Chapter

Markov Chain Monte Carlo in a Dynamical System of Information Theoretic Particles

Tokunbo Ogunfunmi and Manas Deb

Abstract

In Bayesian learning, the posterior probability density of a model parameter is estimated from the likelihood function and the prior probability of the parameter. The posterior probability density estimate is refined as more evidence becomes available. However, any non-trivial Bayesian model requires the computation of an intractable integral to obtain the probability density function (PDF) of the evidence. Markov Chain Monte Carlo (MCMC) is a well-known algorithm that solves this problem by directly generating the samples of the posterior distribution without computing this intractable integral. We present a novel perspective of the MCMC algorithm which views the samples of a probability distribution as a dynamical system of Information Theoretic particles in an Information Theoretic field. As our algorithm probes this field with a test particle, it is subjected to Information Forces from other Information Theoretic particles in this field. We use Information Theoretic Learning (ITL) techniques based on Rényi's α-Entropy function to derive an equation for the gradient of the Information Potential energy of the dynamical system of Information Theoretic particles. Using this equation, we compute the Hamiltonian of the dynamical system from the Information Potential energy and the kinetic energy. The Hamiltonian is used to generate the Markovian state trajectories of the system.

Keywords: Hamiltonian Monte Carlo (HMC), information theoretic learning, Kernel density estimator (KDE), Markov chain Monte Carlo, Parzen window, Rényi's entropy, information potential

1. Introduction

Bayesian learning involves estimating the PDF of a model parameter from the likelihood function and the prior probability of the parameter. Bayesian inference incorporates the concept of belief where the parameter estimate is refined as more data or evidence becomes available. The posterior PDF of the model parameter θ with the PDF of the evidence X denoted as $P(X)$, is expressed by the following well-known Bayes' equation:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{1}$$

$P(X)$ is the integral of the probability of all possible values of θ weighted by the likelihood function:

$$P(X) = \int_{\theta} P(X|\theta)P(\theta)d\theta \tag{2}$$

This is an intractable integration for most non-trivial Bayesian inference problems and makes it impossible to compute the posterior probability. The MCMC algorithm described in [1] provides a solution to this problem by directly generating samples of the posterior PDF without computing this intractable integral. The shape of the posterior PDF and other statistics can be inferred from these samples.

The MCMC algorithm requires knowledge of a function that is proportional to the unknown posterior PDF. It uses this function to generate sample proposals of the unknown PDF. Usually, this function is the product of the likelihood function and the prior probability. In practical applications, one often encounters a system whose outputs are observable, but the process within the system that generated these outputs are unknown. We present a novel perspective on the MCMC method to solve these types of practical problems, where instead of generating the samples of the unknown PDF, it uses the samples of the unknown distribution to estimate the PDF. In this chapter we use the Hamiltonian MCMC (HMC) method described in [2–4] and ITL concepts to show how the samples of the unknown distribution can be viewed as Information Theoretic particles of a dynamical system. The sample space of the given probability distribution is explored by computing trajectories corresponding to the state transition of this dynamical system. The evolution or state transition of the dynamical system is governed by equations which use the total energy or the Hamiltonian of the system of Information Theoretic particles. Each such particle has an inherent Information Potential by virtue of its position with respect to the other particles of the system. The system of Information Theoretic particles creates an Information Field which enables each particle to exert an Information Force on the other particles. We use ITL techniques [5] based on Rényi's α-Entropy function to derive an equation for the gradient of the Information Potential energy of this dynamical system. This equation is one of the main contributions of our work and it is used to compute the Hamiltonian of the system to explore the probability space of the Information Theoretic particles.

In this work, we implement an iterative PDF estimator of an unknown sample distribution, using the HMC method. At every iteration of the estimator, the HMC generates samples such that the mutual information between the generated samples and the given unknown distribution is large. To do this, it uses the Information Potential, the Information Force and the kinetic energy of an Information Theoretic "probe" particle. To compute the Information Potential and the Information Force, the algorithm uses a non-parametric Kernel Density Estimator (KDE). The bandwidth of the KDE determines how close the generated samples are from the unknown sample distribution. At the end of each iteration, the Kullback–Leibler (K–L) divergence of the samples generated by the estimator from the given distribution is computed. The iteration continues until the K-L divergence falls below a specified threshold. We have derived an equation to adapt the kernel bandwidth for each iteration, based on the invariant point theorem. Before starting the next iteration, this equation is used to adapt the kernel bandwidth before generating the next set of samples.

