Mixing Times of MCMC algorithms

Abstract

We study the dependence of mixing times of different MCMC algorithms on dimension and error tolerance, when the target is a multivariate Gaussian. We numerically verify the theoretical results for dependence of mixing time vs dimension.

1 Introduction

Markov chain Monte Carlo methods provide an expeditious way of sampling from known distributions. The behaviour of these methods in high dimensions is of great interest because of potential applications in a plethora of fields such as statistics, operations research, computational materials science, etc.
In this article we study the dependence of mixing times of different MCMC algorithms on the dimension and error tolerance.

2 Theory

2.1 Mixing Time

Let (π,ν) denote a Markov chain with kernel π and initial distribution ν with state space $\mathbb{R}^d$. Assume μ is the unique invariant distribution of the chain.
The **mixing time** is then defined as,
$T^{\nu}_{mix}(\delta) := min\{n : \|\nu\pi^n - \mu\|_{TV} < \delta\}$
where $\|.\|_{TV}$ is the total variation norm.
When the initial distribution is clear from the context we abuse notation and write $T^{\nu}_{mix}(\delta) := T_{mix}(\delta)$

2.2 Warm start

Let μ and ν be measures on $\mathbb{R}^d$.
ν is said to be β**-warm** with respect to μ, if there exists a $\beta \geq 0$ such that,

$$\sup_A \left(\frac{\nu(A)}{\mu(A)}\right) \leq \beta$$

where the supremum is taken over all measurable sets $A \subset \mathbb{R}^d$.
Let π denote a Markov kernel on $\mathbb{R}^d$ with a unique invariant distribution μ. We say that a measure ν is β-warm with respect to π if ν is β-warm with respect to μ.

Lemma 1 Let μ, ν be measures on $\mathbb{R}^d$. Let dx denote the Lebesgue measure on $\mathbb{R}^d$. Assume μ and ν are absolutely continuous with respect to dx with Radon-Nikodym derivatives $f := \frac{d\mu}{dx}$ and $g := \frac{d\nu}{dx}$. Also assume $g(x) > 0$, for all $x \in \mathbb{R}^d$. If there exists $\beta \geq 0$ such that,

$$\sup_{x \in \mathbb{R}^d} \frac{f(x)}{g(x)} \leq \beta$$

Then μ is β-warm with respect to ν.

Proof Let A be a measurable subset of $\mathbb{R}^d$ with positive measure.
By the definition of Radon-Nikodym derivative,
$\mu(A) = \int_A f dx$ and $\nu(A) = \int_A g dx$.
But, $\mu(A) = \int_A f dx \leq \int_A \beta g dx$.
Dividing by $\int_A g dx > 0$ on both sides.
We get $\dfrac{\int_A f dx}{\int_A g dx} \leq \beta$

□

Theorem 1 *Let μ be a d dimensional multivariate Gaussian measure with 0 mean and the covariance matrix, Σ, a $d \times d$ diagonal matrix as follows,*

$$\Sigma = \begin{pmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & a_d \end{pmatrix}$$

such that for all $i \leq d$, $1 \leq a_i$. Let ν be a d dimensional multivariate Gaussian measure with 0 mean and the covariance matrix I_d.
Then there exists $\beta \geq 0$ such that ν is β-warm with respect to μ. Furthermore, β can be chosen to be $\det(\Sigma)^{1/2}$.

Proof We wish to use 1 and thus fix $x \in \mathbb{R}^d$.

$$f(x) := \frac{d\mu}{dx}(x) = (2\pi)^{-d/2} exp(-\frac{\|x\|^2}{2})$$

and

$$g(x) := \frac{d\nu}{dx}(x) = (2\pi)^{-d/2} exp(-\frac{x^T \Sigma^{-1} x}{2})(\prod_{i=1}^{d} a_i)^{-1/2}$$

$$\frac{f(x)}{g(x)} = exp(-\frac{1}{2}(\sum_{i=1}^{d} x_i^2(1-\frac{1}{a_i}))(\prod_{i=1}^{d} a_i)^{1/2}$$

Since for all $i \leq d$, $a_i \geq 1$ we have,

$$exp(-\frac{1}{2}(\sum_{i=1}^{d} x_i^2(1-\frac{1}{a_i})) \leq 1$$

Thus we see that,

$$\frac{f(x)}{g(x)} \leq (\prod_{i=1}^{d} a_i)^{1/2} = \det(\Sigma)^{1/2}$$

Since x was arbitrary, we observe for $\beta := \det(\Sigma)^{1/2}$, ν is β-warm with respect to μ. □

2.3 ULA mixing time

We derive the mixing time for ULA with a Gaussian target and a standard Gaussian initial distribution.

Theorem 2 *Let μ be a d dimensional multivariate Gaussian measure with 0 mean and the covariance matrix, Σ, a $d \times d$ diagonal matrix as follows,*

$$\Sigma = \begin{pmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & a_d \end{pmatrix}$$

such that for all $i \leq d$, $1 \leq a_i \leq C$, let π_h be the transition kernel for ULA, with a fixed stepsize $h \leq 1$ and the target measure μ. Let ν be a standard d dimensional Gaussian measure. Then,

$$\|\nu\pi^n - \mu\|_{TV} \leq \sqrt{d}h(3/2)(1 + 3h^{-1}(1-h/C)^{2n}) \tag{1}$$

Proof : The measure μ is absolutely continuous with respect to the Lebesgue measure. For all $x \in \mathbb{R}^d$,

$$\frac{d\mu}{dx}(x) = exp(-U(x))C$$

where

$$U(x) := (1/2)x^T\Sigma^{-1}x \text{ and } c \in \mathbb{R}$$

Let $\nabla : \mathbf{C}^\infty(\mathbb{R}^d) \to (\mathbf{C}^\infty(\mathbb{R}^d))^d$ be the gradient operator, then $\nabla : f \mapsto \left(\frac{\partial f}{\partial x_i}\right)_{i=1}^{d}$
We wish to find $\nabla U(x)$, as it is used in the ULA transition step.
We compute $\nabla U(x) = \Sigma^{-1}x$.
Let $(X_0, X_1, ...)$ denote the steps of a Markov chain with transition kernel π_h and initial distribution ν.
We first compute the distribution of X_n.

We know that the ULA transition step is given as,

$$X_{n+1} = X_n - h\nabla U(X_n) + \sqrt{2h}Z_{n+1} = (I - h\Sigma^{-1})X_n + \sqrt{2h}Z_{n+1} \tag{2}$$

where for all i $\in \mathbb{N}$, Z_i are iid d dimensional standard Gaussian random variables such that for all $k \in \mathbb{N}$, Z_i is independent of X_k.
Since sums of independent Gaussians is a Gaussian and a linear transformation of a Gaussian is a Gaussian, we see from (2) that if X_n is a Gaussian, then so is X_{n+1}.
Since $X_0 \sim \nu$ is Gaussian, we get that for all $n \in \mathbb{N}$, X_n is a multivariate Gaussian.
Further, by taking expectation on both sides of the equality in (2), we see that for all $n \in \mathbb{N}$, $E[X_{n+1}] = E[X_n]$.
Hence for all $n \in \mathbb{N}$, $E[X_n] = E[X_0] = (0, 0, ..0)$.
Thus, $X_n \sim N(0, C_n)$, where $C_n := Var(X_n)$.
By taking the variance on both sides of the equality in (2), we see that,

$$C_{n+1} = AC_nA^t + 2hI \tag{3}$$

where $A := (I - h\Sigma^{-1})$ and $C_0 = I$.
Note that in (3), if C_n is a diagonal matrix then so is C_{n+1}.
Since $C_0 = I$, we see that for all $n \in \mathbb{N}$, C_n is a diagonal matrix.
After solving the recurrence relation in (3), we see that,

$$C_n = A^{2n} + 2h(I - A^{2n})[(I - A^2)]^{-1})$$

Thus, C_n is a d dimensional diagonal matrix with,

$$(C_n)_{ii} = (1 - h/a_i)^{2n} + 2h\frac{1 - (1 - h/a_i)^{2n}}{1 - (1 - h/a_i)^2}$$

Noting that $\frac{2h}{(h/a_i)(2 - h/a_i)} = \frac{a_i}{1 - h/2a_i}$, we get,

$$(C_n)_{ii} = (1 - h/a_i)^{2n} + a_i\frac{1 - (1 - h/a_i)^{2n}}{1 - h/2a_i}$$

Now that we know that X_n is a Gaussian random variable with a simple covariance matrix, we can compute an upper bound on the TV distance between the target distribution and the distribution of X_n.
According to [1, Thm 1.1], the TV distance between two d dimensional Gaussians, $N(0, \Sigma_1)$ and $N(0, \Sigma_2)$ is bounded by the L^2 norm of the eigenvalues of $\Sigma_1^{-1}\Sigma_2 - I$ times $\frac{3}{2}$.
Taking Σ as Σ_1 and C_n as Σ_2, we get,

$$\begin{aligned}\|\nu\pi^n - \mu\|_{TV} &\leq (3/2)\sqrt{\sum_{i=1}^{d}[(1 - h/a_i)^{2n}/a_i + \frac{1 - (1 - h/a_i)^{2n}}{1 - h/2a_i} - 1]^2} \\ &= (3/2)\sqrt{\sum_{i=1}^{d}[\frac{h/2a_i}{1 - h/2a_i} + (1 - h/a_i)^{2n}[1/a_i - 1/(1 - h/2a_i)]]^2}\end{aligned} \tag{4}$$

We know that,

$$\begin{aligned}&\left|\frac{h/2a_i}{1 - h/2a_i} + (1 - h/a_i)^{2n}[1/a_i - 1/(1 - h/2a_i)]\right| \\ &\leq \frac{h/2a_i}{1 - h/2a_i} + (1 - h/a_i)^{2n}[1/a_i + 1/(1 - h/2a_i)]\end{aligned}$$

Noting that,

$$1/a_i \leq 1,$$
$$h \leq 1 \implies (1-h/2a_i)^{-1} \leq (1-h/2)^{-1} \leq 2,$$
$$(1-h/a_i) \leq (1-h/C),$$

we see that,

$$\frac{h/2a_i}{1-h/2a_i} \leq 1$$

and

$$1/a_i + 1/(1-h/2a_i) \leq 3$$

we then observe that,

$$\frac{h/2a_i}{1-h/2a_i} + (1-h/a_i)^{2n}[1/a_i + 1/(1-h/2a_i)] \leq h + 3(1-h/C)^{2n}$$

