On Convergence of the Alternating Directions Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) Algorithms

Anonymous authors
Paper under double-blind review

Abstract

We study convergence rates of practical Hamiltonian Monte Carlo (HMC) style algorithms where the Hamiltonian motion is approximated with leapfrog integration and where gradients of the log target density are accessed via a stochastic gradient (SG) oracle. Importantly, our analysis extends to allowing the use of general auxiliary distributions via a novel HMC procedure of alternating directions (AD).

The convergence analysis is based on the investigation of the Dirichlet forms associated with the underlying Markov chain driving the algorithms. For this purpose, we provide a detailed analysis on the error of the leapfrog integrator for Hamiltonian motions when both the kinetic and potential energy functions are in general form. We characterize the explicit dependence of the convergence rates on key parameters such as the problem dimension, functional properties of the target and auxiliary distributions and the quality of the SG oracle. Our analysis also identifies a crucial derivative condition on the log density of the auxiliary distribution, and we show that Gaussians (auxiliaries for standard HMC) as well as common choices of general auxiliaries for ADHMC satisfy this condition.

1 Introduction

The Hamiltonian Monte Carlo (HMC) algorithm has its humble beginning in physics, and recently it has seen much wider application in modern statistical analysis (inference and learning) and artificial intelligence. this has also subsequently generated a much deeper understanding of the method. Given a *target* distribution which is known to be proportional to a given positive, integrable function $\mathfrak{f}$, Markov chain Monte Carlo (MCMC) algorithms are commonly employed to provide either estimations of the normalizing constant (aka partition function) to $\mathfrak{f}$ or samples from the target distribution. HMC, a member of the MCMC family, utilizes the invariance and ergodic properties of Hamiltonian motion to offer additional benefits in performance comparing generic MCMC algorithms. With the help of a user chosen *auxiliary* distribution $\mathfrak{g}$ the algorithm generates a (Hamiltonian) motion while preserving the energy $\mathcal{H}(q,p) = U(q)+V(p)$, where $U(q) = -\log \mathfrak{f}(q)$ represents the potential energy and $V(p) = -\log \mathfrak{g}(p)$ the kinetic energy. The chief advantage lies in how well the motion dynamics are implemented; an exact implementation preserves the (joint) density even when making large moves and hence does not require a Metropolis-Hastings (MH) style acceptance/rejection step to ensure proper convergence. In practical instances where the motion is numerically approximated, excellent discrete motion implementations such as the Leafprog symplectic integrator ensure that rejection probability is low in the necessary MH correction even in high dimensional settings.

The literature focuses on analysis of HMC with Gaussians as auxiliary distribution, which corresponds to a rather simple quadratic kinetic energy function. There are several different approaches for both qualitative and quantitative analysis on the key question of the convergence and performance of the HMC algorithms with Gaussian auxiliaries, see,e.g. Zou & Gu (2021); Li et al. (2019); Gao et al. (2021). As HMC is increasingly used in more complex applications, for example large language models, it becomes desirable that the algorithm, as well as its analysis, can be extended to allow more flexibility in the selection of the auxiliary distributions. This requires a certain novel algorithm called the Alternating Direction HMC Ghosh et al. (2025); we describe it in the sequel. As demonstrated in Ghosh et al. (2025), the careful selection of

non-Gaussian auxiliary forms can in fact considerably improve the performance of the HMC algorithm. In many practical situations, the gradient of the potential energy $U(q)$ of the target density function, which is essential for the running of HMC, is not available or difficult to compute. One approach in this case is to substitute the exact calculation of the function $\nabla U(q)$ by an (unbiased) estimator $\tilde{G}_r(q,\xi)$ with an independent random variable ξ. Generally speaking, the computation complexity for calculating $\tilde{G}_r(q,\xi)$ is considered to be significantly less than that of $\nabla U(q)$, especially in high dimensions. Examples of such estimators include Mini-batch stochastic gradient, Stochastic variance reduced gradient, Stochastic averaged gradient and Control variate gradient, as summarized in Zou & Gu (2021).

Our primary purpose is to present a unified quantitative convergence analysis of the family of Alternating Direction HMC algorithms that allow for *arbitrary* not necessarily symmetric auxiliary distributions. The analysis presented here also allows for stochastic (inexpensive) oracles that estimate the gradient $\nabla U(q)$ of the potential of the target distribution, and where Hamiltonian motion is implemented with leapfrog symplectic integrators (requiring additional MH correction). We take an analytical approach based on the analysis of the Dirichlet forms that are defined by the underlying Markov chain, and are able to quantitatively characterize convergence rates of Alternating Direction SGHMC under mild conditions on the target density, the (arbitrary) auxiliary density and the stochastic gradient oracle implementations. These, to the best of our knowledge, are the first set of results on the convergence rates of HMC with this kind of generality.

As a special instance of MCMC, an HMC algorithm is driven by a Markov chain in a general state space. Hence, the analysis of its convergence relies on the analysis of this Markov chain. Convergence of Markov chains, or more general Markov processes, is a central topic in probability theory. The variety of different approaches is larger than what a few books, see, e.g. Meyn & Tweedie (1993) and Levin et al. (2009), can cover. The main approach taken in this paper is based on the Dirichlet form, a systematic treatment of which can be found in Fukushima et al. (2010). Intuitively, this analysis focuses on establishing functional relationships that quantitatively characterize the evolution of the Markov chain, thus facilitating the convergence analysis. In an abstract and general sense, a Dirichlet form is a non-negative definite symmetric bilinear form defined on a Hilbert space, which furthermore is both Markovian and closed. For each Markov chain, a specific Dirichlet form can be defined naturally through a Markov operator on the Hilbert space, which defines this Markov chain. Moreover, due to the celebrated result of Jeff Cheeger, a quantitative relationship between the Dirichlet form and the variance term characterizes the spectral gap of the Markov chain, which in turn is directly related to the convergence rate. These concepts will be introduced in Sec. 2. The main technical portion of the paper will address the estimation of the Dirichlet form.

1.1 Literature review.

The research on the convergence and convergence rates of HMC has been concentrated on the case of the auxiliary distribution $\mathfrak{g}(p)$ being a (conditional) Gaussian. Theoretical understanding of geometric convergence have been developed for these cases, via analytical methods (including comparison theorems for differential equations in Chen & Vempala (2019)), or probabilistic methods, such as Harris recurrence techniques Bou-Rabee & Sanz-Serna (2017) and coupling Bou-Rabee et al. (2020). For HMC with general auxiliary distributions, qualitative results are obtained in Ghosh et al. (2022a) and Ghosh et al. (2022b).

The problem we are dealing with here appears to be one instance of the general problem of HMC with a stochastic gradient. While there are existing results for its convergence, see e.g.Zou & Gu (2021); Li et al. (2019); Gao et al. (2021), our analysis presented in this paper with explicit convergence rate estimates under very general assumptions (general auxiliary distribution, stochastic gradient implementation and alternating direction) are certainly innovative results.

Meanwhile, there are also extensive quantitative studies on establishing the dependence of convergence rates on parameters of the algorithms, including, the dimension of the underlying space, the function properties of the target distribution and the quality of the numerical integrators. For performance of the unadjusted HMC (HMC with numerical integrators but without a Metropolis-Hastings step), in Wasserstein distances, see e.g. Gouraud et al. (2023), Shen & Lee (2019), Cao et al. (2021), Bou-Rabee & M. (2022). Similarly, results in Mangoubi & Vishnoi (2018), Chen et al. (2020), Beskos et al. (2013), and Chen & Gatmiry (2023)

quantifies "gradient complexity"(the amount of gradient calculation required) for HMC with Metropolis-Hastings adjustment.

1.2 Summary of our contributions

1.2.1 Error Estimation:

We present a detailed and comprehensive analysis in Lemmata 4.1 and 4.2 on the quality of Leapfrog implementations of the symplectic integration for Hamiltonian equations with general kinetic energy. This is the key for the analysis of HMC with general auxiliary distributions and stochastic gradient. It not only serves as the main technical component for convergence results in this paper, but it can also be used as a building block for the analysis of many other variations of HMC, such as AD-HMC seen in this paper, as well as other systems where sympletic integrations and estimation are required.

1.2.2 Convergence:

Quantitative bound on the performance for SGHMC algorithms with general auxiliary distributions are derived. To our best knowledge, this is the first such results with such kind of generality. In addition, these bounds are expressed in explicit forms of the system parameters including the dimension. Our analysis also identifies a crucial derivative condition (Assumption 4) on the kinetic energy of general auxiliary distributions that allows for geometric convergence, and we show that Gaussians (auxiliaries for standard HMC) as well as common choices of general auxiliaries for ADHMC (as studied in Ghosh et al. (2025)) satisfy this condition.

1.2.3 Methods:

The method we used here consists of Dirichlet form and functional inequalities. They offer clearness in concepts and flexibility in analysis, and appear to be promising in achieving both qualitative and quantitative results, and we hope that they would find more applications within this community. We also aim to remove some of the restrictions and apply them to more general systems in the future.

1.3 Organization

The rest of the paper is organized as follow: in Sec. 2, we introduce the HMC algorithm and provide details on its various implementation, and list the assumptions on the functions; geometric convergence is discussed in Sec. 3; some ramifications will be presented in Sec. 4; and the paper concludes in Sec. 5.

2 Algorithms and Assumptions

2.1 Definitions, Notations and Assumptions

For any $\mathfrak{q} \in \mathbb{R}^d$, and $p \in \mathbb{Z}_+$, the p-norm is defined as $\|q\|_p = (\sum_{i=1}^d q_i^p)^{1/p}$. For a random variable defined on $\mathbb{R}^d$, this *can be extended to* $|||q|||_p := (\mathbb{E}\|q\|_2^p)^{1/p}$. For a $d \times d$ matrix A, the operator norm (aka spectral norm) is defined as $\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2$ and Frobenius norm as $\|A\|_F = \sqrt{\sum_{i,j=1}^d A_{ij}^2}$. For any function f, $\nabla^3 f$ can be treated as a tensor, and

$$\|\nabla^3 f\| = \sup\left\{ \left| \sum_{i,j,k=1}^d \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k} u_i v_j w_k \right| : \|u\|_2, \|v\|_2, \|w\|_2 \le 1 \right\}.$$

One of the key assumptions in Ghosh et al. (2022b) under which the *geometric convergence* of HMC is established is the *uniform strongly logarithmic concavity* of both the target and auxiliary distributions. This is equivalent to making an assumption on the convexity and derivative-Lipshitzness conditions on the energy functions.

Definition 1 ($\mathcal{S}_{\ell,L}(\mathbb{R}^d)$ **class).** *A function $W : \mathbb{R}^d \to \mathbb{R}$ is called to be a class $\mathcal{S}_{\ell,L}$ for some $\ell, L > 0$ if the following holds for any $x_1, x_2 \in \mathbb{R}^d$ and $t \in [0, 1]$,*

$$W((1-t)x_1 + tx_2) \leq (1-t)W(x_1) + tW(x_2) - \frac{\ell}{2}t(1-t)\|x_1 - x_2\|^2,$$

and $\|\nabla W(x_2) - \nabla W(x_1)\| \leq L\|x_2 - x_1\|$.

