

1 Comparison over different target distributions

Different sampling methods, namely Metropolis-Hastings with a Gaussian random walk proposal (RW), MALA and ULA, are used to sample repeatedly from different target distributions. The target distributions are mixtures of Gaussians. For a certain point in time, the Wasserstein metric between the true distribution and the sampling distribution is approximated empirically by computing the Wasserstein metric between the uniform distribution over k consecutive samples from the chain and k independent samples from the target distribution. The above is repeated at various points in time to illustrate how the distance changes with time.

Since each run is highly variable, the above is repeated for a number of independent runs to give a confidence interval for the estimated distance. On each run, the chain is initialised at a random point picked uniformly from $(-10, 0)$. The banded region shows the region which is within one standard deviation of the mean distance.

To investigate the behaviour at different points in time, different plots are made for the burnin period (first 200 samples) and the 1000 samples after that. The choice of k , number of samples used to estimate the Wasserstein metric, is different for the burnin period and the next 1000 samples so the distances cannot be compared directly.



It seems that across the three targets, random walk performs the best, followed closely by ULA. MALA seems to perform substantially worse.

This is a bit surprising because

- MALA and ULA are similar in the sense that they are both gradient based
- MALA and RW are similar in the sense that they are metropolis-based

Maybe the difference is that MALA gets stuck at local maxima as shown later.

TODO:

1. Check whether the method of computing empirical Wasserstein metric using consecutive samples from the same run actually make sense

2 Trajectory of different sampling methods

In this section, the trajectory of different sampling methods are compared. The target distribution is the rightmost target pdf from the previous section:

$$\frac{1}{7}\mathcal{N}(-10, 1) + \frac{2}{7}\mathcal{N}(-5, 1) + \frac{4}{7}\mathcal{N}(0, 1)$$

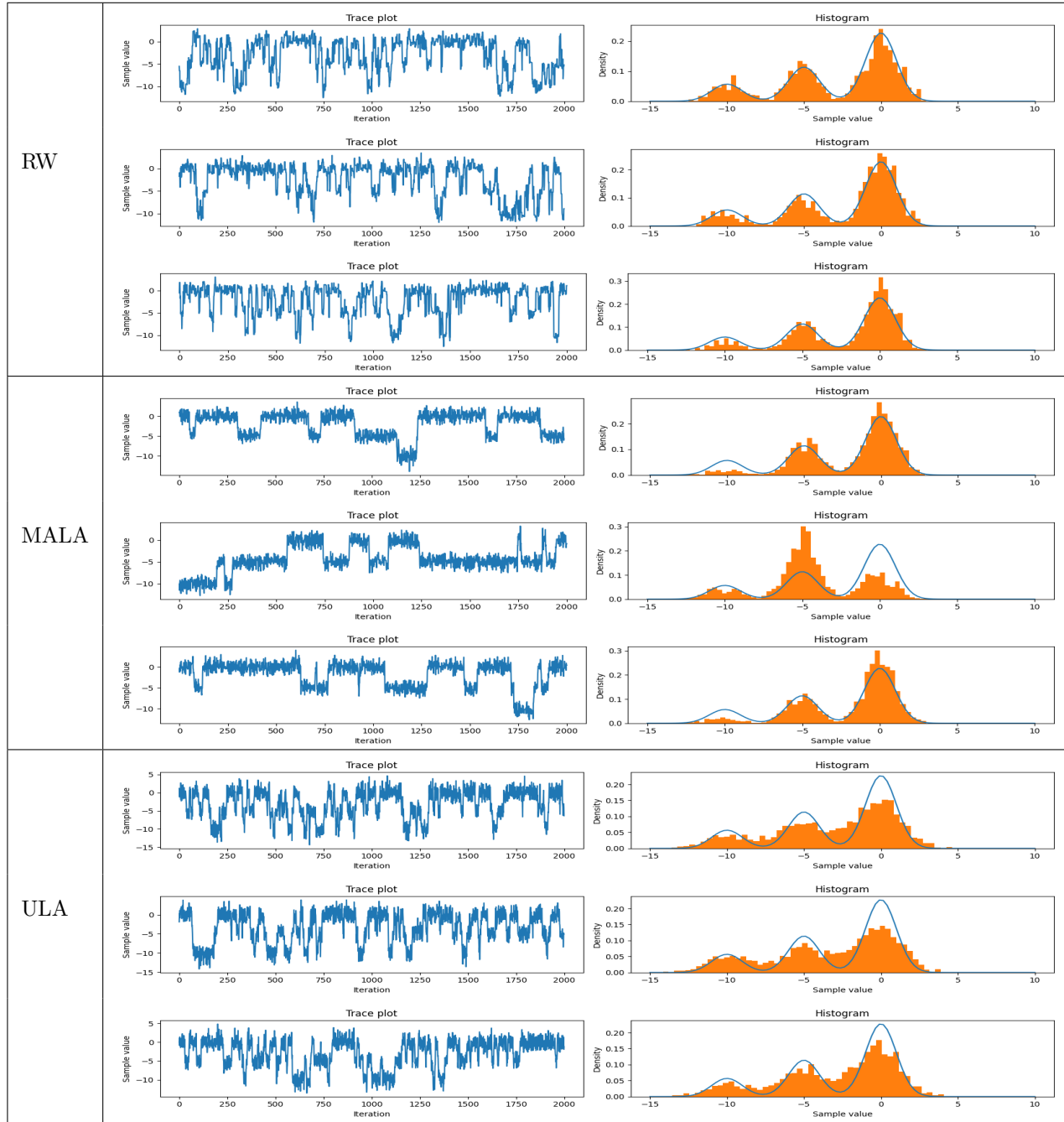
Trajectories of different sampling methods are plotted below. To ensure a fair comparison, all chains start at the mode of the leftmost peak (i.e. $x = -10$).

