Electronic Journal of Probability

Electron. J. Probab. **29** (2024), article no. 89, 1–27.
ISSN: 1083-6489 https://doi.org/10.1214/24-EJP1150

Mixing of Metropolis-adjusted Markov chains via couplings: The high acceptance regime[*]

Nawaf Bou-Rabee[†] Stefan Oberdörster[‡]

Abstract

We present a coupling framework to upper bound the total variation mixing time of various Metropolis-adjusted, gradient-based Markov kernels in the 'high acceptance regime'. The approach uses a localization argument to boost local mixing of the underlying unadjusted kernel to mixing of the adjusted kernel when the acceptance rate is suitably high. As an application, mixing time guarantees are developed for a non-reversible, adjusted Markov chain based on the kinetic Langevin diffusion, where little is currently understood.

Keywords: Metropolis-Hastings; couplings; mixing time; non-reversible Markov chain; high acceptance regime.
MSC2020 subject classifications: Primary 60J05, Secondary 60J60; 60J25.
Submitted to EJP on January 15, 2024, final version accepted on May 24, 2024.

1 Introduction

A nearly universal ingredient to gradient-based Markov chain Monte Carlo (MCMC) kernels are time discretizations of measure-preserving SDEs or PDMPs such as the kinetic Langevin diffusion and Andersen dynamics [74, 28, 60, 37, 11, 42, 26, 52, 7]. These kernels are gradient-based in the sense that they incorporate and rely on evaluation of the gradient of the log-density of the target distribution. In practice, the asymptotic bias due to time discretization is either incurred (leading to *unadjusted* kernels) or eliminated by a Metropolis-Hastings filter (leading to *adjusted* kernels). In either case, a question that is both fundamental mathematically and crucial to applications is [29, 63, 27, 57, 83, 68, 30]: *Starting from a distribution ν, how many steps*

[*] The work of N. Bou-Rabee has been partially supported by the National Science Foundation under Grant No. DMS-2111224. The work of S. Oberdörster has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2047/1 – 390685813.

[†] Department of Mathematical Sciences, Rutgers University Camden, 311 N 5th Street, Camden, NJ 08102, USA. E-mail: nawaf.bourabee@rutgers.edu

[‡] Institute for Applied Mathematics, University of Bonn, Endenicher Allee 60, D-53115 Bonn, Germany. E-mail: oberdoerster@uni-bonn.de

$n \in \mathbb{N}$ *are sufficient for the* n*-step distribution of the Markov chain to be an* ε*-accurate approximation of the stationary distribution in total variation?* The smallest such number of steps is the so-called ε*-mixing time* of the Markov chain from the initial distribution ν.

Recently, there has been significant progress in quantifying the mixing time of unadjusted, gradient-based kernels including the unadjusted Langevin algorithm [33, 23, 34, 43], unadjusted HMC [12, 8, 67], and various unadjusted chains based on the kinetic Langevin diffusion [21, 24, 65, 67]; see [31] for a unified and comprehensive treatment of unadjusted MCMC methods. These works give explicit upper bounds on the mixing time and complexity, which reveal that the time step size required to adequately resolve the asymptotic bias depends substantially on the accuracy ε. This potentially costly dependence motivates Metropolis adjustment, which eliminates the asymptotic bias by employing a Metropolis-Hastings filter. Intuitively speaking, it ensures the proportion of steps the adjusted chain spends in a given region equals the measure of that region with respect to the stationary distribution [62, 49, 29, 82, 4, 27, 2]. As a consequence, though, the adjusted chain involves a complex interplay between the transition step of the unadjusted kernel and the stationary distribution; to quote Bilera & Diaconis [2001], *"for many people... the Metropolis-Hastings algorithm seems like a magic trick. It is hard to see where it comes from or why it works."* Needless to say, the mixing time analysis of adjusted kernels is mathematically more delicate than of unadjusted kernels.

Intrinsically capturing the interplay described above, the notion of conductance has played a significant role in quantifying the mixing time of adjusted kernels. Classical conductance arguments are commonly used to identify bottlenecks, which yield mixing time lower bounds [54, 53, 22]. For adjusted kernels that in addition are reversible, conductance arguments can be adapted to obtain mixing time upper bounds; see, e.g., for MALA and HMC [22, 85, 19]. While these works make mild assumptions on the stationary distribution (e.g. isoperimetric inequalities) and often yield sharp mixing time upper bounds, a warm start assumption is inevitable. In particular, these mixing time upper bounds typically depend logarithmically on the L^∞-norm of the relative density of the initial to the stationary distribution; see [18, 53] for progress towards double-logarithmic dependence and [1] for sampling from a warm start. Beyond asymptotic scaling limits [71], current mathematical tools are limited in their ability to obtain rigorous quantitative mixing time upper bounds for non-reversible adjusted kernels, even from warm starting distributions.

As a step towards filling the gap in capability outlined above, in this work we introduce a new coupling framework to obtain mixing time guarantees for Metropolis-adjusted, gradient-based Markov chains. Let $\varepsilon > 0$ be the desired total variation (TV) accuracy. The underlying idea is to fix an epoch $\mathfrak{E} > 0$ of steps such that two copies of the unadjusted chain given by the kernel π^u starting from different initial conditions x and $\tilde{x}$ meet with probability at least $1-(3e)^{-1}$ after $\mathfrak{E}$ steps, i.e.,

$$\|\delta_x (\pi^u)^{\mathfrak{E}} - \delta_{\tilde{x}} (\pi^u)^{\mathfrak{E}}\|_{\mathsf{TV}} \;\le\; (3e)^{-1} , \tag{1.1}$$

where we used the coupling characterization of the TV distance $\|\cdot\|_{\mathsf{TV}}$. A standard way to ensure (1.1) is to use a contractive coupling for $\mathfrak{E}-1$ steps, followed by a one-shot coupling [72, 59, 42, 65, 8]. The time step size is then tuned such that the probability of a rejection occurring in this epoch is at most $2(3e)^{-1}$, and crucially, this tuning is at most logarithmic in $1/\varepsilon$. Hence, after one epoch, the adjusted kernel π satisfies

$$\|\delta_x \pi^{\mathfrak{E}} - \delta_{\tilde{x}} \pi^{\mathfrak{E}}\|_{\mathsf{TV}} \;\le\; e^{-1} . \tag{1.2}$$

Therefore, after $\lceil \log(1/\varepsilon) \rceil$ epochs, it immediately follows that there exists a coupling of the adjusted kernel which meets with probability at least $1-\varepsilon$. Iterating the epochs is an important step in this new coupling approach, and without this iteration, as Monmarché noted in [65], the aforementioned proof fails to capture a logarithmic scaling of the mixing time with respect to $1/\varepsilon$.

Stated precisely in Theorem 2.3, the main result of this paper provides a broadly applicable coupling framework to obtain mixing time upper bounds for Metropolis-adjusted, gradient-based Markov chains without imposing restrictive assumptions on either the stationary or the starting distribution. In essence, the theorem uses a localization argument to boost local mixing of the unadjusted kernel to mixing of the adjusted kernel when the Metropolis filter intervenes over each epoch with sufficiently low probability, i.e., in the *high acceptance regime*: a notion that is made precise in §2.1. The *low acceptance regime*, allowing for more frequent rejection, falls beyond the scope of this work. As a nontrivial application of Theorem 2.3, in §3 we develop mixing time upper bounds for a non-reversible, adjusted Markov chain based on the kinetic Langevin diffusion.

Complementary literature

Here we briefly highlight some complementary literature on related but different probabilistic techniques for mixing time analysis. In recent years, there has been progress in developing couplings for a variety of Metropolis-adjusted, gradient-based chains whose stationary distributions display high-dimensionality and/or non-logconcavity. In particular, dimension-free upper bounds in Wasserstein distance have been developed for a variant of MALA suitable for perturbations of Gaussian measures in high dimensions [38]. Moreover, a coupling of adjusted HMC that is contractive in non-logconcave settings was introduced in [9]; this coupling offers flexibility for extensions/applications [50, 6, 12]. For MALA and related Markov chains, coupling-based contractivity results are also available in distances that interpolate between L^1-Wasserstein and TV [42]. Moreover, a variety of couplings tailored to Metropolis-Hastings kernels, including maximal couplings, have recently been proposed for MCMC convergence analysis in high dimensions [50, 51, 84, 70]. In addition, there is a considerable and growing body of work devoted to Harris Ergodic Theorem, which is a very powerful tool for verifying geometric ergodicity of Markov chains [63, 64, 76, 73, 45, 36]; for a simple and elegant proof see [46]. Over the years there have been many successful applications of this tool including [61, 75, 60, 81, 10, 48, 15, 32, 58, 36], just to cite a few. There have also been significant advances in refining Harris Ergodic Theorem to obtain more explicit quantitative bounds under more easily verifiable conditions [47, 41, 25, 35, 86].

2 Main result

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let S be a Polish state space with metric d and Borel σ-algebra $\mathcal{B}$. Denote by $\mathcal{P}(S)$ the set of probability distributions on $(S, \mathcal{B})$. Let $\mu \in \mathcal{P}(S)$. A standard way to construct a gradient-based, ergodic Markov chain with stationary distribution μ is to first construct a μ-preserving, ergodic Markov chain with transition kernel π^{exact} from the exact flow of a μ-preserving SDE or PDMP. Both for theoretical purposes and for implementability in applications, it can be desirable to replace the exact flow in π^{exact} by an approximate flow based on time-discretization, which yields an unadjusted Markov transition kernel π^u. However, this unadjusted kernel has the significant drawback that $\mu\pi^u \neq \mu$. Resolving the resulting asymptotic bias in applications can be infeasible. Metropolis-adjustment provides a tool for correcting the stationary distribution and produces an adjusted transition kernel π satisfying $\mu\pi = \mu$. More precisely, we consider transition steps $X \sim \pi(x, \cdot)$ that for $\omega \in \Omega$ are of the general form

$$X(\omega) \;=\; \Phi(\omega, x)\, \mathbf{1}_{A(x)}(\omega) \;+\; \Psi(\omega, x)\, \mathbf{1}_{A(x)^c}(\omega)\;, \tag{2.1}$$

where $\Phi, \Psi : \Omega \times S \to S$ are product measurable and such that $\Phi(\cdot, x) \sim \pi^u(x, \cdot)$ and $\Psi(\cdot, x) \sim \pi^r(x, \cdot)$ for all $x \in S$, where π^r, like π^u, is a probability kernel on $(S, \mathcal{B})$. Hereafter, we omit ω from the notation, writing $\Phi(\cdot, x) = \Phi(x)$ and $\Psi(\cdot, x) = \Psi(x)$. The indicator function of the event $A(x) = \{\mathcal{U} \le \alpha(x, \Phi(x))\} \subseteq \Omega$ with an independent $\mathcal{U} \sim \mathrm{Unif}(0, 1)$ indicates that the proposal $\Phi(x)$ is accepted. Otherwise the proposal is rejected, in which case the chain is allowed to move according to Ψ.

2.1 The high acceptance regime

We now introduce the *high acceptance regime*, in which acceptance occurs sufficiently often such that the adjusted kernel inherits mixing properties of the exact kernel. The TV-mixing time of π^{exact} started in the distribution $\eta \in \mathcal{P}(S)$ to a specified accuracy $\delta > 0$ is defined by

$$\tau_{\mathrm{mix}}^{exact}(\delta, \eta) \;=\; \inf\bigl\{n \ge 0 \,:\, \|\eta(\pi^{exact})^n - \mu\|_{\mathsf{TV}} \le \delta\bigr\}\;. \tag{2.2}$$

The *high acceptance regime* is characterized by the reject probabilities being suitably controlled over a time scale set by the mixing time of the exact kernel. By comparison, this will yield a mixing time upper bound for the adjusted kernel.

Definition 2.1. *On a collection $\mathcal{C} \subseteq \mathcal{P}(S)$ such that $\{\eta\pi : \eta \in \mathcal{C}\} \subseteq \mathcal{C}$, π is in the* high acceptance regime*, if*

$$\sup_{\eta \in \mathcal{C}} \tau_{\mathrm{mix}}^{exact}\bigl((3e)^{-1}, \eta\bigr) \;\cdot\; \sup_{\eta \in \mathcal{C}} \mathbb{P}_{x\sim\eta}\bigl(A(x)^c\bigr) \;\le\; (3e)^{-1}\;, \tag{2.3}$$

where we integrate over both $x \sim \eta$ and $\mathcal{U}$ in $\mathbb{P}_{x\sim\eta}$.

A key feature of Definition 2.1 is that the restrictiveness of the condition (2.3) strongly depends on the choice of $\mathcal{C}$: the larger the collection $\mathcal{C}$, the more restrictive (2.3) becomes. In one extreme $\mathcal{C} = \{\mu\}$, the adjusted kernel is *always* in the high acceptance regime since the left hand side of (2.3) trivially vanishes. This work is concerned with the other extreme: cold start distributions corresponding to $\mathcal{C}$ including distributions which may not even be absolutely continuous with respect to μ. This feature of the definition is what motivates formulating the high acceptance regime in terms of π^{exact}.

2.2 Mixing in the high acceptance regime

Assumption 2.2 stated below is geared towards the high acceptance regime defined in Definition 2.1 with $\mathcal{C}$ including cold start distributions. Under Assumption 2.2, Theorem 2.3 gives mixing time upper bounds for the adjusted kernel. To better understand Assumption 2.2, a brief description is provided.

The possibility of cold start distributions motivates using pointwise acceptance probability bounds for the adjusted chain. However, since such bounds often degenerate at infinity, Assumption 2.2 *(iv)* is introduced to localize the adjusted chain to a bounded domain $D \subseteq S$ with sufficiently high probability. By association, the underlying unadjusted chain is similarly localized to D.

In this domain, and intuitively speaking, Assumption 2.2 *(i)* and *(ii)* require that the underlying unadjusted kernel admits a *locally* successful coupling. More precisely, Assumption 2.2 *(i)* assumes there exists a coupling for π^u that is *locally* contractive in D; and Assumption 2.2 *(ii)* assumes there exists a *local* one-shot coupling for π^u in D.

Although stated in a slightly different way, the main idea underlying Assumption 2.2 *(iii)* is (2.3). Indeed, the epoch $\mathfrak{E}$ of transition steps appearing in *(iii)* is defined in such a way that by *(i)* and *(ii)*, there exists a coupling of two copies of the unadjusted chain starting at two different initial conditions within D that induces meeting with probability at least $1 - (3e)^{-1}$; therefore, this epoch $\mathfrak{E}$ is analogous to $\sup_{\eta\in\mathcal{C}} \tau_{\mathrm{mix}}^{exact}\bigl((3e)^{-1}, \eta\bigr)$ in (2.3).

Denote by Δ the diagonal in the product space $S \times S$. The couplings appearing in Assumption 2.2 are all assumed to be *faithful*. Recall that a coupling Π is faithful if $\Pi((x,x),\Delta) = 1$ for all $x \in S$. Couplings of the adjusted kernel inherit this property from couplings of the unadjusted kernel if a synchronous coupling of the underlying uniform random variables in the Metropolis filter is used.