An important application of our algorithm is in machine learning where sometimes the dataset is either too large to fit in the memory of a computer or too small to obtain an accurate inference model. The dataset can be resampled to the desired size using the PDF estimator and the HMC equations derived in this chapter.

The sections in this chapter are organized in the following manner: In Section 2 we review the MCMC algorithm. Section 3 provides an overview of the Hamiltonian MCMC algorithm. Rényi's Entropy and the concept of Information Theoretic particles are introduced in Section 4. In Section 5 we show how the Hamiltonian MCMC algorithm can be used with Information Theoretic particles and derive a key equation for the system potential gradient. Section 6 describes a method to iteratively estimate the PDF of the target distribution using HMC. In this section we derive an equation to adapt the Information Potential energy estimator bandwidth for each iteration. The simulation results of the HMC algorithm on a system of Information Theoretic particles are listed in Section 7 and we summarize our conclusions in Section 8 of this chapter.

2. Review of the MCMC algorithm

The core principle underlying MCMC techniques is that an ergodic, reversible Markov chain reaches a stationary state. MCMC models the sampling from a distribution as an ergodic and reversible Markov process. When this process reaches a stationary state, the probability distribution of the states of the Markov chain becomes invariant and matches the given probability distribution. The sampling operation in the MCMC is a Markov process that satisfies the following detailed balance equation:

$$\pi_i \mathrm{P}(X_{t-1} = i, X_t = j) = \pi_j \mathrm{P}(X_{t-1} = j, X_t = i) \quad \forall i, j \tag{3}$$

In the detailed balance equation, π_i and π_j are the stationary probability distribution of being in states i and j respectively and $X_0, X_1, X_2, \ldots X_t \ldots$ are a sequence of random variables at discrete time indices $0, 1, 2, \ldots t-1, t, \ldots$. The Monte Carlo part of the MCMC algorithm is used to generate random "proposal" samples from a known probability distribution $Q(X)$. The proposal sample for the next time step of the MCMC algorithm is dependent on the current proposal sample and the transition probability for the new sample is enforced by an acceptance function. The proposal distribution is usually symmetric to ensure the reversibility of the Markov chain:

$$Q(x_t | x_{t-1}) = Q(x_{t-1} | x_t) \tag{4}$$

Symmetric distributions like the Gaussian distribution or the Uniform distribution centered around the current sample value can be used to generate the proposal sample. There are cases where asymmetric distributions are used but we will focus on symmetric distributions to illustrate our algorithm, without any loss of generality.

To lay the groundwork for the HMC, we review the simple Metropolis-Hastings (MH) MCMC [6] in this section. The simplest MH algorithm is the Random-Walk MH which uses a symmetrical proposal distribution. It comprises of the following 3 parts:

1. Generate a proposal sample for the posterior probability from a known symmetric distribution. The new proposal sample is based on the current proposal sample: $x_{proposal} \sim Q(x_i | x_{i-1})$. For example, if $Q(X)$ is a Gaussian distribution, it is centered at sample x_{i-1} to generate sample x_i

2. Calculate the acceptance probability by passing this sample through the posterior density function using:

$$\begin{aligned} P(\theta|X) &= \frac{1}{Z} P(X|\theta)P(\theta) \\ \text{where } Z &= \int\limits_{\theta} P(Z|\theta)P(\theta)d\theta \end{aligned} \tag{5}$$

3. Accept the candidate sample with probability α or reject it with probability $1 - \alpha$ where α is defined in (Eq. (8))

If the proposal density function is symmetric, we have:

$$Q\left(x_{i-1} \middle| x_{proposal}\right) = Q\left(x_{proposal} \middle| x_{i-1}\right) \tag{6}$$