Thus,

$$\left|[\frac{h/2a_i}{1-h/2a_i} + (1-h/a_i)^{2n}[1/a_i - 1/(1-h/2a_i)]]\right|^2 \leq [h + 3(1-h/C)^{2n}]^2$$

$$\sum_{i=1}^{d}[\frac{h/2a_i}{1-h/2a_i} + (1-h/a_i)^{2n}[1/a_i - 1/(1-h/2a_i)]]^2 \leq d[h + 3(1-h/C)^{2n}]^2$$

Taking the square root and multiplying by (3/2) we see,

$$(4) \leq (3/2)h\sqrt{d}(1 + 3h^{-1}(1-h/C)^{2n})$$

□

From (20), we see that in order to be δ-close to our target measure, we can choose

$$h = (1/6)d^{-(\frac{1}{2}+a)}\delta^{1+b} \tag{5}$$

where $a, b \geq 0$, and we get δ-accuracy if $h^{-1}(1-h/C)^{2n} \leq 1$, therefore,

$$T_{mix}(\delta) \leq \lceil (1/2)ln(h^{-1})/ln((1-h/C)^{-1}) \rceil$$

$$\in O(h^{-1}ln(h^{-1})) = O(d^{\frac{1}{2}+a}\delta^{-(1+b)}ln(d\delta^{-1})) \tag{6}$$

2.4 MRW and MALA mixing time

From [2], we cite mixing time results for the Metropolised random walk(MRW) and Metropolis adjusted Langevin algorithm(MALA).
Let μ be the target measure on $\mathbb{R}^d$ which is absolutely continuous with respect to the Lebesgue measure with density, $\frac{d\mu}{dx}(x) \propto \exp(-U(x))$, for $x \in \mathbb{R}^d$ and U, a smooth function from $\mathbb{R}^d \to \mathbb{R}$.
Further, assume there exists constants $m \geq 0$ and $L \geq 0$, such that

$$m \leq \|Hess(U(x))\|_{op} \leq L$$

for all $x \in \mathbb{R}^d$

where $Hess : C^\infty(\mathbb{R}^d) \to \mathbb{M}_{d\times d}(C^\infty(\mathbb{R}^d))$, $f \mapsto \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j}$ and

$\|.\|_{op}$: $\mathbb{M}_{d\times d}(\mathbb{R}) \to \mathbb{R}_+$, $A \mapsto \sup_{x\neq 0}\left(\frac{\|Ax\|_2}{\|x\|_2}\right)$,

where $\|.\|_2$ denotes the L^2 norm on $\mathbb{R}^d$ and $\mathbb{M}_{m\times n}(R)$ denotes the set of all $m \times n$ matrices over a ring R.

We define, r and w which will be useful providing bounds on the mixing time,

$$r(s) := 2 + 2\max\left\{\frac{1}{d^{0.25}} log^{0.25}\left(\frac{1}{s}\right), \frac{1}{d^{0.5}} log^{0.5}\left(\frac{1}{s}\right)\right\}$$

$$w(s) := \min\left\{\frac{\sqrt{m}}{r(s)L\sqrt{dL}}, \frac{1}{Ld}\right\}, \text{ for } s \in \left(0, \frac{1}{2}\right)$$

Note that

$$r(s) \leq 4 \text{ for } s \geq e^{-d} \tag{7}$$

From [2, Theorem 1], we have the following theorem,

Theorem 3 *For any β-warm initial distribution ν and any error tolerance $\delta \in (0, 1]$, the Metropolis adjusted Langevin algorithm with target μ, stepsize $h = cw(\delta/(2\beta))$ and transition kernel π_h satisfies the bound $\|\nu\pi_h^k - \mu\|_{TV} \leq \delta$ for all iteration numbers,*

$$k \geq c' \log\left(\frac{2\beta}{\delta}\right) \max\left\{d\kappa, d^{0.5}\kappa^{1.5} r\left(\frac{\delta}{2\beta}\right)\right\}$$

where c' is a universal constant in $\mathbb{R}$.

From [2, Theorem 2], we have the following theorem,

Theorem 4 *For any β-warm initial distribution ν and any error tolerance $\delta \in (0, 1]$,the Metropolised random walk with target μ, step size $h = \frac{cm}{dL^2 r(\delta/2\beta)}$ and transition kernel π_h, satisfies the bound $\|\nu\pi_h^k - \mu\|_{TV} \leq \delta$ for all iteration numbers,*

$$k \geq c' d\kappa^2 r\left(\frac{\delta}{2\beta}\right) \log\left(\frac{2\beta}{\delta}\right)$$

where c', c are universal constants in $\mathbb{R}$.

2.5 Unadjusted Hamiltonian Monte Carlo

We plan to derive the mixing time of unadjusted Hamiltonian Monte Carlo (uHMC). Before doing so, we recall the Hamiltonian dynamics.

2.5.1 Hamiltonian dynamics

Let $U \in \mathbf{C}^2(\mathbb{R}^{2d})$ and $U(x) \geq 0$, for all $x \in \mathbb{R}^d$.

Let X_t and V_t denote the position and velocity of a particle with unit mass

at time $t \geq 0$ in $\mathbb{R}^d$.
Hamiltonian dynamics with a potential U, initial state x_0 and velocity v_0, gives us the following set of differential equations :-

$$\begin{aligned} \dot{X}_t &= V_t \\ \dot{V}_t &= -\nabla U(X_t) \\ X_0 &= x_0 \\ V_0 &= v_0 \end{aligned}$$

Let Θ_T:$\mathbb{R}^{2d} \to \mathbb{R}^{2d}$ be the flow associated with the Hamiltonian dynamics, i.e $\Theta_T(x, v) = (X_T, V_T)$, where the initial state is x and initial velocity is v.
For most functions U, Θ_T cannot be solved analytically. Thus when trying to simulate the dynamics on a computer, we use a numerical approximation of Θ_T. We now describe how to approximate Θ_T using a (velocity) Verlet scheme. Let h be the stepsize and assume $\frac{T}{h} \in \mathbb{N}$.
Define $\phi_h(x, v) := (x', v')$ with,

$$x' = x + hv - \frac{h^2}{2}\nabla U(x) \tag{8}$$

$$v' = v - \frac{h}{2}\nabla U(x) - \frac{h}{2}\nabla U(x').$$

Then $\Theta_T(x, v)$ is approximated by $\phi_h^k := \phi_h \circ \phi_h ... \circ \phi_h$ - evaluated k times, where $k := \frac{T}{h} \in \mathbb{N}$.
We define τ_h

2.5.2 uHMC algorithm

We recall the algorithm for the Unadjusted Hamiltonian Monte Carlo method with a deterministic termination time $T > 0$. This is directly taken from [3, Algorithm 5].

Algorithm 1 Unadjusted Hamiltonian Monte Carlo (uHMC)

Input: Probability measure ν on $\mathbb{R}^d$, friction parameter $\gamma \in [0, 1]$, stepsize $h \geq 0$, termination time T, with $\frac{T}{h} \in \mathbb{Z}$ and number of steps $N \in \mathbb{N}$.
Output: Samples $\{(X_n, V_n)\}_{n=0}^N$, from a Markov chain with initial law $\nu \otimes \mathcal{N}(0, I_d)$.

$n \leftarrow 0$; $X_0 \leftarrow$ Sample(ν); $V_0 \leftarrow$ Sample($\mathcal{N}(0, I_d)$);
$K \leftarrow \frac{T}{h}$;
while $n < N$ **do**
    $Z_{n+1} \leftarrow$ Sample($\mathcal{N}(0, I_d)$);
    $(X_{n+1}, V_{n+1}) \leftarrow \phi_h^K(X_n, e^{-\gamma T} V_n + \sqrt{1 - e^{-2\gamma T}} Z_{n+1})$;
    $n \leftarrow n + 1$;
end while
Return $((X_0, V_0), (X_1, V_1), ..., (X_N, V_N))$

We wish to derive the mixing time for the unadjusted Hamiltonian Monte Carlo (uHMC) with velocity refreshments when the target is a Gaussian and the initial distribution is a standard Gaussian.

2.5.3 uHMC mixing time - coupling based argument

We wish to derive upper bounds on the mixing time for the unadjusted Hamiltonian Monte Carlo (uHMC) with velocity refreshments when the target is a Gaussian.
We assume the target density is proportional to $\exp(-U(x))$, where $U(x) := (1/2)x^T\Sigma^{-1}x$, where Σ is diagonal, with $a_i := \Sigma_{ii}$. Then the distribution that uHMC aims to sample from is the Boltzmann-Gibbs distribution on phase space, $\mathbb{R}^{2d}$ and is given by $N(0,\Sigma) \otimes N(0, I_d)$.
Let π be the u-HMC transition kernel with target as above and hyperparameters h, γ and k as the stepsize, friction coefficient and number of leapfrog steps taken respectively.
(k is usually $\frac{T}{h}$, where T is the termination time)
We wish to show that $\pi : \mathcal{P}(\mathbb{R}^{2d}) \to \mathcal{P}(\mathbb{R}^{2d})$ acts as a contraction with respect to the TV norm.
By Kantorovich-Rubenstein duality, it is enough to show, that there exists $c < 1$ such that,

$$\|\pi((x,v),.) - \pi((y,w),.)\|_{TV} \leq c, \text{ for all } (x,v) \in \mathbb{R}^{2d}, (y,w) \in \mathbb{R}^{2d} \quad (9)$$

