# Notes on Sampling: Basics & MCMC

Shoichiro Yamanishi

May 29, 2020

## 1 Overview

This is a personal expository material on sampling mainly for my future self to quickly refresh the topics. It also contains explanations to the topics that are unclear to me in the books and articles. The following topics are covered.

- Basic sampling : from uniform distribution to a particular distribution.
- Rejection sampling
- Importance sampling
- Uni/Bivariate Gaussian Distribution : Box-Muller algorithm
- Univariate Gaussian Distribution with rejectio sampling : Ziggurat algorithm
- Multivariate full-covariance Gaussian distribution
- MCMC (Metropolis-Hastings)
- Gibbs Sampling
- MCMC with Hamiltonian dynamics

## 2 Why Sampling?

Sampling is an effective means to numerically calculate expectations of random variables over a given distributoin.

$$\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[f(\boldsymbol{x})\right] = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{L}\sum_{i=1}^{L} f(\boldsymbol{x}_i) \tag{1}$$

where $\boldsymbol{x}_i$ are i.i.d. drawn from the given $p(\boldsymbol{x})$.

## 2.1 Theoretical Back-up

Barber [3] briefly touches on it in Chap 27. We denote the RHS of equation 1 by $\boldsymbol{\mu}_f$. We take the limit of $\boldsymbol{\mu}_f$ as in $L \to \infty$, then:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\mu}_f\right] &= \mathbb{E}\left[\frac{1}{L}\sum_{i=1}^{L} f(\boldsymbol{x}_i)\right] \\ &= \frac{1}{L}\mathbb{E}\left[\sum_{i=1}^{L} f(\boldsymbol{x}_i)\right] \\ &= \frac{1}{L}\mathbb{E}\left[Lf(\boldsymbol{x})\right] \text{ because } \boldsymbol{x}_i \text{ are i.i.d.} \\ &= \mathbb{E}\left[f(\boldsymbol{x})\right]\end{aligned} \tag{2}$$

and the variance of each sampled $f(\boldsymbol{x}_i)$ is

$$\begin{aligned}\mathbb{E}\left[\frac{1}{L}\sum_{i=1}^{L}\left(f(\boldsymbol{x}_i) - \boldsymbol{\mu}_f\right)^2\right] &= \left(\frac{L-1}{L}\right)\left(\mathbb{E}\left[f(\boldsymbol{x}_i)^2\right] - \mathbb{E}\left[f(\boldsymbol{x}_i)\right]^2\right) \\ &= \left(\frac{L-1}{L}\right)\mathbb{E}\left[\left(f(\boldsymbol{x}_i) - \mathbb{E}\left[f(\boldsymbol{x}_i)\right]\right)^2\right]\end{aligned} \tag{3}$$

so the variance of sampled $f(\boldsymbol{x}_i)$ approaches the variane of $f(\boldsymbol{x})$ according to $p(\boldsymbol{x})$ as $L \to \infty$. Equation 3 comes from a similar argument for PRML [5] (1.58), for which I have arranged a separate document titled *PRML (1.58) : Why* $\mathbb{E}[\sigma^2_{sample}] = \frac{N-1}{N}\sigma^2$ *?* for deduction.

**NOTE** The equation (27.1.4) in Barber [3] quoted following is wrong.

$$\langle \hat{f}^2 \rangle - \langle \hat{f} \rangle^2 = \frac{1}{L}\left(\langle f^2(x) \rangle_{p(x)} - \langle f(x) \rangle^2_{p(x)}\right) \tag{4}$$

According to this equation, the variance vanishes as $L \to \infty$, which is clearly wrong. The deductions around (1.54) - (1.58) in PRML[5] gives the correct argument for a similar situation. PRML gives an argument for $\boldsymbol{x}$ but it can be easily applicable to $f(\boldsymbol{x})$.

# 3 Basic Univariate Sampling

Section 11.1. of PRML[5] gives a good overview, but I have made further annotation as it was still not clear to me. We assume a mechanism to draw a sample $x_i$ from a uniform distribution is available.

$$p(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

We want to draw samples $y_i$ from a probability density distribution $p(y)$, which is differentiable everywhere. We want to find a function $x \mapsto y,\ y = h(x)$

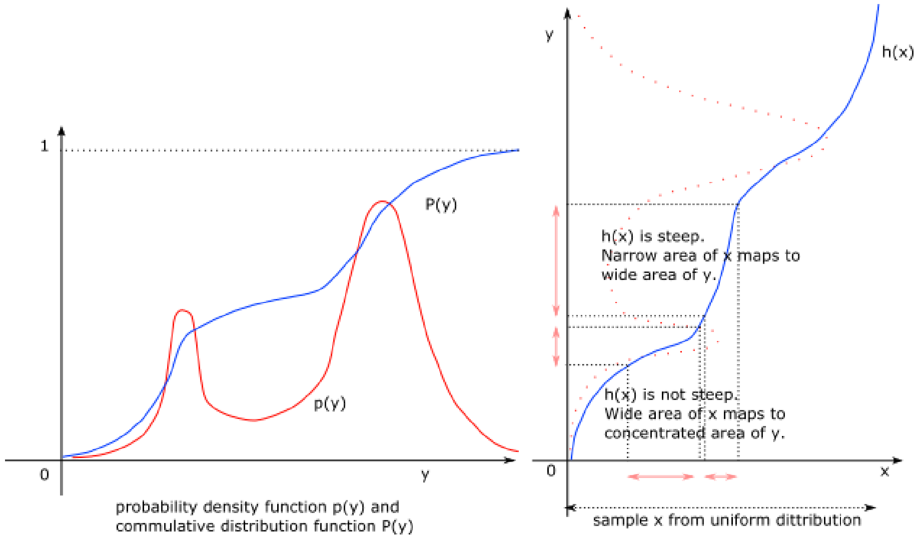

Figure 1: Graphical Interpretation of h(x)

so that we can draw $x$ and then convert it to $y$ that follows the distribuion $p(y)$. We show $y = h(x)$ is an inverse of the cummulative distribution function of $P(y)$.

To see this, for $0 \leq x \leq 1$,

$$p(x) = p(y)\left|\frac{dy}{dx}\right| = 1 \tag{6}$$

so

$$p(x)dx = p(y)dy \tag{7}$$

The cummulative distribution function $P(x)$ is

$$P(x) = \int_{-\infty}^{x} p(x)dx = x, \quad x \leq 1 \tag{8}$$

and the cummulative distribution function $P(y)$ is

$$\begin{aligned} P(y) = \int_{-\infty}^{y} p(y)dy &= \int_{-\infty}^{h^{-1}(y)} p(x)dx \quad \text{(from equation 8)} \\ &= \int_{-\infty}^{h^{-1}(y)} dx = h^{-1}(y) \end{aligned} \tag{9}$$

so $h(x) = P^{-1}(y)$. The following charts give you some intuition into this conversion. This is basically annotated and clearer version of Figure 11.2 in PRML [5].

## 4 Rejection Sampling

Please see 11.1.2 of PRML [5]. Following is a quick refresher. Rejecton Shsampling is applicable for the following case.

- $p(\boldsymbol{z})$ is of low dimension.
- evaluation of an unnormalized $\hat{p}(\boldsymbol{z})$ tractable.
- easy to find a proposal distribution $q(\boldsymbol{z})$ not far from $p(\boldsymbol{z})$, and for all $\boldsymbol{z}$ in the domain there is $K$ such that $Kq(\boldsymbol{z}) \geq \hat{p}(\boldsymbol{z})$.