Similarly to (2.2), define the TV-mixing time of the adjusted kernel with initial distribution $\eta \in \mathcal{P}(S)$ and accuracy $\delta > 0$ to be

$$\tau_{\mathrm{mix}}(\delta,\eta) \;=\; \inf\big\{n \geq 0 \,:\, \|\eta\pi^n - \mu\|_{\mathsf{TV}} \leq \delta\big\}\;. \tag{2.4}$$

We are now prepared to state Assumption 2.2 and then immediately afterwards the main result of the paper, followed by its proof.

Assumption 2.2. *Let $\varepsilon > 0$ be the accuracy, $\nu \in \mathcal{P}(S)$ be the initial distribution, and $D \subseteq S$ be a domain such that $\operatorname{diam}_d(D) \leq R$ for some $R > 0$.*

Regarding the unadjusted transition kernel, we require:

(i) There exists $\rho > 0$ and for all $x, \tilde{x} \in D$ a coupling $\Pi^u_{Contr}((x,\tilde{x}),\cdot)$ of $\pi^u(x,\cdot)$ and $\pi^u(\tilde{x},\cdot)$ such that the contractivity

$$\mathbb{E}d(X^u,\widetilde{X}^u) \;\leq\; (1-\rho)d(x,\tilde{x})$$

holds for $(X^u,\widetilde{X}^u) \sim \Pi^u_{Contr}((x,\tilde{x}),\cdot)$.

(ii) There exists $C_{Reg} > 0$ and for all $x, \tilde{x} \in D$ a coupling $\Pi^u_{Reg}((x,\tilde{x}),\cdot)$ of $\pi^u(x,\cdot)$ and $\pi^u(\tilde{x},\cdot)$ satisfying the regularization

$$\Pi^u_{Reg}\big((x,\tilde{x}),\Delta^c\big) \;\leq\; C_{Reg}d(x,\tilde{x})\;.$$

Regarding the adjusted transition kernel, we require:

(iii) Set the length of an epoch of transition steps at

$$\mathfrak{E} \;=\; \big\lceil \rho^{-1}\log(3eC_{Reg}R)\big\rceil + 1$$

and suppose

$$\mathfrak{E}\,\sup_{x\in D}\mathbb{P}(A(x)^c) \;\leq\; (3e)^{-1}\;.$$

(iv) To reduce to the local properties fixed hitherto, we require control of the exit probability from D over the total number of transition steps

$$\mathfrak{H} \;=\; \mathfrak{E}\,\lceil\log(2/\varepsilon)\rceil$$

consisting of sufficiently many epochs to conclude mixing to ε accuracy. Therefore let $T = \inf\{k \geq 0 \,:\, X_k \notin D\}$ and presume

$$\mathbb{P}\big(T \;\leq\; \mathfrak{H}\big) \;\leq\; \varepsilon/4$$

both for $X_0 \sim \nu$ and $X_0 \sim \mu$.

Theorem 2.3. *Suppose Assumption 2.2 holds for $\varepsilon > 0$ and $\nu \in \mathcal{P}(S)$. Then*

$$\tau_{\mathrm{mix}}(\varepsilon,\nu) \;\leq\; \mathfrak{H}\;.$$

Remark 2.4 (Scope of Coupling Framework)**.** A remarkable feature of the coupling framework presented in this section is that it uses localization to boost local mixing of the unadjusted kernel to mixing of the adjusted kernel. This feature is enabled by Assumption 2.2 *(iv)* which localizes the entire coupling argument to the domain D. The assumption that the unadjusted kernel admits a locally contractive coupling and a local one-shot coupling (i.e., Assumption 2.2 *(i)* and *(ii)*) does not impose global restrictions, such as regularity or convexity, on the stationary distribution. Therefore, this new coupling framework is broadly applicable including, in particular, to stationary distributions whose log-density is non-globally gradient or Hessian Lipschitz, non-globally concave, or even singular, e.g., potentials of Coulomb-type.

Remark 2.5. Denoting the L^1-Wasserstein distance based on d by $\mathcal{W}^1_d$, Assumption 2.2 *(i)* and *(ii)*, respectively, can be alternatively written as

$$\mathcal{W}^1_d(\pi^u(x,\cdot),\pi^u(\tilde{x},\cdot)) \ \le\ (1-\rho)d(x,\tilde{x})$$

and

$$\|\pi^u(x,\cdot)-\pi^u(\tilde{x},\cdot)\|_{\mathsf{TV}} \ \le\ C_{Reg}d(x,\tilde{x})$$

for $\rho, C_{Reg}>0$ and all $x,\tilde{x}\in D$.

Proof of Theorem 2.3. It is notationally convenient to introduce the epoch $m+1=\mathfrak{E}$ of transition steps and the total number of epochs $k=\lceil\log(2/\varepsilon)\rceil$ that will be needed to attain ε accuracy. The total number of transitions hence amounts to $k(m+1)=\mathfrak{H}$.

On the same probability space, consider two copies of the adjusted chain $X_n\sim\nu\pi^n$ and $\widetilde{X}_n\sim\mu\pi^n=\mu$, one of which in stationarity. The copies are coupled via the couplings in Assumption 2.2 *(i)* and *(ii)* of the unadjusted kernels extended to the adjusted kernels by synchronously coupling the Metropolis-steps. One then composes epochs of m transitions of the former followed by one transition of the latter, c.f. (2.7). Denote by T, $\widetilde{T}$ the first exit times from D of X_n and $\widetilde{X}_n$ respectively, and let $\mathfrak{T}=\min(T,\widetilde{T})$.

To see that $k(m+1)$ transition steps of the adjusted chain do indeed suffice, below we will use Assumption 2.2 *(i)* – *(iii)* to prove that over each epoch the following bound holds for all $x,\tilde{x}\in D$:

$$\mathbb{P}_{(x,\tilde{x})}\Big(\{X_{m+1}\neq\widetilde{X}_{m+1}\}\cap\bigcap_{l=0}^{m}\{X_l,\widetilde{X}_l\in D\}\Big)\ \le\ e^{-1}\,, \tag{2.5}$$

where $\mathbb{P}_{(x,\tilde{x})}$ is the distribution conditioned on $X_0=x$ and $\widetilde{X}_0=\tilde{x}$. Iterating (2.5) k times will then yield the desired TV-convergence to ε-accuracy. Indeed, by the coupling characterization of the TV-distance, note that the TV-distance to stationarity after $k(m+1)$ transition steps satisfies

$$\begin{aligned}\|\nu\pi^{k(m+1)}-\mu\|_{TV}\ &\le\ \mathbb{P}\big(X_{k(m+1)}\neq\widetilde{X}_{k(m+1)}\big)\\ &\le\ \mathbb{P}\big(X_{k(m+1)}\neq\widetilde{X}_{k(m+1)},\mathfrak{T}\ge k(m+1)\big)\ +\ \mathbb{P}\big(\mathfrak{T}\le k(m+1)\big)\,. \end{aligned}\tag{2.6}$$

The second term in (2.6) describes the probability that at least one copy exits D within $k(m+1)$ transition steps, and by Assumption 2.2 *(iv)* satisfies

$$\mathbb{P}\big(\mathfrak{T}\le k(m+1)\big)\ \le\ \mathbb{P}\big(T\le k(m+1)\big)+\mathbb{P}\big(\widetilde{T}\le k(m+1)\big)\ \overset{(iv)}{\le}\ \varepsilon/2\,.$$

On the other hand, in the first term in (2.6) neither chain exits D. Denote by $\mathcal{F}_n$ the σ-algebra generated by both copies up to transition step n. Now, by (2.5) and the Markov property, it holds that

$$
\begin{aligned}
&\mathbb{P}\big(X_{k(m+1)} \neq \widetilde{X}_{k(m+1)}, \mathfrak{T} \geq k(m+1)\big) \\
&\quad= \mathbb{E}\Big[\mathbb{P}\Big(\{X_{k(m+1)} \neq \widetilde{X}_{k(m+1)}\} \cap \bigcap_{l=0}^{m} \{X_{(k-1)(m+1)+l}, \widetilde{X}_{(k-1)(m+1)+l} \in D\}\Big| \\
&\qquad\qquad \mathcal{F}_{(k-1)(m+1)}\Big) \,;\, X_{(k-1)(m+1)} \neq \widetilde{X}_{(k-1)(m+1)}, \mathfrak{T} \geq (k-1)(m+1)\Big] \\
&\quad= \mathbb{E}\Big[\mathbb{P}_{(X_{(k-1)(m+1)}, \widetilde{X}_{(k-1)(m+1)})}\Big(\{X_{m+1} \neq \widetilde{X}_{m+1}\} \cap \bigcap_{l=0}^{m} \{X_l, \widetilde{X}_l \in D\}\Big) \,; \\
&\qquad\quad X_{(k-1)(m+1)} \neq \widetilde{X}_{(k-1)(m+1)}, \mathfrak{T} \geq (k-1)(m+1)\Big] \\
&\quad\overset{(2.5)}{\leq} e^{-1}\, \mathbb{P}\Big(X_{(k-1)(m+1)} \neq \widetilde{X}_{(k-1)(m+1)}, \mathfrak{T} \geq (k-1)(m+1)\Big) \\
&\quad\leq \cdots \leq e^{-k}\, \mathbb{P}(X_0 \neq \widetilde{X}_0) \leq e^{-k} \leq \varepsilon/2 \,,
\end{aligned}
$$

where we used $\{X_{k(m+1)} \neq \widetilde{X}_{k(m+1)}\} \subseteq \{X_{(k-1)(m+1)} \neq \widetilde{X}_{(k-1)(m+1)}\}$ in the first equation, which holds by faithfulness, and the choice of k in the last. Since the TV-distance to stationarity $\|\nu\pi^{k(m+1)} - \mu\|_{TV}$ is non-increasing, this shows that $k(m+1)$ transition steps of the adjusted chain suffice for ε accuracy.

We are left to show (2.5) by using Assumption 2.2 *(i)* – *(iii)*. Let $x, \tilde{x} \in D$. Denote the accept events in the $(n+1)$-th transition, i.e. from X_n to X_{n+1} and $\widetilde{X}_n$ to $\widetilde{X}_{n+1}$, by A_{n+1} and $\widetilde{A}_{n+1}$ respectively. Let X_n^u and $\widetilde{X}_n^u$ be the corresponding copies of the underlying unadjusted chain and note that $X_n = X_n^u$ on $\bigcap_{l=0}^{n-1} A_{l+1}$. Considering just one epoch consisting of $m+1$ transition steps, *(iii)* allows to restrict to the case that the Metropolis filter does not intervene over the epoch so that the probability that there exists a coupling of the adjusted chains which induces meeting is determined by the corresponding probability for the underlying unadjusted chains. More precisely,

$$
\begin{aligned}
&\mathbb{P}_{(x,\tilde{x})}\Big(\{X_{m+1} \neq \widetilde{X}_{m+1}\} \cap \bigcap_{l=0}^{m} \{X_l, \widetilde{X}_l \in D\}\Big) \\
&\quad\leq \mathbb{P}_{(x,\tilde{x})}\Big(\{X_{m+1}^u \neq \widetilde{X}_{m+1}^u\} \cap \bigcap_{l=0}^{m} (A_{l+1} \cap \widetilde{A}_{l+1}) \cap \bigcap_{l=0}^{m} \{X_l, \widetilde{X}_l \in D\}\Big) \\
&\qquad+ \sum_{l=0}^{m} \Big[\mathbb{P}_x\big(A_{l+1}^c \cap \{X_l \in D\}\big) + \mathbb{P}_{\tilde{x}}\big(\widetilde{A}_{l+1}^c \cap \{\widetilde{X}_l \in D\}\big)\Big]
\end{aligned}
$$

with the second term bounded by $2(m+1)\sup_{x\in D} \mathbb{P}(A(x)^c) \leq 2(3e)^{-1}$ by *(iii)*. For the first term, we employ m steps of the contractive coupling in *(i)* which brings the two copies of the unadjusted chain sufficiently close together for one step of the regularizing coupling in *(ii)* to induce exact meeting. This yields

$$
\begin{aligned}
&\mathbb{P}_{(x,\tilde{x})}\Big(\{X_{m+1}^u \neq \widetilde{X}_{m+1}^u\} \cap \bigcap_{l=0}^{m} (A_{l+1} \cap \widetilde{A}_{l+1}) \cap \bigcap_{l=0}^{m} \{X_l, \widetilde{X}_l \in D\}\Big) \qquad (2.7) \\
&\quad\leq \mathbb{P}_{(x,\tilde{x})}\Big(\Pi_{Reg}^u\big((X_m^u, \widetilde{X}_m^u), \Delta^c\big) \,;\, \bigcap_{l=0}^{m} \{X_l^u, \widetilde{X}_l^u \in D\}\Big) \\
&\quad\overset{(ii)}{\leq} C_{Reg}\, \mathbb{E}_{(x,\tilde{x})}\Big(d(X_m^u, \widetilde{X}_m^u) \,;\, \bigcap_{l=0}^{m-1} \{X_l^u, \widetilde{X}_l^u \in D\}\Big) \\
&\quad\overset{(i)}{\leq} C_{Reg}\, e^{-\rho m} d(x, \tilde{x}) \leq C_{Reg}\, R e^{-\rho m} \leq (3e)^{-1} \,,
\end{aligned}
$$

where in the last two steps $\mathrm{diam}(D) \leq R$ and the definition of m were used respectively. □

Remark 2.6. A key ingredient in the proof of Theorem 2.3 is a multi-step (local) minorisation condition for the adjusted kernel. Moreover, Assumption 2.2 *(iv)* – that the chain is likely to remain in the domain D for long times with high probability – is ultimately proven with a Lyapunov-type argument, which is closely related to drift conditions. This use of drift and minorisation techniques is reminiscent of Harris ergodic theorem [63, 64, 76, 73, 45, 36]. However, there are two differences that are worth highlighting. First, the so-called “small set” for the multi-step minorization condition consists of the entire domain D. Second, the drift condition needed for the exit probability estimate given in Assumption 2.2 *(iv)* is very mild; e.g. in the application considered in this work, a Foster-Lyapunov function that does not diverge too fast suffices; c.f. §4.4.

3 Application to a non-reversible, adjusted Markov chain

Although there are numerous non-asymptotic convergence results for kinetic Langevin diffusions [21, 20, 24, 40, 17] and their unadjusted discretizations [21, 20, 24, 80, 65], quantitative mixing time guarantees for *adjusted* discretizations are comparatively scarce. In view of this underdevelopment, and as an application of Theorem 2.3, mixing time guarantees for a non-reversible, adjusted Markov chain based on a discretization of the kinetic Langevin diffusion are given in Theorem 3.6 of this section.