For $x, v, g \in \mathbb{R}^d$, we define,

$$\Theta(x,v,g) := \phi_h^k(x, e^{-\gamma T}v + \sqrt{1-e^{-2\gamma T}}g)$$

where $\phi_h : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ as in (8). Assuming we start at (x,v), the updated state (x',v'), after one step of the uHMC algorithm is then given by,

$$(x',v') = \Theta(x,v,Z), \text{ where } Z \sim \mathcal{N}(0, I_d),$$

Noting that since $\nabla U(x) = \Sigma^{-1}x$, we see that ϕ_h acts as a linear map on the phase space with its matrix given as,

$$\phi_h = \begin{pmatrix} (I_d - \frac{h^2}{2}\Sigma^{-1}) & hI_d \\ -h\Sigma^{-1}\left(1 - \frac{h^2}{4}\Sigma^{-1}\right) & (1 - \frac{h^2}{2}\Sigma^{-1}) \end{pmatrix}$$

Noting that

$$\Theta(x,v,g) = \phi_h^k(\begin{pmatrix} I_d & 0 \\ 0 & \exp(-\gamma T) \end{pmatrix} I_d \begin{pmatrix} x \\ v \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{1-\exp(-2\gamma T)I_d} \end{pmatrix} \begin{pmatrix} 0 \\ g \end{pmatrix})$$

We get that,

$$(x', v') \sim N(\phi_h^k(\begin{pmatrix} I_d & 0_d \\ 0_d & \exp(-\gamma T) I_d \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix}), (1 - \exp(-2\gamma T))\phi_h^k P(\phi_h^k)^t) \tag{10}$$

where $P := \begin{pmatrix} 0 & 0 \\ 0 & I_d \end{pmatrix}$, be the projection matrix.
We couple two chains starting from (x, v) and (y, w) in $\mathbb{R}^{2d}$ with the same noise, i.e for $(Z_n)_{n\in\mathbb{N}}$ being a sequence of iid standard d dimensional Gaussian random variables, we have,

$$\forall n \ (x_{n+1}, v_{n+1}) := \Theta(x_n, v_n, Z_n), \ (y_{n+1}, w_{n+1}) := \Theta(x_n, v_n, Z_n)$$

$$(x_0, v_0) := (x, v), \ (y_0, w_0) := (y, w),$$

From (10), we get that,

$$(x_1, v_1) \sim N(\phi_h^k(\begin{pmatrix} I_d & 0_d \\ 0_d & \exp(-\gamma T) I_d \end{pmatrix}\begin{pmatrix} x_0 \\ v_0 \end{pmatrix}), (1 - \exp(-2\gamma T))\phi_h^k P(\phi_h^k)^t) \tag{11}$$

and

$$(y_1, w_1) \sim N(\phi_h^k(\begin{pmatrix} I_d & 0_d \\ 0_d & \exp(-\gamma T) I_d \end{pmatrix}\begin{pmatrix} y_0 \\ w_0 \end{pmatrix}), (1 - \exp(-2\gamma T))\phi_h^k P(\phi_h^k)^t) \tag{12}$$

Note:
(x_1, v_1) and (y_1, w_1) are multivariate Gaussians with different means and the same singular covariance matrix. This means there exists a vector $l \in \mathbb{R}^{2d}, l \neq 0$ such that $< (x_1, v_1), l >=< Mean((x_1, v_1)), l >$, almost surely and $< (y_1, w_1), l >=< Mean((y_1, w_1)), l >$, almost surely. So when projected to the vector l the random variables, (x_1, v_1) and (y_1, w_1) will be almost surely not equal to one another. So we can't employ the parallel coupling to find an upper bound on the total variation mixing time. We note some helpful computations in the next para.

2.5.4 uHMC- 1 step distribution

We now assume that the dimension $d = 1$ and the target density is $\mathcal{N}(0, a)$. Recall, the uHMC transition step is given as

$$(x, v) \leftarrow \phi_h^k(\begin{pmatrix} 1 & 0 \\ 0 & \exp(-\gamma T) \end{pmatrix}\begin{pmatrix} x \\ v \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{1-\exp(-2\gamma T)}1 \end{pmatrix}\begin{pmatrix} 0 \\ Z \end{pmatrix})$$

where $Z \sim \mathcal{N}(0, 1)$, γ is the friction coefficient and

$$\phi_h := \begin{pmatrix} 1 - \frac{h^2}{2a} & h \\ -\frac{h}{a}\left(1 - \frac{h^2}{4a}\right) & 1 - \frac{h^2}{2a} \end{pmatrix}$$

In order to enable easy computation of ϕ_h^k, following [4, 4.9], we find θ_h and χ_h such that ϕ_h can be written as,

$$\phi_h = \begin{pmatrix} cos(\theta_h) & \chi_h sin(\theta_h) \\ -\chi_h^{-1} sin(\theta_h) & cos(\theta_h) \end{pmatrix}$$

where $\theta_h \in (0, \pi)$ chosen such that $\cos(\theta_h) = 1 - \frac{h^2}{2a}$, then $\sin(\theta_h) = \frac{h\sqrt{4a - h^2}}{2a}$. Then $\chi_h = \frac{2a}{\sqrt{4a - h^2}}$. We then see that,

$$\phi_h^k = \begin{pmatrix} cos(k\theta_h) & \chi_h sin(k\theta_h) \\ -\chi_h^{-1} sin(k\theta_h) & cos(k\theta_h) \end{pmatrix}$$

We compute the mean of (x_1, v_1) which is given by,

$$\phi_h^k(\begin{pmatrix} 1 & 0 \\ 0 & \exp(-\gamma T) \end{pmatrix}\begin{pmatrix} x_0 \\ v_0 \end{pmatrix}) = \begin{pmatrix} x_0 \cos(k\theta) + v_0 \exp(-\gamma T)\chi_h \sin(k\theta) \\ -x_0\chi_h^{-1} \sin(k\theta) + v_0 \exp(-\gamma T) \cos(k\theta) \end{pmatrix} \quad (13)$$

We compute the covariance matrix of (x_1, v_1) which is given by,

$$\phi_h^k P(\phi_h^k)^t = \begin{bmatrix} \chi_h^2 \sin^2(k\theta_h) & \chi_h \sin(k\theta_h)\cos(k\theta_h) \\ \chi_h \sin(k\theta_h)\cos(k\theta_h) & \cos^2(k\theta_h) \end{bmatrix}$$

Note that for $l = (\cos(k\theta_h), -\chi_h \sin(k\theta_h))^t$, we have $\phi_h^k P(\phi_h^k)^t l = 0$.
Thus we get that $l.(x_1, v_1)$ has zero variance (as $\mathrm{Var}(l.(x_1, v_1)) = l^t Cov(x_1, v_1) l = 0$)
Therefore $< (x_1, v_1), l > = < Mean(x_1, v_1), l >$ almost surely and the value is computed by taking the dot product of the RHS of (13) with l, it is given by x_0.
So if we have two different copies of HMC starting at (x, v) and (y, w) respectively and $x \neq y$, the TV distance between the distributions in phase space i.e of (x_1, v_1) and (y_1, w_1), after 1 step will always be 1.
This is true since,

$$\mathcal{P}((x_1, v_1) = (y_1, w_1)) \leq \mathcal{P}(< (x_1, v_1), l > = < (y_1, w_1), l >) = \mathcal{P}(x = y) = 0,$$

regardless of the coupling between (x_1, v_1) and (y_1, w_1).
So, we cannot hope to get a contraction in phase space.

2.6 Computing the transition kernel for uHMC in position space

(8) gives the u-HMC transition step in phase space.
We want to compute the transition kernel of uHMC in position space, with full velocity randomization and target density $\mathcal{N}(0, \Sigma)$, where Σ is a $d \times d$ diagonal matrix as follows

$$\Sigma = \begin{pmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & a_d \end{pmatrix}$$

The uHMC transition step in position space is given by

$$x \leftarrow P\phi_h^k(x, Z)$$

where $x \in \mathbb{R}^d, Z \sim \mathcal{N}(0, I_d), \phi_h$ as in (8) and $P : \mathbb{R}^{2d} \to \mathbb{R}^d$, $(x, v) \mapsto x$ is the projection matrix from phase space to position space.
We now compute $P\phi_h^k(x, Z)$.
Since $\nabla U(x) = \Sigma^{-1}x$, where the target measure $\mu(dx) \propto e^{-U(x)}dx$, we see that ϕ_h acts as a linear map on the phase space, with its matrix in the standard basis (i.e position coordinates followed by velocity coordinates, $(x_1, x_2, .., x_d, v_1, v_2, ..., v_d)$) given as,

$$\phi_h = \begin{pmatrix} (I_d - \frac{h^2}{2}\Sigma^{-1}) & hI_d \\ -h\Sigma^{-1}\left(1 - \frac{h^2}{4}\Sigma^{-1}\right) & (1 - \frac{h^2}{2}\Sigma^{-1}) \end{pmatrix}$$

In order to enable easy computation of ϕ_h^k, we reorder the basis vectors of ϕ_h from $(x_1, x_2, .., x_d, v_1, v_2, .., v_d)$ to $(x_1, v_1, x_2, v_2, .., x_d, v_d)$, the matrix for ϕ_h is now a block diagonal with entries as below

$$\phi_h = \begin{pmatrix} 1 - \frac{h^2}{2a_1} & h & 0 & 0 & \cdots & 0 & 0 \\ -\frac{h}{a_1}(1 - \frac{h^2}{4a_1}) & 1 - \frac{h^2}{2a_1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 - \frac{h^2}{2a_2} & h & \cdots & 0 & 0 \\ 0 & 0 & -\frac{h}{a_2}(1 - \frac{h^2}{4a_2}) & 1 - \frac{h^2}{2a_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 - \frac{h^2}{2a_d} & h \\ 0 & 0 & 0 & 0 & \cdots & -\frac{h}{a_d}(1 - \frac{h^2}{4a_d}) & 1 - \frac{h^2}{2a_d} \end{pmatrix}$$

The diagonal blocks are defined for all $i \in \{1, 2, ..d\}$ as

$$A_i = \begin{pmatrix} 1 - \frac{h^2}{2a_i} & h \\ -\frac{h}{a_i}(1 - \frac{h^2}{4a_i}) & 1 - \frac{h^2}{2a_i} \end{pmatrix}$$