Remark 2.1. *The function class $\mathcal{S}_{\ell,L}$ is the same as that of $\mathcal{S}^{1,1}_{\ell,L}(\mathbb{R}^d)$ class in Nesterov (2003). It should be easy to see that $\ell\,\mathrm{Id} \preceq \nabla^2 W \preceq L\,\mathrm{Id}$, if $\nabla^2 W$ exists. For any two matrices A and B, $A \preceq B$ means that $B - A$ is positive semidefinite.*

Assumption 1. *There exist $0 < \ell_U \leq L_U < \infty$ and $0 < \ell_V \leq L_V < \infty$ such that, $U \in \mathcal{S}_{\ell_U, L_U}(\mathbb{R}^d)$, and $V \in \mathcal{S}_{\ell_V, L_V}(\mathbb{R}^d)$.*

Assumption 2. *Both U and V have third derivatives, and there exist $0 < T_U, T_V < \infty$ such that $\sup_{q\in\mathbb{R}^d} \|\nabla^3 U(q)\| \leq T_U$ and $\sup_{q\in\mathbb{R}^d} \|\nabla^3 V(q)\| \leq T_V$.*

2.2 Dirichlet Form and Spectral Gap.

Dirichlet form, as a generalization of the Laplace operator, is an important concept in analysis, a systematic treatment of its connection to probability theory, especially the symmetric Markov processes can be found in Fukushima et al. (2010).

Definition 2. *A symmetric bilinear form $\mathcal{E}(\cdot,\cdot)$ on the Hilbert space $L^2(X, m)$ with X being a metric space and m a Borel measure is* Markovian *if for any $\epsilon > 0$, there exists a real function $\phi_\epsilon(t) : \mathbb{R}a\mathbb{R}$ satisfying $\phi_\epsilon(t) = t$ for $t \in [0, 1]$, $\phi_\epsilon(t) \in [-\epsilon, 1 + \epsilon]$, and $0 \leq \phi_\epsilon(t') - \phi_\epsilon(t) \leq t' - t$ whenever $t < t'$, such that $\mathcal{E}(\phi_\epsilon(u), \phi_\epsilon(u)) \leq \mathcal{E}(u, u)$.*

Definition 3. *A symmetric bilinear form is a* Dirichlet form *if it is both Markovian and closed.*

For a reversible Markov chain on $\mathbb{R}^d$ with invariant measure $\pi(x)$ and transition kernel $P(x, A)$, , such as the ones we considered here in this paper, the following gives a natural Dirichlet form on $L^2(\mathbb{R}^d, \pi)$ (without causing confusion, we will write L^2 in the sequel),

$$\mathcal{E}(g, h) = \int_X \int_X [g(x) - g(y)][h(x) - h(y)]\pi(dx)P(x, dy).$$

Moreover, the spectral gap of such Markov chain, $1 - \lambda_2$, has the following representation,

$$1 - \lambda_2 = \inf_{h \text{ not constant}} \frac{\mathcal{E}(h, h)}{\mathsf{Var}_\pi(h)},$$

with λ_2 represents the second largest eigenvalue and $\mathsf{Var}_\pi(h) := \int_X \int_X (h(x) - h(y))^2 \pi(dx)\pi(dy)$. The Dirichlet form approach on the convergence rate is closely related the study of conductance originated by Jeff Cheeger Cheeger (1969) and carried out by a series of subsequent studies. A detailed exposition of the results and basic arguments can be found in Lawler & Sokal (1988). For Markov chain generated by the HMC algorithm with leapfrog implementation, the presence of invariant measure(up to a constant) and explicit form of transition make the Dirichlet form approach very appealing.

2.3 Hamiltonian Monte-Carlo Algorithms

2.3.1 Basic Algorithms

A generic HMC algorithm, see Algorithm 1, on Euclidean space usually consists of three operations at each step, with a starting point $q \in \mathbb{R}^d$ is given: (1) "lift", it is also called "spread " in literature, where a sample p is drawn from the auxiliary distribution with density $\mathfrak{g}(\cdot)$, (q, p) will be the point in the (symplectic) space of $\mathbb{R}^{2d}$; 2) "rotation" to a new point, $(\hat{q}, \hat{p})$, is identified by the Hamiltonian trajectory with energy $\mathcal{H}(q, p) = -\log[\mathfrak{f}(q)\mathfrak{g}(p)] = U(q) + V(p)$, represented here by its potential and kinetic components; 3) "projection", $\hat{q}$ will be the starting point of the next step. More specifically, Algorithm 1 presents the standard

HMC with Gaussian auxiliaries (and kinetic term $V(p) = p^2/2$) and utilizes a stochastic oracle to obtain noisy estimates $\nabla^\omega U(\cdot)$ for the gradient of the potential $U(q)$. It also uses K steps of the symplectic integrator each with size η to implement Hamiltonian motion. **Leapfrog implementation of the symplectic integration.** The exact Hamiltonian integration to get from (q, p) to $(\hat{q}, \hat{p})$ is in general expensive to calculate, and the following well-known Leapfrog approximation is considered here.

$$\hat{q} = q + \eta \nabla V \left(p - \frac{1}{2} \eta \nabla U(q) \right), \qquad \hat{p} = p - \eta \frac{\nabla U(q) + \nabla U(\hat{q})}{2}. \tag{1}$$

2.3.2 Stochastic Gradient HMC

As discussed in Zou & Gu (2021), while the functions of the auxiliary distribution can be more easily calculated since it is chosen by the users, calculations of gradients of the target density are not always readily available or can be attained with low costs. Therefore, they have been approximated in practice, here are a few examples: *Mini-batch stochastic gradient:* in this case, $U(\mathfrak{q}) = \sum_{i=1}^n U_i(\mathfrak{q})$, ξ is a random variable for uniformly randomly pick a size-B subset $\mathcal{I}$ of $[n]$, then the gradient of $U(\mathfrak{q})$ is estimated by the following unbiased estimator, $\tilde{G}_r(\mathfrak{q}, \xi) = \frac{n}{B} \sum_{i \in \mathcal{I}} \nabla U_i(\mathfrak{q})$. Variations of the above method include *stochastic variance reduced gradient*, *stochastic averaged gradient*, and *control variate gradient*. Details can be found in Zou & Gu (2021) and references therein. We start with an assumption on the quality of the SG oracle.

Assumption 3. *The approximate calculation of ∇U is treated as a distribution, denoted as ∇U^ω. Furthermore, there exist $0 < \underline{\ell} \leq \bar{L} < \infty$ and $\bar{T} > 0$, such that, $U^\omega \in \mathcal{S}_{\ell^\omega, L^\omega}(\mathbb{R}^d)$ almost surely, with $\ell^\omega \geq \underline{\ell} > 0$ and $L^\omega \leq \bar{L} < \infty$ and $\|\nabla^3 U^\omega\| \leq \bar{T}$ almost surely.*

An Acceptance/Rejection step. At each step of the HMC implementation, leapfrog procedure 1 will be invoked K time, produce a proposal (q_K, p_K) from the initial state (q_0, p_0) formed by the initial position and p_0 sampled from the auxiliary distribution. Then the proposal is accepted with probability $A_{K,\eta} = \min\{1, \mathfrak{f}(q_K)\mathfrak{g}(p_K)/\mathfrak{f}(q_0)\mathfrak{g}(p_0)\}$, that is

$$\log A_{K,\eta}(q_0, p_0) := \min\left\{0, \mathcal{H}(q_0, p_0) - \mathcal{H}(q_K, p_K)\right\}. \tag{2}$$

Algorithm 1 SGHMC

Initialization: stochastic oracle for $\nabla U^{\omega}(q)$ for potential energy $U(q)$ gradient; kinetic energy $V(p) = p^2/2$ with gradient $\nabla V(p) = p$; initial iterate q_0; K steps of size η for total trajectory length $K\eta$
for $n = 1, \ldots$ **do**
  Set $q_0 \leftarrow q_{n-1}$
  Sample: $p_0 \sim \mathfrak{g}(p)$
  Lift: $(q_0, p_0) \leftarrow q_0$
  Move:
  Start with sample of $\nabla U^{\omega}(q_0)$
  for $k = 0, \ldots, K-1$ **do**
    Set $p_{k+\frac{1}{2}} \leftarrow p_k - \frac{\eta}{2}\nabla U^{\omega}(q_k)$
    Set $q_{k+1} \leftarrow q_k + \eta \nabla V(p_{k+\frac{1}{2}})$
    Sample $\nabla U^{\omega}(q_{k+1})$
    Set $p_{k+1} \leftarrow p_{k+\frac{1}{2}} - \frac{\eta}{2}\nabla U^{\omega}(q_{k+1})$.
  end for
  Sample $Z \sim$ Uniform $(0, 1)$.
  if $Z \leq \frac{\mathfrak{f}(q_K)\mathfrak{g}(p_K)}{\mathfrak{f}(q_0)\mathfrak{g}(p_0)}$ **then**
    Project: $q_n \leftarrow (q_K, p_K)$
  else
    Set $q_n \leftarrow q_0$
  end if
end for

Algorithm 2 Stochastic Gradient AD-HMC

Initialization: stochastic oracle for $\nabla U^{\omega}(q)$ for potential energy $U(q)$ gradient; kinetic energy $V(p)$ with gradient oracle $\nabla V(p)$; initial iterate q_0; K steps of size η for total trajectory length $K\eta$
for $n = 1, \ldots, N$ **do**
  Set $q_0 \leftarrow q_{n-1}$
  $\{$*forward motion*$\}$
  Sample: $p_0 \sim \mathfrak{g}(p)$
  Lift: $(q_0, p_0) \leftarrow q_0$
  Move:
  Start with sample of $\nabla U^{\omega}(q_0)$
  for $k = 0, \ldots, K-1$ **do**
    Set $p_{k+\frac{1}{2}} \leftarrow p_k - \frac{\eta}{2}\nabla U^{\omega}(q_k)$
    Set $q_{k+1} \leftarrow q_k + \eta \nabla V(p_{k+\frac{1}{2}})$
    Sample $\nabla U^{\omega}(q_{k+1})$
    Set $p_{k+1} \leftarrow p_{k+\frac{1}{2}} - \frac{\eta}{2}\nabla U^{\omega}(q_{k+1})$.
  end for
  Project: $q'_0 \leftarrow (q_K, p_K)$
  $\{$*backward motion*$\}$
  Sample: $p'_0 \sim \mathfrak{g}(p)$
  Lift: $(q'_0, p'_0) \leftarrow q'_0$
  Move:
  Start with sample of $\nabla U^{\omega}(q'_0)$
  for $k = 0, \ldots, K-1$ **do**
    Set $p'_{k+\frac{1}{2}} \leftarrow p'_k + \frac{\eta}{2}\nabla U^{\omega}(q'_k)$
    Set $q'_{k+1} \leftarrow q'_k - \eta \nabla V(p'_{k+\frac{1}{2}})$
    Sample $\nabla U^{\omega}(q'_{k+1})$
    Set $p'_{k+1} \leftarrow p'_{k+\frac{1}{2}} + \frac{\eta}{2}\nabla U^{\omega}(q'_{k+1})$.
  end for
  Sample $Z \sim$ Uniform $(0, 1)$.
  if $Z \leq \frac{\mathfrak{f}(q'_K)\mathfrak{g}(p_K)\mathfrak{g}(p'_K)}{\mathfrak{f}(q_0)\mathfrak{g}(p_0)\mathfrak{g}(p'_0)}$ **then**
    Project: $q_n \leftarrow (q'_K, p'_K)$
  else
    Set $q_n \leftarrow q_0$
  end if
end for

2.3.3 AD-HMC

The possibility of utilizing general asymmetric auxiliary distributions $\mathfrak{g}(\mathfrak{p})$ affords us a modification of SGHMC Algorithm by a procedure alternating Hamiltonian motion in forward and backward directions for the same length T. The modified is called the Alternating Direction HMC (AD-HMC). One of the motivations for such modification is to produce symmetry with asymmetric auxiliary distribution, which corresponding to self-adjointness of the underlying operator.