And the following are the steps.

- Draw $\boldsymbol{z}^*$ from $q(\boldsymbol{z})$.
- Draw $u^*$ from the uniform distribution in $[0, Kq(\boldsymbol{z}^*)]$.
- Accept $\boldsymbol{z}^*$ if $u^* \leq p(\boldsymbol{z}^*)$. Discard it otherwise.

See figure 11.4 in pp 529 of PRML for the reason why it works. This is not suitable if the area for rejection is too big.

## 5 Importance Sampling

Please see 11.1.4 of PRML [5] for more info as it is well written. This does not draw samples directly but approximate expectations using a proposal distribution. you don't need actual samples but only the approximation to some expectations. Imporance sampling is applicable for the following case.

- $p(\boldsymbol{z})$ is of low dimension.
- evaluation of an unnormalized $\hat{p}(\boldsymbol{z})$ tractable.
- you don't need samples but only approxmiation to some expectations.

The following is the basic idea.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] &= \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \\
&= \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})d\boldsymbol{z} \qquad q(\boldsymbol{z}) \text{ is a proposal distribution.} \\
&\approx \frac{1}{L}\sum_{i=1}^{L}\left(f(\boldsymbol{z}_i)\frac{p(\boldsymbol{z}_i)}{q(\boldsymbol{z}_i)}\right) \qquad \boldsymbol{z}_i \sim q(\boldsymbol{z})
\end{aligned} \tag{10}$$

And we can use unnormalized version as follows.

$$\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] \approx \frac{Z_q}{Z_p}\frac{1}{L}\sum_{i=1}^{L}\left(f(\boldsymbol{z}_i)\frac{\hat{p}(\boldsymbol{z}_i)}{\hat{q}(\boldsymbol{z}_i)}\right) \tag{11}$$

where $p(\boldsymbol{z}) = \frac{1}{Z_p}\hat{p}(\boldsymbol{z})$, and $q(\boldsymbol{z}) = \frac{1}{Z_q}\hat{q}(\boldsymbol{z})$, and $\frac{Z_q}{Z_p}$ can be approximated as follows.

$$\begin{aligned} Z_p &= \int \hat{p}(\boldsymbol{z}) d\boldsymbol{z} \\ &= \int \frac{\hat{p}(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z} \\ &= \int Z_q \frac{\hat{p}(\boldsymbol{z})}{\hat{q}(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z} \\ &= Z_q \int \frac{\hat{p}(\boldsymbol{z})}{\hat{q}(\boldsymbol{z})} q(\boldsymbol{z}) d\boldsymbol{z} \\ &\approx Z_q \sum_{i=1}^{L} \left( \frac{\hat{p}(\boldsymbol{z}_i)}{\hat{q}(\boldsymbol{z}_i)} \right) \qquad \boldsymbol{z}_i \sim q(\boldsymbol{z}) \end{aligned} \tag{12}$$

# 6 Uni/Bivariate Gaussian Distribution : Box-Muller Algorithm

We can't simply draw a sample from $\mathcal{N}(0,1)$ because the cummulative probability function of the normal distribution, expressed by $P(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$, $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\text{dt}$ is not tractable, and hence inverse $x = P^{-1}(y)$ is difficult to compute.

Box-Muller algorithm [6] draws a pair of samples $(x_1, x_2)$ using some clever tricks. There are two flavors of Box-Muller Algorithm: $\chi^2$-form and the polar form. 11.1.1 of PRML [5] explains the polar form. Here we describe the $\chi^2$-form, which is simpler and more elegant.

First, we draw $z_1$ and $z_2$ from a uniform distribution in $[0,1]$, and

$$\begin{aligned} x_1 &= \sqrt{-2\ln z_1}\cos(2\pi z_2) \\ x_2 &= \sqrt{-2\ln z_1}\sin(2\pi z_2) \end{aligned} \tag{13}$$

Then we obtain samples $x_1$ and $x_2$ drawn from $\mathcal{N}(0,1)$.

Here's the derivation. Assume $x_1$ and $x_2$ are drawn from $\mathcal{N}(0,1)$, and let

$$R^2 = x_1^2 + x_2^2 = -2\ln z_1.$$

Then $R^2$ follows the $\chi^2$-distribution of degree 2. The PDF of $\chi^2$-distribution is

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2},$$

and when the degree $k = 2$, it is simply an exponential distributoin $\frac{1}{2}e^{-x/2}$, so $p(R^2) = \frac{1}{2}e^{-R^2/2}$, which suggests R follows a normal distribution. To be precise,

$$z_1 = \exp\left(\frac{x_1^2 + x_2^2}{2}\right) \tag{14}$$

and

$$\frac{x_2}{x_1} = \tan(2\pi z_2) \qquad z_2 = \frac{1}{2\pi}\tan^{-1}(\frac{x_2}{x_1}) \tag{15}$$

Taking the partial derivatives of them,

$$\begin{aligned}
\frac{\partial z_1}{\partial x_1} &= -x_1 \exp\left(\frac{x_1^2 + x_2^2}{2}\right) \\
\frac{\partial z_1}{\partial x_2} &= -x_2 \exp\left(\frac{x_1^2 + x_2^2}{2}\right)
\end{aligned} \tag{16}$$

and ($\frac{d}{dx}\tan^{-1}(x) = \frac{1}{x^2+1}$),

$$\begin{aligned}
\frac{\partial z_2}{\partial x_1} &= \frac{1}{2\pi}\frac{-x_2/x_1^2}{(x_2/x_1)^2+1} = -\frac{1}{2\pi}\frac{x_2}{(x_1^2+x_2^2)} \\
\frac{\partial z_2}{\partial x_2} &= \frac{1}{2\pi}\frac{x_1^{-1}}{(x_2/x_1)^2+1} = \frac{1}{2\pi}\frac{x_1}{(x_1^2+x_2^2)}
\end{aligned} \tag{17}$$

and from the conversion formula of the PDF,

$$\begin{aligned}
p(x_1, x_2) &= p(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(x_1, x_2)}\right| \\
&= p(z_1)p(z_2)\left|\frac{\partial z_1}{\partial x_1}\frac{\partial z_2}{\partial x_2} - \frac{\partial z_2}{\partial x_1}\frac{\partial z_1}{\partial x_2}\right| \\
&= \left|-\frac{1}{2\pi}\frac{x_1^2}{x_1^2+x_2^2}\exp\left(-\frac{x_1^2+x_2^2}{2}\right) - \frac{1}{2\pi}\frac{x_2^2}{x_1^2+x_2^2}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)\right| \\
&= \left|-\frac{1}{2\pi}\frac{x_1^2+x_2^2}{x_1^2+x_2^2}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)\right| \\
&= \left|-\frac{1}{2\pi}\exp\left(-\frac{x_1^2}{2}\right)\exp\left(-\frac{x_2^2}{2}\right)\right| \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_1^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x_2^2}{2}\right) \\
&= \mathcal{N}(x_1; 0, 1)\mathcal{N}(x_2; 0, 1)
\end{aligned} \tag{18}$$

# 7 Univariate Gaussian Distribution : Ziggurat Algorithm

This is a combination of table look-up and rejection sampling for speed. This framework is applicable to any monotone-decreasing or increasing PDF. Here we apply the positive half of the unit normal distribution. $\mathcal{N}(x; 0, 1), x \geq 0$.

