IMPERIAL

Sampling from log-concave distributions through Markov Chain Monte Carlo methods

Analysis and evaluation of unadjusted and underdamped Langevin algorithms

Peiyi Zhou

June 12, 2024

Table of Contents

Table of Contents

Introduction

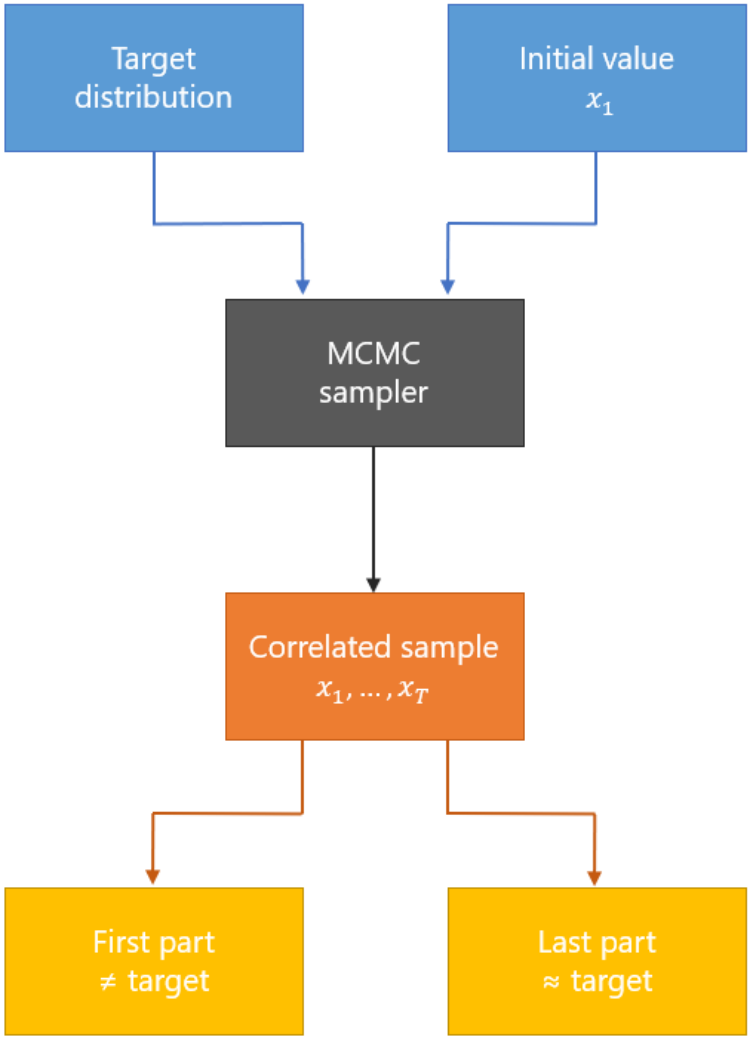


Figure: Flowchart showing how MCMC works. [1]

Table of Contents

Overdamped Langevin stochastic differential equation (SDE)

Let us assume that we want to sample a target distribution $p_\star$ where

$$p_\star(x)\mathrm{d}x \propto \exp(-U(x))\mathrm{d}x \tag{1}$$

An efficient method for obtaining a sample from (1) is simulating the overdamped Langevin SDE [2] by

$$\mathrm{d}X_t = -h(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \tag{2}$$

Settings of overdamped Langevin SDE

- A random initial condition, or a starting point, $X_0 = x_0$.
- The function $h = \nabla U$, as the gradient of U.
- U is strongly convex and Lipschitz-smooth (L-smooth). This means $U : \mathbb{R}^d \to \mathbb{R}$ should be continuously differentiable, and its gradient ∇U should be Lipschitz continuous, with Lipschitz constant L [2]

$$\|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|$$

- $\{B_t\}_{t\geq 0}$ is a d-dimensional (standard) Brownian motion.
- $\{X_t\}_{t\geq 0}$ is the Markov chain generated, which each state in the chain represents a sample from (1).

Rise of unadjusted Langevin algorithm (ULA)

The Euler–Maruyama discretisation scheme of the Langevin SDE defined for $t \in \mathbb{N}$ by [3]

$$X_{t+1} = X_t - \gamma \nabla U(X_t) + \sqrt{2\gamma} Z_{t+1}, \qquad X_0 = x_0 \tag{3}$$

where

- $x_0 \in \mathbb{R}^d$ as the starting point.
- The stepsize $\gamma > 0$.
- $(Z_t)_{t \in \mathbb{N}}$ are i.i.d. standard d-dimensional Gaussian variables, i.e. $Z_t \sim \mathcal{N}(0, \mathbf{I}_d)$.

Another way for writing this is to set $v_t = \sqrt{2\gamma} Z_t \sim \mathcal{N}(0, 2\gamma \mathbf{I}_d)$,

$$X_{t+1} = X_t - \gamma \nabla U(X_t) + v_{t+1}, \qquad X_0 = x_0 \tag{4}$$

Rise of unadjusted Langevin algorithm (ULA)

Algorithm 1: Unadjusted Langevin algorithm (ULA)

Input : N, the number of samples we want to take
$X_0 = x_0$, as the initial value
n, the number of burnin samples need to discard

Output: Approximately uncorrelated samples of target distribution $p_\star$

1 **for** $t = 1, 2, \cdots, N$ **do**

2 Take a random sample x' from the proposal

$$X' \sim q(x'|x_{t-1}) = \mathcal{N}(x'; x_{t-1} - \gamma \nabla U(x_{t-1}), 2\gamma \mathbf{I}_d)$$

which is same as

$$x' = x_{t-1} - \gamma \nabla U(x_{t-1}) + v_t, \quad \text{for } v_t \sim \mathcal{N}(\mathbf{0}, 2\gamma \mathbf{I}_d)$$

3 Update the new state by $x_t = x'$;

4 **end**

5 **return** Discard first n burnin samples and return the remaining samples.

Specific setting: 1D Gaussian target distribution

The most common and trivial setting for a strongly convex and L-smooth $U(x)$: **quadratic function with a positive leading coefficient!**

$$U(x) = \frac{1}{2}ax^2 + bx + c \implies p_\star(x) = \mathcal{N}\left(x; -\frac{b}{a}, \frac{1}{a}\right) \tag{5}$$

Also, put this gradient of $U(x)$ back to our ULA chain in (4), we would have

$$X_{t+1} = X_t - \gamma(aX_t + b) + v_{t+1} = (1 - a\gamma)X_t - \gamma b + v_{t+1}, \qquad X_0 = x_0 \tag{6}$$

which becomes an AR(1) process!

Stationary analysis

Definition (Stationary/covariance-stationary/weakly stationary)

A process is said to be covariance stationary, or weakly stationary, if its first and second moments are time invariant. i.e. for a covariance-stationary process Y_t, we would have

$$\begin{aligned} \mathbb{E}(Y_t) &= \mathbb{E}(Y_{t-1}) && \forall t \\ \operatorname{cov}(Y_t, Y_{t+\tau}) &= s_\tau && \forall t, \tau > 0 \\ \operatorname{var}(Y_t) &= s_0 < \infty && \forall t \end{aligned}$$

where s_τ $(\tau \geq 0)$ is called the autocovariance sequence (acvs) that is independent of t. [4]

Stationary analysis

Theorem (Conditions for stationarity, AR(p) process)

For an AR(p) process written in the form as

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \mu + \epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ *as the white noise process. We say this process is stationary/stable if the roots of the lag polynomial, also called the characteristic polynomial*

$$1 - \phi_1 z - \cdots - \phi_p z^p = 0$$

lie outside the unit circle/disk $|z| < 1$. *[4]*

Stationary analysis

Therefore, for the sequence (6) generated by ULA, we have the characteristic equation

$$1 - (1 - a\gamma)\, z = 0 \implies z = \frac{1}{1 - a\gamma}$$

Thus our Markov chain (6) generated by ULA is stationary if we have appropriate stepsize γ to satisfy

$$\left| \frac{1}{1 - a\gamma} \right| > 1 \tag{7}$$

Since the ULA works based on the recursion of the chain (6), it is really important to choose the appropriate stepsize γ that satisfies (6) to hold the stationarity.

Stationary analysis: mean and variance

We can derive the mean of the sequence (6) by applying expectation on both sides. If we set $\mathbb{E}(X_t) = \mu$, then

$$\mu = (1 - a\gamma)\,\mu - \gamma b \implies \mathbb{E}(X_t) = \mu = -\frac{b}{a} \tag{8}$$

which is exactly the mean of the target distribution (5)!
We can also apply the variance on both sides, as when the process is stationary, its variance is time invariant. If we assume $\mathrm{var}(X_t) = \sigma^2$, then

$$\sigma^2 = (1 - a\gamma)^2\,\sigma^2 + 2\gamma \implies \sigma^2 = \frac{2\gamma}{1 - (1 - a\gamma)^2} = \frac{2}{a(2 - a\gamma)} \tag{9}$$

Stationary distribution of ULA

Asymptotically we can now deduce the stationary distribution of ULA chain (6):

$$X \sim \mathcal{N}\left(\mu, \sigma^2\right) = \mathcal{N}\left(-\frac{b}{a}, \frac{2}{a(2-a\gamma)}\right) \tag{10}$$

This is biased, due to absence of Metropolis step, compared with the target distribution $p_\star$ as in (5), but this distribution (10) of chain generated by ULA is really close when γ is not too big - so $a\gamma$ would be somewhat negligible.