One step of the proposed AD-HMC algorithm for asymmetrical auxiliaries $\mathfrak{g}$ starts from a $q_0 \in \mathbb{R}^d$ by generating a sample $p_0 \in \mathbb{R}^d$ and applying forward Hamiltonian motion that then carries the pair (q_0, p_0) to some (q_1, p_{01}). Then, another momentum $p_{12} \in \mathbb{R}^d$ is sampled and the backward Hamiltonian motion carries (q_1, p_{12}) to (q_2, p_2), yielding the candidate q_2 for the next state. Similarly, should we start AD-HMC with q_2, the pair of momentum vectors p_2 and p_{01} will take us back to q_0 through q_1. So, we will accept the

proposed move to q_2 with probability

$$\mathcal{P}(q_0, q_1, q_2) = \min\left\{1, \frac{\mathfrak{f}(q_2)\mathfrak{g}(\Pi_{q_1}^{-b}(q_0))\mathfrak{g}(\Pi_{q_2}^{-f}(q_1))}{\mathfrak{f}(q_0)\mathfrak{g}(\Pi_{q_0}^{-f}(q_1))\mathfrak{g}(\Pi_{q_1}^{-b}(q_2))}\right\}, \tag{3}$$

where $\Pi_{q_0}^{-f}(q_1)$ denotes the momentum of the forward motion carries q_0 to q_1, and $\Pi_{q_1}^{-b}(q_2)$ the momentum of the backward motion carries q_1 to q_2. The transition probability of the AD-HMC Markov chain with the Hamiltonian motion augmented with the MH rejection step using equation 3 is equal to $\mathsf{P}(q_0, q_2) = \int_{\mathbb{R}^d} \mathcal{P}(q_0, q_1, q_2)\mathfrak{g}(\Pi_{q_0}^{-f}(q_1))\mathfrak{g}(\Pi_{q_1}^{-b}(q_2))\, dq_1$. It is established in Ghosh et al. (2025) that the underlying Markov chain of the augmented AD-HMC procedure possesses the desired time reversibility.

Algorithm 2 presents the Alternating Direction HMC algorithm that can be used with arbitrary auxiliaries (with kinetic term $V(p)$). A stochastic oracle $\nabla U^{\omega}(q)$ provides an estimate of the gradient of the potential $U(q)$. Two sets of Hamiltonian motions are implemented using K steps of the symplectic integrator of size η : the first implements forward motion, and the second implements backward motion. Note that a new sample p_0' of the momentum is used for the latter set of motion. The MH correction steps involve both sets of initial momenta p_0 and p_0' as well as the last momenta p_K and p_K'.

2.4 Preliminary Results.

Lemma 2.1. *Under the Assumption 1, we have, for any integer $p > 1$,*

$$\left[\mathbb{E}_{q\sim\exp(-U)}\|\nabla U(q)\|_2^{2p}\right]^{\frac{1}{p}} \leq (d + 2p - 2)L_U, \tag{4}$$

$$\left[\mathbb{E}_{p\sim\exp(-V)}\|\nabla V(p)\|_2^{2p}\right]^{\frac{1}{p}} \leq (d + 2p - 2)L_V. \tag{5}$$

Proof. Assumption 1 implies that $Tr(\nabla^2 U) \leq L_U d$, equation 4 follows from Lemma 9 of Chen & Gatmiry (2023), note that the subexponential condition is naturally satisfied. Meanwhile, by an application of the Green's formula and Hölder's inequality, similar to that in the proof of Lemma 9 in Chen & Gatmiry (2023), equation 5 follows. □

Lemma 2.2. *If, in addition to Assumption 1, the auxiliary distribution satisfies $\mathbb{E}[p_i] = 0$, $\mathbb{E}[p_i p_j] = \delta_{ij}\sigma_i^2$, and there exists $\iota > 0$ such that $\mathbb{E}[p_i^4] \leq \Sigma_4$ and $\sigma_i^2 \leq \Sigma_2$ for any $i, j = 1, 2, \ldots, d$. we have, $\mathbb{E}_{p\sim\exp(-V)}[(p^T\nabla^2 U(q)p)^2] \leq (\Sigma_2 + \Sigma_4)dL_U^2$.*

Proof. Direct calculations give us

$$\begin{aligned}\mathbb{E}_{p\sim\exp(-V)}[(p^T\nabla^2 U(q)p)^2] =& \mathbb{E}\sum_{i,j=1,i\neq j}^{d}\left[\frac{\partial^2}{\partial q_i \partial q_j}U(q)\right]^2\sigma_i^2\sigma_j^2 + \sum_{i=1}^{d}\left[\frac{\partial^2}{\partial q_i^2}U(q)\right]^2\mathbb{E}[p_i^4] \\ \leq& (\Sigma_2 + \Sigma_4)\|\nabla^2 U(q)\|_F^2 \leq (\Sigma_2 + \Sigma_4)dL_U^2.\end{aligned}$$

The last inequality follows from the Assumption 1. More specifically, for any two real positive semi-definite matrices A and B satisfying $A \preceq B$, then we know that $\|A\|_F \leq \|B\|_F$. (A quick proof $Tr(B^2) \geq 2Tr(AB) - Tr(A^2) = Tr(A^2) + 2(Tr(A(B - A)) \geq Tr(A^2)$ where the first inequality is due to $Tr(A - B)^2 \geq 0$ and the second one since both A and $B - A$ are symmetric and positive semidefinite.) □

3 Geometric Convergence of SGHMC implementations in Euclidean spaces

In this section, explicit geometric convergence rates are estimated for general HMC algorithms, including features such as general auxiliary distribution (ADHMC) and stochastic gradient estimation(SGHMC). We are using a *functional approach.*

A basic argument for geometric convergence of Markov chain, treated as iterations driven by the Markov operator, including explicit estimation of convergence rates, based on analyzing functional displacement of

the operator is presented in Lovász & Simonovits (1993). The key result is that given a *time-reversible* Markov chain, with conductance Φ, the inequality for the inner product $\langle h, Mh\rangle \leq \left(1 - \frac{\Phi^2}{2}\right) \|h\|^2$ holds for every mean zero, non constant $h \in L^2$, where Mh represents the image of h under the Markov operator, more precisely, $Mf(q) = \int_{\mathbb{R}^d} f(q)P(q, dq')dq'$. Rewriting the inequality, we have $\|h\|^2 - \langle h, Mh\rangle \geq \frac{\Phi^2}{2}\|h\|^2$, which says that the norm of the displacement of the Markov chain is lower bounded by norm of the preimage up to a constant. Subsequently, the convergence rate of the Markov chain to its invariant measure can be quantified utilizing the above inequality, see e.g. Roberts & Rosenthal (1997), Roberts & Tweedie (2001) and Madras & Randall (2002). More specifically,

Lemma 3.1. *If $\|h\|^2 - \langle h, Mh\rangle \geq \frac{\Phi^2}{2}\|h\|^2$ is satisfied by any mean zero $h \in L^2$, the time reversible Markov chain converges to its stationary distribution at a rate at least $1 - \frac{\Phi^2}{2}$ in total variational distance.*

Definition 4. *For a $r \in (0, 1)$, a Markov chain is said to converge exponentially to its stationary distribution with rate at least r if there exists a $C > 0$ such that for all $n \geq 1$, we have $d(\pi_n, \pi_\infty) \leq Cr^n$ for certain distance between (probability) measures.*

We impose the following additional assumption on the curvature of $V(p)$ that will play an important role in deriving explicit expressions for the geometric convergence rates of the HMC variants.

Assumption 4. *The auxiliary $g(p) \propto \exp(-V(p))$ satisfies the derivative condition (taking expectation w. r. t. $g(p)$):*

$$\mathbb{E}\left[\nabla V(p)\ \nabla V(p)^\top - \nabla^2 V(p)\right] = 0.$$

Assumption 4 bears a superficial resemblance to the celebrated information matrix equality (c.f. Ch 1.3.2 in Amemiya (1985)) that arises in estimation of parameters θ of a density $g_\theta(p)$ by maximizing the likelihood $L_\theta(p) = \log g_\theta(p)$. However, they are distinct because the derivatives taken there are with respect to parameters θ while the derivatives in Assumption 4 are w.r.t. p. In particular, the proof of the equality $\mathbb{E}\left[\nabla_\theta L_\theta(p) \nabla_\theta L_\theta(p)^\top\right] = -\mathbb{E}\left[\nabla^2_\theta L_\theta(p)\right]$ depends on a crucial interchange between integral over p and differential over θ, which does not apply in the case of Assumption 4.

Proposition 3.1. *Suppose $\{g_m(\cdot), m \in [M]\}$ are a finite collection of densities, each of which satisfy Assumption 4. Let $\{\gamma_m, m \in [M]\}$ be a collection of real numbers such that $\gamma_m \geq 0$ and $\sum_m \gamma_m = 1$. Then, the mixture density $g(p) = \sum_m \gamma_m g_m(p)$ also satisfies Assumption 4.*

Proof. Let $V_m(p) = -\log g_m(p)$. The first derivative of $V(p) = -\log g(p)$ is

$$\nabla V(p) \ = \ -\frac{1}{g(p)} \sum_m \gamma_m \nabla g_m(p) \ = \ \sum_m \gamma_m \frac{g_m(p)}{g(p)} \nabla V_m(p).$$

Applying derivatives of products, we get the second derivative as

$$\begin{aligned}
\nabla^2 V(p) = & \sum_m \gamma_m \left[\frac{1}{g(p)} \nabla V_m(p) \nabla g_m(p)^\top \ - \ \frac{g_m(p)}{g^2(p)} \nabla V_m(p) \nabla g(p)^\top \ + \ \frac{g_m(p)}{g(p)} \nabla^2 V_m(p)\right] \\
= & \sum_m \gamma_m \left[\frac{g_m(p)}{g(p)} \left(\nabla^2 V_m(p) - \nabla V_m(p) \nabla V_m(p)^\top\right) \ + \ \frac{g_m(p)}{g(p)} \nabla V_m(p) \nabla V(p)^\top\right] \\
= & \sum_m \gamma_m \frac{g_m(p)}{g(p)} \left(\nabla^2 V_m(p) - \nabla V_m(p) \nabla V_m(p)^\top\right) \ + \ \left(\sum_m \gamma_m \frac{g_m(p)}{g(p)} \nabla V_m(p)\right) \nabla V(p)^\top
\end{aligned}$$

From the expression for the first derivative $\nabla V(p)$, we see that the last term simplifies to $\nabla V(p) \nabla V(p)^\top$. The expectation w.r.t. $g(p)$ of the first term yields a zero since each density $g_m(p)$ satisfies Assumption 4. Thus, the mixture $g(p)$ also satisfies Assumption 4. □

We now observe that multivariate Gaussians, the auxiliaries used in standard HMC, as well as mixtures of Gaussians, the nonsymmetric auxiliaries studied in Ghosh et al. (2025) with ADHMC, satisfy Assumption 4.

