we have $f'(x) = -2x$, and then $\partial f(x) = \{-2x\}$.
- For $x = 1$, the right limit as $x_i \to 1^+$ yields $f'(x_i) = 2x_i \to 2$, while the left limit as $x_i \to 1^-$ gives $f'(x_i) = -2x_i \to -2$. Consequently, the Clarke subdifferential is $\partial f(1) = \text{conv}\{-2, 2\} = [-2, 2]$.
- For $x = -1$, the right limit as $x_i \to -1^+$ yields $f'(x_i) = -2x_i \to 2$, while the left limit as $x_i \to -1^-$ gives $f'(x_i) = 2x_i \to -2$. Consequently, the Clarke subdifferential is $\partial f(-1) = \text{conv}\{-2, 2\} = [-2, 2]$.

Thus, the Clarke subgradient is given by

$$\partial f(x) = \begin{cases} \{2x\} & \text{if } x < -1 \text{ or } x > 1, \\ [-2, 2] & \text{if } x = \pm 1, \\ \{-2x\} & \text{if } -1 < x < 1. \end{cases}$$

D.2.4 The Strength of the MH Correction

The multi-valued nature of the proximal operator hinders the application of proximal-based methods, however we can still compare with the algorithmic version without the MH correction to numerically visualize the strength of the MH correction. Difonzo et al. (2022) focused on the theoretical analysis of stochastic Langevin differential inclusions of continuous-time while they ran unadjusted Langevin dynamics with the Euler-Maruyama discretization. Here, we refer it as Unadjusted Subdifferential Langevin Algorithm (USLA), which simply draws a subgradient from the set-valued field at each iteration:

$$\breve{X}_{n+1} = \breve{X}_n - \gamma\beta(\breve{X}_n) + \sqrt{2\gamma}\, Z_{n+1}, \qquad \beta(\breve{X}_n) \in D_U(\breve{X}_n), \tag{D.3}$$

We initialized both USLA and MASLA chains at $x = 0$, ran them for 100,000 iterations with a step size of $\tau = 0.1$, and discarded the first 20% of iterations as burn-in. Figure 4 compares the performance of USLA and MASLA chains against the true density. MASLA closely aligns with the true density, whereas USLA deviates significantly. The total variation (TV) distance errors are 0.116761 for USLA and 0.014363 for MASLA, while the Wasserstein-2 distance errors are 0.092183 for USLA and 0.008199 for MASLA. The computation times are 0.9976 seconds for USLA and 1.5361 seconds for MASLA, measured on a personal laptop equipped with an Apple M1 Pro chip, featuring 10 cores and 16 GB of memory.

D.3 A high-dimensional nonsmooth nonconvex quartic target

The two-dimensional convex benchmark in Section D.1 and the one-dimensional nonconvex example in Section D.2 permit direct visualization of the target distribution. To complement these examples, we compare MASLA and USLA on a 100-dimensional target with a superlinearly growing drift. This experiment examines numerical escape of the unadjusted discretization and the effect of the MH correction in a higher-dimensional setting.

Target and subgradient selection. For $d = 100$, let

$$\pi_d(x) = Z_d^{-1} e^{-U_d(x)}, \qquad U_d(x) = \sum_{i=1}^{d} u(x_i), \qquad u(t) = \frac{t^4}{4} - t^2 + |t|. \tag{D.4}$$

The potential is locally Lipschitz and semialgebraic, and is nonsmooth on the coordinate hyperplanes. It is nonconvex: at points with all coordinates nonzero, its Hessian is $\text{diag}(3x_i^2 - 2)$, which has a negative eigenvalue whenever some $0 < |x_i| < \sqrt{2/3}$. The quartic terms ensure normalizability; indeed, $u(t) \geq t^4/8 - 2$. Each coordinate potential has strict local minima at 0 and ± 1,

giving 3^d strict local modes of the product density. We use the Clarke conservative field and its Borel selection

$$\partial U_d(x) = \prod_{i=1}^{d} \big(x_i^3 - 2x_i + \partial|x_i|\big), \qquad \beta_i(x) = x_i^3 - 2x_i + \operatorname{sign}(x_i), \tag{D.5}$$

where $\partial|0| = [-1, 1]$ and $\operatorname{sign}(0) = 0$. Although the target factorizes, both algorithms update the entire 100-dimensional vector at each iteration.

Simulation protocol. For each $\gamma \in \{0.05, 0.10, 0.20\}$, we simulate 16 independent replicates of each algorithm, initialized at $X_0 = 0.1\mathbf{1}_d$, for up to 10,000 updates. Both use proposals $Y = X - \gamma\beta(X) + \sqrt{2\gamma}\, Z$, with $Z \sim N(0, I_d)$. USLA accepts every proposal, as in (D.3). MASLA applies one joint MH decision using the full target and Gaussian proposal densities in (3.8); the log acceptance ratio is evaluated directly for numerical stability. Independent random streams for the algorithms and step sizes are generated from NumPy's `SeedSequence` with seed 42.