ULA performance analysis

We shall now analyse ULA, from its stationary distribution derived above with respect to the target distribution $p_\star$, in the perspective of

- Sampling performance
- Convergence behaviour

Distance measure

Two measures that quantitatively determine 'the difference between one probability distribution to another':

- Kullback-Leibler divergence/relative entropy
- 2-Wasserstein distance

Kullback-Leibler (KL) divergence

Definition (Kullback-Leibler divergence/relative entropy)

Consider some unknown distribution $p(x)$, and suppose that we have modelled this using an approximating distribution $q(x)$. The Kullback-Leibler divergence (KL divergence), or known as the relative entropy, between the distributions $p(x)$ and $q(x)$, is calculated as [5]

$$\mathrm{KL}(p||q) = -\int p(x) \log \frac{q(x)}{p(x)} \mathrm{d}x = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \tag{11}$$

Kullback-Leibler (KL) divergence

Lemma (KL divergence on two univariate normal distributions p, q)

Assume that $p(x) = \mathcal{N}(x; \mu_p, \sigma_p^2)$ *and* $q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2)$ *as two normal probability density functions, then their KL divergence defined in (11) can be written as [6]*

$$\mathrm{KL}(p||q) = \frac{1}{2}\left(\log\frac{\sigma_q^2}{\sigma_p^2} + \frac{\sigma_p^2}{\sigma_q^2} + \frac{(\mu_p - \mu_q)^2}{\sigma_q^2} - 1\right) \tag{12}$$

$$= \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \tag{13}$$

2-Wasserstein distance

Definition (2-Wasserstein distance, on two normal distributions)

For two non-degenerate, d-dimensional normal distributions $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ (so Σ_0, Σ_1 are positive definite matrices), their 2-Wasserstein distance is defined as [7]

$$W_2^2(\mu, \nu) = \|m_1 - m_0\|^2 + \operatorname{tr}\left(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\right) \tag{14}$$

2-Wasserstein distance

Lemma (2-Wasserstein distance on two univariate normal distributions)

More specifically, if we treat univariate normal distributions $p(x) = \mathcal{N}(x; \mu_p, \sigma_p^2)$ *and* $q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2)$*, then the 2-Wasserstein distance becomes*

$$W_2^2(p, q) = (\mu_q - \mu_p)^2 + (\sigma_p - \sigma_q)^2 \tag{15}$$

which also implies the symmetric behaviour of 2-Wasserstein metric for the univariate normal case.

ULA sampling performance analysis

Now, with the notation of $p_\star(x) = \mathcal{N}\left(x; -\frac{b}{a}, \frac{1}{a}\right)$ as the unknown target distribution, and $q(x) = \mathcal{N}\left(x; -\frac{b}{a}, \frac{2}{a(2-a\gamma)}\right)$ as the stationary distribution generated by ULA as in (10), we can deduce

$$\mathrm{KL}(p_\star||q) = \frac{1}{2}\log\left(\frac{2}{a(2-a\gamma)}\right) - \frac{1}{2}\log\frac{1}{a} - \frac{a\gamma}{4} \tag{16}$$

$$W_2^2(p_\star, q) = \left(\frac{1}{a} - \sqrt{\frac{2}{a(2-a\gamma)}}\right)^2 \tag{17}$$

Notice that both metrics are independent of the mean.

Plot for distance measures

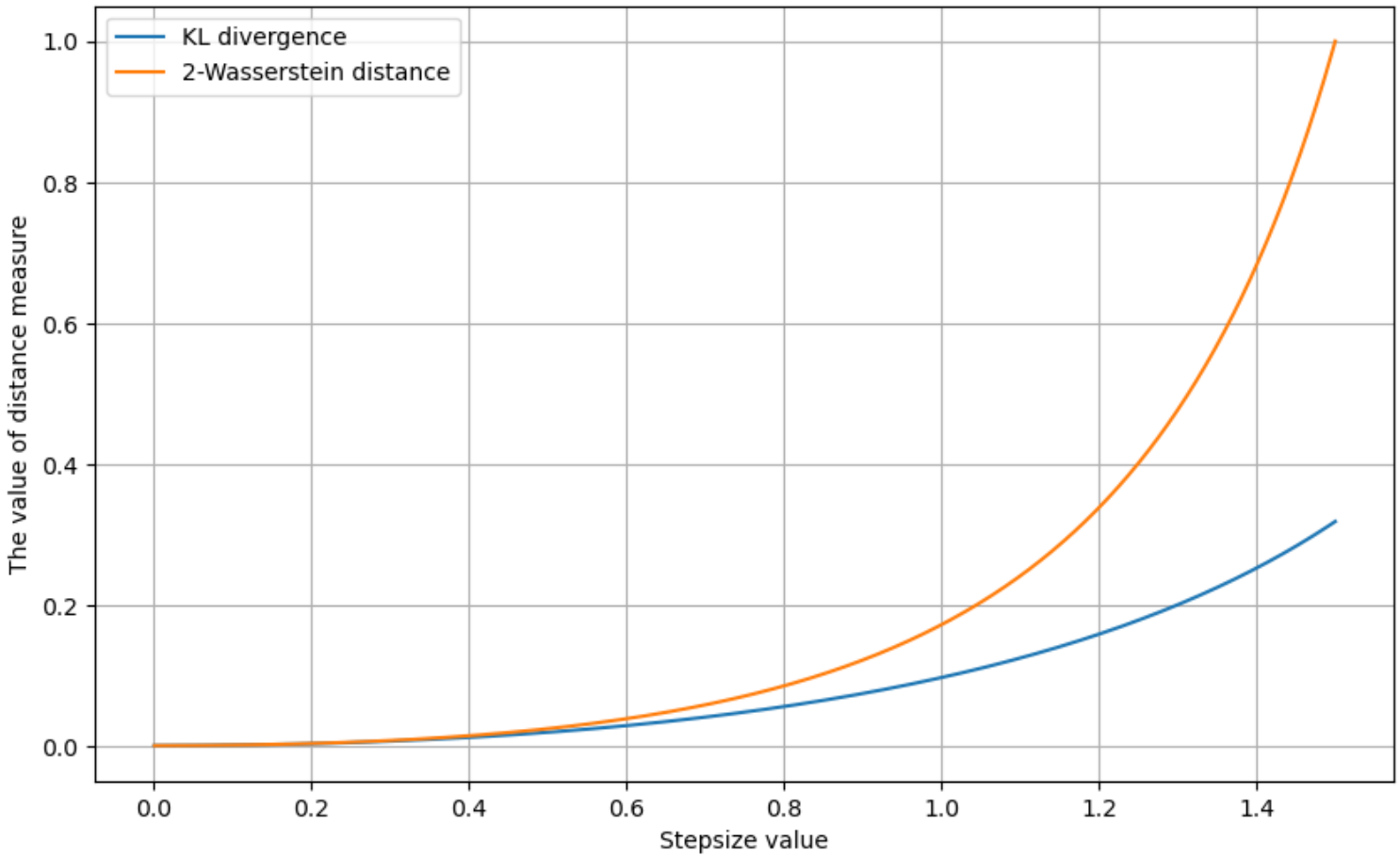
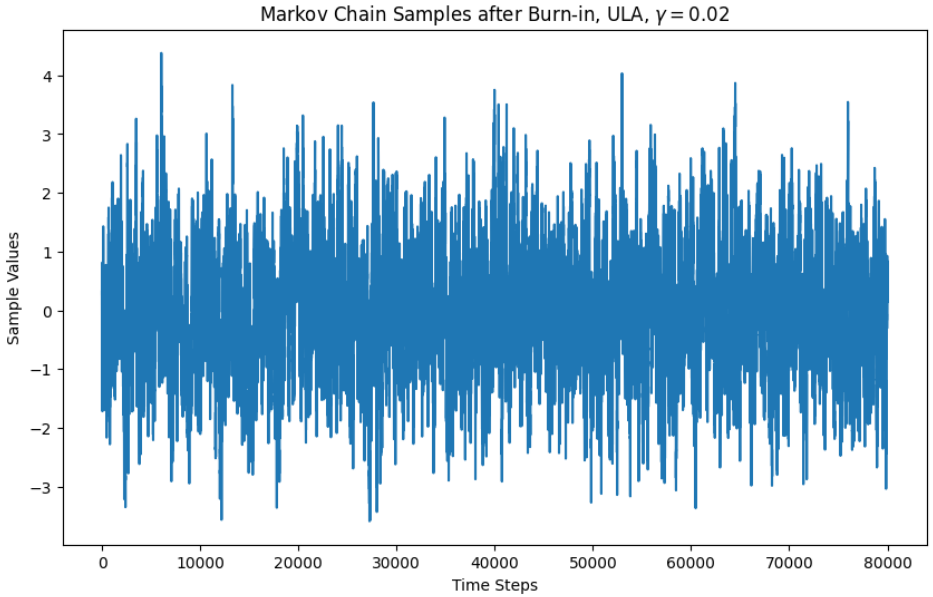


Figure: Graph for plotting the KL divergence (blue) and 2-Wasserstein distance (orange) on the special target $p_\star = \mathcal{N}(\mu, 1)$ (i.e. $a = 1$) and stationary distribution generated by ULA q as in (10).