Corollary 1. *Suppose $g_m(p)$ is a multivariate Gaussian centered at p^m with covariance Σ_m. Then, Assumption 4 is satisfied by $g_m(p)$. Further, mixtures $g(p) = \sum_m \gamma_m g_m(p)$ of multivariate Gaussian distributions also satisfy Assumption 4.*

Proof. We have that $V_m(p) = \frac{1}{2}(p-p^m)^\top \Sigma_m^{-1}(p-p^m)$ with derivatives $\nabla V_m(p) = \Sigma_m^{-1}(p-p^m)$ and $\nabla^2 V_m(p) = \Sigma^{-1}$. We get the first assertion by noting that $\mathbb{E}[(p-p^m)(p-p^m)^\top] = \Sigma_m$. The second assertion follows by applying Proposition 3.1. □

Utilizing Lemma 3.1, we have:

Lemma 3.2. *Under Assumptions 1 and 4, suppose that M_H represents the Markov operator generated by the HMC algorithm with K leapfrog steps. Then,*

$$\|h\|^2 - \langle h, M_H h\rangle \geq K\eta^2 \left(\frac{C_1\sigma_V^2}{2} - A_3\eta\right)\|h\|^2. \tag{6}$$

where C_1 is a constant determined by Poincaré inequality for general measure, $\sigma_V^2 := \int_{\mathbb{R}^d} \|\nabla V(p)\|_2 g(p)dp$, and the constant A_3 is defined in Lemma 4.5.

The proof of Lemma 3.2 can be found in Sec. C. It leads to the following result:

Theorem 3.1. *Under Assumptions 1 and 4, for $\eta < \frac{C_1\sigma_V^2}{4A_3}$, the Markov chain generated by the HMC algorithm converges at a rate at least $\frac{K\eta^3\sigma_V^2}{4}$. More precisely, we have that*

$$d_{TV}(\hat{\pi}^n, \pi) \leq \left(1 - \frac{K\eta^3\sigma_V^2}{4}\right)^n d_{TV}(\hat{\pi}, \pi),$$

with $\hat{\pi}$ denotes the initial distribution of the Markov chain and $\hat{\pi}^n$ denotes its distribution after n transitions.

Proof. The theorem follows from Lemmata 3.1 and 3.2. □

In case of SGHMC, we have:

Lemma 3.3. *Under Assumptions 1, 3 and 4, suppose that M_{SG} represents the Markov operator generated by the SGHMC algorithm with K leapfrog steps. Then,*

$$\|h\|^2 - \langle h, M_{SG} h\rangle \geq K\eta^2 \left(\frac{C_1\sigma_V^2}{2} - A_3^{SG}\eta\right)\|h\|^2,$$

where A_3^{SG} is the constant from Lemma 4.6.

Proof. The only difference from Lemma 3.2 are the constants estimated by Lemma 4.2. □

Corollary 2. *Under Assumptions 1, 3 and 4, for $\eta < \frac{C_1\sigma_V^2}{4A_3^{SG}}$, the Markov chain generated by the SGHMC algorithm converge at a rate at least $\frac{K\eta^3\sigma_V^2}{4}$. More precisely,*

$$d_{TV}(\hat{\pi}^n, \pi) \leq \left(1 - \frac{K\eta^3\sigma_V^2}{4}\right)^n d_{TV}(\hat{\pi}, \pi).$$

As we can see from Lemma 3.2, Theorem 3.1 and their proofs, the key for determining the convergence rates of the SGHMC algorithms lies in quantifying the closeness of the *symplectic integrator* to exact solution to the Hamiltonian system. These *quantifications* have been summarized in the next section, and detailed calculations are presented in Sec. A. Both Theorem 3.1 and Corollary 2 apply to AD-HMC algorithms directly. For background and derivations, as well as estimates of constant for the *Poincaré inequality* for a general family of measures that include log-concave case, see, e.g. Bakry et al. (2008) and Villani (2009). Due to this connection, as well as conditions on the moments of the auxiliary distributions, our convergence

rates are less restricted by large dimension, comparing to for example those in Mangoubi & Vishnoi (2018) and Chen & Gatmiry (2023). It is desirable to obtain more precise results on the rate, this will be depend on more sophisticated on advances in quantitative results on functional inequalities including the Poincaré's inequality.

4 Technical Results and Ramifications

In this section, we present quantitative results on some key aspects of general HMC algorithms. First, we provide a range of *quantitative estimations* of potential numerical errors in leapfrog implementations of the algorithms. Second, we will characterize the statistical distance in *KL divergence* between two proposed HMC steps with respect to the distance between their initial states. Lastly, bounds on the (Metropolis-Hastings) *acceptance probability* are obtained. From the dependence of our main results in last section on some of these results, we can see that that these quantities are crucial for the performance of SGHMC. In addition, The results presented here can also be utilized for establishing quantitative convergence results through other arguments, for example through *conductance* type arguments as presented in Chen & Gatmiry (2023).

4.1 Leapfrog vs Exact

One of the keys to the success of HMC algorithms is the effectiveness of the numerical symplectic integration of the Hamiltonian differential equations. Extensive efforts have been devoted to such studies for HMC Gaussian auxiliary distribution across the existing literature. Allowing *general auxiliary distributions* certainly has made the analysis more complicated; the introduction of stochastic gradient estimation leads to new difficulties in this problem. In a series of technical results, we are able to provide sharp estimates on the error produced in these numerical procedures.

Lemma 4.1. *Under Assumptions 1 and 2,when* $\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\| < \infty$ *and* $\sup_{p\in\mathbb{P}}\left\|\nabla^3 V(p)\right\| < \infty$*, we have,*

$$
\begin{aligned}
|||Q(\eta)-\hat{q}|||_2 &\leq \left\{\frac{\sup_{p\in\mathbb{P}}\left\|\nabla^3 V(p)\right\|_{op}(d+2)L_U}{24}+\frac{L_V L_U[(d+2)L_V]^{\frac{1}{2}}}{6}\right\}\eta^3,\\
|||P(\eta)-\hat{p}|||_2 &\leq (L_V(d+2))^{1/2}\times\left\{\frac{\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\|(L_U)^{3/2}(d+2)^{1/2}+(L_U)^{3/2}(L_V)^{1/2}}{6}\right.\\
&\qquad\left.+\ \frac{\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\|}{12}+\frac{(L_U)^{3/2}(L_V)^{1/2}}{4}\right\}\eta^3.
\end{aligned}
$$

In case of stochastic gradient HMC, Assumptions 3 allows the exchange of limit and expectation, hence,

Lemma 4.2. *Under Assumptions 3, when* $\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\| < \infty$ *and* $\sup_{p\in\mathbb{P}}\left\|\nabla^3 V(p)\right\| < \infty$*, we have,*

$$
\begin{aligned}
|||Q(\eta)-\hat{q}|||_2 &\leq \left\{\frac{\sup_{p\in\mathbb{P}}\left\|\nabla^3 V(p)\right\|_{op}(d+2)\mathbb{E}[L_U^\omega]}{24}+\frac{L_V\mathbb{E}[L_U^\omega][(d+2)L_V]^{\frac{1}{2}}}{6}\right\}\eta^3,\\
|||P(\eta)-\hat{p}|||_2 &\leq (L_V(d+2))^{1/2}\times\left\{\frac{\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\|\mathbb{E}[(L_U^\omega)^{3/2}](d+2)^{1/2}+\mathbb{E}[(L_U^\omega)^{3/2}](L_V)^{1/2}}{6}\right.\\
&\qquad\left.+\frac{\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\|}{12}+\frac{\mathbb{E}[(L_U^\omega)^{3/2}](L_V)^{1/2}}{4}\right\}\eta^3.
\end{aligned}
$$

4.2 Continuity of the Step in Probability Space

A key observation for HMC is that for a pair of starting points q_1 and q_2, the distance of probability measures, as measured for instance by the Kullback-Leibler (KL) divergence, for the next step of the algorithm is bounded by the linear order of the distance between q_1 and q_2. Therefore, when these two points are close, the next step distributions are also similar. Hence, it is easy to see that this is a key step in a conductance based argument for geometric convergence, as seen in Chen & Gatmiry (2023).

For a fixed $\mathfrak{q} \in \mathbb{R}^d$, the probability measure $\mathcal{P}_q$ of the image $Q \in \mathbb{R}^d$ can be viewed as a pushforward of the auxiliary probability measure via the integrator. Its density is given by the expression (7), where $\mathfrak{p}(\mathfrak{q}, Q)$ denotes the inverse of the integrator (as $\Pi_Q^{-f}(\mathfrak{q})$ in equation 3):

$$\mathfrak{g}(\mathfrak{p}(\mathfrak{q}, Q)) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}, Q)}{\partial Q} \right) . \tag{7}$$

For any pair $\mathfrak{q}_1, \mathfrak{q}_2 \in \mathbb{R}^d$, the Kullback-Leibler(KL) divergence $KL(\mathcal{P}_{\mathfrak{q}_1} || \mathcal{P}_{\mathfrak{q}_2})$ can be written as,

$$KL(\mathcal{P}_{\mathfrak{q}_1} || \mathcal{P}_{\mathfrak{q}_2}) = \int_{\mathbb{R}^d} \left\{ \log \mathfrak{g}(\mathfrak{p}) - \log[\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})))] - \log \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})}{\partial Q} \right) \right\} \mathfrak{g}(\mathfrak{p}) d\mathfrak{p},$$

We present Lemmata 4.3, 4.4, which are not directly employed in obtaining the main results of this paper, but they are key elements of various other approaches of convergence analysis for HMC and its variants, such as the *conductance analysis* in Chen & Gatmiry (2023) or the *probabilistic analysis* in Mangoubi & Vishnoi (2018).

Lemma 4.3. *For HMC with general auxiliary distribution, we have,*

$$KL(\mathcal{P}_{\mathfrak{q}_1} || \mathcal{P}_{\mathfrak{q}_2}) \leq \frac{\eta \|\nabla^3 V\| L_U}{2} \|q_1 - q_2\|.$$

In addition, we have the similar result for stochastic gradient estimate,

Lemma 4.4. *For SGHMC, we have*

$$KL(\mathcal{P}_{\mathfrak{q}_1} || \mathcal{P}_{\mathfrak{q}_2}) \leq \frac{\eta \|\nabla^3 V\| \mathbb{E}[L_U^\omega]}{2} \|q_1 - q_2\|.$$

The proof of Lemmata 4.3 and 4.4 can be found in Sec.D.

