Condition-number-independent convergence rate of Riemannian Hamiltonian Monte Carlo with numerical integrators

Yunbum Kook YB.KOOK@GATECH.EDU
Georgia Institute of Technology

Yin Tat Lee YINTAT@UW.EDU
University of Washington and Microsoft Research

Ruoqi Shen SHENR3@CS.WASHINGTON.EDU
University of Washington

Santosh Vempala VEMPALA@GATECH.EDU
Georgia Institute of Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We study the convergence rate of discretized Riemannian Hamiltonian Monte Carlo on sampling from distributions in the form of $e^{-f(x)}$ on a convex body $\mathcal{M} \subset \mathbb{R}^n$. We show that for distributions in the form of $e^{-\alpha^\top x}$ on a polytope with m constraints, the convergence rate of a family of commonly-used integrators is independent of $\|\alpha\|_2$ and the geometry of the polytope. In particular, the implicit midpoint method (IMM) and the generalized Leapfrog method (LM) have a mixing time of $\widetilde{O}\left(mn^3\right)$ to achieve ϵ total variation distance to the target distribution. These guarantees are based on a general bound on the convergence rate for densities of the form $e^{-f(x)}$ in terms of parameters of the manifold and the integrator. Our theoretical guarantee complements the empirical results of Kook et al. (2022), which shows that RHMC with IMM can sample ill-conditioned, non-smooth and constrained distributions in very high dimension efficiently in practice.

Keywords: Sampling, Markov Chain Monte Carlo, Riemannian Hamiltonian Monte Carlo

1. Introduction

Efficient sampling from high dimensional distributions is a fundamental question that arises in many fields such as statistics, machine learning, and theoretical computer science. One class of distributions that arises in many applications is constrained distributions, where the distribution is defined on a constrained set. Sampling from such distribution can be an efficient way to study the geometric properties of the constrained set when direct computation is not feasible. For instance, in systems biology, a metabolic network is defined by a set of equalities and inequalities that represents feasible steady state reaction rates (Lewis et al., 2012; Thiele et al., 2013). For large metabolic networks, sampling from the constraint set can be an efficient way to simulate the biochemical network and evaluate its capacity. In mathematics, computing the volume of the Birkhoff polytope plays a key role in several areas, including algebraic geometry, and probability. However, computing the volume exactly using algebraic representations can take years even for a small dimension $n = 11$. On the other hand, a sampling based-algorithm can compute the volume efficiently up to dimension half-million (Kook et al., 2022).

Traditional samplers The current primary approach for sampling is Markov Chain Monte Carlo (MCMC) method, which for many problems is the only known method with provable efficiency guarantees. For general non-smooth distributions, the traditional class of samplers is the zeroth-order samplers, which query the density of the distributions to determine the algorithm's trajectory. This class of samplers includes Ball walk (Lovász and Simonovits, 1993; Kannan et al., 1997), its affine-invariant version Dikin walk (Kannan and Narayanan, 2012; Laddha et al., 2020) and Hit-and-Run (Smith, 1984; Lovász, 1999), which avoids an explicit step size. However, this class of sampler is inefficient in practice because it intrinsically needs a step size smaller than $O(1/\sqrt{n})$, where n is the dimension, to avoid stepping outside the constraint set, which leads to a bottleneck of quadratic mixing time in dimension. Moreover, without putting the convex body into an isotropic position, which requires expensive computation in practice, the mixing time of Ball walk and Hit-and-Run, $\widetilde{O}(n^2R^2)$, depends on the condition number R of the convex body. The condition number of the distributions appearing in practical applications can be large, *e.g.,* the condition number of RECON1 (King et al., 2016), a human metabolic network, can be as large as 10^6. Using Hit-and-Run to sample from metabolic networks can be over 100 times slower than the better algorithms on this problem (Cousins and Vempala, 2016). Sampling from the Birkhoff polytope can be prohibitively expensive for any dimension higher than $n = 20$ (Cousins and Vempala, 2016).

Another class of samplers commonly used is the first-order samplers, which update the Markov chain based on the gradient information. The mixing time of the continuous processes as well as the various discretization methods of this class of samplers has been studied in a long line of recent works. The most well-studied first-order samplers include Langevin algorithm (Dalalyan, 2017; Dwivedi et al., 2018; Durmus et al., 2019; Vempala and Wibisono, 2019; Chewi et al., 2021, 2022), its variant Underdamped Langevin algorithm (Cheng et al., 2018; Shen and Lee, 2019), and Hamiltonian Monte Carlo (HMC) (Chen and Vempala, 2022; Chen et al., 2020; Lee et al., 2020). The mixing time of this class of samplers also suffers from dependence on the condition number of the distributions. Moreover, this class of samplers cannot be applied to constrained distributions directly because their Markov chain can easily step outside the constraint set. Currently, popular sampling packages such as Stan (Stan Development Team, 2020) and Pyro (Bingham et al., 2019) that are based on this class of samplers are not able to handle constrained distributions, despite their effectiveness in other settings.

Non-Euclidean Samplers Given the limitations of the traditional samplers, researchers have sought to extend these methods to non-Euclidean samplers, which leverage the local geometry of distributions to speed up the samplers. For instance, Riemannian Hamiltonian Monte Carlo (RHMC) extends the traditional HMC by considering the dynamics on a Riemannian manifold that uses a non-Euclidean metric corresponding to the distribution's local geometry. When combined with a local metric induced by the Hessian of a self-concordant barrier function, RHMC can sample from ill-conditioned and non-smooth distributions efficiently. A recent work (Kook et al., 2022) showed that RHMC can achieve a 1000-fold acceleration on the benchmark dataset RECON3D (King et al., 2016), the largest published human metabolic network, compared to previous methods. While RHMC has demonstrated superior practical performance, the convergence rate of discretized RHMC remains open. Lee and Vempala (2018) bounded the convergence rate of continuous RHMC in terms of the isoperimetry and natural smoothness parameters of the associated Riemannian manifold. However, to implement RHMC, sophisticated integrators such as implicit midpoint integrator (IMM) or the generalized Leapfrog integrator (LM) are necessary to maintain measure-preservation and time reversibility. Simple integrators, such as the naive Leapfrog method, are not suitable for RHMC

as they are no longer symplectic on general Riemannian manifolds (Cobb et al., 2019). These sophisticated integrators provide accurate discretization and efficient convergence in practice, but their theoretical analysis is challenging. In particular, there is no theoretical guarantee that the convergence rate of RHMC remains independent of the condition number after discretization, which is the main motivation for using non-Euclidean samplers in our case.

In fact, analyzing discretized non-Euclidean samplers has been a persistent challenge in many recent works. Another commonly studied class of non-Euclidean samplers is the Riemannian Langevin algorithm (RLA) (Girolami and Calderhead, 2011), which extends the Langevin algorithm to non-Euclidean space. A closely related process is the Mirror Langevin diffusion (MLD) (Zhang et al., 2020), which is a special case of RLA when the metric is given by the Hessian of a Legendre-type convex potential ϕ. Many recent works have focused on obtaining the convergence rate of discretized MLD or RLA, but many of them require strong assumptions or oracles for accurate discretization. The analysis of Zhang et al. (2020); Jiang (2021); Li et al. (2022) and the empirical results in Jiang (2021) suggest that unless a strong regularity assumption between the target distribution and ϕ is satisfied, the naive integrators can lead to a bias term that exists even when the step size tends to zero. This bias arises from the third-order error terms resulting from non-Euclidean geometries and is hard to control. Ahn and Chewi (2021) circumvented this issue by proposing an alternative discretization method that uses the exact solution to the Brownian motion term, but it remains unclear whether the discretization is feasible for general ϕ. Similarly, Gatmiry and Vempala (2022) analyzed the convergence rate of RLA using an oracle to sample from the natural Brownian motion on the manifold. Given the current limitations in our understanding of the integrators for non-Euclidean samplers, we believe it is crucial to investigate the integrators more thoroughly and explore alternative integrators.

Contribution We provide (to our knowledge) the first convergence rate of discretized RHMC on a class of numerical integrators. We consider a general class of constrained distributions that can be written as

$$e^{-f(x)} \text{ subject to } x \in \mathcal{M}, \tag{1.1}$$

where we assume f is a convex function and $\mathcal{M} \subset \mathbb{R}^n$ is a convex body with a (highly) self-concordant barrier. We give theoretical guarantees showing that a large class of integrators can maintain smoothness and condition number independence when sampling from distributions in the form of $e^{-\alpha^\top x}$ on a polytope with m constraints. In fact, many applications can be written in this form because any log-concave density in the form of (1.1) can be reduced to

$$e^{-t} \text{ subject to } (x, t) \in \mathcal{M}', \tag{1.2}$$

where $\mathcal{M}' = \{(x, t) : f(x) \leq t, x \in \mathcal{M}\}$ is convex in (x, t). We show for distributions in the form of $e^{-\alpha^\top x}$, the implicit midpoint method (IMM) and the generalized Leapfrog method (LM) have a mixing time of $\widetilde{O}\left(mn^3\right)$ to achieve ϵ total variation distance to the target distribution. In addition, we give a general convergence result on sampling from distributions in the form of $e^{-f(x)}$ on a convex body in terms of parameters of the manifold and the integrator, which can be useful for future works that analyze the convergence rate on other integrators or distributions.

While numerical integration is a rich and active field (Hairer et al., 2006), and the study of the local convergence of numerical estimators is quite sophisticated, we are not aware of global polynomial-time mixing time guarantees based on commonly-used numerical integrators such as IMM and LM. Our convergence result is the theoretical foundation of Kook et al. (2022) and extends

Lee and Vempala (2018) to settings of practical importance. Our results apply to not only IMM and LM, but also a more general class of symplectic and time-reversible integrators that satisfies a sensitivity condition, which advances our understanding of integrators for RHMC and the more general non-Euclidean samplers.

Moreover, in our algorithm, we use a Metropolis filter to correct the distribution, which is a crucial step for high-accuracy sampling. To address the discretization issues of RLA and MLD, applying a Metropolis filter to correct the bias is one potential solution. Nevertheless, to the best of our knowledge, there is no analysis of general-purpose metropolized non-Euclidean Langevin algorithm in the literature. We believe that our analysis of metropolized RHMC can provide valuable insights into the design and analysis of future metropolized non-Euclidean Langevin algorithms.

It is important for readers to be aware that although the convergence rate we obtain is independent of the condition number, the convergence rate is likely to be far from optimal due to the complicated analysis of the integrators used. To couple the discretized and ideal RHMC in our analysis, we need a step size much smaller than what is typically required in practice. Kook et al. (2022) demonstrated that RHMC with IMM can achieve sublinear mixing times in dimension on metabolic networks and structured polytopes including hypercubes, simplices, and Birkhoff polytopes. We believe a tighter convergence bound is possible with more advanced analysis.

1.1. Prior work

The convergence rate of MCMC methods in sampling from a convex body has been a topic of active research for decades (see Lee and Vempala (2022) for a more detailed discussion). The mixing time of ball walk on isotropic log-concave density is bounded by $\widetilde{O}(n^2)$ from a warm start (Kannan et al., 1997), where a convex body can be put into a near isotropic position in $\widetilde{O}(n^3)$ membership queries (Jia et al., 2021). Dikin walk uses the local geometry to improve the mixing rate to $O(mn)$ on polytopes, where m is the number of constraints. Moreover, due to its affine invariance, there is no need to put the polytope into an isotropic position. With an LS barrier (Lee and Sidford, 2014), Dikin walk can achieve a mixing rate of $\widetilde{O}(n^2)$ for any polytope (Laddha et al., 2020). Geodesic walk utilizes non-Euclidean geometry by taking a random walk on a manifold. Geodesic walk with an exact exponential map and a Metropolis filter can converge to the uniform density in $O(mn^{3/4})$ steps (Lee and Vempala, 2017). Continuous RHMC avoids the use of a Metropolis filter due to its measure preservation and time reversibility, which further improves the mixing time to $O(mn^{2/3})$ (Lee and Vempala, 2018) on uniform density. Our paper extends the mixing time result to discretized RHMC with feasible integrators on more general distributions. Note that even an extension to distribution $e^{-\alpha^\top x}$ needs nontrivial work to avoid dependence on quantities such as the domain diameter.

2. RHMC with numerical integrators

2.1. Basics of RHMC

Hamiltonian Monte Carlo (HMC) is one of the most widely used MCMC methods and is the default sampler implementation in many sampling packages (Stan Development Team (2020); Salvatier et al. (2016); Bingham et al. (2019); Kook et al. (2022)). HMC introduces an auxiliary velocity variable v in addition to the position x, defines a joint density on (x, v), and determines its trajectory according to the *Hamiltonian dynamics*. The Hamiltonian dynamics is characterized by the *Hamiltonian equations*, the first-order differential equations of the *Hamiltonian* H with respect to x and v. The

Hamiltonian has a natural interpretation as the total energy of a particle consisting of the kinetic and potential energy at position x with velocity v.

The dynamic can be naturally generalized to the setting of Riemannian manifold with local metric $\{g(x)\}_{x\in\mathcal{M}}$. A natural extension of the Hamiltonian is given by

$$H(x,v) = f(x) + \frac{1}{2}v^\top g(x)^{-1}v + \frac{1}{2}\log\det g(x),$$

with $g(x)$ viewed as a positive-definite matrix. For later use, we split H into two parts $H_1(x,v) = f(x) + \frac{1}{2}\log\det g(x)$ and $H_2(x,v) = \frac{1}{2}v^\top g(x)^{-1}v$. A curve $(x(t),v(t)) \in \mathcal{M}\times T_x\mathcal{M} \subset \mathbb{R}^n\times\mathbb{R}^n$ is called the *Hamiltonian curve* if it is the solution to the Hamiltonian equations:

$$\begin{aligned}
\frac{dx}{dt} &= \frac{\partial H}{\partial v}(x,v) = g(x)^{-1}v,\\
\frac{dv}{dt} &= -\frac{\partial H}{\partial x}(x,v) = -\left(\underbrace{\nabla f(x) + \frac{1}{2}\mathrm{Tr}\left[g(x)^{-1}Dg(x)\right]}_{\frac{\partial H_1}{\partial x}} + \underbrace{\left(-\frac{1}{2}Dg(x)\left[\frac{dx}{dt},\frac{dx}{dt}\right]\right)}_{\frac{\partial H_2}{\partial x}}\right).
\end{aligned} \tag{2.1}$$

When clear from context, the Hamiltonian curve refers to $x(t)\in\mathcal{M}$ only. The Hamiltonian curves $(x(t),v(t))$ have several geometric properties. For a map $F_t : (x,v)\mapsto(x(t),v(t))$,

1. Hamiltonian preservation: $\frac{d}{dt}H(x(t),v(t)) = 0$.
2. Symplectic: $DF_t(x,v)^\top\cdot J\cdot DF_t(x,v) = J$ for any $t\geq 0$ and $J = \begin{bmatrix} 0 & I_n \\ -I_n & 0\end{bmatrix}$.
3. Measure-preservation: $\det(DF_t(x,v)) = 1$ for any $t\geq 0$. Note that measure-preservation immediately follows from symplecticity.
4. Time-reversible: $F_t(x(t),-v(t)) = (x,-v)$.

Just as the Hamiltonian dynamics can be extended to the Riemannian setting, Girolami and Calderhead (2011) extended HMC to a Riemannian version called Riemannian Hamiltonian Monte Carlo (RHMC); see Algorithm 1 for its one-step description. In fact, HMC can be recovered from RHMC using the Euclidean metric (i.e., $g(x) = I$).

Our goal is to sample from a probability density proportional to $e^{-f(x)}$ supported on a convex body. To this end, we use RHMC with the Hamiltonian $H : \mathcal{M}\times\mathbb{R}^n\subset\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$, viewing the convex body as a Riemannian manifold $\mathcal{M}$ with a local metric g.

2.2. Notation and setting

We use $(\mathcal{M},g)$ to denote a connected and compact Riemannian manifold with a boundary and a metric g on which a target distribution is supported. For a function $f:\mathcal{M}\subset\mathbb{R}^n\to\mathbb{R}$, we denote a target distribution by $\pi(x)$ whose density is proportional to $e^{-f(x)}$ (i.e., $\frac{d\pi}{dx}\sim e^{-f(x)}$). We use $T_x\mathcal{M}$ to denote the tangent space of $\mathcal{M}$ at $x\in\mathcal{M}$. We denote by π_x the projection map onto x-space (i.e., $\pi_x(x,v)\stackrel{\text{def}}{=}x$) and by i_x the inclusion map (i.e., $i_x(v)\stackrel{\text{def}}{=}(x,v)$). We reserve h for the step size of RHMC.

Algorithm 1: RIEMANNIAN HAMILTONIAN MONTE CARLO

Input: Initial point x, step size h

`// Step 1: Sample an initial velocity` v

Sample $v \sim \mathcal{N}(0, g(x))$.

`// Step 2: Solve the Hamiltonian equations`

Solve the Hamiltonian equations (2.1) to obtain $(x(t), v(t))$.

`// Step 3: Metropolis-filter (skipped for ideal RHMC)`

Accept $x(h)$ with probability $\min\left(1, \frac{e^{-H(x(h),v(h))}}{e^{-H(x,v)}}\right)$. Otherwise, stay at x.

With both manifold $\mathcal{M}$ and tangent space $T_x\mathcal{M}$ endowed with the Euclidean metric, we define a map $F_t : \mathcal{M} \times T_x\mathcal{M} \to \mathcal{M} \times \bigcup_{z\in\mathcal{M}} T_z\mathcal{M}$ by $F_t(x, v) \stackrel{\text{def}}{=} (x(t), v(t))$, where $(x(t), v(t))$ is the solution to the Hamiltonian equations at time t with an initial condition (x, v). In particular, we define $T_{x,h} : T_x\mathcal{M} \to \mathcal{M}$ by $T_{x,h}(v) \stackrel{\text{def}}{=} (\pi_x \circ F_h \circ i_x)(v) = x(h)$. When both $\mathcal{M}$ and $T_x\mathcal{M}$ are endowed with the local metric g, we instead use $\mathrm{Ham}_{x,t} : T_x\mathcal{M} \to \mathcal{M}$ defined by $\mathrm{Ham}_{x,t}(v) \stackrel{\text{def}}{=} x(t)$.

When a numerical integrator with step size h outputs $(\bar{x}_h, \bar{v}_h)$ by solving the Hamiltonian equations with an initial condition (x, v), we denote $\overline{F}_h(x, v) \stackrel{\text{def}}{=} (\bar{x}_h, \bar{v}_h)$ for a function $\overline{F}_h : \mathcal{M} \times T_x\mathcal{M} \to \mathcal{M} \times \bigcup_{z\in\mathcal{M}} T_z\mathcal{M}$, where the domain and range are endowed with the Euclidean metric. We define $\overline{T}_{x,h} : T_x\mathcal{M} \to \mathcal{M}$ (endowed with the Euclidean metric) by $\overline{T}_{x,h}(v) = (\pi_x \circ \overline{F}_h \circ i_x)(v) = \bar{x}_h$. We drop h from $T_{x,h}, \overline{F}_h$ and $\overline{T}_{x,h}$ if the step size is clear from context.

We assume that the domain $\mathcal{M} \subset \mathbb{R}^n$ with a boundary is convex and has a (highly) self-concordant barrier $\phi : \mathcal{M} \subset \mathbb{R}^n \to \mathbb{R}$ (Definition 54), and that the metric g is induced by the Hessian of the barrier (i.e., $g(x) = \nabla^2\phi(x)$). We denote the local norm of a vector v by $\|v\|_x$ or $\|v\|_{g(x)}$, and the Riemannian distance by d_ϕ (Definition 56). We use $a \lesssim b$ to indicate that $a \leq cb$ for some universal constant $c > 0$.

2.3. Discretized RHMC

We use *ideal* RHMC to denote the algorithm when the Hamiltonian equations in Step 2 is accurately solved without any error. However, we cannot expect such an accurate ODE solver to always exist in reality, so numerical integrators with solutions that approximate the accurate ODE solutions are necessary. We use *discretized* RHMC to denote the algorithm when Step 2 of RHMC is solved by a numerical integrator and a Metropolis-filter is used to correct the distribution.

We now define a condition of numerical integrators that plays an important role in our convergence-rate analysis.

Definition 1 *For a numerical integrator $\overline{F}$ and $(x, v) \in \mathcal{M} \times T_x\mathcal{M} \subset \mathbb{R}^n \times \mathbb{R}^n$, we call $\overline{F}$* sensitive *at (x, v) if there exists step size $h_0(x, v)$ such that the numerical integrator with step size h less than h_0 satisfies*

$$\frac{\left|D\overline{T}_{x,h}(v)\right|}{|DT_{x,h}(v')|} \geq 0.998,$$

where v' satisfies $T_{x,h}(v') = \overline{T}_{x,h}(v)$ and the Jacobian DT is taken with respect to the velocity variable. In other words, the solution of the numerical integrator changes almost as fast as the ideal solution does. Unless specified otherwise, a sensitive integrator is additionally assumed to be measure-preserving (i.e., $\det(D\overline{F}_h(x,v)) = 1$) and time-reversible (i.e., $\overline{F}_h(\overline{x}_h, -\overline{v}_h) = (x, -v)$).

As a time-reversible numerical integrator is even-order, second-orderness automatically follows. That is, for sufficiently small step size $h > 0$, $d_g(\bar{x}_h, x_h) \le C_x(x,v)h^2$ and $\|\bar{v}_h - v_h\|_{g(x)^{-1}} \le C_v(x,v)h^2$ for some functions of x and v, C_x and C_v. In other words, the errors of the numerical integrator $\overline{F}_h$ with respect to the exact ODE solver F_h grow at most quadratically in the step size h.

This family of numerical integrators turns out to cover many commonly used integrators in practice. For example, the implicit midpoint integrator (IMM) (Algorithm 2) and the generalized Leapfrog integrator (LM) (Algorithm 3) satisfy symplecticity, time-reversibility, and sensitivity (as shown in Section C). Measure-preservation gives the simple formula of the acceptance probability in Step 3 of Algorithm 1. Measure-preservation together with time-reversibility plays an important role in showing that the discretized RHMC converges to its stationary distribution with density proportional to $e^{-f(x)}$ (see Theorem 8 in Kook et al. (2022)).

3. Our results

We analyze the mixing time of RHMC discretized by numerical integrators commonly used in practice, with the Hamiltonian set to be $H(x,v) = f(x) + \frac{1}{2}v^\top g(x)^{-1}v + \frac{1}{2}\log\det g(x)$. Previous analysis of RHMC was based on high accuracy numerical integrators, which are not always achievable in practice (Lee and Vempala, 2018), and the complexity bounds were derived for uniform density on a polytope. We extend the setting to sampling exponential densities with practically feasible integrators. In the next theorem, we denote $\mathcal{M}_\rho := \left\{x \in \mathcal{M} : \|\alpha\|^2_{g(x)^{-1}} \le 10n^2\log^2\frac{1}{\rho}\right\}$ for $\rho > 0$.

Theorem 2 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S\subset\mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T steps of RHMC discretized by a sensitive integrator on $\mathcal{M}$. For any $\varepsilon > 0$, if for $x \in \mathcal{M}_{\frac{\varepsilon}{2\Lambda}}$ and $v \in \mathbb{R}^n$ randomly drawn from $\mathcal{N}(0, g(x))$, we have that with probability at least 0.99, step size $h \le h_0(x,v)$,*

$$h \le \frac{10^{-20}}{n^{7/12}\log^{1/2}\frac{\Lambda}{\varepsilon}},\ hC_x(x,v) \le \frac{10^{-20}}{\sqrt{n}},\ h^2C_x(x,v) \le \frac{10^{-10}}{n\log\frac{\Lambda}{\varepsilon}} \text{ and } h^2C_v(x,v) \le \frac{10^{-10}}{\sqrt{n\log\frac{\Lambda}{\varepsilon}}},$$

then $d_{TV}(\pi_T, \pi) \le \varepsilon$ for $T = O\left(mh^{-2}\log\frac{\Lambda}{\varepsilon}\right)$.

By setting $C_x = C_v = 0$, we can obtain the following corollary for the mixing time of the ideal RHMC in this setting.

Corollary 3 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S\subset\mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of the ideal RHMC on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{7/12}\log^{1/2}\frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^{7/6}\log^2\frac{\Lambda}{\varepsilon}\right)$ such that $d_{TV}(\pi_T, \pi) \le \varepsilon$.*

After we compute the parameters C_x and C_v of IMM and LM (see Section C and D), and identify the sufficient conditions on the step size for their sensitivity, the following mixing times of RHMC discretized by IMM or LM immediately follow.

Corollary 4 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of RHMC discretized by IMM on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{3/2} \log \frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ such that $d_{TV}(\pi_T, \pi) \leq \varepsilon$.*

Corollary 5 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of RHMC discretized by LM on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{3/2} \log \frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ such that $d_{TV}(\pi_T, \pi) \leq \varepsilon$.*

In fact, Theorem 2 comes from a general result on the mixing time of RHMC for density e^{-f} on a convex body $\mathcal{M} \subset \mathbb{R}^n$. We provide its informal version here and defer its full statement (Theorem 24) to Section B.

Theorem *(Informal) Let π be a target distribution on a convex set $\mathcal{M} \subset \mathbb{R}^n$ and $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let $\mathcal{M}$ be the Hessian manifold with its metric induced by the Hessian of a highly self-concordant barrier and π_T the distribution obtained after T steps of RHMC discretized by a sensitive integrator on $\mathcal{M}$. For any $\varepsilon > 0$, let $\mathcal{M}_{\frac{\varepsilon}{2\Lambda}} \subset \mathcal{M}$ be a convex subset of measure at least $1 - \frac{\varepsilon}{2\Lambda}$. There is a step size bound h_0, defined in terms of smoothness parameters of the manifold and the integrator, so that for any step size $h \leq h_0$, there exists $T = O\left(\left(h\psi_{\mathcal{M}_{\frac{\varepsilon}{2\Lambda}}}\right)^{-2} \log \frac{\Lambda}{\varepsilon}\right)$ where $\psi_{\mathcal{M}_{\frac{\varepsilon}{2\Lambda}}}$ is the isoperimetry of $\mathcal{M}_{\frac{\varepsilon}{2\Lambda}}$, such that $d_{TV}(\pi_T, \pi) \leq \varepsilon$.*

3.1. Discussion

Suboptimal mixing time In our analysis of RHMC with numerical integrators, the main technical bottleneck that limits the step size to $O(n^{-3/2})$ is the coupling argument between the ideal RHMC and the discretized RHMC. To illustrate, suppose that the ideal RHMC maps (x, v_1) to (x', v') and discretized RHMC maps (x, v_2) to (x', v''). In the coupling of the two processes, we need to ensure that the ratio of the density functions of Gaussian estimated at v_1 and v_2 is close to 1, which comes down to $\left|\|v_1\|_{g^{-1}}^2 - \|v_2\|_{g^{-1}}^2\right| = O(1)$. To prove this, we use the triangle inequality as follows: $\left|\|v_1\|_{g^{-1}}^2 - \|v_2\|_{g^{-1}}^2\right| \leq \|v_1 - v_2\|_{g^{-1}} (\|v_1\|_{g^{-1}} + \|v_2\|_{g^{-1}})$. As $(\|v_1\|_{g^{-1}} + \|v_2\|_{g^{-1}}) = \Theta(\sqrt{n})$ w.h.p., we should take step size small enough so that $\|v_1 - v_2\|_{g^{-1}} = O(1/\sqrt{n})$. Thus, controlling $\left|\|v_1\|_{g^{-1}}^2 - \|v_2\|_{g^{-1}}^2\right|$ without the triangle inequality is potentially a way to improve the dependency on n. An alternative approach to improve the mixing time is to directly couple two discretized RHMC processes. However, this approach is technically challenging to carry out, so we use a detour through the ideal RHMC by first coupling the ideal and the discretized RHMC and then relying on the coupling result between two ideal processes.

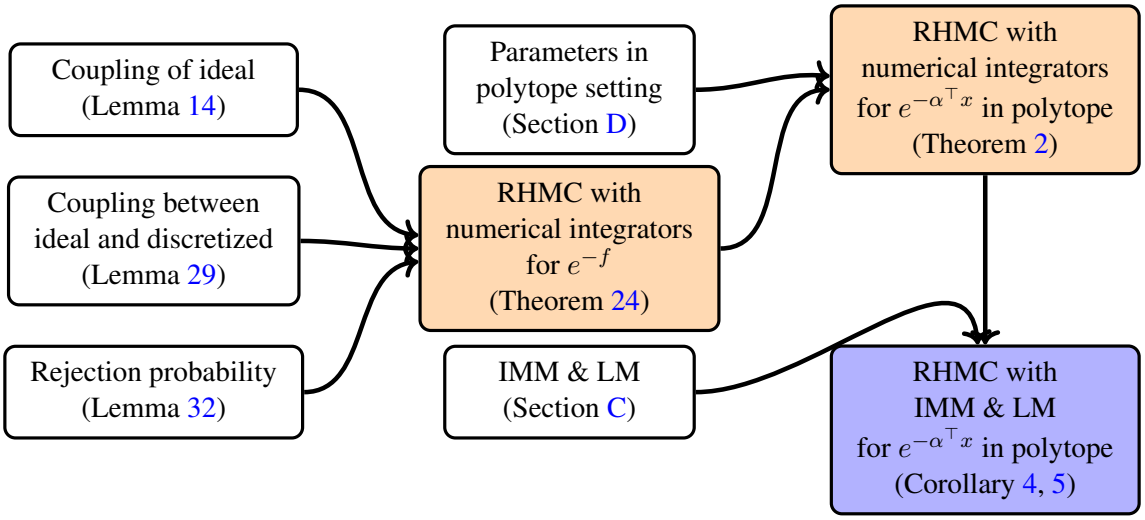

Figure 4.1: Proof outline

Initialization (warmness parameter) There are several ways to initialize efficiently and avoid a warm start penalty. For instance, one can run an $O^*(n^3)$ algorithm by Jia et al. (2021) that uses Gaussian cooling (see Section 6.2 in Cousins and Vempala (2018)) to generate an $O(1)$-warm start for the uniform distribution. In detail, the cooling algorithm involves a sequence of Gaussians truncated on the convex body, with variances increasing from small to large. The initial Gaussian is almost contained in the body, and the last one is almost close to the uniform distribution over the body. In each phase of the cooling algorithm, a sampler such as the Ball walk or Hit-and-Run is run. This sample serves as a warm start to the next phase (i.e., Gaussian with larger variance).

In fact, Gaussian Cooling can be integrated into RHMC itself (see Section 6 in Lee and Vempala (2018)), removing the dependency on $\log M$. Specifically, we consider a sequence of target distributions $\propto e^{-(f+\alpha_i\phi)}$ with $\alpha_1 = 1$ and the barrier ϕ. In the first phase, with the minimizer x^* of $f + \phi$, the Gaussian with mean x^* and covariance $C \cdot (\nabla^2\phi(x^*))^{-1}$ for some constant $C > 0$ provides a good warm start for RHMC to sample from the density $\propto e^{-(f+\phi)}$. Then in each subsequent phase, RHMC is started at the sample from the previous phase and runs with target $e^{-(f+\alpha_i\phi)}$ while decreasing α_i toward 0. Since the isoperimetry of $e^{-(f+\alpha_i\phi)}$ is $\gtrsim \sqrt{\alpha}$, the mixing times of RHMC in earlier phases are smaller than those in later phases and the total running time of RHMC with cooling is dominated by RHMC in the last phase, which is the complexity of RHMC with $M = O(1)$.