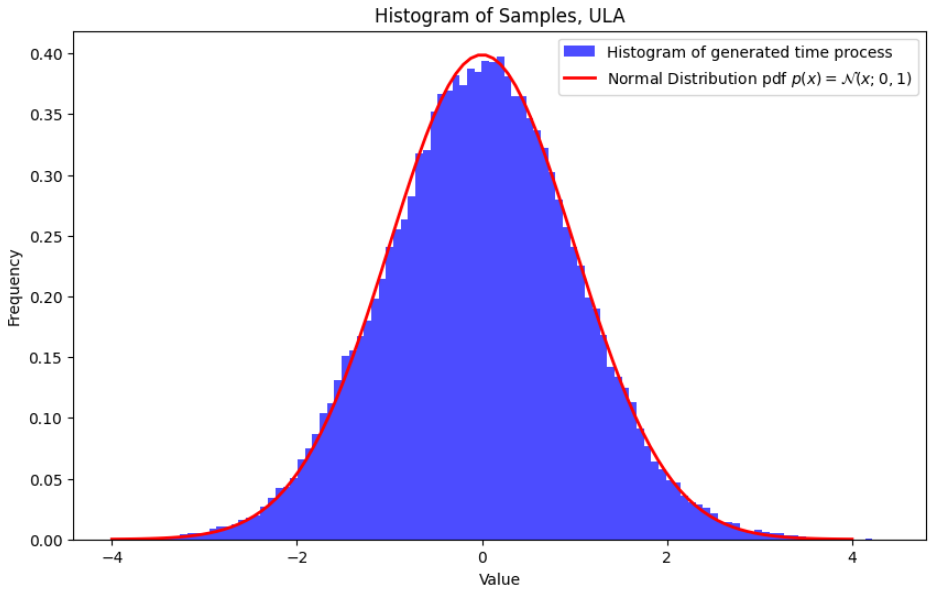
ULA sampling performance analysis

Quantitatively, the larger the value of distance measure, the stationary distribution would act more different compared with the target distribution, leading a worse performance of ULA for this stepsize. Therefore, **we would like to set a stepsize so that the distance measure is close to 0.**

ULA performance, with chosen stepsize $\gamma = 0.02$



(a) Markov chain samples, $\gamma = 0.02$

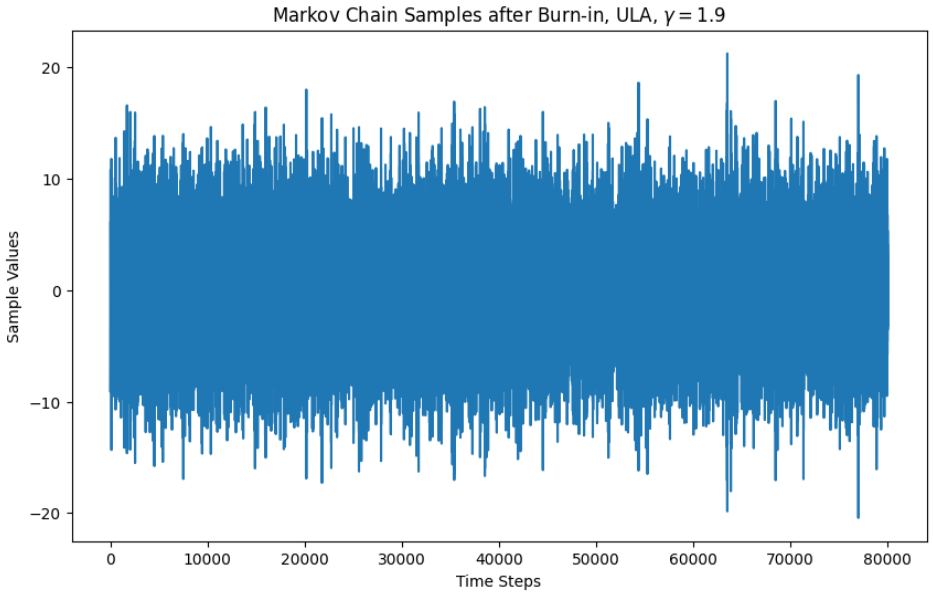


(b) Histogram and density plot, $\gamma = 0.02$

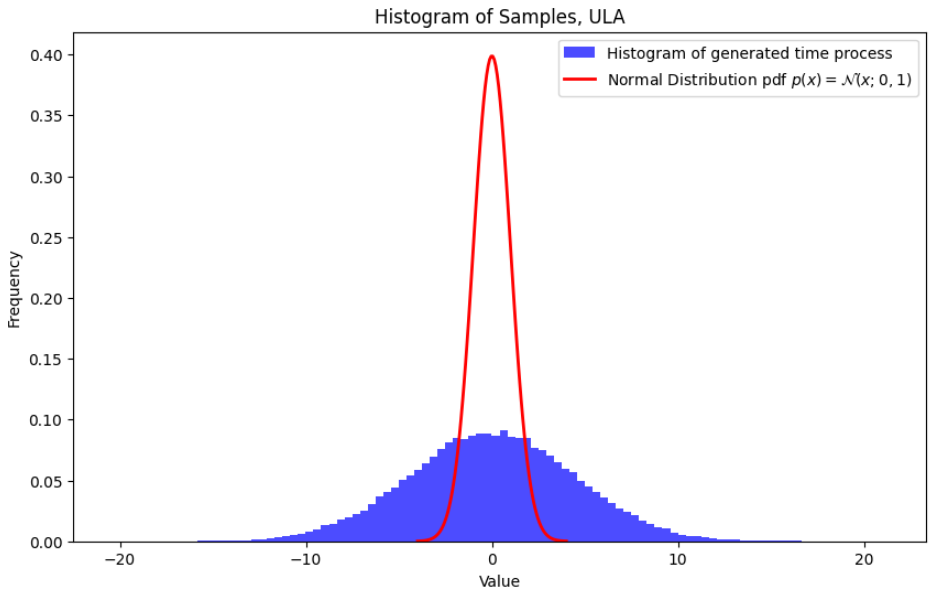
Calculated distance measure values:

- KL divergence: 2.517×10^{-5}
- 2-Wasserstein distance: 2.538×10^{-5}

ULA performance, with chosen stepsize $\gamma = 1.9$



(a) Markov chain samples for ULA, $\gamma = 1.9$



(b) Histogram and density plot, $\gamma = 1.9$

Calculated distance measure values:

- KL divergence: 1.023
- 2-Wasserstein distance: 12.06

ULA convergence behaviour

We would discuss the rate of convergence of ULA with a range of stepsizes, focus on the distribution at each time and measure the distance measures to the target distribution $p_\star$ in the form of (5), with certain stepsize γ. Recall the chain generated by ULA is in the form of (6), with the substitution $m = 1 - a\gamma$ and $n = -\gamma b$:

$$\begin{aligned} X_t = mX_{t-1} + n + v_t &= m(mX_{t-2} + n + v_{t-1}) + n + v_t \\ &= m^2 X_{t-2} + n(1+m) + (v_t + mv_{t-1}) \\ &\vdots \\ &= m^t x_0 + n(1 + m + \cdots + m^{t-1}) \\ &\quad + (v_t + mv_{t-1} + \cdots + m^{t-1} v_1) \\ &\sim \mathcal{N}\left(m^t x_0 + n \sum_{i=0}^{t-1} m^i, 2\gamma \sum_{i=0}^{t-1} m^{2i} \right) \end{aligned}$$

ULA convergence behaviour

Therefore, the distribution of the chain generated by ULA at time t is written as

$$X_t \sim \mathcal{N}\left(\left(1-a\gamma\right)^t x_0 - \gamma b \sum_{i=0}^{t-1}\left(1-a\gamma\right)^i, 2\gamma \sum_{i=0}^{t-1}\left(1-a\gamma\right)^{2i}\right) \tag{18}$$

ULA convergence behaviour: KL divergence

If we denote this distribution as q_t, then the KL divergence between this distribution and the target distribution would be

$$\begin{aligned}\mathrm{KL}(p_\star||q_t) = &\frac{1}{2}\log\left(2a\gamma\sum_{i=0}^{t-1}(1-a\gamma)^{2i}\right)\\ &+\frac{1}{2}\frac{1}{2a\gamma\sum_{i=0}^{t-1}(1-a\gamma)^{2i}}\\ &+\frac{1}{2}\frac{\left[-\frac{b}{a}-\left((1-a\gamma)^t x_0-\gamma b\sum_{i=0}^{t-1}(1-a\gamma)^i\right)\right]^2}{2\gamma\sum_{i=0}^{t-1}(1-a\gamma)^{2i}}-\frac{1}{2}\end{aligned} \tag{19}$$

ULA convergence behaviour: 2-Wasserstein distance

The 2-Wasserstein distance between this distribution and the target distribution is

$$W_2^2(p_\star, q_t) = \left(-\frac{b}{a} - \left[(1-a\gamma)^t x_0 - \gamma b \sum_{i=0}^{t-1} (1-a\gamma)^i \right] \right)^2 + \left(\sqrt{\frac{1}{a}} - \sqrt{2\gamma \sum_{i=0}^{t-1} (1-a\gamma)^{2i}} \right)^2 \tag{20}$$

We may then use the simpler 2-Wasserstein distance for our following discussion.

For illustration, we again use the example by setting the target distribution be $p_\star(x) = \mathcal{N}(x; 0, 1)$ with $a = 1, b = 0$, and we initialise the ULA chain by setting $x_0 = 1$.