4.3 Lower bounding the Acceptance Probability

Another key component in all convergence analysis of HMC with numerical integrator is the estimation (bounding) of the acceptance probability of the proposed motion. This is eventually reduced to the estimation of the deviation of the numerical integrator from the exact solution in terms of the function value. More specifically, we have,

Lemma 4.5. *Under Assumptions 1 and 2, we have* $\mathbb{E}|U(\hat{q}) - U(Q(\eta))| + \mathbb{E}|V(\hat{p}) - V(P(\eta))| \leq A_3 \eta^3$*, with*

$$\begin{aligned} A_3 := & \left[L_U |||q|||_2 + (dL_U)^{1/2} \right] \left\{ \frac{(d+2) T_V L_U}{24} + \frac{L_V L_U [(d+2) L_V]^{\frac{1}{2}}}{6} \right\} + \\ & + \left[L_V |||p|||_2 + (dL_U)^{1/2} \right] (L_V (d+2))^{1/2} \left\{ \frac{T_U (L_U)^{3/2} (d+2)^{1/2} + (L_U)^{3/2} (L_V)^{1/2}}{6} + \right. \\ & \left. + \frac{T_U}{12} + \frac{(L_U)^{3/2} (L_V)^{1/2}}{4} \right\}. \end{aligned}$$

Similarly, for SGHMC, we have,

Lemma 4.6. *Under Assumptions 1 and 2 for V and Assumptions 3 for the stochastic implementations, we have $\mathbb{E}|U(\hat{q}) - U(Q(\eta))| + \mathbb{E}|V(\hat{p}) - V(P(\eta))| \leq A_3^{SG}\eta^3$, with*

$$
\begin{aligned}
A_3^{SG} := & \left\{ \frac{(d+2)T_V[\mathbb{E}[(L_U^\omega)^2] + d^{1/2}\mathbb{E}[L_U^\omega)^{3/2}]}{24} + \frac{L_V[\mathbb{E}[(L_U^\omega)^2] + d^{1/2}\mathbb{E}[(L_U^\omega)^{3/2}][(d+2)L_V]^{\frac{1}{2}}}{6} \right\} \\
& + L_V|||p|||_2(L_V(d+2))^{1/2}\left\{ \frac{\bar{T}\mathbb{E}[L_U^\omega)^{3/2}]](d+2)^{1/2} + \mathbb{E}[L_U^\omega)^{3/2}](L_V)^{1/2}}{6} \right. \\
& \left. + \frac{\bar{T}}{12} + \frac{\mathbb{E}[L_U^\omega)^{3/2}](L_V)^{1/2}}{4} \right\} \\
& + d^{1/2}\left\{ \frac{\bar{T}\mathbb{E}[(L_U^\omega)^2](d+2)^{1/2} + \mathbb{E}[(L_U^\omega)^2](L_V)^{1/2}}{6} + \frac{\bar{T}\mathbb{E}[L_U^\omega)^{1/2}]}{12} + \frac{\mathbb{E}[(L_U^\omega)^2](L_V)^{1/2}}{4} \right\}.
\end{aligned}
$$

The proofs of these Lemmata can be found in Sec.B. This naturally leads to the following result.

Proposition 4.1. *Under Assumptions 1 and 2 (Assumptions 1 and 2 for V and Assumptions 3 for SGHMC): For any $\varrho, \delta \in (0,1)$, one can choose K and η such that for subset $D \subseteq \mathbb{R}^d \times \mathbb{R}^d$ and $\mathsf{P}[(q,p) \in D] \geq 1 - \delta$, the acceptance probability is lower bounded by ϱ.*

5 Conclusions

In this paper, we analyzed the convergence of stochastic gradient HMC with alternating direction and leapfrog implementations. As more applications of HMC emerge from different areas of machine learning, we expect these results to allow the presented algorithms to be adapted more readily and with higher confidence. We also expect the analytic methods developed in the paper can be more extensively utilized in the analysis of algorithms in this domain.

References

T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13(none): 60 – 66, 2008. doi: 10.1214/ECP.v13-1352. URL `https://doi.org/10.1214/ECP.v13-1352`.

A Beskos, N Pillai, G Roberts, J. Sanz-Serna, and A. Stuart. Optimal tuning of the hybrid monte carlo algorithm. *Bernoulli*, 19:1501–1534, 2013.

N. Bou-Rabee and Marsden M. Unadjusted hamiltonian mcmc with stratified monte carlo time integration. *arXiv:2211.11003*, 2022.

Nawaf Bou-Rabee and Jesús María Sanz-Serna. Randomized hamiltonian monte carlo. *Ann. Appl. Probab.*, 27(4):2159–2194, 08 2017. doi: 10.1214/16-AAP1255. URL `https://doi.org/10.1214/16-AAP1255`.

Nawaf Bou-Rabee, Andreas Eberle, and Raphael Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *Ann. Appl. Probab.*, 30(3):1209–1250, 06 2020. doi: 10.1214/19-AAP1528. URL `https://doi.org/10.1214/19-AAP1528`.

Y. Cao, J. Lu, and L. Wang. Complexity of randomized algorithms for underdamped langevin dynamics. *Communications in Mathematical Sciences*, 19:1827–1853, 2021.

Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. In *Proceedings of the Princeton conference in honor of Professor S. Bochner*, pp. 195–199, 1969.

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*, 21:1–71, 2020.

Yuansi Chen and Khashayar Gatmiry. When does metropolized hamiltonian monte carlo provably outperform metropolis-adjusted langevin algorithm? *arXiv:2304.04724*, 2023.

Zongcheng Chen and Santosh S. Vempala. Optimal convergence rate of hamiltonian monte carlo for strongly logconcave distributions. *RANDOM*, 2019.

Masatoshi Fukushima, Yoichi Oshima, and Masayoshi Takeda. *Dirichlet Forms and Symmetric Markov Processes*. De Gruyter, Berlin, New York, 2010. ISBN 9783110218091. doi: doi:10.1515/9783110218091. URL https://doi.org/10.1515/9783110218091.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient hamiltonian monte carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 0(0):null, 2021. doi: 10.1287/opre.2021.2162. URL https://doi.org/10.1287/opre.2021.2162.

S. Ghosh, Y. Lu, and T. Nowicki. Hamiltonian Monte Carlo with asymmetrical momentum distributions. *Physica D: Nonlinear Phenomena*, to appear, arXiv:2110.12907, 2025.

Soumyadip Ghosh, Yingdong Lu, and Tomasz Nowicki. On L_2 convergence of the Hamiltonian Monte Carlo. *Applied Mathematics Letters*, 127:107811, 2022a. ISSN 0893-9659. doi: https://doi.org/10.1016/j.aml.2021.107811. URL https://www.sciencedirect.com/science/article/pii/S0893965921004377.

Soumyadip Ghosh, Yingdong Lu, and Tomasz Nowicki. On L^q convergence of the hamiltonian monte carlo. *Journal of Applied Analysis*, 2022b. doi: doi:10.1515/jaa-2022-1006. URL https://doi.org/10.1515/jaa-2022-1006.

N. Gouraud, P. Le Bris, A. Majka, and P. Monmarché. Hmc and underdamped langevin united in the unadjusted convex smooth case. *arXiv:2202.00977*, 2023.

Gregory F. Lawler and Alan D. Sokal. Bounds on the l^2 spectrum for markov chains and markov processes: A generalization of cheeger's inequality. *Transactions of the American Mathematical Society*, 309(2):557–580, 1988. ISSN 00029947. URL http://www.jstor.org/stable/2000925.

D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Soc., 2009. ISBN 9780821886274. URL https://books.google.com/books?id=6Cg5Nq5sSv4C.

Zhize Li, Tianyi Zhang, Shuyu Cheng, Jun Zhu, and Jian Li. Stochastic gradient hamiltonian monte carlo with variance reduction for bayesian inference. *Machine Learning*, 108(8):1701–1727, 2019. doi: 10.1007/s10994-019-05825-y. URL https://doi.org/10.1007/s10994-019-05825-y.

E.H. Lieb and M. Loss. *Analysis*. CRM Proceedings & Lecture Notes. American Mathematical Society, 2001. ISBN 9780821827833. URL https://books.google.com/books?id=Eb_7oRorXJgC.

L. Lovász and M. Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993. doi: https://doi.org/10.1002/rsa.3240040402. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.3240040402.

Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *The Annals of Applied Probability*, 12(2):581 – 606, 2002. doi: 10.1214/aoap/1026915617. URL https://doi.org/10.1214/aoap/1026915617.

Oren Mangoubi and Nisheeth K. Vishnoi. Dimensionally tight bounds for second-order hamiltonian monte carlo. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6030–6040, Red Hook, NY, USA, 2018. Curran Associates Inc.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer US, 2003. ISBN 9781402075537. URL https://books.google.com/books?id=VyYLem-l3CgC.

Gareth Roberts and Jeffrey Rosenthal. Geometric Ergodicity and Hybrid Markov Chains. *Electronic Communications in Probability*, 2:13 – 25, 1997. doi: 10.1214/ECP.v2-981. URL https://doi.org/10.1214/ECP.v2-981.

Gareth O. Roberts and Richard L. Tweedie. Geometric l2 and l1 convergence are equivalent for reversible markov chains. *Journal of Applied Probability*, 38(A):37–41, 2001. doi: 10.1239/jap/1085496589.

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

C. Villani. *Hypocoercivity*. Number nos. 949-951 in Hypocoercivity. American Mathematical Society, 2009. ISBN 9780821844984. URL https://books.google.com.jm/books?id=JtrNAwAAQBAJ.

Difan Zou and Quanquan Gu. On the convergence of hamiltonian monte carlo with stochastic gradients. *Proceedings of the 38th International Conference on Machine Learning*, 139:13012–13022, 2021.

In the appendix, we present some fundamental estimations that are both crucial for establishing the main results in the paper and of independent interests. First, in Sec. A, we present a detailed estimation between the exact Hamiltonian calculation and the leapfrog integrator; then, in Sec. B, we provide a lower bound to the acceptance probability. Then, the proof of the key Lemma 3.2 is presented in C. In Sec. D we quantify the dependence of the leapfrog implementations upon the initial state in terms of divergence of probability measure.

A Error Estimation for Leapfrog Gradient Calculations

In this section, we quantify the errors between the leapfrog implementation, including the case with stochastic gradient, and the exact Hamiltonian solutions. These results will be used later in multiple occasions, such as the estimation of the lower bound on acceptance probabilities and eventual convergence rates.

Error Estimation for Exact Gradient Calculation with General Auxiliary in One Leapfrog Step

In this section, we will carry out a similar error estimation for leapfrog symplectic integrator for general auxiliary distributions, but just an one-step leapfrog symplectic integrator for the Hamiltonian equation. Recall that the update takes the following form,

$$\hat{q} = q + \eta \nabla V \left(p - \frac{1}{2}\eta \nabla U(q) \right), \tag{8}$$

$$\hat{p} = p - \frac{1}{2}\eta \nabla U(q) - \frac{1}{2}\eta \nabla U(\hat{q}). \tag{9}$$

Meanwhile, the exact trajectory follows,

$$\dot{Q} = \nabla V(P), \quad \dot{P} = -\nabla U(Q); \qquad Q(0) = q, \quad P(0) = p.$$

Hence, we have,

$$Q(\eta) = q + \int_0^\eta \nabla V(P(t))dt \tag{10}$$

$$P(\eta) = p - \int_0^\eta \nabla U(Q(t))dt. \tag{11}$$

Examination of the difference between $\hat{q}$ and $Q(\eta)$ We have:

$$\hat{q} - Q(\eta) = \int_0^\eta \left[\nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right) - \nabla V(P(t))\right] dt$$

follows from equations equation 8 and equation 10

$$= \int_0^\eta \int_0^1 \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right)\left[\left(p - \frac{1}{2}\eta\nabla U(q)\right) - P(t)\right] ds\, dt$$

an application of Newton-Leibnitz

$$= \int_0^\eta \int_0^1 \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right)\left(\int_0^t \nabla U(Q(\tau))d\tau - \frac{1}{2}\eta\nabla U(q)\right) ds\, dt,$$

follows from equations equation 9 and equation 11

$$= \underbrace{\int_0^\eta \int_0^1 \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right)\left(\int_0^t \nabla[U(Q(\tau)) - \nabla U(q)]d\tau\right) ds\, dt}_{A_1} +$$

$$+ \underbrace{\int_0^\eta \int_0^1 \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right)\left(t - \frac{1}{2}\eta\right)\nabla U(q)\, ds\, dt}_{A_2}$$

write the term $\frac{1}{2}\eta\nabla U(q)$ as $\frac{1}{2}\eta\nabla U(q) - \int_0^t \nabla U(q)d\tau + t\nabla U(q)$.