Beforehand, the PDF is split into strips of equal area horizontally. See Figure 2. Each Strip consists of the red rectangle region and the yellow region to the right, except for the top strip ($R9$ in the example), which consists of a yellow region only, and the bottom strip $R0$, which is not rectanglar but horizontall unbounded. For $R0$, the gray rectangle to the right has the same height as $R0$ and the same area as the yellow unbounded tail.

Each sampling of $x_i$ goes through the following steps.

- draw a strip $RS_i$ from a discrete uniform distribution.
- draw a horizontal extent $x_i$ from a uniform distribution $[0, x_{S_i}]$.
- if $x_i$ falls into the red region, then accept $x_i$.
- otherwise, if $RS_i$ is not the bottom strip, then draw a horizontal extent $y_i$ from $[y_{S_i}, y_{S_i+1}]$ and if $p(x_i) \geq y_i$, accept $x_i$. Reject it otherwise.
- if $RS_i$ is the bottom strip and if $x_i$ is in the gray region $[x_1, x_0]$, then we ignore $x_i$ and go through the following extra steps.

The extra steps for the bottom unbounded tail are as follows. Here I follow the origina article [8], which is a declassified defense document, as its explanation is the clerest for me. As the horizontal extent is unbounded, we can't draw $x_i$ from a uniform distribution. We reverse the role of $x$-axis and $p(x)$-axis as in Figure 3. In this figure, $a$ corresponds to $x_1$ in Figure 2. And we introduce a new variable $U_1$ for the horizontal axis. And we scale the horizontal axis by $\frac{1}{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{a}{2}\right\}}$ so that the range of $U_1$ becomes $[0, 1]$. (The original article [8] is made that way, so I follow it.) Then the curve is expressed by $x = \sqrt{a^2 - 2\ln(U_1)}$.

- draw a sample $U_1^*$ from a uniform distribution from $[0, 1]$.
- $a^* = \sqrt{a^2 - 2\ln(U_1^*)}$.
- apply rejection sampling. Accept $a^*$ with the probability of $a/a^*$. For that draw another sample $U_2$ from a uniform distribution from $[0, 1]$, and accept $a^*$ if $U_2 \leq a/a^*$.

# 8 Multivariate Full-Covariance Gaussian Distribution

To draw a sample $\boldsymbol{x} \in R^N, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$, then first we take a cholesky decomposition $\Sigma = LL^T$, then we draw a sample $\boldsymbol{z} \in R^N$, such that each element $z_i$ is drawn from unit normal distribution $\mathcal{N}(z_i; 0, 1)$. Then $\boldsymbol{x} = \boldsymbol{\mu} + L\boldsymbol{z}$.

# 9 Basic MCMC

The best study material I find is the secton 3 of *An Introduction to MCMC for Machine Learning*[1] available free from Springer. The interpretation of the

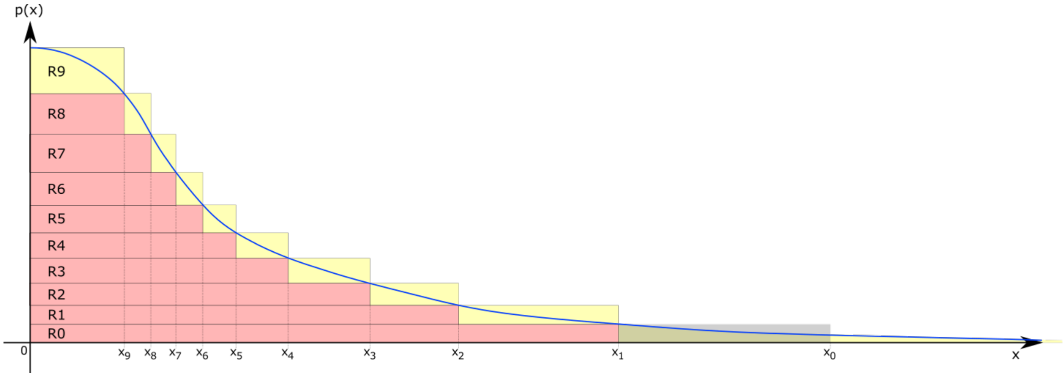

Figure 2: Horizontal Layers of Equal Area

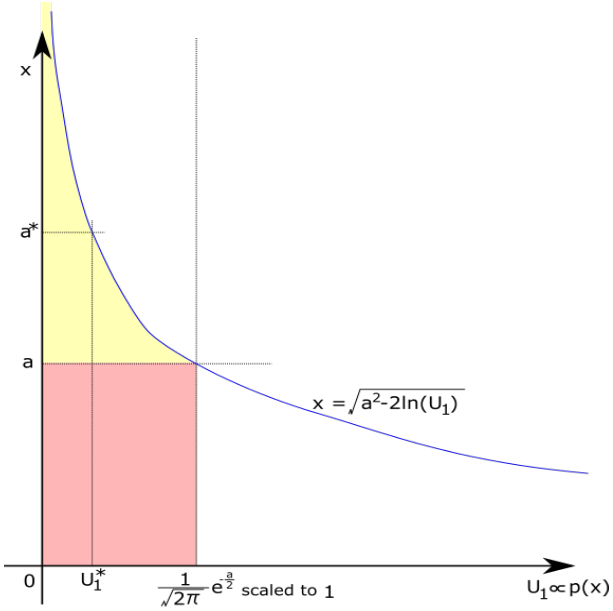
x
a*
a
$x = \sqrt{a^2 - 2\ln(U_1)}$
0
$U_1^*$
$\frac{1}{\sqrt{2\pi}} e^{-\frac{a}{2}}$ scaled to 1
$U_1 \propto p(x)$

Figure 3: Layer R0 Arranged Upright and Scaled Horizontally such that $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) = 1$.

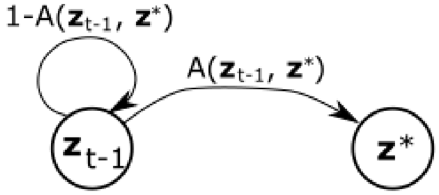
1-A($\mathbf{z}_{t-1}$, $\mathbf{z}^*$)
A($\mathbf{z}_{t-1}$, $\mathbf{z}^*$)
$\mathbf{z}_{t-1}$
$\mathbf{z}^*$

Figure 4: Transition from $\boldsymbol{z}_{t-1}$ to $\boldsymbol{z}_t$

transition probability was not clear to me, so I have added extra subsection below.

MCMC, or Markov chain Monte Carlo, is suitable for the following conditions.

- Intractable to draw samples directly from $p(\boldsymbol{z})$.
- Intractable to draw samples even with a proposal distribution. (Dominant rejection space due to high demensionality, mismatch between the proposal distribution and the actual distribution etc.)
- Tractable to evaluate unnormalized distribution $\hat{p}(\boldsymbol{z})$

The basic idea is to draw sequence of samples $\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots$ from an ergodic Markov chain with transition probability $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$, and expect it to become $p(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = p(\boldsymbol{z}_t)$ for any $\boldsymbol{z}_0$ as $t \to \infty$. For ergodicity, please see the course notes [10] available from Dartmouth University, which gives a very nice presentation of ergodicity for a dicrete case using power of a matrix.

Basically it states the following. Let $P$ be the transtion matrix for a regular chain, then $P^n$ approaches a limiting matrix $W$ with all rows the same vector $\boldsymbol{w}$, which is a positive probability vector.