ULA rate of convergence analysis: $\gamma = 0.02$

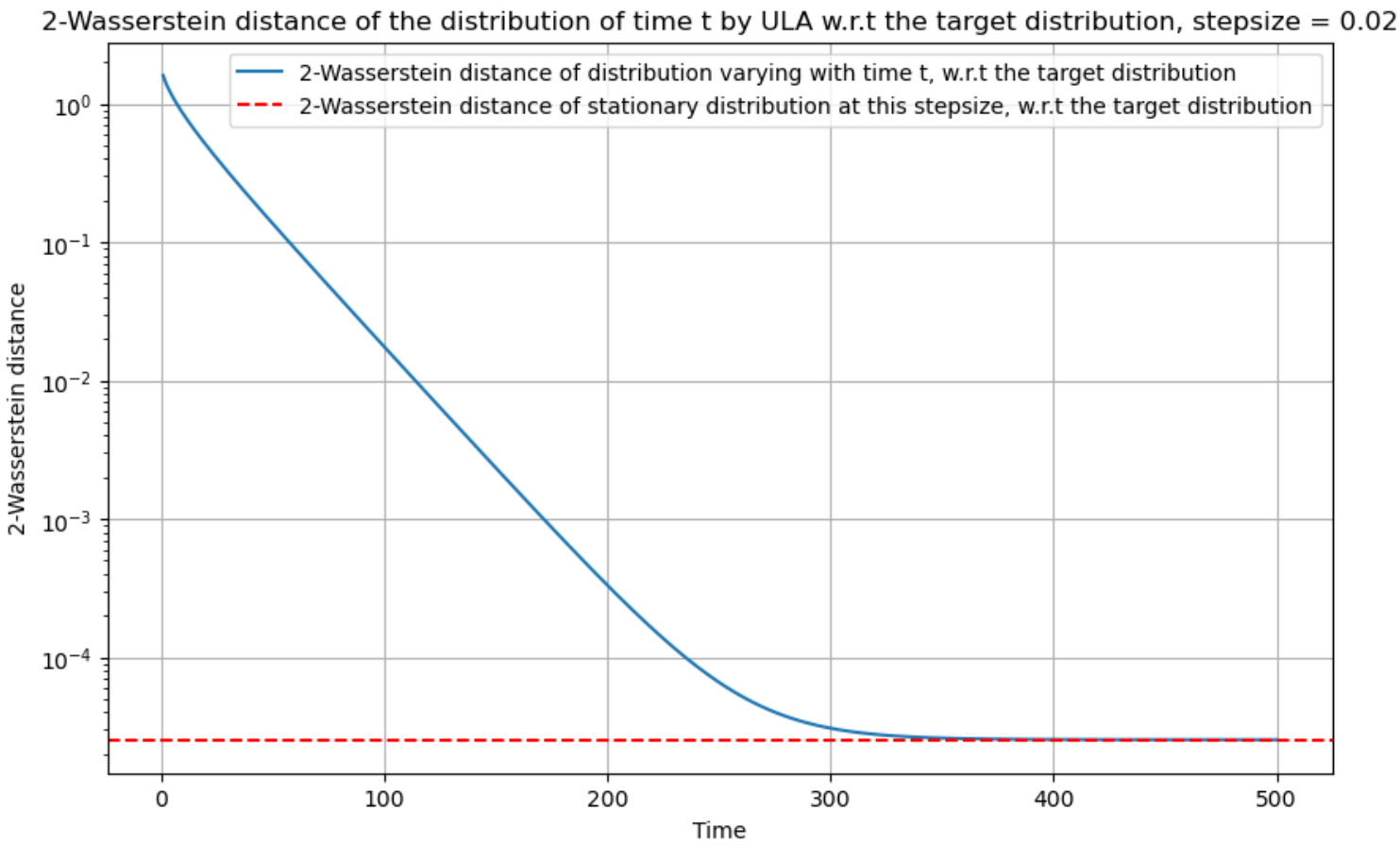


Figure: 2-Wasserstein distance of the distribution (18) with respect to the target $p_\star$, with $\gamma = 0.02$. Convergence to stationary distribution occurs at around $t = 400$.

ULA rate of convergence analysis: $\gamma = 0.001$

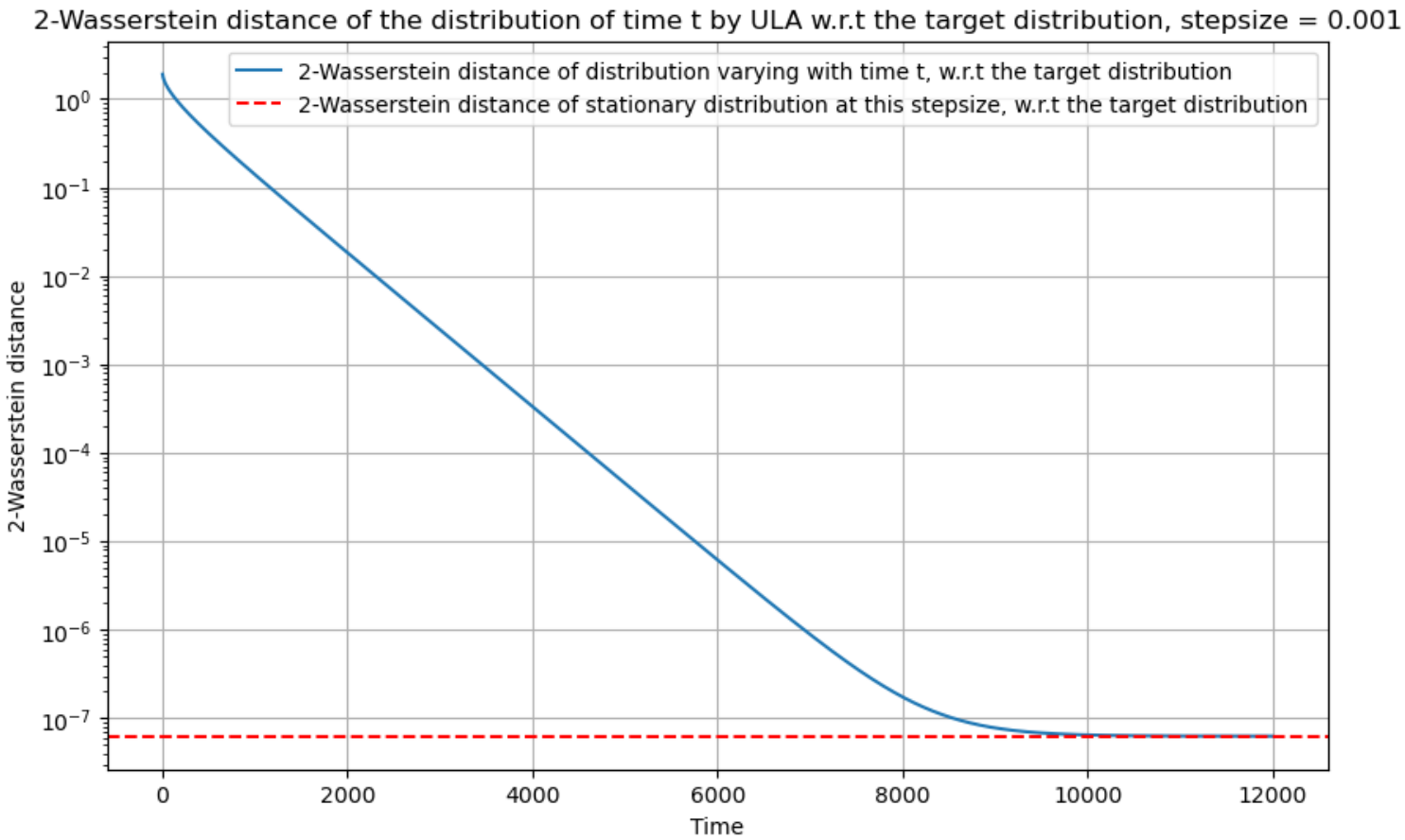


Figure: 2-Wasserstein distance of the distribution (18) with respect to the target $p_\star$, with $\gamma = 0.001$. Convergence to stationary distribution occurs after $t = 10000$.

Summary

'Trade-off'!

- Larger stepsize: lower sampling cost with faster convergence $\longleftrightarrow$ worse sampling performance.
- Smaller stepsize: higher sampling cost with slower convergence $\longleftrightarrow$ better sampling performance.

Table of Contents

Introduction: underdamped Langevin sampler algorithm

Used as a more common sampling algorithm in many areas.
In some sampling problems, the underdamped Langevin algorithm serves as an **'accelerated'** version of the overdamped Langevin system, which typically achieves better convergence rates by incorporating a momentum process $\mathbf{V}_t$.

Underdamped Langevin SDE

Underdamped Langevin algorithm comes from the underdamped Langevin SDE [8], which is given as a system of two SDEs

$$\begin{cases} \mathrm{d}V_t = \underbrace{-\eta V_t \mathrm{d}t}_{\text{friction}} \underbrace{-h(X_t)\mathrm{d}t}_{\text{acceleration}} + \sqrt{\frac{2\eta}{\beta}}\mathrm{d}B_t \\ \mathrm{d}X_t = V_t \mathrm{d}t \end{cases} \tag{21}$$

the above system of equations (21) is also known as the **kinetic Langevin diffusion**, or the **second-order Langevin process** [9], where

- $\{X_t, V_t\}_{t\geq 0}$ are called position and momentum process respectively, and we desire $\{X_t\}$, for sampling target density $p_\star(x)$.
- $h = \nabla U$, and U needs to be a strong convex and L-smooth function.

Euler-Maruyama discretisation

The Euler-Maruyama discretisation would convert (21) into the following system, where we set $\beta = 1$ without loss of generality.

$$\begin{cases} V_{t+1} = (1-\gamma\eta)V_t - \gamma\nabla U(X_t) + \sqrt{2\gamma\eta}\epsilon_{t+1} \\ X_{t+1} = X_t + \gamma V_t \end{cases} \tag{22}$$

If we set $U(x) = \frac{1}{2}ax^2 + bx + c$ as before, we already know $p_\star(x) = \mathcal{N}\left(x; -\frac{b}{a}, \frac{1}{a}\right) = \mathcal{N}(x; \mu_\star, \sigma^2)$, with substitution $-\frac{b}{a} = \mu_\star$ and $\frac{1}{a} = \sigma^2$, then (22) can be written as

$$\begin{pmatrix} V_{t+1} \\ X_{t+1} \end{pmatrix} = \begin{pmatrix} 1-\gamma\eta & -\frac{\gamma}{\sigma^2} \\ \gamma & 1 \end{pmatrix} \begin{pmatrix} V_t \\ X_t \end{pmatrix} + \begin{pmatrix} \frac{\gamma\mu_\star}{\sigma^2} \\ 0 \end{pmatrix} + \begin{pmatrix} \sqrt{2\gamma\eta} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \epsilon_{t+1} \\ \zeta_{t+1} \end{pmatrix}$$
$$\implies \mathbf{Z}_{t+1} = \mathbf{A}\mathbf{Z}_t + \mathbf{b} + \mathbf{C}\mathbf{W}_{t+1} \tag{23}$$

Euler-Maruyama discretisation

by letting

$$\mathbf{Z}_t = (V_t, X_t)^T, \quad \mathbf{A} = \begin{pmatrix} 1-\gamma\eta & -\frac{\gamma}{\sigma^2} \\ \gamma & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \frac{\gamma\mu_\star}{\sigma^2} \\ 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \sqrt{2\gamma\eta} & 0 \\ 0 & 0 \end{pmatrix} \tag{24}$$

and under the normality assumption, we would assume $\epsilon_t \sim \mathcal{N}(0,1), \zeta_t \sim \mathcal{N}(0,1)$, so then

$$\mathbf{W}_t = \begin{pmatrix} \epsilon_t \\ \zeta_t \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2) \tag{25}$$

Underdamped Langevin sampler algorithm

Algorithm 2: Underdamped Langevin sampler algorithm (underdamped LA, or ULD)

Input : N, the number of samples we want to take
$\mathbf{Z}_0 = \mathbf{z}_0$, as the initial momentum-position process vector
n, the number of burnin samples need to discard

Output: Approximately uncorrelated samples of target distribution $p_\star$

1 **for** $t = 1, 2, \cdots, N$ **do**

2 Take a random sample z' from the proposal

$$\mathbf{Z}' \sim q(\mathbf{z}'|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}'; \mathbf{A}\mathbf{z}_{t-1} + \mathbf{b}, \mathbf{C}\mathbf{C}^T)$$

which is equivalently to generate z' by $\mathbf{z}' = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{b} + \mathbf{C}\mathbf{w}_t$, for $\mathbf{W}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$;

3 Update the new state by $\mathbf{z}_t = \mathbf{z}'$.

4 **end**

5 **return** Discard first n burnin samples and return the remaining samples. The second entry of $\{\mathbf{Z}_t\}_{t\geq 0}$ is the required position process $\{X_t\}_{t\geq 0}$ that for sampling the target distribution $p_\star(x)$.