The plots are split into short run and long run, which aim to compare the non-asymptotic behaviour and the asymptotic behaviour respectively.

For short runs, three independent chains are sampled for each method. For each chain, 3000 iterations are ran and the first 1000 are discarded.

For long runs, one chain is sampled for each method and 11000 iterations are ran, the first 1000 of which are discarded.

2.1 Short run



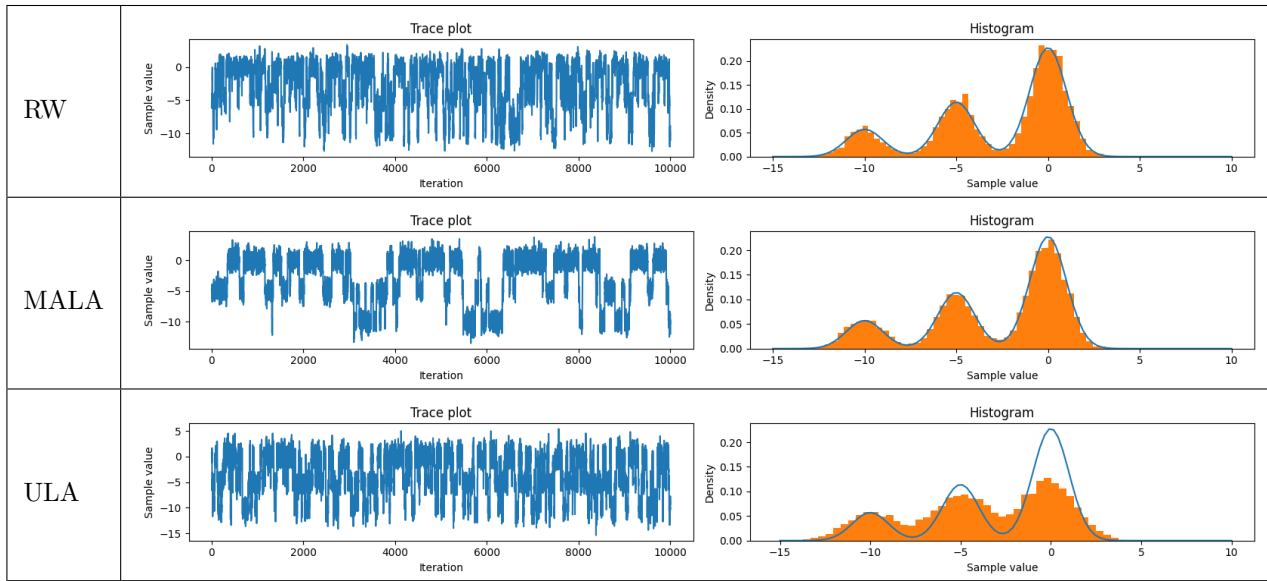
2.1.1 Observations

1. MALA showed a bias towards local maxima of the pdf
2. MALA shows long intervals of similar values on the trace plot
3. MALA shows the greatest variation among different runs
4. ULA showed a bias towards local minima of the pdf

The first two observations are due to the gradient informed nature of MALA, which favours local maxima and makes it difficult to move from one maximum to another since it involves traversing a local minimum. The third observation is related to the above in the sense that the trajectory heavily depends on which peak the simulation gets stuck on.