4. Technical overview

In this section, we provide a summary of the key proof ingredients that gives the convergence rate of RHMC with numerical integrators to samples from density $e^{-f(x)}$ on a convex body; see Figure 4.1 for the roadmap. In Section 4.1, we review a general technique using s-conductance for bounding the mixing time of a Markov chain. In Section 4.2, we summarize a refined analysis of the ideal RHMC (Section A) and the technique to couple the ideal and discretized RHMC (Section B). Finally in Section 4.3, we describe the high-level ideas of our analysis of the numerical integrators (Section C),

IMM and LM, and how to get the results for sampling from $e^{-\alpha^\top x}$ on the Hessian manifolds of polytopes (Section D).

4.1. Mixing time via s-conductance: isoperimetry and one-step coupling

Consider a Markov chain with a state space $\mathcal{M}$, a transition distribution $\mathcal{T}_x$ and stationary distribution π. We consider a *lazy* Markov chain to avoid a uniqueness issue of the stationary distribution. At each step, the lazy version of the Markov chain does nothing with probability $\frac{1}{2}$ (*i.e.*, stays at where it is). Note that this change for the purpose of proof worsens the mixing time only by a factor of 2.

We use a standard conductance-based argument in Vempala (2005) to bound the mixing time, which consists of two main ingredients – *the isoperimetry* and *the total variation (TV) distance coupling of one-step distributions* (Definition 57) starting from two close points.

Definition 6 (s-conductance) *Consider a Markov chain with a state space $\mathcal{M}$, a transition distribution $\mathcal{T}_x$ and stationary distribution π. For any $s \in [0, 1/2)$, the* s-conductance *of the Markov chain is*

$$\Phi_s \stackrel{\text{def}}{=} \inf_{\pi(S)\in(s,1-s)} \frac{\int_S \mathcal{T}_x(S^c)\pi(x)dx}{\min(\pi(S)-s, \pi(S^c)-s)}.$$

As shown by Lovász and Simonovits (1993), a lower bound on the s-conductance of a Markov chain leads to an upper bound on the mixing time of the Markov chain.

Lemma 7 (Lovász and Simonovits (1993)) *Let π_t be the distribution obtained after t steps of a lazy reversible Markov chain with the stationary distribution π. It follows that*

$$d_{TV}(\pi_t, \pi) \le H_s + \frac{H_s}{s}\left(1 - \frac{\Phi_s^2}{2}\right)^t,$$

where $H_s = \sup\{|\pi_0(A) - \pi(A)| : A \subset \mathcal{M},\ \pi(A) \le s\}$ with $0 < s \le \frac{1}{2}$.

We now define the isoperimetry of a subset of $\mathcal{M}$.

Definition 8 (Isoperimetry) *Let $(\mathcal{M}, g)$ be a Riemannian manifold and $\mathcal{M}'$ a measurable subset of $\mathcal{M}$ with $\pi(\mathcal{M}') > \frac{1}{2}$. The* isoperimetry ψ *of the subset with stationary distribution π is defined by*

$$\psi_{\mathcal{M}'} = \inf_{S\subset\mathcal{M}'} \frac{\lim_{\delta\to 0^+} \frac{1}{\delta}\int_{\{x\in\mathcal{M}' : 0<d_g(S,x)\le\delta\}} \pi(x)dx}{\min(\pi(S), \pi(\mathcal{M}'\backslash S))}.$$

The following illustrates how one-step coupling with the isoperimetry leads to a lower bound on the s-conductance. It can be proved similarly as Lemma 13 in Lee and Vempala (2018).

Proposition 9 *For a Riemannian manifold $(\mathcal{M}, g)$, let π be the stationary distribution of a reversible Markov chain on $\mathcal{M}$ with a transition distribution $\mathcal{T}_x$. Let $\mathcal{M}' \subset \mathcal{M}$ be a subset with $\pi(\mathcal{M}') \ge 1 - \rho$ for some $\rho < \frac{1}{2}$. We assume the following one-step coupling: if $d_g(x, x') \le \Delta \le 1$ for $x, x' \in \mathcal{M}'$, then $d_{TV}(\mathcal{T}_x, \mathcal{T}_{x'}) \le 0.9$. Then for any $\rho \le s < \frac{1}{2}$, the s-conductance is bounded below by*

$$\Phi_s \ge \Omega\left(\psi_{\mathcal{M}'}\Delta\right).$$

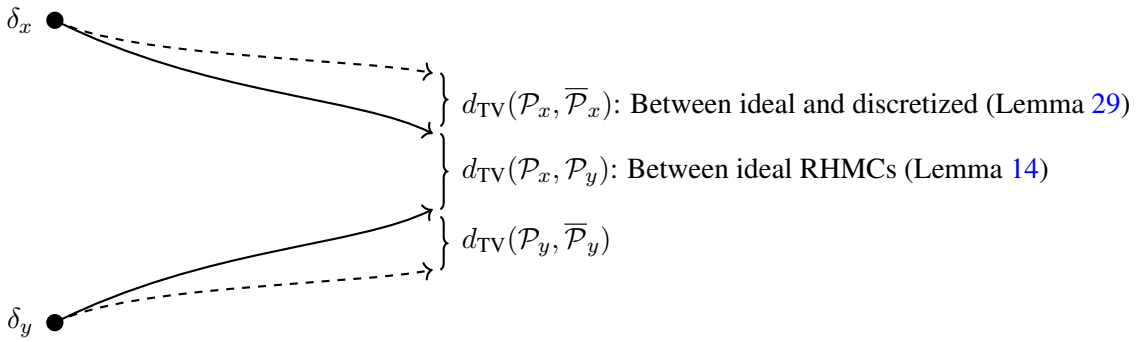

Figure 4.2: An illustration of our approach to one-step coupling. The thick line indicates the ideal RHMC, and the dashed line indicates the discretized RHMC.

Proof Let $S \subseteq \mathcal{M}'$ be a mesurable subset with $\pi(S) \in (s, 1-s)$. Consider a partition $\{S_1, S_2, S_3\}$ of $\mathcal{M}'$ defined by

$$\begin{aligned} S_1 &= \left\{x \in S \cap \mathcal{M}' : \mathcal{T}_x(S^c) < 0.05\right\}, \\ S_2 &= \left\{x \in S^c \cap \mathcal{M}' : \mathcal{T}_x(S) < 0.05\right\}, \\ S_3 &= \mathcal{M}' \backslash (S_1 \cup S_2). \end{aligned}$$

Since $\pi(\mathcal{M}') \geq 1 - \rho \geq 1 - s$, it follows that $\pi(S \cap \mathcal{M}') \geq \pi(S) - s$ and $\pi(S^c \cap \mathcal{M}') \geq \pi(S^c) - s$, so it suffices to show $\int_S \mathcal{T}_x(S^c) d\pi \gtrsim \psi_{\mathcal{M}'} \Delta \min\{\pi(S \cap \mathcal{M}'), \pi(S^c \cap \mathcal{M}')\}$. The stationarity of a reversible Markov chain implies $\int_S \mathcal{T}_x(S^c) d\pi = \frac{1}{2}\left(\int_S \mathcal{T}_x(S^c) d\pi + \int_{S^c} \mathcal{T}_x(S) d\pi\right)$.

We may assume $\pi(S_1) \geq \frac{1}{2}\pi(S \cap \mathcal{M}')$ and $\pi(S_2) \geq \frac{1}{2}\pi(S^c \cap \mathcal{M}')$; otherwise, $\int_S \mathcal{T}_x(S') d\pi(x) \gtrsim \pi(S \cap \mathcal{M}')$ or $\pi(S^c \cap \mathcal{M}')$ and thus $\Phi_S = \Omega(1)$. Note that

$$\begin{aligned} \int_S \mathcal{T}_x(S^c) d\pi &\geq \frac{1}{2}\left(\int_{S \cap \mathcal{M}' \backslash S_1} \mathcal{T}_x(S^c) d\pi + \int_{S^c \cap \mathcal{M}' \backslash S_2} \mathcal{T}_x(S) d\pi\right) \\ &\gtrsim \pi(S_3) \geq \psi_{\mathcal{M}'} \Delta \min\{\pi(S_1), \pi(S_2)\}, \end{aligned}$$

where the last follows from the definition of the isoperimetry. By assumptions on $\pi(S_1)$ and $\pi(S_2)$, we conclude $\int_S \mathcal{T}_x(S^c) d\pi \gtrsim \psi_{\mathcal{M}'} \Delta \min\{\pi(S \cap \mathcal{M}'), \pi(S^c \cap \mathcal{M}')\}$. ■

4.2. One-step coupling of discretized RHMC

In light of Proposition 9, we can focus on coupling the one-step distributions of the discretized RHMC starting from two close-by points. Let $\mathcal{P}_x$ and $\overline{\mathcal{P}}_x$ be the one-step distributions on $\mathcal{M}$ of the ideal and the discretized RHMC starting from x, respectively. We use $\overline{\mathcal{P}}'_x$ to denote the discretized RHMC without the Metropolis filter. As illustrated in Figure 4.2, for two close points x and y, the triangle inequality leads to

$$\begin{aligned} d_{\mathrm{TV}}(\overline{\mathcal{P}}_x, \overline{\mathcal{P}}_y) &\leq d_{\mathrm{TV}}(\overline{\mathcal{P}}_x, \mathcal{P}_x) + d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) + d_{\mathrm{TV}}(\mathcal{P}_y, \overline{\mathcal{P}}_y) \\ &\leq \left(d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \mathcal{P}_x) + d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) + d_{\mathrm{TV}}(\mathcal{P}_y, \overline{\mathcal{P}}'_y)\right) + \left(d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x) + d_{\mathrm{TV}}(\overline{\mathcal{P}}'_y, \overline{\mathcal{P}}_y)\right). \end{aligned}$$

Hence, it suffices to bound $d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y)$, $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \mathcal{P}_x)$ and $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x)$, respectively. We bound in Section A the first term $d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y)$, the TV distance of one-step distributions of the ideal RHMC. For the remaining terms, when numerical integrators do not preserve the Hamiltonian, a Metropolis filter is necessary to ensure that the discretized RHMC converges to a target distribution. Due to the filter, we need to handle a point-mass distribution at x. We address this by first bounding the second term $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \mathcal{P}_x)$, the TV distance between the ideal and discretized RHMC *without* the Metropolis filter in Section B.2. We then separately bound the rejection probability $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x)$ in Section B.3.

4.2.1. COUPLING OF IDEAL RHMC

We summarize how to bound $d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y)$ here (see Section A for the full version).

Lemma *(Informal, Lemma 14) For most of x and y, and step size h small enough, if $d_\phi(x, y) \le \frac{1}{100}$, then $d_{TV}(\mathcal{P}_x, \mathcal{P}_y) \le O\left(\frac{1}{h}\right) d_\phi(x, y) + \frac{1}{25}$.*

Previous approach Lee and Vempala (2018) provided a general framework for computing the mixing rate of RHMC on a manifold embedded in $\mathbb{R}^n$, in terms of the isoperimetry and smoothness parameters depending on the manifold and step size. One of the major proof ingredients is one-step coupling: for two close points x and y, the one-step distributions at x and y have large overlap.

They use the notion of a 'regular' Hamiltonian curve, which enables them to handle this task in low level, where the regularity can be understood as *average* behavior of Hamiltonian curves with high probability and is quantified by some auxiliary functions. As the starting point of a regular Hamiltonian curve changes from x to y along a length-minimizing geodesic $c(s)$ joining x and y, they find a one-to-one correspondence between regular Hamiltonian curves started at x and y, and bound $\left|\frac{d}{ds} d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_{c(s)})\right|$ over s. They achieve this by quantifying the rate of changes of the probability density (see (A.1)).

It is daunting to directly work with the exact density function, so they make use of the following techniques: (1) Show that the determinant of Jacobian is close to h^n up to small step size by applying a matrix-ODE theory to the second-order ODE of the Hamiltonian equation (see Lemma 59). It allows them to work with an approximate but simpler density with the Jacobian replaced by h^n (see (A.3)). (2) Establish the one-to-one correspondence along variations of Hamiltonian curves by the implicit function theorem; for a given endpoint z, as the starting point of a Hamiltonian curve moves along $c(s)$, there exists a unique initial velocity $v_{c(s)}$ at each point on $c(s)$ that brings $c(s)$ to the endpoint z in step size h (i.e., $\mathrm{Ham}_{c(s),h}(v_{c(s)}) = z$). At the same time, by using the matrix-ODE theory again they show that the regularity of Hamiltonian curves does not blow up along $c(s)$ and quantify how much the proper initial velocity changes.

Refined analysis Lee and Vempala (2018) bounded $\left|\frac{d}{ds} d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_{c(s)})\right|$ in terms of smoothness parameters, supremum bounds on some quantities defined over the regular Hamiltonian curves starting from *any* point in $\mathcal{M}$. However, considering all starting points leads to a weaker coupling in the end. In fact, this makes sampling from an exponential density have dependence on the condition number, since one of the smoothness parameters requires the supremum bound on $\|\alpha\|_{g(x)^{-1}}$ over $x \in \mathcal{M}$, which can be as large as $\|\alpha\|_2$ times the diameter of the convex body.

To achieve a condition-number independent mixing time, we work in a convex subset $\mathcal{M}_\rho$ (call a good region) instead of $\mathcal{M}$, which requires refinement of the framework by generalizing the smoothness parameters (Section A.1) and theorems in their paper accordingly. It allows us to obtain

a stronger coupling by only considering Hamiltonian curves starting from $\mathcal{M}_\rho$. This region is the region $\mathcal{M}'$ in Proposition 9.

This simple change, however, yields technical difficulties in following how Lee and Vempala (2018) proceeds with the original parameters. Recall in the one-step coupling, they consider a Hamiltonian variation along a geodesic joining two points, but the geodesic might step out of the good region. To address this issue, we use the straight line between the points instead of the geodesic, as the straight line is contained in $\mathcal{M}_\rho$ due to the convexity. We elaborate on how the technical details of the previous approach can be modified accordingly under the redefined parameters and new variation curve in order to get valid one-step coupling on this smaller region in Section A.2.

4.2.2. COUPLING BETWEEN IDEAL AND DISCRETIZED RHMC & REJECTION PROBABILITY

We provide a summary of Section B, where we prove the following lemma and Theorem 24.

Lemma *(Informal, Lemma 29 and Lemma 32) For most of (x, v), if step size h is small enough and falls under a sensitivity regime at (x, v) of a numerical integrator, then $d_{TV}(\overline{\mathcal{P}}'_x, \mathcal{P}_x) < \frac{1}{10}$ and $d_{TV}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x) < \frac{1}{10^3}$.*

For the former (bound on $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \mathcal{P}_x)$), we show that the densities of the ideal and discretized RHMC are similar by relating two velocities v and v^*, where $\overline{T}_x(v) = T_x(v^*)$. It can be reduced to establishing a constant lower bound on $\frac{p^*_x(v^*)}{p^*_x(v)} \frac{|D\overline{T}_x(v)|}{|DT_x(v^*)|}$ for the probability density p^*_x of Gaussian $\mathcal{N}(0, g(x))$.

We first define numerical integrators' analogues of the smoothness parameters. Then, we elaborate the idea above in Section B.2, where we study the dynamics of the ideal and discretized RHMC. In particular, we show the existence of v^* for a given v by the Banach fixed-point theorem and a one-to-one correspondence between them, together with an upper bound on $\|v - v^*\|_{g^{-1}}$. This upper bound allows us to bound the ratio of $p^*_x(v^*)/p^*_x(v)$. We note that these results heavily rely on the stability of local norm (see Section B.1), which follows from that the local metric is given by the Hessian of self-concordant barriers. The ratio of the Jacobian follows from the sensitivity of the integrators.

For the latter (bound on $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x)$), we observe that the acceptance probability comes down to bounding the difference of the Hamiltonian at ideal and numerical solutions. To bound this, we heavily use the stability of local norm as well as the quantitative relationships between the ideal and discretized RHMC established above. Putting these pieces together, we can obtain the mixing-time bound of the discretized RHMC in Theorem 24.

4.3. Analysis of numerical integrators & Parameter estimation in polytopes

To apply the framework established so far, we analyze in Section C two practical numerical integrators, IMM (Section C.1) and LM (Section C.2), by estimating second-order parameters C_x and C_v and quantifying their sensitivity regimes, and then apply these estimations to sampling from distribution $e^{-\alpha^\top x}$ on a polytope in Section D.

Lemma *(Informal, adapted to polytope setting) For most of (x, v), both IMM and LM have $C_x(x, v) = \widetilde{O}(n)$ and $C_v(x, v) = \widetilde{O}(n^{3/2})$ for step size $h = \widetilde{O}(1/\sqrt{n})$. The sensitivity region is $h = \widetilde{O}(1/n)$ for IMM and $h = \widetilde{O}(1/n^{3/2})$ for LM.*

To analyze one-step process of each numerical integrator, we need to keep track all the quantities explicitly to obtain the condition-number independence. However, the implicit nature of both integrators lead to *coupled* equations for x and v, making the analysis complicated. To address this, we handle these coupled equations parallelly by moving back and forth between local norms at different points, where we use the stability of local norm due to self-concordance. We remark that our approach to analyze each integrator depends on the specific implementation of the integrator, so each integrator requires slightly different techniques in this task.

For the sensitivity, we apply implicit differentiation to the one-step equation of each integrator, obtaining a matrix equation in the form of $(I - hE)D\overline{F}_h = hC$ for matrices $E, C \in \mathbb{R}^{2n\times 2n}$. We use matrix-perturbation theory to quantify a sufficient condition on the step size h that ensures the invertibility of $(I - hE)$, obtaining an equation of the form $D\overline{F}_h = \sum_{i=0}^{\infty}(hC')^i$ for a matrix $C' \in \mathbb{R}^{2n\times 2n}$. By extracting the upper-right $n \times n$ block matrix, it follows that $D\overline{T}_h = I + E'$ for $E' = \sum_{i=1}^{\infty}(hC^*)^i$ with a matrix $C^* \in \mathbb{R}^{n\times n}$. Using the self-concordance of the local metric, we get upper bounds on matrix quantities of C^* including the trace, two-norm and Frobenius norm. With E' viewed as perturbation, we apply matrix-perturbation theory again to estimate a lower bound on $\left|D\overline{T}_h\right|$.

Lastly in Section D, we show that $\mathcal{M}_\rho = \left\{x \in \mathcal{M} : \|\alpha\|^2_{g(x)^{-1}} \le 10n^2 \log^2 \frac{1}{\rho}\right\}$ is convex by checking the second-order condition and that $\mathcal{M}_\rho$ has large measure by using a functional inequality. Then we compute all parameters discussed so far – isoperimetry, smoothness parameters of the manifold and numerical integrator – for the polytope setting, putting them together to obtain the results in Section 3.

Acknowledgments

This work was supported in part by NSF awards CCF-1909756, CCF-2007443 and CCF-2134105.

References

Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:28405–28418, 2021.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *The Journal of Machine Learning Research (JMLR)*, 20: 28:1–28:6, 2019. URL http://jmlr.org/papers/v20/18-403.html.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *The Journal of Machine Learning Research (JMLR)*, 21:92–1, 2020.

Zongchen Chen and Santosh S Vempala. Optimal Convergence Rate of Hamiltonian Monte Carlo for Strongly Logconcave Distributions. *Theory of Computing*, 18(1):1–18, 2022.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: a non-asymptotic analysis. In *Conference on Learning Theory (COLT)*, pages 300–323. PMLR, 2018.

Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory (COLT)*, pages 1260–1300. PMLR, 2021.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. In *Conference on Learning Theory (COLT)*, pages 1–2. PMLR, 2022.

Adam D Cobb, Atılım Güneş Baydin, Andrew Markham, and Stephen J Roberts. Introducing an explicit symplectic integration scheme for Riemannian manifold Hamiltonian Monte Carlo. *arXiv preprint arXiv:1910.06243*, 2019.

Ben Cousins and Santosh Vempala. A practical volume algorithm. *Mathematical Programming Computation*, 8(2):133–160, 2016.

Ben Cousins and Santosh Vempala. Gaussian Cooling and $o^*(n^3)$ Algorithms for Volume and Gaussian Volume. *SIAM Journal on Computing*, 47(3):1237–1273, 2018.

Arnak Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory (COLT)*, pages 678–689. PMLR, 2017.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *The Journal of Machine Learning Research (JMLR)*, 20(1):2666–2711, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory (COLT)*, pages 793–797. PMLR, 2018.

Khashayar Gatmiry and Santosh S Vempala. Convergence of the Riemannian Langevin Algorithm. *arXiv preprint arXiv:2204.10818*, 2022.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric numerical integration. *Oberwolfach Reports*, 3(1):805–882, 2006.

He Jia, Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Reducing isotropy and volume to KLS: an $O^*(n^3\psi^2)$ volume algorithm. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 961–974, 2021.

Qijia Jiang. Mirror Langevin Monte Carlo: the case under isoperimetry. *Advances in Neural Information Processing Systems*, 34:715–725, 2021.

Adam Tauman Kalai and Santosh Vempala. Simulated annealing for convex optimization. *Mathematics of Operations Research*, 31(2):253–266, 2006.

Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

Ravindran Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

Zachary A King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2016.

Yunbum Kook, YinTat Lee, Ruoqi Shen, and Santosh Vempala. Sampling with Riemannian Hamiltonian Monte Carlo in a Constrained Space. In *Advances in Neural Information Processing Systems*, 2022.

Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Strong self-concordance and sampling. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1212–1222, 2020.

Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $O(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 424–433. IEEE, 2014.

Yin Tat Lee and Santosh S Vempala. Geodesic walks in polytopes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on theory of Computing (STOC)*, pages 927–940, 2017.

Yin Tat Lee and Santosh S Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1115–1121, 2018.

Yin Tat Lee and Santosh S Vempala. The manifold joys of sampling. In *49th International Colloquium on Automata, Languages, and Programming (ICALP)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized Hamiltonian Monte Carlo. In *Conference on Learning Theory (COLT)*, pages 2565–2597. PMLR, 2020.

Nathan E Lewis, Harish Nagarajan, and Bernhard O Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10 (4):291–305, 2012.

Ruilin Li, Molei Tao, Santosh S Vempala, and Andre Wibisono. The mirror Langevin algorithm converges with vanishing bias. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 718–742. PMLR, 2022.

László Lovász. Hit-and-run mixes fast. *Mathematical programming*, 86:443–461, 1999.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Yurii E Nesterov, Michael J Todd, et al. On the Riemannian geometry defined by self-concordant barriers and interior-point methods. *Foundations of Computational Mathematics*, 2(4):333–361, 2002.

Van Hoang Nguyen. Dimensional variance inequalities of Brascamp-Lieb type and a local approach to dimensional Prékopa's theorem. *arXiv preprint arXiv:1302.4589*, 2013.

John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Robert L Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

Stan Development Team. RStan: the R interface to Stan, 2020. URL http://mc-stan.org/. R package version 2.21.2.

Ines Thiele, Neil Swainston, Ronan MT Fleming, Andreas Hoppe, Swagatika Sahoo, Maike K Aurich, Hulda Haraldsdottir, Monica L Mo, Ottar Rolfsson, Miranda D Stobbe, et al. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5):419–425, 2013.

Pravin M Vaidya. A new algorithm for minimizing convex functions over convex sets. *Mathematical programming*, 73(3):291–341, 1996.

Santosh Vempala. Geometric random walks: a survey. *Combinatorial and computational geometry*, 52(573-612):2, 2005.

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems (NeurIPS)*, 32, 2019.

Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror Langevin Monte Carlo. In *Conference on Learning Theory (COLT)*, pages 3814–3841. PMLR, 2020.

Appendix A. Convergence rate of ideal RHMC

Lee and Vempala (2018) provided a general framework for computing the mixing rate of RHMC on a manifold embedded in $\mathbb{R}^n$. They represent the mixing rate in terms of the isoperimetry and smoothness parameters depending on the manifold and step size. In particular, they explicitly compute those parameters and isoperimetry for the uniform distribution on a polytope with m constraints, concluding that the mixing rate of RHMC on the Hessian manifold induced by the logarithmic barrier of the polytope is $O\left(mn^{2/3}\right)$. Notably, this mixing rate is independent of the condition number of the polytope. Independence of the condition number is desirable in practice, since real-world instances are highly skewed and thus make it challenging for sampling algorithms to sample efficiently.

Going beyond uniform sampling, we would like to obtain the condition-number-independence of RHMC for more densities. However, even an extension to an exponential density needs care to avoid dependence on a condition number (such as the diameter of the domain).

In this section, we refine this framework by working on a subset $\mathcal{M}_\rho$ instead of $\mathcal{M}$ and extending the smoothness parameters and theorems developed in their paper accordingly. It enables us to couple the one-step distributions of the ideal RHMC starting at two close points by bounding the TV distance in terms of the smoothness parameters.

A.1. Auxiliary function and smoothness parameters

We redefine those smoothness parameters in Lee and Vempala (2018) that depend on a subset $\mathcal{M}_\rho$ of manifold (internally parameterized by $\rho > 0$) and step size h, pointing out how ours differ from the original ones. We then develop the theory for one-step coupling based on the new parameters.

A.1.1. WORKING IN HIGH PROBABILITY REGION

When defining smoothness parameters, Lee and Vempala (2018) pays attention to "well-behaved" Hamiltonian curves γ starting at *any* point in $\mathcal{M}$, where the well-behavedness may be viewed as the average behavior of Hamiltonian curves with *high probability* and is quantified by some auxiliary function. Then the smoothness parameters are estimated by bounding some quantities along the curves. To do so, they should give supremum bounds on those parameters over all points in $\mathcal{M}$, which lead to a weaker mixing rate in the end.

For a refined analysis, we apply a high-probability idea once again to starting points of curves this time. In other words, we consider well-behaved Hamiltonian curves starting only from a good region that has high probability. Then we couple the one-step distributions at two close-by points only in this region. This region will serve as $\mathcal{M}'$ in Proposition 9.

This simple change, however, turns out to yield technical difficulties in following how Lee and Vempala (2018) proceeds with the original parameters. In bounding the overlap of the one-step distributions, they deal with Hamiltonian curves and Hamiltonian variations, starting points of which are on a geodesic between two points, but the geodesic might step out of the good region. Hence, it leads to us considering a different path joining two points instead of the geodesic. We choose the straight line between two points instead and carefully check if the original approach to one-step coupling still goes through. In addition to this, we have to redefine each of the smoothness parameters and modify most of the statements proven in Lee and Vempala (2018) accordingly, as we work in the region smaller than the entire domain. We now formalize this approach.

Definition 10 *Let π be a target distribution on $\mathcal{M}$ such that $\frac{d\pi}{dx} \sim e^{-f(x)}$. Given $\rho > 0$, we call a measurable subset $\mathcal{M}_\rho$ of $\mathcal{M}$ a* good region *if it is convex and has measure $\pi(\mathcal{M}_\rho) \geq 1 - \rho$.*

Good region for exponential density. As mentioned earlier, the necessity of a refined analysis naturally arises in attempts to obtain a condition-number-independent mixing rate of RHMC for $f(x) = e^{-\alpha^\top x}$. One of parameters in Lee and Vempala (2018) depends on $\sup_{x\in\mathcal{M}} \|\nabla f(x)\|^2_{g(x)^{-1}} = \sup_{x\in\mathcal{M}} \|\alpha\|^2_{g(x)^{-1}}$, but this supremum bound can be worsened by scaling up α, and even for fixed α it can be as large as the diameter of $\mathcal{M}$. To address this issue, for given $\rho > 0$ we work on a smaller convex region that has probability at least $1 - \rho$, in which the quantity only depends on the dimension n and ρ, and set it to be a good region. To be precise, we will take $\mathcal{M}_\rho = \left\{x \in \mathcal{M} : \|\nabla f(x)\|^2_{g(x)^{-1}} \leq 10n^2 \log^2 \frac{1}{\rho}\right\}$ for the exponential densities.

A.1.2. AUXILIARY FUNCTION ℓ WITH PARAMETERS ℓ_0 AND ℓ_1

Initial velocities of Hamiltonian trajectories drawn from $\mathcal{N}(0, g(x)^{-1})$ can be large even though it rarely happens, as seen in the standard concentration inequality for Gaussian distributions. Since those worst-case trajectories lead to a weaker coupling, Lee and Vempala (2018) focuses on "well-behaved" Hamiltonian trajectories rather than all trajectories. They formalize this idea by defining an auxiliary function ℓ, which measures how regular a Hamiltonian trajectory is, along with two parameters ℓ_0 and ℓ_1.

Definition 11 *An auxiliary function ℓ with parameters ℓ_0 and ℓ_1 is a function that assigns a non-negative real value to any Hamiltonian curve with step size h, such that*

- *For any $x \in \mathcal{M}_\rho$, we have*

$$\mathbf{P}_\gamma\left(\ell(\gamma) > \frac{1}{2}\ell_0\right) < \frac{1}{100}\min\left(1, \frac{\ell_0}{\ell_1 h}\right),$$

  where γ is a Hamiltonian trajectory starting at x with an initial random velocity drawn from $\mathcal{N}(0, g(x)^{-1})$.

- *For any variations γ_s starting from $\mathcal{M}_\rho$ with $\ell(\gamma_s) \leq \ell_0$, we have*

$$\left|\frac{d}{ds}\ell(\gamma_s)\right| \leq \ell_1 \cdot \left(\left\|\frac{d}{ds}\gamma_s(0)\right\|_{\gamma_s(0)} + \delta \left\|D_s\gamma_s'(0)\right\|_{\gamma_s(0)}\right),$$

  where the variations γ_s satisfy the Hamiltonian equations, and D_s denotes the covariant derivative of the velocity field $\gamma_s'(0)$ along a curve of starting points of the variations.

In the original definitions, the parameter ℓ_0 is defined over the Hamiltonian curves starting from any $x \in \mathcal{M}$, and the parameter ℓ_1 is defined over the variations starting from any $x \in \mathcal{M}$ with $\ell(\gamma_s) \leq \ell_0$.

Intuitions behind these parameters can be understood in the following way. The auxiliary function ℓ measures how regular Hamiltonian trajectories are, and ℓ_0 serves as a threshold that allows us to consider only Hamiltonian curves with regularity below the threshold, while it is large enough to capture most trajectories.

To see the role of ℓ_1, we run through a high-level idea for one-step coupling. For a given endpoint z, we consider the set of regular Hamiltonian curves γ_x stating at x with $\ell(\gamma_x) \leq \frac{1}{2}\ell_0$, which takes into account most trajectories due to the definition of ℓ_0. Along the straight line joining x and y, we smoothly vary the starting point of the Hamiltonian curve to obtain a Hamiltonian curve γ_y starting at y with the same endpoint z and then find a correspondence between γ_x and γ_y. In doing so, it is desirable to maintain the regularity of Hamiltonian curves. In other words, the auxiliary function should not change rapidly so that $\ell(\gamma_y)$ is still bounded by ℓ_0. We enforce this situation via the parameter ℓ_1 that bounds the rate of change of the auxiliary function, $\frac{d}{ds}\ell(\gamma_s)$, along the straight line.