The acceptance function is derived as follows:

$$\begin{aligned} \frac{P\left(\theta \middle| X = x_{proposal}\right)}{P\left(\theta \middle| X = x_{i-1}\right)} \frac{Q\left(x_{i-1} \middle| x_{proposal}\right)}{Q\left(x_{proposal} \middle| x_{i-1}\right)} &= \frac{\cancel{\frac{1}{Z}} P\left(X = x_{proposal} \middle| \theta\right) P(\theta)}{\cancel{\frac{1}{Z}} P\left(X = x_{i-1} \middle| \theta\right) P(\theta)} \frac{\cancel{Q\left(x_{i-1} \middle| x_{proposal}\right)}}{\cancel{Q\left(x_{proposal} \middle| x_{i-1}\right)}} \\ &= \frac{P\left(X = x_{proposal} \middle| \theta\right) P(\theta)}{P\left(X = x_{i-1} \middle| \theta\right) P(\theta)} \\ &= \frac{P\left(X = x_{proposal}, \theta\right)}{P\left(X = x_{i-1}, \theta\right)} \end{aligned} \tag{7}$$

It is evident from (Eq. (7)) that since the acceptance function is a ratio of the posterior probability, the intractable integral to compute the value of Z is completely bypassed. The acceptance probability of a sample proposal of the MH-MCMC is:

$$\alpha = \min\left\{1, \frac{P(X = x_{proposal}, \theta)}{P(X = x_{i-1}, \theta)}\right\} \tag{8}$$

The transition probability of each state of the Markov chain is defined by the acceptance probability. In the stationary state, the product of the Markov chain state probability and the transition probability matrix remains stationary and matches the posterior PDF of the model parameter. The sample points x_i generated by this MCMC in the stationary state of the Markov chain are therefore the sample points of the posterior PDF.

3. The Hamiltonian MCMC algorithm

Instead of the random-walk method of the Metropolis-Hastings algorithm, this MCMC technique uses Hamiltonian dynamics to sample from the posterior PDF. The random-walk method of the Metropolis-Hastings algorithm is inefficient and converges slowly to the target posterior distribution. Instead of randomly generating "proposal" samples from a known probability distribution, the Hamiltonian method uses the dynamics of a physical system to generate these samples. This enables the system to explore the target posterior probability space more efficiently, which in turn results in faster convergence compared to random-walk methods.

Hamiltonian dynamics is a concept borrowed from statistical mechanics where the energy of a dynamic system changes from potential energy to kinetic energy and back. The Hamiltonian represents the total energy of the system, which for a closed system, is the sum of its potential and kinetic energy.

As described in [2, 3], Hamiltonian dynamics operates on an N dimensional position vector $\mathbf{q}$ and an N dimensional momentum vector $\mathbf{p}$ and the dynamic system is described by the Hamiltonian $H(\mathbf{q}, \mathbf{p})$. The partial derivatives of the Hamiltonian define how the system evolves with time:

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \qquad i = 1, 2, \dots, N \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} \end{aligned} \tag{9}$$

Given the state of the system at time t, these equations can be used to determine the state of the system at time $t + T$ where $T = 1, 2, 3, \dots$. For the time evolution of the dynamical system, we use the following Hamiltonian:

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + K(\mathbf{p}) \tag{10}$$

In (10), $U(\mathbf{q})$ is the potential energy and $K(\mathbf{p})$is the kinetic energy of the system. The position vector $\mathbf{q}$ corresponds to the model parameter and the PDF of $\mathbf{q}$ is the target posterior PDF that we want to estimate. The potential energy of the Hamiltonian system is expressed as the negative log of the probability of $\mathbf{q}$:

$$U(\mathbf{q}) = -\log(P(\mathbf{q})) \tag{11}$$

To relate the Hamiltonian $H(\mathbf{q}, \mathbf{p})$ to the target posterior probability, we use a basic concept from statistical mechanics known as the canonical ensemble. If there are several microstates of a physical system contained in the vector $\boldsymbol{\theta}$ and there is an energy function $E(\boldsymbol{\theta})$ defined for these microstates, then the canonical probability distribution of the microstates is expressed as:

$$p(\theta) = \frac{1}{Z} e^{-\frac{E(\theta)}{T}} \tag{12}$$

where T is the temperature of the system and the variable Z is a normalizing constant called the partition function. Z scales the canonical probability distribution such that it sums to one. For a system described by Hamiltonian dynamics, the energy function is:

$$E(\boldsymbol{\theta}) = H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + K(\mathbf{p}) \tag{13}$$