Stationary analysis

We are now interested in the stationary analysis of the chain

$$\mathbf{Z}_t = \mathbf{A}\mathbf{Z}_{t-1} + \mathbf{b} + \mathbf{C}\mathbf{W}_t \tag{26}$$

Such process looks similar to the a vector autoregressive process VAR(1) once we treat $\mathbf{CW_t} = \gamma_t$, and we may also 'center' this process by removing $(\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$ on both sides *(later we would show this is the mean of this process)*, so (26) can be written as the centered VAR(1) process

$$\boldsymbol{\xi}_t = \mathbf{A}\boldsymbol{\xi}_{t-1} + \gamma_t \tag{27}$$

where $\boldsymbol{\xi}_t = \mathbf{Z}_t - (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$. Meanwhile, by (25), we have

$$\mathbb{E}(\gamma_t) = \mathbf{0}, \qquad \operatorname{cov}(\gamma_i, \gamma_j) = \mathbb{E}(\gamma_i \gamma_j^T) = \delta_{ij}\mathbf{C}\mathbf{C}^T \tag{28}$$

where δ_{ij} is the Kronecker delta.

Stationary analysis: VAR(p) process

The p-th order vector autoregression VAR(p) is in the form of

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\epsilon}_t \tag{29}$$

where

- $\mathbf{c}$ denotes an $(n \times 1)$ vector of constants.
- $\mathbf{\Phi}_j$ an $(n \times n)$ matrix of autoregressive coefficients for $j = 1, 2, \cdots, p$.
- The $(n \times 1)$ vector $\boldsymbol{\epsilon}_t$ is a vector of white noise: $\mathbb{E}(\boldsymbol{\epsilon}_t) = 0$ and $\mathbb{E}(\boldsymbol{\epsilon_i} \boldsymbol{\epsilon_j^T}) = \delta_{ij} \mathbf{\Omega}$, for $\mathbf{\Omega}$ represents the covariance for the noise $\boldsymbol{\epsilon}_t$. [10]

Theorem (Conditions for stationarity, VAR(p) process)

A VAR(p) process defined as (29) is stationary as long as all values of z satisfying

$$|\mathbf{I}_n - \mathbf{\Phi}_1 z - \mathbf{\Phi}_2 z^2 - \cdots - \mathbf{\Phi}_p z^p| = 0$$

lie outside the unit disk $|z| < 1$ [10].

Stationary analysis

Therefore,

$$|\mathbf{I}_2 - \mathbf{A}z| = \left(1 - \gamma\eta + \frac{\gamma^2}{\sigma^2}\right) z^2 + (\gamma\eta - 2)z + 1 = 0$$

$$\implies z = \frac{2 - \eta\gamma \pm \sqrt{(\gamma\eta - 2)^2 - 4\left(1 - \gamma\eta + \frac{\gamma^2}{\sigma^2}\right)}}{2\left(1 - \gamma\eta + \frac{\gamma^2}{\sigma^2}\right)}$$

hence (26) is stationary if we have appropriate η, γ, σ to satisfy

$$\left| \frac{2 - \eta\gamma \pm \sqrt{(\gamma\eta - 2)^2 - 4\left(1 - \gamma\eta + \frac{\gamma^2}{\sigma^2}\right)}}{2\left(1 - \gamma\eta + \frac{\gamma^2}{\sigma^2}\right)} \right| > 1 \tag{30}$$

Stationary analysis: mean

Assume

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{Z}_t) = \begin{pmatrix} \mathbb{E}(V_t) \\ \mathbb{E}(X_t) \end{pmatrix}$$

since the process (26) is stationary, we can take expectations of both sides [10] and derive this expectation by

$$\mathbb{E}(\mathbf{Z}_t) = \mathbf{A}\mathbb{E}(\mathbf{Z}_{t-1}) + \mathbf{b} \implies \boldsymbol{\mu} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \implies \boldsymbol{\mu} = (\mathbf{I}_2 - \mathbf{A})^{-1}\mathbf{b} \quad (31)$$

Explicitly,

$$\boldsymbol{\mu} = (\mathbf{I}_2 - \mathbf{A})^{-1}\mathbf{b} = \frac{\sigma^2}{\gamma^2} \begin{pmatrix} 0 & -\frac{\gamma}{\sigma^2} \\ \gamma & \gamma\eta \end{pmatrix} \begin{pmatrix} \frac{\gamma\mu_\star}{\sigma^2} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \mu_\star \end{pmatrix} = \begin{pmatrix} \mathbb{E}(V_t) \\ \mathbb{E}(X_t) \end{pmatrix} \quad (32)$$

Stationary analysis: autocovariance sequence (acv)

We then consider the autocovariance sequence (acv): $s_\tau = \mathrm{cov}(\mathbf{Z}_t, \mathbf{Z}_{t+\tau})$ for this process. As stated, we know that our target process (26) is equivalent to the centered VAR(1) process (27).
Since $\boldsymbol{\xi}_t = \mathbf{Z}_t - (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \mathbf{Z}_t - \boldsymbol{\mu}$ and $\boldsymbol{\mu}$ is a constant vector, we shall directly have

$$\mathrm{cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+\tau}) = \mathrm{cov}(\mathbf{Z}_t, \mathbf{Z}_{t+\tau}) = s_\tau$$

Now by (27), we can easily expand $\boldsymbol{\xi}_{t+\tau}$ recursively as

$$\begin{aligned}
\boldsymbol{\xi}_{t+\tau} = \mathbf{A}\boldsymbol{\xi}_{t+\tau-1} + \boldsymbol{\gamma}_{t+\tau} &= \mathbf{A}(\mathbf{A}\boldsymbol{\xi}_{t+\tau-2} + \boldsymbol{\gamma}_{t+\tau-1}) + \boldsymbol{\gamma}_{t+\tau} \\
&= \mathbf{A}^2\boldsymbol{\xi}_{t+\tau-2} + (\mathbf{A}\boldsymbol{\gamma}_{t+\tau-1} + \boldsymbol{\gamma}_t) \\
&= \cdots \\
&= \mathbf{A}^\tau \boldsymbol{\xi}_t + \sum_{k=0}^{\tau-1} \mathbf{A}^k \boldsymbol{\gamma}_{t+\tau-k}
\end{aligned}$$

Stationary analysis: autocovariance sequence (acv)

Therefore, the acv can be derived as

$$s_\tau = \mathrm{cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t+\tau}) = \mathrm{cov}\left(\boldsymbol{\xi}_t, \mathbf{A}^\tau \boldsymbol{\xi}_t + \sum_{k=0}^{\tau-1} \mathbf{A}^k \boldsymbol{\gamma}_{t+\tau-k}\right) \tag{33}$$

$$= \mathrm{cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t)(\mathbf{A}^\tau)^T + \mathrm{cov}\left(\boldsymbol{\xi}_t, \sum_{k=0}^{\tau-1} \mathbf{A}^k \boldsymbol{\gamma}_{t+\tau-k}\right) \tag{34}$$

$$= \mathrm{cov}(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t)(\mathbf{A}^\tau)^T = s_0(\mathbf{A}^\tau)^T \tag{35}$$

because $\boldsymbol{\xi}_t = \sum_{k=0}^{\infty} \mathbf{A}^k \boldsymbol{\gamma}_{t-k}$, and with (28) and the linearity of covariance, we would have $\mathrm{cov}\left(\boldsymbol{\xi}_t, \sum_{k=0}^{\tau-1} \mathbf{A}^k \boldsymbol{\gamma}_{t+\tau-k}\right) = \mathbf{0}$.