## 9.1 Steps

Each sample $\boldsymbol{z}_t$ is sampled by the following steps. $\boldsymbol{z}_0$ is chosen arbitrarily.

- Draw $\boldsymbol{z}^*$ from a proposal distribution $q(\boldsymbol{z}^*|\boldsymbol{z}_{t-1})$.
- Accept the transition with the probability given by a function $A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*\right)$.
- If accepted, then $\boldsymbol{z}_t := \boldsymbol{z}^*$, if not, then $\boldsymbol{z}_t := \boldsymbol{z}_{t-1}$.

This is illustrated in Figure 4.

## 9.2 Construction and Interpretation of the Transition Probability $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$

The transition probability is expressed as follows.

$$\begin{aligned} T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) = & \, q\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) \\ & + \delta_{\boldsymbol{z}_{t-1}}\left(\boldsymbol{z}_t\right) \int q\left(\boldsymbol{z}^*|\boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*\right)\right) d\boldsymbol{z}^* \end{aligned} \tag{19}$$

This is difficult to grasp but it can be explained as follows. First, we show the validity of $\int T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$.

$$\begin{aligned} \int T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z}_t = & \int q\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) d\boldsymbol{z}_t \\ & + \int \delta_{\boldsymbol{z}_{t-1}}\left(\boldsymbol{z}_t\right)\left(\int q\left(\boldsymbol{z}^*|\boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*\right)\right) d\boldsymbol{z}^*\right) d\boldsymbol{z}_t \\ = & \int q\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) d\boldsymbol{z}_t \\ & + \int q\left(\boldsymbol{z}^*|\boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*\right)\right) d\boldsymbol{z}^* \\ = & \int q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}\right) d\boldsymbol{z} + \int q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}\right)\right) d\boldsymbol{z} \\ = & \int q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z} = 1 \end{aligned} \tag{20}$$

Also, it is easy to see that $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) \geq 0$ for any $\boldsymbol{z}_t$ in the domain. Therefore $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$ is a valid probability density function.

Next, we interpret $\int T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$, for that we take an integral of $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$ for an infinitesimal region $\Delta\boldsymbol{z}$. If $\boldsymbol{z}_{t-1} \notin \Delta\boldsymbol{z}$, then

$$\int_{\Delta\boldsymbol{z}} T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z}_t = \int_{\Delta\boldsymbol{z}} q\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) d\boldsymbol{z}_t \tag{21}$$

This means the probability of transition to the new region specified by $\Delta\boldsymbol{z}$ which is represented by $\boldsymbol{z}_t$. On the other hand, if $\boldsymbol{z}_{t-1} \in \Delta\boldsymbol{z}$, then

$$\begin{aligned}
\int_{\Delta\boldsymbol{z}} T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z}_t &= \int_{\Delta\boldsymbol{z}} q\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) d\boldsymbol{z}_t \\
&\quad + \int q\left(\boldsymbol{z}^*|\boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*\right)\right) d\boldsymbol{z}^* \\
&= \int_{\Delta\boldsymbol{z}} q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}\right) d\boldsymbol{z} \\
&\quad + \int q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z} - \int q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}\right) d\boldsymbol{z} \\
&= \int q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z} - \int_{\backslash\Delta\boldsymbol{z}} q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}\right) d\boldsymbol{z} \\
&= \int_{\Delta\boldsymbol{z}} q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right) d\boldsymbol{z} + \int_{\backslash\Delta\boldsymbol{z}} q\left(\boldsymbol{z}|\boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}\right)\right) d\boldsymbol{z}
\end{aligned} \tag{22}$$

where $\backslash\Delta\boldsymbol{z}$ means the domain of $\boldsymbol{z}$ excluding $\Delta\boldsymbol{z}$. The first term of equation 22 means the probability of staying in the region represented by $\boldsymbol{z}_{t-1}$. The second term means the probability of trying to transit to another region but rejected, and eventually staying in the current region represented by $\boldsymbol{z}_{t-1}$.

## 9.3 How to Find an Ergodic $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$?

We want to find $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right)$ such that $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) = p(\boldsymbol{z}_t)$ as $t \to \infty$ for any $\boldsymbol{z}_0$. This is achieved if $T\left(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}\right) = p(\boldsymbol{z}_t)$ is ergodic, and a sufficient condition for the ergodicity is called *detailed balance*, for which equation 23 must hold for any $\boldsymbol{z}_a$, and $\boldsymbol{z}_b$.

$$p(\boldsymbol{z}_b)T(\boldsymbol{z}_a|\boldsymbol{z}_b) = p(\boldsymbol{z}_a)T(\boldsymbol{z}_b|\boldsymbol{z}_a) \tag{23}$$

It has the following invariant property.

$$\begin{aligned}
\int p(\boldsymbol{z}_a)T(\boldsymbol{z}_b|\boldsymbol{z}_a)d\boldsymbol{z}_a &= \int p(\boldsymbol{z}_b)T(\boldsymbol{z}_a|\boldsymbol{z}_b)d\boldsymbol{z}_a \\
&= p(\boldsymbol{z}_b)\int T(\boldsymbol{z}_a|\boldsymbol{z}_b)d\boldsymbol{z}_a \\
&= p(\boldsymbol{z}_b)
\end{aligned} \tag{24}$$

Hatings[7] found in 1970 one such $A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right)$, which satisfies detailed balance.

$$A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) = \min\left\{1, \frac{p(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)}{p(\boldsymbol{z}_{t-1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})}\right\} \tag{25}$$

Please note that we can use unnormalized probability density functions as follows.

$$A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) = \min\left\{1, \frac{\tilde{p}(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)}{\tilde{p}(\boldsymbol{z}_{t-1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})}\right\} \tag{26}$$

Following shows that the acceptance function in equation 25 meets the detailed balance. If $\boldsymbol{z}_{t-1} = \boldsymbol{z}_t$, then it trivially holds. If $\boldsymbol{z}_{t-1} \neq \boldsymbol{z}_t$, then,

$$\begin{aligned} p(\boldsymbol{z}_{t-1})T(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) &= p(\boldsymbol{z}_{t-1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})A(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t) \quad \text{(from equation 19)} \\ &= \min\left\{\, p(\boldsymbol{z}_{t-1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}),\ \ p(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) \,\right\} \\ &= \min\left\{\, p(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t),\ \ p(\boldsymbol{z}_{t-1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) \,\right\} \\ &= p(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)A(\boldsymbol{z}_t, \boldsymbol{z}_{t-1}) \\ &= p(\boldsymbol{z}_t)T(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) \end{aligned} \tag{27}$$

This is called *Metropolis-Hastings* algorithm. Usually a multivariabe normal distribution, whose mean is $\boldsymbol{z}_{t-1}$, is used as a proposal distribution $q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})$. The drawback of using this proposal distribution is that $\boldsymbol{z}_0, \boldsymbol{z}_1, \cdots$ becomes a random walk, and the distance is in the order of $\sqrt{t}$.