3.1 Metropolis-adjusted kinetic Langevin algorithm (MAKLA)

Consider an absolutely continuous probability distribution on $\mathbb{R}^d$ of the form

$$\mu_{target}(dx) \propto e^{-U(x)} dx \ ,$$

where $U : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable potential energy function. Here we analyze the mixing of an MCMC method aimed at μ_{target} based on the kinetic Langevin diffusion

$$\mathrm{d}X_t \ = \ V_t \, \mathrm{d}t \ , \quad \mathrm{d}V_t \ = \ -\nabla U(X_t) \, \mathrm{d}t - \gamma V_t \, \mathrm{d}t + \sqrt{2\gamma} \, \mathrm{d}B_t \ , \tag{3.1}$$

where B_t is a standard d-dimensional Brownian motion and $\gamma > 0$ is the friction. Let I_d be the $d \times d$ identity matrix. A key property of (3.1) is that it leaves invariant the probability measure

$$\mu(dz) \ = \ \mu_{target} \otimes \mathcal{N}(0, I_d)(dx \ dv) \ \propto \ e^{-H(z)} \ dz \tag{3.2}$$

on phase space $z = (x, v) \in \mathbb{R}^{2d}$ with energy

$$H(z) \ = \ \frac{1}{2}|v|^2 \ + \ U(x) \ .$$

A variety of discretizations of (3.1) can be Metropolis-adjusted [79, 56, 13, 5, 71] and fit the framework (2.1). Here we focus on a symmetric Strang splitting [14, 5, 3], where the splitting components are given by

1. the Ornstein-Uhlenbeck (OU) flow

$$\mathrm{d}X_t \ = \ 0 \ , \quad \mathrm{d}V_t \ = \ -\gamma V_t \, \mathrm{d}t + \sqrt{2\gamma} \, \mathrm{d}B_t \ ,$$

2. the purely potential flow

$$\mathrm{d}X_t \ = \ 0 \ , \quad \mathrm{d}V_t \ = \ -\nabla U(X_t) \, \mathrm{d}t \ , \quad \text{and}$$

3. the purely kinetic flow

$$\mathrm{d}X_t \;=\; V_t\,\mathrm{d}t\;, \quad \mathrm{d}V_t \;=\; 0\;.$$

The corresponding discretized flows are for

1. the OU-substep

$$O_h(\mathsf{b})(x,v) \;=\; (x, e^{-\gamma h}v + (1-e^{-2\gamma h})^{\frac{1}{2}}\,\mathsf{b})\;, \quad \mathsf{b}\in\mathbb{R}^d\;, \tag{3.3}$$

2. the B-substep for the *kick* due to the potential part

$$\theta_h^{(B)}(x,v) \;=\; \Big(x,\; v-h\nabla U(x)\Big)\;, \quad \text{and} \tag{3.4}$$

3. the A-substep for the *drift* due to the kinetic part

$$\theta_h^{(A)}(x,v) \;=\; \Big(x+hv, v\Big)\;. \tag{3.5}$$

Combining these flow maps in the following palindromic fashion yields the *unadjusted kinetic Langevin algorithm* (UKLA) with transition step given by

$$(X_1^u, V_1^u) \;=\; O_{h/2}(\xi_2)\circ\theta_{h/2}^{(A)}\circ\theta_h^{(B)}\circ\theta_{h/2}^{(A)}\circ O_{h/2}(\xi_1)(X_0^u, V_0^u)\;, \tag{3.6}$$

where ξ_1, ξ_2 are i.i.d. random variables with distribution $\mathcal{N}(0, I_d)$. This discretization is commonly referred to as "OABAO" where each letter refers to either (3.3), (3.4) or (3.5). For the sequel, it is convenient to introduce

$$\theta_h \;=\; \theta_{h/2}^{(A)}\circ\theta_h^{(B)}\circ\theta_{h/2}^{(A)}\;. \tag{3.7}$$

By construction, θ_h is both volume-preserving and reversible [44, 11]. The transition kernel of UKLA is given by $\pi^u = \Xi\Theta\Xi$, where

$$\begin{aligned}\Xi((x,v),\,\cdot\,) \;&=\; \delta_x\otimes\mathcal{N}\big(e^{-\gamma h/2}v, (1-e^{-\gamma h})\,I_d\big)\;,\\ \Theta((x,v),\,\cdot\,) \;&=\; \delta_{\theta_h(x,v)}\;.\end{aligned}$$

Due to asymptotic bias, UKLA does not leave μ invariant, i.e., $\mu\pi^u \neq \mu$. This failure is not surprising, since although the OU steps leave μ invariant and θ_h is volume-preserving, the time discretization induces an energy error under θ_h, i.e., $(H\circ\theta_h - H)\not\equiv 0$, which is the root cause of the asymptotic bias.

The OABAO scheme can be readily Metropolis-adjusted by simply adjusting θ_h, which is possible since θ_h is both volume-preserving and reversible [11, Prop. 5.1]; see also [82, Theorem 2]. The resulting algorithm is called the *Metropolis-adjusted kinetic Langevin algorithm* (MAKLA) with transition step

$$(X_1, V_1) \;=\; O_{h/2}(\xi_2)\circ\hat{\theta}_h(\mathcal{U})\circ O_{h/2}(\xi_1)(X_0, V_0)\;, \tag{3.8}$$

where $\mathcal{U}\sim\mathrm{Unif}(0,1)$ is independent of the other random variables and the state of the chain, and the Metropolis-adjusted integrator is defined through the mapping

$$\hat{\theta}_h(\mathsf{u})(x,v) \;=\; \begin{cases}\theta_h(x,v) & \text{if } \mathsf{u}\le\alpha((x,v),\theta_h(x,v)),\\ \mathcal{S}(x,v) & \text{else,}\end{cases} \tag{3.9}$$

where $\alpha((x,v),(x',v')) = \exp(-(H(x',v')-H(x,v))^+)$ is the accept probability, $\mathcal{S}(x,v) = (x,-v)$ is the velocity flip involution, and $[\cdot]^+ = \max(0,\cdot)$. The transition kernel of MAKLA is $\pi = \Xi\widetilde{\Theta}\Xi$ with

$$\begin{aligned}\widetilde{\Theta}((x,v), dx'\,dv') \;&=\; \alpha((x,v),(x',v'))\,\delta_{\theta_h(x,v)}(dx'\,dv')\\ &\quad+\big(1-\alpha((x,v),(x',v'))\big)\,\delta_{\mathcal{S}(x,v)}(dx'\,dv')\;.\end{aligned}$$

It is easily verified that π leaves μ invariant, i.e., $\mu\pi = \mu$, and therefore, the x-marginal of the corresponding Markov chain can be used to sample from μ_{target}.

3.2 Assumptions & additional notation

For simplicity, we focus on strongly log-concave target distributions with gradient Lipschitz log-densities. More precisely, we fix the following assumptions:

Assumption 3.1. *Suppose U is K-strongly convex, i.e., there exists $K > 0$ such that*

$$\big(\nabla U(x) - \nabla U(y)\big) \cdot (x-y) \ \geq \ K|x-y|^2 \quad \textit{for all } x, y \in \mathbb{R}^d.$$

Assumption 3.2. *Suppose U has a global minimum at 0, $U(0) = 0$, and U is L-gradient Lipschitz continuous, i.e., there exists $L > 0$ such that*

$$\big|\nabla U(x) - \nabla U(y)\big| \ \leq \ L\,|x-y| \quad \textit{for all } x, y \in \mathbb{R}^d.$$

Below it is sometimes convenient to write the results and conditions in terms of the *condition number* of the target distribution defined in the usual way by $\kappa = L/K$. Define the third derivative via the trilinear product

$$\nabla^3 U(x)(a \otimes b \otimes c) \ = \ \sum\nolimits_{i,j,k=1}^{d} \partial^3_{ijk} U(x)\, a_i b_j c_k \quad \text{for } x, a, b, c \in \mathbb{R}^d.$$

Assumption 3.3. *Suppose $U \in C^3(\mathbb{R}^d)$ is L_H-Hessian Lipschitz, i.e., there exists $L_H \geq 0$ such that*

$$\big|\nabla^3 U(x)(a \otimes b \otimes c)\big| \ \leq \ L_H |a|\,|b|\,|c| \quad \textit{for all } x, a, b, c \in \mathbb{R}^d.$$

Define the sets of model parameters and user-specified hyperparameters to be $\mathcal{M} = \{d, K, L, L_H\}$ and $\mathcal{H} = \{\varepsilon, \nu, \gamma, h\}$, respectively. Since we mainly care about the non-logarithmic dependencies of the mixing time on the underlying model parameters, and for the sake of legibility of expressions, we often suppress logarithmic dependencies on parameters in $\mathcal{M}$ by using the notation: for two quantities $\mathsf{x}, \mathsf{y} \in \mathbb{R}$, we write $\mathsf{x} = \widetilde{\mathcal{O}}(\mathsf{y})$ if there exists $C > 0$ depending at most logarithmically on any parameter in $\mathcal{M}$ such that $\mathsf{x} \leq C\mathsf{y}$. The symbol $\mathcal{O}$ is defined similarly except that it expresses all logarithmic dependencies.

Assumption 3.4. *Regarding the user-tuned hyperparameters, let $0 < \varepsilon \leq 1/2$ and suppose $\nu \in \mathcal{P}(\mathbb{R}^d)$ such that $\log \nu(e^{H/8})$ depends at most polynomially on the model parameters, i.e., there exist constants $n_1, n_2, n_3, n_4 \in \mathbb{Z}$ such that*

$$\log \nu(e^{H/8}) \ = \ \widetilde{\mathcal{O}}\big(d^{n_1} K^{n_2} L^{n_3} L_H^{n_4}\big)\ . \tag{3.10}$$

Further, let $\gamma, h > 0$ satisfy

$$L^{1/2}\gamma^{-1} \ \leq \ 1/10 \quad \textit{and} \quad 2\gamma h \ \leq \ 1.5936\ ,$$

as well as $\log(1/h) = \widetilde{\mathcal{O}}(1)$.

Note that (3.10) and the last part of 3.4 pose no relevant restriction because exponential dependencies on model parameters of the quantities of interest are unrealistic.

Remark 3.5 (Possibilities to Relax the Assumptions)**.** There are several possibilities the assumptions made above can be relaxed while sustaining the mixing guarantees of Theorem 3.6. First, the global strong convexity assumption in 3.1 can be relaxed to asymptotic strong convexity by employing a more sophisticated coupling in Assumption 2.2 *(i)* as developed in [9, 20, 12]. However, the resulting contraction rates will depend on underlying parameters in a more intricate way. Second, as emphasized in Remark 2.4, both the global gradient and Hessian Lipschitz continuity in 3.2 and 3.3 as well as the global convexity in 3.1 can be replaced with local versions. In particular, convexity in a suitable shell suffices.

3.3 Mixing guarantees for MAKLA

We are now in position to state upper bounds on the mixing time of MAKLA as defined in (2.4) with μ given by (3.2).

Theorem 3.6. *Suppose Assumptions 3.1-3.4 hold. Then there exists $\overline{h} > 0$ with*

$$\begin{aligned}(\overline{h})^{-1} \;=\; \widetilde{\mathcal{O}}\Big[&(L^{1/2}\gamma^{-1})^{-1/2}\kappa\log(1/\varepsilon)\\ &\times\Big(L_H^{1/2}K^{-1/4}d^{3/4}\max\big((L^{1/2}\gamma^{-1})^{-2},(\kappa d)^{-1}\log\nu(e^{H/8})\big)^{3/4}\\ &\quad+L^{1/2}d^{1/2}\max\big((L^{1/2}\gamma^{-1})^{-2},(\kappa d)^{-1}\log\nu(e^{H/8})\big)^{1/2}\Big)\Big]\end{aligned}$$

such that for all $h \le \overline{h}$, it holds that

$$\tau_{\mathrm{mix}}(\varepsilon,\nu) \;=\; \widetilde{\mathcal{O}}\big(h^{-1}K^{-1}\gamma\log(1/\varepsilon)\big)\;.$$

For a fixed step size $h \le \overline{h}$, Theorem 3.6 guarantees that starting in ν, $\tau_{\mathrm{mix}}(\varepsilon,\nu)$ transition steps of MAKLA suffice to ensure ε-accuracy in TV. The assumptions on the initial distribution are minimal. In particular, cold start distributions are covered, i.e., $\nu = \delta_z$ for some $z \in \mathbb{R}^{2d}$.

Remark 3.7 (Mixing Guarantee)**.** Note that if

$$\gamma \;=\; \mathcal{O}(L^{1/2}) \qquad \text{and} \qquad \log\nu(e^{H/8}) \;=\; \widetilde{\mathcal{O}}(\kappa d)\;,$$

Theorem 3.6 asserts that for $h = \overline{h}$,

$$\tau_{\mathrm{mix}}(\varepsilon,\nu) \;=\; \widetilde{\mathcal{O}}\Big[\kappa^{3/2}\max\big(L_H^{1/2}(d/K)^{3/4},\, L^{1/2}(d/K)^{1/2}\big)\log^2(1/\varepsilon)\Big] \tag{3.11}$$

since in this case

$$(\overline{h})^{-1} \;=\; \widetilde{\mathcal{O}}\Big[\kappa^{1/2}L^{1/2}\max\big(L_H^{1/2}(d/K)^{3/4},\, L^{1/2}(d/K)^{1/2}\big)\log(1/\varepsilon)\Big]\;. \tag{3.12}$$

This choice of γ minimizes the mixing time upper bound while still satisfying 3.4. Moreover, the assumption on ν is mild; e.g., it is satisfied by all cold starts in $z \in \mathbb{R}^{2d}$ such that $H(z)/8 = \log\delta_z(e^{H/8}) = \mathcal{O}(\kappa d)$. To put this in perspective, note that the Gaussian measure $\nu = \mathcal{N}(0,A^{-1})\otimes\mathcal{N}(0,I_d)$ with energy $H(z) = \frac{1}{2}|v|^2 + \frac{1}{2}|A^{1/2}x|^2$ amounts to $\log\nu(e^{H/8}) = d\log(8/7)$.

Remark 3.8 (Dimension Dependence)**.** Remarkably, the dimension scaling obtained in (3.11) is optimal in the *high acceptance regime*, cf. Definition 2.1, from a cold start distribution as illustrated by

$$U(x) \;=\; \frac{1}{2}x\cdot\operatorname{diag}\big(2,1,\dots\big)x - \sin(x_1)\;. \tag{3.13}$$

Denote by e_1 the unit vector in the first component and consider the collection $\mathcal{C} = \{\delta_{(0,d^{1/2}e_1)}\pi^n : n \ge 0\}$ corresponding to a cold start in $(0,d^{1/2}e_1) \in \mathbb{R}^{2d}$. According to (2.3), π being in the high acceptance regime on $\mathcal{C}$ requires

$$h^{-1}\,\mathbb{P}\big(A(0,d^{1/2}e_1)^c\big) \;=\; \mathcal{O}(1)\;, \tag{3.14}$$

where we used that the mixing time $\tau_{\mathrm{mix}}^{exact}$, cf. (2.2), of the transition kernel of the kinetic Langevin diffusion over time h is of order h^{-1}. Expanding (4.34) shows that the energy error to leading order is

$$(H\circ\theta_h - H)(x,v) \;=\; \frac{h^3}{24}\big(\nabla^3U(x)v^{\otimes 3} - 6v\cdot\nabla^2U(x)\nabla U(x)\big) + \mathcal{O}(h^4)\;.$$

For U as in (3.13), it hence holds that

$$(H\circ\theta_h - H)(0, d^{1/2}e_1) \;=\; \frac{h^3}{24}\big(d^{3/2}+12d^{1/2}\big)+\mathcal{O}(h^4)\ .$$

Since the reject probability from cold start in $(0, d^{1/2}e_1)$ is given by

$$\mathbb{P}\big(A(0,d^{1/2}e_1)^c\big) \;=\; 1-\mathbb{E}e^{-(H\circ\theta_h-H)\circ O_{h/2}(\xi_1)(0,d^{1/2}e_1)^+}\ ,$$

and the OU step to leading order in h equals the identity, (3.14) implies

$$h^2d^{3/2} \;=\; \mathcal{O}(1)\ .$$

Remark 3.9 (Condition Number Dependence)**.** In Lemma 4.1, UKLA is shown to converge to its stationary distribution with rate $\rho\propto K\gamma^{-1}h$, which under 3.4 is at best κ^{-1} for h^{-1} of order $L^{1/2}$. This rate differs from the optimal rate obtained for the kinetic Langevin diffusion under warm start [17]. Passing to MAKLA via Theorem 2.3 further increases scaling in condition number. In (3.11), the additional $\kappa^{1/2}$ in front of the maximum is expected for Assumption 2.2 *(iii)* to hold with ρ and the energy error bounds of Lemma 4.6. However, due to the linear appearance of $\gamma h\mathfrak{H}$ in (3.15) (cf. Remark 4.8), Assumption 2.2 *(iii)* requires the extra $K^{-3/4}$ and $K^{-1/2}$ inside the maximum. At present, the optimal condition number dependence for either UKLA or MAKLA from a cold start distribution is not known.