On the other hand, ULA has a bias towards local minima while MALA does not. The metropolis step in MALA rejects candidates which have a low probability under the target distribution and keeps the chain away from local minima. For ULA, no such correction is done so samples are relatively more likely to reach local minima.

2.2 Long run



2.2.1 Observations

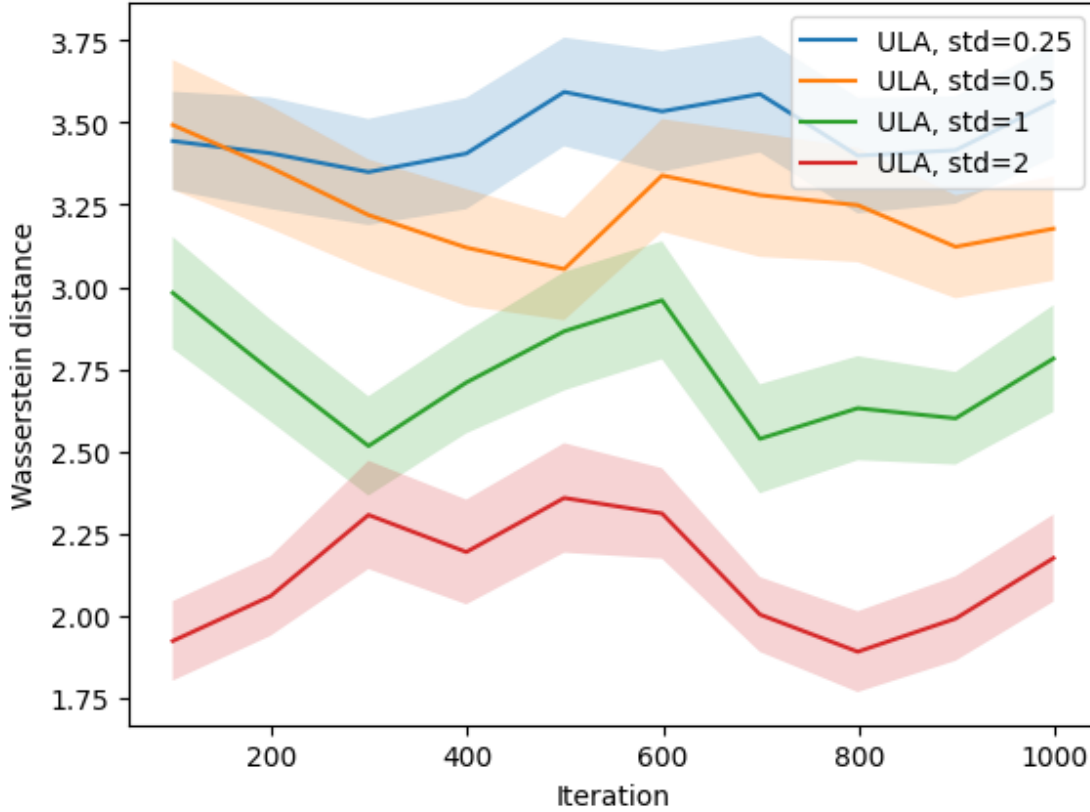
Both Metropolis-based algorithms have empirical distributions which converge to the target pdf. ULA on the other hand does not converge to the target pdf, which is probably due to the discretisation error

3 Comparing different step sizes for ULA

In the previous section, it is noted that the discretisation error may cause ULA to converge to a biased distribution. This section aims to investigate the impact of the step size on the behaviour of the ULA sampler.