In order to enable easy computation of A_i^k, following [4, 4.9], we find θ_i and χ_i such that A_i can be written as,

$$A_i = \begin{pmatrix} cos(\theta_i) & \chi_i sin(\theta_i) \\ -\chi_i^{-1} sin(\theta_i) & cos(\theta_i) \end{pmatrix}$$

where $\theta_i \in (0, \pi)$ chosen such that,

$$\cos(\theta_i) = 1 - \frac{h^2}{2a_i} \tag{14}$$

Then,

$$\sin(\theta_i) = \frac{h\sqrt{4a_i - h^2}}{2a_i} \tag{15}$$

We see that,

$$\chi_i = \frac{2a_i}{\sqrt{4a_i - h^2}} \tag{16}$$

We then see that,

$$A_i^k = \begin{pmatrix} \cos(k\theta_i) & \chi_i \sin(k\theta_i) \\ -\chi_i^{-1} \sin(k\theta_i) & \cos(k\theta_i) \end{pmatrix}$$

Thus

$$\phi_h^k = \begin{pmatrix} A_1^k & 0 & \cdots & 0 \\ 0 & A_2^k & \cdots & 0 \\ 0 & 0 & \cdots & A_d^k \end{pmatrix}$$

The projection $P : \mathbb{R}^{2d} \to \mathbb{R}^d, (x, v) \mapsto x$ in the new $(x_1, v_1, x_2, v_2, ..., x_d, v_d)$ basis is given as

$$P_{d\times 2d} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

We then see that,

$$P\phi_h^k(x_1, v_1, x_2, v_2, ..., x_d, v_d) = (x_1 \cos(k\theta_1) + v_1\chi_1 \sin(k\theta_1), x_2 \cos(k\theta_2) + v_2\chi_2 \sin(k\theta_2), ..., x_d \cos(k\theta) + v_d\chi_d \sin(k\theta_d))$$

This can be seen by noting that,

$$A_i^k \begin{pmatrix} x_i \\ v_i \end{pmatrix} = \begin{pmatrix} x_i \cos(k\theta_i) + v_i \chi_i \sin(k\theta_i) \\ -x_i \chi_i^{-1} \sin(k\theta_i) + v_i \cos(k\theta_i) \end{pmatrix}$$

and P filters out only the odd indexed coordinates.
Finally, we compute the uHMC transition for the position,

$$(x_1, x_2, ..., x_d) \leftarrow (x_1 \cos(k\theta_i) + Z_1 \chi_i \sin(k\theta_i), x_2 \cos(k\theta_i) + Z_2 \chi_i \sin(k\theta_i), ..., x_d \cos(k\theta_i) + Z_d \chi_i \sin(k\theta_i))$$

where for all $(Z_i)_{i=1}^d$ are iid standard d dimensional Gaussian random variables. This can be succintly written as,

$$X \leftarrow \cos(k\Theta) X + \chi \sin(k\Theta) Z \tag{17}$$

where $X \in \mathbb{R}^d$,$Z \sim N(0, I_d)$ is independent of X, $\cos(k\Theta)$ is a $d \times d$ diagonal matrix with,

$$\cos(k\Theta)_{ii} = \cos(k\theta_i) \tag{18}$$

$\chi \sin(k\Theta)$ is a $d \times d$ diagonal matrix with,

$$(\chi \sin(k\Theta))_{ii} = \chi_i \sin(k\theta_i) \tag{19}$$

2.6.1 uHMC Mixing Time - Full Velocity Randomization

Theorem 5 *Let μ be a d dimensional multivariate Gaussian measure with 0 mean and the covariance matrix, Σ, a $d \times d$ diagonal matrix as follows,*

$$\Sigma = \begin{pmatrix} a_1 & 0 & 0 & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & 0 \\ 0 & 0 & a_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & a_d \end{pmatrix}$$

such that for all $i \leq d$, $1 \leq a_i \leq C$, let π be the transition kernel for the position vector of u-HMC with full velocity randomization, with a fixed stepsize $h \leq 1$, a fixed number of leapfrog steps k and the target measure μ. Let ν be a standard d dimensional Gaussian measure. Then,

$$\|\nu\pi^n - \mu\|_{TV} \leq h^2 \sqrt{d} \left(1 + \frac{3\|\cos(k\Theta)\|^{2n}}{2}\right) \tag{20}$$

where $\cos(k\Theta)$ as in (18).

Proof Let $(X_0, X_1, ...)$ denote the steps of a Markov chain with transition kernel π and an initial standard Gaussian distribution ν.
We compute the distribution of X_n.

From (17), we know that the u-HMC transition step is given as,

$$X_{n+1} = \cos(k\Theta)X_n + \chi sin(k\Theta)Z_{n+1} \tag{21}$$

where for all $n \in \mathbb{N}$, $Z_n \sim N(0, I_d)$ are iid and for all $k \leq n$, Z_n is independent of X_k. $\cos(k\Theta), \chi sin(k\Theta)$ are given by (18),(19) respectively. For all $i \leq d$, θ_i, χ_i are given by (14),(16) respectively.
Since sums of independent Gaussians is a Gaussian and a linear transformation of a Gaussian is a Gaussian, we see from (21) that if X_n is a Gaussian, then so is X_{n+1}. Since $X_0 \sim \nu$ is Gaussian, we get that for all $n \in \mathbb{N}$, X_n is a multivariate Gaussian. Further, by taking expectation on both sides of the equality in (2), we see that for all $n \in \mathbb{N}$, $E[X_{n+1}] = E[X_n]$.
Hence for all $n \in \mathbb{N}$, $E[X_n] = E[X_0] = (0, 0, ..0)$.
Thus, $X_n \sim N(0, C_n)$, where $C_n := Var(X_n)$.
By taking the variance on both sides of the equality in (2), we see that,

$$C_{n+1} = \cos(k\Theta)C_n \cos(k\Theta)^t + (\chi \sin(k\Theta))^2 \tag{22}$$

where $C_0 = I$.
Note that in (22), if C_n is a diagonal matrix then so is C_{n+1}.
Since $C_0 = I$, we see that for all $n \in \mathbb{N}$, C_n is a diagonal matrix.
After solving the recurrence relation in (22), we see that,

$$C_n = \cos(k\Theta)^{2n} + (\chi \sin(k\Theta))^2(I - \cos(k\Theta)^{2n})[(I - \cos(k\Theta)^2)]^{-1}$$

We compute the diagonal entries of C_n.

$$(C_n)_{ii} = \cos^{2n}(k\theta_i) + \chi_i^2 \sin^2(k\theta_i)\frac{1 - \cos^{2n}(k\theta_i)}{1 - \cos^2(k\theta_i)}$$

This can be simplified as

$$(C_n)_{ii} = \cos^{2n}(k\theta_i) + \chi_i^2(1 - \cos^{2n}(k\theta_i))$$

Knowing the distribution of $X_n \sim \mathcal{N}(0, C_n)$, we now compute an upper bound on the TV distance between X_n and the target $\mu \sim \mathcal{N}(0, \Sigma)$ using [1, Theorem 1].
The upper bound is given as 1.5 times the L^2 norm of the eigenvalues of $\Sigma^{-1}C_n - I_d$.
$\Sigma^{-1}C_n - I_d$ is a d dimensional, diagonal matrix with diagonal entries as,

$$(\Sigma^{-1}C_n - I_d)_{ii} = \frac{\cos^{2n}(k\theta_i) + \chi_i^2(1 - \cos^{2n}(k\theta_i))}{a_i} - 1$$

Plugging in $\chi_i = \dfrac{2a_i}{\sqrt{4a_i - h^2}}$, we see

$$(\Sigma^{-1}C_n - I_d)_{ii} = \frac{\cos^{2n}(k\theta_i)}{a_i} + \frac{4a_i}{4a_i - h^2}(1 - \cos^{2n}(k\theta_i)) - 1$$

After simplification we get,

$$(\Sigma^{-1}C_n - I_d)_{ii} = \frac{\cos^{2n}(k\theta_i)}{a_i} + \left(1 + \frac{h^2}{4a_i - h^2}\right)(1 - \cos^{2n}(k\theta_i)) - 1$$

Further simplification yields,

$$(\Sigma^{-1}C_n - I_d)_{ii} = \frac{h^2}{4a_i - h^2} + \cos^{2n}(k\theta_i)\left(\frac{1}{a_i} - \frac{h^2}{4a_i - h^2}\right)$$

We wish to bound the modulus of the diagonal entry, noting that $\frac{h^2}{4a_i - h^2} \geq 0$, since $a_i \geq 1$ and $h \leq 1$, we get

$$|(\Sigma^{-1}C_n - I_d)_{ii}| \leq \frac{h^2}{4a_i - h^2} + \frac{\cos^{2n}(k\theta_i)}{a_i} + \frac{h^2\cos^{2n}(k\theta_i)}{4a_i - h^2}$$

Noting that $\cos^{2n}(k\theta_i) \leq 1$ and $4a_i - h^2 \geq 3$, we get

$$|(\Sigma^{-1}C_n - I_d)_{ii}| \leq \frac{2h^2}{3} + \frac{\cos^{2n}(k\theta_i)}{a_i}$$

Noting that $a_i \geq 1$, we see

$$|(\Sigma^{-1}C_n - I_d)_{ii}| \leq \frac{2h^2}{3} + \cos^{2n}(k\theta_i) \tag{23}$$

Assume $\|A\|$ defines the operator norm of a matrix A. We then have,

$$|(\Sigma^{-1}C_n - I_d)_{ii}| \leq \frac{2h^2}{3} + \|\cos(k\Theta)\|^{2n}$$

We then have that

$$\|\nu\pi^n - \mu\|_{TV} \leq (3/2)\sqrt{\sum_{i=1}^{d}(|(\Sigma^{-1}C_n - I_d)_{ii}|^2)} \leq (3/2)\sqrt{\sum_{i=1}^{d}(\frac{2h^2}{3} + (1-\epsilon)^{2n})^2}$$

This finally gives,

$$\|\nu\pi^n - \mu\|_{TV} \leq (3/2)\sqrt{d}\left(\frac{2h^2}{3} + \|\cos(k\Theta)\|^{2n}\right)$$

Which when using the fact that $h \leq 1$,

$$\|\nu\pi^n - \mu\|_{TV} \leq h^2\sqrt{d}\left(1 + \frac{3\|\cos(k\Theta)\|^{2n}}{2}\right)$$

□

In order to get δ-close to the target distribution, we can choose n such that $\frac{3\|\cos(k\Theta)\|^{2n}}{2} < 1$, this implies

$$n \geq \frac{\ln\left(\frac{3}{2}\right)}{-2\ln(\|\cos(k\Theta)\|)} \tag{24}$$

and we choose h such that $h^2\sqrt{d} \leq \delta/2$,

$$h \leq d^{-\frac{1}{4}}\sqrt{\delta}(1/\sqrt{2}) \tag{25}$$

Note that in the continous time limit $k \to \infty$ and $h \to 0$ and $kh = t$, using (15), we get that that,

$$n \geq \frac{\ln\left(\frac{3}{2}\right)}{-2\ln\left(\max_{i\in[d]}\left|\cos\left(\frac{t}{\sqrt{a_i}}\right)\right|\right)} \tag{26}$$

The way to compare the mixing time of the different MCMC algorithms is to measure the number of gradient evaluations that we need to perform. This is given by nk, where $k = t/h$.
Thus, we see that to get δ close to the target distribution, we need to make $\frac{td^{1/4}\delta^{-1/2}\ln\left(\frac{3}{2}\right)}{-\sqrt{2}\ln\left(\max_{i\in[d]}\left|\cos\left(\frac{t}{\sqrt{a_i}}\right)\right|\right)}$ many gradient evaluations.