A.1.3. SMOOTHNESS PARAMETERS R_1, R_2, R_3

In relating the regular Hamiltonian curves γ_x and γ_y, some quantities naturally arise from the proof. We begin with the definition of Riemannian curvature tensor and then define three important parameters that govern those quantities.

Definition 12 *The Riemannian curvature tensor is a map $R : V(\mathcal{M}) \times V(\mathcal{M}) \times V(\mathcal{M}) \to V(\mathcal{M})$ for $V(\mathcal{M})$, the collection of vector fields on $\mathcal{M}$, defined by*

$$R(u,v)w = \nabla_u \nabla_v w - \nabla_v \nabla_u w - \nabla_{[u,v]} w \quad \textit{for } u, v, w \in V(\mathcal{M}),$$

where ∇ is the Levi-Civita connection on $\mathcal{M}$, and $[u,v] \stackrel{\text{def}}{=} \nabla_u v - \nabla_v u$ is the Lie bracket of the vector fields u and v.

Definition 13 *Given an auxiliary function ℓ with parameters ℓ_0 and ℓ_1 and the operator $\Phi(\gamma, t) : V(\mathcal{M}) \to V(\mathcal{M})$ defined by $\Phi(\gamma,t)u \stackrel{\text{def}}{=} D_u \mu(\gamma(t)) - R(u, \gamma'(t))\gamma'(t)$,*

- *R_1 is a parameter such that for any $t \in [0,h]$ and any Hamiltonian curves γ starting from $\mathcal{M}_\rho$ with step size h and $\ell(\gamma) \leq \ell_0$*

$$\|\Phi(\gamma,t)\|_{F,\gamma(t)} \stackrel{\text{def}}{=} \sqrt{\mathbb{E}_{v,w\sim\mathcal{N}(0,g(\gamma(t))^{-1})} \langle v, \Phi(\gamma,t)w\rangle^2_{g(\gamma(t))}} \leq R_1.$$

- *R_2 is a parameter such that for any $t \in [0,h]$, any Hamiltonian curves γ starting from $\mathcal{M}_\rho$ with step size h and $\ell(\gamma) \leq \ell_0$, any curve $c(s)$ starting from $\gamma(t)$ and any vector field $v(s)$ along the curve $c(s)$ with $v(0) = \gamma'(t)$,*

$$\left|\frac{d}{ds}\mathrm{Tr}\Phi(v(s))\Big|_{s=0}\right| \leq R_2 \cdot \left(\left\|\frac{dc}{ds}\Big|_{s=0}\right\|_{\gamma(t)} + h\left\|D_s v(s)|_{s=0}\right\|_{\gamma(t)}\right),$$

  where $v(s)$ in Φ indicates a Hamiltonian curve at time t starting with an initial condition $(c(s), v(s))$.

- *R_3 is a parameter such that for any Hamiltonian curves γ starting from $\mathcal{M}_\rho$ with step size h and $\ell(\gamma) \leq \ell_0$, if $\zeta(t) \in T_{\gamma(t)}\mathcal{M}$ is the parallel transport of the vector $\gamma'(0)$ along γ, then*

$$\sup_{t\in[0,h]} \|\Phi(\gamma,t)\zeta(t)\|_{\gamma(t)} \leq R_3.$$

A.2. One-step coupling and convergence rate

In this section, we bound the TV distance of two one-step distributions of the ideal RHMC starting at two close points in terms of the redefined parameters and step size. The following result is a slight tweak of Theorem 29 in Lee and Vempala (2018).

Lemma 14 *For $x, y \in \mathcal{M}_\rho$ and step size $h \le \min\left(\frac{1}{10^5 R_1^{1/2}}, \left(\frac{\ell_0}{10^3 R_1^2 \ell_1}\right)^{1/5}\right)$, if $d_\phi(x,y) \le \frac{1}{100}\min\left(1, \frac{\ell_0}{\ell_1}\right)$, then*

$$d_{TV}(\mathcal{P}_x, \mathcal{P}_y) \le O\left(\frac{1}{h} + h^2 R_2 + h R_3\right) d_\phi(x,y) + \frac{1}{25}.$$

This provides the convergence rate of the ideal RHMC for a general density e^{-f}, which is a slight generalization of Theorem 30 in Lee and Vempala (2018).

Proposition 15 *Let π_T be the distribution obtained after T steps of a lazy ideal RHMC with the stationary distribution π satisfying $\frac{d\pi}{dx} \sim e^{-f(x)}$. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of an initial distribution π_0. For any $\varepsilon > 0$, let $\rho = \frac{\varepsilon}{2\Lambda}$ and $\mathcal{M}_\rho$ a good region. If step size h satisfies*

$$h^2 \le \frac{1}{10^{10} R_1},\ h^5 \le \frac{\ell_0}{10^3 R_1^2 \ell_1},\ h^3 R_2 + h^2 R_3 \le \frac{1}{10^{10}} \text{ and } h \le \frac{1}{10^{10}} \min\left(1, \frac{\ell_0}{\ell_1}\right),$$

where the parameters are defined in Definition 11 and 13, then for the isoperimetry $\psi_{\mathcal{M}_\rho}$ of $\mathcal{M}_\rho$ there exists $T = O\left(\left(h\psi_{\mathcal{M}_\rho}\right)^{-2} \log \frac{1}{\rho}\right)$ such that $d_{TV}(\pi_T, \pi) \le \varepsilon$.

Toward this result, we walk through how each lemma and theorem should change so that they can be put together well, along with the modified smoothness parameters and auxiliary function. We start with the formula of the probability density of the one-step distribution at x.

Lemma 16 (Lee and Vempala (2018), Lemma 10) *The probability density of one-step distribution of RHMC at $x \in \mathcal{M} \subset \mathbb{R}^n$ is*

$$p_x(z) = \sum_{v_x : \mathrm{Ham}_{x,h}(v_x) = z} \underbrace{\left|D\mathrm{Ham}_{x,h}(v_x)\right|^{-1} \sqrt{\frac{|g(z)|}{(2\pi)^n}} \exp\left(-\frac{1}{2}\|v_x\|_x^2\right)}_{\stackrel{\text{def}}{=} p_x^0(v_x)}. \tag{A.1}$$

Note that the velocity v_x is normalized by $g(x)^{-1}$, since the domain of $\mathrm{Ham}_{x,h}$ is endowed with the local metric $g(x)$. In the Euclidean coordinate, the density can be rewritten as

$$\begin{aligned} p_x(z) &= \sum_{v_x' : T_{x,h}(v_x') = z} \left|DT_{x,h}(v_x')\right|^{-1} \frac{1}{\sqrt{(2\pi)^n |g(x)|}} \exp\left(-\frac{1}{2}\left\|v_x'\right\|_{g(x)^{-1}}^2\right) \\ &= \sum_{v_x' : T_{x,h}(v_x') = z} \left|DT_{x,h}(v_x')\right|^{-1} p_x^*(v_x'), \end{aligned} \tag{A.2}$$

where p_x^* is the probability density of the Gaussian distribution $\mathcal{N}(0, g(x))$. This relation follows from $v'_x = g(x)v_x$ and $|DT_{x,h}(v'_x)| = \frac{|D\mathrm{Ham}_{x,h}(v_x)|}{\sqrt{|g(x)||g(x')|}}$ (see the proof of Proposition 30) for $x' = \mathrm{Ham}_{x,h}(v_x) = T_{x,h}(v'_x)$.

We can derive (A.2) in the following way as well. Intuitively, the probability of moving from x to z through one step of RHMC is the summation of the probability of choosing a proper initial velocity that brings x to z, which is the probability density function of $\mathcal{N}(0, g(x))$ divided by $|DT_{x,h}(v'_x)|$. This Jacobian term comes from the change of variables used when moving back from the position space z to the velocity space v'_x.

High-level idea. We are ready to run through a high-level idea of the one-step coupling proof in Lee and Vempala (2018). For two close-by points, it is plausible that the probability densities at x and y are similar, and one should relate those two densities to quantify how close they are, which in turn results in a bound on the overlap of two one-step distributions. It is a Hamiltonian curve that enables them to handle this task in low level. Then they find a one-to-one correspondence between the set of regular Hamiltonian curves from x and y.

As one varies a starting point of a regular Hamiltonian curve from x to y along a curve $c(s)$ joining x and y, one should quantify how fast each term in (A.1) changes. To this end, they first prove that the determinant of Jacobian is close to h^n and that $\mathrm{Ham}_{x,h}$ is locally injective, which makes it possible to work with an approximate but simpler density

$$\tilde{p}_x(z) \stackrel{\text{def}}{=} \sum_{v_x:\mathrm{Ham}_{x,h}(v_x)=z} \sqrt{\frac{|g(z)|}{(2\pi h^2)^n}} \exp\left(-\frac{1}{2}\|v_x\|_x^2\right). \tag{A.3}$$

Next, they prove the following on variations of Hamiltonian curves; for a given endpoint z, as the starting point of a Hamiltonian curve moves along $c(s)$, there exists a unique initial velocity $v_{c(s)}$ at each point on $c(s)$ that brings $c(s)$ to the fixed endpoint z in step size h (i.e., $\mathrm{Ham}_{c(s),h}(v_{c(s)}) = z$). At the same time, they prove that the regularity of Hamiltonian curves, $\ell(\gamma_{c(s)})$, does not change too rapidly along $c(s)$, quantifying how much the proper initial velocity changes as well. These are enough to achieve one-step coupling in terms of only R_1. For further improvement, a more accurate estimate (A.4) of the determinant of Jacobian is used instead, leading to an improved bound via R_2 and R_3.

Following this approach, we elaborate on how each of the proof ingredients can be formalized under the redefined parameters. The first ingredient about the local injectivity of $\mathrm{Ham}_{x,h}$ and an approximation of its Jacobian follows from Lemma 22 in Lee and Vempala (2018) by restricting starting points of Hamiltonian curves to $\mathcal{M}_\rho$ in the statement.

Lemma 17 *Let $\gamma(t) = \mathrm{Ham}_{x,t}(v_x)$ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \le \ell_0$ and step size h satisfying $h^2 \le 1/R_1$. Then $D\mathrm{Ham}_{x,h}$ is invertible and*

$$\left|\log\left|\frac{1}{h}D\mathrm{Ham}_{x,h}(v_x)\right| - \int_0^h \frac{t(h-t)}{h}\mathrm{Tr}\Phi(t)dt\right| \le \frac{(h^2R_1)^2}{10}.$$

As a corollary, we obtain the following estimate on the Jacobian of the Hamiltonian map.

Corollary 18 *Let $(x(t), v(t))$ be the Hamiltonian curve starting with $(x, v) \in \mathcal{M}_\rho \times T_x\mathcal{M}$, where $T_x\mathcal{M} \subset \mathbb{R}^n$ is endowed with the local metric $g(x)$. For step size h with $h^2 \le \frac{1}{10^5\sqrt{n}R_1}$, and*

$v \in T_x\mathcal{M}$ with $\ell(\mathrm{Ham}_{x,t}(v)) \le \ell_0$, we have

$$h^n e^{-\frac{1}{600}} \le |D\mathrm{Ham}_{x,h}(v)| \le h^n e^{\frac{1}{600}}.$$

Namely, $\left|\frac{1}{h^n}|D\mathrm{Ham}_{x,h}(v)| - 1\right| \le 0.002$

The next one is the local uniqueness and existence of Hamiltonian variations, obtained by adjusting Lemma 23 in Lee and Vempala (2018).

Lemma 19 *Let $\gamma(t) = \mathrm{Ham}_{x,t}(v_x)$ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \le \ell_0$ and step size h satisfying $h^2 \le 1/R_1$. Let $x = \gamma(0)$ and $z = \gamma(h)$ be its endpoints.*

- *For a neighborhood U of $x \in \mathcal{M}_\rho$ and neighborhood V of v_x, there exists a unique smooth invertible vector field $v : U \to V$ such that $v(x) = v_x$ and $z = \mathrm{Ham}_{x,h}(v(y))$ for any $y \in U$.*
- *For $\eta \in T_x\mathcal{M}$, we have that $\|\nabla_\eta v(x)\|_x \le \frac{5}{2h}\|\eta\|_x$ and $\left\|\frac{1}{h}\eta + \nabla_\eta v(x)\right\|_x \le \frac{3}{2}R_1 h\|\eta\|_x$.*
- *Let $\gamma_s(t) = \mathrm{Ham}_{c(s),h}(v(c(s))$ be a variation of γ along a path $c(s)$ in U with $c(0) = x$ and $c'(0) = \eta$. For $t \in [0, h]$, we have $\left\|\frac{\partial\gamma_s(t)}{\partial s}\big|_{s=0}\right\|_{\gamma(t)} \le 5\|\eta\|_x$ and $\left\|D_s\gamma_s'(t)\big|_{s=0}\right\|_{\gamma(t)} \le \frac{10}{h}\|\eta\|_x$.*

The first item reveals the local uniqueness and existence of proper initial velocities at any starting point around x. The second item bounds how fast the initial velocity at x change in a given direction η. The last item extends the second result to each point on $\gamma(t)$.

The corresponding result in Lee and Vempala (2018) is made for all regular Hamiltonian curves from $\mathcal{M}$. Since its proof relies on Lemma 17 to apply the implicit function theorem to $f(y, w) = \mathrm{Ham}_{y,h}(w)$ and also on the definition of R_1 for the second result, the statement should be restricted to Hamiltonian curves starting from $\mathcal{M}_\rho$.

We can now prove that regular Hamiltonian curves starting at x with an endpoint z can be smoothly varied along the straight line between x and y, with the regularity of variations almost preserved.

Lemma 20 *Let $\gamma(t) = \mathrm{Ham}_{x,t}(v_x)$ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \le \frac{1}{2}\ell_0$ and step size h satisfying $h^2 \le 1/R_1$. Let $x = \gamma(0)$ and $z = \gamma(h)$ be its endpoints. For $y \in \mathcal{M}_\rho$ and $\beta = \frac{y-x}{\|y-x\|_x}$, let $c(s) = s\beta + x$ be a straight line joining x and y with $c(0) = x$ and $c(s') = y$. Let $s' \le \frac{1}{100}\min\left(1, \frac{\ell_0}{\ell_1}\right)$.*

- *There exists a unique velocity field v along c such that $z = \mathrm{Ham}_{c(s),h}(v(c(s)))$. Furthermore, this vector field is also uniquely determined by $c(s)$ and $v(c(s))$ on $c(s)$.*
- *$\ell(\mathrm{Ham}_{c(s),h}(v(c(s)))) \le \ell_0$ for all s.*

Compared to the original result, we use a straight line instead of a unit-speed geodesic between x and y, as the geodesic might escape the good region $\mathcal{M}_\rho$. The first item implies that there exists an initial velocity $v(c(s))$ at each point $c(s)$ such that we can reach the fixed endpoint z via the Hamiltonian trajectory with the initial condition $(c(s), v(c(s))$. The second item indicates that the regularity of such Hamiltonian trajectories is preserved up to constant along the straight line.

Proof The first result can be proven similarly as in Lee and Vempala (2018). For the second, we denote by γ_s the Hamiltonian trajectory starting at $c(s)$ with the proper initial velocity $v(c(s))$. We note that $s' = \|y - x\|_x$ and $\frac{dc(s)}{ds} = \beta$. By self-concordance of g, we have that $\|\beta\|_{c(s)} \leq (1 + \|x - c(s)\|_x) \|\beta\|_x \leq 1 + s'$. Thus by Lemma 19,

$$\|D_s v(s)\|_{c(s)} \leq \frac{5}{2h} \|\beta\|_{c(s)} \leq \frac{5}{2h} (1 + \|x - c(s)\|_x),$$

and

$$\begin{aligned}
\ell(\gamma_y) &\leq \ell(\gamma_x) + \int_0^{s'} \left| \frac{d}{ds} \ell(\gamma_s) \right| ds \leq \frac{1}{2}\ell_0 + \ell_1 \int_0^{s'} \left(\|\beta\|_{c(s)} + \frac{5}{2} \|\beta\|_{c(s)} \right) ds \\
&\leq \frac{1}{2}\ell_0 + \ell_1 s' \cdot \frac{7}{2}(1 + s') \leq \frac{1}{2}\ell_0 + \ell_1 s' \cdot 4 \\
&\leq \ell_0,
\end{aligned}$$

where we used that $1 + s' \leq 1.01$ and $s' \leq \frac{1}{100}\frac{\ell_0}{\ell_1}$. ■

The next two lemmas provide bounds on some quantities via R_2 and R_3, which are modifications of Lemma 34 and 32 in Lee and Vempala (2018).

Lemma 21 *Let γ_s be a family of Hamiltonian curves joining $c(s)$ and z defined in Lemma 20 with $\ell(\gamma_s) \leq \ell_0$ and step size h satisfying $h^2 \leq 1/R_1$. Then,*

$$\left| \int_0^h \frac{t(h-t)}{h} \frac{d}{ds} \mathrm{Tr}\Phi\left(\gamma_s'(t)\right) dt \right| \leq O\left(h^2 R_2\right).$$

Recall that γ_s given in Lemma 20 has a starting point in $\mathcal{M}_\rho$ with $\ell(\gamma_s) \leq \ell_0$. Since its original proof uses the definition of R_2 and Lemma 19, and they are applicable to regular Hamiltonian curves starting from $\mathcal{M}_\rho$ with $\ell(\gamma_s) \leq \ell_0$, the original proof of this lemma still works with our new definitions of the parameters.

Lemma 22 *Let $\gamma(t) = \mathrm{Ham}_{x,t}(v_x)$ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \leq \ell_0$ and step size h satisfying $h^2 \leq 1/R_1$. Then,*

$$\frac{h}{2} \left| \nabla_\eta \|v(x)\|_x^2 \right| \leq |\langle v_x, \eta \rangle_x| + 3h^2 R_3 \|\eta\|_x.$$

We can follow its original proof by using Lemma 19 and the definition of R_3, as the regular Hamiltonian curve considered starts at $x \in \mathcal{M}_\rho$. We are now ready to prove Lemma 14.

Proof of Lemma 14 Let $c(s)$ be the straight line joining x and y, contained in $\mathcal{M}_\rho$ due to the convexity of $\mathcal{M}_\rho$. We denote $\tilde{\ell} \stackrel{\text{def}}{=} \min\left(1, \frac{\ell_0}{\ell_1 h}\right)$. For $x \in \mathcal{M}_\rho$, let V_x be the set of velocities v_x such that $\ell(\mathrm{Ham}_{x,h}(v_x)) \leq \frac{1}{2}\ell_0$. Note that $\mathcal{P}_x^*(V_x^c) \leq \frac{1}{100}\tilde{\ell}$ by the definition of ℓ_0, where $\mathcal{P}_x^*$ is the one-step distribution over velocities (not position) at x. Since $c(s)$ is contained in $\mathcal{M}_\rho$ and $\gamma(t) = \mathrm{Ham}_{x,t}(v_x)$ has regularity at most $\frac{1}{2}\ell_0$, Lemma 20 guarantees the existence of a family of Hamiltonian variations $\gamma_s(t)$ joining $c(s)$ and $\gamma(h)$ with $\ell(\gamma_s) \leq \ell_0$ for all $s \in [0, \|y - x\|_x]$.

We define an approximate probability density $\tilde{p}_{c(s)}$ of $p_{c(s)}$, where $p_{c(s)}$ is the probability density of $\mathcal{P}_{c(s)}$. Driven by Lemma 17, for $z \in \mathcal{M}$ we define

$$\tilde{p}_{c(s)}(z) \stackrel{\text{def}}{=} \sum_{v:\,\mathrm{Ham}_{c(s),h}(v)=z} \underbrace{\sqrt{\frac{\left|g\left(\mathrm{Ham}_{c(s),h}(v)\right)\right|}{(2\pi h^2)^n}} \exp\left(-\int_0^h \frac{t(h-t)}{h}\mathrm{Tr}\Phi(\gamma_s,t)dt\right) \cdot \exp\left(-\frac{1}{2}\left\|v\right\|_{c(s)}^2\right)}_{\stackrel{\text{def}}{=}\tilde{p}^0_{c(s)}(v)}, \tag{A.4}$$

which is obtained by using $\exp\left(-\int_0^h \frac{t(h-t)}{h}\mathrm{Tr}\Phi(t)dt\right)$ in place of $\left|D\mathrm{Ham}_{c(s),h}(v)\right|^{-1}$ in $p_x(z)$ (see (A.1)).

We now relate $\tilde{p}_{c(s)}$ to $p_{c(s)}$. Note that the ratio of the summand of $\tilde{p}_{c(s)}(z)$ and $p_{c(s)}(z)$ is equal to $\tilde{p}^0_{c(s)}(v)/p^0_{c(s)}(v) = \frac{\frac{1}{h^n}\exp\left(-\int_0^h \frac{t(h-t)}{h}\mathrm{Tr}\Phi(\gamma_s,t)dt\right)}{\left|D\mathrm{Ham}_{c(s),h}(v)\right|^{-1}}$ (see (A.1)). Due to $\ell(\gamma_s) \le \ell_0$, we can apply Lemma 17 to $\gamma_s(t)$, obtaining

$$\exp\left(-\frac{(h^2R_1)^2}{10}\right) \le \frac{\frac{1}{h^n}\exp\left(-\int_0^h \frac{t(h-t)}{h}\mathrm{Tr}\Phi(\gamma_s,t)dt\right)}{\left|D\mathrm{Ham}_{c(s),h}(v)\right|^{-1}} \le \exp\left(\frac{(h^2R_1)^2}{10}\right).$$

Using the conditions on the step size h, we can show that for $C \stackrel{\text{def}}{=} 1 + \frac{1}{10^3}\tilde{\ell}$

$$\exp\left(\frac{(h^2R_1)^2}{10}\right) \le 1 + 2\frac{(h^2R_1)^2}{10} \le 1 + 2\min\left(\frac{1}{10^{10}}, \frac{\ell_0}{10^3\ell_1}\right) \le C.$$

Thus, the ratio is bounded below by C^{-1} and above by C, and it implies that

$$C^{-1} \cdot p^0_{c(s)}(v) \le \tilde{p}^0_{c(s)}(v) \le C \cdot p^0_{c(s)}(v). \tag{A.5}$$

By Lemma 20, for each $v_x \in V_x$ with $\mathrm{Ham}_{x,h}(v_x) = z$ there is a one-to-one correspondence between v_x and v_y, where v_y satisfies $\mathrm{Ham}_{y,h}(v_y) = z$. For this v_y, (A.5) leads to

$$\begin{aligned} p^0_x(v_x) - p^0_y(v_y) &\le C \cdot \tilde{p}^0_x(v_x) - C^{-1} \cdot \tilde{p}^0_y(v_y) \\ &= \left(C^2 - 1\right) C^{-1}\tilde{p}^0_x(v_x) + C^{-1}\left(\tilde{p}^0_x(v_x) - \tilde{p}^0_y(v_y)\right) \\ &\le \left(C^2 - 1\right) p^0_x(v_x) + C^{-1}\left(\tilde{p}^0_x(v_x) - \tilde{p}^0_y(v_y)\right). \end{aligned} \tag{A.6}$$

In a similar way, we can show that

$$\left(C^{-2} - 1\right) p^0_x(v_x) + C\left(\tilde{p}^0_x(v_x) - \tilde{p}^0_y(v_y)\right) \le p^0_x(v_x) - p^0_y(v_y). \tag{A.7}$$

Using this,

$$\begin{aligned} p_x(z) - p_y(z) &= \sum_{v_x:\mathrm{Ham}_{x,h}(v_x)=z} p^0_x(v_x) - \sum_{v_y:\mathrm{Ham}_{y,h}(v_y)=z} p^0_y(v_y) \\ &\le \sum_{v_x \notin V_x:\mathrm{Ham}_{x,h}(v_x)=z} p^0_x(v_x) + \sum_{v_x \in V_x:\mathrm{Ham}_{x,h}(v_x)=z} \left(p^0_x(v_x) - p^0_y(v_y)\right), \end{aligned} \tag{A.8}$$

where in the inequality we only left v_y such that $\mathrm{Ham}_{y,h}(v_y) = z$ and that v_y is the counterpart of $v_x \in V_x$ given by the one-to-one correspondence.

We now bound the TV distance between $\mathcal{P}_x$ and $\mathcal{P}_y$ as follow:

$$\begin{aligned}
d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) &= \frac{1}{2}\int |p_x(z) - p_y(z)|\, dz \\
&\underset{\text{(A.8)}}{\leq} \int_z \sum_{v_x \notin V_x : \mathrm{Ham}_{x,h}(v_x)=z} p_x^0(v_x)dz + \int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} \left|p_x^0(v_x) - p_y^0(v_y)\right| dz \\
&\underset{\text{(A.6), (A.7)}}{\leq} \mathcal{P}_x^*(V_x^c) + \left(C^2 - 1\right)\int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} p_x^0(v_x)dz \\
&\qquad + 2\int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} \left|\tilde{p}_x^0(v_x) - \tilde{p}_y^0(v_y)\right| dz \\
&\leq \frac{\tilde{\ell}}{100} + \frac{\tilde{\ell}}{100}\int_{V_x} p_x^*(v)dv + 2\int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} \int_s \left|\frac{d}{ds}\tilde{p}_{c(s)}^0(v_{c(s)})\right| ds dz \\
&\leq \frac{\tilde{\ell}}{50} + 2\underbrace{\int_s \int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} \left|\frac{d}{ds}\tilde{p}_{c(s)}^0(v_{c(s)})\right| dz\, ds}_{\stackrel{\text{def}}{=} F_s}, \qquad \text{(A.9)}
\end{aligned}$$

where we used that $\int_{V_x} p_x^*(v)dv \leq 1$ in the last inequality, and $v_{c(s)}$ is the initial velocity at $c(s)$ corresponding to $v_x \in V_x$ (via the one-to-one correspondence).

Let us bound F_s in terms of the parameters. From direct computation

$$\frac{d}{ds}\tilde{p}_{c(s)}^0(v_{c(s)}) = \left(-\int_0^h \frac{t(h-t)}{h}\frac{d}{ds}\mathrm{Tr}\Phi(\gamma_s'(t))dt - \frac{1}{2}\frac{d}{ds}\left\|v_{c(s)}\right\|_{c(s)}^2\right)\tilde{p}_{c(s)}^0(v_{c(s)}).$$

Due to $\tilde{p}_{c(s)}^0(v_{c(s)}) \leq 2p_{c(s)}^0(v_{c(s)})$, we have

$$\left|\frac{d}{ds}\tilde{p}_{c(s)}^0(v_{c(s)})\right| \leq 2\left(\left|\int_0^h \frac{t(h-t)}{h}\frac{d}{ds}\mathrm{Tr}\Phi(\gamma_s'(t))dt\right| + \frac{1}{2}\left|\frac{d}{ds}\left\|v_{c(s)}\right\|_{c(s)}^2\right|\right)p_{c(s)}^0(v_{c(s)}).$$

As $\ell(\gamma_s) \leq \ell_0$ due to Lemma 20, it follow from Lemma 21 that

$$\begin{aligned}
F_s &\leq 4\int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} \left(\left|\int_0^h \frac{t(h-t)}{h}\frac{d}{ds}\mathrm{Tr}\Phi(\gamma_s'(t))dt\right| + \frac{1}{2}\left|\frac{d}{ds}\left\|v_{c(s)}\right\|_{c(s)}^2\right|\right)p_{c(s)}^0(v_{c(s)})dz \\
&\leq O\left(h^2 R_2\right)\int_{V_x} p_v^*(v)dv + 2\int_z \sum_{v_x \in V_x : \mathrm{Ham}_{x,h}(v_x)=z} \left|\frac{d}{ds}\left\|v_{c(s)}\right\|_{c(s)}^2\right| p_{c(s)}^0(v_{c(s)})dz \\
&\leq O\left(h^2 R_2\right) + 2\underbrace{\int_{\left\{v\,:\,\ell\left(\mathrm{Ham}_{c(s),h}(v)\right)\leq \ell_0\right\}} \left|\frac{d}{ds}\left\|v\right\|_{c(s)}^2\right| p_{c(s)}^*(v)dv}_{\stackrel{\text{def}}{=} S},
\end{aligned}$$

where we used that $\int_{V_x} p_v^*(v)dv \leq 1$ again for the first term and that $\ell(\gamma_s) \leq \ell_0$ as well as the change of variable with $z = \mathrm{Ham}_{c(s),h}(v(c(s)))$ for the second term.

We now bound S in terms of R_3. As $\ell(\gamma_s) = \ell(\mathrm{Ham}_{c(s),h}(v(c(s)))) \le \ell_0$, we use Lemma 22 to show that

$$\begin{aligned} S &= \mathbb{E}_{\ell(\gamma_s)\le\ell_0}\left|\frac{d}{ds}\|v\|^2_{c(s)}\right| \\ &\le \frac{2}{h}\mathbb{E}_{\ell(\gamma_s)\le\ell_0}\left|\left\langle v, \frac{d}{ds}c(s)\right\rangle_{c(s)}\right| + 6hR_3\mathbb{E}_{\ell(\gamma_s)\le\ell_0}\left\|\frac{d}{ds}c(s)\right\|_{c(s)}. \end{aligned}$$

We recall from the proof of Lemma 20 that $\left\|\frac{d}{ds}c(s)\right\|_{c(s)} \le 1.01$. In addition to this, as v is a Gaussian vector with respect to the local metric, $\left|\left\langle v, \frac{d}{ds}c(s)\right\rangle_{c(s)}\right| = O(1)$ with high probability, which easily follows from the standard concentration inequality for the Gaussian distributions. Therefore,

$$S \le O\left(\frac{1}{h}\right) + 7hR_3.$$

Substituting this back to the inequality for F_s, we have

$$F_s \le O\left(h^2R_2 + \frac{1}{h} + hR_3\right).$$

Putting this to (A.9), it follows that

$$d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) \le O\left(h^2R_2 + \frac{1}{h} + hR_3\right)\|x-y\|_x + \frac{\tilde{\ell}}{50}.$$

Due to $\|x-y\|_{g(x)} \le 2d_\phi(x,y)$ by Lemma 64 and $\tilde{\ell} \le \frac{1}{100}$, it follows that

$$d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) \le O\left(h^2R_2 + \frac{1}{h} + hR_3\right)d_\phi(x,y) + \frac{1}{5000}.$$

■

Using this one-step coupling, we can prove Proposition 15 on the mixing time of the ideal RHMC for a general density e^{-f}.

Proof Due to the assumptions on the step size h, Lemma 14 implies that if $d_\phi(x,y) \lesssim h$, then $d_{\mathrm{TV}}(\mathcal{P}_x, \mathcal{P}_y) \le \frac{1}{1000}$. By Proposition 9 with $\rho = s = \frac{\epsilon}{2\Lambda}$, we can obtain the following lower bound on the s-conductance:

$$\Phi_s = \Omega\left(h\psi_{\mathcal{M}_\rho}\right).$$

By Lemma 7, we have

$$d_{\mathrm{TV}}(\pi_t, \pi) \le s\Lambda + \Lambda\left(1 - \frac{\Phi_s^2}{2}\right)^t.$$

Therefore, it suffices to choose $T = O\left(\left(h\psi_{\mathcal{M}_\rho}\right)^{-2}\log\frac{1}{\rho}\right)$ to ensure $d_{\mathrm{TV}}(\pi_T, \pi) \le \varepsilon$. ■

Appendix B. Convergence rate of discretized RHMC

We bound the remaining two terms, $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \mathcal{P}_x)$ in Section B.2 and $d_{\mathrm{TV}}(\overline{\mathcal{P}}_x, \overline{\mathcal{P}}'_x)$ in Section B.3, obtaining a result on the one-step coupling of RHMC discretized by a numerical integrator with parameters C_x and C_v. To analyze the convergence rate of the discretized RHMC, we define additional parameters.