Proof of Theorem 3.6. To invoke Theorem 2.3, it suffices to verify Assumption 2.2 for MAKLA, which as described below, relies on ingredients developed in §4.

Let $S=\mathbb{R}^{2d}$ with metric induced by the twisted norm $\|\cdot\|_{\mathsf{tw}}$ defined in (4.1). Define the domain $D=\{\mathcal{E}(z)\le R_U\}$ for some $R_U\ge 2$ to be determined momentarily and where $\mathcal{E}$ is the energy-like function defined in (4.33). Note that $K|x|\le|\nabla U(x)|$ by 3.1 and 3.2, and hence,

$$\|z\|_{\mathsf{tw}}^2 \overset{(4.5)}{\le} \frac{17}{16}\big(|v|^2+\gamma^2|x|^2\big) \;\le\; \frac{17}{16}\frac{L\gamma^2}{K^2}\mathcal{E}(z)\ ,$$

where in the last step we used 3.4 to factor out $L\gamma^2/K^2\ge 36\kappa^2\ge 1$. Thus,

$$\operatorname{diam}_{\|\cdot\|_{\mathsf{tw}}}(D) \;\le\; 2\sup_{z\in D}\|z\|_{\mathsf{tw}} \;\le\; 3\frac{L^{1/2}\gamma}{K}R_U^{1/2} \;=:\; R\ ,$$

which specifies R in Assumption 2.2.

Regarding the unadjusted transition kernel,

- Assumption 2.2 *(i)* holds by Lemma 4.1 and 3.4 with rate

$$\rho \;=\; \frac{\gamma h}{34\sqrt{e}}\min\big(K\gamma^{-2},1\big) \;=\; \frac{K\gamma^{-1}h}{34\sqrt{e}}\ ;\quad\text{and,}$$

- Assumption 2.2 *(ii)* holds by Lemma 4.2 and (4.5) with

$$C_{Reg} \;=\; 14\big((\gamma h)^{-3/2}+\gamma^{-1}L_Hd^{1/2}h^2\big)\ .$$

This completes the verification of Assumption 2.2 *(i)* and *(ii)*.

It remains to verify Assumption 2.2 *(iii)* and *(iv)*, which concern the adjusted transition kernel. To this end, the epoch of transition steps $\mathfrak{E}$ and the total number of transition steps $\mathfrak{H}$ play a pivotal role. Assumption 3.4 implies $\log(\gamma C_{Reg})=\widetilde{\mathcal{O}}(1)$. Thus,

$$\mathfrak{E} \;=\; \widetilde{\mathcal{O}}\big(K^{-1}\gamma h^{-1}\log R_U\big) \quad\text{and}\quad \mathfrak{H} \;=\; \widetilde{\mathcal{O}}\big(K^{-1}\gamma h^{-1}\log R_U\log(1/\varepsilon)\big)\ .$$

Since *(iii)* depends on R_U, which needs to be chosen sufficiently large for the exit probability bound in *(iv)* to hold, we first verify *(iv)*.

- To verify Assumption 2.2 *(iv)*, we invoke Lemma 4.7 as follows. By Lemma 4.6, $C_{\Delta H} = 4L$ and $k = 2$ in Lemma 4.7. Moreover, (4.41) holds due to 3.4. Define $\overline{h}_1 > 0$ to saturate the bound $400L\mathfrak{H}h^2 \leq 1$ and let $h \leq \overline{h}_1$. We now select R_U to counter-saturate the bound
$$R_U \;\geq\; 32\Big[\gamma h \mathfrak{H} d + \log\Big(\frac{4}{\varepsilon}\max\big(\nu(e^{H/8}), (2\kappa)^{d/2}\big)\Big)\Big]\,, \tag{3.15}$$
where we inserted $\mu(e^{H/8}) \leq (2\kappa)^{d/2}$ by 3.1 and 3.2. By Lemma 4.7, this choice of R_U ensures Assumption 2.2 *(iv)* to hold starting from both ν and μ. Since the right hand side of (3.15) depends logarithmically on R_U, note that
$$R_U \;=\; \widetilde{\mathcal{O}}\Big(\max\big(K^{-1}\gamma^2 d, \log\nu(e^{H/8})\big)\log(1/\varepsilon)\Big)\,,$$
which implies $(\overline{h}_1)^{-1} \;=\; \widetilde{\mathcal{O}}\big(\kappa\gamma\log R_U \log(1/\varepsilon)\big) \;=\; \widetilde{\mathcal{O}}\big(\kappa\gamma\log(1/\varepsilon)\big)$.
- Finally, we verify Assumption 2.2 *(iii)*. Leveraging: (a) the higher order bound of Lemma 4.6; (b) the bounds
$$\begin{aligned} &\mathbb{E}\,\mathcal{E}(O_{h/2}(\xi_1)(z)) \;\leq\; \mathcal{E}(z) + \gamma h d\,, \quad \text{as well as}\\ &\mathbb{E}\,\mathcal{E}(O_{h/2}(\xi_1)(z))^{3/2} \;\leq\; 4\big(\mathcal{E}(z)^{3/2} + 3(\gamma h d)^{3/2}\big) \end{aligned}$$
that each hold for all $z \in \mathbb{R}^{2d}$; and (c) the definition of D yields
$$\begin{aligned} \mathfrak{E}\sup_{z\in D}\mathbb{P}(A(z)^c) \;&\leq\; \mathfrak{E}\sup_{z\in D}\mathbb{E}|\Delta H \circ O_{h/2}(\xi_1)(z)|\\ &\leq\; \mathfrak{E}h^3\sup_{z\in D}\big(2L_H\mathbb{E}\,\mathcal{E}(O_{h/2}(\xi_1)(z))^{3/2} + L^{3/2}\mathbb{E}\,\mathcal{E}(O_{h/2}(\xi_1)(z))\big)\\ &\leq\; \mathfrak{E}h^3\Big(8L_H\big(R_U^{3/2} + 3(\gamma h d)^{3/2}\big) + L^{3/2}\big(R_U + \gamma h d\big)\Big)\,. \end{aligned}$$
Inserting $\mathfrak{E}$ and R_U, and using that $\gamma h d = \mathcal{O}(K^{-1}\gamma^2 d)$ which holds by 3.4 and $K \leq L$, shows that there exists $\overline{h}_2 > 0$ such that the last display is bounded by $(3e)^{-1}$ for all $h \leq \overline{h}_2$ with
$$\begin{aligned} (\overline{h}_2)^{-1} \;=\; \widetilde{\mathcal{O}}\Big[&L_H^{1/2}K^{-1/4}\kappa d^{3/4}\log^{3/4}(1/\varepsilon)(L^{1/2}\gamma^{-1})^{-1/2}\\ &\times\max\big((L^{1/2}\gamma^{-1})^{-2}, (\kappa d)^{-1}\log\nu(e^{H/8})\big)^{3/4}\\ &+ L^{1/2}\kappa d^{1/2}\log^{1/2}(1/\varepsilon)(L^{1/2}\gamma^{-1})^{-1/2}\\ &\times\max\big((L^{1/2}\gamma^{-1})^{-2}, (\kappa d)^{-1}\log\nu(e^{H/8})\big)^{1/2}\Big]\,. \end{aligned}$$

This completes the verification of Assumption 2.2 *(iii)* and *(iv)*. To finish, set $\overline{h} = \min(\overline{h}_1, \overline{h}_2)$ which satisfies, by 3.4,
$$\begin{aligned} (\overline{h})^{-1} \;=\; \widetilde{\mathcal{O}}\Big[&(L^{1/2}\gamma^{-1})^{-1/2}\kappa\log(1/\varepsilon)\\ &\times\Big(L_H^{1/2}K^{-1/4}d^{3/4}\max\big((L^{1/2}\gamma^{-1})^{-2}, (\kappa d)^{-1}\log\nu(e^{H/8})\big)^{3/4}\\ &+ L^{1/2}d^{1/2}\max\big((L^{1/2}\gamma^{-1})^{-2}, (\kappa d)^{-1}\log\nu(e^{H/8})\big)^{1/2}\Big)\Big]\,. \end{aligned}$$

Since Assumption 2.2 holds, Theorem 3.6 now follows by invoking Theorem 2.3. □

4 Key ingredients of the Proof of Theorem 3.6

The ingredients needed in the proof of Theorem 3.6 are developed in this section.

4.1 Verifying Assumption 2.2 *(i)*: contractive coupling for UKLA

Here we use a coupling Π^u_{Contr} of two copies of UKLA starting from different initial distributions to demonstrate L^1-Wasserstein contractivity with respect to a *twisted metric* induced by the *twisted norm* on $\mathbb{R}^{2d}$

$$\|(x,v)\|^2_{\mathsf{tw}} := \alpha\,|x|^2 + \beta\langle x,v\rangle + |v|^2 \text{ where } \alpha = \frac{\beta}{h}\sinh(\frac{\gamma h}{2}) \text{ and } \beta = \frac{\gamma}{4}\,. \tag{4.1}$$

By using the elementary inequality

$$\mathsf{x} \le \sinh(\mathsf{x}) \le \frac{6}{5}\mathsf{x} \quad \text{valid for all } \mathsf{x}\in[0,1], \tag{4.2}$$

note that

$$\frac{1}{8}\gamma^2 \le \alpha \le \frac{3}{20}\gamma^2\,. \tag{4.3}$$

Hence, the twisted norm compares to the *untwisted norm*

$$\|(x,v)\|^2 := \gamma^2|x|^2 + |v|^2 \tag{4.4}$$

via

$$\frac{1}{16}\|(x,v)\|^2 \le \|(x,v)\|^2_{\mathsf{tw}} \le \frac{17}{16}\|(x,v)\|^2\,. \tag{4.5}$$

Lemma 4.1. *Suppose that Assumptions 3.1, 3.2, and 3.4 hold. Then, for all* $z=(x,v),\tilde{z}=(\tilde{x},\tilde{v})\in\mathbb{R}^{2d}$ *and* $\mathsf{a}_1,\mathsf{a}_2\in\mathbb{R}^d$*, it holds that*

$$\begin{aligned}&\left\|O_{h/2}(\mathsf{a}_2)\circ\theta_h\circ O_{h/2}(\mathsf{a}_1)(z) - O_{h/2}(\mathsf{a}_2)\circ\theta_h\circ O_{h/2}(\mathsf{a}_1)(\tilde{z})\right\|_{\mathsf{tw}}\\ &\quad\le (1-c\,h)\ \|z-\tilde{z}\|_{\mathsf{tw}} \quad \textit{where} \quad c = \frac{1}{34e^{\frac{1}{2}}}\min\left(K\gamma^{-1},\gamma\right)\,.\end{aligned}$$

Under similar assumptions, variants of Lemma 4.1 have beeen proven elsewhere in the literature; see, e.g., [65, 78, 55]. For the convenience of the reader, however, a complete proof is given below. As emphasized in previous works, a key ingredient in the proof of Lemma 4.1 is the *co-coercivity* property of ∇U, which in terms of the potential force F can be written as

$$|F(x_1)-F(x_2)|^2 \le -L\,\langle F(x_1)-F(x_2), x_1-x_2\rangle \quad \text{for all } x_1,x_2\in\mathbb{R}^d\,. \tag{4.6}$$

This holds if U is continuously differentiable, convex, and L-gradient Lipschitz [69, Theorem 2.1.10]. Additionally, the following elementary inequality is used

$$e^{-\mathsf{x}} \le 1-\frac{\mathsf{x}}{2} \qquad \text{valid for all } \mathsf{x}\in[0,1.5936]. \tag{4.7}$$

Proof. It is notationally convenient to define

$$\begin{aligned}&Z := O_{h/2}(\mathsf{a}_2)\circ\tilde{\theta}_h\circ O_{h/2}(\mathsf{a}_1)(z)\,,\quad \tilde{Z} := O_{h/2}(\mathsf{a}_2)\circ\tilde{\theta}_h\circ O_{h/2}(\mathsf{a}_1)(\tilde{z})\,,\\ &\Delta := \|Z-\tilde{Z}\|^2_{\mathsf{tw}} - \|z-\tilde{z}\|^2_{\mathsf{tw}}\,,\quad \zeta := x-\tilde{x}\,,\quad \omega := v-\tilde{v}\,,\\ &v_O := e^{-\frac{\gamma h}{2}}\,v + \sqrt{1-e^{-\gamma h}}\mathsf{a}_1\,,\quad \tilde{v}_O := e^{-\frac{\gamma h}{2}}\,\tilde{v} + \sqrt{1-e^{-\gamma h}}\mathsf{a}_1\,,\quad \hat{\omega} := e^{-\frac{\gamma h}{2}}\,\omega\,,\\ &\zeta^\star := \zeta + \frac{h}{2}\hat{\omega}\,,\ \ \Delta F^\star := F(x+\frac{h}{2}v_O) - F(\tilde{x}+\frac{h}{2}\tilde{v}_O)\,,\ \text{ and } \Phi^\star := -\langle\Delta F^\star,\zeta^\star\rangle\,.\end{aligned}$$

By definition of the OABAO scheme in (3.6),

$$Z-\tilde{Z} = \left(\zeta + h\hat{\omega} + \frac{h^2}{2}\Delta F^\star,\ e^{-\frac{\gamma h}{2}}(\hat{\omega}+h\Delta F^\star)\right)\,.$$

Inserting this difference into Δ yields

$$\begin{aligned}\Delta \;=\; &\alpha\left(-h^2\Phi^\star + h^2|\hat\omega|^2 + \frac{h^3}{2}\langle\hat\omega,\Delta F^\star\rangle + \frac{h^4}{4}|\Delta F^\star|^2\right)\\ &+\beta e^{-\frac{\gamma h}{2}}\left(-h\Phi^\star + h|\hat\omega|^2 + h^2\langle\hat\omega,\Delta F^\star\rangle + \frac{h^3}{2}|\Delta F^\star|^2\right)\\ &+e^{-\gamma h}\left((e^{-\gamma h}-e^{\gamma h})|\omega|^2 + 2h\langle\hat\omega,\Delta F^\star\rangle + h^2|\Delta F^\star|^2\right)\;.\end{aligned} \tag{4.8}$$