2.7 One shot coupling

We first discuss the one shot coupling in general. Before doing so, recall the existence of optimal couplings with respect to the total variation norm. This is borrowed from [5, Section 4].

Theorem 6 *Let ν and μ be probability measures on $\mathbb{R}^d$. Then there exists a coupling of the measures, (Z_1, Z_2) such that $Z_1 \sim \nu$, $Z_2 \sim \mu$ and*

$$\mathbb{P}(Z_1 \neq Z_2) = \|\mu - \nu\|_{TV}$$

Proof [3, Lemma 3.4] or [5, Lemma 1] □

We now define the one shot coupling.

Theorem 7 *Let ν and μ be probability measures on $\mathbb{R}^d$. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a bijective C^1 function. Then there exists a coupling of the measures, (Z_1, Z_2) such that $Z_1 \sim \nu$, $Z_2 \sim \mu$ and*

$$\mathbb{P}(F(Z_1) \neq Z_2) = \|\mu - \nu \circ F\|_{TV}$$

Proof We apply 6 to the probability measures μ and $\nu \circ F$. We then get $Y_1 \sim \mu$ and $Y_2 \sim \nu \circ F$, such that,

$$\mathbb{P}(Y_1 \neq Y_2) = \|\mu - \nu \circ F\|_{TV}$$

Define $Z_1 := Y_1$ and $Z_2 := F(Y_2)$, then $Z_1 \sim \mu$ and $Z_2 \sim \nu$. Further,

$$\mathbb{P}(F(Z_1) \neq Z_2) = \mathbb{P}(F(Y_1) \neq F(Y_2)) = \mathbb{P}(Y_1 \neq Y_2) = \|\mu - \nu \circ F\|_{TV}$$

Hence (Z_1, Z_2) is the required coupling. □

2.7.1 uHMC transition on state space with full velocity randomization

We now recall the uHMC transition step in state space from (8). We consider full velocity randomization and assume the target density $\propto e^{-U(x)}dx$. The transition step is then given by,

$$x \leftarrow P \circ \phi_h^k(x, Z)$$

where ϕ_h is the transition step for uHMC in phase space, $P : \mathbb{R}^{2d} \to \mathbb{R}^d$ projects $(x, v) \to x$ and $Z \sim \mathcal{N}(0, I_d)$ is independent of x.
For ease of notation, we denote $P \circ \phi_h^k$ as q. Hence the uHMC transition step is given by,

$$x \leftarrow q(x, Z) \tag{27}$$

Let π denote the transition kernel for uHMC.

2.7.2 One shot coupling for uHMC

We wish to obtain a succesful coupling for uHMC as in [6].
For $x, y \in \mathbb{R}^d$, we wish to compute an upper bound for $\|\pi(x, .) - \pi(y, .)\|_{TV}$.
Note that $\pi(x, .) = q(x, Z_1)$ and $\pi(y, .) = q(x, Z_2)$, where $Z_1, Z_2 \sim \mathcal{N}(0, I_d)$.
We find a coupling of the two standard Gaussian random variables, Z_1 and Z_2, such that after one step, we maximize the probability that $q(x, Z_1) = q(y, Z_2)$.
Note that by properties of the Hamiltonian dynamics, there exists a C^1 bijection $F : \mathbb{R}^d \to \mathbb{R}^d$, such that for all $v \in \mathbb{R}^d$

$$q(x, v) = q(y, F(v))$$

We couple the velocities using the one shot coupling from 7, using the bijection F. Thus we obtain (Z_1, Z_2) such that $Z_1, Z_2 \sim \mathcal{N}(0, I_d)$ and

$$\mathbb{P}(F(Z_1) \neq Z_2) = TV(\mathcal{N}(0, I_d), F(\mathcal{N}(0, I_d)))$$

Since, if $F(Z_1) = Z_2$, $q(x, Z_1) = q(x, Z_2)$, we get

$$\mathbb{P}(q(x, Z_1) \neq q(x, Z_2)) \leq \mathbb{P}(F(Z_1) \neq Z_2) = TV(\mathcal{N}(0, I_d), F(\mathcal{N}(0, I_d)))$$

Thus we obtain,

$$\|\pi(x, .) - \pi(y, .)\|_{TV} \leq TV(\mathcal{N}(0, I_d), F(\mathcal{N}(0, I_d))) \tag{28}$$

The authors of [6] use conditions on U and Pinsker's inequality to bound $TV(\mathcal{N}(0, I_d), F(\mathcal{N}(0, I_d)))$ by $\|x - y\|_2$. We restrict ourselves to the Gaussian case, where we can explicitly compute F.

2.7.3 Gaussian one shot coupling

Recall from 17, the uHMC transition step for a Gaussian target.

$$x \leftarrow q(x, Z) := xcos(k\Theta) + Z\chi sin(k\Theta)$$

where $x \in \mathbb{R}^d$ and $Z \sim N(0, I_d)$ is independent of x.
We wish to find $F : \mathbb{R}^d \to \mathbb{R}^d$, such that for $x, y \in \mathbb{R}^d$,

$$q(x, v) = q(y, F(v)) \text{ for all } v \in \mathbb{R}^d$$

We compute F as,

$$F(v) = v + \chi^{-1}\cot(k\Theta)(x-y)$$

From (28), we need to bound $TV(\mathcal{N}(0,I_d),\mathcal{N}(\chi^{-1}\cot(k\Theta)(x-y),I_d))$.
From [7, Theorem 1], we get that,

$$TV(\mathcal{N}(0,I_d),\mathcal{N}(\chi^{-1}\cot(k\Theta)(x-y),I_d)) \leq 2\Phi\left(\frac{1}{2}\|\chi^{-1/2}\cot(k\Theta)\,(x-y)\|_2\right)-1$$

where $\Phi(z) := (2\pi)^{-1/2}\int_{-\infty}^{z} e^{-t^2/2}dt$ is the standard normal cumulative distribution function, $\|\cdot\|_2$ denotes the appropriate 2-norm (spectral for symmetric $n\times n$-matrices, Euclidean for n-vectors)
Using, $\Phi(z) \leq 1/2 + (2\pi)^{-1/2}z$ for $z \geq 0$, the RHS of the above inequality can be bounded and we get,

$$\|\pi(x,.) - \pi(y,.)\|_{TV} \leq (2\pi)^{-1/2}\|\chi^{-1/2}\cot(k\Theta_h)\|_2\,\|x-y\|_2 \tag{29}$$

We now obtain the TV/W_1 regularization result.
Let $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ and let (X,Y) be the optimal W_1 coupling for the measures μ and ν, i.e $X\sim\mu, Y\sim\nu$ and $E[\|(X-Y)\|_2] = W_1(\mu,\nu)$.

$$\begin{aligned} TV(\nu\pi,\mu\pi) &\leq \mathbb{E}[TV((\pi(X,.),\pi(Y,.)))] \\ &\leq (2\pi)^{-1/2}\|\chi^{-1/2}\cot(k\Theta)\|_2\, d_{W_1}(\nu,\mu)\,, \end{aligned} \tag{30}$$

where the last line used the inequality in (29). Define,

$$C_{TV/W_1} := (2\pi)^{-1/2}\|\chi^{-1/2}\cot(k\Theta)\|_2 \tag{31}$$

Notice that if we show that π acts as a contraction on $\mathcal{P}(\mathbb{R}^d)$ with respect to the W_1 metric, we can show geometric ergodicity in the total variation metric. This can be seen by assuming there exists $C_{W_1} < 1$, such that for all $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$,

$$W_1(\mu\pi,\nu\pi) \leq C_{W_1}W_1(\mu,\nu)$$

Iterating the bound gives us,

$$W_1(\mu\pi^n,\nu\pi^n) \leq C_{W_1}^n W_1(\mu,\nu)$$

But we have,

$$TV(\mu\pi^n,\nu\pi^n) \leq C_{TV/W_1}W_1(\mu\pi^{n-1},\nu\pi^{n-1}) \leq C_{TV/W_1}C_{W_1}^{n-1}W_1(\mu,\nu) \tag{32}$$

The RHS goes to 0 as $n\to\infty$ and we establish an exponential convergence to the stationary distribution.
We now show that π acts as a contraction on $P(\mathbb{R}^d)$ with respect to the W_1

Wasserstein metric.
We want to compute $0 \leq C_{W_1} < 1$, such that for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

$$W_1(\nu\pi, \mu\pi) \leq C_{W_1} W_1(\nu, \mu)$$

By Kantorvich-Rubenstein duality, it suffices to consider the case when the initial distributions are Dirac delta measures.
Let $x, y \in \mathbb{R}^d$. We use a synchronous coupling to couple $\pi(x, .)$ and $\pi(y, .)$. In particular, let $X_1 = q(x, Z)$ and $Y_1 = q(y, Z)$, where $Z \sim \mathcal{N}(0, I_d)$. Then we have,

$$W_1(\pi(x, .), \pi(y, .)) \leq E[\|X_1 - Y_1\|_2] \leq \|\cos(k\Theta)x - \cos(k\Theta)y\|_2 \leq \|\cos(k\Theta)\|_2 \|x - y\|_2$$

where for the second inequality we used that for $x, v \in \mathbb{R}^d$, $q(x, v) := x cos(k\Theta) + \chi \sin(k\Theta)v$.
So we have that $C_{W_1} = \|\cos(k\Theta)\|_2$.
From (32), we see that we are δ close to the invariant distribution μ_h, with the initial distribution μ, if

$$n \geq 1 + \frac{\ln(C_{TV/W_1}) + \ln(W_1(\mu, \mu_h)) - \ln(\delta)}{-\ln(C_{W_1})} \tag{33}$$