# 10 Gibbs Sampling

This is a MCMC with a particular proposal distribution, in which one element $z^i$ of $\boldsymbol{z} = (z^1, z^2, \cdots, z^N)$ is conditioned on the rest $\boldsymbol{z}^{\setminus i}$, such that $q(z_t^i) = p(z_t^i|\boldsymbol{z}_{t-1}^{\setminus i})$ In this case, the transition is always accepted as the following acceptance probability function implies.

$$\begin{aligned} A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) &= \min\left\{1, \frac{p(\boldsymbol{z}_t)q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)}{p(\boldsymbol{z}_{t-1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1})}\right\} \\ &= \min\left\{1, \frac{p(\boldsymbol{z}_t)p(z_{t-1}^i|\boldsymbol{z}_t^{\setminus i})}{p(\boldsymbol{z}_{t-1})p(z_t^i|\boldsymbol{z}_{t-1}^{\setminus i})}\right\} \\ &= \min\left\{1, \frac{p(z_t^i, \boldsymbol{z}_t^{\setminus i})p(z_{t-1}^i|\boldsymbol{z}_t^{\setminus i})}{p(z_{t-1}^i, \boldsymbol{z}_{t-1}^{\setminus i})p(z_t^i|\boldsymbol{z}_{t-1}^{\setminus i})}\right\} \\ &= \min\left\{1, \frac{p(z_t^i|\boldsymbol{z}_t^{\setminus i})p(\boldsymbol{z}_t^{\setminus i})p(z_{t-1}^i|\boldsymbol{z}_t^{\setminus i})}{p(z_{t-1}^i|\boldsymbol{z}_{t-1}^{\setminus i})p(\boldsymbol{z}_{t-1}^{\setminus i})p(z_t^i|\boldsymbol{z}_{t-1}^{\setminus i})}\right\} \\ &= \min\left\{1, \frac{p(z_t^i|\boldsymbol{z}_{t-1}^{\setminus i})p(\boldsymbol{z}_{t-1}^{\setminus i})p(z_{t-1}^i|\boldsymbol{z}_{t-1}^{\setminus i})}{p(z_{t-1}^i|\boldsymbol{z}_{t-1}^{\setminus i})p(\boldsymbol{z}_{t-1}^{\setminus i})p(z_t^i|\boldsymbol{z}_{t-1}^{\setminus i})}\right\} \qquad \left(\boldsymbol{z}_t^{\setminus i} = \boldsymbol{z}_{t-1}^{\setminus i}\right) \\ &= \min\left\{1, 1\right\} = 1 \end{aligned} \tag{28}$$

The index $i$ for the element $z^i$ is cyclically iterated or randomly chosen.

# 11 MCMC with Hamiltonian dynamics

Here the emphasis is on the regorous formation of MCMC using Hamiltonian proposal distribution, as almost all the books and articles focus on the Hamiltonian dynamics and numerical integration. 11.5 of PRML [5], which gives a good overview. For deeper understanding, read the chapter on handbook by Neal [9], which is freely available online. It is also useful to read section 3 “The foundations of Hamiltonian Monte Carlo” and 4 “Efficient Hamiltonion Monte Carlo” of the article by Betancourt [4]. All the articles and books I have read put emphasis on Hamiltonian dynamics and its numerical integration, but more rigorous formation of MCMC using Hamiltonian dynamics is often neglected. Especially, formation of a proposal function, an acceptance function, transition probability distribution and ergodicity are not well discussed. This section focus on those topics, rather than the Hamiltonion dynamics and the numerical simulation, as I’m already well versed with physics simulation.

This is an attempt to overcome the limited travel of random walk by drawing a sample from an ordinary proposal distribution $q(\boldsymbol{z})$. The basic idea is to introduce a new variable (kind of latent variable) $\boldsymbol{r}$ and to form a joint distribution $p(\boldsymbol{z}, \boldsymbol{r}) = p_{model}(\boldsymbol{z})q(\boldsymbol{r})$. $q(\boldsymbol{r})$ is a multivariate Gaussian distribution, which is also considered a negative of kinetic energy in Hamiltonian dynamics. At each step, we have a previously drawn sample $(\boldsymbol{z}_i, \boldsymbol{r}_i)$. First we draw a sample $\boldsymbol{r}_i^*$ from $q(\boldsymbol{r})$, and then we deterministically find the next sample $(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1})$ from $(\boldsymbol{z}_i, \boldsymbol{r}_i^*)$ such that $p(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1}) = p(\boldsymbol{z}_i, \boldsymbol{r}_i^*)$, hoping $\boldsymbol{z}_{i+1}$ has travelled further from $\boldsymbol{z}_i$ using Hamiltonian dynamics. The first part is considered Gibbs sampling, and hence the detailed balance holds. The second deterministic part is carefully designed to hold the detailed balance. However, due to the determinism, neither irreducibility or aperiodicity holds. Hence the ergodicity of the whole process must be proven separately.

## 11.1 Augmentation of the model distribution $p_{model}(\boldsymbol{z})$ with momentum $\boldsymbol{r}$

Given a distribution $p_{model}(\boldsymbol{z})$, $\boldsymbol{z} \in \mathcal{R}^d$, from which you want to draw samples, we augment it with an additional variable $\boldsymbol{r}$, $\boldsymbol{r} \in \mathcal{R}^d$, which works as a momemtum. The joint distribution of $\boldsymbol{z}$ and $\boldsymbol{r}$ is defined as follows.

$$p(\boldsymbol{z}, \boldsymbol{r}) = p_{model}(\boldsymbol{z})\, p_{\boldsymbol{r}}(\boldsymbol{r}) \tag{29}$$

where $p_{\boldsymbol{r}}(\boldsymbol{r})$ is a Gaussian distribution with zero mean as follows.

$$p_{\boldsymbol{r}}(\boldsymbol{r}) = \frac{1}{Z_{\boldsymbol{r}}} \exp\left(-\frac{1}{2}\boldsymbol{r}^T M^{-1} \boldsymbol{r}\right) \tag{30}$$

where $M$ is a diagonal mass matrix. The original distribution $p_{model}(\boldsymbol{z})$ corresponds to a potential energy.

$$U(\boldsymbol{z}) = -\ln\left(p_{model}(\boldsymbol{z})\right) \tag{31}$$

And the kinetic energy is defined as follows.

$$K(\boldsymbol{r}) \propto -\ln\left(p_{\boldsymbol{r}}(\boldsymbol{r})\right) = -\frac{1}{2}\boldsymbol{r}^T M^{-1} \boldsymbol{r} \tag{32}$$

In the simplest form $K(\boldsymbol{r}) = \frac{1}{2}\sum_{i=1}^{d} r_i^2$, and $\boldsymbol{r}$ works as velocity. Finally the Hamiltonian is defined as follows.

$$H(\boldsymbol{z}, \boldsymbol{r}) = U(\boldsymbol{z}) + K(\boldsymbol{r}) \tag{33}$$

The motion dynamics of $\boldsymbol{z}$ and $\boldsymbol{r}$ over continuous time $t$ is as follows.

$$\frac{dz_i}{dt} = \frac{\partial H}{\partial r_i} = \left[M^{-1}\boldsymbol{r}\right]_i \tag{34}$$

$$\frac{dr_i}{dt} = -\frac{\partial H}{\partial z_i} = -\frac{\partial U}{\partial z_i} \tag{35}$$

From the equations 34 & 35, we can determine $(\boldsymbol{z}(t_0 + \Delta t), \boldsymbol{r}(t_0 + \Delta t))$ from $(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))$ and $\Delta t$, considering $\boldsymbol{z}$ and $\boldsymbol{r}$ functions of $t$. We use three laws regarding the Hamiltonian motion dynamics.