Note that the cross-terms involving $\langle\zeta,\omega\rangle$ vanish by definition of α in (4.1). Applying Young's inequality with parameters $\delta_1,\delta_2,\delta_3>0$ yields

$$\begin{aligned}\Delta \;\le\; &\alpha\left(-h^2\Phi^\star + \left(h^2+\frac{h^3}{4\delta_1}\right)|\hat\omega|^2 + \left(\frac{h^3\delta_1}{4}+\frac{h^4}{4}\right)|\Delta F^\star|^2\right)\\ &+\beta e^{-\frac{\gamma h}{2}}\left(-h\Phi^\star + \left(h+\frac{h^2}{2\delta_2}\right)|\hat\omega|^2 + \left(\frac{h^2\delta_2}{2}+\frac{h^3}{2}\right)|\Delta F^\star|^2\right)\\ &+e^{-\gamma h}\left((e^{-\gamma h}-e^{\gamma h})|\omega|^2 + \frac{h}{\delta_3}|\hat\omega|^2 + \left(h\delta_3+h^2\right)|\Delta F^\star|^2\right)\;.\end{aligned} \tag{4.9}$$

By the co-coercivity property in (4.6) evaluated at $x_1 = x+\frac{h}{2}v_O$ and $x_2=\tilde x+\frac{h}{2}\tilde v_O$, it holds that $|\Delta F^\star|^2\le L\,\Phi^\star$. Inserting this bound into (4.9) yields

$$\begin{aligned}\Delta \;&\le\; \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}\,,\quad\text{where} \qquad (4.10)\\ \mathrm{I} \;&:=\; -\frac12\beta h e^{-\frac{\gamma h}{2}}\Phi^\star - \sinh(\gamma h)|\hat\omega|^2\,,\\ \mathrm{II} \;&:=\; \alpha\left(\frac{Lh^4}{4}+\frac{Lh^3\delta_1}{4}-h^2\right)\Phi^\star\,,\\ \mathrm{III} \;&:=\; \left(\frac{\beta e^{-\frac{\gamma h}{2}}}{2}\left(Lh^3+Lh^2\delta_2\right) + e^{-\gamma h}L(h^2+h\delta_3) - \frac12\beta h e^{-\gamma h/2}\right)\Phi^\star\,,\\ \mathrm{IV} \;&:=\; \left(\alpha\left(h^2+\frac{h^3}{4\delta_1}\right)+\beta e^{-\frac{\gamma h}{2}}\left(h+\frac{h^2}{2\delta_2}\right)+e^{-\gamma h}\frac{h}{\delta_3}-\sinh(\gamma h)\right)|\hat\omega|^2\,.\end{aligned}$$

Below in (4.11)-(4.13), we show that II-IV are non-positive, and hence, $\Delta\le\mathrm{I}$. In particular, choosing $\delta_1=h$ yields

$$\mathrm{II} \;=\; \alpha h^2\left(Lh^2\left(\frac14+\frac{\delta_1}{2h}\right)-1\right)\Phi^\star \;\overset{3.4}{\le}\; \alpha h^2\left(\frac16\left(\frac14+\frac12\right)-1\right)\Phi^\star \;\le\; 0\,. \tag{4.11}$$

Choosing $\delta_2=2h$, $\delta_3=2\gamma^{-1}$, and by definition of α and β in (4.1),

$$\begin{aligned}\mathrm{III} \;&\le\; \frac{\gamma h}{4}e^{-\gamma h/2}\left(\frac32 Lh^2 + 4L\gamma^{-2}(\gamma h+2) - \frac12\right)\Phi^\star\\ &\overset{3.4}{\le}\; \frac{\gamma h}{4}e^{-\gamma h/2}\left(L\gamma^{-2}\left(\frac32+12\right)-\frac12\right)\Phi^\star \;\overset{3.4}{\le}\; \frac{\gamma h}{4}e^{-\gamma h/2}\left(\frac1{36}\frac{27}{2}-\frac12\right)\Phi^\star \;\le\; 0\,,\end{aligned} \tag{4.12}$$

as well as

$$\begin{aligned}\mathrm{IV} \;&\overset{(4.3)}{\le}\; \gamma h\left(\frac{3}{20}\gamma h\left(1+\frac{h}{4\delta_1}\right)+\frac14\left(1+\frac{h}{2\delta_2}\right)+\frac{\gamma^{-1}}{\delta_3}-1\right)|\hat\omega|^2\\ &\le\; \gamma h\left(\frac{3}{20}\gamma h\left(1+\frac14\right)+\frac14\left(1+\frac14\right)-\frac12\right)|\hat\omega|^2 \;\overset{3.4}{\le}\; 0\,.\end{aligned} \tag{4.13}$$

Combining the above, and by definition of β in (4.1),

$$\begin{aligned} \Delta \ \leq \ \mathrm{I} &\overset{3.1}{\leq} -\frac{1}{2}\beta h e^{-\frac{\gamma h}{2}} K|\zeta^\star|^2 - \sinh(\gamma h) e^{-\gamma h}|\omega|^2 \\ &\overset{(4.7)}{\leq} -\min\Big(K\gamma^{-1} h e^{-\frac{\gamma h}{2}}, \gamma h\Big) \big(\frac{1}{8}\gamma^2|\zeta^\star|^2 + \frac{1}{2}|\omega|^2\big) \\ &\leq -\min\Big(K\gamma^{-1} h e^{-\frac{\gamma h}{2}}, \gamma h\Big) \big(\frac{1}{16}\gamma^2|\zeta|^2 + \left(\frac{1}{2} - \frac{1}{16}\gamma^2 h^2\right)|\omega|^2\big) \\ &\overset{3.4}{\leq} -\frac{1}{16} e^{-1/2} \min\left(K\gamma^{-1} h, \gamma h\right) \left(\gamma^2|\zeta|^2 + |\omega|^2\right) \\ &\overset{(4.5)}{\leq} -\frac{1}{17} e^{-1/2} \min\left(K\gamma^{-1}, \gamma\right) h \left\|(\zeta,\omega)\right\|_{\mathsf{tw}}^2 \ . \end{aligned} \tag{4.14}$$

Thus, $\left\|Z - \tilde{Z}\right\|_{\mathsf{tw}} \leq (1 - c\,h)\left\|z - \tilde{z}\right\|_{\mathsf{tw}}$ with $c = (1/34)e^{-1/2}\min\left(K\gamma^{-1}, \gamma\right)$ — as required. □

4.2 Verifying Assumption 2.2 *(ii)*: one-shot coupling for UKLA

By using a one-shot coupling Π^u_{Reg}, cf. [72, 59, 42, 65, 8], the next lemma proves that the transition kernel of UKLA has a regularizing effect. Throughout this section, for any $z = (x, v), \tilde{z} = (\tilde{x}, \tilde{v}) \in \mathbb{R}^{2d}$, $\mathsf{a}_1, \mathsf{a}_2 \in \mathbb{R}^d$, let $\Phi : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ be the *one-shot* map $\Phi : (\mathsf{a}_1, \mathsf{a}_2) \mapsto (\tilde{\mathsf{a}}_1, \tilde{\mathsf{a}}_2)$ implicitly defined by

$$O_{h/2}(\mathsf{a}_2) \circ \theta_h \circ O_{h/2}(\mathsf{a}_1)(z) = O_{h/2}(\tilde{\mathsf{a}}_2) \circ \theta_h \circ O_{h/2}(\tilde{\mathsf{a}}_1)(\tilde{z}) \ . \tag{4.15}$$

Note that Φ depends on both z and $\tilde{z}$.

Lemma 4.2. *Suppose Assumptions 3.2, 3.3, and 3.4 hold. Let $\xi_1, \xi_2 \sim \mathcal{N}(0, I_d)$ be independent. Then, for all $z, \tilde{z} \in \mathbb{R}^{2d}$, it holds that*

$$\begin{aligned} \|\delta_z \Pi^u_{Reg} - \delta_{\tilde{z}} \Pi^u_{Reg}\|_{\mathsf{TV}} \ &\leq \ \|\operatorname{Law}(\xi_1, \xi_2) - \operatorname{Law}(\Phi(\xi_1, \xi_2))\|_{\mathsf{TV}} & (4.16) \\ &\leq \ \frac{7}{2}\left(\frac{1}{(\gamma h)^{3/2}} + d^{1/2}\frac{L_H h^3}{\gamma h}\right)\|z - \tilde{z}\| \ . & (4.17) \end{aligned}$$

A closely related one-shot coupling result has recently been developed for an “OBABO” discretization of kinetic Langevin dynamics; see Proposition 3 and Proposition 22 of [65], and for extensions see [66, 16]. Although the upper bound in (4.17) degenerates as $h \searrow 0$, this degeneration manifests only logarithmically in the mixing time results for MAKLA; see Assumption 2.2 *(iii)*.

Proof. Since the map $(\xi_1, \xi_2) \mapsto O_{h/2}(\xi_2) \circ \theta_h \circ O_{h/2}(\xi_1)(z)$ is deterministic and measurable [59, Lemma 3],

$$\|\delta_z \Pi^u_{Reg} - \delta_{\tilde{z}} \Pi^u_{Reg}\|_{\mathsf{TV}} \ \leq \ \|\operatorname{Law}(\xi_1, \xi_2) - \operatorname{Law}(\Phi(\xi_1, \xi_2))\|_{\mathsf{TV}} \ ,$$

which gives (4.16). Inserting Lemmas 4.3, 4.4, and 4.5 into (4.16) gives (4.17). □

As already indicated, the following lemmas are used in the proof of Lemma 4.2.

Lemma 4.3. *Let $\xi \sim \mathcal{N}(0, \Sigma)$ where Σ is an $n \times n$ matrix and suppose that $\mathsf{F} : \mathbb{R}^n \to \mathbb{R}^n$ is an invertible and differentiable map. Then*

$$\begin{aligned} &\|\operatorname{Law}(\xi) - \operatorname{Law}(\mathsf{F}(\xi))\|_{\mathsf{TV}} \ \leq \\ &\frac{1}{2}\sqrt{\mathbb{E}\Big[|\Sigma^{-\frac{1}{2}}(\mathsf{F}(\xi) - \xi)|^2 + 2\operatorname{tr}(D\mathsf{F}(\xi) - I_n) - 2\log|\det D\mathsf{F}(\xi)|\Big]} \ . \end{aligned} \tag{4.18}$$

Proof of Lemma 4.3. The proof of this result is an extension of the proof of Lemma 15 in [8] to the case where the covariance matrix of the reference Gaussian measure is Σ. Since this is a small modification, a proof is omitted. □

Lemma 4.4. *Suppose Assumptions 3.2, 3.3, and 3.4 hold. Then, for any* $z = (x, v), \tilde{z} = (\tilde{x}, \tilde{v}) \in \mathbb{R}^{2d}$ *and* $\mathsf{a}_1, \mathsf{a}_2 \in \mathbb{R}^d$*,*

$$|\Phi(\mathsf{a}_1, \mathsf{a}_2) - (\mathsf{a}_1, \mathsf{a}_2)|^2 \;\le\; \frac{44}{\gamma^3 h^3}\, \|z - \tilde{z}\|^2 \,. \tag{4.19}$$

Lemma 4.5. *Suppose Assumptions 3.2, 3.3, and 3.4 hold. Then, for any* $z = (x, v), \tilde{z} = (\tilde{x}, \tilde{v}) \in \mathbb{R}^{2d}$ *and* $\mathsf{a}_1, \mathsf{a}_2 \in \mathbb{R}^d$*,*

$$\operatorname{tr}(D\Phi(\mathsf{a}_1, \mathsf{a}_2) - I_{2d}) - \log|\det D\Phi(\mathsf{a}_1, \mathsf{a}_2)| \;\le\; 2d\frac{(L_H h^3)^2}{\gamma^2 h^2}\|z - \tilde{z}\|^2 \,. \tag{4.20}$$

For the proofs of Lemmas 4.4 and 4.5, it is notationally convenient to define

$$\begin{aligned}
&\zeta \;:=\; x - \tilde{x} \,, \quad \omega \;:=\; v - \tilde{v} \,, \\
&v_O \;:=\; e^{-\frac{\gamma h}{2}}\, v + \sqrt{1 - e^{-\gamma h}}\mathsf{a}_1 \,, \quad \tilde{v}_O \;:=\; e^{-\frac{\gamma h}{2}}\, \tilde{v} + \sqrt{1 - e^{-\gamma h}}\tilde{\mathsf{a}}_1 \,, \\
&x^\star \;:=\; x + \frac{h}{2} v_O \,, \quad \tilde{x}^\star \;:=\; x + \frac{h}{2} v_O \,, \quad \zeta^\star \;:=\; x^\star - \tilde{x}^\star \,.
\end{aligned}$$

This elementary inequality is used in the proofs: by 3.4 and (4.7),

$$1 - e^{-\gamma h} \ge \frac{\gamma h}{2} \implies (1 - e^{-\gamma h})^{-1} \le \frac{2}{\gamma h} \,. \tag{4.21}$$

Since $\zeta^\star = \zeta + \frac{h}{2} e^{-\frac{\gamma h}{2}} \omega + \frac{h}{2}\sqrt{1 - e^{-\gamma h}}(\mathsf{a}_1 - \tilde{\mathsf{a}}_1)$,

$$|\zeta^\star|^2 \;\le\; 3\left(|\zeta|^2 + \frac{h^2}{4}|\omega|^2 + \frac{h^2}{4}(1 - e^{-\gamma h})|\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2\right) \,. \tag{4.22}$$

By inserting (4.24) from the calculation below into (4.22), and using the elementary inequality $1 - e^{-\mathsf{x}} \le \mathsf{x}$ valid for $\mathsf{x} > -1$, we obtain

$$|\zeta^\star|^2 \le 3\left(|\zeta|^2 + \frac{h^2}{4}|\omega|^2 + \frac{\gamma h^3}{4}|\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2\right) \overset{3.4}{\le} 8\gamma^{-2}\left(\gamma^2|\zeta|^2 + |\omega|^2\right) \,. \tag{4.23}$$

Proof of Lemma 4.4. By definition of the one-shot map in (4.15),

$$\begin{aligned}
|\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 &= \frac{1}{1 - e^{-\gamma h}}|\frac{1}{h}\zeta + e^{-\frac{\gamma h}{2}}\omega - \frac{h}{2}H_U^\star \zeta^\star|^2 \\
&\le\; \frac{3}{1 - e^{-\gamma h}}\left(\frac{1}{h^2}|\zeta|^2 + |\omega|^2 + \frac{h^2}{4}\,\|H_U^\star\|_{\mathsf{op}}^2\,|\zeta^\star|^2\right) \\
&\overset{3.2}{\le} \frac{3}{1 - e^{-\gamma h}}\left(\frac{1}{h^2}|\zeta|^2 + |\omega|^2 + \frac{L^2 h^2}{4}|\zeta^\star|^2\right) \\
&\overset{(4.22),(4.21)}{\le} \left(\frac{6}{\gamma^3 h^3} + \frac{9}{2}L^2 h \gamma^{-3}\right)\gamma^2|\zeta|^2 + \left(\frac{6}{\gamma h} + \frac{9}{8}L^2 h^3 \gamma^{-1}\right)|\omega|^2 + \frac{9}{16}L^2 h^4 |\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 \\
&\overset{3.4}{\implies} |\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 \;\le\; \frac{181}{30\gamma^3 h^3}\left(\gamma^2|\zeta|^2 + |\omega|^2\right)
\end{aligned} \tag{4.24}$$