3.1 Wasserstein metric

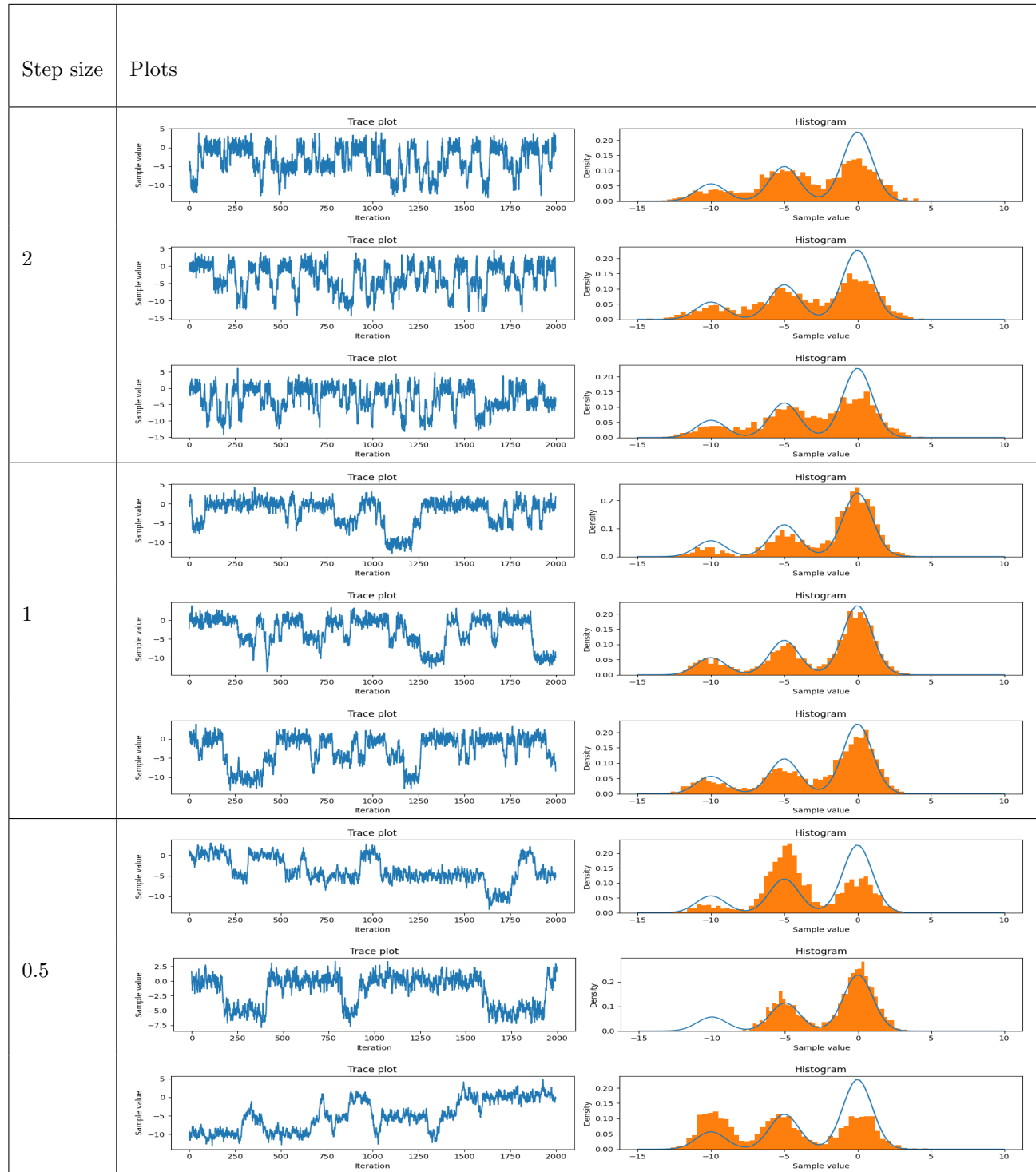
The Wasserstein metric for different step sizes / standard deviation is computed as before.