Now, examine A_2

$$A_2 = \int_0^\eta \int_0^1 \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right)\left(t - \frac{1}{2}\eta\right)\nabla U(q)\, ds\, dt,$$

$$= \int_0^1 \int_0^\eta \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right)\left(t - \frac{1}{2}\eta\right)\nabla U(q)\, dt\, ds$$

exchange the order of integrations

$$= \int_0^1 \int_0^\eta \left[\nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right) - \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)p\right)\right] \cdot$$

$$\cdot \left(t - \frac{1}{2}\eta\right)\nabla U(q)\, dt\, ds +$$

$$+ \int_0^1 \int_0^\eta \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)p\right)\left(t - \frac{1}{2}\eta\right)\nabla U(q)\, dt\, ds,$$

$$= \int_0^1 \int_0^\eta \left[\nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(t)\right) - \nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)p\right)\right] \cdot$$

$$\cdot \left(t - \frac{1}{2}\eta\right)\nabla U(q)\, dt\, ds$$

because $\nabla^2 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)p\right)$ is independent of t and $\int_0^\eta (t - \frac{1}{2}\eta)dt = 0$

$$= \int_0^1 \int_0^\eta \int_0^t \nabla^3 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(\tau)\right)\nabla U(Q(\tau))\left(t - \frac{1}{2}\eta\right)\nabla U(q)\, d\tau\, dt\, ds$$

an application of Newton-Leibnitz

$$= \int_0^1 \int_0^\eta \int_\tau^\eta \nabla^3 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(\tau)\right)\nabla U(Q(\tau))\left(t - \frac{1}{2}\eta\right)\nabla U(q)\, dt\, d\tau\, ds$$

exchange the order of integrations with respect to t and τ

$$= \int_0^1 \int_0^\eta \nabla^3 V\left(s\left(p - \frac{1}{2}\eta\nabla U(q)\right) + (1-s)P(\tau)\right)\nabla U(Q(\tau))\frac{\tau}{2}(\eta - \tau)\nabla U(q)\, d\tau\, ds$$

integrate out t

The above calculations can be summarized as

Lemma A.1. *The difference between the leapfrog update with step η and the trajectory at time η is equal to:*

$$\begin{aligned}
&\hat{q} - Q(\eta) \\
=&\int_0^\eta \int_0^1 \nabla^2 V \left(s \left(p - \frac{1}{2}\eta \nabla U(q) \right) + (1-s) P(t) \right) \left(\int_0^t \nabla [U(Q(\tau)) - \nabla U(q)] d\tau \right) \, ds \, dt \\
&+ \int_0^1 \int_0^\eta \nabla^3 V \left(s \left(p - \frac{1}{2}\eta \nabla U(q) \right) + (1-s) P(\tau) \right) \nabla U(Q(\tau)) \frac{\tau}{2} (\eta - \tau) \nabla U(q) \, d\tau \, ds
\end{aligned} \tag{12}$$

This will for the basic for quantifying $|||Q(\eta) - \hat{q}|||_2$.

Lemma A.2. *Under Assumptions 1 and 2, we have,*

$$|||Q(\eta) - \hat{q}|||_2 \leq \left\{ \frac{(d+2) L_U T_V}{24} + \frac{L_V L_U [(d+2) L_V]^{\frac{1}{2}}}{6} \right\} \eta^3, \tag{13}$$

Proof. Let us look at the two terms in equation 12.

$$\begin{aligned}
&\left| \left| \left| \int_0^\eta \int_0^1 \nabla^2 V \left(s \left(p - \frac{1}{2}\eta \nabla U(q) \right) + (1-s) P(t) \right) \left(\int_0^t \nabla [U(Q(\tau)) - \nabla U(q)] d\tau \right) \, ds \, dt \right| \right|_2 \\
=&\left| \left| \left| \int_0^\eta \int_0^1 \nabla^2 V \left(s \left(p - \frac{1}{2}\eta \nabla U(q) \right) + (1-s) P(t) \right) \left(\int_0^t \int_0^\tau \nabla^2 U(Q(w)) \nabla V(P(w)) dw; d\tau \right) \, ds \, dt \right| \right|_2 \\
&\text{an application of Newton-Leibnitz} \\
\leq &\int_0^\eta \int_0^1 \int_0^t \int_0^\tau \left| \left| \left| \nabla^2 V \left(s \left(p - \frac{1}{2}\eta \nabla U(q) \right) + (1-s) P(t) \right) \nabla^2 V(P(w)) \nabla V(P(w)) \right| \right| \right|_2 dw; d\tau \, ds \, dt \\
\leq &\int_0^\eta \int_0^1 \int_0^t \int_0^\tau L_U L_V \left| \left| \left| \nabla V(P(w)) \right| \right| \right|_2 dw \, d\tau \, ds \, dt \\
&\text{due to assumption 1} \\
\leq &L_V L_U [(d+2) L_V]^{\frac{1}{2}} / 6. \\
&\text{due to Lemma2.1 and the fact that the joint probability is invariant under the Hamilton motion}
\end{aligned}$$

$$\begin{aligned}
&\left| \left| \left| \int_0^1 \int_0^\eta \nabla^3 V \left(s \left(p - \frac{1}{2}\eta \nabla U(q) \right) + (1-s) P(\tau) \right) \nabla U(Q(\tau)) \frac{\tau}{2} (\eta - \tau) \nabla U(q) \, d\tau \, ds \right| \right|_2 \\
\leq &\int_0^1 \int_0^\eta \|\nabla^3 V\| \frac{\tau}{2} (\eta - \tau) \mathbb{E} \left[\left| \left| \nabla U(Q(\tau)) \right| \right|_2 \cdot \left| \left| \nabla U(q) \right| \right|_2 \right] \, d\tau \, ds \\
&\text{assumption on the operator norm of the third degree tensor} \\
\leq &\int_0^1 \int_0^\eta \|\nabla^3 V\| \frac{\tau}{2} (\eta - \tau) \left| \left| \left| \nabla U(Q(\tau)) \right| \right| \right|_2 \cdot \left| \left| \left| \nabla U(q) \right| \right| \right|_2 d\tau \, ds \\
&\text{Hölder's inequality} \\
\leq &\|\nabla^3 V\| L_U (d+2) / 24.
\end{aligned}$$

□

Examine the difference between $\hat{p}$ and $P(\eta)$

Again, let us start with a decomposition of the term $P(\eta) - \hat{p}$,

$$\begin{aligned}
\hat{p} - P(\eta) &= \int_0^\eta \nabla U(Q(t)) - \frac{1}{2}\nabla U(q) - \frac{1}{2}\nabla U(\hat{q})\, dt \\
&\qquad \text{follows from equations equation 9 and equation 11} \\
&= \int_0^\eta [\nabla U(Q(t)) - \nabla U(q)]\, dt - \frac{1}{2}\int_0^\eta [\nabla U(q) - \nabla U(\hat{q})]\, dt \\
&= \int_0^\eta \int_0^t \nabla^2 U(Q(s))\nabla V(P(s))\, ds\, dt - \frac{1}{2}\int_0^\eta [\nabla U(q) - \nabla U\left(q + \eta \nabla V\left(p - \frac{1}{2}\eta \nabla U(q)\right)\right)]\, dt \\
&\qquad \text{an application of Newton-Leibnitz} \\
&= \underbrace{\int_0^\eta \int_0^t \nabla^2 U(Q(s))\nabla V(P(s))\, ds\, dt - \frac{1}{2}\int_0^\eta [\nabla U(q) - \nabla U(q + \eta\nabla V(p))]\, dt}_{B_1} \\
&\quad + \underbrace{\frac{1}{2}\int_0^\eta \left[\nabla U(q + \eta\nabla V(p)) - \nabla U\left(q + \eta\nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right)\right)\right]\, dt}_{B_2}.
\end{aligned}$$

$$\begin{aligned}
B_2 =& \frac{1}{2}\int_0^\eta \left[\nabla U(q + \eta\nabla V(p)) - \nabla U\left(q + \eta\nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right)\right)\right]\, dt \\
=& \frac{1}{2}\int_0^\eta \int_0^\eta \left[\nabla^2 U\left(q + s\nabla V(p) + (1-s)\nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right)\right)\left[\nabla V(p) - \nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right)\right]\right]\, ds\, dt. \\
&\text{an application of Newton-Leibnitz}
\end{aligned}$$

B_2 is clearly a η^3 term.

$$\begin{aligned}
B_1 &= \int_0^\eta \int_0^t \nabla^2 U(Q(s))\nabla V(P(s))\, ds\, dt - \frac{1}{2}\int_0^\eta [\nabla U(q) - \nabla U(q + \eta\nabla V(p))]\, dt \\
&= \int_0^\eta \int_0^t \nabla^2 U(Q(s))\nabla V(P(s))\, ds\, dt - \frac{1}{2}\int_0^\eta \int_0^\eta [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\, dt \\
&\qquad \text{an application of Newton-Leibnitz} \\
&= \underbrace{\int_0^\eta \int_0^t \nabla^2 U(Q(s))\nabla V(P(s))\, ds\, dt - \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\, ds\, dt}_{B_{11}} \\
&\quad + \underbrace{\int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\, ds\, dt - \frac{1}{2}\int_0^\eta \int_0^\eta [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\, dt}_{B_{12}}.
\end{aligned}$$

For B_{11}, we have,

$$\begin{aligned}
B_{11} &= \int_0^\eta \int_0^t \nabla^2 U(Q(s))\nabla V(P(s))\, ds\, dt - \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\, ds\, dt \\
&= \int_0^\eta \int_0^t [\nabla^2 U(Q(s)) - \nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\, ds\, dt
\end{aligned}$$

$$
\begin{aligned}
B_{12} =& \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\ ds\ dt - \frac{1}{2}\int_0^\eta \int_0^\eta [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ dt \\
=& \underbrace{\int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\ ds\ dt - \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ ds\ dt}_{B_{121}} \\
&+ \underbrace{\int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ ds\ dt - \frac{1}{2}\int_0^\eta \int_0^\eta [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ dt}_{B_{122}}.
\end{aligned}
$$

Hence, we have the following expression,

Lemma A.3.

$$
\begin{aligned}
& P(\eta) - \hat{p} \\
=& \int_0^\eta \int_0^t [\nabla^2 \nabla^2 U(Q(s)) - U(q + s\nabla V(p))]\nabla V(P(s))\ ds\ dt \\
&+ \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))\ ds\ dt - \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ ds\ dt \\
&+ \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ ds\ dt - \frac{1}{2}\int_0^\eta \int_0^\eta [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ dt \\
&+ \frac{1}{2}\int_0^\eta \int_0^\eta \left[\nabla^2 U\left(q + s\nabla V(p) + (1-s)\nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right)\right)\left[\nabla V(p) - \nabla V\left(p - \frac{1}{2}\eta\nabla U(q)\right)\right]\right]\ ds\ dt.
\end{aligned}
$$

Lemma A.4. *Under Assumptions 1 and 2, we have,*

$$
\begin{aligned}
& |||P(\eta) - \hat{p}|||_2 \\
\leq& (L_V(d+2))^{1/2}\left\{\frac{T_U (L_U)^{3/2}(d+2)^{1/2} + (L_U)^{3/2}(L_V)^{1/2}}{6} + \frac{T_U}{12} + \frac{(L_U)^{3/2}(L_V)^{1/2}}{4}\right\}\eta^3. \qquad (14)
\end{aligned}
$$