### 11.1.1 Reversibility of Motion

This means we can find $(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))$ from $(\boldsymbol{z}(t_0 + \Delta t), \boldsymbol{r}(t_0 + \Delta t))$ and $-\Delta t$.

### 11.1.2 Conservation of Energy

This means $H\left(\boldsymbol{z}(t_0 + \Delta t), \boldsymbol{r}(t_0 + \Delta t)\right) = H\left(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)\right)$ This is proven as follows.

$$\begin{aligned} \frac{dH}{dt} &= \sum_{i=1}^{d} \left[\frac{dz_i}{dt}\frac{\partial H}{\partial z_i} + \frac{dr_i}{dt}\frac{\partial H}{\partial r_i}\right] \\ &= \sum_{i=1}^{d} \left[\frac{\partial H}{\partial r_i}\frac{\partial H}{\partial z_i} - \frac{\partial H}{\partial z_i}\frac{\partial H}{\partial r_i}\right] = 0 \end{aligned} \tag{36}$$

### 11.1.3 Conservation of Volume

Let $R(t_0) \subset \mathcal{R}^{2d}$ be a closed bounded volume whose points represent $(\boldsymbol{z}, \boldsymbol{r})$, and $R(t_0 + \Delta t)$ be the same volume at time $t + \Delta t$. Then $||R(t_0)|| = ||R(t_0 + \Delta t)||$. This is comes from something called Liouville's Theorem for symplectic geometry. Most of the MCMC articles and books skip the proof. The cleanest and clearest explanation for the proof that I know is given in the physics book by Arnold[2], Chap. 15 *Hamilton's equations*, and Chap 16 *Liouville's theorem*. Following gives an outline.

Let the $2d$-dimensional space determined by $\boldsymbol{z}, \boldsymbol{r}$ denoted by *phrase space*. Let the group of transformations over time $\Delta t$ from $t_0$ denoted by *phase flow*.

$$g^{\Delta t} : (\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)) \mapsto (\boldsymbol{z}(t_0 + \Delta t), \boldsymbol{r}(t_0 + \Delta t))$$

[2] leaves the proof that $\{g^{\Delta t}\}$ forms a group, but basically we need to provide the following

- Closure for addition: Addition of two phase flows is another phase flow.

- Associativity: $(g^{\Delta t_a} + g^{\Delta t_b}) + g^{\Delta t_c} = g^{\Delta t_a} + (g^{\Delta t_b} + g^{\Delta t_c})$
- Identity element: $g^{\Delta t_0}$ is a flow such that
$$(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)) = (\boldsymbol{z}(t_0 + \Delta t), \boldsymbol{r}(t_0 + \Delta t))$$
- Inverse element: For each $g^{\Delta t_0}$ there is an inverse flow such that
$$-g^{\Delta t} : (\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)) \mapsto (\boldsymbol{z}(t_0 - \Delta t), \boldsymbol{r}(t_0 - \Delta t))$$

Let $(\dot{\boldsymbol{z}}, \dot{\boldsymbol{r}}) = \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{r})$ be the ODE of Hamiltonian dynamics such that

$$\boldsymbol{f}(\boldsymbol{z}, \boldsymbol{r}) = \left[\begin{array}{c} \frac{\partial H}{\partial \boldsymbol{r}} \\ \hline -\frac{\partial H}{\partial \boldsymbol{z}} \end{array}\right] \tag{37}$$

Then the phase flow is approximated as follows.

$$g^{\Delta t}(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)) = (\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)) + \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{r})|_{\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)} \Delta t + O(\Delta t^2) \tag{38}$$

Then we can express the volumes as folows.

$$R(t_0 + \Delta t) = g^{\Delta t_a}(R(t_0)) \tag{39}$$

Then the formula for changing variables in a multiple integral gives. ( $(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0)) \mapsto (\boldsymbol{z}(t_0 + \Delta t), \boldsymbol{r}(t_0 + \Delta t))$ can be considered a change in the variables.)

$$|R(t_0 + \Delta t)| = \int_{R(t_0)} \det \left( \frac{\partial g^{\Delta t}(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))}{\partial(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))} \right) \partial(\boldsymbol{z}(t_0) \boldsymbol{r}(t_0)) \tag{40}$$

From equation 38,

$$\frac{\partial g^{\Delta t}(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))}{\partial(\boldsymbol{z}(t_0) \boldsymbol{r}(t_0))} = I_{2d} + \frac{\partial \boldsymbol{f}}{\partial(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))} \Delta t + O(\Delta t^2) \tag{41}$$

Then

$$\begin{aligned} \det \left( \frac{\partial g^{\Delta t}(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))}{\partial(\boldsymbol{z}(t_0) \boldsymbol{r}(t_0))} \right) &= 1 + \operatorname{tr} \left( \frac{\partial \boldsymbol{f}}{\partial(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))} \right) \Delta t + O(\Delta t^2) \\ &= 1 + \boldsymbol{\nabla} \cdot \boldsymbol{f} \Delta t + O(\Delta t^2) \end{aligned} \tag{42}$$

And the divergence is zero as follows. From equation 37,

$$\boldsymbol{\nabla} \cdot \boldsymbol{f} = \frac{\partial}{\partial \boldsymbol{z}} \frac{\partial H}{\partial \boldsymbol{r}} - \frac{\partial}{\partial \boldsymbol{r}} \frac{\partial H}{\partial \boldsymbol{z}} = 0 \tag{43}$$

From equations 42 & 43,

$$\det \left( \frac{\partial g^{\Delta t}(\boldsymbol{z}(t_0), \boldsymbol{r}(t_0))}{\partial(\boldsymbol{z}(t_0) \boldsymbol{r}(t_0))} \right) = 1 + O(\Delta t^2) \tag{44}$$

From equations 40 and 44,

$$\begin{aligned} \frac{d|R(t_0+\Delta t)|}{d\Delta t} &= \frac{d}{d\Delta t}\int_{R(t_0)} \left(1+O(\Delta t^2)\right)\ \partial(\boldsymbol{z}(t_0)\boldsymbol{r}(t_0)) \\ &= \int_{R(t_0)} \frac{d}{d\Delta t}\left(1+O(\Delta t^2)\right)\ \partial(\boldsymbol{z}(t_0)\boldsymbol{r}(t_0)) \\ &= 0 \end{aligned} \tag{45}$$

Hence $|R(t_0+\Delta t)| = |R(t_0)|$, and the convervation of the volume holds.

The volume preservation and reversibility over time implies that the phase flow is a bijection (one-to-one mapping).

## 11.2 Hamiltonian Dynamics for a Step

This is a numerical integration of phase flow ODE over time $t$ defined in equations 37 & 38. It seems an Euler variant called the leap frog method is used as follows.

$$\begin{aligned} r_i(t+\epsilon/2) &= r_i(t) - (\epsilon/2)\frac{\partial U}{\partial z_i} z_i(t) \\ z_i(t+\epsilon) &= z_i(t) + \epsilon\frac{r_i(t+\epsilon/2)}{m_i} \\ r_i(t+\epsilon) &= r_i(t+\epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial z_i} z_i(t+\epsilon) \end{aligned} \tag{46}$$

It is an explicit method, but neither $U(\boldsymbol{z})$ or $K(\boldsymbol{r})$ is steep enough to cause stiffness.