Definition 23 *Given an auxiliary function ℓ, a good region $\mathcal{M}_\rho$ and step size h, we define new parameters M_1, M_1^*, M_2, M_2^* and $\bar{\ell}_0, \bar{\ell}_1, \bar{R}_1$.*

- *M_1 is a parameter such that for any $t \in [0, h]$ and any Hamiltonian curve γ starting at $x \in \mathcal{M}_\rho$ with step size h and $\ell(\gamma) \le \ell_0$*

$$n \le M_1 \quad \text{and} \quad \|\nabla f(\gamma(t))\|^2_{g(x)^{-1}} \le M_1.$$

- *M_2 is a parameter such that for any $t \in [0, h]$ and any two Hamiltonian curves γ_1, γ_2 starting at $x \in \mathcal{M}_\rho$ with step size h and $\ell(\gamma_i) \le \ell_0$ for $i = 1, 2$*

$$\frac{\|\nabla f(\gamma_1(t)) - \nabla f(\gamma_2(t))\|_{g(x)^{-1}}}{\|\gamma_1(t) - \gamma_2(t)\|_x} \le M_2.$$

- *Let γ be any Hamiltonian curve γ starting from $(x, v) \in \mathcal{M}_\rho \times T_x\mathcal{M}$ with step size h and $\ell(\gamma) \le \ell_0$. Let $\bar{x}_j$'s be intermediate points produced by a numerical integrator with step size h and an initial condition (x, v). We define M_1^* to be the smallest number such that for any $t \in [0, h]$*

$$\frac{|f(\gamma(t)) - f(\bar{x}_j)|}{\|\gamma(t) - \bar{x}_j\|_x} \le \sqrt{M_1^*} \text{ for all } j.$$

  We define M_2^* to be the smallest number such that for any $t \in [0, h]$

$$\frac{\|\nabla f(\gamma(t)) - \nabla f(\bar{x}_j)\|_{g(x)^{-1}}}{\|\gamma(t) - \bar{x}_j\|_x} \le M_2^* \text{ for all } j.$$

- *Let $\overline{\mathcal{M}_\rho}$ be a convex subset of $\mathcal{M}$ that contains $\bar{x}_h$ and $\gamma(h)$. We call an auxiliary function $\bar{\ell}$ symmetric if $\bar{\ell}(\mathrm{Ham}_{x,h}(v)) = \bar{\ell}(\mathrm{Ham}_{x',h}(-v'))$ for $F_h(x, v) = (x', v')$. For a symmetric auxiliary function $\bar{\ell}$, the parameters $\bar{\ell}_0, \bar{\ell}_1$ and $\bar{R}_1$ are defined as in Definition 11 and 13 with $\overline{\mathcal{M}_\rho}$ in place of $\mathcal{M}_\rho$.*

Note that such $\overline{\mathcal{M}_\rho}$ always exists, as $\mathcal{M}$ is convex. We are now ready to formalize the informal statement on the convergence rate of RHMC with a sensitive integrator for a density e^{-f} on the Hessian manifold induced by the highly self-concordant barrier of $\mathcal{M}$.

Theorem 24 *Let π be a target distribution on a convex set $\mathcal{M} \subset \mathbb{R}^n$ and $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let $\mathcal{M}$ be the Hessian manifold with its metric induced by the Hessian of a strongly self-concordant barrier and π_T the distribution obtained after T steps of RHMC discretized by a numerical integrator on $\mathcal{M}$. For any $\varepsilon > 0$, let $\rho = \frac{\varepsilon}{2\Lambda}$ and $\mathcal{M}_\rho$ any good region. If step size h guarantees the sensitivity of the integrator and*

$$h^2 \le \frac{10^{-10}}{\max(R_1, \bar{R}_1)},\ h^5 \le \frac{\ell_0}{10^3 R_1^2 \ell_1},\ h^3 R_2 + h^2 R_3 \le 1,\ h \le \frac{1}{10^{10}} \min\left(1, \frac{\ell_0}{\ell_1}\right),\ h^2 \le \frac{10^{-10}}{n + \sqrt{M_1} + M_2},$$
$$hC_x(x,v) \le \frac{10^{-10}}{\sqrt{n}},\ h^2 C_x(x,v) \le 10^{-10} \min\left(1, \frac{\bar{\ell}_0}{\bar{\ell}_1}, \frac{1}{n + \sqrt{M_1} + \sqrt{M_1^*}}\right),\ h^2 C_v(x,v) \le \frac{10^{-10}}{\sqrt{n + \sqrt{M_1}}}$$

for $x \in \mathcal{M}_\rho$ *and* $v \in V_{good}^x = \left\{v \in \mathbb{R}^n : \|v\|_{g^{-1}} \le 128\sqrt{n},\ \bar{\ell}(\mathrm{Ham}_{x,t}(g(x)^{-1}v)) \le \frac{1}{2}\bar{\ell}_0\right\}$ *(see (B.1)), where the parameters are defined in Definition 11, 13 and 23, then for the isoperimetry* $\psi_{\mathcal{M}_\rho}$ *of* $\mathcal{M}_\rho$ *there exists* $T = O\left(\left(h\psi_{\mathcal{M}_\rho}\right)^{-2} \log \frac{1}{\rho}\right)$ *such that* $d_{TV}(\pi_T, \pi) \le \varepsilon$.

B.1. Stability via self-concordance

We summarize computational lemmas used in coupling one-step distributions and bounding rejection probability. Going forward, the self-concordance of g is repetitively used to relate local metrics g at two close points (see Lemma 25). We recall that $(1-\|x-y\|_{g(x)})^2 g(x) \preceq g(y) \preceq \frac{1}{(1-\|x-y\|_{g(x)})^2} g(x)$ for the local metric g induced by the Hessian of a self-concordant barrier when $\|x-y\|_{g(x)} < c < 1$. It implies that the local norm of a vector with respect to $g(x)$ is within a constant factor of the local norm with respect to $g(y)$ (and vice versa). Namely, for a vector v we have $\|v\|_{g(x)} \le O(1) \cdot \|v\|_{g(y)}$ and $\|v\|_{g(y)} \le O(1) \cdot \|v\|_{g(x)}$. It enables us to move back and forth between the local metric $g(x)$ and $g(y)$ whenever x and y are sufficiently close in the local metric $g(x)$ or $g(y)$.

Lemma 25 *Let* $g(x) = \nabla^2 \phi(x)$ *for some highly self-concordant barrier* ϕ.

- $(1 - \|y-x\|_{g(x)})^2 g(x) \preceq g(y) \preceq \frac{1}{(1-\|y-x\|_{g(x)})^2} g(x)$.
- $\|Dg(x)[v,v]\|_{g(x)^{-1}} \le 2\|v\|_{g(x)}^2$.
- $\|Dg(x)[v,v] - Dg(y)[v,v]\|_{g(x)^{-1}} \le \frac{6}{(1-\|y-x\|_{g(x)})^3} \|v\|_{g(x)}^2 \|y-x\|_{g(x)}$.
- $\|Dg(x)[v,v] - Dg(x)[w,w]\|_{g(x)^{-1}} \le 2\|v-w\|_{g(x)} \|v+w\|_{g(x)}$.

Proof The first fact follows from Theorem 4.1.6 in Nesterov (2003). The second fact follows from Lemma 4.1.2 in Nesterov (2003). To be precise,

$$\|Dg(x)[v,v]\|_{g(x)^{-1}} = \max_{\|v\|_{g(x)}=1} Dg(x)[v,v,u] \le 2\|v\|_{g(x)}^2.$$

The third fact is from the following calculation:

$$\begin{aligned}
&\|Dg(y)[v,v] - Dg(x)[v,v]\|_{g(x)^{-1}} \\
\le &\int_0^1 \|D^2 g(x + t(y-x))[v,v,y-x]\|_{g(x)^{-1}} dt \\
\le &\int_0^1 \frac{1}{1 - t\|y-x\|_{g(x)}} \|D^2 g(x + t(y-x))[v,v,y-x]\|_{g(x+t(y-x))^{-1}} dt \\
\le &\int_0^1 \frac{6}{1 - t\|y-x\|_{g(x)}} \|v\|_{g(x+t(y-x))}^2 \|y-x\|_{g(x+t(y-x))} dt
\end{aligned}$$

$$\leq \int_0^1 \frac{6}{(1-t\|y-x\|_{g(x)})^4} dt \cdot \|v\|_{g(x)}^2 \|y-x\|_{g(x)}$$
$$\leq \frac{6}{(1-\|y-x\|_{g(x)})^3} \|v\|_{g(x)}^2 \|y-x\|_{g(x)}.$$

where the third and fifth line above follow from the first fact, and the fourth line follows from Proposition 9.1.1 in Nesterov and Nemirovskii (1994).

The fourth fact is from the following calculation:

$$\begin{aligned}
&\|Dg(x)[v,v] - Dg(x)[w,w]\|_{g(x)^{-1}} \\
=& \max_{\|u\|_{g(x)}=1} Dg(x)[v,v,u] - Dg(x)[w,w,u] \\
=& \max_{\|u\|_{g(x)}=1} Dg(x)[v-w,v,u] + Dg(x)[w,v-w,u] \\
=& \max_{\|u\|_{g(x)}=1} Dg(x)[v-w,v,u] + Dg(x)[v-w,w,u] \\
=& \max_{\|u\|_{g(x)}=1} Dg(x)[v-w,v+w,u] \\
\leq& \max_{\|u\|_{g(x)}=1} 2\|v-w\|_{g(x)} \|v+w\|_{g(x)} \|u\|_{g(x)} \\
\leq& 2\|v-w\|_{g(x)} \|v+w\|_{g(x)}.
\end{aligned}$$

■

Lemma 26 *For $x, x' \in \mathcal{M}$, let $g = g(x)$ and $g' = g(x')$. Let $\delta_x := \|x - x'\|_g < 0.99$ and $\delta_v := \|v - v'\|_{g^{-1}}$.*

1. $(1 - O(\delta_x))g \preceq g' \preceq (1 + O(\delta_x))g$.
2. $(1 - O(\delta_x))g' \preceq g \preceq (1 + O(\delta_x))g'$.
3. $(1 - O(\delta_x))g^{-1} \preceq g'^{-1} \preceq (1 + O(\delta_x))g^{-1}$.
4. $(1 - O(\delta_x))g'^{-1} \preceq g^{-1} \preceq (1 + O(\delta_x))g'^{-1}$.
5. $-O(\delta_x)I \preceq I - g^{\frac{1}{2}} g'^{-1} g^{\frac{1}{2}} \preceq O(\delta_x)I$.
6. $-O(\delta_x)I \preceq I - g'^{\frac{1}{2}} g^{-1} g'^{\frac{1}{2}} \preceq O(\delta_x)I$.
7. $\left\|g'^{-\frac{1}{2}} g^{\frac{1}{2}}\right\|_2 \leq 1 + O(\delta_x) \quad \& \quad \left\|g'^{\frac{1}{2}} g^{-\frac{1}{2}}\right\|_2 \leq 1 + O(\delta_x)$.
8. $\left\|g^{\frac{1}{2}} g'^{-\frac{1}{2}}\right\|_2 \leq 1 + O(\delta_x) \quad \& \quad \left\|g^{-\frac{1}{2}} g'^{\frac{1}{2}}\right\|_2 \leq 1 + O(\delta_x)$.
9. $\left\|(g^{-1} - g'^{-1})p\right\|_g \lesssim \delta_x \|p\|_{g^{-1}}$.
10. $\left\|g^{-1}p - g'^{-1}q\right\|_g \leq \|p - q\|_{g^{-1}} + O(\delta_x) \|q\|_{g^{-1}}$.

11. $\left\|\frac{\partial H}{\partial v}(x,v) - \frac{\partial H}{\partial v}(x',v')\right\|_g \le \delta_v + O(\delta_x)\left\|v'\right\|_{g^{-1}}$.

12. $\left\|\frac{\partial H}{\partial x}(x,v) - \frac{\partial H}{\partial x}(x',v')\right\|_{g^{-1}} \lesssim (\delta_v + \delta_x \left\|v\right\|_{g^{-1}})(\left\|v\right\|_{g^{-1}} + \left\|v'\right\|_{g^{-1}}) + n\delta_x + \left\|\nabla f(x) - \nabla f(x')\right\|_{g^{-1}}$.

Proof The first four lemmas follow from Lemma 25-1. For 5 (and 6, 7, 8 similarly), using 3

$$(1 - O(\delta_x))I \preceq g^{\frac{1}{2}} g'^{-1} g^{\frac{1}{2}} \preceq (1 + O(\delta_x))I.$$

Thus $-O(\delta_x)I \preceq I - g^{\frac{1}{2}} g'^{-1} g^{\frac{1}{2}} \preceq O(\delta_x)I$. Also by the definition of two-norm, it follows that

$$\left\|g'^{-\frac{1}{2}} g^{\frac{1}{2}}\right\|_2 \le 1 + O(\delta_x).$$

Fact 9 follows from the following computation:

$$\left\|(g^{-1} - g'^{-1})p\right\|_g = \left\|(I - g^{\frac{1}{2}} g'^{-1} g^{\frac{1}{2}}) g^{-\frac{1}{2}} p\right\|_2 \le O(\delta_x)\left\|p\right\|_{g^{-1}}. \quad \text{(Fact 5)}$$

Fact 10 follows from the following computation:

$$\left\|g^{-1}p - g'^{-1}q\right\|_g \le \left\|g^{-1}(p-q) + (g^{-1} - g'^{-1})q\right\|_g \le \left\|p - q\right\|_{g^{-1}} + \underbrace{O(\delta_x)\left\|q\right\|_{g^{-1}}}_{\text{Fact 9}}.$$

Fact 11 follows from the following computation and Fact 10:

$$\left\|\frac{\partial H}{\partial v}(x,v) - \frac{\partial H}{\partial v}(x',v')\right\|_g = \left\|g^{-1}v - g'^{-1}v'\right\|_g \le \delta_v + O(\delta_x)\left\|v'\right\|_{g^{-1}}.$$

For Fact 12, we note that

$$\begin{aligned}
\frac{\partial H}{\partial x}(x,v) - \frac{\partial H}{\partial x}(x',v') &= \left(\nabla f(x) - \nabla f(x')\right) \\
&\quad - \frac{1}{2}\left(Dg\left[g^{-1}v, g^{-1}v\right] - Dg'\left[g'^{-1}v, g'^{-1}v\right]\right) + \frac{1}{2}\left(\mathrm{Tr}(g^{-1}Dg) - \mathrm{Tr}(g'^{-1}Dg')\right) \\
&= -\frac{1}{2}\left(\underbrace{Dg\left[g^{-1}v, g^{-1}v\right] - Dg'\left[g^{-1}v, g^{-1}v\right]}_{F} + \underbrace{Dg'\left[g^{-1}v, g^{-1}v\right] - Dg'\left[g'^{-1}v, g'^{-1}v\right]}_{S}\right) \\
&\quad + \frac{1}{2}\left(\underbrace{\mathrm{Tr}(g^{-1}Dg - g'^{-1}Dg)}_{T} + \underbrace{\mathrm{Tr}(g'^{-1}Dg - g'^{-1}Dg')}_{R}\right) + \left(\nabla f(x) - \nabla f(x')\right).
\end{aligned}$$

For F, by the third fact in Lemma 25

$$\left\|F\right\|_{g^{-1}} \lesssim \frac{1}{(1-\delta_x)^3}\left\|g^{-1}v\right\|_g^2 \left\|x - x'\right\|_g = \frac{1}{(1-\delta_x)^3}\left\|v\right\|_{g^{-1}}^2 \delta_x \lesssim \delta_x \left\|v\right\|_{g^{-1}}^2.$$

For S, by the fourth fact in Lemma 25

$$\begin{aligned}
\left\|S\right\|_{g^{-1}} \lesssim \left\|S\right\|_{g'^{-1}} &\lesssim \left\|g^{-1}v - g'^{-1}v'\right\|_{g'} \left\|g^{-1}v + g'^{-1}v'\right\|_{g'} \\
&\lesssim \left(\left\|v - v'\right\|_{g'^{-1}} + O(\delta_x)\left\|v\right\|_{g'^{-1}}\right)\left(\left\|v\right\|_{g'^{-1}} + \left\|v'\right\|_{g'^{-1}}\right) \\
&\lesssim \left(\delta_v + \delta_x \left\|v\right\|_{g^{-1}}\right)\left(\left\|v\right\|_{g'^{-1}} + \left\|v'\right\|_{g'^{-1}}\right).
\end{aligned}$$

For T, using the stochastic estimator of trace

$$\begin{aligned}
\left\|\mathrm{Tr}\left(g^{-1}Dg - g'^{-1}Dg\right)\right\|_{g^{-1}} &= \max_{\|u\|_g=1} \mathrm{Tr}\left((g^{-1}-g'^{-1})Dg[u]\right) \\
&= \max_{\|u\|_g=1} \mathrm{Tr}\left(g^{\frac{1}{2}}(g^{-1}-g'^{-1})Dg[u]g^{-\frac{1}{2}}\right) \\
&= \max_{\|u\|_g=1} \mathbb{E}_{z\sim\mathcal{N}(0,I)}\left[z^\top g^{\frac{1}{2}}(g^{-1}-g'^{-1})Dg[u]g^{-\frac{1}{2}}z\right] \\
&= \max_{\|u\|_g=1} \mathbb{E}Dg\left[u, g^{-\frac{1}{2}}z, (g^{-1}-g'^{-1})g^{\frac{1}{2}}z\right] \\
&\le 2\mathbb{E}\max_{\|u\|_g=1}\|u\|_g\left\|g^{-\frac{1}{2}}z\right\|_g \underbrace{\left\|\left(g^{-1}-g'^{-1}\right)g^{\frac{1}{2}}z\right\|_g}_{\text{Fact 9}} \\
&\le O(\delta_x)\mathbb{E}\|z\|_2\left\|g^{\frac{1}{2}}z\right\|_{g^{-1}} = O(\delta_x)\mathbb{E}\|z\|_2^2 \\
&= O(n\delta_x).
\end{aligned}$$

For R, in a similar way that we bounded $\|T\|_{g^{-1}}$

$$\begin{aligned}
\left\|\mathrm{Tr}\left(g'^{-1}Dg - g'^{-1}Dg'\right)\right\|_{g^{-1}} &= \left\|\mathbb{E}_{z\sim\mathcal{N}(0,I)}\left(Dg[g'^{-\frac{1}{2}}z, g'^{-\frac{1}{2}}z] - Dg'[g'^{-\frac{1}{2}}z, g'^{-\frac{1}{2}}z]\right)\right\|_{g^{-1}} \\
&\le \mathbb{E}\underbrace{\left\|Dg[g'^{-\frac{1}{2}}z, g'^{-\frac{1}{2}}z] - Dg'[g'^{-\frac{1}{2}}z, g'^{-\frac{1}{2}}z]\right\|_{g^{-1}}}_{\text{Use Lemma 25}} \\
&\le O(\delta_x)\mathbb{E}\left\|g'^{-\frac{1}{2}}z\right\|_{g'}^2 \le O(\delta_x)\mathbb{E}\|z\|_2^2 \\
&= O(n\delta_x).
\end{aligned}$$

By adding up these bounds, we obtain

$$\left\|\tfrac{\partial H}{\partial x}(x,v) - \tfrac{\partial H}{\partial x}(x',v')\right\|_{g^{-1}} \lesssim \left(\delta_v + \delta_x\|v\|_{g^{-1}}\right)\left(\|v\|_{g^{-1}} + \|v'\|_{g^{-1}}\right) + n\delta_x + \|\nabla f(x) - \nabla f(x')\|_{g^{-1}}.$$

■

We now bound the partial derivatives of H with respect to x and v. For H_1 and H_2 given by

$$H_1(x,v) = f(x) + \frac{1}{2}\log\det g(x) \;\text{ and }\; H_2(x,v) = \frac{1}{2}v^\top g(x)^{-1}v,$$

we recall from (2.1) that

$$\begin{aligned}
&\frac{\partial H_1}{\partial x}(x,v) = \nabla f(x) + \frac{1}{2}\mathrm{Tr}(g^{-1}Dg), \\
&\frac{\partial H_2}{\partial x}(x,v) = -\frac{1}{2}Dg[g^{-1}v, g^{-1}v] \;\text{ and }\; \frac{\partial H_2}{\partial v}(x,v) = g^{-1}v.
\end{aligned}$$

Lemma 27 *For $x\in\mathcal{M}$ and $g := g(x)$, the following inequalities hold.*

$$\left\|\frac{\partial H_1(x,v)}{\partial x}\right\|_{g^{-1}} \le \|\nabla f(x)\|_{g^{-1}} + n,$$

$$\left\|\frac{\partial H_2(x,v)}{\partial v}\right\|_g \le \|v\|_{g^{-1}} \quad\&\quad \left\|\frac{\partial H_2(x,v)}{\partial x}\right\|_{g^{-1}} \le \|v\|_{g^{-1}}^2$$

Proof For $\frac{\partial H_1(x,v)}{\partial x}$,

$$\left\|\frac{\partial H_1(x,v)}{\partial x}\right\|_{g^{-1}} \le \left\|\nabla f(x) + \frac{1}{2}\mathrm{Tr}(g^{-1}Dg)\right\|_{g^{-1}} \le \|\nabla f(x)\|_{g^{-1}} + \left\|\frac{1}{2}\mathrm{Tr}(g^{-1}Dg)\right\|_{g^{-1}}.$$

Note that

$$\left\|\frac{1}{2}\mathrm{Tr}(g^{-1}Dg)\right\|_{g^{-1}} = \frac{1}{2}\max_{\|u\|_g=1}\mathrm{Tr}(g^{-1}Dg[u])$$

By self-concordance, for any $h \in \mathbb{R}^n$ we have $h^\top Dg[u]h \le 2\|h\|_g^2$ and thus $Dg[u] \preceq 2g$, resulting in $g^{-\frac{1}{2}}Dg[u]g^{-\frac{1}{2}} \preceq 2I$. Then

$$\mathrm{Tr}(g^{-1}Dg[u]) \le 2\mathrm{Tr}(I) \le 2n.$$

For $\frac{\partial H_2(x,v)}{\partial v}$,

$$\left\|\frac{\partial H_2(x,v)}{\partial v}\right\|_g = \left\|g^{-1}v\right\|_g = \|v\|_{g^{-1}}.$$

For $\frac{\partial H_2(x,v)}{\partial x}$,

$$\left\|\frac{\partial H_2(x,v)}{\partial x}\right\|_{g^{-1}} \le \left\|\frac{1}{2}Dg\left[g^{-1}v, g^{-1}v\right]\right\|_{g^{-1}} \le \left\|g^{-1}v\right\|_g^2 = \|v\|_{g^{-1}}^2,$$

where the second step follows from Lemma 25. ■

B.2. Coupling between ideal and discretized RHMC

We bound $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \mathcal{P}_x)$, the TV distance between the one-step distributions of the ideal RHMC and the discretized RHMC without the rejection step. We use $\overline{\mathcal{P}}_x$ to indicate $\overline{\mathcal{P}}'_x$ for simplicity in this section only. We denote by p_x and $\overline{p}_x$ the probability density functions of $\mathcal{P}_x$ and $\overline{\mathcal{P}}_x$ respectively. We let $g = g(x)$ and $g_t = g(x_t)$.

Let us elaborate on our approach. We work with the Euclidean metric this time, as we find it easier to handle numerical integrators with the Euclidean representation. As mentioned in (A.2), the one-step distributions $\mathcal{P}_x$ and $\overline{\mathcal{P}}_x$ of the ideal and discretized RHMC on $\mathcal{M}$ are the pushforwards by T_x and $\overline{T}_x$ of the Gaussian distribution of initial velocities on the tangent space $T_x\mathcal{M}$. Thus, it follows by the change of variables that for $z = T_x(v^*) = \overline{T}_x(v)$ these two probability densities on the different spaces (one on $\mathcal{M}$ and another on $T_x\mathcal{M}$) are related as follows. For p_x^* the probability density function of $\mathcal{N}(0, g(x))$,

$$p_x(z) = \sum_{v^*:T_x(v^*)=z} \frac{p_x^*(v^*)}{|DT_x(v^*)|} \quad \text{and} \quad \overline{p}_x(z) = \sum_{v:\overline{T}_x(v)=z} \frac{p_x^*(v)}{\left|D\overline{T}_x(v)\right|}.$$

We aim to couple these v^* and v on $T_x\mathcal{M}$. In this coupling, we can exclude 'bad' velocities, as long as such velocities have small measure. To see this, let V_{bad}^x be a set of bad initial velocities of

measure $\varepsilon < 1$ and V^x_{good} be the rest. Assuming a one-to-one correspondence between v and v^* for $v \in V^x_{\text{good}}$, we have that for $x \in \mathcal{M}_\rho$

$$\begin{aligned}
d_{\text{TV}}(\overline{\mathcal{P}}_x, \mathcal{P}_x) &= \sup_{A \subset \mathcal{M}} \int_A (\overline{p}_x(z) - p_x(z)) dz \\
&\le \sup_{A \subset \mathcal{M}} \int_A \left(\sum_{v : \overline{T}_x(v) = z} \frac{p^*_x(v)}{\left| D\overline{T}_x(v) \right|} - \sum_{v^* : T_x(v^*) = z} \frac{p^*_x(v^*)}{|DT_x(v^*)|} \right) dz \\
&\le \int_{V^x_{\text{bad}}} p^*_x(v) dv + \sup_{A \subset \mathcal{M}} \int_A \sum_{v \in V^x_{\text{good}} : \overline{T}_x(v) = z} \left(\frac{p^*_x(v)}{\left| D\overline{T}_x(v) \right|} - \frac{p^*_x(v^*)}{|DT_x(v^*)|} \right) dz \\
&= P^*_x(V^x_{\text{bad}}) + \sup_{A \subset \mathcal{M}} \int_A \sum_{v \in V^x_{\text{good}} : \overline{T}_x(v) = z} \frac{p^*_x(v)}{\left| D\overline{T}_x(v) \right|} \left(1 - \frac{p^*_x(v^*)}{p^*_x(v)} \frac{\left| D\overline{T}_x(v) \right|}{|DT_x(v^*)|} \right) dz,
\end{aligned}$$

where P^*_x is $\mathcal{N}(0, g(x))$, and in the third line we used the one-to-one correspondence between v and v^* to pair them in the summation. If we show that on $v \in V^x_{\text{good}}$ the term of $1 - \frac{p^*_x(v^*)}{p^*_x(v)} \frac{\left| D\overline{T}_x(v) \right|}{|DT_x(v^*)|}$ is bounded by a small constant (say, η), then

$$\begin{aligned}
d_{\text{TV}}(\overline{\mathcal{P}}_x, \mathcal{P}_x) &\le P^*_x(V^x_{\text{bad}}) + \eta \sup_{A \subset \mathcal{M}} \int_A \sum_{v \in V^x_{\text{good}} : \overline{T}_x(v) = z} \frac{p^*_x(v)}{\left| D\overline{T}_x(v) \right|} dz \\
&\le P^*_x(V^x_{\text{bad}}) + \eta \int_{V^x_{\text{good}}} p^*_x(v) dv \\
&= P^*_x(V^x_{\text{bad}}) + \eta P^*_x(V^x_{\text{good}}) \\
&\le \eta + (1 - \eta)\varepsilon.
\end{aligned}$$

By taking ε sufficiently small, we can bound $d_{\text{TV}}(\overline{\mathcal{P}}_x, \mathcal{P}_x)$ smaller then $1/10$.

For each $x \in \mathcal{M}$, our bad set V^x_{bad} of velocities is the union of the following sets:

$$\begin{aligned}
V_1 &= \left\{ v \in \mathbb{R}^n : \|v\|_{g^{-1}} > 128\sqrt{n} \right\}, \\
V_2 &= \left\{ v \in \mathbb{R}^n : \bar{\ell}(\text{Ham}_{x,t}(g(x)^{-1}v)) > \frac{1}{2}\bar{\ell}_0 \right\},
\end{aligned}$$

and thus

$$V^x_{\text{good}} = \left\{ v \in \mathbb{R}^n : \|v\|_{g^{-1}} \le 128\sqrt{n},\ \bar{\ell}(\text{Ham}_{x,t}(g(x)^{-1}v)) \le \frac{1}{2}\bar{\ell}_0 \right\}. \tag{B.1}$$

We remark that a velocity $v \in \mathbb{R}^n$ should be normalized by $g(x)^{-1}$ before feeding into $\text{Ham}_{x,t}$, since the domain $T_x\mathcal{M}$ of $\text{Ham}_{x,t}$ is endowed with the local metric. Since the standard concentration inequality for the Gaussian distributions implies that $P^*_x(V_1) < \frac{1}{100}$, and the definition $\bar{\ell}_0$ implies that $P^*_x(V_2) < \frac{1}{100}$, it follows that $P^*_x(V^x_{\text{bad}}) < 0.02$.

B.2.1. DYNAMICS OF IDEAL AND DISCRETIZED RHMC

We study the dynamics of the ideal and discretized RHMC.

Proposition 28 *For $x \in \mathcal{M}_\rho$ and $v, v' \in V^x_{good}$, let $g := g(x)$ and h step size satisfying*

$$\underbrace{h^2 \le \frac{10^{-10}}{n+\sqrt{M_1}+M_2}}_{①}, \underbrace{h^2 \le \frac{10^{-10}}{\bar{R}_1}}_{②}, \underbrace{C_x(x,v)h^2 \le \frac{1}{10^{10}}\min\left(1, \frac{\bar{\ell}_0}{\bar{\ell}_1}\right)}_{③}, \underbrace{C_x(x,v)h \le \frac{\sqrt{n}}{10^{10}}}_{④}.$$

For $t \in [0,h]$, we let (x_t, v_t) and (x'_t, v'_t) be the Hamiltonian curves of the ideal RHMC at time t with initial conditions (x,v) and (x,v'), respectively. Let $(\overline{x}, \overline{v})$ be the point obtained from RHMC with a sensitive second-order numerical integrator with the step size h and initial condition (x,v). Let $\phi \stackrel{\text{def}}{=} \sup_{t\in[0,h]} \|x_t - x'_t\|_{g(x_t)}$, $\psi \stackrel{\text{def}}{=} \sup_{t\in[0,h]} \|v_t - v'_t\|_{g(x_t)^{-1}}$ and $\Gamma_t(v) \stackrel{\text{def}}{=} g(x_t)^{-1}(v_t - v)$.