In MCMC, the Hamiltonian is an energy function of the states of both the position $\mathbf{q}$ and the momentum $\mathbf{p}$. Therefore, the canonical probability distribution of a Hamiltonian system can be expressed as:

$$\begin{aligned} P(\mathbf{q}, \mathbf{p}) &= \frac{1}{Z} e^{-\frac{H(\mathbf{q}, \mathbf{p})}{T}} \\ &= \frac{1}{Z} e^{-\frac{U(\mathbf{q}) + K(\mathbf{p})}{T}} \\ &= \frac{1}{Z} \exp\left(-\frac{U(\mathbf{q})}{T}\right) \exp\left(-\frac{K(\mathbf{p})}{T}\right) \end{aligned} \tag{14}$$

This equation shows that $\mathbf{q}$ and $\mathbf{p}$ are independent and each have canonical distributions with energy functions $U(\mathbf{q})$ and $K(\mathbf{p})$. The probability density of $\mathbf{q}$ is the posterior probability density of the model parameter θ and is the product of the likelihood function of $\boldsymbol{\theta}$ given the data $\mathbf{D}$ and the prior probability of $\boldsymbol{\theta}$. An important point to note here is that the momentum variable $\mathbf{p}$ has been introduced in the probability distribution in (Eq. (14)) so that we can use Hamiltonian dynamics. Since $\mathbf{p}$ is independent of $\mathbf{q}$, we can choose any distribution for this variable. In our HMC algorithms use a zero-mean multivariate Gaussian distribution for the momentum vector $\mathbf{p}$. The temperature $T = 1$ in this discussion on the HMC.

The kinetic energy of the dynamical system for a unit mass is expressed as:

$$K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T\mathbf{p} \tag{15}$$

On applying the Hamiltonian partial derivatives in (9) to the definition of the HMC in (10) we get the following differential equations which describe the time evolution of the dynamical system:

$$\begin{aligned} \frac{d\mathbf{q}}{dt} &= \frac{\partial H}{\partial \mathbf{p}} = \frac{\partial[U(\mathbf{q}) + K(\mathbf{p})]}{\partial \mathbf{p}} = \frac{\partial}{\partial \mathbf{p}}\left(\frac{1}{2}\mathbf{p}^T\mathbf{p}\right) = \mathbf{p} \\ \frac{d\mathbf{p}}{dt} &= -\frac{\partial H}{\partial \mathbf{q}} = -\frac{\partial[U(\mathbf{q}) + K(\mathbf{p})]}{\partial \mathbf{q}} = -\frac{\partial U(\mathbf{q})}{\partial \mathbf{q}} \end{aligned} \tag{16}$$

Since the Hamiltonian equations for the time evolution of the system are differential equations, computer simulation of the HMC must discretize time. A popular scheme to implement this discretization is the “Leapfrog” algorithm [4]. The HMC algorithm uses the leapfrog algorithm to update the momentum and the position while computing the trajectory towards the next sample proposal in the distribution. The Leapfrog integrator has 2 main advantages:

1. It is time reversible. A Leapfrog integration by N steps in the forward direction and then in the backward direction results in the same starting position

2. It is symplectic in nature. In other words, it conserves the energy of dynamical systems

The steps of the Hamiltonian MCMC algorithm are:

1. At every time step t, determine a trajectory of the system potential and kinetic energy. To do that, generate a random value from a standard normal distribution for the momentum variable.

2. Execute the Leapfrog algorithm to update the position and momentum variables according to the differential equations in (Eq. (16)). This determines the trajectory of the system towards the next sample proposal

3. Compute the potential and kinetic energy $(U(\mathbf{q}_{t-1}), K(\mathbf{p}_{t-1}))$of the system at the beginning of the trajectory and at the end $\left(U\left(\mathbf{q}_{proposed}\right), K\left(\mathbf{p}_{proposed}\right)\right)$ of the proposed trajectory