Plugging in the values of C_{TV/W_1} and C_{W_1}, we get,

$$n \geq 1 + \frac{\ln((2\pi)^{-1/2} \|\chi^{-1/2} \cot(k\Theta)\|_2\ ) + \ln(W_1(\mu, \mu_h)) - \ln(\delta)}{-\ln(\|\cos(k\Theta)\|_2)} \tag{34}$$

In the continuous limit, $k \to \infty$, $h \to 0$ and $kh \to t$, from (16), we get $\chi \to \sqrt{a}$, from (15), we see that $\|\cos(k\Theta)\|_2 \to \max_{i \in [d]} \left|\cos\left(\frac{t}{\sqrt{a_i}}\right)\right|$, we get

$$n \geq 1 + \frac{\ln\left(\max\limits_{i \in [d]} \left|a_i^{1/4} \cot\left(\frac{t}{\sqrt{a_i}}\right)\right|\right) + \ln(W_1(\mu, \mu_h)) - \ln(\delta)}{-\ln(\max\limits_{i \in [d]} \left|\cos\left(\frac{t}{\sqrt{a_i}}\right)\right|)} \tag{35}$$

2.8 mHMC

[8] [9]

3 Numerical Experiments

We vary either the dimension or the error tolerance and estimate the mixing time of the various MCMC methods.

3.1 Experimental setup

The target is a d dimensional Gaussian distribution with zero mean and covariance matrix Σ defined as follows,

$$\Sigma_{d\times d} = \begin{pmatrix} 2 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{36}$$

The potential function, $U(x) := \frac{1}{2}x^t\Sigma x$.
The upper bounds and lower bounds on the operator norm of the Hessian of the potential are, $m \leq \|Hess(U(x))\|_{op} \leq L$, where $m = \frac{1}{2}$ and $L = 1.0$ for this problem. The condition number $\kappa := L/m = 2$.

3.2 Theoretical results for mixing time vs stepsize

We wish to apply 3 and 4 for finding bounds on the mixing time of MALA and MRW respectively.
In our experimental setup, using 1, we see that $N(0, I_d)$ is β-warm with respect to $N(0, \Sigma)$, where Σ as in (36), with $\beta = \det(\Sigma)^{1/2} = \sqrt{2}$.
Using (7), we see $r\left(\frac{\delta}{2\beta}\right) \leq 4$ for $\delta \geq 0.1$ and $d \geq 5$.
Thus we can treat $r\left(\frac{\delta}{2\beta}\right)$ as a constant independent of d and δ as long as $\delta \geq 0.1$.
Further notice that the condition number κ is fixed for our experiments and hence can also be treated as a constant.
The relation between mixing time and stepsize is summarized in the following table.

"426A30Cc"426A30Cc"426A30Cc"426A30Cc

Algorithm	Stepsize	Mixing time
ULA (6)	$\Theta(d^{\frac{-1}{2}}\delta)$	$O(d^{\frac{1}{2}}\delta^{-1}ln(d\delta^{-1}))$
ULA (6)	$\Theta(d^{-1}\delta)$	$O(d\delta^{-1}ln(d\delta^{-1}))$
MRW [2, Cor 3]	$\Theta(d^{-1})$	$O(dln(\delta^{-1}))$
MALA [2, Cor 3]	$O(d^{-1})$	$O(dln(\delta^{-1})))$

Table 1: Theoretical results for mixing times vs stepsizes for various MCMC algorithms with the initial distribution, $\nu = N(0, I_d)$ and target density $\mu = N(0, \Sigma)$, where Σ is a diagonal matrix with all but the first entry equal to 1, with the first entry strictly greater than 1.
δ denotes the error tolerance and d denotes the dimension.

We say that a Markov chain "starts from the origin", if the initial distribution is a a Dirac delta measure centered at the origin.
"426A30Cc"426A30Cc"426A30Cc"426A30Cc

Algorithm	Stepsize	Mixing time
ULA [10, Discussion after Thm 14]	$\Theta(d^{\frac{-1}{2}}\delta)$	$O(d^{\frac{1}{2}}\delta^{-1}ln(d\delta^{-1}))$
ULA [10, Discussion after Thm 14]	$\Theta(d^{-1}\delta)$	$O(d\delta^{-1}ln(d\delta^{-1}))$

Table 2: Theoretical results for mixing times vs stepsizes for ULA when we start from the origin with a target density $\mu = N(0, \Sigma)$, where Σ is a diagonal matrix with bounded diagonal entries that are greater than or equal to 1.
δ denotes the error tolerance and d denotes the dimension.

When we start from the origin, the stepsizes of MALA/MRW should be $O(d^{\frac{-1}{2}})$[11, Thm 3.17] / $O(d^{-1})$[12, Lecture 6] respectively, for T_{mix} to not exponentially increase with dimension.

3.3 Estimation of mixing time

We describe how we sample the statistic used to estimate the mixing time.
Let (π,ν) denote a Markov chain with kernel π and initial distribution ν with state space $\mathbb{R}^d$. Assume μ is the unique invariant distribution of the chain.
We aim to estimate the mixing time of (π,ν).
Let $\delta \geq 0$ denote the error tolerance and d the dimension.
Let $i \leq d$ be the coordinate for which we check for mixing. For a multivariate Gaussian target with independent components, i would be chosen to be the coordinate on which the target has the most variance as this exhibits the slowest mixing.
The statistic $\hat{T}(\delta)$ is defined to be

$$\hat{T}(\delta) := min\{n : |q_{\frac{3}{4}}(X_0^{[i]}, X_1^{[i]}, ..., X_n^{[i]}) - q_{\frac{3}{4}}(\mu^{[i]})| < \delta\}$$

where $X_0^{[i]}, X_1^{[i]}, ..., X_n^{[i]}$ are the ith components of the states of the Markov chain,
$q_{\frac{3}{4}}(a_1, a_2, a_3, ...a_n) :=$ empirical $\frac{3}{4}$th quantile of the vector $(a_1, a_2, a_3, ...a_n)$,
$q_{\frac{3}{4}}(\mu^{[i]})$ is the $\frac{3}{4}$th quantile of μ projected to the ith coordinate.
Let $N \in \mathbb{N}$ be the number of independent samples of $\hat{T}(\delta)$ we wish to draw.
The statistic we use to estimate $T_{mix}(\delta)$ is defined as

$$\hat{T}_{mix}(\delta) := \sum_{j=1}^{N} \hat{T}_j(\delta)$$

where $\hat{T}_1(\delta), \hat{T}_2(\delta), ..., \hat{T}_N(\delta)$ are iid draws of the statistic $\hat{T}(\delta)$.

3.4 Mixing Time vs Dimension

We look at the behaviour of mixing time with dimensions after fixing an error tolerance. The dimensions we consider are all odd numbers in the range of 5 to 23.

The initial distribution is $N(0, L^{-1}I_d)$, unless stated otherwise, where L denotes the upper bound on the operator norm of the Hessian of the potential function.

Note that by 1, $N(0, I_d)$ is β-warm for the target, $N(0, \Sigma)$, where Σ as from 36.

We sample $\hat{T}_{mix}(\delta)$ as described in 3.3, with the following hyperparameter values.

δ is set as 0.4, after experimenting with various values. This value is large enough to achieve quick convergence and the estimator. $\hat{T}(\delta)$ has a low empirical variance, when compared to smaller values of δ.

i, the coordinate considered for mixing in 3.3 is set as 1, as the eigenvalue of Σ is highest in the first coordinate.

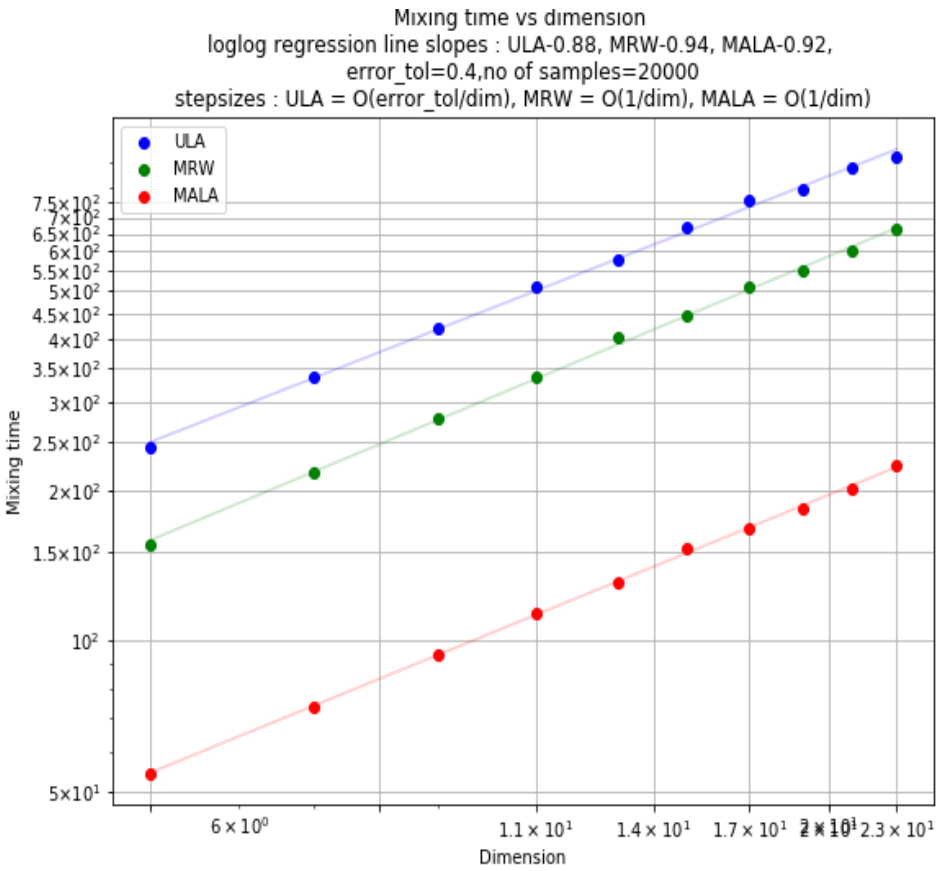
N, the no of samples of we average over is set as 20,000.