1. $\|x - x_t\|_g = O\left(t\sqrt{n} + t^2(n+\sqrt{M_1})\right) < \frac{1}{4}$ and $\|v - v_t\|_{g^{-1}} = O\left(t(n+\sqrt{M_1})\right)$.
2. $(1-o(1))\|v_h - v'_h\|_{g^{-1}} \le \psi \le (1+o(1))\|v - v'\|_{g^{-1}}$.
3. $(1-o(1))\|T_x(v) - T_x(v')\|_g \le \phi \le (1+o(1))h\psi \le (1+o(1))h\|v - v'\|_g$.
4. $\|\Gamma_t(v) - \Gamma_t(v')\|_g \le L\|v - v'\|_{g^{-1}}$ for some $L < 1/10$.
5. For $z = \overline{T}_x(v)$, there exists $v^* \in \mathbb{R}^n$ with $z = T_x(v^*)$ such that $\bar{\ell}(\mathrm{Ham}_{x,t}(g(x)^{-1}v^*)) \le \bar{\ell}_0$ and $\|v - v^*\|_{g^{-1}} = O\left(\frac{1}{h}\right)\left\|T_x(v) - \overline{T}_x(v)\right\|_g$. Moreover, there is a one-to-one correspondence between v and v^*.

Proof of 1. For $0 \le t \le h$, let us define $\phi(t) := \|x - x_t\|_g$ and $\psi(t) := \|v - v_t\|_{g^{-1}}$. Note that

$$\begin{aligned} 2\|x - x_t\|_g \frac{d\phi(t)}{dt} = \frac{d\phi^2(t)}{dt} &= \frac{d}{dt}(x_t - x)^\top g(x_t - x) \\ &= 2\left(\frac{\partial H}{\partial v}(x_t, v_t)\right)^\top g(x_t - x). \end{aligned}$$

Hence,

$$\left|\|x - x_t\|_g \phi'(t)\right| = \left|\left(\frac{\partial H}{\partial v}(x_t, v_t)\right)^\top g(x_t - x)\right| \le \left\|\frac{\partial H}{\partial v}(x_t, v_t)\right\|_g \|x_t - x\|_g,$$

and $|\phi'(t)| \le \left\|\frac{\partial H}{\partial v}(x_t, v_t)\right\|_g$. When $\|x - x_t\|_g < \frac{1}{4}$ for $0 \le t \le h$, since the local norms at x and x_t are within a small constant factor as follows, we have

$$\begin{aligned} \left\|\frac{\partial H}{\partial v}(x_t, v_t)\right\|_g \le 2\left\|\frac{\partial H}{\partial v}(x_t, v_t)\right\|_{g_t} &= 2\|v_t\|_{g_t^{-1}} \quad \text{(Lemma 27)} \\ &\le 4\|v_t\|_{g^{-1}} \le 4(\|v_t - v\|_{g^{-1}} + \|v\|_{g^{-1}}), \end{aligned}$$

and thus

$$\phi'(t) \le 10^3\sqrt{n} + 4\psi(t) \quad \text{if } \|x - x_t\|_g < \frac{1}{4}. \tag{B.2}$$

Similarly, we can obtain

$$\begin{aligned} 2\left\|v-v_t\right\|_{g^{-1}}\frac{d\psi}{dt} &= \frac{d\psi^2(t)}{dt} = \frac{d}{dt}(v_t-v)^\top g^{-1}(v_t-v) \\ &= -2\left(\frac{\partial H}{\partial x}(x_t,v_t)\right)^\top g^{-1}(v_t-v), \end{aligned}$$

and thus $|\psi'(t)| \le \left\|\frac{\partial H}{\partial x}(x_t,v_t)\right\|_{g^{-1}}$. If $\left\|x-x_t\right\|_g < \frac{1}{4}$ for $0 \le t \le h$, then by Lemma 27

$$\begin{aligned} \left\|\frac{\partial H}{\partial x}(x_t,v_t)\right\|_{g^{-1}} &\le 2(\|v_t\|_{g_t^{-1}}^2 + \sqrt{M_1}+n) \quad (\because x \in \mathcal{M}_\rho) \\ &\le 2\left(4\left\|v_t\right\|_{g^{-1}}^2 + \sqrt{M_1}+n\right) \\ &\le 2\left(8(\|v\|_{g^{-1}}^2 + \|v_t-v\|_{g^{-1}}^2) + \sqrt{M_1}+n\right) \quad (\because (a+b)^2 \le 2(a^2+b^2)) \\ &\le 10^6 n + 16\psi^2(t) + 2\sqrt{M_1}, \end{aligned}$$

and thus

$$\psi'(t) \le 10^6 n + 16\psi^2(t) + 2\sqrt{M_1} \quad \text{if } \left\|x-x_t\right\|_g < \frac{1}{4}. \tag{B.3}$$

Now let us solve the coupled inequalities (B.2) and (B.3). When $\psi(t) \le 1000t(n+\sqrt{M_1})$, (B.3) becomes

$$\psi'(t) \le 10^6 n + 2\sqrt{M_1} + 16\cdot 10^6 t^2 (n+\sqrt{M_1})^2,$$

and this inequality holds up until h satisfying $\int_0^h \left(10^6 n + 2\sqrt{M_1} + 16\cdot 10^6 t^2(n+\sqrt{M_1})^2\right) dt \le 1000h(n+\sqrt{M_1})$. We can check that for any $h \le \frac{10^{-10}}{\sqrt{n+\sqrt{M_1}}}$ (i.e., condition ①) this integral inequality is satisfied. Recall that we have to ensure that $\phi(t) < \frac{1}{4}$ for $t \le h$. By substituting $\psi(t) \le 1000t(n+\sqrt{M_1})$ into (B.2), we have

$$\phi'(t) \le 10^3\sqrt{n} + 4000t(n+\sqrt{M_1}),$$

as long as $\phi(t) = \left\|x_t - x\right\|_g < \frac{1}{4}$. It is straightforward to see that for $t \le h$ one has

$$\begin{aligned} \phi(t) &\le 10^3\sqrt{n}t + 2000t^2(n+\sqrt{M_1}) \le 10^{-7}\sqrt{\frac{n}{n+\sqrt{M_1}}} + \frac{2000}{10^{10}} \\ &< \frac{1}{10^6}. \end{aligned}$$

Proof of 2. From the first item, $\left\|x_t - x_t'\right\|_g \le \left\|x-x_t\right\|_g + \left\|x-x_t'\right\|_g < 10^{-5}$. Due to Lemma 26, we can switch the local norms among $g, g_t = g(x_t)$ and $g_t' = g(x_t')$ by losing a multiplicative constant like $1+10^{-4}$.

For $\delta_x = \left\|x_t - x_t'\right\|_{g_t}$ and $\delta_v = \left\|v_t - v_t'\right\|_{g_t^{-1}}$,

$$\begin{aligned} \left\|v_t - v_t'\right\|_{g_t} &= \left\|v - v' + \int_0^t \left(\frac{\partial H}{\partial x}(x_s,v_s) - \frac{\partial H}{\partial x}(x_s',v_s')\right) ds\right\|_{g_t^{-1}} \\ &\le \left\|v-v'\right\|_{g_t^{-1}} + O(h)\sup_{t\in[0,h]}\left\|\frac{\partial H}{\partial x}(x_t,v_t) - \frac{\partial H}{\partial x}(x_t',v_t')\right\|_{g_t^{-1}} \\ &\le \left\|v-v'\right\|_{g_t^{-1}} + O(h)\sup_{t\in[0,h]}\left((\delta_v + \delta_x\left\|v_t\right\|_{g^{-1}})(\left\|v_t\right\|_{g^{-1}} + \left\|v_t'\right\|_{g^{-1}}) + (n+M_2)\delta_x\right), \end{aligned}$$

where the last step follows from Lemma 26-12. By the first item and ①, we have $\|v_t\|_{g^{-1}}, \|v'_t\|_{g^{-1}} \le 7\sqrt{n+\sqrt{M_1}}$ and thus

$$\begin{aligned}\left\|v_t - v'_t\right\|_{g_t^{-1}} &\le \left\|v - v'\right\|_{g_t^{-1}} + O(h)\left(\psi\sqrt{n+\sqrt{M_1}} + \phi\left(n+\sqrt{M_1}+M_2\right)\right)\\ &\le (1+o(1))\left\|v-v'\right\|_{g^{-1}} + O(h)\psi\left(\sqrt{n+\sqrt{M_1}} + h\left(n+\sqrt{M_1}+M_2\right)\right),\end{aligned}$$

where we used $\phi \le (1+o(1))O(h)\psi$ that we prove in the next item. Taking the supremum over $t \in [0,h]$, we obtain

$$\left(1 - O(h)\left(\sqrt{n+\sqrt{M_1}} + h\left(n+\sqrt{M_1}+M_2\right)\right)\right)\psi \le (1+o(1))\left\|v-v'\right\|_{g^{-1}}.$$

Taking a sufficiently small constant in h and using ①, it follows that $\psi \le (1+o(1))\left\|v-v'\right\|_{g^{-1}}$.

Proof of 3. By Lemma 26-11,

$$\begin{aligned}\left\|x_t - x'_t\right\|_{g_t} &= \left\|x + \int_0^t \frac{\partial H}{\partial v}(x_s, v_s)ds - \left(x + \int_0^t \frac{\partial H}{\partial v}(x'_s, v'_s)ds\right)\right\|_{g_t}\\ &\le O(h)\sup_{t\in[0,h]}\left\|\frac{\partial H}{\partial v}(x_t,v_t) - \frac{\partial H}{\partial v}(x'_t,v'_t)\right\|_{g_t}\\ &\le O(h)\left(\psi + \phi\sqrt{n+\sqrt{M_1}}\right).\end{aligned}$$

By using ① and taking the supremum over $t \in [0,h]$ and a sufficiently small constant in h, we obtain the inequality of $\phi \le (1+o(1))O(h)\psi$ as we promised, and the second item implies that

$$\phi \le (1+o(1))O(h)\left\|v-v'\right\|_{g^{-1}}.$$

Proof of 4. By Lemma 26-12,

$$\begin{aligned}\left\|\Gamma_t(v) - \Gamma_t(v')\right\|_{g_t} &= \left\|g_t^{-1}(v_t - v) - g_t'^{-1}(v'_t - v')\right\|_{g_t}\\ &\le \left\|v_t - v - v'_t + v'\right\|_{g_t^{-1}} + O(1)\underbrace{\left\|x_t - x'_t\right\|_{g_t}}_{\le\phi}\underbrace{\left\|v'_t - v'\right\|_{g_t^{-1}}}_{\le O(h(n+\sqrt{M_1}))}\\ &\le \left\|\int_0^t\left(\frac{\partial H}{\partial x}(x_s,v_s) - \frac{\partial H}{\partial x}(x'_s,v'_s)\right)ds\right\|_{g_t^{-1}} + O\left(h\left(n+\sqrt{M_1}\right)\right)\phi\\ &\le O(h)\sup_{t\in[0,h]}\left\|\frac{\partial H}{\partial x}(x_t,v_t) - \frac{\partial H}{\partial x}(x'_t,v'_t)\right\|_{g_t^{-1}} + O\left(h\left(n+\sqrt{M_1}\right)\right)\phi.\end{aligned}$$

We can bound the first term by $O(h^2)\left(\sqrt{n+\sqrt{M_1}} + h\left(n+\sqrt{M_1}+M_2\right)\right)\psi$ by following the proof of the second item. Using the second and third items with the condition ①, and taking a sufficiently small constant in h, for some $L < 1/10$

$$\begin{aligned}\left\|\Gamma_t(v) - \Gamma_t(v')\right\|_g &\lesssim h^2\sqrt{n+\sqrt{M_1}} + h^3\left(n+\sqrt{M_1}+M_2\right) + h^2\left(n+\sqrt{M_1}\right)\left\|v-v'\right\|_{g^{-1}}\\ &\le L\left\|v-v'\right\|_{g^{-1}}.\end{aligned}$$

Proof of 5. Let $z = \overline{T}_x(v)$ for $v \in V^x_{\text{good}}$. We show that the map defined on $u \in V_{\text{dom}} = \{v' \in \mathbb{R}^n : \|v - v'\|_{g^{-1}} \le 4\sqrt{n}\}$ by

$$\Upsilon(u) = u - \frac{1}{h} g T_x(u) + \frac{1}{h} g z,$$

is Lipschitz in u with respect to the local norm g^{-1}, and then apply the Banach fixed-point theorem to obtain the unique fixed-point v^*. Note that it satisfies $g(T_x(v^*) - \overline{T}_x(v)) = 0$ and thus $T_x(v^*) = \overline{T}_x(v)$.

For Lipschitzness, let (x_t, u_t) and (x'_t, u'_t) be the Hamiltonian curves of the ideal RHMC starting from (x, u) and (x, u') for $u, u' \in V_{\text{dom}}$, respectively. Observe that

$$\begin{aligned}
\left\|\Upsilon(u) - \Upsilon(u')\right\|_{g^{-1}} &= \left\|u - u' - \frac{1}{h} g \left(T_x(u) - T_x(u')\right)\right\|_{g^{-1}} \qquad \text{(B.4)}\\
&= \left\|u - u' - \frac{1}{h}\int_0^h g\left(g_t^{-1} u_t - g_t^{-1} u'_t\right) dt\right\|_{g^{-1}}\\
&= \left\|\left(I - \frac{1}{h} g \int_0^h g_t^{-1} dt\right) u - \left(I - \frac{1}{h} g \int_0^h g_t'^{-1} dt\right) u' - \frac{1}{h}\int_0^h g\left(\Gamma_t(u) - \Gamma_t(u')\right) dt\right\|_{g^{-1}}\\
&= \left\|\underbrace{\frac{1}{h}\left(\int_0^h (I - g g_t^{-1}) dt\right)}_{I_u} u - \underbrace{\frac{1}{h}\left(\int_0^h (I - g g_t'^{-1}) dt\right)}_{I_{u'}} u' - \frac{1}{h}\int_0^h g\left(\Gamma_t(u) - \Gamma_t(u')\right) dt\right\|_{g^{-1}}\\
&\le \left\|I_u(u - u') + (I_u - I_{u'})u'\right\|_{g^{-1}} + \frac{1}{h}\left\|\int_0^h g\left(\Gamma_t(u) - \Gamma_t(u')\right) dt\right\|_{g^{-1}}\\
&\le \left\|I_u(u - u')\right\|_{g^{-1}} + \left\|(I_u - I_{u'})u'\right\|_{g^{-1}} + \sup_{t\in[0,h]} \left\|\Gamma_t(u) - \Gamma_t(u')\right\|_g\\
&\le \underbrace{\left\|I_u(u - u')\right\|_{g^{-1}}}_{F} + \underbrace{\left\|(I_u - I_{u'})u'\right\|_{g^{-1}}}_{S} + L\left\|u - u'\right\|_{g^{-1}}, \qquad \text{(B.5)}
\end{aligned}$$

where the last inequality follows from the fourth item.

For F, let $p = u - u'$ and observe that

$$\begin{aligned}
\|I_u p\|_{g^{-1}} &\le \frac{1}{h}\int_0^h \left\|(I - g g_t^{-1})p\right\|_{g^{-1}} dt \le \sup_{t\in[0,h]} \left\|(I - g g_t^{-1})p\right\|_{g^{-1}}\\
&\le O(1) \sup_{t\in[0,h]} \left\|I - g^{\frac{1}{2}} g_t^{-1} g^{\frac{1}{2}}\right\|_2 \|p\|_{g^{-1}} \le O(1) \sup_{t\in[0,h]} \|x - x_t\|_g \|p\|_{g^{-1}}\\
&\lesssim \left(h\sqrt{n} + h^2\left(n + \sqrt{M_1}\right)\right)\left\|u - u'\right\|_{g^{-1}}. \quad \text{(First item)}
\end{aligned}$$

For S, we can bound it as follows:

$$\begin{aligned}\left\|(I_u - I_{u'})u'\right\|_{g^{-1}} &\le \frac{1}{h}\int_0^h \left\|g(g_t'^{-1} - g_t^{-1})u'\right\|_{g^{-1}} dt \le \sup_{t\in[0,h]} \left\|(g_t'^{-1} - g_t^{-1})u'\right\|_g \\ &\lesssim \sup_{t\in[0,h]} \left\|(g_t'^{-1} - g_t^{-1})u'\right\|_{g_t'} \lesssim \sup_{t\in[0,h]} \left\|x_t - x_t'\right\|_{g_t} \left\|u'\right\|_{g_t'} \\ &\lesssim \phi\sqrt{n+\sqrt{M_1}} \lesssim h\sqrt{n+\sqrt{M_1}} \left\|u - u'\right\|_{g^{-1}}. \quad \text{(Third item)}\end{aligned}$$

Substituting the bounds on F and S into (B.5) with a small constant in h taken, we can conclude that

$$\left\|\Upsilon(u) - \Upsilon(u')\right\|_{g^{-1}} < \frac{1}{3}\left\|u - u'\right\|_{g^{-1}}.$$

Next, we show that the image of Υ is included in V_{dom}. For $u \in V_{\text{dom}}$,

$$\begin{aligned}\|\Upsilon(u) - v\|_{g^{-1}} &= \left\|u - v - \frac{1}{h}g(T_x(u) - \overline{T}_x(v))\right\|_{g^{-1}} \\ &= \left\|u - v - \frac{1}{h}g(T_x(u) - T_x(v)) + \frac{1}{h}g(T_x(v) - \overline{T}_x(v))\right\|_{g^{-1}} \\ &\le \left\|u - v - \frac{1}{h}g(T_x(u) - T_x(v))\right\|_{g^{-1}} + \frac{1}{h}\left\|T_x(v) - \overline{T}_x(v)\right\|_g.\end{aligned}$$

Repeating the proof for the first item[1] (see (B.4)), we can bound the first term by $\frac{1}{2}\|u - v\|_{g^{-1}}$ and thus by $2\sqrt{n}$, due to $u \in V_{\text{dom}}$. By Lemma 64 and ④, we can bound the second term by

$$\frac{1}{h}\left\|T_x(v) - \overline{T}_x(v)\right\|_g \le \frac{2}{h}d_g(T_x(v), \overline{T}_x(v)) \le \frac{2}{h}C_x(x,v)h^2 \le 2\sqrt{n}.$$

Putting them together, we obtain $\|\Upsilon(u) - v\|_{g^{-1}} \le 4\sqrt{n}$.

By the Banach fixed-point theorem, there is a unique fixed point v^* of Υ such that $T_x(v^*) = \overline{T}_x(v)$ and

$$\|\Upsilon(v) - v^*\|_{g^{-1}} < \frac{1}{3}\|v - v^*\|_{g^{-1}}.$$

Moreover, $\|\Upsilon(v) - v^*\|_{g^{-1}} = \left\|v - v^* - \frac{1}{h}g(T_x(v) - z)\right\|_{g^{-1}} \ge \|v - v^*\|_{g^{-1}} - \frac{1}{h}\left\|T_x(v) - \overline{T}_x(v)\right\|_g$. Relating these two inequalities, we obtain

$$\|v - v^*\|_{g^{-1}} \lesssim \frac{1}{h}\left\|T_x(v) - \overline{T}_x(v)\right\|_{g^{-1}}.$$

We now show a one-to-one correspondence between $v \in V_{\text{good}}^x$ and v^*. Let $F_h(x,v) = (x_2, v_2)$ and $F_h(x, v^*) = (z, v')$. By the reversibility of the Hamiltonian trajectories, we have a one-to-one correspondence between v^* and v' in a sense that $F_h(x, v^*) = (z, v')$ and $F_h(z, -v') = (x, -v^*)$. Similarly, we also have a one-to-one correspondence between v and v_2. Thus, it suffices to show a one-to-one correspondence between $v_2 \in T_{x_2}\mathcal{M}$ and $v' \in T_z\mathcal{M}$.

1. For $u \in V_{\text{dom}}$, we might have $\|u\|_{g^{-1}} \ge 128\sqrt{n}$ though, it is still bounded above by $132\sqrt{n}$. The proofs of the second to fifth items can be exactly reproduced for $V_{\text{relaxed}} \stackrel{\text{def}}{=} \{v' \in \mathbb{R}^n : \|v'\|_{g^{-1}} \le 132\sqrt{n}\}$, leading to a similar conclusion like $\|\Upsilon(u) - \Upsilon(u')\|_{g^{-1}} < (\frac{1}{3} + \epsilon)\|u - u'\|_{g^{-1}}$ for a small constant $\epsilon > 0$.

Consider the straight line between z and x_2. We have that $\|z - x_2\|_g = \left\|\overline{T}_x(v) - T_x(v)\right\|_g \leq 2C_x(x,v)h^2 \leq 10^{-9}\min\left(1, \frac{\bar{\ell}_0}{\bar{\ell}_1}\right)$ by ③ and $\bar{\ell}(\mathrm{Ham}_{x_2,h}(-g(x_2)^{-1}v_2)) = \bar{\ell}(\mathrm{Ham}_{x,h}(g(x)^{-1}v)) \leq \bar{\ell}_0/2$ by the symmetry of $\bar{\ell}$. Due to ②, we can apply Lemma 20 to x_2 with an initial velocity $-g(x_2)^{-1}v_2$. Thus, a one-to-one correspondence between v' and v_2 follows, and we also have that $\bar{\ell}_0 \geq \bar{\ell}(\mathrm{Ham}_{z,h}(-g(z)^{-1}v')) = \bar{\ell}(\mathrm{Ham}_{x,h}(g(x)^{-1}v^*)$. ■

B.2.2. ONE-STEP COUPLING

As elaborated in Section B.2, it suffices to prove that for $v \in V^x_{\text{good}}$ the term of $1 - \frac{p^*_x(v^*)}{p^*_x(v)}\frac{\left|D\overline{T}_x(v)\right|}{|DT_x(v^*)|}$ is bounded by a constant smaller than 1.

Lemma 29 *For $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$, let step size h guarantee the sensitivity of a numerical integrator at (x,v), and satisfy*

$$\underbrace{h^2 \leq \frac{10^{-10}}{n + \sqrt{M_1} + M_2}}_{①}, \underbrace{h^2 \leq \frac{10^{-10}}{\bar{R}_1}}_{②}, \underbrace{C_x(x,v)h^2 \leq \frac{1}{10^{10}}\min\left(1, \frac{\bar{\ell}_0}{\bar{\ell}_1}\right)}_{③}, \underbrace{C_x(x,v)h \leq \frac{1}{10^{10}\sqrt{n}}}_{④}.$$

Then $d_{TV}(\overline{\mathcal{P}}_x, \mathcal{P}_x) \leq \frac{1}{10}$.

Proof For given $v \in V^x_{\text{good}}$, Proposition 28-5 (① $\sim$ ④ required) and the order of the numerical integrator ensure that there exists $v^* \in T_x\mathcal{M}$ such that $T_x(v^*) = \overline{T}_x(v)$ and $\|v - v^*\|_{g^{-1}} \lesssim C_x(x,v)h \leq \frac{1}{10^{10}\sqrt{n}}$ by ④. As p^*_x is the probability density function of $\mathcal{N}(0, g(x))$,

$$\left|\log\left(\frac{p^*_x(v^*)}{p^*_x(v)}\right)\right| = \left|\|v^*\|^2_{g^{-1}} - \|v\|^2_{g^{-1}}\right| \leq \|v^* - v\|_{g^{-1}}\left(\|v\|_{g^{-1}} + \|v^*\|_{g^{-1}}\right) \leq \frac{1}{10^5},$$

and thus the ratio of $\frac{p^*_x(v^*)}{p^*_x(v)}$ is bounded below by 0.999. Also, the sensitivity of the numerical integrator yields $\frac{\left|D\overline{T}_x(v)\right|}{|DT_x(v^*)|} \geq 0.998$. Hence, for any $v \in V^x_{\text{good}}$

$$1 - \frac{p^*_x(v^*)}{p^*_x(v)}\frac{\left|D\overline{T}_x(v)\right|}{|DT_x(v^*)|} \leq 0.003,$$

and the claim follows. ■

We finish this section by providing a sufficient condition on the step size for the sensitivity of a numerical integrator, which we find useful later when checking the sensitivity of IMM and LM.

Proposition 30 *Let $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$. Let step size h satisfy $h^2 \leq \frac{1}{10^5\sqrt{n}\bar{R}_1}$ in addition to the step-size conditions in Proposition 28. A numerical integrator $\overline{T}_{x,h}$ is sensitive at (x,v) if $\left|D\overline{T}_x(v)\right| \geq \frac{(1-10^{-6})h^n}{\sqrt{|g(x')||g(x)|}}$ for $x' = \overline{T}_x(v)$.*

Proof By Proposition 28-5, there exists v^* such that $T_x(v^*) = \overline{T}_x(v)$ and $\bar{\ell}(\mathrm{Ham}_{x,t}(g(x)^{-1}v^*)) \leq \bar{\ell}_0$. Let us estimate $|DT_x(v^*)|$. Recall that $\mathrm{Ham}_{x,h}$ is the Hamiltonian map from $T_x\mathcal{M}$ to $\mathcal{M}$, where

both spaces are endowed with the local metric g. Even though T_x has the same domain and range, these spaces are endowed with the Euclidean metric. Therefore, we can relate T_x to $\mathrm{Ham}_{x,h}$ by

$$T_x(v^*) = \left(\mathrm{id}_{\mathcal{M}\to\mathbb{R}^n} \circ \mathrm{Ham}_{x,h} \circ \mathrm{id}_{\mathbb{R}^n\to T_x\mathcal{M}}\right)(g(x)^{-1}v^*),$$

where $\mathrm{id}_{\mathcal{M}\to\mathbb{R}^n}$ is the embedding with transition of metric from $g(x)$ to the Euclidean, and $\mathrm{id}_{\mathbb{R}^n\to T_x\mathcal{M}}$ is the embedding with transition of metric from the Euclidean to $g(x)$. Note that we have to normalize v^* by $g(x)^{-1}$ before $\mathrm{Ham}_{x,h}$ takes it as input. Using this formula and the chain rule,

$$\begin{aligned}
\left|DT_x(v^*)\right| &= \left|D\mathrm{id}_{\mathcal{M}\to\mathbb{R}^n}(x')\right| \left|D\mathrm{Ham}_{x,h}(g(x)^{-1}v^*)\right| \left|D\mathrm{id}_{\mathbb{R}^n\to T_x\mathcal{M}}(g(x)^{-1}v^*)\right| \\
&\le \left|g(x')\right|^{-\frac{1}{2}} \cdot h^n \left(1 + \frac{2}{1000}\right) \cdot |g(x)|^{-1} \cdot |g(x)|^{\frac{1}{2}} \\
&= \frac{h^n}{\sqrt{|g(x')|\,|g(x)|}} \left(1 + \frac{2}{1000}\right),
\end{aligned}$$

where we used Corollary 18 for $\bar{\ell}$ in the second line. Hence, if $\left|D\overline{T}_x(v)\right| \ge \frac{(1-10^{-6})h^n}{\sqrt{|g(x')||g(x)|}}$, then

$$\frac{\left|D\overline{T}_x(v)\right|}{\left|DT_x(v^*)\right|} \ge \frac{1-10^{-6}}{1+0.002} \ge 0.998.$$

■

B.3. Bound on rejection probability

Lastly, we bound the rejection probability $d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x)$.

Lemma 31 *For $x, x' \in \mathcal{M}$, let $g = g(x)$ and $g' = g(x')$. If for some $0 < \delta_x < 1$ and $0 < \delta_v$ we have $\|x - x'\|_g \le \delta_x$ and $\|v - v'\|_{g^{-1}} \le \delta_v$, then*

$$\left|-H(x',v') + H(x,v)\right| \le \Theta\left(\left|f(x) - f(x')\right| + \left(\delta_x \|v\|_{g^{-1}}^2 + \delta_v^2 + \delta_v \|v\|_{g^{-1}}\right) + n\delta_x\right),$$

where the Hamiltonian is $H(x,v) = f(x) + \frac{1}{2}v^\top g(x)^{-1}v + \frac{1}{2}\log\det g(x)$.

Proof We consider each term separately. For the second term,

$$\left|\frac{1}{2}v^\top g^{-1}v - \frac{1}{2}v'^\top g'^{-1}v\right| \le \frac{1}{2}\underbrace{\left|v^\top g^{-1}v - v^\top g'^{-1}v\right|}_{F} + \frac{1}{2}\underbrace{\left|v^\top g'^{-1}v - v'^\top g'^{-1}v'\right|}_{S}.$$

For F, we have $F \le O(\delta_x)\|v\|_{g^{-1}}^2$ by Lemma 26-3. For S, it follows that

$$\begin{aligned}
S &= \left|\|v\|_{g'^{-1}}^2 - \|v'\|_{g'^{-1}}^2\right| \le \|v - v'\|_{g'^{-1}} \|v + v'\|_{g'^{-1}} \\
&\le O(1)\|v - v'\|_{g^{-1}} \left(\|v\|_{g^{-1}} + \|v'\|_{g^{-1}}\right) \\
&\le O(1)\delta_v(\delta_v + \|v\|_{g^{-1}}).
\end{aligned}$$

Therefore, the second term is bounded by $O(1)\left(\delta_x \|v\|_{g^{-1}}^2 + \delta_v^2 + \delta_v \|v\|_{g^{-1}}\right)$.

For the third term,

$$\left|\frac{1}{2}\left(\log\det g(x) - \log\det g(x')\right)\right| = \frac{1}{2}\left|\log\det g'^{-\frac{1}{2}} g g'^{-\frac{1}{2}}\right| \le O(n\delta_x),$$

where the inequality follows from Lemma 26 and the fact that the determinant is the product of eigenvalues. ■

Since the ideal RHMC preserves the Hamiltonian along its Hamiltonian curve, $H(x,v) = H(x_h, v_h)$. Hence, we can obtain a lower bound on the acceptance probability by computing either

$$\min\left(1, \frac{e^{-H(\bar{x}_h,\bar{v}_h)}}{e^{-H(x,v)}}\right) \text{ or } \min\left(1, \frac{e^{-H(\bar{x}_h,\bar{v}_h)}}{e^{-H(x_h,v_h)}}\right).$$

Lemma 32 *Let (x_h, v_h) and $(\bar{x}, \bar{v})$ be the points obtained by the ideal RHMC and discretized RHMC with a sensitive numerical integrator starting at $x \in \mathcal{M}_\rho$ with $v \in V_{good}^x$. If the step size h satisfies*

$$h^2 \le \frac{10^{-10}}{n + \sqrt{M_1} + M_2},\ h^2 C_x(x,v) \le \frac{10^{-10}}{n + \sqrt{M_1} + \sqrt{M_1^*}},\ h^2 C_v(x,v) \le \frac{10^{-10}}{\sqrt{n + \sqrt{M_1}}},$$

then the rejection probability of the Metropolis filter is bounded by 10^{-3}.

Proof We use the first condition on the step size to obtain $\|v_h\|_{g^{-1}} = O\left(h\left(n + \sqrt{M_1}\right)\right) = O\left(\sqrt{n + \sqrt{M_1}}\right)$ by Proposition 28-1. Then the claims follows from

$$\begin{aligned} d_{\mathrm{TV}}(\overline{\mathcal{P}}'_x, \overline{\mathcal{P}}_x) &\le 10^5\left(\delta_x\sqrt{M_1^*} + \delta_x\left(n + \|v_h\|_{g^{-1}}^2\right) + \delta_v\left(\delta_v + \|v_h\|_{g^{-1}}\right)\right) \\ &\le 10^6 h^2\left(C_x(x,v)\left(n + \sqrt{M_1} + \sqrt{M_1^*}\right) + C_v(x,v)\left(C_v(x,v)h^2 + \sqrt{n + \sqrt{M_1}}\right)\right) \\ &\le 10^{-4} + 10^{-20} + 10^{-4} \le 10^{-3}, \end{aligned}$$

where we used the second and third step-size conditions in the last inequality. ■

Putting three main parts together, we obtain the result on the mixing rate of RHMC discretized by a sensitive numerical integrator.