Similarly, from (4.15),

$$
\begin{aligned}
|\mathsf{a}_2 - \tilde{\mathsf{a}}_2|^2 &= \frac{1}{1-e^{-\gamma h}} \left| e^{-\frac{\gamma h}{2}} \left(e^{-\frac{\gamma h}{2}} \omega + h\sqrt{1-e^{-\gamma h}} (\mathsf{a}_1 - \tilde{\mathsf{a}}_1) - h H_U^\star \zeta^\star \right) \right|^2 \\
&\le \frac{3}{1-e^{-\gamma h}} \left(|\omega|^2 + (1-e^{-\gamma h}) |\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 + h^2 |H_U^\star \zeta^\star|^2 \right) \\
&\overset{3.2}{\le} \frac{3}{1-e^{-\gamma h}} \left(L^2 h^2 |\zeta^\star|^2 + |\omega|^2 \right) + 3 |\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 \\
&\overset{(4.23)}{\le} \frac{3}{1-e^{-\gamma h}} \left(8 L^2 h^2 \gamma^{-2} \left(\gamma^2 |\zeta|^2 + |\omega|^2 \right) + |\omega|^2 \right) + 3 |\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 \\
&\overset{(4.21)}{\le} \frac{6}{\gamma h} \left(8 L^2 h^2 \gamma^{-2} \left(\gamma^2 |\zeta|^2 + |\omega|^2 \right) + |\omega|^2 \right) + 3 |\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 \\
&\overset{(4.24)}{\le} \frac{6}{\gamma h} \left(8 L^2 h^2 \gamma^{-2} \left(\gamma^2 |\zeta|^2 + |\omega|^2 \right) + |\omega|^2 \right) + \frac{181}{10 \gamma^3 h^3} \left(\gamma^2 |\zeta|^2 + |\omega|^2 \right) \\
&\overset{3.4}{\le} \frac{25}{\gamma^3 h^3} \left(\gamma^2 |\zeta|^2 + |\omega|^2 \right) \ . \qquad (4.25)
\end{aligned}
$$

Combining (4.24) and (4.25) yields,

$$
|\mathsf{a}_1 - \tilde{\mathsf{a}}_1|^2 + |\mathsf{a}_2 - \tilde{\mathsf{a}}_2|^2 \le \frac{44}{\gamma^3 h^3} \left(\gamma^2 |\zeta|^2 + |\omega|^2 \right)
$$

as required. □

Proof of Lemma 4.5. Since, by definition of the one-shot map in (4.15),

$$
\frac{\partial \tilde{\mathsf{a}}_2}{\partial \mathsf{a}_2} = I_d \qquad \text{and} \qquad \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_2} = 0 \ ,
$$

it follows that $D\Phi = \begin{pmatrix} \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} & 0 \\ \frac{\partial \tilde{\mathsf{a}}_2}{\partial \mathsf{a}_1} & I_d \end{pmatrix}$, and hence,

$$
\det D\Phi(\mathsf{a}_1, \mathsf{a}_2) = \det \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} \cdot \det \frac{\partial \tilde{\mathsf{a}}_2}{\partial \mathsf{a}_2} = \det \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} \ , \quad \text{and} \tag{4.26}
$$

$$
\operatorname{tr}(D\Phi(\mathsf{a}_1, \mathsf{a}_2) - I_{2d}) = \operatorname{tr}(\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d) \ . \tag{4.27}
$$

Combining (4.26) and (4.27) yields

$$
\begin{aligned}
&\operatorname{tr}(D\Phi(\mathsf{a}_1, \mathsf{a}_2) - I_{2d}) - \log |\det D\Phi(\mathsf{a}_1, \mathsf{a}_2)| \\
&\quad = \operatorname{tr}(\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d) - \log |\det \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1}| \ .
\end{aligned} \tag{4.28}
$$

This observation motivates estimating $\left\| \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d \right\|_{\mathsf{op}}^2$.

From (4.15), note that

$$
\begin{aligned}
\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d &= \frac{1}{\sqrt{1-e^{-\gamma h}}} \frac{\partial}{\partial \mathsf{a}_1} \left[-\frac{h}{2} (\nabla U(x^\star) - \nabla U(\tilde{x}^\star)) \right] \\
&= -\frac{h^2}{4} \left(D^2 U(x^\star) - D^2 U(\tilde{x}^\star) \frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} \right) \\
&= -\frac{h^2}{4} \left(D^2 U(x^\star) - D^2 U(\tilde{x}^\star) \right) + \frac{h^2}{4} D^2 U(\tilde{x}^\star) \left(\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d \right) \ . \qquad (4.29)
\end{aligned}
$$

On the one hand, by 3.2,

$$\begin{aligned}&\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2 \le \frac{3h^4}{16}\sup_x \left\|D^2U(x)\right\|_{\mathsf{op}}^2 \left(2 + \left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2\right)\\ &\overset{3.2}{\le} \frac{3(Lh^2)^2}{16}\left(2 + \left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2\right) \overset{3.4}{\Longrightarrow} \left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}} \le \frac{1}{2}\ .\end{aligned} \tag{4.30}$$

On the other hand, by 3.3,

$$\begin{aligned}&\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2 \overset{3.3}{\le} \frac{h^4}{8}\left(L_H^2|\zeta^\star|^2 + \sup_x \left\|D^2U(x)\right\|_{\mathsf{op}}^2 \left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2\right)\\ &\overset{3.2}{\le} \frac{1}{8}\left(L_H^2h^4|\zeta^\star|^2 + (Lh^2)^2\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2\right)\\ &\overset{3.4}{\Longrightarrow} \left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2 \le \frac{1}{7}L_H^2h^4|\zeta^\star|^2 \overset{(4.23)}{\le} \frac{8}{7}\frac{(L_Hh^3)^2}{\gamma^2h^2}\left(\gamma^2|\zeta|^2 + |\omega|^2\right)\ .\end{aligned} \tag{4.31}$$

Combining (4.30) and (4.31) yields

$$\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}} \le \min\left(\frac{1}{2}, \sqrt{\frac{8}{7}}\frac{L_Hh^3}{\gamma h}\|z-\tilde{z}\|\right)\ .$$

Since $\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}} \le 1/2$, the spectral radius of $\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d$ does not exceed $1/2$. Therefore, we can invoke Theorem 1.1 of [77], to obtain

$$\begin{aligned}\operatorname{tr}\left(\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right) - \log\left|\det\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1}\right| &\le \frac{\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_F^2/2}{1-\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}}\\ &\le \left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_F^2 \le d\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}^2 \le 2d\frac{(L_Hh^3)^2}{\gamma^2h^2}\|z-\tilde{z}\|^2\ ,\end{aligned} \tag{4.32}$$

where in the second to last step we used $\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_F \le \sqrt{d}\left\|\frac{\partial \tilde{\mathsf{a}}_1}{\partial \mathsf{a}_1} - I_d\right\|_{\mathsf{op}}$. Inserting (4.32) into (4.28) gives the required result. □

4.3 Verifying Assumption 2.2 *(iii)*: energy error estimates

The following Lemma provides upper bounds for the energy error in terms of the energy-like function $\mathcal{E}:\mathbb{R}^{2d}\to\mathbb{R}$ defined by

$$\mathcal{E}(z) \ = \ |v|^2 + L^{-1}|\nabla U(x)|^2\ . \tag{4.33}$$

As the isotropic Gaussian case suggests, where $U(x)=(L/2)|x|^2$, the scaling in (4.33) is natural, since in that case: if $(X,V)\sim\mu$ then $L^{-1/2}\nabla U(X)=L^{1/2}X$ and V are both standard normally distributed.

Lemma 4.6. *Suppose that Assumption 3.2 holds and let $Lh^2\le 1$. Then, the energy error $\Delta H = H\circ\theta_h - H$ with θ_h as in (3.7) satisfies*

$$|\Delta H(z)| \ \le \ 4Lh^2\mathcal{E}(z) \quad \textit{for all } z\in\mathbb{R}^{2d}.$$

If additionally Assumption 3.3 holds, then

$$|\Delta H(z)| \ \le \ 2L_Hh^3\mathcal{E}(z)^{3/2} + L^{3/2}h^3\mathcal{E}(z) \quad \textit{for all } z\in\mathbb{R}^{2d}.$$

Proof. For $t \in [0, h]$, introduce the linear interpolation

$$(x_t, v_t) \;=\; \big(x + tv - \frac{th}{2}\nabla U(x^*),\; v - t\nabla U(x^*)\big)\;,$$

where $x^* = x + \frac{h}{2}v$. Note that $(x_0, v_0) = (x, v)$ and $(x_h, v_h) = \theta_h(x, v)$. Therefore, the energy error can be written as

$$\Delta H(x,v) \;=\; (H \circ \theta_h - H)(x,v) \;=\; \int_0^h \frac{\mathrm{d}}{\mathrm{d}t} H(x_t, v_t)\,\mathrm{d}t\;. \tag{4.34}$$

Expanding the integrand using $H(x,v) = |v|^2/2 + U(x)$ yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(x_t,v_t) \;&=\; v_t \cdot \frac{\mathrm{d}}{\mathrm{d}t}v_t + \nabla U(x_t)\cdot\frac{\mathrm{d}}{\mathrm{d}t}x_t \\ &= -\big(v - t\nabla U(x^*)\big)\cdot\nabla U(x^*) + \nabla U\big(x + tv - \frac{th}{2}\nabla U(x^*)\big)\cdot\big(v - \frac{h}{2}\nabla U(x^*)\big)\;.\end{aligned} \tag{4.35}$$

Let $t \geq 0$ and $a \in \mathbb{R}^d$. To further simplify the last display, we expand

$$\nabla U(x + ta) \;=\; \nabla U(x) + \int_0^t \nabla^2 U(x + sa)a\,\mathrm{d}s$$

and

$$\nabla U(x + ta) \;=\; \nabla U(x) + t\nabla^2 U(x)a + \int_0^t (t-s)\nabla^3 U(x+sa) : a^{\otimes 2}\,\mathrm{d}s\;.$$

In particular,

$$\nabla U(x^*) \;=\; \nabla U(x) + I_1^* \tag{4.36}$$

with $I_1^* = \frac{1}{2}\int_0^h \nabla^2 U\big(x + \frac{s}{2}v\big)v\,\mathrm{d}s$ satisfying $|I_1^*| \leq \frac{1}{2}Lh|v|$, and

$$\nabla U(x^*) \;=\; \nabla U(x) + \frac{h}{2}\nabla^2 U(x)v + I_2^* \tag{4.37}$$

with $I_2^* = \frac{1}{4}\int_0^h (h-s)\nabla^3 U\big(x + \frac{s}{2}v\big) : v^{\otimes 2}\,\mathrm{d}s$ satisfying $|I_2^*| \leq \frac{1}{8}L_H h^2 |v|^2$. Further,

$$\nabla U\big(x + tv - \frac{th}{2}\nabla U(x^*)\big) \;=\; \nabla U(x) + I_1(t) \tag{4.38}$$

with $I_1(t) = \int_0^t \nabla^2 U\big(x + sv - \frac{sh}{2}\nabla U(x^*)\big)\big(v - \frac{h}{2}\nabla U(x^*)\big)\,\mathrm{d}s$ satisfying $|I_1(t)| \leq Lt\big|v - \frac{h}{2}\nabla U(x^*)\big|$, and

$$\nabla U\big(x + tv - \frac{th}{2}\nabla U(x^*)\big) \;=\; \nabla U(x) + t\nabla^2 U(x)\big(v - \frac{h}{2}\nabla U(x^*)\big) + I_2(t) \tag{4.39}$$

with $I_2(t) = \int_0^t (t-s)\nabla^3 U\big(x + sv - \frac{sh}{2}\nabla U(x^*)\big) : \big(v - \frac{h}{2}\nabla U(x^*)\big)^{\otimes 2}\,\mathrm{d}s$ satisfying $|I_2(t)| \leq \frac{1}{2}L_H t^2 \big|v - \frac{h}{2}\nabla U(x^*)\big|^2$.

Using the higher and lower order expansions will give the higher and lower order bound, respectively, due to more or less cancellation. For the lower order bound, inserting (4.36) and (4.38) into (4.35) yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H(x_t,v_t) \;&=\; (t - h/2)|\nabla U(x)|^2 - I_1^*\cdot\big(v + \frac{1}{2}(h - 4t)\nabla U(x)\big) + t|I_1^*|^2 \\ &\quad + I_1(t)\cdot\big(v - \frac{h}{2}\nabla U(x^*)\big)\;.\end{aligned}$$

Inserting this expression and the bounds on I_1^* and $I_1(t)$ back into (4.34) shows

$$\begin{aligned}|\Delta H(x,v)| \;&\le\; \frac{1}{2}Lh^2|v|\big|v-\frac{h}{2}\nabla U(x)\big| + \frac{1}{8}L^2h^4|v|^2 + \frac{1}{2}Lh^2\big|v-\frac{h}{2}\nabla U(x^*)\big|^2 \\ &\le\; 4Lh^2\mathcal{E}(z)\ ,\end{aligned}$$

where, besides $Lh^2 \le 1$, we used that due to (4.36)

$$\big|v-\frac{h}{2}\nabla U(x^*)\big| \;\le\; |v| + \frac{h}{2}|\nabla U(x^*)| \overset{(4.36)}{\le} \big(1+\frac{1}{4}Lh^2\big)|v| + \frac{h}{2}|\nabla U(x)| \;\le\; 2\mathcal{E}(z)^{1/2}$$

and that a similar bound holds for $\big|v-\frac{h}{2}\nabla U(x)\big|$. Repeating the calculation with the higher order expansions (4.37) and (4.39) gives

$$\begin{aligned}|\Delta H(x,v)| \;\le\;& -h\big(v-\frac{h}{2}\nabla U(x^*)\big)\cdot I_2^* - \frac{h^3}{4}v\cdot\nabla^2U(x)\nabla U(x^*) \\ &+ \frac{h^4}{8}\nabla U(x^*)\cdot\nabla^2U(x)\nabla U(x^*) + \int_0^h I_2(t)\,\mathrm{d}t\cdot\big(v-\frac{h}{2}\nabla U(x^*)\big) \\ \le&\ \big(\frac{4}{3}+\frac{1}{4}\big)L_Hh^3\mathcal{E}(z)^{3/2} + \frac{1}{4}Lh^3\big(|v|+\frac{h}{2}|\nabla U(x^*)|\big)|\nabla U(x^*)| \\ \le&\ 2L_Hh^3\mathcal{E}(z)^{3/2} + L^{3/2}h^3\mathcal{E}(z)\ ,\end{aligned}$$

as required. □

4.4 Verifying Assumption 2.2 *(iv)*: exit probability estimates for MAKLA

We now turn to the exit probability bound required in Assumption 2.2 *(iv)*. For some suitably large R_U and $\mathcal{E}$ as in (4.33), we show that the exit probability from

$$D \;=\; \big\{\mathcal{E}(z) \le R_U\big\}$$

is small over the total number of steps $\mathfrak{H}$ required to attain the desired TV convergence. More precisely, let $(Z_k)_{k\ge 0}$ be a copy of MAKLA started in an initial distribution ν and define the first exit time of the chain from D to be

$$T \;=\; \inf\big\{k\ge 0\ :\ Z_k\notin D\big\}\ .$$

The following lemma is general in the sense that it only assumes an energy error bound satisfied by, amongst other discretizations, θ_h as in (3.7).