One step $(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1}) = \phi(\boldsymbol{z}_i, \boldsymbol{r}_i)$ consists of $L$ numerical integration of time step $\epsilon$. There is one tweek to make the step symmetry. At the last step, we flip the sign of each $r_i$ so that the following also holds. $(\boldsymbol{z}_i, \boldsymbol{r}_i) = \phi(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1})$. Flipping of the sing of $r_i$ has the same effect as numerical integration over negative time $-\epsilon$ toward past.

## 11.3 Proposal Distribution

The hamitonian dynamics at one step is given as follows. $(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1}) = \boldsymbol{\phi}(\boldsymbol{z}_i, \boldsymbol{r}_i)$. This is a symmetric process as in $(\boldsymbol{z}_i, \boldsymbol{r}_i) = \boldsymbol{\phi}(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1})$. This is achieved by flipping the polarity of $\boldsymbol{r}$ after the numerical integration finished as stated in the previous section..

The proposal distribution is a deterministic process whose density function is given as follows.

$$q(\boldsymbol{z}, \boldsymbol{r}|\boldsymbol{z}_i, \boldsymbol{r}_i) = \delta\left(\boldsymbol{z} - \boldsymbol{\phi}(\boldsymbol{z}_i)\right)\delta\left(\boldsymbol{r} - \boldsymbol{\phi}(\boldsymbol{r}_i)\right) \tag{47}$$

Please note that $q(\boldsymbol{z}, \boldsymbol{r}|\boldsymbol{z}_i, \boldsymbol{r}_i)$ is everywehere 0 except for $(\boldsymbol{\phi}(\boldsymbol{z}_i), \boldsymbol{\phi}(\boldsymbol{r}_i))$. This means the detailed balance of the transition probability $T(\boldsymbol{z}, \boldsymbol{r}|\boldsymbol{z}_i, \boldsymbol{r}_i)$ is not enouch to prove the ergodicity. This will be discussed later.

## 11.4 Acceptance Function

The acceptance function must be defined on volume, not on point. For that, we extend the hamiltonian dynamics function from the point to volume as follows. Let $\partial R \subseteq \mathcal{R}^{2d}$ be a closed bounded volume defined in $(\boldsymbol{z}, \boldsymbol{r})$ phase space. Let $\partial R_{i+1} = \boldsymbol{\phi}(\partial R_i)$ The symmetry still holds. $\partial R_i = \boldsymbol{\phi}(\partial R_{i+1})$

The acceptance probability is defined as follows, provided $|\partial R_a| = |\partial R_b|$.

$$A(\partial R_a, \partial R_b) = \begin{cases} p(\boldsymbol{z}_b, \boldsymbol{r}_b)/p(\boldsymbol{z}_a, \boldsymbol{r}_a) & \partial R_b = \boldsymbol{\phi}(\partial R_a) \\ \text{undefined} & \text{otherwise} \end{cases} \tag{48}$$

where $(\boldsymbol{z}_a, \boldsymbol{r}_a) \in \partial R_a$, and $(\boldsymbol{z}_b, \boldsymbol{r}_b) = \boldsymbol{\phi}(\boldsymbol{z}_a, \boldsymbol{r}_a) \in \partial R_b$, and we assume $p(\boldsymbol{z}, \boldsymbol{r})$ is constant in a minute volume $\partial R$.

The following derives equation 48. First, we form the Metropolis acceptance from its definition.

$$A(\partial R_a, \partial R_b) = \min\left\{1, \frac{\int_{\partial R_a}\int_{\partial R_b} p(\boldsymbol{z}_b, \boldsymbol{r}_b) q(\boldsymbol{z}_a, \boldsymbol{r}_a | \boldsymbol{z}_b, \boldsymbol{r}_b) \partial(\boldsymbol{z}_b \boldsymbol{r}_b) \partial(\boldsymbol{z}_a \boldsymbol{r}_a)}{\int_{\partial R_b}\int_{\partial R_a} p(\boldsymbol{z}_a, \boldsymbol{r}_a) q(\boldsymbol{z}_b, \boldsymbol{r}_b | \boldsymbol{z}_a, \boldsymbol{r}_a) \partial(\boldsymbol{z}_a \boldsymbol{r}_a) \partial(\boldsymbol{z}_b \boldsymbol{r}_b)}\right\} \tag{49}$$

The numerator of the second term in the min is expanded into:

$$\begin{aligned} \text{term1} &= \int_{\partial R_a}\int_{\partial R_b} p(\boldsymbol{z}_b, \boldsymbol{r}_b)\delta\left(\boldsymbol{z}_a - \boldsymbol{\phi}(\boldsymbol{z}_b)\right)\delta\left(\boldsymbol{r}_a - \boldsymbol{\phi}(\boldsymbol{r}_b)\right)\partial(\boldsymbol{z}_b \boldsymbol{r}_b)\partial(\boldsymbol{z}_a \boldsymbol{r}_a) \\ &= \int_{\partial R_b} p(\boldsymbol{z}_b, \boldsymbol{r}_b)\int_{\partial R_a} \delta\left(\boldsymbol{z}_a - \boldsymbol{\phi}(\boldsymbol{z}_b)\right)\delta\left(\boldsymbol{r}_a - \boldsymbol{\phi}(\boldsymbol{r}_b)\right)\partial(\boldsymbol{z}_a \boldsymbol{r}_a)\partial(\boldsymbol{z}_b \boldsymbol{r}_b) \\ &= \begin{cases} 0 & \partial R_b \cap \boldsymbol{\phi}(\partial R_a) = \emptyset \\ \int_{\partial R_b} p(\boldsymbol{z}_b, \boldsymbol{r}_b)\partial(\boldsymbol{z}_b \boldsymbol{r}_b) & \partial R_b = \boldsymbol{\phi}(\partial R_a) \end{cases} \\ &= \begin{cases} 0 & \partial R_b \cap \boldsymbol{\phi}(\partial R_a) = \emptyset \\ |\partial R_a|\, p(\boldsymbol{z}_b, \boldsymbol{r}_b) & \partial R_b = \boldsymbol{\phi}(\partial R_a) \end{cases} \end{aligned} \tag{50}$$

In the same way, the denominator of the second term in the min is expanded into:

$$\text{term2} = \begin{cases} 0 & \partial R_b \cap \boldsymbol{\phi}(\partial R_b) = \emptyset \\ |\partial R_b|\, p(\boldsymbol{z}_a, \boldsymbol{r}_a) & \partial R_a = \boldsymbol{\phi}(\partial R_b) \end{cases} \tag{51}$$

Because of the conservation of the volume and the reversibility of $\phi$,

$$\partial R_a = \boldsymbol{\phi}(\partial R_b) \Leftrightarrow \partial R_b = \boldsymbol{\phi}(\partial R_a)$$

Due to the determinism of $\phi$, the undefined case in equation 48 never occurs. Also, due to conservation of energy, the first case in equations 48 & 49 is theoretically 1, but it can be slightly different due to numerical error in the integration.

## 11.5 Transition Probability and Detailed Balance

The transition probability is defined in equation 19. We restate it as follows.

$$\begin{aligned} T\left(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}\right) &= q\left(\boldsymbol{z}_t | \boldsymbol{z}_{t-1}\right) A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}_t\right) \\ &\quad + \delta\left(\boldsymbol{z}_t - \boldsymbol{z}_{t-1}\right) \int q\left(\boldsymbol{z}^* | \boldsymbol{z}_{t-1}\right)\left(1 - A\left(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*\right)\right) d\boldsymbol{z}^* \end{aligned} \tag{52}$$

The first term indicates the probability in which the proposed $\boldsymbol{z}_t$ was accepted. The second term indicates the collective probability in which the proposed $\boldsymbol{z}^*$ was rejected, and it has to stay in $\boldsymbol{z}_{t-1}$.