Proof. From above derivations, we know that, $|||P(\eta) - \hat{p}|||_2 \leq |||B_{11}|||_2 + |||B_{121}|||_2 + |||B_{122}|||_2 + |||B_2|||_2$.Therefore,

$$
\begin{aligned}
|||B_{11}|||_2 \leq& \int_0^\eta \int_0^t |||\nabla^2 U(Q(s))\nabla V(P(s)) - [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s))|||_2\ ds\ dt \\
\leq& \sup_{q\in\mathbb{Q}} \left\|\nabla^3 U(q)\right\| \int_0^\eta \int_0^t \int_0^s \mathbb{E}\left[\left\|\nabla V(P(\tau)) - \nabla V(p)\right\|_2 \cdot \left\|\nabla V(P(s))\right\|_2\right]\ ds\ dt \\
& \text{assumption on the operator norm of the third degree tensor} \\
\leq& \sup_{q\in\mathbb{R}^d} \left\|\nabla^3 U(q)\right\| \int_0^\eta \int_0^t \int_0^s \left|\left|\left|\nabla V(P(\tau)) - \nabla V(p)\right|\right|\right|_2 \cdot \left|\left|\left|\nabla V(P(s))\right|\right|\right|_2\ ds\ dt \\
& \text{Cauchy-Schwartz inequality} \\
\leq& \frac{\sup_{q\in\mathbb{R}^d}\left\|\nabla^3 U(q)\right\|(L_U)^{3/2}(L_V)^{1/2}(d+2)}{6}\eta^3. \\
& \text{Lemma 2.1}
\end{aligned}
$$

$$
\begin{aligned}
|||B_{121}|||_2 \leq & \int_0^\eta \int_0^t \left|\left|\left| [\nabla^2 U(q + s\nabla V(p))]\nabla V(P(s)) - \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(p) \right|\right|\right|_2 ds\ dt \\
\leq & L_U \int_0^\eta \int_0^t \left|\left|\left| \nabla V(P(s)) - \nabla V(p) \right|\right|\right|_2 ds\ dt \\
& \text{assumption on } U \\
\leq & L_U L_V \int_0^\eta \int_0^t \left|\left|\left| P(s) - p \right|\right|\right|_2 ds\ dt \\
& \text{assumption on } V \\
\leq & L_U L_V \int_0^\eta \int_0^t \int_0^s |||\nabla U(\tau)|||_2 \, d\tau\ ds\ dt \\
& \text{an application of Newton-Leibnitz} \\
\leq & \frac{(L_U)^{3/2} L_V (d+2)^{1/2}}{6} \eta^3.
\end{aligned}
$$

Apply Lemma A.5, we have,

$$
\begin{aligned}
|||B_{122}|||_2 \leq & \left|\left|\left| \int_0^\eta \int_0^t [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ ds\ dt - \frac{1}{2} \int_0^\eta \int_0^\eta [\nabla^2 U(q + s\nabla V(p))]\nabla V(p)\ dt \right|\right|\right|_2 \\
\leq & \frac{\sup_{q \in \mathbb{R}^d} \left\| \nabla^3 U(q) \right\| (L_V)^{1/2} (d+2)^{1/2}}{12} \eta^3.
\end{aligned}
$$

$$
\begin{aligned}
|||B_2|||_2 \leq & \frac{1}{2} \int_0^\eta \int_0^\eta \left|\left|\left| \nabla^2 U \left(q + s\nabla V(p) + (1-s)\nabla V \left(p - \frac{1}{2}\eta \nabla U(q) \right) \right) \left[\nabla V(p) - \nabla V \left(p - \frac{1}{2}\eta \nabla U(q) \right) \right] \right|\right|\right|_2 ds\ dt \\
\leq & \frac{1}{2} \int_0^\eta \int_0^\eta L_U \left|\left|\left| \left[\nabla V(p) - \nabla V \left(p - \frac{1}{2}\eta \nabla U(q) \right) \right] \right|\right|\right|_2 ds\ dt \\
\leq & \frac{(L_U)^{3/2} L_V (d+2)^{1/2}}{4} \eta^3.
\end{aligned}
$$

□

The following technical lemma and its proof are included for completeness.

Lemma A.5. *For a locally integrable function $f(\cdot)$, we have,*

$$
\int_0^\eta \int_0^t f(s)\ ds\ dt - \frac{1}{2} \int_0^\eta \int_0^\eta f(s)\ ds\ dt = \int_0^\eta \frac{\tau}{2} (\tau - \eta) f'(\tau)\ d\tau.
$$

Proof.

$$
\begin{aligned}
&\int_0^\eta \int_0^t f(s)\, ds\, dt - \frac{1}{2}\int_0^\eta \int_0^\eta f(s)\, ds\, dt \\
=&\int_0^\eta \int_s^\eta f(s)\, dt\, ds - \frac{\eta}{2}\int_0^\eta f(s)\, ds \\
=&\int_0^\eta f(s)(\eta - s)\, ds - \frac{\eta}{2}\int_0^\eta f(s)\, ds \\
=&\int_0^\eta f(s)(\frac{\eta}{2} - s)\, ds \\
=&\int_0^\eta [f(s) - f(0)](\frac{\eta}{2} - s)\, ds \\
=&\int_0^\eta \int_0^s f'(\tau)\, d\tau(\frac{\eta}{2} - s)\, ds \\
=&\int_0^\eta \int_\tau^\eta f'(\tau)(\frac{\eta}{2} - s)\, ds\, d\tau \\
=&\int_0^\eta \frac{\tau}{2}(\tau - \eta) f'(\tau)\, d\tau.
\end{aligned}
$$

□

B Lower Bounding the Acceptance Probability

From the expression of the acceptance/rejection probability calculationequation 2, we know that it suffices to show that there exists $a > 0$, such that, $|U(q_K) + V(p_K) - U(q_0) - V(p_0)| \leq a$. Meanwhile, since the Hamiltonian is invariant for the exact solution, hence this quantity become the difference between the exact Hamiltonian and that of the symplectic integrator. In essence, we need to estimate $|U(\hat{q}) - U(Q(\eta))|$ and $|V(\hat{p}) - V(P(\eta)|$. Lemmata A.2 and A.4 imply that these terms are of order η^3. Then the desired results follows. This is a similar result to Chen & Gatmiry where they say that, for general subexponential target probability distribution, there exists a compact set $\Lambda \in \mathbb{R}^d \times \mathbb{R}^d$, such that when the initial point $(q_0, p_0) \in \Lambda$, the acceptance can be bounded from below.

$$
\begin{aligned}
\mathbb{E}|U(\hat{q}) - U(Q(\eta))| \leq& \mathbb{E}\left| \int_0^1 \nabla U(q_s) \cdot [\hat{q} - Q(\eta)] ds \right| \\
\leq& \mathbb{E}\left| \int_0^1 [\nabla U(q_s) - \nabla U(q)] \cdot [\hat{q} - Q(\eta)] ds \right| + \int_0^1 \mathbb{E}\left| \nabla U(q) \cdot [\hat{q} - Q(\eta)] ds \right| \\
\leq& L_U \mathbb{E}\left| \int_0^1 \int_0^s [q_\tau - q] \cdot [\hat{q} - Q(\eta)]\, d\tau\, ds \right| + \int_0^1 \mathbb{E}\left| \nabla U(q) \cdot [\hat{q} - Q(\eta)] ds \right| \\
\leq& L_U \int_0^1 \int_0^s |||q_\tau - q|||_2 |||[\hat{q} - Q(\eta)|||_2\, d\tau\, ds + |||\nabla U(q)|||_2 |||[\hat{q} - Q(\eta)|||_2 \\
\leq& (L_U |||q|||_2 + |||\nabla U(q)|||_2)|||)|||\hat{q} - Q(\eta)|||_2 \\
\leq& [L_U |||q|||_2 + (dL_U)^{1/2}] \left\{ \frac{T_V (d+2) L_U}{24} + \frac{L_V L_U [(d+2)L_V]^{\frac{1}{2}}}{6} \right\} \eta^3,
\end{aligned}
$$

with $q_s = (s\hat{q} + (1-s)Q(\eta))$. All the quantities are available through Lemmata 2.1 and A.2. By the same arguments, we can obtain the following estimate for $\mathbb{E}|V(\hat{p}) - V(P(\eta)|$ which can be bounded further with

Lemmata 2.1 and A.4

$$\begin{aligned}
&\mathbb{E}|V(\hat{p}) - V(P(\eta))| \\
\leq&(L_V|||p|||_2 + |||\nabla V(q)|||_2)|||)|||\hat{p} - P(\eta)|||_2 \\
\leq&[L_V|||p|||_2 + (dL_U)^{1/2}](L_V(d+2))^{1/2}\left\{\frac{T_U(L_U)^{3/2}(d+2)^{1/2} + (L_U)^{3/2}(L_V)^{1/2}}{6}\right. \\
&\left.+\frac{T_U}{12} + \frac{(L_U)^{3/2}(L_V)^{1/2}}{4}\right\}\eta^3.
\end{aligned}$$

C Proof of Lemma 3.2

Proof of Lemma 3.2. We will first show that equation 6 holds for Schwartz functions $h \in L^2(\mathbb{R}^d, \mathfrak{f}(q)dq)$ with uniform constants and $K = 1$, then a mollification procedure, see e.g. Lieb & Loss (2001) establishes equation 6 for all $h \in L^2(\mathbb{R}^d, \mathfrak{f}(q)dq)$. Repeat the arguments, equation 6 can be obtained for general K. First,

$$\begin{aligned}
&\int_{\mathbb{R}^d} h(q)\int_{\mathbb{R}^d}[h(q) - h(\hat{q})]g(p)dpdq \\
=&\int_{\mathbb{R}^d} h(q)\int_{\mathbb{R}^d}\left[h(q) - h(q + \eta\nabla V(p - \frac{\eta}{2}\nabla U(q)))\right]g(p)dpdq \\
=&\underbrace{\int_{\mathbb{R}^d} h(q)\int_{\mathbb{R}^d}[h(q) - h(q + \eta\nabla V(p))]g(p)dpdq}_{III_1} \\
&+\underbrace{\int_{\mathbb{R}^d} h(q)\int_{\mathbb{R}^d}\left[h(q + \eta\nabla V(p)) - h(q + \eta\nabla V(p - \frac{\eta}{2}\nabla U(q)))\right]g(p)dpdq}_{III_2}.
\end{aligned}$$

For III_1 and III_2

$$\begin{aligned}
III_1 =&\int_{\mathbb{R}^d} h(q)\int_{\mathbb{R}^d}[h(q) - h(q + \eta\nabla V(p))]g(p)dpdq \\
\overset{(1)}{=}&\underbrace{-\frac{\eta^2}{2}\int_{\mathbb{R}^d}\nabla V(p)\cdot\left[\int_{\mathbb{R}^d} h(q)\nabla^2 h(q)f(q)dq\right]\cdot\nabla V(p)g(p)dp}_{III_{11}} \\
&+\underbrace{\frac{\eta^2}{2}\int_{\mathbb{R}^d}\nabla V(p)\cdot\left[\int_{\mathbb{R}^d} h(q)[\nabla^2 h(q) - \nabla^2 h(\tilde{q})]f(q)dq\right]\cdot\nabla V(p)g(p)dp}_{III_{12}}.
\end{aligned}$$

with (1) is due to the zero mean assumption of p and mean value theorem with $\tilde{q}$.