Lemma 4.7. *Suppose Assumptions 3.1 and 3.2 hold and that the energy error $\Delta H = H\circ\theta_h - H$ satisfies*

$$|\Delta H(z)| \;\le\; C_{\Delta H}h^k\mathcal{E}(z) \tag{4.40}$$

for some $C_{\Delta H} > 0$, $k\ge 2$ and all $z\in\mathbb{R}^{2d}$. Let $h,\gamma > 0$ be such that

$$(1+25C_{\Delta H}h^k)^2\max(\gamma h, 1) \;\le\; 4\ . \tag{4.41}$$

Then, for $\mathfrak{H}, R_U > 0$, it holds that

$$\mathbb{P}\big(T\le\mathfrak{H}\big) \;\le\; \exp\Big(\frac{1+25C_{\Delta H}h^k}{4}\gamma h\mathfrak{H}d - \frac{1-50C_{\Delta H}\mathfrak{H}h^k}{16}R_U\Big)\nu\big(e^{H/8}\big)\ . \tag{4.42}$$

If additionally $100C_{\Delta H}\mathfrak{H}h^k\le 1$, then $\mathbb{P}(T\le\mathfrak{H})\le\varepsilon/4$ for $\varepsilon > 0$ if

$$R_U \;\ge\; 32\big[\gamma h\mathfrak{H}d + \log\big(4\nu(e^{H/8})/\varepsilon\big)\big]\ . \tag{4.43}$$

Although the energy error bound (4.40) is assumed to hold globally for simplicity, it can be relaxed to hold in a neighborhood of D; cf. Remark 2.4.

Remark 4.8 (Effect of Velocity Flip)**.** The MAKLA transition step involves a velocity flip involution in the event of a rejection; cf. (3.8). This makes it tricky to construct a Foster-Lyapunov function exploiting the contractivity of the unadjusted kernel by incorporating a cross-term $x \cdot v$. The function used below does not involve such a cross-term, and as a consequence, the time horizon $h\mathfrak{H}$ enters linearly into (4.43). In contrast, similar bounds for the MALA transition step only require the radius to depend logarithmically on the time horizon [38, §6]. However, due to the wide availability of energy error bounds such as (4.40), the Foster-Lyapunov function presented here is a robust alternative.

Proof. Below, we will show that the Lyapunov function $e^{H/8}$ solves

$$\begin{cases} \mathcal{L}_h e^{H(z)/8} \ \le \ \big(e^{\lambda} - 1\big)\, e^{H(z)/8} & \text{for } z \in D, \\ e^{H(z)/8} \ \ge \ e^{R_U/16} & \text{for } z \in \partial D, \end{cases} \tag{4.44}$$

where $\mathcal{L}_h = \pi - \mathrm{id}$ is the generator of MAKLA and

$$\lambda \ = \ \frac{1}{8}[2(1 + 25C_{\Delta H}h^k)\gamma h d + 25 C_{\Delta H} h^k R_U] \ .$$

By Chernoff's inequality,

$$\begin{aligned} \mathbb{P}\big(T \le \mathfrak{H}\big) &\le \ \mathbb{E}\exp\Big(- \lambda(T - \mathfrak{H})\Big) \le \ \exp\Big(\lambda\mathfrak{H} - R_U/16\Big)\nu\big(e^{H/8}\big) \\ &\le \ \exp\Big(\frac{1}{4}(1 + 25C_{\Delta H}h^k)\gamma h\mathfrak{H}d - \frac{1}{16}(1 - 50C_{\Delta H}\mathfrak{H}h^k)R_U\Big)\nu\big(e^{H/8}\big) \ , \end{aligned}$$

where in the second to last step we used a maximum principle to upper bound the Laplace transform of the first exit time T by the solution $e^{H/8}$ of the boundary value problem in (4.44), i.e., $\mathbb{E}e^{-\lambda T}e^{R_U/16} \le \nu(e^{H/8})$; for details see, e.g., [39]. Therefore, if $100C_{\Delta H}\mathfrak{H}h^k \le 1$, it follows that $\mathbb{P}(T \le \mathfrak{H}) \le \varepsilon/4$ if

$$R_U \ \ge \ 32\big[\gamma h\mathfrak{H}d + \log\big(4\nu(e^{H/8})/\varepsilon\big)\big] \ .$$

We now turn to the proof that $e^{H/8}$ solves (4.44). On ∂D, the lower bound holds since $U(x) \ge \frac{1}{2L}|\nabla U(x)|^2$ by 3.2, and hence,

$$e^{H(z)/8} \ = \ e^{(|v|^2/2 + U(x))/8} \ \ge \ e^{\mathcal{E}(z)/16} \ = \ e^{R_U/16} \ .$$

Let $z = (x, v) \in D$ and set $\delta = 1/8$. Then

$$\mathcal{L}_h e^{H(z)/8} \ = \Big[\mathbb{E}\exp\Big(\delta\big[H \circ O_{h/2}(\xi_2) \circ \hat{\theta}_h(\mathcal{U}) \circ O_{h/2}(\xi_1)(z) - H(z)\big]\Big) - 1\Big]e^{H(z)/8} \ .$$

Therefore, the upper bound in (4.44) holds if

$$\begin{aligned} &\mathbb{E}\exp\Big(\delta\big[H \circ O_{h/2}(\xi_2) \circ \hat{\theta}_h(\mathcal{U}) \circ O_{h/2}(\xi_1)(z) - H(z)\big]\Big) \\ &\quad \le \ \exp\Big(\delta\big[2\big(1 + (1 + 3/\delta)C_{\Delta H}h^k\big)\gamma h d + (1 + 3/\delta)C_{\Delta H}h^k R_U\big]\Big) \ = \ e^{\lambda} \ . \end{aligned} \tag{4.45}$$

Treating the two outcomes of the Metropolis step separately yields

$$\begin{aligned} &\mathbb{E}\exp\Big(\delta\big[H \circ O_{h/2}(\xi_2) \circ \hat{\theta}_h(\mathcal{U}) \circ O_{h/2}(\xi_1)(z) - H(z)\big]\Big) \overset{(3.8)}{\le} \mathrm{I} + \mathrm{I\!I} \quad \text{where} \\ &\mathrm{I} = \mathbb{E}\exp\Big(\delta\big[H \circ O_{h/2}(\xi_2) \circ \theta_h \circ O_{h/2}(\xi_1)(z) - H(z)\big]\Big) \\ &\mathrm{I\!I} = \mathbb{E}\Big[\exp\Big(\delta\big[H \circ O_{h/2}(\xi_2) \circ \mathcal{S} \circ O_{h/2}(\xi_1)(z) - H(z)\big]\Big) \, ; \, A\big(O_{h/2}(\xi_1)(z)\big)^c\Big] \ . \end{aligned} \tag{4.46}$$

To bound I in (4.46), we use the elementary bound

$$\mathbb{E}_{\xi\sim\mathcal{N}(0,I_d)}\exp\bigl(b\cdot\xi+c|\xi|^2\bigr)\ \le\ \exp\bigl(|b|^2+2cd\bigr) \tag{4.47}$$

valid for $c\in[0,1/4]$ and $b\in\mathbb{R}^d$. In particular, applying (4.47) with $c=\gamma h\delta$ which is possible since $c\le 1/4$ by (4.41), implies that for all $z'=(x',v')\in\mathbb{R}^{2d}$ and for $\xi\sim\mathcal{N}(0,I_d)$

$$\begin{aligned}
&\mathbb{E}\exp\Bigl(\delta\bigl[H\circ O_{h/2}(\xi)(z')-H(z')\bigr]\Bigr)\\
&=\ \mathbb{E}\exp\Bigl(\delta\bigl[-\frac{1}{2}(1-e^{-\gamma h})|v'|^2+e^{-\gamma h/2}\sqrt{1-e^{-\gamma h}}v'\cdot\xi+\frac{1}{2}(1-e^{-\gamma h})|\xi|^2\bigr]\Bigr)\\
&\overset{(4.47)}{\le}\ \exp\Bigl(\delta\bigl[-\frac{1}{2}(1-2\delta e^{-\gamma h})(1-e^{-\gamma h})|v'|^2+(1-e^{-\gamma h})d\bigr]\Bigr)\ \le\ e^{\delta\gamma hd}\,.
\end{aligned} \tag{4.48}$$

Therefore, using $\Delta H=H\circ\theta_h-H$,

$$\begin{aligned}
\mathrm{I}\ &=\ \mathbb{E}\exp\Bigl(\delta\bigl[H\circ O_{h/2}(\xi_2)\circ\theta_h\circ O_{h/2}(\xi_1)(z)-H\circ\theta_h\circ O_{h/2}(\xi_1)(z)\\
&\qquad\qquad+\Delta H\circ O_{h/2}(\xi_1)(z)+H\circ O_{h/2}(\xi_1)(z)-H(z)\bigr]\Bigr)\\
&\overset{(4.48)}{\le}\ e^{\delta\gamma hd}\mathbb{E}\exp\Bigl(\delta\bigl[\Delta H\circ O_{h/2}(\xi_1)(z)+H\circ O_{h/2}(\xi_1)(z)-H(z)\bigr]\Bigr)\,.
\end{aligned} \tag{4.49}$$

Inserting the energy error bound (4.40) into the exponent of (4.49) then yields,

$$\begin{aligned}
&\Delta H\circ O_{h/2}(\xi_1)(z)+H\circ O_{h/2}(\xi_1)(z)-H(z)\\
&\le\ C_{\Delta H}h^k\mathcal{E}(O_{h/2}(\xi_1)(z))+H\circ O_{h/2}(\xi_1)(z)-H(z)\\
&\le\ C_{\Delta H}h^k\mathcal{E}(z)-\frac{1}{2}(1-e^{-\gamma h})|v|^2+(1+2C_{\Delta H}h^k)e^{-\gamma h/2}\sqrt{1-e^{-\gamma h}}v\cdot\xi_1\\
&\quad+\frac{1}{2}(1+2C_{\Delta H}h^k)\gamma h|\xi_1|^2\,.
\end{aligned}$$

Hence, inserting this bound back into (4.49), using $\mathcal{E}(z)\le R_U$, and applying (4.47) with $c=\delta(1+2C_{\Delta H}h^k)\gamma h/2$ which is possible since $c\le 1/4$ holds by (4.41), shows

$$\begin{aligned}
\mathrm{I}\ &\le\ \exp\Bigl(\delta\bigl[2(1+C_{\Delta H}h^k)\gamma hd+C_{\Delta H}h^kR_U\\
&\qquad-\frac{1}{2}\bigl(1-2\delta(1+2C_{\Delta H}h^k)^2e^{-\gamma h}\bigr)(1-e^{-\gamma h})|v|^2\bigr]\Bigr)\\
&\le\ \exp\Bigl(\delta\bigl[2(1+C_{\Delta H}h^k)\gamma hd+C_{\Delta H}h^kR_U\bigr]\Bigr)
\end{aligned} \tag{4.50}$$

since $2\delta(1+2C_{\Delta H}h^k)^2\le 1$ by (4.41). For II in (4.46), using $H\circ\mathcal{S}=H$,

$$\begin{aligned}
\mathrm{II}\ &=\ \mathbb{E}\Bigl[\mathbb{E}_{\xi_2}\exp\Bigl(\delta\bigl[H\circ O_{h/2}(\xi_2)\bigl(\mathcal{S}\circ O_{h/2}(\xi_1)(z)\bigr)-H\bigl(\mathcal{S}\circ O_{h/2}(\xi_1)(z)\bigr)\bigr]\Bigr)\\
&\qquad\times\exp\Bigl(\delta\bigl[H(O_{h/2}(\xi_1)(z))-H(z)\bigr]\Bigr)\,;\,A\bigl(O_{h/2}(\xi_1)(z)\bigr)^c\Bigr]\\
&\overset{(4.48)}{\le}\ e^{\delta\gamma hd}\mathbb{E}\Bigl[\exp\Bigl(\delta\bigl[H(O_{h/2}(\xi_1)(z))-H(z)\bigr]\Bigr)\,;\,A\bigl(O_{h/2}(\xi_1)(z)\bigr)^c\Bigr]\,.
\end{aligned}$$

We continue estimating the last display using Cauchy-Schwarz inequality combined with $\mathbb{P}(A(z)^c)=1-e^{\Delta H(z)^+}\le|\Delta H(z)|$, (4.48), and (4.40) to obtain

$$\begin{aligned}
\mathrm{II}\ &\le\ e^{\delta\gamma hd}\Bigl(\mathbb{E}e^{2\delta[H(O_{h/2}(\xi_1)(z))-H(z)]}\Bigr)^{1/2}\bigl(\mathbb{E}|\Delta H(O_{h/2}(\xi_1)(z))|^2\bigr)^{1/2}\\
&\overset{(4.48)}{\le}\ e^{2\delta\gamma hd}C_{\Delta H}h^k\bigl(\mathbb{E}\,\mathcal{E}(O_{h/2}(\xi_1)(z))^2\bigr)^{1/2}\le 3e^{2\delta\gamma hd}C_{\Delta H}h^k(R_U+2\gamma hd)\,,
\end{aligned} \tag{4.51}$$

where in the last step we used

$$\mathbb{E}\,\mathcal{E}(O_{h/2}(\xi_1)(z))^2 \;\le\; 4\mathbb{E}\big(\mathcal{E}(z)+\gamma h|\xi_1|^2\big)^2 \;\le\; 8\big(R_U^2+3(\gamma hd)^2\big)\ .$$

Inserting (4.50) and (4.51) into (4.46) and simplifying yields

$$\begin{aligned}
&\mathbb{E}\exp\Big(\delta\big[H\circ O_{h/2}(\xi_2)\circ\hat{\theta}_h(\mathcal{U})\circ O_{h/2}(\xi_1)(z)-H(z)\big]\Big)\\
&\le\; \exp\Big(\delta\big[2(1+C_{\Delta H}h^k)\gamma hd+C_{\Delta H}h^kR_U\big]\Big)+3e^{2\delta\gamma hd}C_{\Delta H}h^k(R_U+2\gamma hd)\\
&\le\; \exp\Big(\delta\big[2(1+C_{\Delta H}h^k)\gamma hd+C_{\Delta H}h^kR_U\big]\Big)\cdot\big(1+3C_{\Delta H}h^k(R_U+2\gamma hd)\big)\\
&\le\; \exp\Big(\delta\big[2(1+C_{\Delta H}h^k)\gamma hd+C_{\Delta H}h^kR_U\big]+\log\big(1+3C_{\Delta H}h^k(R_U+2\gamma hd)\big)\Big)\\
&\le\; \exp\Big(\delta\big[2\big(1+(1+3/\delta)C_{\Delta H}h^k\big)\gamma hd+(1+3/\delta)C_{\Delta H}h^kR_U\big]\Big)\;=\;e^{\lambda}\ ,
\end{aligned}$$

where we used $\log(1+\mathsf{x})\le\mathsf{x}$ valid for $\mathsf{x}\ge 0$. This proves (4.45) holds, and hence, $\exp(H(z)/8)$ solves (4.44), as required. □

References

[1] J. M. Altschuler and S. Chewi, *Faster high-accuracy log-concave sampling via algorithmic warm starts*, arXiv preprint arXiv:2302.10249 (2023). MR4720376

[2] C. Andrieu, A. Lee, and S. Livingstone, *A general perspective on the metropolis-hastings kernel*, arXiv preprint arXiv:2012.14881 (2020).