However, in the case of Hamiltonian transition, there is only one deterministic transition and it is simplified as follows.

$$T\left(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1} | \boldsymbol{z}_i, \boldsymbol{r}_i\right) = \begin{cases} \min\left\{1, \frac{p(\boldsymbol{\phi}(\boldsymbol{z}_i), \boldsymbol{\phi}(\boldsymbol{r}_i))}{p(\boldsymbol{z}_i, \boldsymbol{r}_i)}\right\} & \boldsymbol{z}_{i+1} = \boldsymbol{\phi}(\boldsymbol{z}_i) \wedge \boldsymbol{r}_{i+1} = \boldsymbol{\phi}(\boldsymbol{r}_i) \\ 1 - \min\left\{1, \frac{p(\boldsymbol{\phi}(\boldsymbol{z}_i), \boldsymbol{\phi}(\boldsymbol{r}_i))}{p(\boldsymbol{z}_i, \boldsymbol{r}_i)}\right\} & \boldsymbol{z}_{i+1} = \boldsymbol{z}_i \wedge \boldsymbol{r}_{i+1} = \boldsymbol{r}_i \\ 0 & \text{otherwise} \end{cases} \tag{53}$$

The detailed balance holds as follows. First, if $\boldsymbol{z}_{i+1} = \boldsymbol{\phi}(\boldsymbol{z}_i) \wedge \boldsymbol{r}_{i+1} = \boldsymbol{\phi}(\boldsymbol{r}_i)$, then from the symmetry of $\phi$, $\boldsymbol{z}_i = \boldsymbol{\phi}(\boldsymbol{z}_{i+1}) \wedge \boldsymbol{r}_i = \boldsymbol{\phi}(\boldsymbol{r}_{i+1})$, and

$$\begin{aligned} p\left(\boldsymbol{z}_i, \boldsymbol{r}_i\right) T\left(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1} | \boldsymbol{z}_i, \boldsymbol{r}_i\right) &= \min\left\{p(\boldsymbol{z}_i, \boldsymbol{r}_i), p(\boldsymbol{\phi}(\boldsymbol{z}_i), \boldsymbol{\phi}(\boldsymbol{r}_i))\right\} \\ &= \min\left\{p(\boldsymbol{\phi}(\boldsymbol{z}_{i+1}), \boldsymbol{\phi}(\boldsymbol{r}_{i+1})), p(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1})\right\} \\ &= p\left(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1}\right) T\left(\boldsymbol{z}_i, \boldsymbol{r}_i | \boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1}\right) \end{aligned} \tag{54}$$

So the it holds. Second, if $\boldsymbol{z}_{i+1} = \boldsymbol{z}_i \wedge \boldsymbol{r}_{i+1} = \boldsymbol{r}_i$, then the it holds. Otherwise, $T\left(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1} | \boldsymbol{z}_i, \boldsymbol{r}_i\right) = 0$, and it trivially holds.

## 11.6 Additional Gibb Sampling per Step and Ergodicity

The Hamiltonian step described above itself does not ensure ergodicity. In fact, without additional step, it oscillates between $(\boldsymbol{z}_i, \boldsymbol{r}_i)$ and $(\boldsymbol{z}_{i+1}, \boldsymbol{r}_{i+1})$ due to the symmetry of $\phi$. For ergodicity, we draw a fresh $\boldsymbol{r}$ before each Hamiltonian step. So, one step to draw a sample consists of the following.

- Before the step, we have $(\boldsymbol{z}_{i-1}, \boldsymbol{r}_{i-1})$, and corresponding energy $H_{i-1} = U(\boldsymbol{z}_{i-1}) + K(\boldsymbol{r}_{i-1})$.
- Draw $\boldsymbol{r}^*_{i-1} \sim \mathcal{N}\left(\boldsymbol{r}^*_{i-1}; 0, M^{-1}\right)$. Energy changes to $H^*_{i-1} = U(\boldsymbol{z}_{i-1}) + K(\boldsymbol{r}^*_{i-1})$.
- Make transition to $(\boldsymbol{z}_i, \boldsymbol{r}_i) = \left(\boldsymbol{\phi}(\boldsymbol{z}_{i-1}), \boldsymbol{\phi}(\boldsymbol{r}^*_{i-1})\right)$ if it is accepted. The energy does not change theorecitally, but might change due to numerical error. $H^*_i = U(\boldsymbol{z}_i) + K(\boldsymbol{r}_i) \approx H^*_{i-1}$

We then consider the probability distribution in terms of the energy $H$. Please see figure 5. When $\boldsymbol{r}^*_{i-1}$ is drawn from a normal distribution, the probability distribution of the new kinetic energy will be $\chi^2$-distribution of degree $d$,

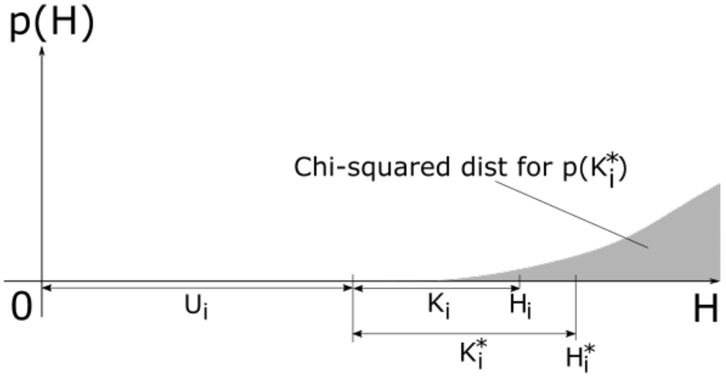

Figure 5: Probability distribution over energy

as $U(\boldsymbol{r}_{i-1}^*) \propto \sum_{j=1}^{d} r_j^2$, and each $r_j$ is drawn from a normal distribution. The new $H_{i-1}^*$ is chosen accordingly.

In the Hamiltonian step, $H_i$ does not change much from $H_{i-1}^*$, but each of $U(\boldsymbol{z}_i)$ and $K(\boldsymbol{r}_i)$ change. Since it is a deterministic process dictated by the laws of physics, it is not possible to express the change in probability. However, $U(\boldsymbol{z}_i)$ has a tendency to move toward 0, i.e., from higher energy to lower energy, and it has a tendency to stay in the lower potential energy area unless external kinetic energy is given. This ensures $H$ covers lower energy area.

# References

[1] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1):5–43, 2003.

[2] V.I. Arnold. *Mathematical Methods of Classical Mechanics.* Springer-Verlag, 2 edition, 1989.

[3] David Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2011.

[4] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. https://arxiv.org/abs/1701.02434, 01 2017. Accessed: 2020-03-30.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 1 edition, 2007.

[6] G. E. P. Box and Mervin E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 1958.

[7] W.K. Hatings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 50(1):97–109, 1970.

[8] George Marsaglia. Generating a variable from the tail of the normal distribution, mathemacital note no. 322, dtic accession number ad0423993. Technical report, Boeing Scientific Research Labs, 09 1963.

[9] Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[10] Dartmouth University. Course notes ergodic markov chains 8/14/2006. https://math.dartmouth.edu/archive/m20x06/public_html/Lecture15.pdf. Accessed: 2020-03-30.