Proof of Theorem 24 By Lemma 14, 29 and 32, we have $d_{\mathrm{TV}}(\overline{\mathcal{P}}_x, \overline{\mathcal{P}}_y) \le \frac{9}{10}$ if $d_\phi(x,y) \le h$ for $x, y \in \mathcal{M}_\rho$. Then the claim follows by reproducing the proof of Proposition 15. ■

Appendix C. Numerical integrators

We examine two numerical integrators commonly used in practice, the implicit midpoint method (IMM) in Section C.1 and the generalized Leapfrog method (LM) in Section C.2. To this end, we bound C_x and C_v, the second-orderness parameters, and then find a condition on step size for the sensitivity. We note that these integrators are symplectic (so measure-preserving) and time-reversible (see Hairer et al. (2006)).

C.1. Implicit midpoint method

For an initial condition (x, v) and step size h, the implicit midpoint method attempts to find the solution (x', v') for the following implicit equation:

$$x' = x + h\frac{\partial H}{\partial v}\left(\frac{x+x'}{2}, \frac{v+v'}{2}\right),\ v' = v - h\frac{\partial H}{\partial x}\left(\frac{x+x'}{2}, \frac{v+v'}{2}\right).$$

In general, these implicit equations require several iterations so that an initial guess for this equation converges to the fixed point (x', v').

In this section, we consider a variant of IMM in Algorithm 2 instead. It has computational benefits over the original IMM, since iterations for finding the fixed point of the integrator run with a simpler Hamiltonian $H_2(x, v) = \frac{1}{2}v^\top g(x)^{-1}v$ instead of $H = H_1 + H_2$. We then prove that if for $x \in \mathcal{M}_\rho$ and $v \in V^x_{\text{good}}$ step size h satisfies $h^2\left(n + \sqrt{M_1}\right) \le 10^{-10}$, then IMM is second-order with $C_x(x, v) = O\left(n + \sqrt{M_1}\right)$ and $C_v(x, v) = O\left(\sqrt{n + \sqrt{M_1}}\left(n + \sqrt{M_1} + M_2^*\right)\right)$. Moreover, if the step size h satisfies $h^2 \le \min\left(\frac{10^{-10}}{\left(n+\sqrt{M_1}\right)^2}, \frac{10^{-5}}{\sqrt{n}\bar{R}_1}\right)$ in addition to the step-size conditions in Proposition 28, then IMM is sensitive at $x \in \mathcal{M}_\rho$ and $v \in V^x_{\text{good}}$.

Algorithm 2: `Implicit Midpoint Method`

Input: Initial point x, velocity v, step size h

`// Step 1:` Solve $\frac{dx}{dt} = \frac{\partial H_1(x,v)}{\partial v}, \frac{dv}{dt} = -\frac{\partial H_1(x,v)}{\partial x}$

Set $x_{\frac{1}{3}} \leftarrow x$ and $v_{\frac{1}{3}} \leftarrow v - \frac{h}{2}\frac{\partial H_1(x,v)}{\partial x}$.

`// Step 2:` Solve $\frac{dx}{dt} = \frac{\partial H_2(x,v)}{\partial v}, \frac{dv}{dt} = -\frac{\partial H_2(x,v)}{\partial x}$ `(Implicit)`

Find $(x_{\frac{2}{3}}, v_{\frac{2}{3}})$ such that

$$x_{\frac{2}{3}} = x_{\frac{1}{3}} + h\frac{\partial H_2}{\partial v}\left(\frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}, \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}\right),$$
$$v_{\frac{2}{3}} = v_{\frac{1}{3}} - h\frac{\partial H_2}{\partial x}\left(\frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}, \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}\right).$$

`// Step 3:` Solve $\frac{dx}{dt} = \frac{\partial H_1(x,v)}{\partial v}, \frac{dv}{dt} = -\frac{\partial H_1(x,v)}{\partial x}$

Set $x_1 \leftarrow x_{\frac{2}{3}}$ and $v_1 \leftarrow v_{\frac{2}{3}} - \frac{h}{2}\frac{\partial H_1}{\partial x}\left(x_{\frac{2}{3}}, v_{\frac{2}{3}}\right)$.

Output: x_1, v_1

C.1.1. SECOND-ORDER

Lemma 33 *For $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$, let $g = g(x)$ and h step size of IMM with $h^2\left(n + \sqrt{M_1}\right) \le 10^{-10}$. Let $(\overline{x}, \overline{v})$ be the point obtained from RHMC discretized by IMM with the step size h and initial condition (x, v).*

1. $\|x - \overline{x}\|_g = O\left(h\sqrt{n} + h^2\left(n + \sqrt{M_1}\right)\right)$ *and* $\|v - \overline{v}\|_{g^{-1}} = O\left(h\left(\|\nabla f(\overline{x})\|_{g^{-1}} + n + \sqrt{M_1}\right)\right)$.

2. $C_x(x,v) = O\left(n + \sqrt{M_1}\right)$.

3. $C_v(x,v) = O\left(\sqrt{n+\sqrt{M_1}}\left(n + \sqrt{M_1} + M_2^*\right)\right)$.

Proof of 1. Let $\bar{x} = x_{\frac{2}{3}} = \overline{T}_x(v)$ and $v_{\frac{1}{3}}, v_{\frac{2}{3}}, \bar{v}$ be the velocity points obtained when starting with (x,v). Then x_{mid} and v_{mid} satisfy

$$\begin{aligned} x_{\frac{2}{3}} &= x_{\frac{1}{3}} + h g_{\text{mid}}^{-1} v_{\text{mid}}, \\ v_{\frac{2}{3}} &= v_{\frac{1}{3}} + \frac{h}{2} Dg_{\text{mid}}\left[g_{\text{mid}}^{-1} v_{\text{mid}}, g_{\text{mid}}^{-1} v_{\text{mid}}\right], \end{aligned} \tag{C.1}$$

where $x_{\text{mid}} = \frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}$, $v_{\text{mid}} = \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}$ and $g_{\text{mid}} = g\left(x_{\text{mid}}\right)$. Since $\|\bar{x} - x\|_{g_{\text{mid}}} = \left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} \to 0$ as $h \to 0$, we can take $h_0 > 0$ such that $\left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} \le \frac{1}{1000}$ for $h \le h_0$ with the equality held at $h = h_0$, or $\left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} \le \frac{1}{1000}$ for any $h > 0$.

We start with the former. By adding $v_{\frac{1}{3}}$ to the second line of (C.1) and dividing by 2, we have

$$h_0 v_{\text{mid}} = h_0 v_{\frac{1}{3}} + \frac{h_0^2}{4} Dg_{\text{mid}}\left[g_{\text{mid}}^{-1} v_{\text{mid}}, g_{\text{mid}}^{-1} v_{\text{mid}}\right]. \tag{C.2}$$

As $\left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} = h_0 \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}$ from the first line of (C.1), taking the g_{mid}^{-1}-norm on both sides of (C.2) and using Lemma 27 yield

$$\begin{aligned} \frac{1}{1000} &= h_0 \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}} \\ &\le h_0 \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} + \frac{h_0^2}{2} \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}^2 \le h_0 \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} + \frac{1}{2000}, \end{aligned}$$

and we obtain $\frac{1}{2000} \le h_0 \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}}$. Recall that $\|\bar{x} - x\|_{g_{\text{mid}}} \le 1/1000$ for $h \le h_0$, so we can swap the local norms between x_{mid} and x due to Lemma 26, losing a multiplicative constant like 1.001. Using Lemma 27 on the first step of the numerical integrator,

$$\begin{aligned} \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} &\le 1.001 \left\|v_{\frac{1}{3}}\right\|_{g^{-1}} \le 1.001 \left(\|v\|_{g^{-1}} + h_0\left(n + \sqrt{M_1}\right)\right) \\ &\le 200\sqrt{n} + 2h_0\left(n + \sqrt{M_1}\right). \end{aligned} \tag{C.3}$$

Due to $1/2000 \le h_0 \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}}$, it follows that

$$\frac{1}{2000} \le h_0 \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} \le 200\sqrt{n} h_0 + 2h_0^2\left(n + \sqrt{M_1}\right),$$

and solving this for h_0 we have $h_0 \ge \frac{1}{10^4\sqrt{n+\sqrt{M_1}}}$. For the case of $\left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} \le \frac{1}{1000}$ for any $h > 0$, we can simply think of h_0 as ∞.

Now for $h \le h_0$, we can obtain from (C.1)

$$\begin{aligned} \left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} &\le h \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}, \\ \left\|v_{\frac{2}{3}} - v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} &\le h \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}^2. \end{aligned} \tag{C.4}$$

Using this and $h \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}} = \left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} \le \frac{1}{1000}$, we can bound $\left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}$ by

$$\begin{aligned} \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}} &= \left\|v_{\frac{1}{3}} + \frac{v_{\frac{2}{3}} - v_{\frac{1}{3}}}{2}\right\|_{g_{\text{mid}}^{-1}} \le \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} + \frac{1}{2} h \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}^2 \\ &\le \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} + \frac{1}{2000} \left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}, \end{aligned}$$

and thus

$$\left\|v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}} \le \frac{2000}{1999} \left\|v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} \underset{\text{(C.3)}}{\le} 200\sqrt{n} + 2h\left(n + \sqrt{M_1}\right). \tag{C.5}$$

Putting this back to (C.4) for step size $h \le \frac{1}{10^5\sqrt{n+\sqrt{M_1}}}$, we have

$$\begin{aligned} \left\|x_{\frac{2}{3}} - x\right\|_{g_{\text{mid}}} &\le 200h\sqrt{n} + 2h^2\left(n + \sqrt{M_1}\right), \\ \left\|v_{\frac{2}{3}} - v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}} &\le 125h\left(n + \sqrt{M_1}\right). \end{aligned}$$

Hence by substituting the step size into above and switching local norms properly, we have $\left\|x_{\frac{2}{3}} - x\right\|_g \le 10^{-8}$.

By applying Lemma 27 to $\overline{v} = v_{\frac{2}{3}} - \frac{h}{2}\frac{\partial H_1}{\partial x}\left(x_{\frac{2}{3}}, v_{\frac{2}{3}}\right)$ in the third step, we also have

$$\begin{aligned} \left\|\bar{v} - v\right\|_{g^{-1}} &\le \left\|\bar{v} - v_{\frac{2}{3}}\right\|_{g^{-1}} + \left\|v_{\frac{2}{3}} - v_{\frac{1}{3}}\right\|_{g^{-1}} + \left\|v_{\frac{1}{3}} - v\right\|_{g^{-1}} \\ &\le 1.001h\left(\left\|\nabla f(x_{\frac{2}{3}})\right\|_{g^{-1}} + n + 125\left(n + \sqrt{M_1}\right) + \left(n + \sqrt{M_1}\right)\right) \\ &\le 200h\left(\left\|\nabla f(x_{\frac{2}{3}})\right\|_{g^{-1}} + n + \sqrt{M_1}\right). \end{aligned}$$

In conclusion,

$$\begin{aligned} \left\|\bar{x} - x\right\|_g &\le 200h\sqrt{n} + 3h^2\left(n + \sqrt{M_1}\right), \\ \left\|\bar{v} - v\right\|_{g^{-1}} &\le 200h\left(\left\|\nabla f(\bar{x})\right\|_{g^{-1}} + n + \sqrt{M_1}\right). \end{aligned}$$

Proof of 2. For $t \in [0, h]$, let (x_t, v_t) be the Hamiltonian curve of the ideal RHMC at time t starting from (x, v). Recall that for $g_t = g(x_t)$

$$\begin{aligned} T_x(v) &= x + \int_0^h \frac{\partial H}{\partial v}(x_t, v_t)dt = x + \int_0^h g_t^{-1} v_t dt, \\ \overline{T}_x(v) &= \overline{x} = x + h g_{\text{mid}}^{-1} v_{\text{mid}}. \end{aligned}$$

Thus,

$$\begin{aligned}\left\|T_x(v)-\overline{T}_x(v)\right\|_{g_{\text{mid}}} &= \left\|\left(x+\int_0^h g_t^{-1}v_t dt\right)-\left(x+hg_{\text{mid}}^{-1}v_{\text{mid}}\right)\right\|_{g_{\text{mid}}}\\ &= \left\|\int_0^h \left(g_t^{-1}v_t-g_{\text{mid}}^{-1}v_{\text{mid}}\right)dt\right\|_{g_{\text{mid}}}\\ &\le h\max_{t\in[0,h]}\left\|g_t^{-1}v_t-g_{\text{mid}}^{-1}v_{\text{mid}}\right\|_{g_{\text{mid}}}\end{aligned}$$

By Lemma 26-11,

$$\begin{aligned}\left\|g_t^{-1}v_t-g_{\text{mid}}^{-1}v_{\text{mid}}\right\|_{g_{\text{mid}}} &\lesssim \left\|v_t-v_{\text{mid}}\right\|_{g_{\text{mid}}^{-1}}+\left\|x_t-x_{\text{mid}}\right\|_{g_{\text{mid}}}\left\|v_t\right\|_{g_{\text{mid}}^{-1}}\\ &\le \frac{1}{2}\left(\left\|v_t-v_{\frac{1}{3}}\right\|_{g_{\text{mid}}^{-1}}+\left\|v_t-v_{\frac{2}{3}}\right\|_{g_{\text{mid}}^{-1}}\right)\\ &\quad+\frac{1}{2}\left(\left\|x_t-x\right\|_{g_{\text{mid}}}+\left\|x_t-\overline{x}\right\|_{g_{\text{mid}}}\right)\left\|v_t\right\|_{g_{\text{mid}}^{-1}}\\ &\lesssim \left(\left\|v-v_t\right\|_{g^{-1}}+\left\|v-v_{\frac{1}{3}}\right\|_{g^{-1}}+\left\|v_t-v\right\|_{g^{-1}}+\left\|v-v_{\frac{2}{3}}\right\|_{g^{-1}}\right)\\ &\quad+\left(\left\|x_t-x\right\|_g+\left\|x_t-x\right\|_g+\left\|\overline{x}-x\right\|_g\right)\left(\left\|v_t-v\right\|_{g^{-1}}+\left\|v\right\|_{g^{-1}}\right)\\ &\lesssim \left(\left\|v-v_t\right\|_{g^{-1}}+\left\|v-v_{\frac{1}{3}}\right\|_{g^{-1}}+\left\|v_{\frac{1}{3}}-v_{\frac{2}{3}}\right\|_{g^{-1}}\right)\\ &\quad+\left(\left\|x_t-x\right\|_g+\left\|\overline{x}-x\right\|_g\right)\left(\left\|v_t-v\right\|_{g^{-1}}+\left\|v\right\|_{g^{-1}}\right).\end{aligned}$$

Using our bounds on $\left\|\bar{x}-x\right\|_g, \left\|x_t-x\right\|_g$ and $\left\|v\right\|_{g^{-1}}, \left\|v_t-v\right\|_{g^{-1}}, \left\|v-v_{\frac{1}{3}}\right\|_{g^{-1}}, \left\|v_{\frac{1}{3}}-v_{\frac{2}{3}}\right\|_{g^{-1}}$, we conclude that $\max_{t\in[0,h]}\left\|g_t^{-1}v_t-g_{\text{mid}}^{-1}v_{\text{mid}}\right\|_{g_{\text{mid}}} \le 10^4 h\left(n+\sqrt{M_1}\right)$, and thus

$$\left\|T_x(v)-\overline{T}_x(v)\right\|_g \le 10^4 h^2\left(n+\sqrt{M_1}\right).$$

Proof of 3. From the algorithm,

$$\begin{aligned}v_h &= v-\int_0^h \frac{\partial H}{\partial x}(x_t,v_t)dt\\ &= v-\int_0^h \frac{\partial H_1}{\partial x}(x_t,v_t)dt-\int_0^h \frac{\partial H_2}{\partial x}(x_t,v_t)dt,\\ \bar{v} &= v_{\frac{2}{3}}-\frac{h}{2}\frac{\partial H_1}{\partial x}(x_{\frac{2}{3}},v_{\frac{2}{3}}) = v_{\frac{1}{3}}-h\frac{\partial H_2}{\partial x}(x_{\text{mid}},v_{\text{mid}})-\frac{h}{2}\frac{\partial H_1}{\partial x}(x_{\frac{2}{3}},v_{\frac{2}{3}})\\ &= v-\frac{h}{2}\left(\frac{\partial H_1}{\partial x}(x,v)+\frac{\partial H_1}{\partial x}(x_{\frac{2}{3}},v_{\frac{2}{3}})\right)-h\frac{\partial H_2}{\partial x}(x_{\text{mid}},v_{\text{mid}}).\end{aligned}$$

Thus,

$$
\begin{aligned}
&\|v_h - \bar{v}\|_{g_{\text{mid}}^{-1}} \qquad \text{(C.6)}\\
&= \left\|\int_0^h \left(\frac{\partial H_1}{\partial x}(x_t, v_t) - \frac{1}{2}\left(\frac{\partial H_1}{\partial x}(x, v) + \frac{\partial H_1}{\partial x}(x_{\frac{2}{3}}, v_{\frac{2}{3}})\right)\right) dt + \int_0^h \left(\frac{\partial H_2}{\partial x}(x_t, v_t) - \frac{\partial H_2}{\partial x}(x_{\text{mid}}, v_{\text{mid}})\right) dt\right\|_{g_{\text{mid}}^{-1}}\\
&\leq h \underbrace{\max_{t\in[0,h]} \left\|\frac{\partial H_1}{\partial x}(x_t, v_t) - \frac{1}{2}\left(\frac{\partial H_1}{\partial x}(x, v) + \frac{\partial H_1}{\partial x}(x_{\frac{2}{3}}, v_{\frac{2}{3}})\right)\right\|_{g_{\text{mid}}^{-1}}}_{F}\\
&\quad + h \underbrace{\max_{t\in[0,h]} \left\|\frac{\partial H_2}{\partial x}(x_t, v_t) - \frac{\partial H_2}{\partial x}(x_{\text{mid}}, v_{\text{mid}})\right\|_{g_{\text{mid}}^{-1}}}_{S}.
\end{aligned}
$$

We separately bound F and S. For F, by following the proof of Proposition 26-12, we have

$$
F \lesssim n\left(\|v_t - v\|_{g^{-1}} + \left\|v_t - v_{\frac{2}{3}}\right\|_{g^{-1}}\right) + M_2^*\left(\|x_h - x\|_g + \|x_h - \bar{x}\|_g\right).
$$

Using our bounds on $\|x_h - \bar{x}\|_g$, $\|x_h - x\|_g$ and $\|v_t - v\|_{g^{-1}}$, $\left\|v - v_{\frac{2}{3}}\right\|_{g^{-1}}$, we obtain

$$
F \lesssim h\left(n + \sqrt{M_1}\right)^{3/2} + h\sqrt{n + \sqrt{M_1}}M_2^*.
$$

We remark that the smoothness of f guarantees that M_2^* is bounded by some constant for all sufficiently small h.

Similarly for S, we have that for $\delta_v = \|v_t - v_{\text{mid}}\|_{g_{\text{mid}}^{-1}}$ and $\delta_x = \|x_t - x_{\text{mid}}\|_{g_{\text{mid}}}$

$$
S \lesssim \max_{t\in[0,h]} \left(\delta_v + \delta_x \|v_{\text{mid}}\|_{g_{\text{mid}}^{-1}}\right)\left(\|v_{\text{mid}}\|_{g_{\text{mid}}^{-1}} + \|v_t\|_{g_{\text{mid}}^{-1}}\right).
$$

Using our bounds on $\|\bar{x} - x\|_g$, $\|x_t - x\|_g$ and $\|v\|_{g^{-1}}$, $\|v_t - v\|_{g^{-1}}$, $\left\|v - v_{\frac{1}{3}}\right\|_{g^{-1}}$, $\left\|v_{\frac{1}{3}} - v_{\frac{2}{3}}\right\|_{g^{-1}}$, it follows that

$$
S \lesssim h\left(n + \sqrt{M_1}\right)^{3/2}.
$$

Substituting the bounds on F and S into (C.6) we can conclude that

$$
\|v_h - \bar{v}\|_{g^{-1}} \leq 10^{10}h^2\sqrt{n + \sqrt{M_1}}\left(n + \sqrt{M_1} + M_2^*\right).
$$

C.1.2. SENSITIVITY

We use Proposition 30 to show that IMM is sensitive. Here we assume that $\log\det g(x)$ is convex in $\mathcal{M}$, which is the case for the logarithmic barriers of polytopes.

Lemma 34 *For $x \in \mathcal{M}_\rho$ and $v \in V_{good}^x$, if $\log\det g(x)$ is convex in x, then IMM is sensitive at (x, v) for step size h satisfying $h^2 \leq \min\left(\frac{10^{-10}}{\left(n+\sqrt{M_1}\right)^2}, \frac{10^{-5}}{\sqrt{n}\bar{R}_1}\right)$ and the step-size conditions in Proposition 28.*

Proof Let $\overline{F}(x,v) = (\bar{x}, \bar{v})$ and $\overline{T}_x(v) = \bar{x}$. We lower bound $\left|D\overline{T}_x(v)\right|$. Recall that one iteration of IMM consists of three steps with input (x,v) and output (x_1, v_1), as described in the following diagram:

$$(x,v) \xrightarrow{X} \left(x_{\frac{1}{3}}, v_{\frac{1}{3}}\right) \xrightarrow{Y} \left(x_{\frac{2}{3}}, v_{\frac{2}{3}}\right) \xrightarrow{Z} (x_1, v_1),$$

where each of the maps X, Y and Z is defined by

$$\begin{aligned}
X(x,v) &= \left(x, v - \frac{h}{2}\frac{\partial H_1(x,v)}{\partial x}\right), \\
Y\left(x_{\frac{1}{3}}, v_{\frac{1}{3}}\right) &= \left(x_{\frac{1}{3}} + h\frac{\partial H_2(x_{\text{mid}}, v_{\text{mid}})}{\partial v}, v_{\frac{1}{3}} - h\frac{\partial H_2(x_{\text{mid}}, v_{\text{mid}})}{\partial x}\right), \\
Z\left(x_{\frac{2}{3}}, v_{\frac{2}{3}}\right) &= \left(x_{\frac{2}{3}}, v_{\frac{2}{3}} - \frac{h}{2}\frac{\partial H_1(x_{\frac{2}{3}}, v_{\frac{2}{3}})}{\partial x}\right),
\end{aligned}$$

for $x_{\text{mid}} = \frac{x_{\frac{1}{3}} + x_{\frac{2}{3}}}{2}$ and $v_{\text{mid}} = \frac{v_{\frac{1}{3}} + v_{\frac{2}{3}}}{2}$. Due to $\overline{T}_x(v) = \pi_x \circ (Z \circ Y \circ X)(x,v)$, it follows that $D\overline{T}_x(v)$ is the upper-right $n \times n$ submatrix of $D(Z \circ Y \circ X)(x,v)$. From direct computation, we have

$$\begin{aligned}
DX(x,v) &= \begin{bmatrix} I & 0 \\ * & I \end{bmatrix}, \\
DY(x_{\frac{1}{3}}, v_{\frac{1}{3}}) &= \begin{bmatrix} P & Q \\ R & S \end{bmatrix}, \\
DZ(x_{\frac{2}{3}}, v_{\frac{2}{3}}) &= \begin{bmatrix} I & 0 \\ * & I \end{bmatrix},
\end{aligned}$$

and due to $D(Z \circ Y \circ X) = DZ \cdot DY \cdot DX$ we have $D\overline{T}_x(v) = Q$. Thus, it suffices to focus on the second step only (i.e., the map Y).

Now let us represent the map Y in a compact way. With two symbols

$$r = \begin{bmatrix} x \\ v \end{bmatrix} \in \mathbb{R}^{2n} \text{ and } J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix} \in \mathbb{R}^{2n \times 2n},$$

the second step can be rewritten as

$$r_{\frac{2}{3}} = r_{\frac{1}{3}} + hJ\nabla_{(x,v)} H_2(r_{\text{mid}}),$$

where $r_* = \begin{bmatrix} x_* \\ v_* \end{bmatrix}$ for $* \in \left\{\frac{1}{3}, \frac{2}{3}, \text{mid}\right\}$ and $H_2(x,v) = \frac{1}{2} v^\top g(x)^{-1} v$. Differentiating both sides by $r_{\frac{1}{3}}$,

$$\frac{\partial r_{\frac{2}{3}}}{\partial r_{\frac{1}{3}}} = I_{2n} + hJ\nabla^2 H_2(r_{\text{mid}}) \left(\frac{1}{2} I_{2n} + \frac{1}{2}\frac{\partial r_{\frac{2}{3}}}{\partial r_{\frac{1}{3}}}\right).$$

As $DY\left(r_{\frac{1}{3}}\right) = \frac{\partial r_{\frac{2}{3}}}{\partial r_{\frac{1}{3}}}$, we have that

$$\left(I_{2n} - \frac{h}{2} J\nabla^2 H_2(r_{\text{mid}})\right) DY\left(r_{\frac{1}{3}}\right) = I_{2n} + \frac{h}{2} J\nabla^2 H_2(r_{\text{mid}}).$$

For $G(x) \stackrel{\text{def}}{=} \begin{bmatrix} g(x)^{\frac{1}{2}} & 0 \\ 0 & g(x)^{-\frac{1}{2}} \end{bmatrix}$, we have

$$\begin{aligned} &G(x_{\mathrm{mid}}) \left(I_{2n} - \frac{h}{2} J \nabla^2 H_2(r_{\mathrm{mid}}) \right) G(x_{\mathrm{mid}})^{-1} G(x_{\mathrm{mid}}) DY \left(r_{\frac{1}{3}} \right) G(x_{\mathrm{mid}})^{-1} \\ &\quad = G(x_{\mathrm{mid}}) \left(I_{2n} + \frac{h}{2} J \nabla^2 H_2(r_{\mathrm{mid}}) \right) G(x_{\mathrm{mid}})^{-1} \end{aligned}$$

and

$$\begin{aligned} &\left(I_{2n} - \frac{h}{2} G(x_{\mathrm{mid}}) J \nabla^2 H_2(r_{\mathrm{mid}}) G(x_{\mathrm{mid}})^{-1} \right) G(x_{\mathrm{mid}}) DY \left(r_{\frac{1}{3}} \right) G(x_{\mathrm{mid}})^{-1} \\ &= I_{2n} + \frac{h}{2} G(x_{\mathrm{mid}}) J \nabla^2 H_2(r_{\mathrm{mid}}) G(x_{\mathrm{mid}})^{-1}. \end{aligned} \tag{C.7}$$

Let us look into the term $B \stackrel{\text{def}}{=} G(x_{\mathrm{mid}}) J \nabla^2 H_2(r_{\mathrm{mid}}) G(x_{\mathrm{mid}})^{-1}$. By direct computation, for block matrices B_1, B_2, B_3, B_4 of size $n \times n$ we have

$$\begin{aligned} B &\stackrel{\text{def}}{=} \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} \\ &= \begin{bmatrix} g(x_{\mathrm{mid}})^{\frac{1}{2}} & \\ & g(x_{\mathrm{mid}})^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \frac{\partial^2 H_2}{\partial v \partial x}(r_{\mathrm{mid}}) & \frac{\partial^2 H_2}{\partial v^2}(r_{\mathrm{mid}}) \\ -\frac{\partial^2 H_2}{\partial x^2}(r_{\mathrm{mid}}) & -\left(\frac{\partial^2 H_2}{\partial x \partial v}(r_{\mathrm{mid}}) \right)^\top \end{bmatrix} \begin{bmatrix} g(x_{\mathrm{mid}})^{-\frac{1}{2}} & \\ & g(x_{\mathrm{mid}})^{\frac{1}{2}} \end{bmatrix} \end{aligned}$$

and thus

$$\begin{aligned} B_1 &= g(x_{\mathrm{mid}})^{-\frac{1}{2}} Dg(x_{\mathrm{mid}}) \left[g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \right] g(x_{\mathrm{mid}})^{-\frac{1}{2}}, \\ B_2 &= I_n, \\ B_3 &= g(x_{\mathrm{mid}})^{-\frac{1}{2}} \Big(- v_{\mathrm{mid}}^\top g(x_{\mathrm{mid}})^{-1} Dg(x_{\mathrm{mid}}) g(x_{\mathrm{mid}})^{-1} Dg(x_{\mathrm{mid}}) g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \\ &\quad + \frac{1}{2} v_{\mathrm{mid}}^\top g(x_{\mathrm{mid}})^{-1} D^2 g(x_{\mathrm{mid}}) g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \Big) g(x_{\mathrm{mid}})^{-\frac{1}{2}} \\ &= -B_1^2 + \frac{1}{2} g(x_{\mathrm{mid}})^{-\frac{1}{2}} D^2 g(x_{\mathrm{mid}}) \left[g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \right] g(x_{\mathrm{mid}})^{-\frac{1}{2}}, \\ B_4 &= -B_1^\top. \end{aligned} \tag{C.8}$$

Now we bound the operator norm of B_i for each $i \in [4]$ as follows.

$$\begin{aligned} \|B_1\| &= \|B_4\| \\ &= \max_{p,q : \|p\|_2, \|q\|_2 \le 1} p^\top g(x_{\mathrm{mid}})^{-\frac{1}{2}} Dg(x_{\mathrm{mid}}) \left[g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \right] g(x_{\mathrm{mid}})^{-\frac{1}{2}} q \\ &= \max_{p,q} Dg(y_{\mathrm{mid}}) \left[g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-\frac{1}{2}} p, g(x_{\mathrm{mid}})^{-\frac{1}{2}} q \right] \\ &\le 2 \max_{p,q} \left\| g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \right\|_{g(x_{\mathrm{mid}})} \left\| g(x_{\mathrm{mid}})^{-\frac{1}{2}} p \right\|_{g(x_{\mathrm{mid}})} \left\| g(x_{\mathrm{mid}})^{-\frac{1}{2}} q \right\|_{g(x_{\mathrm{mid}})} \\ &= 2 \max_{p,q} \left\| g(x_{\mathrm{mid}})^{-1} v_{\mathrm{mid}} \right\|_{g(x_{\mathrm{mid}})} \|p\|_2 \|q\|_2 \le 2 \|v_{\mathrm{mid}}\|_{g_{\mathrm{mid}}^{-1}} \\ &\le O\left(\sqrt{n + \sqrt{M_1}} \right), \end{aligned}$$

where we used (C.5) guaranteed by the condition of $h^2\left(n+\sqrt{M_1}\right) \le 10^{-10}$ (① in Proposition 28). For B_2 and B_3, we have

$$\begin{aligned}
&\|B_2\| = 1,\\
&\|B_3\| \le \|B_1\|^2 + \frac{1}{2}\max_{p,q:\|p\|_2,\|q\|_2\le 1} D^2 g(x_{\mathrm{mid}})\left[g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-\frac{1}{2}}p, g(x_{\mathrm{mid}})^{-\frac{1}{2}}q\right]\\
&\quad \le O\left(n+\sqrt{M_1}\right) + 3\max_{p,q}\left\|g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}\right\|_{g(x_{\mathrm{mid}})}^2 \left\|g(x_{\mathrm{mid}})^{-\frac{1}{2}}p\right\|_{g(x_{\mathrm{mid}})} \left\|g(x_{\mathrm{mid}})^{-\frac{1}{2}}q\right\|_{g(x_{\mathrm{mid}})}\\
&\quad = O\left(n+\sqrt{M_1}\right) + O\left(n+\sqrt{M_1}\right)\max_{p,q}\|p\|_2\|q\|_2\\
&\quad = O\left(n+\sqrt{M_1}\right),
\end{aligned}$$

where the second inequality for $\|B_3\|$ follows from the highly self-concordance of ϕ. Due to $\|B\| \le \sum_{i=1}^4 \|B_i\|$, we have $\left\|\frac{h}{2}B\right\| = O\left(h\left(n+\sqrt{M_1}\right)\right)$. Hence, the condition of $h^2\left(n+\sqrt{M_1}\right)^2 \le 10^{-10}$ ensures that the inverse of $I_{2n}-\frac{h}{2}B$ exists, and it can be written as a series of matrices,

$$\left(I_{2n}-\frac{h}{2}B\right)^{-1} = \sum_{i=0}^{\infty}(hB/2)^i.$$

By substituting this series into (C.7),

$$\begin{aligned}
G(x_{\mathrm{mid}})DY\left(r_{\frac{1}{3}}\right)G(x_{\mathrm{mid}})^{-1} &= \sum_{i=0}^{\infty}(hB/2)^i\left(I_{2n}+\frac{h}{2}B\right) = \sum_{i=0}^{\infty}(hB/2)^i + \sum_{i=1}^{\infty}(hB/2)^i\\
&= I_{2n} + 2\sum_{i=1}^{\infty}(hB/2)^i.
\end{aligned}$$