[3] A. Beskos, F. J. Pinski, J. M. Sanz-Serna, and A. M. Stuart, *Hybrid Monte-Carlo on Hilbert spaces*, Stochastic Processes and their Applications **121** (2011), no. 10, 2201–2230. MR2822774

[4] L. J. Billera and P. Diaconis, *A Geometric Interpretation of the Metropolis-Hastings Algorithm*, Statistical Science **16** (2001), no. 4, 335 – 339. MR1888448

[5] N. Bou-Rabee, *Time integrators for molecular dynamics*, Entropy **16** (2014), 138–162.

[6] N. Bou-Rabee and A. Eberle, *Two-scale coupling for preconditioned hamiltonian monte carlo in infinite dimensions*, Stochastics and Partial Differential Equations: Analysis and Computations **9** (2021), no. 1, 207–242. MR4218791

[7] N. Bou-Rabee and A. Eberle, *Couplings for Andersen Dynamics in High Dimension*, Ann. Inst. H. Poincaré Probab. Statist **58** (2022), no. 2, 916–944. MR4421613

[8] N. Bou-Rabee and A. Eberle, *Mixing time guarantees for unadjusted hamiltonian monte carlo*, Bernoulli **29** (2023), no. 1, 75–104. MR4497240

[9] N. Bou-Rabee, A. Eberle, and R. Zimmer, *Coupling and convergence for hamiltonian monte carlo*, Ann. Appl. Probab. **30** (2020), no. 3, 1209–1250. MR4133372

[10] N. Bou-Rabee and M. Hairer, *Non-asymptotic mixing of the MALA algorithm*, IMA J of Numer Anal **33** (2013), 80–110. MR3020951

[11] N. Bou-Rabee and J. M. Sanz-Serna, *Geometric integrators and the hamiltonian Monte Carlo method*, Acta Numerica **27** (2018), 113–206. MR3826507

[12] N. Bou-Rabee and K. Schuh, *Convergence of unadjusted Hamiltonian Monte Carlo for mean-field models*, Electronic Journal of Probability **28** (2023), no. none, 1 – 40. MR4610714

[13] N. Bou-Rabee and E. Vanden-Eijnden, *Pathwise accuracy and ergodicity of Metropolized integrators for SDEs*, Comm Pure and Appl Math **63** (2010), 655–696. MR2583309

[14] G. Bussi, D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling*, J Chem Phys **126** (2007), 014101.

[15] O. Butkovsky, *Subgeometric rates of convergence of Markov processes in the Wasserstein metric*, The Annals of Applied Probability **24** (2014), no. 2, 526 – 552. MR3178490

[16] E. Camrud, A. Durmus, P. Monmarché, and G. Stoltz, *Second order quantitative bounds for unadjusted generalized hamiltonian monte carlo*, arXiv preprint arXiv:2306.09513 (2023).

[17] Yu Cao, Jianfeng Lu, and Lihan Wang, *On explicit l^2-convergence rate estimate for underdamped langevin dynamics*, Archive for Rational Mechanics and Analysis **247** (2023), 90. MR4632836

[18] Y. Chen, R. Dwivedi, M. J. Wainwright, and B. Yu, *Fast mixing of metropolized hamiltonian monte carlo: Benefits of multi-step gradients*, Journal of Machine Learning Research **21** (2020), no. 92, 1–72. MR4119160

[19] Z. Chen and K. Gatmiry, *When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm?*, arXiv preprint arXiv:2304.04724 (2023).

[20] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan, *Sharp convergence rates for langevin dynamics in the nonconvex setting*, arXiv preprint arXiv:1805.01648 (2018).

[21] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, Conference On Learning Theory, 2018, pp. 300–323.

[22] S. Chewi, C. Lu, K. Ahn, X. Cheng, T. L. Gouic, and P. Rigollet, *Optimal dimension dependence of the metropolis-adjusted langevin algorithm*, arXiv preprint arXiv:2012.12810 (2020).

[23] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, Journal of the Royal Statistical Society Series B: Statistical Methodology **79** (2017), no. 3, 651–676. MR3641401

[24] A. S. Dalalyan and L. Riou-Durand, *On sampling from a log-concave density using kinetic langevin diffusions*, Bernoulli **26** (2020), no. 3, 1956–1988. MR4091098

[25] V. De Bortoli and A. Durmus, *Convergence of diffusions and their discretizations: from continuous to discrete processes and back*, arXiv preprint arXiv:1904.09808 (2019).

[26] G. Deligiannidis, D. Paulin, A. Bouchard-Côté, and A. Doucet, *Randomized Hamiltonian Monte Carlo as scaling limit of the bouncy particle sampler and dimension-free convergence rates*, The Annals of Applied Probability **31** (2021), no. 6, 2612 – 2662. MR4350970

[27] P. Diaconis, *The Markov Chain Monte Carlo Revolution*, Bulletin of the American Mathematical Society **46** (2009), no. 2, 179–205. MR2476411

[28] P. Diaconis, S. Holmes, and R. M. Neal, *Analysis of a nonreversible Markov chain sampler*, Annals of Applied Probability (2000), 726–752. MR1789978

[29] P. Diaconis and L. Saloff-Coste, *What do we know about the metropolis algorithm?*, Proceedings of the twenty-seventh annual ACM symposium on Theory of computing, 1995, pp. 112–129.

[30] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov chains*, Springer, 2018. MR3889011

[31] A. Durmus and A. Eberle, *Asymptotic bias of inexact markov chain monte carlo methods in high dimension*, arXiv:2108.00682 [math.PR], 2021.

[32] A. Durmus, G. Fort, and E. Moulines, *Subgeometric rates of convergence in Wasserstein distance for Markov chains*, Ann. Inst. H. Poincaré Probab. Statist **52** (2016), no. 4, 1799 – 1822. MR3573296

[33] A. Durmus and E. Moulines, *Nonasymptotic convergence analysis for the unadjusted langevin algorithm*, Annals of Applied Probability **27** (2017), no. 3, 1551–1587. MR3678479

[34] A. Durmus and E. Moulines, *High-dimensional bayesian inference via the unadjusted langevin algorithm*, Bernoulli **25** (2019), no. 4A, 2854–2882. MR4003567

[35] A. Durmus and E. Moulines, *On the geometric convergence for MALA under verifiable conditions*, arXiv preprint arXiv:2201.01951 (2022).

[36] A. Durmus, E. Moulines, and E. Saksman, *Irreducibility and geometric ergodicity of Hamiltonian Monte Carlo*, The Annals of Statistics **48** (2020), no. 6, 3545 – 3564. MR4185819

[37] W. E and D. Li, *The Andersen thermostat in molecular dynamics*, CPAM **61** (2008), 96–136. MR2361305

[38] A. Eberle, *Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions*, The Annals of Applied Probability **24** (2014), no. 1, 337 – 377. MR3161650

[39] A. Eberle, *Bonn University Lecture Notes: Markov Processes*, 4 2023.

[40] A. Eberle, A. Guillin, and R. Zimmer, *Couplings and quantitative contraction rates for Langevin dynamics*, Ann. Probab. **47** (2019), no. 4, 1982–2010. MR3980913

[41] A. Eberle, A. Guillin, and R. Zimmer, *Quantitative harris-type theorems for diffusions and mckean–vlasov processes*, Transactions of the American Mathematical Society **371** (2019), no. 10, 7135–7173. MR3939573

[42] A. Eberle and M. B. Majka, *Quantitative contraction rates for Markov chains on general state spaces*, Electronic Journal of Probability **24** (2019), no. none, 1 – 36. MR3933205

[43] M. A. Erdogdu and R. Hosseinzadeh, *On the convergence of langevin monte carlo: The interplay between tail growth and smoothness*, Conference on Learning Theory, PMLR, 2021, pp. 1776–1822.

[44] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration*, Springer, 2010. MR2840298

[45] M. Hairer, *Convergence of markov processes*, Lecture notes (2010).

[46] M. Hairer and J. C. Mattingly, *Yet another look at harris' ergodic theorem for markov chains*, Seminar on Stochastic Analysis, Random Fields and Applications VI: Centro Stefano Franscini, Ascona, May 2008, Springer, 2011, pp. 109–117. MR2857021

[47] M. Hairer, J. C. Mattingly, and M. Scheutzow, *Asymptotic coupling and a general form of harris' theorem with applications to stochastic delay equations*, Probability theory and related fields **149** (2011), 223–259. MR2773030

[48] M. Hairer, A. M. Stuart, and S. J. Vollmer, *Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions*, Ann. Appl. Probab. **24** (2014), no. 6, 2455–2490. MR3262508

[49] W. K. Hastings, *Monte-Carlo methods using Markov chains and their applications*, Biometrika **57** (1970), 97–109. MR3363437

[50] J. Heng and P. E. Jacob, *Unbiased hamiltonian monte carlo with couplings*, Biometrika **106** (2019), no. 2, 287–302. MR3949304

[51] P. E. Jacob, J. O'Leary, and Y. F. Atchadé, *Unbiased markov chain monte carlo with couplings (with discussion)*, J. R. Stat. Soc. Ser. B **82** (2020), 543–600. MR4112777

[52] T. S. Kleppe, *Connecting the dots: Numerical randomized hamiltonian monte carlo with state-dependent event rates*, Journal of Computational and Graphical Statistics **31** (2022), no. 4, 1238–1253. MR4513384

[53] Y. T. Lee, R. Shen, and K. Tian, *Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo*, Proceedings of Thirty Third Conference on Learning Theory, vol. 125, 2020, pp. 2565–2597.

[54] Y. T. Lee, R. Shen, and K. Tian, *Lower bounds on metropolized sampling methods for well-conditioned distributions*, Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 18812–18824.

[55] B. Leimkuhler, D. Paulin, and P. A. Whalley, *Contraction and convergence rates for discretized kinetic langevin dynamics*, arXiv preprint arXiv:2302.10684 (2023). MR4748799

[56] T. Lelièvre, M. Rousset, and G. Stoltz, *Free energy computations: A mathematical perspective*, 1st ed., Imperial College Press, 2010. MR2681239

[57] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*, American Mathematical Society, 2009. MR2466937

[58] S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami, *On the geometric ergodicity of Hamiltonian Monte Carlo*, Bernoulli **25** (2019), no. 4A, 3109–3138. MR4003576

[59] N. Madras and D. Sezer, *Quantitative bounds for markov chain convergence: Wasserstein and total variation distances*, Bernoulli **16** (2010), no. 3, 882–908. MR2730652

[60] J. C. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise*, Stoch. Proc. Appl. **101** (2002), no. 2, 185–232. MR1931266

[61] K. L. Mengersen and R. L. Tweedie, *Rates of convergence of the Hastings and Metropolis algorithms*, Ann Stat **24** (1996), 101–121. MR1389882

[62] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, J Chem Phys **21** (1953), 1087–1092.

[63] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer-Verlag, 1993. MR1287609

[64] S. P. Meyn and R. L. Tweedie, *Computable Bounds for Geometric Convergence Rates of Markov Chains*, The Annals of Applied Probability **4** (1994), no. 4, 981 – 1011. MR1304770

[65] P. Monmarché, *High-dimensional mcmc with a standard splitting scheme for the underdamped langevin diffusion.*, Electronic Journal of Statistics **15** (2021), no. 2, 4117–4166. MR4309974

[66] P. Monmarché, *An entropic approach for hamiltonian monte carlo: the idealized case*, arXiv preprint arXiv:2209.13405 (2022). MR4728169

[67] P. Monmarché, *Hmc and langevin united in the unadjusted and convex case*, arXiv preprint arXiv:2202.00977 (2022).

[68] R. Montenegro and P. Tetali, *Mathematical aspects of mixing times in Markov chains*, Found. Trends Theor. Comput. Sci. **1** (2006), no. 3, x+121. MR2341319

[69] Y. Nesterov, *Lectures on convex optimization*, vol. 137, Springer, 2018. MR3839649

[70] J. O'Leary and G. Wang, *Metropolis-hastings transition kernel couplings*, arXiv preprint arXiv:2102.00366 (2021).

[71] L. Riou-Durand and J. Vogrinc, *Metropolis Adjusted Langevin Trajectories: a robust alternative to Hamiltonian Monte Carlo*, arXiv preprint arXiv:2202.13230 (2022).

[72] G. Roberts and J. Rosenthal, *One-shot coupling for certain stochastic recursive sequences*, Stochastic processes and their applications **99** (2002), no. 2, 195–208. MR1901153

[73] G. O. Roberts and J. S. Rosenthal, *General state space Markov chains and MCMC algorithms*, Probability Surveys **1** (2004), no. none, 20 – 71. MR2095565

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

[75] G. O. Roberts and R. L. Tweedie, *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms*, Biometrika **1** (1996), 95–110. MR1399158

[76] J. S. Rosenthal, *Minorization conditions and convergence rates for markov chain monte carlo*, Journal of the American Statistical Association **90** (1995), no. 430, 558–566. MR1340509

[77] S. M. Rump, *Estimates of the determinant of a perturbed identity matrix*, Linear algebra and its applications **558** (2018), 101–107. MR3854189

[78] J. M. Sanz-Serna and K. C. Zygalakis, *Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations*, The Journal of Machine Learning Research **22** (2021), no. 1, 11006–11042. MR4329821

[79] A. Scemama, T. Lelièvre, G. Stoltz, E. Cancés, and M. Caffarel, *An efficient sampling algorithm for variational Monte Carlo*, J Chem Phys **125** (2006), 114105.

[80] R. Shen and Y. T. Lee, *The randomized midpoint method for log-concave sampling*, Advances in Neural Information Processing Systems **32** (2019).

[81] D. Talay, *Stochastic Hamiltonian systems: Exponential convergence to the invariant measure, and discretization by the implicit Euler scheme*, Markov Processes and Related Fields **8** (2002), 1–36. MR1924934

[82] L. Tierney, *A note on Metropolis-Hastings kernels for general state spaces*, The Annals of Applied Probability **8** (1998), no. 1, 1 – 9. MR1620401

[83] C. Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008. MR2459454

[84] G. Wang, J. O'Leary, and P. Jacob, *Maximal couplings of the metropolis-hastings algorithm*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1225–1233.

[85] K. Wu, S. Schmidler, and Y. Chen, *Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling*, J. Mach. Learn. Res. **23** (2022). MR4577709

[86] J. Yang and J. S. Rosenthal, *Complexity results for MCMC derived from quantitative bounds*, The Annals of Applied Probability **33** (2023), no. 2, 1459 – 1500. MR4564431

Acknowledgments. The authors are greatly indebted to Andreas Eberle for his guidance & support. They further thank the referees for their comments.