Stationary analysis: variance

If we directly apply variance operator on both sides of (27), then

$$s_0 = \mathbf{A} s_0 \mathbf{A}^T + \mathbf{C}\mathbf{C}^T \tag{36}$$

where $s_0 = \mathrm{cov}(\mathbf{Z}_t, \mathbf{Z}_t)$ is the variance of the chain (26).
In order to solve s_0 explicitly, the **vec** operator can be used to obtain a closed-form solution to (36).

vec operator

Definition (**vec** operator)

If $\mathbf{A}$ is an $(m \times n)$ matrix, then $\mathbf{vec}(\mathbf{A})$ is an $(mn \times 1)$ column vector obtained by stacking the columns of $\mathbf{A}$, one below the other, with the columns ordered from left to right. For example, if

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \in \mathbb{R}^{3\times 2}$$

then $\mathbf{vec}(\mathbf{A}) = (a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32})^T \in \mathbb{R}^6$. [10]

Kronecker product

Definition (Kronecker product)

Let $\mathbf{A} \in \mathbb{R}^{K \times L}$ and $\mathbf{B} \in \mathbb{R}^{M \times N}$ be two matrices, then the Kronecker product between $\mathbf{A}$ and $\mathbf{B}$ is the $(KM \times LN)$ block matrix that

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1L}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{K1}\mathbf{B} & \cdots & a_{KL}\mathbf{B} \end{pmatrix}$$

where a_{ij} represents the ij-th entry of matrix $\mathbf{A}$ as a real scalar, which means that $a_{ij}\mathbf{B}$ is a matrix that its pq-th entry is $a_{ij}b_{pq}$. [11]

Stationary analysis: variance

Theorem

Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *be matrices whose dimensions are such that the product* $\mathbf{ABC}$ *exists. Then*

$$\mathbf{vec}(\mathbf{ABC}) = (\mathbf{C}^T \otimes \mathbf{A})\mathbf{vec}(\mathbf{B})$$

where the symbol $\otimes$ *denotes the Kronecker product. [10]*

Therefore, we can have

$$\mathbf{vec}(s_0) = \mathbf{vec}(\mathbf{A}s_0\mathbf{A}^T + \mathbf{CC^T}) = (\mathbf{A} \otimes \mathbf{A})\mathbf{vec}(s_0) + \mathbf{vec}(\mathbf{CC}^T) \quad (37)$$

with $\mathcal{A} = \mathbf{A} \otimes \mathbf{A}$ and the rearrangement, we can solve the variance s_0 as

$$(\mathbf{I}_4 - \mathcal{A})\mathbf{vec}(s_0) = \mathbf{vec}(\mathbf{CC}^T) \implies \mathbf{vec}(s_0) = (\mathbf{I}_4 - \mathcal{A})^{-1}\mathbf{vec}(\mathbf{CC}^T) \quad (38)$$

Stationary, marginal distribution of position chain

Finally, the closed-form of variance s_0 for the chain (26) is

$$s_0 = \begin{pmatrix} -\frac{4\eta\sigma^4}{2\eta^2\gamma\sigma^4-3\eta\gamma^2\sigma^2-4\eta\sigma^4+\gamma^3+4\gamma\sigma^2} & \frac{2\eta\gamma\sigma^4}{2\eta^2\gamma\sigma^4-3\eta\gamma^2\sigma^2-4\eta\sigma^4+\gamma^3+4\gamma\sigma^2} \\ \frac{2\eta\gamma\sigma^4}{2\eta^2\gamma\sigma^4-3\eta\gamma^2\sigma^2-4\eta\sigma^4+\gamma^3+4\gamma\sigma^2} & \frac{2\eta\sigma^4(\eta\gamma\sigma^2-\gamma^2-2\sigma^2)}{2\eta^2\gamma\sigma^4-3\eta\gamma^2\sigma^2-4\eta\sigma^4+\gamma^3+4\gamma\sigma^2} \end{pmatrix} \tag{39}$$

so the stationary distribution is

$$\mathbf{Z} = (\mathbf{V}, \mathbf{X}) \sim \mathcal{N}(\boldsymbol{\mu}, s_0) \tag{40}$$

and our desired, stationary marginal distribution of the position chain $\mathbf{X}_t$ is

$$\mathbf{X} \sim \mathcal{N}\left(\mu_\star, \frac{2\eta\sigma^4(\eta\gamma\sigma^2-\gamma^2-2\sigma^2)}{2\eta^2\gamma\sigma^4-3\eta\gamma^2\sigma^2-4\eta\sigma^4+\gamma^3+4\gamma\sigma^2}\right) \tag{41}$$

ULD sampling performance analysis

Again, we are going to use KL divergence and 2-Wasserstein distance to measure the difference between the target distribution $p_\star(x) = \mathcal{N}(x; \mu, \sigma^2)$ and the marginal, stationary distribution q we've derived as in (41).
If we set $\sigma_q^2 = \frac{2\eta\sigma^4(\eta\gamma\sigma^2 - \gamma^2 - 2\sigma^2)}{2\eta^2\gamma\sigma^4 - 3\eta\gamma^2\sigma^2 - 4\eta\sigma^4 + \gamma^3 + 4\gamma\sigma^2}$, then we would have

$$\mathrm{KL}(p_\star || q) = \log\left(\frac{\sigma_q}{\sigma}\right) + \frac{\sigma^2}{2\sigma_q^2} - \frac{1}{2} \tag{42}$$

$$W_2^2(p_\star, q) = (\sigma - \sigma_q)^2 \tag{43}$$

As before, both metrics are independent of the mean.

ULD sampling performance analysis: varying stepsize γ

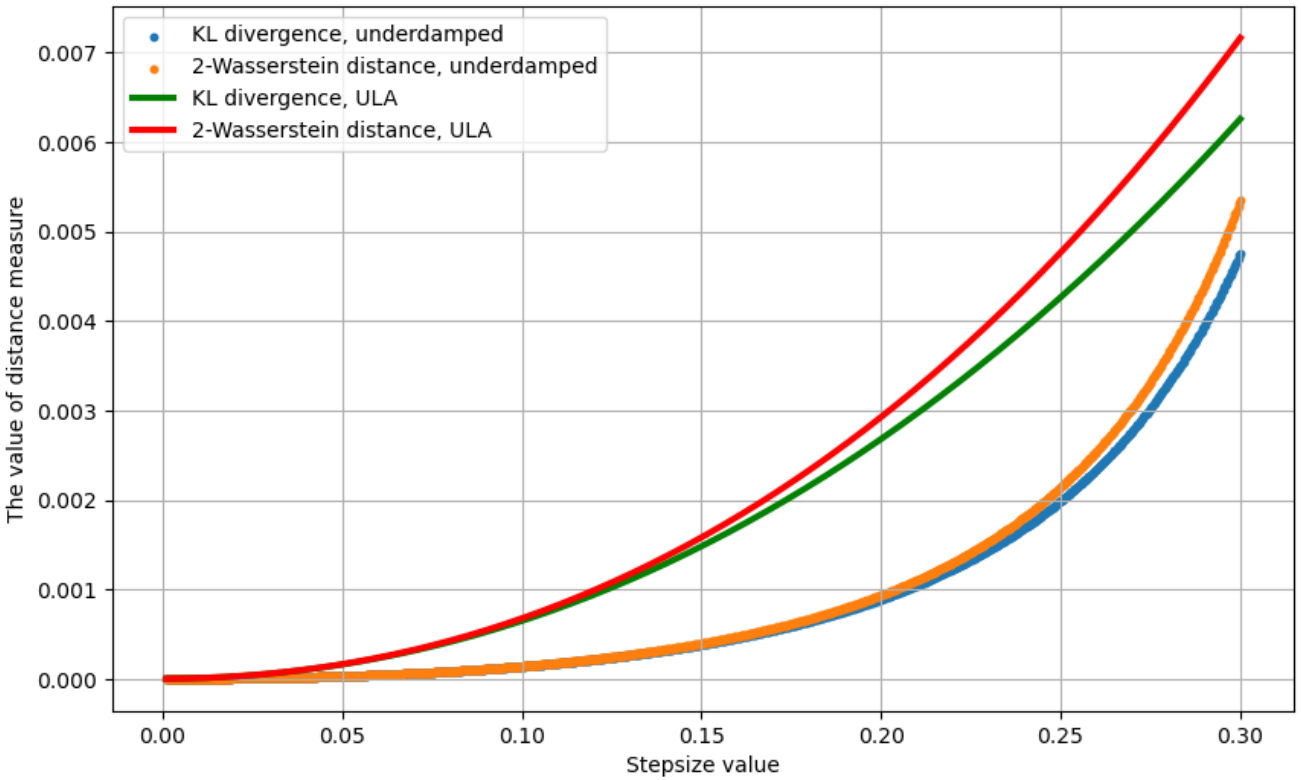
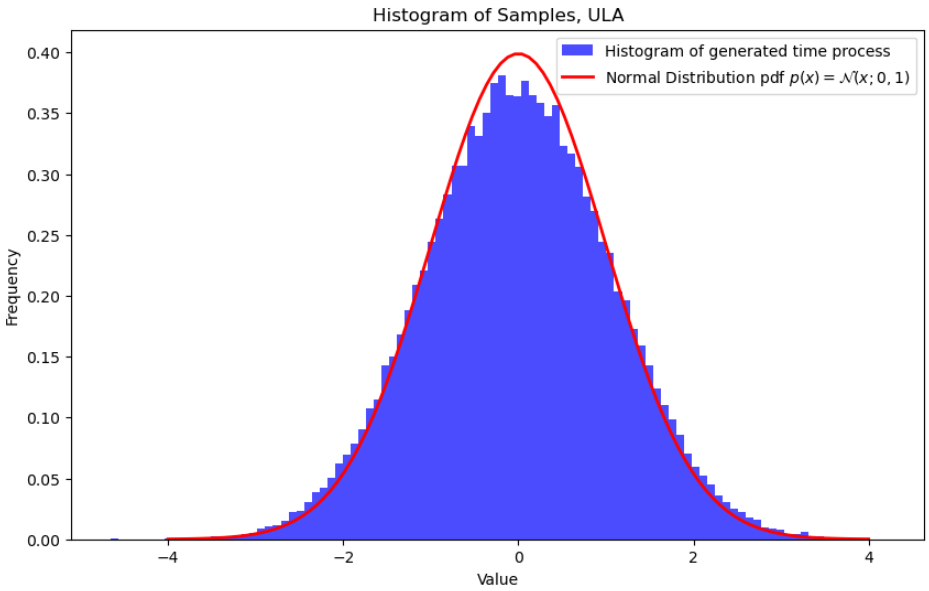
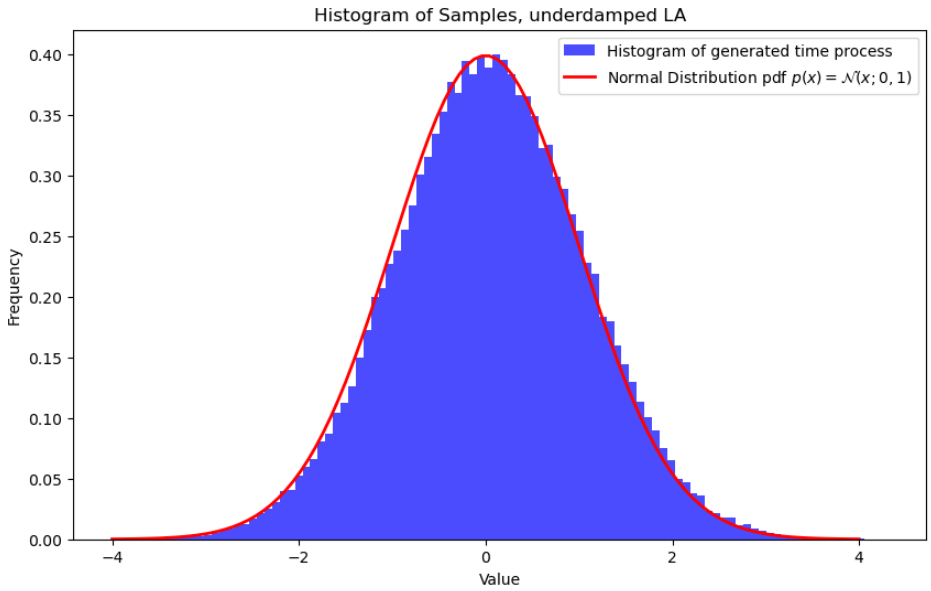