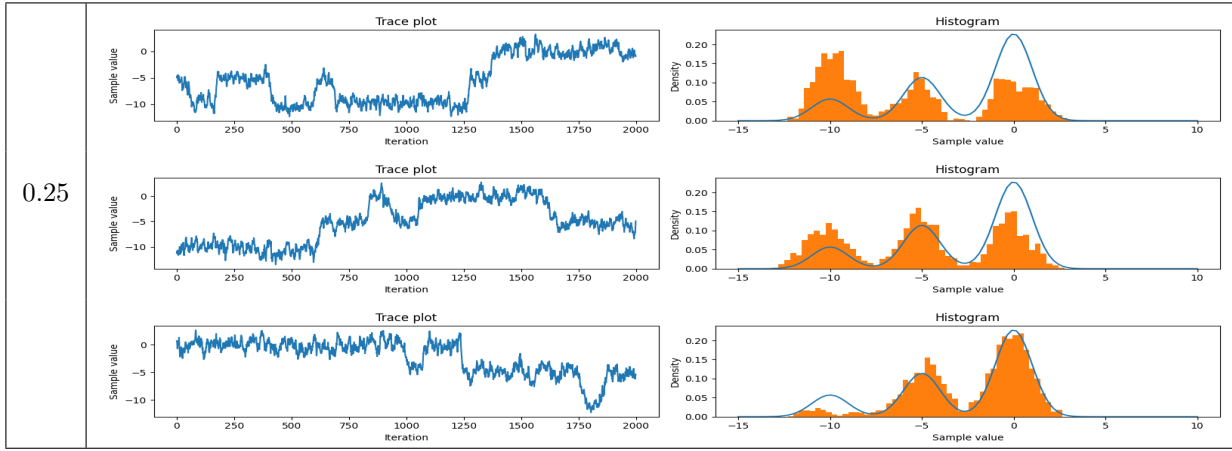


Remark. I am not sure what to make of this plot - it seems that the bigger the step size, the lower the distance. However, since the distance is computed using samples along the same chain, this might just be indicative of the fact that larger step sizes lead to samples which are less correlated along the same chain.

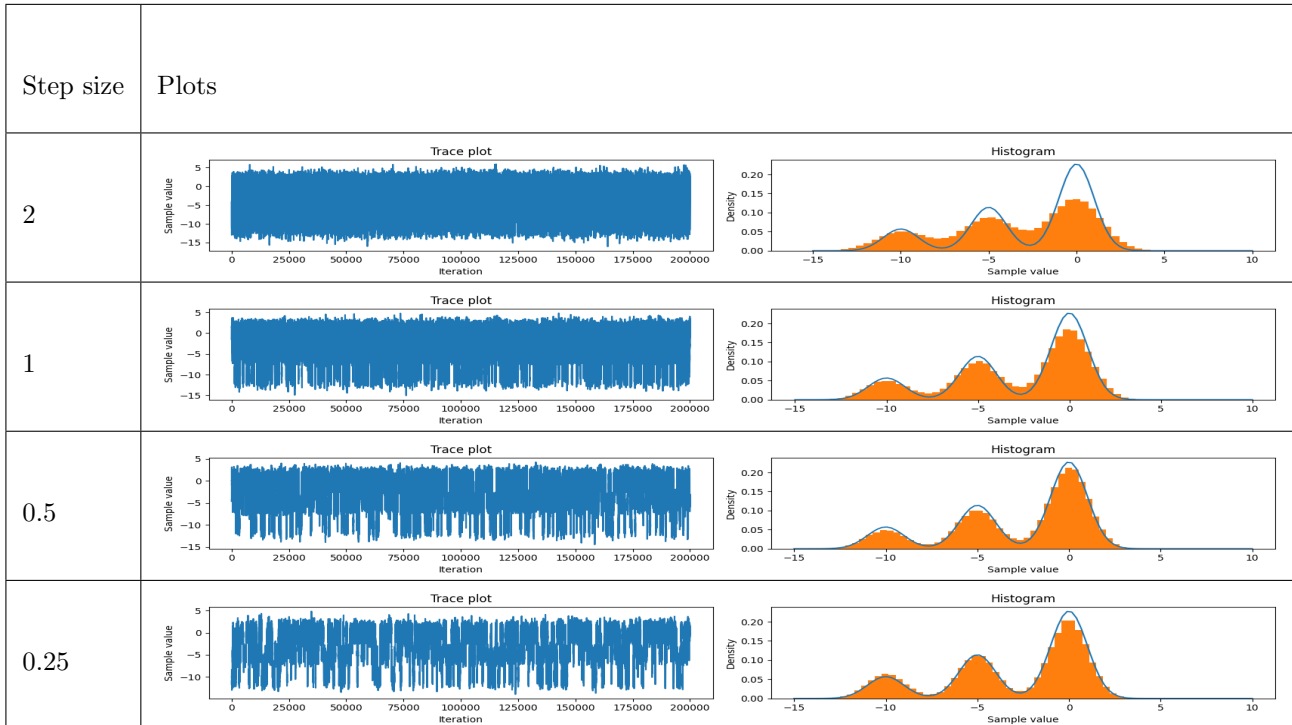
3.2 Trajectories for different step sizes

3.2.1 Short run





3.2.2 Long run



3.3 Observations

From the short run plots, we see that as step size increases, the histogram becomes more similar across different runs. This may suggest that higher step size leads to faster convergence to the asymptotic sampling distribution.

From the long run plots, we see that as step size decreases, the histogram becomes more like the target distribution, which suggests the asymptotic sampling distribution converges to the target distribution as the step size decreases.

4 Github

<https://github.com/justinlau20/Sampling>