By multiplying $\begin{bmatrix} I_n & 0 \end{bmatrix}^\top$ to the left and $\begin{bmatrix} 0 \\ I_n \end{bmatrix}$ to the right on both sides,

$$g(x_{\mathrm{mid}})^{\frac{1}{2}}D\overline{T}_x(v)g(x_{\mathrm{mid}})^{\frac{1}{2}} = 2\sum_{i=1}^{\infty}\begin{bmatrix} I_n & 0 \end{bmatrix}^\top (hB/2)^i \begin{bmatrix} 0 \\ I_n \end{bmatrix}.$$

From (C.8), B is of the form

$$B = \begin{bmatrix} C & I_n \\ -C^2+R & -C \end{bmatrix},$$

where $C \in \mathbb{R}^{n\times n}$ is symmetric and $R \in \mathbb{R}^{n\times n}$, and thus by Lemma 58

$$g(x_{\mathrm{mid}})^{\frac{1}{2}}D\overline{T}_x(v)g(x_{\mathrm{mid}})^{\frac{1}{2}} = h\sum_{i=0}^{\infty}(h^2R/2)^i,$$

where $R = \frac{1}{2}g(x_{\mathrm{mid}})^{-\frac{1}{2}}D^2g(x_{\mathrm{mid}})\left[g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}\right]g(x_{\mathrm{mid}})^{-\frac{1}{2}}$. Thus for $E \overset{\mathrm{def}}{=} \sum_{i=1}^{\infty}(h^2R/2)^i$

$$\frac{1}{h}g(x_{\mathrm{mid}})^{\frac{1}{2}}D\overline{T}_x(v)g(x_{\mathrm{mid}})^{\frac{1}{2}} = I + E. \tag{C.9}$$

We now bound its operator norm, trace and Frobenius norm. It is easy to see that

$$\begin{aligned}
\|E\|_2 &\lesssim \sum_{i\geq 1}\left(\frac{h^2}{2}\left(n+\sqrt{M_1}\right)\right)^i,\\
\mathrm{Tr}(E) &\lesssim \sum_{i\geq 1}\left(\frac{h^2}{2}\mathrm{Tr}(R)\right)^i \leq \sum_{i\geq 1}\left(\frac{h^2}{2}n\left(n+\sqrt{M_1}\right)\right),\\
\|E\|_F &\lesssim \sum_{i\geq 1}\left(\frac{h^2}{2}\sqrt{n}\left(n+\sqrt{M}\right)\right)^i,
\end{aligned}$$

where we used the following estimations

$$\begin{aligned}
\|R\|_2 &\leq O\left(n+\sqrt{M_1}\right)\\
\mathrm{Tr}(R) &= \frac{1}{2}\mathbb{E}_{p\sim\mathcal{N}(0,I)}p^\top g(x_{\mathrm{mid}})^{-\frac{1}{2}}D^2g(x_{\mathrm{mid}})[g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}, g^{-1}v_{\mathrm{mid}}]g(x_{\mathrm{mid}})^{-\frac{1}{2}}p\\
&\leq \mathbb{E}D^2g(x_{\mathrm{mid}})\left[g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-1}v_{\mathrm{mid}}, g(x_{\mathrm{mid}})^{-\frac{1}{2}}p, g(x_{\mathrm{mid}})^{-\frac{1}{2}}p\right]\\
&\leq \mathbb{E}\left\|v_{\mathrm{mid}}\right\|^2_{g^{-1}_{\mathrm{mid}}}\|p\|_2^2 = O\left(n\left(n+\sqrt{M_1}\right)\right),\\
\|R\|_F &\leq \sqrt{n}\,\|R\|_2 = O\left(\sqrt{n}\left(n+\sqrt{M}\right)\right).
\end{aligned}$$

Therefore, the step-size condition of $h^2\left(n+\sqrt{M_1}\right)^2 \leq 10^{-10}$ ensures that these three quantities can be made smaller than 10^{-8}. Applying Lemma 60 to (C.9), we have

$$e^{\mathrm{Tr}(E)}e^{-\|E\|_F^2} \leq \left|\frac{1}{h}g(x_{\mathrm{mid}})^{\frac{1}{2}}D\overline{T}_x(v)g(x_{\mathrm{mid}})^{\frac{1}{2}}\right| \leq e^{\mathrm{Tr}(E)}e^{\|E\|_F^2},$$

and thus

$$\left|D\overline{T}_x(v)\right| \geq (1-10^{-6})\cdot\frac{h^n}{|g(x_{\mathrm{mid}})|}.$$

Since $\log\det g(x) = \log\det\nabla^2\phi(x)$ is convex in x, it follows that

$$\begin{aligned}
\log|g(x_{\mathrm{mid}})| &= \log\left|g\left(\frac{x_{\frac{1}{3}}+x_{\frac{2}{3}}}{2}\right)\right| \leq \frac{1}{2}\left(\log\left|g(x_{\frac{1}{3}})\right|+\log\left|g(x_{\frac{2}{3}})\right|\right) = \frac{1}{2}\log\left|g(x_{\frac{1}{3}})g(x_{\frac{2}{3}})\right|\\
&= \log\sqrt{|g(x)|\,|g(\bar{x})|},
\end{aligned}$$

and thus $\left|D\overline{T}_x(v)\right| \geq \frac{(1-10^{-6})h^n}{\sqrt{|g(x)||g(\bar{x})|}}$. Due to the step-size conditions of $h^2\sqrt{n}\bar{R}_1 \leq 10^{-5}$ and in Proposition 28, we can use Proposition 30 to conclude that $\overline{T}_x(v)$ is sensitive at (x,v). ■

C.2. Generalized Leapfrog method (Störmer–Verlet)

We now analyze the generalized Leapfrog method (Algorithm 3), which is symplectic and reversible in the Riemannian settings. In a similar way we analyzed IMM, we show that if step size h satisfies $h^2\left(n+\sqrt{M_1}\right) \leq 10^{-10}$, then LM is second-order for $x \in \mathcal{M}_\rho$ and $v \in V^x_{\mathrm{good}}$ with

$C_x(x,v) = O\left(n + \sqrt{M_1}\right)$ and $C_v(x,v) = O\left(\sqrt{n+\sqrt{M_1}}\left(n + \sqrt{M_1} + M_2^*\right)\right)$. Next, if the step size h satisfies $h^2 \leq \min\left(\frac{10^{-20}}{n^2(n+\sqrt{M_1})}, \frac{10^{-5}}{\sqrt{n}R_1}\right)$ in addition to the step-size conditions in Proposition 28, then LM is sensitive at $x \in \mathcal{M}_\rho$ and $v \in V^x_{\text{good}}$.

Algorithm 3: `Generalized Leapfrog Method`

Input: Initial point x, velocity v, step size h

`// Step 1: Update v (Implicit)`

Find $v_{\frac{1}{2}}$ such that $v_{\frac{1}{2}} \leftarrow v - \frac{h}{2}\frac{\partial H(x,v)}{\partial x}$.

`// Step 2: Update x (Implicit)`

Find x_1 such that

$$x_1 = x + \frac{h}{2}\left(\frac{\partial H}{\partial v}\left(x, v_{\frac{1}{2}}\right) + \frac{\partial H}{\partial v}\left(x_1, v_{\frac{1}{2}}\right)\right).$$

`// Step 3: Update v (Explicit)`

Set $v_1 \leftarrow v_{\frac{1}{2}} - \frac{h}{2}\frac{\partial H}{\partial x}\left(x_1, v_{\frac{1}{2}}\right)$.

Output: x_1, v_1

C.2.1. SECOND-ORDER

Lemma 35 *For $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$, let $g = g(x)$ and h step size of LM with $h^2\left(n + \sqrt{M_1}\right) \leq 10^{-10}$. Let $(\overline{x}, \overline{v})$ be the point obtained from RHMC discretized by LM with the step size h and initial condition (x,v).*

1. $\|x - \overline{x}\|_g = O\left(h\sqrt{n} + h^2\left(n + \sqrt{M_1}\right)\right)$ *and* $\|v - \overline{v}\|_{g^{-1}} = O\left(h\left(\|\nabla f(\bar{x})\|_{g^{-1}} + n + \sqrt{M_1}\right)\right)$.
2. $C_x(x,v) = O\left(n + \sqrt{M_1}\right)$.
3. $C_v(x,v) = O\left(\sqrt{n+\sqrt{M_1}}\left(n + \sqrt{M_1} + M_2^*\right)\right)$.

Proof of 1. Let $\bar{x} = x_1 = \overline{T}_x(v)$ and $v_1 (= \bar{v}), v_{\frac{1}{2}}$ be the velocity obtained from LM with the initial condition (x,v) and the step size h. Let $g_1 = g(x_1)$. As $v_{\frac{1}{2}} \to v$ as $h \to 0$, we can take $h_0 > 0$ such that $h\left(\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} + \left\|g_1^{-1}v_{\frac{1}{2}}\right\|_g\right) \leq \frac{2}{1000}$ for $h \leq h_0$ with the equality held at $h = h_0$. Thus for $h \leq h_0$ we have $h\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} \leq \frac{1}{500}$.

From the first step of Algorithm 3 and Lemma 27, for step size $h \leq h_0$ it follows from $x \in \mathcal{M}_\rho$ that

$$\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} \leq \|v\|_{g^{-1}} + \frac{h}{2}\left(\sqrt{M_1} + n + \left\|v_{\frac{1}{2}}\right\|^2_{g^{-1}}\right).$$

Multiplying h to both sides, and using $h\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} \leq \frac{1}{500}$ and $v \in V^x_{\text{good}}$,

$$h\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} \leq 150h\sqrt{n} + \frac{h^2}{2}\left(\sqrt{M_1}+n\right) + \frac{h^2}{2}\left\|v_{\frac{1}{2}}\right\|_{g^{-1}}^2$$
$$\leq 150h\sqrt{n} + \frac{h^2}{2}\left(\sqrt{M_1}+n\right) + \frac{1}{1000}h\left\|v_{\frac{1}{2}}\right\|_{g^{-1}},$$

and thus

$$\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} \leq 200\sqrt{n} + h^2\left(n+\sqrt{M_1}\right). \tag{C.10}$$

From the second step of Algorithm 3 and Lemma 27, for step size $h \leq h_0$

$$\|x_1 - x\|_g \leq \frac{h}{2}\left(\left\|\frac{\partial H}{\partial v}\left(x, v_{\frac{1}{2}}\right)\right\|_g + \left\|\frac{\partial H}{\partial v}\left(x_1, v_{\frac{1}{2}}\right)\right\|_g\right) \leq \frac{h}{2}\left(\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} + \left\|g_1^{-1}v_{\frac{1}{2}}\right\|_g\right),$$

and thus its is obvious that $\|x - x_1\|_g \leq \frac{1}{500}$. We now lower bound h_0 as follows:

$$\frac{1}{500} = h_0\left(\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} + \left\|g_1^{-1}v_{\frac{1}{2}}\right\|_g\right) \leq h_0\left(\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} + 1.1\left\|v_{\frac{1}{2}}\right\|_{g^{-1}}\right)$$
$$\leq 3h_0\left\|v_{\frac{1}{2}}\right\|_{g^{-1}}$$
$$\leq 600h_0\sqrt{n} + 3h_0^2\left(n+\sqrt{M_1}\right),$$

where in the first inequality we switched the local norm at from x_1 to x due to $\|x_1 - x\|_g \leq \frac{1}{500}$ and used (C.10) in the last inequality. Therefore, $h_0 \geq \frac{1}{10^4\sqrt{n+\sqrt{M_1}}}$ and for step size $h \leq \frac{1}{10^5\sqrt{n+\sqrt{M_1}}}$ we have

$$\|x - x_1\|_g \leq 600h\sqrt{n} + 31h^2\left(n+\sqrt{M_1}\right),$$
$$\left\|v - v_{\frac{1}{2}}\right\|_{g^{-1}} \leq h\left(20000n + \sqrt{M_1}\right).$$

Similarly, we can bound $\left\|v_1 - v_{\frac{1}{2}}\right\|_{g^{-1}}$ by Lemma 27:

$$\left\|v_1 - v_{\frac{1}{2}}\right\|_{g^{-1}} \leq \frac{h}{2}\left\|\frac{\partial H}{\partial x}\left(x_1, v_{\frac{1}{2}}\right)\right\|_{g^{-1}} \leq \frac{h}{2}\left(\|\nabla f(x_1)\|_{g^{-1}} + n + \left\|v_{\frac{1}{2}}\right\|_{g_{\text{mid}}^{-1}}^2\right)$$
$$\leq 40000h\left(\|\nabla f(x_1)\|_{g^{-1}} + n + \sqrt{M_1}\right),$$

and thus by adding it to the inequality for $\left\|v - v_{\frac{1}{2}}\right\|_{g^{-1}}$ we have

$$\|v - v_1\|_{g^{-1}} \leq 40000\left(\|\nabla f(x_1)\|_{g^{-1}} + n + \sqrt{M_1}\right).$$

Proof of 2. For $t \in [0,h]$, let (x_t, v_t) be the Hamiltonian curve of the ideal RHMC at time t starting from (x,v). Recall that

$$\begin{aligned} T_x(v) &= x + \int_0^h \frac{\partial H}{\partial v}(x_t, v_t)dt = x + \int_0^h g_t^{-1} v_t dt, \\ \overline{T}_x(v) &= \overline{x} = x + \frac{h}{2}\left(g^{-1} + g_1^{-1}\right) v_{\frac{1}{2}}. \end{aligned}$$

Thus,

$$\begin{aligned} \left\|T_x(v) - \overline{T}_x(v)\right\|_g &= \left\| \left(x + \int_0^h g_t^{-1} v_t dt \right) - \left(x + \frac{h}{2}\left(g^{-1} + g_1^{-1}\right) v_{\frac{1}{2}} \right) \right\|_g \\ &= \left\| \int_0^h \left(g_t^{-1} v_t - \frac{1}{2}\left(g^{-1} + g_1^{-1}\right) v_{\frac{1}{2}} \right) dt \right\|_g \\ &\leq h \max_{t\in[0,h]} \left\| g_t^{-1} v_t - \frac{1}{2}\left(g^{-1} + g_1^{-1}\right) v_{\frac{1}{2}} \right\|_g. \end{aligned}$$

By Lemma 26-11,

$$\begin{aligned} \left\| g_t^{-1} v_t - \frac{1}{2}\left(g^{-1} + g_1^{-1}\right) v_{\frac{1}{2}} \right\|_g &\leq \frac{1}{2}\left\| g_t^{-1} v_t - g^{-1} v_{\frac{1}{2}} \right\|_g + \frac{1}{2}\left\| g_t^{-1} v_t - g_1^{-1} v_{\frac{1}{2}} \right\|_g \\ &\lesssim \left\| v_t - v_{\frac{1}{2}} \right\|_{g^{-1}} + \|x_t - x\|_g \|v_t\|_{g^{-1}} \\ &\quad + \left\| v_t - v_{\frac{1}{2}} \right\|_{g^{-1}} + \|x_t - x_1\|_g \|v_t\|_{g^{-1}} \\ &\lesssim \left(\|v_t - v\|_{g^{-1}} + \left\| v - v_{\frac{1}{2}} \right\|_{g^{-1}} \right) \\ &\quad + \left(\|x_t - x\|_g + \|x - x_1\|_g \right) \left(\|v_t - v\|_{g^{-1}} + \|v\|_{g^{-1}} \right). \end{aligned}$$

Using our bounds on $\|x_1 - x\|_g$, $\|x_t - x\|_g$ and $\|v\|_{g^{-1}}$, $\|v_t - v\|_{g^{-1}}$, $\left\| v - v_{\frac{1}{2}} \right\|_{g^{-1}}$, we conclude that
$\max_{t\in[0,h]} \left\| g_t^{-1} v_t - g_{\text{mid}}^{-1} v_{\text{mid}} \right\|_{g_{\text{mid}}} \leq 10^4 h \left(n + \sqrt{M_1}\right)$ and thus

$$\left\|T_x(v) - \overline{T}_x(v)\right\|_g \leq 10^4 h^2 \left(n + \sqrt{M_1}\right).$$

Proof of 3. From the algorithm,

$$\begin{aligned} v_h &= v - \int_0^h \frac{\partial H}{\partial x}(x_t, v_t)dt, \\ \bar{v} &= v_{\frac{1}{2}} - \frac{h}{2}\frac{\partial H}{\partial x}\left(x_1, v_{\frac{1}{2}}\right) \\ &= v - \frac{h}{2}\frac{\partial H}{\partial x}\left(x_1, v_{\frac{1}{2}}\right) - \frac{h}{2}\frac{\partial H(x,v)}{\partial x}. \end{aligned}$$

Thus,

$$\begin{aligned}\|v_h - \bar{v}\|_{g^{-1}} &= \left\| \int_0^h \frac{1}{2}\left(\frac{\partial H}{\partial x}(x_t, v_t) - \frac{\partial H}{\partial x}\left(x_1, v_{\frac{1}{2}}\right)\right) dt + \int_0^h \frac{1}{2}\left(\frac{\partial H}{\partial x}(x_t, v_t) - \frac{\partial H}{\partial x}(x, v)\right) dt \right\|_{g^{-1}} \\ &\leq \underbrace{\frac{h}{2} \max_{t\in[0,h]} \left\| \frac{\partial H}{\partial x}(x_t, v_t) - \frac{\partial H}{\partial x}\left(x_1, v_{\frac{1}{2}}\right)\right\|_{g^{-1}}}_{F} + + \underbrace{\frac{h}{2} \max_{t\in[0,h]} \left\| \frac{\partial H}{\partial x}(x_t, v_t) - \frac{\partial H}{\partial x}(x, v)\right\|_{g^{-1}}}_{S}.\end{aligned}$$

For $\delta_v = \left\| v_t - v_{\frac{1}{2}} \right\|_{g^{-1}}$ and $\delta_x = \|x_t - x_1\|_g$, we use Proposition 26-12 to show that

$$F \lesssim \max_{t\in[0,h]} \left(\delta_v + \delta_x \left\| v_{\frac{1}{2}} \right\|_{g^{-1}}\right) \left(\left\| v_{\frac{1}{2}} \right\|_{g^{-1}} + \|v_t\|_{g^{-1}} \right) + M_2^* \delta_x.$$

In a similar way, S can be bounded as follows:

$$S \lesssim h\left(n + \sqrt{M_1}\right)^{3/2} + M_2^* \delta_x.$$

Using our bounds on $\|x_1 - x\|_g$, $\|x_t - x\|_g$ and $\|v\|_{g^{-1}}$, $\|v_t - v\|_{g^{-1}}$, $\left\| v - v_{\frac{1}{2}} \right\|_{g^{-1}}$,

$$F + S \lesssim h\left(n + \sqrt{M_1}\right)^{3/2} + h\sqrt{n + \sqrt{M_1}} M_2^*.$$

Substituting the bounds on F and S, we can conclude that

$$\|v_h - \bar{v}\|_{g^{-1}} \leq 10^{10} \left(\left(n + \sqrt{M_1}\right)^{3/2} + \sqrt{n + \sqrt{M_1}} M_2^* \right) h^2.$$

C.2.2. SENSITIVITY

We show that for some step size h the generalized Leapfrog integrator is sensitive at (x, v) for $x \in \mathcal{M}_\rho$ and $v \in V^x_{\text{good}}$.

Lemma 36 *For $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$, LM is sensitive at (x, v) if step size h satisfies $h^2 \leq \min\left(\frac{10^{-10}}{n^2\left(n+\sqrt{M_1}\right)}, \frac{10^{-5}}{\sqrt{n}R_1}\right)$ and the step-size conditions in Proposition 28.*

Proof Let $\overline{T}_x(v) = \bar{x} = x_1$. We lower bound $\left|D\overline{T}_x(v)\right|$. It suffices to look into the determinant of composition of first two steps in Algorithm 3, since the third step only changes v. The first two steps are

$$\begin{aligned} v_{\frac{1}{2}} &= v - \frac{h}{2}\frac{\partial H}{\partial x}(x, v_{\frac{1}{2}}), \\ x_1 &= x + \frac{h}{2}\left(\frac{\partial H}{\partial v}(x, v_{\frac{1}{2}}) + \frac{\partial H}{\partial v}(x_1, v_{\frac{1}{2}})\right). \end{aligned}$$

Differentiating the first equation with respect to v, we have

$$\frac{\partial v_{\frac{1}{2}}}{\partial v} = I_n - \frac{h}{2}\frac{\partial^2 H}{\partial x \partial v}\left(x, v_{\frac{1}{2}}\right)\frac{\partial v_{\frac{1}{2}}}{\partial v},$$

and so

$$\left(I + \frac{h}{2}\frac{\partial^2 H}{\partial x \partial v}\left(x, v_{\frac{1}{2}}\right)\right)\frac{\partial v_{\frac{1}{2}}}{\partial v} = I_n. \tag{C.11}$$

Differentiating the second equation with respect to v, we obtain

$$\frac{\partial x_1}{\partial v} = \frac{h}{2}\left(\frac{\partial^2 H}{\partial v^2}\left(x, v_{\frac{1}{2}}\right)\frac{\partial v_{\frac{1}{2}}}{\partial v} + \frac{\partial^2 H}{\partial x \partial v}\left(x_1, v_{\frac{1}{2}}\right)\frac{\partial x_1}{\partial v} + \frac{\partial^2 H}{\partial v^2}\left(x_1, v_{\frac{1}{2}}\right)\frac{\partial v_{\frac{1}{2}}}{\partial v}\right).$$

Collecting all $\partial x_1/\partial v$ terms from this equation, for $g = g(x)$ and $g_1 = g(x_1)$

$$\begin{aligned}\left(I_n - \frac{h}{2}\frac{\partial^2 H}{\partial x \partial v}\left(x_1, v_{\frac{1}{2}}\right)\right)\frac{\partial x_1}{\partial v} &= \frac{h}{2}\left(\frac{\partial^2 H}{\partial v^2}\left(x, v_{\frac{1}{2}}\right) + \frac{\partial^2 H}{\partial v^2}\left(x_1, v_{\frac{1}{2}}\right)\right)\frac{\partial v_{\frac{1}{2}}}{\partial v} \\ &= \frac{h}{2}\left(g^{-1} + g_1^{-1}\right)\left(I + \frac{h}{2}\frac{\partial^2 H}{\partial x \partial v}\left(x, v_{\frac{1}{2}}\right)\right)^{-1},\end{aligned}$$

where we used (C.11). Hence,

$$\begin{aligned}\frac{\partial x_1}{\partial v} &= h\left(I_n - \frac{h}{2}\frac{\partial^2 H}{\partial x \partial v}\left(x_1, v_{\frac{1}{2}}\right)\right)^{-1}\left(\frac{g^{-1} + g_1^{-1}}{2}\right)\left(I_n + \frac{h}{2}\frac{\partial^2 H}{\partial x \partial v}\left(x, v_{\frac{1}{2}}\right)\right)^{-1} \\ &= h\left(I_n - \frac{h}{2}g_1^{-1}Dg_1\left[g_1^{-1}v_{\frac{1}{2}}\right]\right)^{-1}\left(\frac{g^{-1} + g_1^{-1}}{2}\right)\left(I_n + \frac{h}{2}g^{-1}Dg\left[g^{-1}v_{\frac{1}{2}}\right]\right)^{-1} \\ &= hg_1^{\frac{1}{2}}\left(I_n - \frac{h}{2}g_1^{-\frac{1}{2}}Dg_1\left[g_1^{-1}v_{\frac{1}{2}}\right]g_1^{-\frac{1}{2}}\right)^{-1}g_1^{-\frac{1}{2}}\left(\frac{g^{-1} + g_1^{-1}}{2}\right)g^{\frac{1}{2}}\left(I_n + \frac{h}{2}g^{-\frac{1}{2}}Dg\left[g^{-1}v_{\frac{1}{2}}\right]g^{-\frac{1}{2}}\right)^{-1}g^{-\frac{1}{2}}.\end{aligned}$$

Due to the concavity of log-determinant in the set of positive definite matrices, we have

$$\log\left|\frac{g^{-1} + g_1^{-1}}{2}\right| \geq \frac{1}{2}\left(\log\left|g^{-1}\right| + \log\left|g_1^{-1}\right|\right) = \log\frac{1}{\sqrt{|g|\,|g_1|}},$$

and thus

$$\begin{aligned}\left|D\overline{T}_x(v)\right| &= \left|\frac{\partial x_1}{\partial v}\right| \\ &\geq \frac{h^n}{\sqrt{|g|\,|g_1|}}\left|I_n - \frac{h}{2}g_1^{-\frac{1}{2}}Dg_1\left[g_1^{-1}v_{\frac{1}{2}}\right]g_1^{-\frac{1}{2}}\right|^{-1}\left|I_n + \frac{h}{2}g^{-\frac{1}{2}}Dg\left[g^{-1}v_{\frac{1}{2}}\right]g^{-\frac{1}{2}}\right|^{-1}.\end{aligned}$$

For $E \stackrel{\text{def}}{=} \frac{h}{2}g^{-\frac{1}{2}}Dg\left[g^{-1}v_{\frac{1}{2}}\right]g^{-\frac{1}{2}}$, as bounded in (C.8), we have that

$$\begin{aligned}\|E\|_2 &\leq \frac{h}{2}\left\|v_{\frac{1}{2}}\right\|_{g^{-1}}, \\ \operatorname{Tr}(E) &\lesssim hn\left\|v_{\frac{1}{2}}\right\|_{g^{-1}}, \\ \|E\|_F &\leq \frac{h\sqrt{n}}{2}\left\|v_{\frac{1}{2}}\right\|_{g^{-1}}.\end{aligned}$$

Due to $h^2 \leq \frac{10^{-10}}{n^2\left(n+\sqrt{M_1}\right)}$, it follows from (C.10) that $\left\|v_{\frac{1}{2}}\right\|_{g^{-1}} \leq O\left(\sqrt{n+\sqrt{M_1}}\right)$. This condition also allows us to make all these three quantities smaller than 10^{-5}. By Lemma 60,

$$\left|I_n - \frac{h}{2}g_1^{-\frac{1}{2}}Dg_1\left[g_1^{-1}v_{\frac{1}{2}}\right]g_1^{-\frac{1}{2}}\right|^{-1} \geq 1 - 10^{-8}.$$

Similarly, we obtain

$$\left| I_n + \frac{h}{2} g^{-\frac{1}{2}} Dg \left[g^{-1} v_{\frac{1}{2}} \right] g^{-\frac{1}{2}} \right|^{-1} \geq 1 - 10^{-8},$$

and thus $\left| D\overline{T}_x(v) \right| \geq \frac{(1-10^{-6})h^n}{\sqrt{|g(x)||g(\bar{x})|}}$. Using the step-size conditions in Proposition 28, we use Proposition 30 to show that $\overline{T}_x$ is sensitive at (x, v). ■

Appendix D. Convergence rate of RHMC in polytopes

In this section, we present the mixing times of the ideal and discretized RHMC for an exponential density in a polytope. We set $f(x) = \alpha^\top x$ for $\alpha \in \mathbb{R}^n$. For a full-rank matrix $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, the polytope is represented by $\{x \in \mathbb{R}^n : Ax \geq b\}$, equipped with the logarithmic barrier $\phi(x) = -\sum_{i=1}^m \log(a_i^\top x - b_i)$, where a_i is the i^{th} row of A and b_i is the i^{th} entry of b. We can check by direct computation that the logarithmic barriers are highly self-concordant. We view this polytope as the Hessian manifold $\mathcal{M}$ induced by the local norm $g(x) = \nabla^2 \phi(x)$. We denote a slack vector by $s_x = Ax - b \in \mathbb{R}^m$ and its diagonalization by $S_x = \mathrm{Diag}(s_x) \in \mathbb{R}^{m \times m}$. We also define $A_x = S_x^{-1} A$ and $s_v = A_x v$ for $v \in T_x \mathcal{M}$, where $T_x \mathcal{M}$ is endowed with the local metric g. One can check by direct computation that $\nabla^2 \phi(x) = A_x^\top A_x$.

In this setting, we can compute all the parameters we have defined, obtaining the mixing time of RHMC discretized by a sensitive numerical integrator.

D.1. Isoperimetry of convex set

An isoperimetry inequality is one of the two main ingredients for bounding the mixing rate. We use the Riemannian version of some isoperimetry inequality. To state it, we need another distance called *Hilbert distance* in addition to Riemannian distance d_ϕ.

Definition 37 *For a convex body $\mathcal{K}$, the cross-ratio distance $d_{\mathcal{K}}(x, y)$ between x and y is*

$$d_{\mathcal{K}}(x, y) = \frac{|x - y||p - q|}{|p - x||y - q|},$$

where p and q are on the boundary of $\mathcal{K}$ such that p, x, y, q are on the straight line $\overline{xy}$ and are in order. The Hilbert distance d_H between $x, y \in \mathcal{K}$ is

$$d_H(x, y) = \log(1 + d_{\mathcal{K}}(x, y)) = \log\left(1 + \frac{|x - y||p - q|}{|p - x||y - q|}\right).$$

For sets X and Y, we define $d_*(X, Y) = \inf_{x \in X, y \in Y} d_*(x, y)$ for $* \in \{\mathcal{K}, H, \phi\}$.

Lemma 38 (Vempala (2005), Theorem 4.4) *Let π be a log-concave distribution supported on a convex body $\mathcal{K}$. Let S_1, S_2, S_3 be a partition of $\mathcal{K}$. Then,*

$$\pi(S_3) \geq d_{\mathcal{K}}(S_1, S_2)\pi(S_1)\pi(S_2).$$

The following lemma is a generalization of Theorem 26 in Lee and Vempala (2017) to a subset $\mathcal{K}'$.