Figure: Graph for the KL divergence (blue) and 2-Wasserstein distance (orange) on the special target $p_\star = \mathcal{N}(\mu_\star, 1)$ and marginal distribution generated by ULD q as in (41), with the friction coefficient $\eta = 5$, together with KL divergence (green) and 2-Wasserstein distance (red) on $p_\star$ and stationary distribution generated by ULA as in (10).

ULD sampling performance analysis



(a) ULA at $\gamma = 0.2$



(b) ULD with $\eta = 5$ at $\gamma = 0.2$

Figure: The above set of histogram plots shows that the underdamped Langevin diffusion (ULD) would result a better sampling than the ULA with the same stepsize, for a closer match with respect to the true density curve.

ULD sampling performance analysis: varying friction η

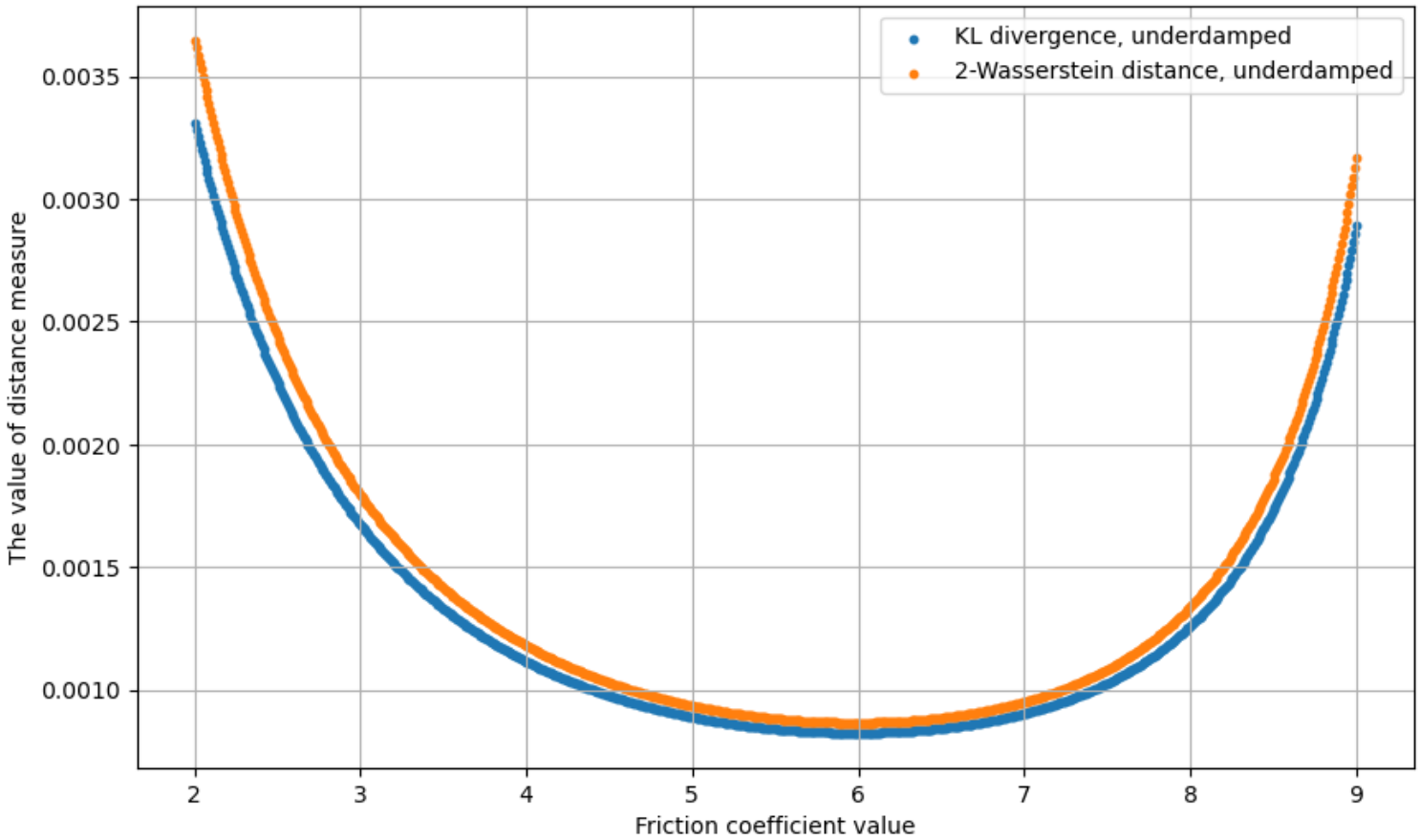


Figure: Graph for scatter plotting the KL divergence (blue) and 2-Wasserstein distance (orange) on the target $p_\star = \mathcal{N}(\mu_\star, 1)$ and stationary, marginal distribution generated by underdamped Langevin algorithm q as in (41), with fixed stepsize $\gamma = 0.2$.

ULD convergence behaviour analysis

Unlike ULA, it is really hard to derive the exact distribution of $\mathbf{Z}_t$ at each time t from the chain (23), along with the marginal distribution of position chain X_t. Therefore, we now instead using computational 1-Wasserstein distance for this purpose, which involves in-built function `wasserstein_distance` in `scipy.stats` in Python.

ULD convergence behaviour, $\gamma = 0.2, \eta = 5$

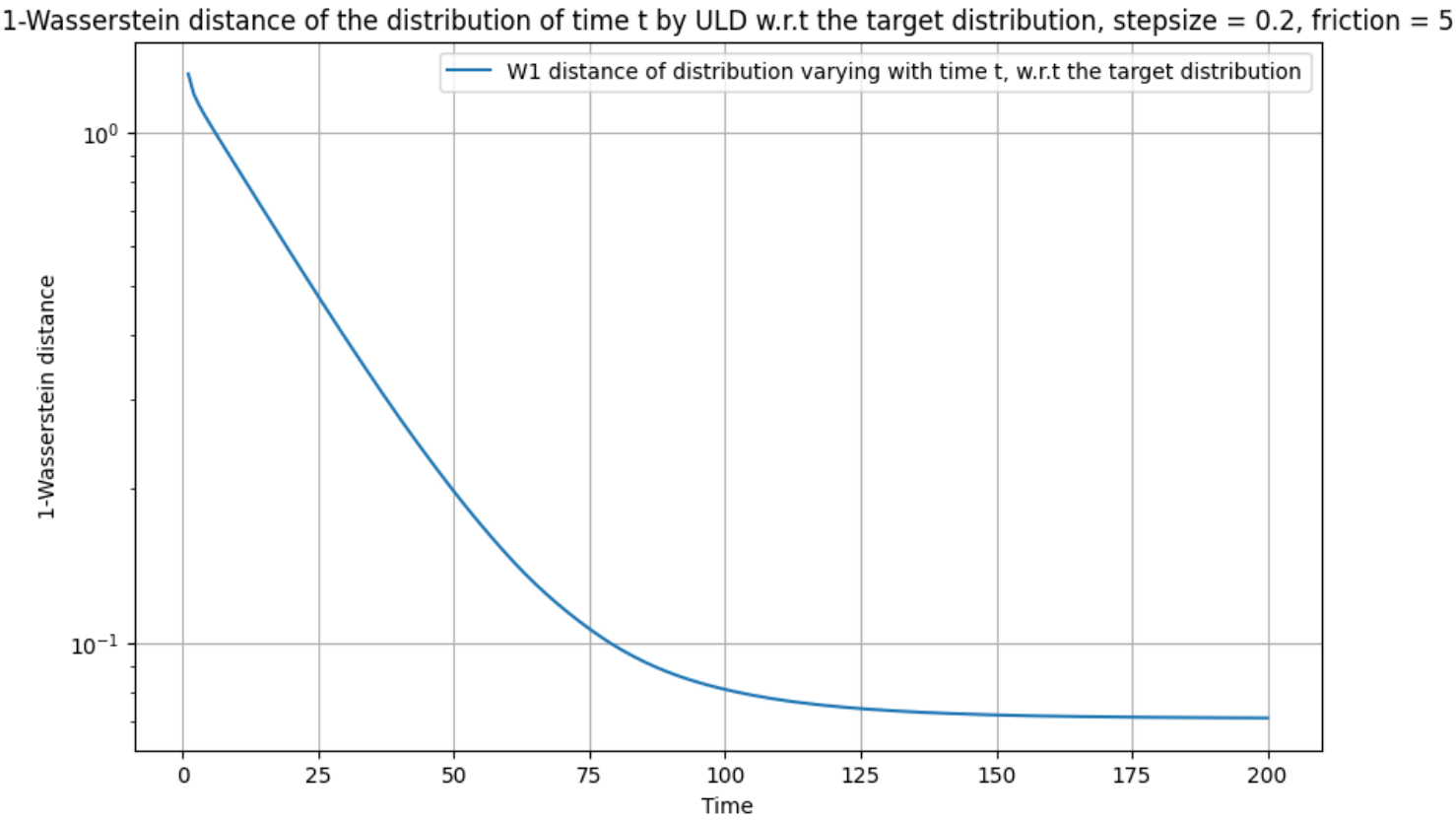


Figure: Computational 1-Wasserstein distance of the position chain X_t from the marginal distribution of ULD, with $\gamma = 0.2, \eta = 5$. Note the convergence of 1-Wasserstein distance approximately occurs after $t = 150$, meaning a convergence to the stationary marginal distribution (41).