Plots

We first look at the case where the stepsizes of ULA and MALA have a $O(d^{-1})$ dependence on dimension.
"426A30Cc"426A30Cc"426A30Cc"426A30Cc

ULA	MRW	MALA
$\delta/d\kappa L$	$1/d\kappa L$	$1/Ld$

Table 3: Stepsizes for the algorithms
δ denotes the error tolerance, d denotes the dimension and L the upper bound on the operator norm of the Hessian



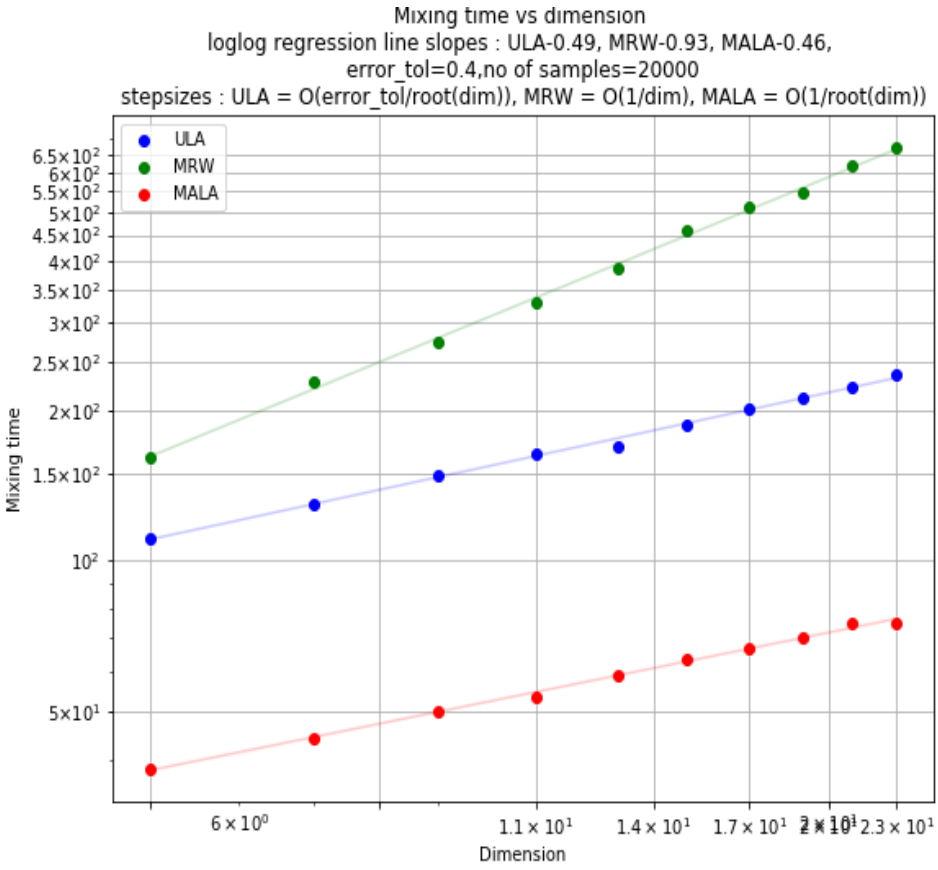
The loglog slopes of 0.88, 0.94 and 0.92 for ULA/MRW/MALA show that the mixing time is $O(d)$ and it agrees with the theoretical estimates from 1.
It is however surprising that MRW performs better than ULA here.

Consider the case where the stepsizes of ULA and MALA have a $O(d^{-1/2})$ dependence on dimension.

"426A30Cc"426A30Cc"426A30Cc"426A30Cc

ULA	MRW	MALA
$\delta/\sqrt{d\kappa L}$	$1/d\kappa L$	$1/\sqrt{d\kappa L}$

Table 4: Stepsizes for the algorithms
δ denotes the error tolerance, d denotes the dimension and L the upper bound on the operator norm of the Hessian

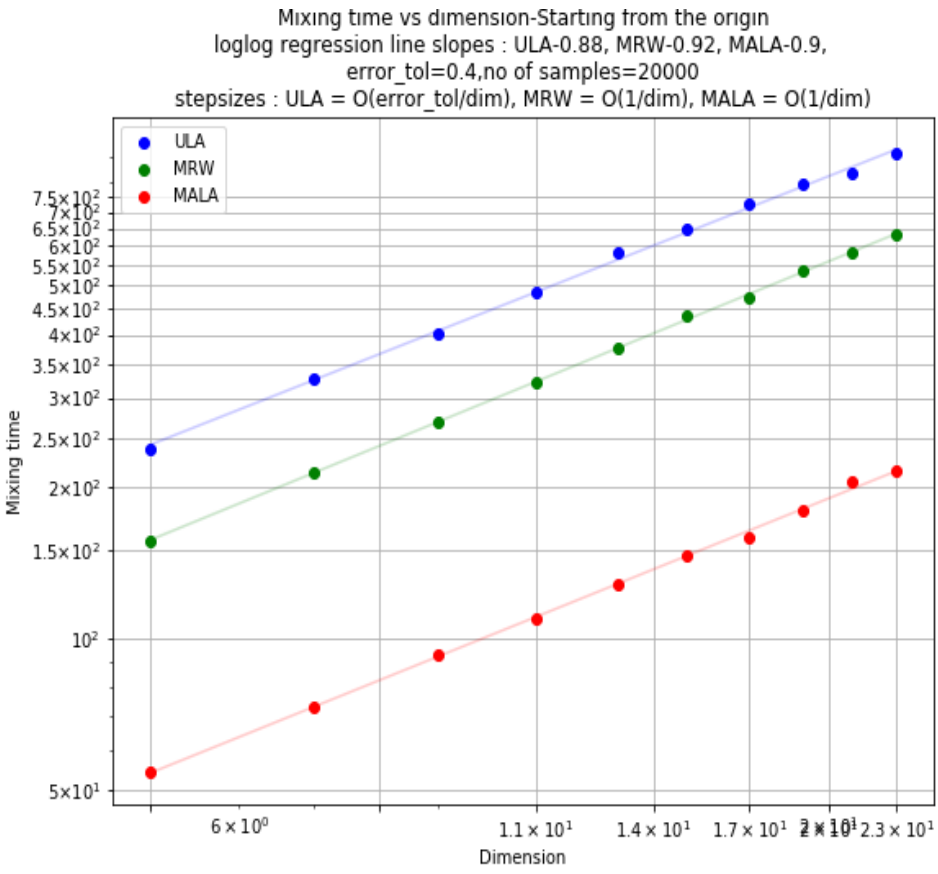


The loglog slope of 0.49 for ULA agrees with the theoretical result from 2 and suggests that mixing time is $O(d^{-1/2})$ dependence.
The loglog slope of 0.94 for MRW is not surprising as before.
The loglog slope of 0.46 for MALA suggests that the mixing time is $O(d^{-1/2})$

Now we look at plots with the initial distribution being a Dirac delta centered at the origin. Let the stepsizes of ULA and MALA have a $O(d^{-1})$ dependence on dimension.
"426A30Cc"426A30Cc"426A30Cc"426A30Cc

ULA	MRW	MALA
$\delta/d\kappa L$	$1/d\kappa L$	$1/Ld$

Table 5: Stepsizes for the algorithms
δ denotes the error tolerance, d denotes the dimension and L the upper bound on the operator norm of the Hessian

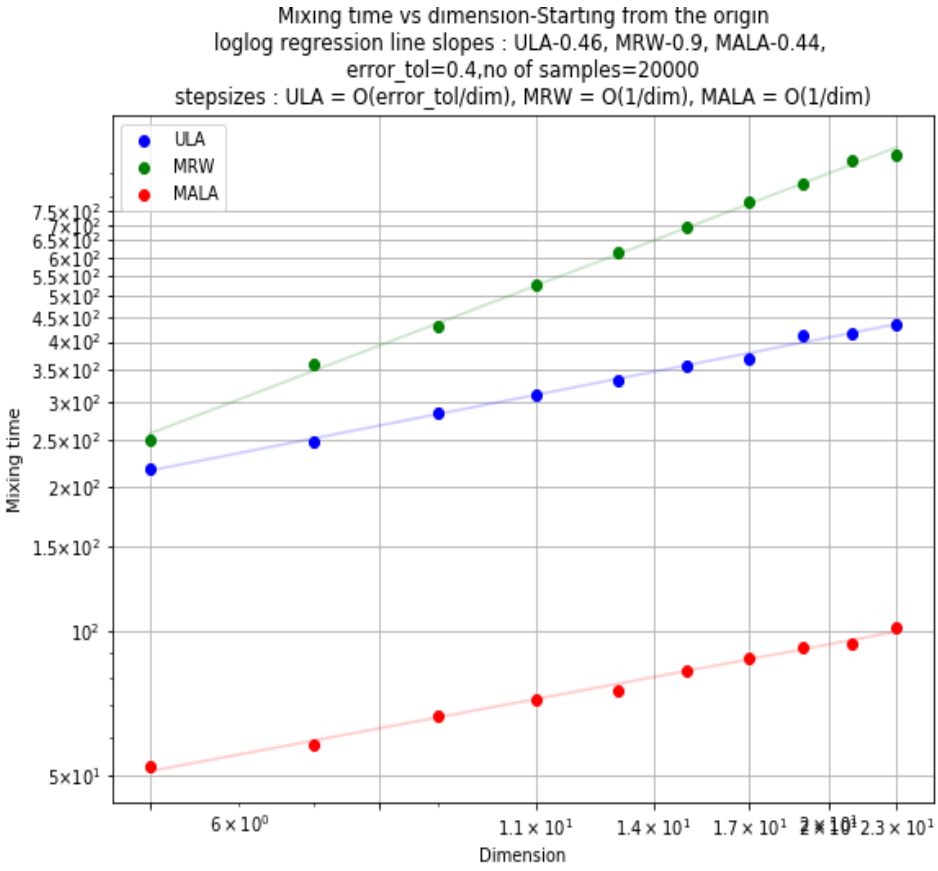


The loglog slopes of 0.88/0.92/0.9 for ULA/MRW/MALA suggest that the mixing time is $O(d^{-1})$ which is similar to the case when we had a warm start. For ULA, the slope is in accordance with the theoretical result from 2.