Lemma 39 *Let π be a log-concave distribution supported on a convex body $\mathcal{K}$, and ϕ a self-concordant barrier of $\mathcal{K}$. Let $\mathcal{K}'$ be a convex subset of $\mathcal{K}$, and S_1, S_2, S_3 a partition of $\mathcal{K}'$. Then*

$$\pi(S_3)\pi(\mathcal{K}') \geq \frac{d_\phi(S_1, S_2)}{G}\pi(S_1)\pi(S_2),$$

where $G = \sup_{x,y\in\mathcal{K}} \frac{d_\phi(x,y)}{d_H(x,y)}$.

Proof Applying Lemma 38 to the distribution $\pi_{\mathcal{K}'}$ defined by π restricted to $\mathcal{K}'$, we have

$$\pi(S_3)\pi(\mathcal{K}') \geq d_{\mathcal{K}'}(S_1, S_2)\pi(S_1)\pi(S_2).$$

Due to $\mathcal{K}' \subseteq \mathcal{K}$, one can check $d_{\mathcal{K}'}(S_1, S_2) \geq d_{\mathcal{K}}(S_1, S_2)$ by simple algebra. As $d_{\mathcal{K}}(x, y) \geq d_H(x, y)$, it follows that

$$\pi(S_3)\pi(\mathcal{K}') \geq \frac{d_\phi(S_1, S_2)}{\frac{d_\phi(S_1,S_2)}{d_H(S_1,S_2)}}\pi(S_1)\pi(S_2) \geq \frac{d_\phi(S_1, S_2)}{G}\pi(S_1)\pi(S_2).$$

■

We now define the symmetric self-concordance parameter of the barrier ϕ.

Definition 40 (Laddha et al. (2020)) *For a convex body $\mathcal{K} \subseteq \mathbb{R}^n$, the symmetric self-concordance parameter $\bar{\nu}_\phi$ of $\mathcal{K}$ is the smallest number such that for any $x \in \mathcal{K}$*

$$D(x) \subseteq \mathcal{K} \cap (2x - \mathcal{K}) \subseteq \sqrt{\bar{\nu}_\phi} D(x),$$

where $D(x) = \left\{y \in \mathbb{R}^n : \|y - x\|_{\nabla^2\phi(x)} \leq 1\right\}$ is the Dikin ellipsoid at x.

In general, it is known that $\bar{\nu}_\phi = O(\nu_\phi^2)$ for the self-concordance parameter ν_ϕ (see Definition 54), but a tighter bound of $\bar{\nu}_\phi = O(\nu_\phi)$ holds for important barriers such as the logarithmic barrier and Lee-Sidford barrier Lee and Sidford (2014).

Lemma 41 (Laddha et al. (2020), Lemma 2.3) *$d_\phi(x, y) \lesssim \sqrt{\bar{\nu}_\phi} d_H(x, y)$ for any $x, y \in \mathcal{K}$.*

Using Lemma 39 and 41 together, we have

$$\pi(S_3)\pi(\mathcal{K}') \geq \frac{d_\phi(S_1, S_2)}{\sqrt{\bar{\nu}_\phi}}\pi(S_1)\pi(S_2),$$

and it implies that the isoperimetry of $\mathcal{K}'$ is at least $1/\sqrt{\bar{\nu}_\phi}$. As $\nu_\phi = O(m)$ for the logarithmic barrier, $\psi_{\mathcal{K}'} \geq 1/\sqrt{m}$ for a convex subset $\mathcal{K}'$.

D.2. Good region $\mathcal{M}_\rho$

Taking a proper good region $\mathcal{M}_\rho$ plays an important role in establishing a condition-number independent mixing rate of RHMC for an exponential density in a polytope. To this end, we set our good region to

$$\mathcal{M}_\rho \stackrel{\text{def}}{=} \left\{x \in \mathcal{M} : \|\alpha\|^2_{g(x)^{-1}} \leq 10n^2 \log^2\left(\frac{1}{\rho}\right)\right\}.$$

To establish the isoperimetry of $\mathcal{M}_\rho$ following Section D.1, we check its convexity in the following lemma. Note that the assumption in the lemma is satisfied by the logarithmic barriers.

Lemma 42 *If the fourth directional derivative of ϕ is positive (i.e., $D^4\phi[a,a,b,b] \geq 0$), then $\mathcal{M}_\rho$ is convex.*

Proof Let $\Upsilon(x) := \alpha^\top g(x)^{-1}\alpha = \alpha^\top(\nabla^2\phi(x))^{-1}\alpha$. It suffices to show that $\Upsilon(x)$ is convex. Note that

$$\frac{\partial\Upsilon(x)}{\partial x_i} = \alpha^\top g(x)^{-1}\frac{\partial g(x)}{\partial x_i}g(x)^{-1}\alpha,$$

and thus its directional derivative in $h = (h_1, ..., h_n)$ is

$$\nabla\Upsilon(x)\cdot h = \sum_i h_i\left(s(x)^\top\frac{\partial g(x)}{\partial x_i}s(x)\right),$$

where $s(x) := g(x)^{-1}\alpha$. Note that

$$\begin{aligned}\frac{\partial}{\partial x_j}\left(s(x)^\top\frac{\partial g(x)}{\partial x_i}s(x)\right) &= s(x)^\top\frac{\partial^2 g(x)}{\partial x_i\partial x_j}s(x) + 2s(x)^\top\frac{\partial g(x)}{\partial x_i}\left(\frac{\partial s(x)}{\partial x_j}\right)\\ &= s(x)^\top\frac{\partial^2 g(x)}{\partial x_i\partial x_j}s(x) + 2s(x)^\top\frac{\partial g(x)}{\partial x_i}g(x)^{-1}\frac{\partial g(x)}{\partial x_j}s(x).\end{aligned}$$

Therefore,

$$\begin{aligned}D^2\Upsilon(x)[h,h] &= \sum_{i,j}h_ih_js(x)^\top\frac{\partial^2 g(x)}{\partial x_i\partial x_j}s(x) + 2\sum_{i,j}\left(\frac{\partial g(x)}{\partial x_i}s(x)h_i\right)^\top g(x)^{-1}\left(\frac{\partial g(x)}{\partial x_j}s(x)h_j\right)\\ &= D^4\phi[h,h,s(x),s(x)] + 2\sum_{i,j}\left(\frac{\partial g(x)}{\partial x_i}s(x)h_i\right)^\top g(x)^{-1}\left(\frac{\partial g(x)}{\partial x_j}s(x)h_j\right).\end{aligned}$$

The first term is non-negative due to the assumption, and the second term is also non-negative since $g(x)^{-1}$ is also positive semi-definite. ■

Next, we show that $\mathcal{M}_\rho$ takes up probability of at least $1-\rho$ over the stationary distribution π, where $\frac{d\pi(x)}{dx} \propto \exp(-\alpha^\top x)$.

Lemma 43 $\pi(\mathcal{M}_\rho) \geq 1-\rho$.

Proof Let $g = g(x)$. For $\|\alpha\|_{g^{-1}}$, note that

$$\begin{aligned}\|\alpha\|_{g(x)^{-1}} =& \max_{\|u\|_{g(x)}=1}\alpha^\top u\\ =&\alpha^\top x - \min_{\|y-x\|_g=1}\alpha^\top y \leq \alpha^\top x - \min_{y\in\mathcal{M}}\alpha^\top y,\end{aligned}$$

where the first equality is due to duality of norms and the last inequality follows from the well-known fact that the Dikin ellipsoid at x is inside $\mathcal{M}$.

By Lemma 61 with $c = \alpha/\|\alpha\|_2$ and $T = 1/\|\alpha\|_2$, we have

$$\mathbb{E}_{x\sim\pi^*}[\alpha^\top x] \leq n + \min_{y\in\mathcal{M}}\alpha^\top y.$$

Then, by Lemma 62 we have $\mathbb{E}[(\alpha^\top x - \min_{y\in\mathcal{M}} \alpha^\top y)^2] - \mathbb{E}[\alpha^\top x - \min_{y\in\mathcal{M}} \alpha^\top y]^2 \le n$ so that $\mathbb{E}[(\alpha^\top x - \min_{y\in\mathcal{M}} \alpha^\top y)^2] \le n + n^2$. By Lemma 63, we have

$$\Pr_{x\sim\pi}\left[\alpha^\top x - \min_{y\in\mathcal{M}} \alpha^\top y > 2\left(\log\frac{1}{\rho}+1\right)n\right] \le \rho.$$

■

D.3. Auxiliary function ℓ and smoothness parameters R

In this region $\mathcal{M}_\rho$ and step size h, the parameters M_1, M_2 and M_1^*, M_2^* (see Definition 23) are computed by

$$\begin{aligned}
M_1 &= \max\left(n, \|\alpha\|^2_{g(x)^{-1}}\right) \le 10n^2\log^2\left(\frac{1}{\rho}\right),\\
M_1^* &\le \|\alpha\|^2_{g(x)^{-1}} \le 10n^2\log^2\left(\frac{1}{\rho}\right),\\
M_2, M_2^* &= 0.
\end{aligned}$$

We use the following auxiliary function ℓ proposed in Lee and Vempala (2018) and symmetric auxiliary function $\bar{\ell}$:

$$\ell(\gamma) = \max_{t\in[0,h]}\left(\frac{\|s_{\gamma'(t)}\|_2}{\sqrt{n}+2M_1^{1/4}} + \frac{\|s_{\gamma'(t)}\|_4}{2M_1^{1/4}} + \frac{\|s_{\gamma'(t)}\|_\infty}{\sqrt{\log n}+2h\sqrt{M_1}}\right) + \frac{\|s_{\gamma'(0)}\|_2}{\sqrt{n}} + \frac{\|s_{\gamma'(0)}\|_4}{n^{1/4}} + \frac{\|s_{\gamma'(0)}\|_\infty}{\sqrt{\log n}},$$

$$\bar{\ell}(\gamma) = \max_{t\in[0,h]}\left(\frac{\|s_{\gamma'(t)}\|_2}{\sqrt{n}+2M_1^{1/4}} + \frac{\|s_{\gamma'(t)}\|_4}{2M_1^{1/4}} + \frac{\|s_{\gamma'(t)}\|_\infty}{\sqrt{\log n}+2h\sqrt{M_1}}\right).$$

This measures how fast a Hamiltonian trajectory approaches the facets of a polytope in the local norm.

We make simple observations based on the self-concordance of g.

Proposition 44 *Let* $\overline{\mathcal{M}_\rho} \stackrel{\text{def}}{=} \left\{x\in\mathcal{M} : \|\alpha\|^2_{g(x)^{-1}} \le 20n^2\log^2\left(\frac{1}{\rho}\right)\right\}$ *and* γ *be any Hamiltonian curve* γ *starting at* $x\in\mathcal{M}_\rho$ *with* $v\in V^x_{good}$. *If step size* h *satisfies* $h^2 \le 10^{-11}\min\left(\frac{1}{n\log\frac{1}{\rho}}, \frac{1}{C_x(x,v)}\right)$, *then* x_h *and* $\bar{x}_h$ *are contained in* $\overline{\mathcal{M}_\rho}$.

Proof Due to the assumption on the step size, we can use Proposition 28-1, obtaining $\|x-\gamma(t)\|_x \le O\left(t\sqrt{n+\sqrt{M_1}}\right) = O\left(t\sqrt{n\log\frac{1}{\rho}}\right) < \frac{1}{4}$. Also, $\|x-\bar{x}_h\|_g \le \|x-\gamma(h)\|_g + \|\gamma(h)-\bar{x}_h\|_g \le \frac{1}{4} + h^2C_x(x,v) \le \frac{1}{3}$. The claim follows from the self-concordance of $g(x)$, due to $\|\alpha\|^2_{g(\gamma(h))^{-1}} \le (1+\|x-\gamma(h)\|_x)\|\alpha\|^2_{g(x)^{-1}} \le 20n^2\log^2\frac{1}{\rho}$ and $\|\alpha\|^2_{g(\bar{x}_h)^{-1}} \le \frac{4}{3}\|\alpha\|^2_{g(x)^{-1}} \le 20n^2\log^2\frac{1}{\rho}$. ■

As in Lee and Vempala (2018), we can represent the parameters ℓ_0, ℓ_1 and the smoothness parameters R_1, R_2, R_3 in terms of M_1. The original proof in Lee and Vempala (2018) relies on the fact that $\|\nabla f(\gamma(t))\|^2_{g(\gamma(t))^{-1}} \le M_1$ for any time $t\in[0,h]$ and any regular Hamiltonian curves. In our setting, $\|\nabla f(\gamma(t))\|^2_{g(\gamma(t))^{-1}} \le 2M_1$ for any time $t\in[0,h]$ if $h^2 \le \frac{10^{-11}}{n\log\frac{1}{\rho}}$, we can simply reproduce Lemma 54~59 by replacing M_1 by $2M_1$.

Lemma 45 *Consider a Hamiltonian trajectory γ starting at $x \in \mathcal{M}_\rho$ with an initial (normalized) velocity randomly chosen from $\mathcal{N}(0, g(x)^{-1})$, with step size h satisfying $h^2 n \log \frac{1}{\rho} \le 10^{-11}$. For n large enough, if s satisfies $sh = O(n)$, then*

$$\mathbf{P}_\gamma(\ell(\gamma) \ge 128) \le \frac{1}{100} \min\left(1, \frac{\ell_0}{sh}\right).$$

As we shortly see in Lemma 49, we have $\ell_1 h = O\left(h^2 M_1^{1/4}\right) = O\left(\frac{1}{\sqrt{n \log \frac{1}{\rho}}}\right)$, and thus ℓ_1 can be used in place of s in this lemma.

Lemma 46 *Let γ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \le \ell_0 \le 256$ and step size h satisfying $h^2 n \log \frac{1}{\rho} \le 10^{-11}$. Then*

$$\sup_{t \in [0,h]} \|\Phi(\gamma, t)\|_{F, \gamma(t)} \le R_1$$

with $R_1 = O\left(\sqrt{M_1}\right)$.

Lemma 47 *Let γ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \le \ell_0 \le 256$ and step size h satisfying $h^2 n \log \frac{1}{\rho} \le 10^{-11}$. For any $t \in [0, h]$, any curve $c(s)$ starting from $\gamma(t)$ and any velocity field $v(c(s))$ on $c(s)$ with $v(c(0)) = v(\gamma(t)) = \gamma'(t)$, we have that*

$$\left| \frac{d}{ds} \mathrm{Tr} \Phi(v(c(s)) \Big|_{s=0} \right| \le R_2 \left(\left\| \frac{dc}{ds} \Big|_{s=0} \right\|_{\gamma(t)} + h \left\| D_s v|_{s=0} \right\|_{\gamma(t)} \right)$$

with $R_2 = O\left(\sqrt{n M_1} + \sqrt{n} M_1 h^2 + \frac{M_1^{1/4}}{h} + \frac{\sqrt{n \log n}}{h}\right)$.

Lemma 48 *Let γ be a Hamiltonian curve starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma) \le \ell_0 \le 256$ and step size h satisfying $h^2 n \log \frac{1}{\rho} \le 10^{-11}$. Let $\zeta(t)$ be the parallel transport of the vector $\gamma'(0)$ to $\gamma(t)$. Then*

$$\sup_{t \in [0,h]} \|\Phi(\gamma, t) \zeta(t)\|_{\gamma(t)} \le R_3$$

with $R_3 = O\left(\sqrt{M_1 \log n} + M_1^{3/4} n^{1/4} h\right)$.

Lemma 49 *Let γ_s be a Hamiltonian variation starting at $x \in \mathcal{M}_\rho$ with $\ell(\gamma_s) \le \ell_0 \le 256$ and step size h satisfying $h^2 n \log \frac{1}{\rho} \le 10^{-11}$. Then*

$$\left| \frac{d}{ds} \ell(\gamma_s) \right| \le O\left(M_1^{1/4} h + \frac{1}{h\sqrt{\log n}}\right) \left(\left\| \frac{d}{ds} \gamma_s(0) \right\|_{\gamma_s(0)} + h \left\| D_s \gamma_s'(0) \right\|_{\gamma_s(0)} \right),$$

and thus $\ell_1 = O\left(M_1^{1/4} h + \frac{1}{h\sqrt{\log n}}\right)$.

For $\bar{\ell}_0, \bar{\ell}_1, \bar{R}_1$, we can repeat the arguments so far for regular Hamiltonian curves starting from $\overline{\mathcal{M}_\rho}$, in which $\|\alpha\|^2_{g(\gamma(t))^{-1}}$ is within a constant factor of M_1. Therefore, these three parameters also have the same bounds in Lemma 45, 46 and 49 up to a multiplicative constant factor.

D.4. Convergence rate of RHMC with numerical integrators

Now that we estimated all the parameters, we can put them together and state the mixing time of RHMC discretized by a sensitive numerical integrator.

Theorem 2 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S\subset\mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T steps of RHMC discretized by a sensitive integrator on $\mathcal{M}$. For any $\varepsilon > 0$, if for $x \in \mathcal{M}_{\frac{\varepsilon}{2\Lambda}}$ and $v \in \mathbb{R}^n$ randomly drawn from $\mathcal{N}(0, g(x))$, we have that with probability at least 0.99, step size $h \le h_0(x, v)$,*

$$h \le \frac{10^{-20}}{n^{7/12}\log^{1/2}\frac{\Lambda}{\varepsilon}},\ hC_x(x,v) \le \frac{10^{-20}}{\sqrt{n}},\ h^2C_x(x,v) \le \frac{10^{-10}}{n\log\frac{\Lambda}{\varepsilon}} \text{ and } h^2C_v(x,v) \le \frac{10^{-10}}{\sqrt{n\log\frac{\Lambda}{\varepsilon}}},$$

then $d_{TV}(\pi_T, \pi) \le \varepsilon$ for $T = O\left(mh^{-2}\log\frac{\Lambda}{\varepsilon}\right)$.

Proof We first note that $V^x_{\text{good}} = \left\{v \in \mathbb{R}^n : \bar{\ell}\left(\text{Ham}_{x,t}\left(g(x)^{-1}v\right)\right) \le 128\right\}$, the measure of which is at least 0.99 by the definition of $\bar{\ell}_0$. We check the conditions on the step size in Theorem 24. Let $\rho = \frac{\varepsilon}{2\Lambda}$. We first bound M_1, M_1^* by $20n^2\log^2\frac{1}{\rho}$ and set M_2 to 0. Substituting these to Lemma 49, 46, 47 and 48, we have

$$\begin{aligned}
\ell_1 &\lesssim h\sqrt{n\log\frac{1}{\rho}} + \frac{1}{h},\\
R_1 &\lesssim n\log\frac{1}{\rho},\\
R_2 &\lesssim n^{3/2}\log\frac{1}{\rho} + h^2n^{5/2}\log^2\frac{1}{\rho} + \frac{\sqrt{n\log\frac{1}{\rho}}}{h} + \frac{\sqrt{n\log n}}{h},\\
R_3 &\lesssim n\sqrt{\log n}\log\frac{1}{\rho} + hn^{7/4}\log^{3/2}\frac{1}{\rho}.
\end{aligned}$$

Due to $h \le \frac{10^{-20}}{n^{7/12}\log^{1/2}\frac{1}{\rho}}$, direct computation leads to $h^2\max\left(R_1, \bar{R}_1\right), h^5R_1^2\ell_1/\ell_0, h^3R_2 + h^2R_3 \lesssim 1$ and $h \lesssim \min\left(1, \frac{\ell_0}{\ell_1}\right)$. The rest of conditions on the step size, $hC_x(x,v) \le \frac{10^{-20}}{\sqrt{n}}$, $h^2C_x(x,v) \le \frac{10^{-10}}{n\log\frac{1}{\rho}}$ and $h^2C_v(x,v) \le \frac{10^{-10}}{\sqrt{n\log\frac{1}{\rho}}}$, guarantee that

$$hC_x(x,v) \le \frac{10^{-20}}{\sqrt{n}},\ h^2C_x(x,v) \le 10^{-10}\min\left(1, \frac{\bar{\ell}_0}{\bar{\ell}_1}, \frac{1}{n+\sqrt{M_1}+\sqrt{M_1^*}}\right),\ h^2C_v(x,v) \le \frac{10^{-10}}{\sqrt{n+\sqrt{M_1}}}.$$

As the isoperimetry is lower bounded by $\frac{1}{\sqrt{m}}$, Theorem 24 results in the mixing time of $T = O\left(mh^{-2}\log\frac{\Lambda}{\varepsilon}\right)$ that ensures $d_{\text{TV}}(\pi_T, \pi) \le \varepsilon$. ■

By setting C_x, C_v to 0, we can obtain the mixing time of the ideal RHMC for exponential densities in polytopes.

Corollary 3 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of the ideal RHMC on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{7/12} \log^{1/2} \frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^{7/6} \log^2 \frac{\Lambda}{\varepsilon}\right)$ such that $d_{TV}(\pi_T, \pi) \leq \varepsilon$.*

D.4.1. IMPLICIT MIDPOINT METHOD

In the polytope setting, we can explicitly compute $C_x(x, v)$ and $C_v(x, v)$ of IMM in terms of n and ρ.

Lemma 50 *For $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$, let h be step size of IMM with $h^2 n \log \frac{1}{\rho} \leq 10^{-11}$. Then*

$$C_x(x, v) = O\left(n \log \frac{1}{\rho}\right), \; C_v(x, v) = O\left(n^{3/2} \log^{3/2} \frac{1}{\rho}\right).$$

Proof By Lemma 33-2, it follows that

$$C_x(x, v) = O\left(n + \sqrt{M_1}\right) \lesssim n + n \log \frac{1}{\rho} = O\left(n \log \frac{1}{\rho}\right).$$

For $C_v(x, v)$, we first note that $M_2^* = 0$ due to $\nabla^2 f(x) = 0$. Thus by Lemma 33-3, we have

$$C_v(x, v) \lesssim \left(n + \sqrt{M_1}\right)^{3/2} = O\left(n^{3/2} \log^{3/2} \frac{1}{\rho}\right).$$

■

We can also specify a sufficient condition on the step size for the sensitivity of IMM in the polytope setting.

Lemma 51 *For $x \in \mathcal{M}_\rho$, $v \in V^x_{good}$ and step size h with $h^2 n^2 \log \frac{1}{\rho} \leq 10^{-10}$, IMM is sensitive at (x, v).*

Proof Note that $\log \det g(x)$ is convex in $\mathcal{M}$, since the volumetric barrier defined by $\log \det \nabla^2 \phi(x)$ is convex in x (Lemma 1~3 in Vaidya (1996)). Thus, the claim follows from Lemma 34. ■

Substituting the estimates of $C_x(x, v)$ and $C_v(x, v)$ as well as the sufficient condition for the sensitivity to Theorem 2, we prove that the mixing rate of RHMC discretized by IMM for an exponential density in a polytope is independent of the condition number and $\|\alpha\|_2$.

Corollary 4 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of RHMC discretized by IMM on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{3/2} \log \frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ such that $d_{TV}(\pi_T, \pi) \leq \varepsilon$.*

Proof We can check that the step size $h = O\left(\frac{1}{n^{3/2} \log \frac{\Lambda}{\varepsilon}}\right)$ satisfies all the conditions in Theorem 2. Hence, it suffices to choose $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ to obtain $d_{\mathrm{TV}}(\pi_T, \pi) \leq \varepsilon$. ■

D.4.2. GENERALIZED LEAPFROG METHOD

We now compute the mixing rate of RHMC discretized by LM. For LM, we have the same results on $C_x(x,v)$ and $C_v(x,v)$ as IMM.

Lemma 52 *For $x \in \mathcal{M}_\rho$ and $v \in V^x_{good}$, let h be step size of LM with $h^2 n \log \frac{1}{\rho} \leq 10^{-10}$. Then*

$$C_x(x,v) = O\left(n \log \frac{1}{\rho}\right),\ C_v(x,v) = O\left(n^{3/2} \log^{3/2} \frac{1}{\rho}\right).$$

For the sensitivity, LM requires a slightly stronger condition on step size compared to IMM, which follows from Lemma 36.

Lemma 53 *For $x \in \mathcal{M}_\rho$, $v \in V^x_{good}$ and step size h with $h^2 n^3 \log \frac{1}{\rho} \leq 10^{-20}$, LM is sensitive at (x,v).*

We prove that the mixing rate of RHMC discretized by LM for an exponential density in a polytope with m constraints is also independent of the condition number.

Corollary 5 *Let π be a target distribution on a polytope with m constraints in $\mathbb{R}^n$ such that $\frac{d\pi}{dx} \sim e^{-\alpha^\top x}$ for $\alpha \in \mathbb{R}^n$. Let $\mathcal{M}$ be the Hessian manifold of the polytope induced by the logarithmic barrier of the polytope. Let $\Lambda = \sup_{S \subset \mathcal{M}} \frac{\pi_0(S)}{\pi(S)}$ be the warmness of the initial distribution π_0. Let π_T be the distribution obtained after T iterations of RHMC discretized by LM on $\mathcal{M}$. For any $\varepsilon > 0$ and step size $h = O\left(\frac{1}{n^{3/2} \log \frac{\Lambda}{\varepsilon}}\right)$, there exists $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ such that $d_{TV}(\pi_T, \pi) \leq \varepsilon$.*

Proof For step size $h = O\left(\frac{1}{n^{3/2} \log \frac{\Lambda}{\varepsilon}}\right)$, LM is sensitive in $\mathcal{M}_\rho \times V_1^c$ by Lemma 36, and we can use the estimates of C_x and C_v proven in Lemma 52. Thus, this step size satisfies all the conditions in Theorem 2. Hence, it suffices to choose $T = O\left(mn^3 \log^3 \frac{\Lambda}{\varepsilon}\right)$ to obtain $d_{\mathrm{TV}}(\pi_T, \pi) \leq \varepsilon$. ■

Appendix E. Definitions

Definition 54 *(Self-concordant barrier) A self-concordant barrier $\phi : K \subset \mathbb{R}^n \to \mathbb{R}$ is a function such that $\phi(x) \to \infty$ as $x \to \partial K$ and that $\left|Df^3(x)[h,h,h]\right| \leq 2\left(D^2 f(x)[h,h]\right)^{3/2}$ for all $x \in K$ and $h \in \mathbb{R}^n$. If $\left|Df^4(x)[h,h,h,h]\right| \leq 6\left(D^2 f(x)[h,h]\right)^2$ is also satisfied for all h, then ϕ is called a highly self-concordant barrier.*

Definition 55 *(Self-concordance parameter) For a self-concordant function ϕ, the self-concordance parameter of ϕ is the smallest non-negative real number ν_ϕ such that*

$$|D\phi(x)[h]|^2 \leq \nu_\phi D^2 \phi(x)[h,h],$$

where $Df(x)[h]$ is the directional derivative of f along direction h and $D^2 f(x)[h_1,h_2]$ is the second-order directional derivative of f along directions h_1 and h_2.

Definition 56 *(Riemannian length and distance) Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a self-concordant function. For all $x \in \mathbb{R}^d$, we define the local norm induced by $\nabla^2\phi(x)$ by*

$$\|h\|_{\nabla^2\phi(x)} = \sqrt{h^\top \nabla^2\phi(x) h}.$$

For any smooth curve $c : [0,1] \to \mathbb{R}^n$, we define the length of the curve as

$$L_\phi(c) = \int_0^1 \left\| \frac{d}{dt} c(t) \right\|_{\nabla^2\phi(c(t))} dt.$$

For any $x, y \in \mathbb{R}^d$, we define the distance $d_\phi(x,y)$ to be the infimum of the lengths of all piecewise smooth curves with $c(0) = x$ and $c(1) = y$.

Definition 57 *(Total variation distance) For probability distributions P and Q supported on K, the total variation distance (TV distance) is defined by*

$$d_{TV}(P,Q) = \sup_{A \subset K} \left(P(A) - Q(A) \right).$$

Appendix F. Lemmas

Lemma 58 *For $n \in \mathbb{N}$ and matrix $X \in \mathbb{R}^{2n \times 2n}$ of the form*

$$X = \begin{bmatrix} C & I_n \\ -C^2 + R & -C \end{bmatrix}$$

with a symmetric matrix $C \in \mathbb{R}^{n \times n}$ and matrix $R \in \mathbb{R}^{n \times n}$, we have

$$X^{2n} = \begin{bmatrix} R^n & 0 \\ R^n C - C R^n & R^n \end{bmatrix},$$
$$X^{2n+1} = \begin{bmatrix} R^n C & R^n \\ R^{n+1} - C R^n C & -C R^n \end{bmatrix}.$$

The claim immediately follows from induction.

Lemma 59 (Lee and Vempala (2018), Lemma 7) *In the Euclidean coordinate, the Hamiltonian equations in (2.1) can be represented via the second-order ODE as follows:*

$$D_t \frac{dx}{dt} = \mu(x),$$
$$\frac{dx}{dt}(0) \sim \mathcal{N}(0, g(x)^{-1}),$$

where D_t is the covariant derivative along the Hamiltonian trajectory $x(t)$ and $\mu(x) \stackrel{\text{def}}{=} -g(x)^{-1}\nabla f(x) - \frac{1}{2} g(x)^{-1} \mathrm{Tr}\left[g(x)^{-1} D g(x) \right]$.

Lemma 60 (Lee and Vempala (2018), Lemma 64) *For matrix $E \in \mathbb{R}^{n \times n}$ with $\|E\|_2 < \frac{1}{4}$, we have*

$$\left| \log\det(I + E) - \mathrm{Tr} E \right| \le \|E\|_F^2.$$

Lemma 61 (Kalai and Vempala (2006), Lemma 4.1) *For a unit vector $c \in \mathbb{R}^n$, constant T and convex set $K \subset \mathbb{R}^n$, we have*

$$\mathbb{E}_{x\sim\pi}\left[c^\top x\right] \leq nT + \min_{x\in K} c^\top x,$$

where π is a probability density proportional to $e^{-\frac{c^\top x}{T}}$.

Lemma 62 (Nguyen (2013), Corollary 6) *Let π be a log-concave density proportional to $\exp(-V)$ on $\mathbb{R}^n$. Then,*

$$Var_{x\sim\pi}\left(V(x)\right) \leq n.$$

Lemma 63 (Lovász and Vempala (2007), Lemma 5.17) *Let $X \in \mathbb{R}^n$ be randomly chosen from a log-concave distribution. Then for any $R > 1$,*

$$\mathbf{P}\left(|X| > R\sqrt{\mathbb{E}X^2}\right) < e^{-R+1}.$$

Lemma 64 (Nesterov et al. (2002), Lemma 3.1) *Suppose $\phi : \mathbb{R}^n \to \mathbb{R}$ is self-concordant and $\mathcal{K} \subset \mathbb{R}^n$ is convex. For any $x, y \in \mathcal{K}$, i*

- *If $d_\phi(x, y) \leq \delta - \delta^2 < 1$ for some $0 < \delta < 1$, then $\|y - x\|_{\nabla^2\phi(x)} \leq \delta$.*
- *If $\delta = \|x - y\|_{\nabla^2\phi(x)} < 1$, then $\delta - \frac{1}{2}\delta^2 \leq d_\phi(x, y) \leq -\log(1 - \delta)$.*