ULD convergence behaviour, $\gamma = 0.02, \eta = 5$

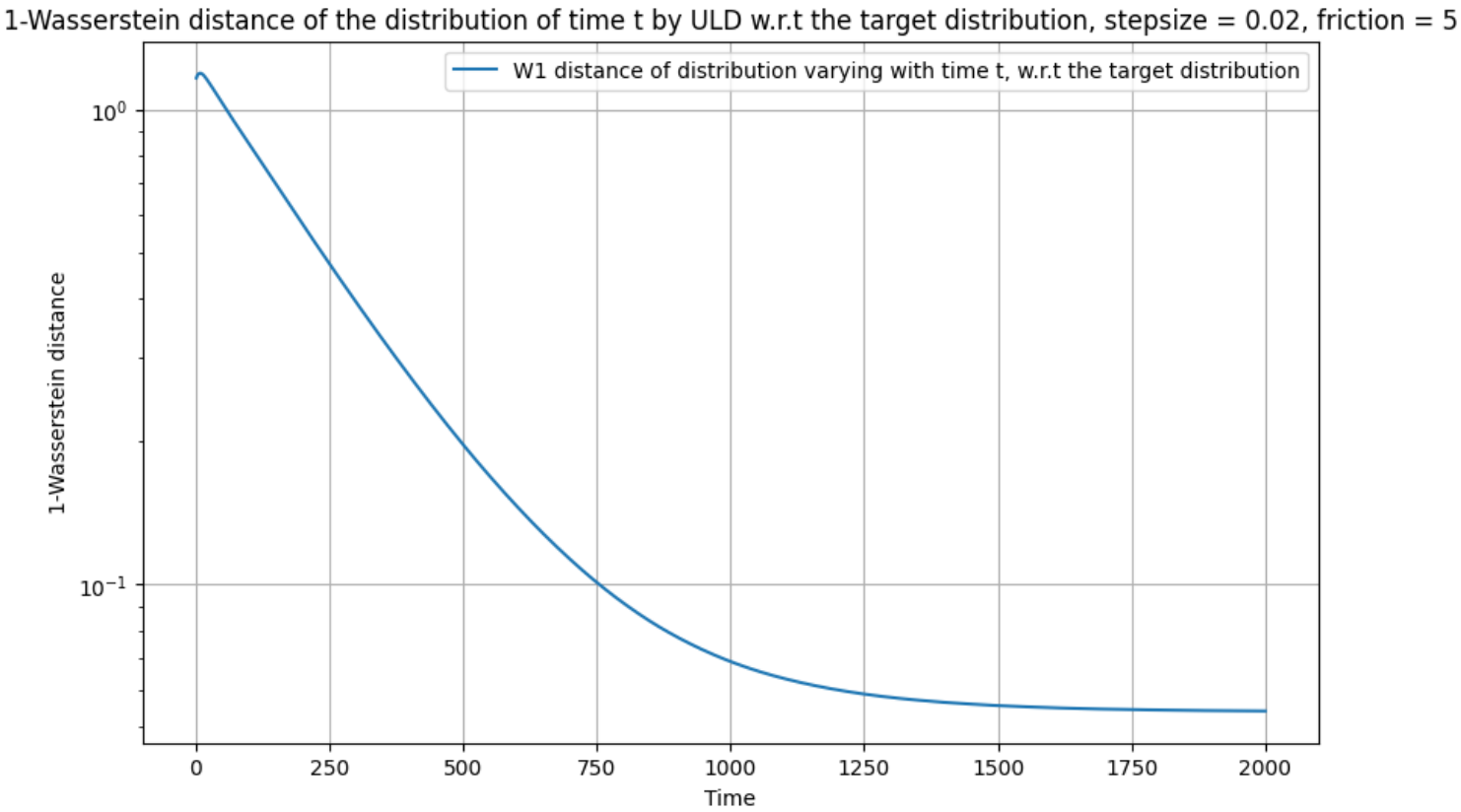


Figure: Computational 1-Wasserstein distance of the position chain X_t from the marginal distribution of ULD, with $\gamma = 0.02, \eta = 5$. Note this time the convergence of 1-Wasserstein distance approximately occurs after $t = 1500$, implying a ten times slower convergence to (41) than the previous one as in Figure 10.

Correlation of ULD samples

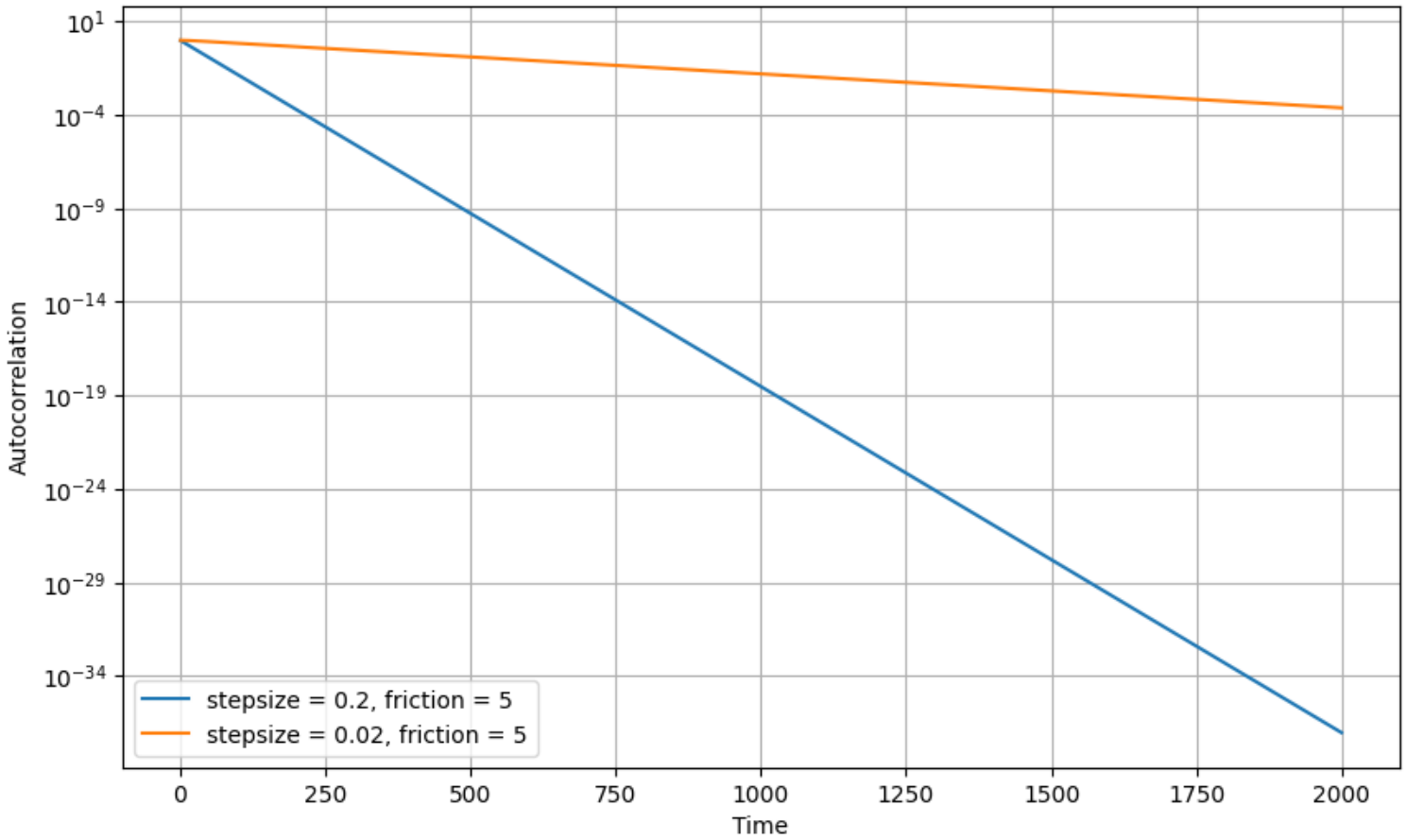


Figure: The autocorrelation of position chain X_t along with time with different parameters. The autocorrelation is in the log-scale.

Table of Contents

Conclusion

- Analyse sampling algorithms from both theoretical and empirical perspectives.
- Derive analytical expressions for the stationary distributions, and examine the asymptotic behaviours for both unadjusted and underdamped Langevin algorithms.
- Evaluate sampling performance and convergence behaviour to the asymptotic, stationary distributions.

Future perspectives

- Other possible discretisations on the underdamped Langevin SDE, with one introduced in the report.
- High-dimensional target distribution analysis.

Thank you for your listening!

Questions?

References

Marco Taboga.
"*Markov Chain Monte Carlo (MCMC) methods*", Lectures on probability theory and mathematical statistics.
https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo, 2021.
[Online; accessed June 2, 2024].

Alain Durmus and Éric Moulines.
"*The Langevin MCMC: Theory and Methods - Course 2*".
https://www.icts.res.in/sites/default/files/paap-2019-08-09-Eric%20Moulines.pdf, August 9, 2019.
[Online; accessed February 27, 2024].

Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis.
The tamed unadjusted langevin algorithm.
Stochastic processes and their applications, 129(10):3638–3663, 2019.

Matthieu Stigler.