Define the mean joint partial derivatives $\mu_{ij} = \int_{\mathbb{R}^d}\frac{\partial V(p)}{\partial p_i}\frac{\partial V(p)}{\partial p_j}\mathfrak{g}(p)dp$ and the mean second derivative $\sigma_{ij} = \int_{\mathbb{R}^d}\frac{\partial^2 V(p)}{\partial p_i \partial p_j}\mathfrak{g}(p)dp$. Note that Assumption 4 gives that $\mu_{ij} = \sigma_{ij}$ for all i, j.

Then, for III_{11} we have

$$\begin{aligned}
III_{11} =&-\frac{\eta^2}{2}\int_{\mathbb{R}^d}\nabla V(p)\cdot\left[\int_{\mathbb{R}^d} h(q)\nabla^2 h(q)f(q)dq\right]\cdot\nabla V(p)g(p)dp \\
=&\frac{\eta^2}{2}\int_{\mathbb{R}^d}\nabla V(p)\cdot\int_{\mathbb{R}^d}[\nabla h(q)\otimes\nabla h(q)f(q)] + h(q)\nabla h(q)\otimes\nabla f(q)]dq\cdot\nabla V(p)g(p)dp \\
=&\frac{\eta^2\sigma_p^2}{2}\int_{\mathbb{R}^d}\left[\|\nabla h(q)\|_2^2 - \sum_{i,j} h(q)\partial_i h(q)\partial_j U(q)\right]f(q)dq\cdot\mu_{ij}.
\end{aligned}$$

For III_{12}, from a direct application of the Hólder's inequality uniform bounded on third order derivatives on the mollified function, see, e.g. Lieb & Loss (2001). we know that, there exists a $C_2 > 0$, such that,

$$\left\| \int_{\mathbb{R}^d} h(q) \|\nabla^2 h(q) - \nabla^2 h(\tilde{q})\|_F f(q) dq \right\| \leq C_2 \eta \|h\|_2^2.$$

For III_2, we have,

$$\int_{\mathbb{R}^d} h(q)[\nabla h(q) \cdot \nabla U(q)] f(q) dq \cdot \int_{\mathbb{R}^d} \nabla^2 V(p) g(p) dp, = \sum_{i,j} \int_{\mathbb{R}^d} h(q)[h(q)\partial_i h(q) \partial_j U(q)] f(q) dq \cdot \sigma_{ij}.$$

Therefore, the cancellation follows from the assumption $\sigma_{ij} = \mu_{ij}$. Meanwhile

$$\begin{aligned} III_2 &= \int_{\mathbb{R}^d} h(q) \int_{\mathbb{R}^d} \left[h(q + \eta \nabla V(p)) - h(q + \eta \nabla V(p - \frac{\eta}{2} \nabla U(q))) \right] g(p) dp dq \\ &= \int_{\mathbb{R}^d} h(q) \nabla h(q) \cdot \nabla U(q)] f(q) dq \cdot \int_{\mathbb{R}^d} \nabla^2 V(p) g(p) dp, \end{aligned}$$

by the convexity assumption on $U(q)$. Therefore, we have, $\int_{\mathbb{R}^d} h(q) \int_{\mathbb{R}^d} [h(q) - h(\hat{q})] g(p) dp dq \geq \frac{\varrho \eta^2}{2} C_1 \sigma_V^2 \|h\|_2$ with C_1 being the optimal Poincaré inequality constant, and $\sigma_V^2 := \int_{\mathbb{R}^d} \|\nabla V(p)\|_2 g(p) dp$.

Denote $A(q,p)$ as the event that the proposal is accepted. $\|h\|_2^2 - \langle h, M_H h \rangle = \int_{\mathbb{R}^d} h(q) \int_{\mathbb{R}^d} [h(q) - h(\hat{q})] f(q) g(p) dp dq + \int_{\mathbb{R}^d} h(q) \int_{\mathbb{R}^d} [1 - \mathbf{1}_{A(q,p)}][h(q) - h(\hat{q})] g(p) dp dq$. We can bound the second term with higher order of $A_3 \eta^3 \|h\|_2^2$ (at least η^3) using Hölder's inequality and Lemma 4.5. Therefore, we have, $\|h\|_2^2 - \langle h, M_H h \rangle \geq \eta^2 (\frac{C_1 \sigma_V^2}{2} - A_3 \eta) \|h\|_2^2$. □

D Density of Pushforward Auxiliary Distributions

Fix $\mathfrak{q} \in \mathbb{R}^d$, the probability measure $\mathcal{P}_q$ of the image $Q \in \mathbb{R}^d$ can be viewed as a pushforward of the auxiliary probability measure via the integrator, therefore, its density bears the following form,

$$\mathfrak{g}(\mathfrak{p}(\mathfrak{q}, Q)) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}, Q)}{\partial Q} \right), \tag{15}$$

with $\mathfrak{p}(\mathfrak{q}, Q)$ denotes the inverse of the integrator.

Kullback-Leibler(KL) divergence calculation

For any pair $\mathfrak{q}_1, \mathfrak{q}_2 \in \mathbb{R}^d$, the Kullback-Leibler(KL) divergence $KL(\mathcal{P}_{\mathfrak{q}_1} || \mathcal{P}_{\mathfrak{q}_2})$ can be written as,

$$\begin{aligned} KL(\mathcal{P}_{\mathfrak{q}_1} || \mathcal{P}_{\mathfrak{q}_2}) &= \int_{\mathbb{R}}^d \mathfrak{g}(\mathfrak{p}(\mathfrak{q}_1, Q)) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_1, Q)}{\partial Q} \right) \log \left(\frac{\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_1, Q)) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_1, Q)}{\partial Q} \right)}{\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_2, Q)) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_2, Q)}{\partial Q} \right)} \right) dQ \\ &\overset{(1)}{=} \int_{\mathbb{P}} \log \left(\frac{\mathfrak{g}(\mathfrak{p}) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_1, Q(\mathfrak{q}_1, \mathfrak{p}))}{\partial Q} \right)}{\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p}))) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})}{\partial Q} \right)} \right) \mathfrak{g}(\mathfrak{p}) d\mathfrak{p} \\ &\overset{(2)}{=} \int_{\mathbb{P}} \log \left(\frac{\mathfrak{g}(\mathfrak{p})}{\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p}))) \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})}{\partial Q} \right)} \right) \mathfrak{g}(\mathfrak{p}) d\mathfrak{p} \\ &= \int_{\mathbb{P}} \left(\log \mathfrak{g}(\mathfrak{p}) - \log[\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})))] - \log \det \left(\frac{\partial \mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})}{\partial Q} \right) \right) \mathfrak{g}(\mathfrak{p}) d\mathfrak{p}, \end{aligned}$$

equation (1) is the result of change of variable from Q to p, and (2) is due to the fact that,

$$\frac{\partial \mathfrak{p}(\mathfrak{q}_1, Q(\mathfrak{q}_1, \mathfrak{p}))}{\partial Q} = Id.$$

Note that, conceptually, the term $\mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p}))$ is treated as perturbation of $\mathfrak{p}$, denoted as $\tilde{\mathfrak{p}} = \mathfrak{p} + \epsilon$, then it can be see that

$$\begin{aligned}
&\int_{\mathbb{P}} \left(\log \mathfrak{g}(\mathfrak{p}) - \log[\mathfrak{g}(\mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})))] - \log\det\left(\frac{\partial \mathfrak{p}(\mathfrak{q}_2, Q(\mathfrak{q}_1, \mathfrak{p})}{\partial Q} \right) \right) \mathfrak{g}(\mathfrak{p}) d\mathfrak{p} \\
=&\int_{\mathbb{P}} \left(\log\left(\frac{\mathfrak{g}(\mathfrak{p})}{\mathfrak{g}(\tilde{\mathfrak{p}})} \right) - \log\det \partial \mathfrak{g}(\tilde{\mathfrak{p}}) \right) \mathfrak{g}(\mathfrak{p}) d\mathfrak{p} \\
=&\int_{\mathbb{P}} \left(\log\left(1 + \epsilon_1\right) - \log\det(I + \epsilon_2) \right) \mathfrak{g}(\mathfrak{p}) d\mathfrak{p}.
\end{aligned}$$

Now, let us examine them more carefully. Recall that $\mathfrak{g}(\mathfrak{p}) = \exp[-V(\mathfrak{p})]$, so,

$$\log\left(\frac{\mathfrak{g}(\mathfrak{p})}{\mathfrak{g}(\tilde{\mathfrak{p}})} \right) = V(\tilde{\mathfrak{p}}) - V(\mathfrak{p}). \tag{16}$$

Let us see how $\tilde{\mathfrak{p}}$ is calculated. We start with the case that only one leapfrog step is taken,

$$\begin{aligned}
\hat{q}_1 =& q_1 + \eta \nabla V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right), \\
\hat{p}_1 =& p - \frac{1}{2}\eta \nabla U(q_1) - \frac{1}{2}\eta \nabla U(\hat{q}_1).
\end{aligned}$$

So $\tilde{\mathfrak{p}}$ satisfies,

$$q_2 + \eta \nabla V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) = q_1 + \eta \nabla V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right).$$

Therefore, we can compute $\frac{\partial \tilde{\mathfrak{p}}}{\partial \mathfrak{p}}$,

$$\eta \nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) \frac{\partial \tilde{\mathfrak{p}}}{\partial \mathfrak{p}} = \eta \nabla^2 V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right),$$

Hence,

$$\frac{\partial \tilde{\mathfrak{p}}}{\partial \mathfrak{p}} = \left(\nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) \right)^{-1} \nabla^2 V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right).$$

Furthermore, we know that,

$$\frac{\partial \mathfrak{p}(\mathfrak{q}_1, Q)}{\partial Q} = \frac{\partial \tilde{\mathfrak{p}}}{\partial \mathfrak{p}} \cdot \frac{\partial \mathfrak{p}}{\partial Q}.$$

These calculations will make an estimation of equation 16 possible. More specifically, Hessian Lipschitz condition will lead to

$$\begin{aligned}
\frac{\partial \tilde{\mathfrak{p}}}{\partial \mathfrak{p}} =& \left(\nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) \right)^{-1} \nabla^2 V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right) \\
=& I - \left(\nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) \right)^{-1} \left[\nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) - \nabla^2 V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right) \right].
\end{aligned}$$

We know that

$$\left(\nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) \right)^{-1}$$

is bounded ($\succeq mI$ by strong convexity). Therefore, we only need to bound

$$\left\| \nabla^2 V \left(\tilde{\mathfrak{p}} - \frac{1}{2}\eta \nabla U(q_2) \right) - \nabla^2 V \left(p - \frac{1}{2}\eta \nabla U(q_1) \right) \right\|.$$

Hessian Lipschitz condition leads to

$$\begin{aligned}&\left\|\nabla^2 V\left(\tilde{\mathfrak{p}}-\frac{1}{2}\eta\nabla U(q_2)\right)-\nabla^2 V\left(p-\frac{1}{2}\eta\nabla U(q_1)\right)\right\|\\ \leq&\frac{\|\nabla^3 V\|\eta}{2}\|\nabla U(q_1)-\nabla U(q_2)\|\leq\frac{\eta\|\nabla^3 V\|L_U}{2}\|q_1-q_2\|.\end{aligned}\tag{17}$$

Inequality equation 17 thus give Lemmata 4.3 and 4.4, similar to Lemma 2 in Chen & Gatmiry (2023).