For each USLA replicate, we record the first threshold-crossing time

$$T_{\text{esc}} = \inf\{n \geq 1 : \|X_n\|_\infty \geq 10^6\}.$$

The crossing value is retained, and that replicate is then stopped to avoid overflow; paths are neither clipped nor restarted. Runs without a crossing are observed through iteration 10,000. Thus the reported escape fraction uses all 16 replicates, including those stopped early. For MASLA, we discard the first 2,000 updates and pool coordinate 1 from the remaining 8,000 states of each replicate, including repeated states from rejections. The resulting histogram contains 128,000 correlated observations. Its reference curve is the marginal density proportional to $e^{-u(t)}$, normalized numerically by the trapezoidal rule on 4,001 equally spaced points in $[-4, 4]$. Acceptance rates are computed over all 10,000 proposals and averaged across replicates.

Results. Figure 6 reports the marginal histograms, individual trajectory magnitudes, and cumulative escape fractions. No USLA replicate crossed the threshold at $\gamma = 0.05$ or 0.10 during the observation period. At $\gamma = 0.20$, all 16 replicates crossed it, with first-crossing times ranging from 10 to 336 iterations. The cubic drift explains the potential for large excursions: in each coordinate the noiseless proposal is $(1+2\gamma)t - \gamma t^3 - \gamma \operatorname{sign}(t)$, whose magnitude grows cubically for large $|t|$.

MASLA remained below the escape threshold in all runs. Its mean acceptance rates were 58.03%, 21.49%, and 1.37% at the three respective step sizes. The marginal histograms agree reasonably with the reference at the smaller step sizes, whereas the low acceptance rate and irregular histogram at $\gamma = 0.20$ indicate poor exploration over the available run length. Thus the MH correction prevents the observed explosive excursions in this experiment, but a large step size still compromises sampling efficiency. These finite simulations provide evidence of numerical instability of USLA.

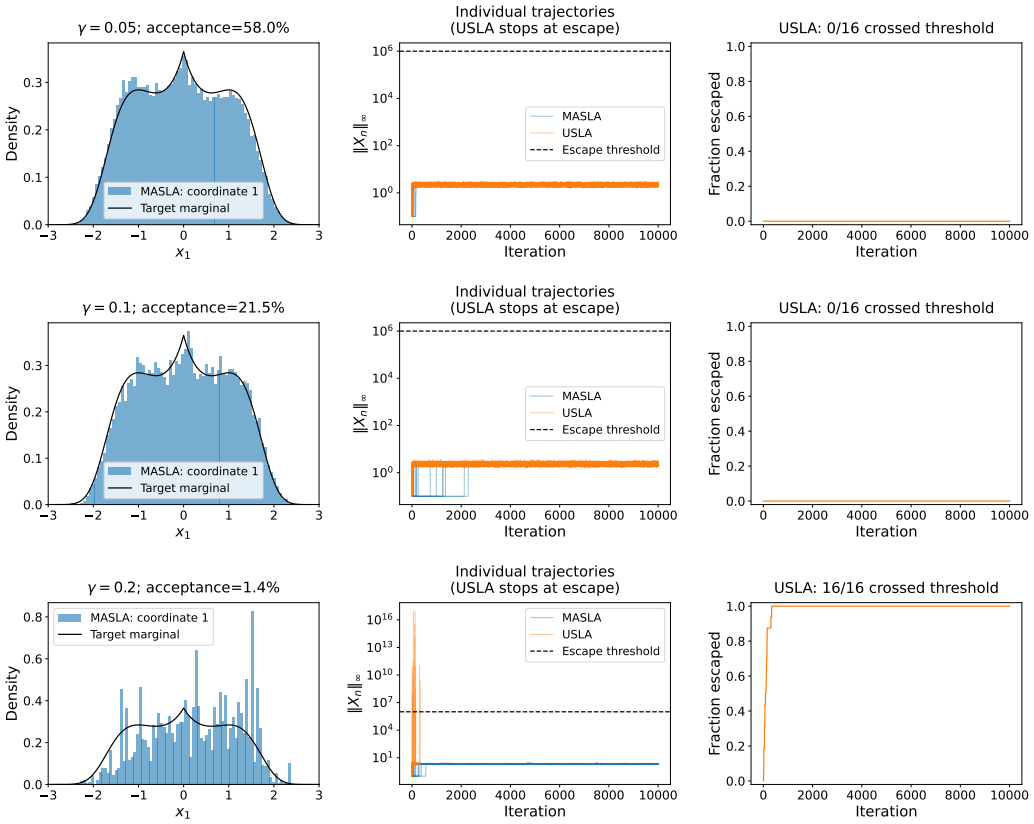

Figure 6: MASLA and USLA for the 100-dimensional nonsmooth nonconvex quartic target (D.4). Rows correspond to $\gamma = 0.05$, 0.10, and 0.20. Left: MASLA coordinate-1 histograms after burn-in and the target marginal. Middle: $\|X_n\|_\infty$ for all replicates, on a logarithmic scale; the dashed line is the 10^6 escape threshold, and USLA trajectories stop after crossing it. Right: the fraction of the 16 USLA replicates that have crossed the threshold by each iteration. All USLA replicates escape at $\gamma = 0.20$, while MASLA has a low acceptance rate at that step size.