Let the initial distribution be a Dirac delta centered at the origin. and let the stepsizes of ULA and MALA have a $O(d^{-1/2})$ dependence on dimension.
"426A30Cc"426A30Cc"426A30Cc"426A30Cc

ULA	MRW	MALA
$\delta/\sqrt{d}\kappa L$	$1/d\kappa L$	$1/\sqrt{d}\kappa L$

Table 6: Stepsizes for the algorithms
δ denotes the error tolerance, d denotes the dimension and L the upper bound on the operator norm of the Hessian



The loglog slopes of 0.46/ 0.44 for ULA and MALA suggests the mixing time is $O(d^{-1/2})$. A loglog slope of 0.9 for MRW suggests that the mixing time is $O(d^{-1})$. These results are similar to the case when we had a warm start. For ULA, the slope is in accordance with the theoretical result from 2.

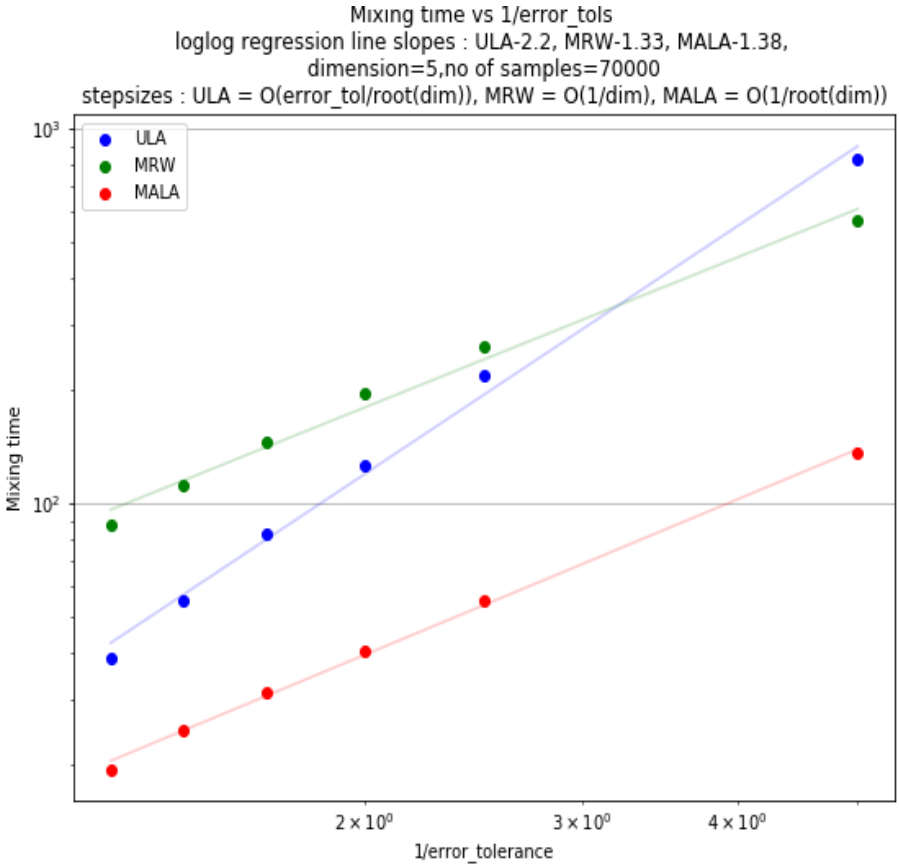
3.5 Mixing Time vs $\frac{1}{\delta}$

We look at the behaviour of mixing time with error tolerance after fixing a certain dimension. The initial distribution is $N(0, L^{-1}I_d)$ unless stated otherwise. Note that by 1, $N(0, I_d)$ is β-warm for the target, $N(0, \Sigma)$, where Σ as from 36. We sample $\hat{T}_{mix}(\delta)$ as described in 3.3 with the following hyperparameter values.
d, the dimension is set as 5, to obtain fast mixing.
i, the coordinate considered for mixing in 3.3 is set as 1, as the eigenvalue of Σ is highest in the first coordinate.
N, the no of samples of we average over.

Plots
"426A30Cc"426A30Cc"426A30Cc"426A30Cc

ULA	MRW	MALA
$\delta/\sqrt{d\kappa L}$	$1/d\kappa L$	$1/\sqrt{d\kappa L}$

Table 7: Stepsizes for the algorithms
δ denotes the error tolerance, d denotes the dimension and L the upper bound on the operator norm of the Hessian.

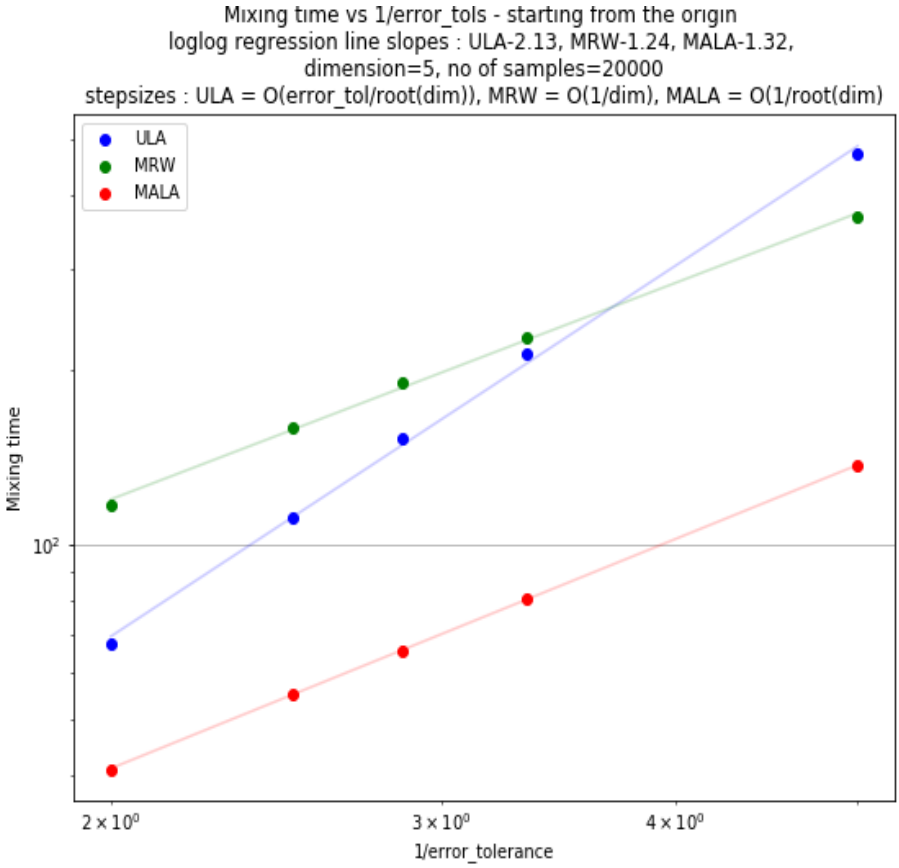


The loglog slopes of 2.2 for ULA is in accordance with what was obtained in [2, Figure 2] but contradicts what was obtained in 1. The slopes of 1.33/1.38 for MRW/MALA suggest a $O(1/\delta)$ dependence, which isn't the case according to 1.

Consider the same with the initial distribution a Dirac delta centered at the origin.
"426A30Cc"426A30Cc"426A30Cc"426A30Cc

ULA	MRW	MALA
$\delta/\sqrt{d\kappa L}$	$1/d\kappa L$	$1/\sqrt{d\kappa L}$

Table 8: Stepsizes for the algorithms
δ denotes the error tolerance, d denotes the dimension and L the upper bound on the operator norm of the Hessian



We get the same results as in the case of a warm start.

References

[1] Devroye, L., Mehrabian, A., Reddad, T.: The total variation distance between high-dimensional Gaussians (2020)

[2] Dwivedi, R., Chen, Y., Wainwright, M.J., Yu, B.: Log-concave sampling: Metropolis-Hastings algorithms are fast (2019)

[3] Eberle, A.: Markov processes. Lecture Notes at University of Bonn (2021)

[4] Bou-Rabee, N., Sanz-Serna, J.M.: Geometric integrators and the hamiltonian monte carlo method. Acta Numerica **27**, 113–206 (2018). https://doi.org/10.1017/s0962492917000101

[5] Roberts, G.O., Rosenthal, J.S.: One-shot coupling for certain stochastic recursive sequences. Stochastic Processes and their Applications **99**(2), 195–208 (2002). https://doi.org/10.1016/S0304-4149(02)00096-0

[6] Bou-Rabee, N., Eberle, A.: Mixing Time Guarantees for Unadjusted Hamiltonian Monte Carlo (2021)

[7] Barsov, S.S., Ulyanov, V.: Estimates of the proximity of gaussian measures. Doklady Mathematics **34**, 462 (1987)

[8] Bou-Rabee, N., Eberle, A., Zimmer, R.: Coupling and convergence for hamiltonian monte carlo. The Annals of Applied Probability **30**(3) (2020). https://doi.org/10.1214/19-aap1528

[9] Chen, Y., Dwivedi, R., Wainwright, M.J., Yu, B.: Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients (2021)

[10] Durmus, A., Moulines, E.: High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm (2018)

[11] Bou-Rabee, N., Eberle, A.: Markov chain Monte Carlo methods (2020)

[12] Bou-Rabee, N.: Slides from course on Markov chain Monte Carlo methods taught at University of Bonn (2021)