4. Calculate the acceptance probability of the new trajectory using the following ratio of probabilities:

$$\begin{aligned}\beta &= \min\left\{1, \frac{P\left(\mathbf{q}_{proposal}, \mathbf{p}_{proposal}\right)}{P(\mathbf{q}_{t-1}, \mathbf{p}_{t-1})}\right\} \\ &= \min\left\{1, \frac{\frac{1}{Z}\exp\left(-U\left(\mathbf{q}_{proposal}\right)\right)\exp\left(-K\left(\mathbf{p}_{proposal}\right)\right)}{\frac{1}{Z}\exp\left(-U(\mathbf{q}_{t-1})\right)\exp\left(-K(\mathbf{p}_{t-1})\right)}\right\} \\ &= \min\left\{1, \exp\left(\begin{array}{c}(U(\mathbf{q}_{t-1}) + K(\mathbf{p}_{t-1}))- \\ \left(U\left(\mathbf{q}_{proposal}\right) + K\left(\mathbf{p}_{proposal}\right)\right)\end{array}\right)\right\}\end{aligned} \tag{17}$$

5. Generate a random number $u \sim Uniform(0, 1)$ to accept or reject the proposal

```
if (β > u) then
    $\mathbf{q}_t \leftarrow \mathbf{q}_{proposal}$   //accept the proposed trajectory
else
    $\mathbf{q}_t \leftarrow \mathbf{q}_{t-1}$   //reject the proposed trajectory
endif
```

4. Rényi's entropy and Information Theoretic particles

The concept of Information Theoretic particles comes from Alfréd Rényi's pioneering work on generalized measures of entropy and information [7]. At the core of Rényi's work is the concept of generalized mean or the Kolmogorov-Nagumo (K-N) mean [8–10]. For numbers $x_1, x_2, \ldots x_N$, the K-N mean is expressed as:

$$\psi^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\psi(x_i)\right) \tag{18}$$

where, $\psi(.)$ is the K-N function. This function is continuous and strictly monotonic implying that it has an inverse. In the general theory of means, the quasi-linear mean of a random variable X which takes the values $x_1, x_2, \ldots x_N$ with probabilities $p_1, p_2, \ldots p_N$ is defined as:

$$E_\psi[X] = \langle X\rangle_\psi = \psi^{-1}\left(\sum_{k=1}^{N}p_k\psi(x_k)\right) \tag{19}$$

From the theorem on additivity of quasi-linear means [11], if $\psi(.)$ is a K-N function and $c$is a real constant, then:

$$\psi^{-1}\left(\sum_{k=1}^{N}p_k\psi(x_k + c)\right) = \psi^{-1}\left(\sum_{k=1}^{N}p_k\psi(x_k)\right) + c \tag{20}$$

if and only if $\psi(.)$ is either linear or exponential.

4.1 Rényi's entropy

Consider a random variable X which takes the values $x_1, x_2, \ldots x_N$ with probabilities $p_1, p_2, \ldots p_N$. The amount of information generated when X takes the value x_k is given by the Hartley [12] information measurement function $I(x_k)$:

$$I(x_k) = \log_2\left(\frac{1}{p_k}\right) \text{ bits} \tag{21}$$

The expected value of $I(x_k)$ yields the expression for Shannon's entropy [13]:

$$H(X) = \sum_{k=1}^{N} p_k I(x_k) = \sum_{k=1}^{N} p_k \log_2\left(\frac{1}{p_k}\right) \tag{22}$$

Rényi replaced the linear mean in (Eq. (22)) with the quasi-linear mean in (Eq. (19)) to obtain a generalized measure of information:

$$H_\psi(X) = \psi^{-1}\left(\sum_{k=1}^{N} p_k \psi\left(\log_2\left(\frac{1}{p_k}\right)\right)\right) \tag{23}$$

For $H_\psi(X)$ to satisfy the additivity property of independent events, it must satisfy$\langle X + c\rangle_\psi = \langle X\rangle_\psi + c$ where c is a constant. From (Eq. (20)), this implies that $\psi(x) = cx$ (linear) or $\psi(x) = c2^{(1-\alpha)x}$ (exponential). Setting $\psi(x) = cx$ reduces (Eq. (23)) to the linear mean and yields Shannon entropy equation. Substituting $\psi(x) = c2^{(1-\alpha)x}$ and the corresponding inverse function $\psi^{-1} = \frac{1}{(1-\alpha)} \log_2$ in (Eq. (23)) yields the expression for Rényi's $\alpha-$entropy:

$$H_\alpha(X) = \frac{1}{(1-\alpha)} \log_2\left(\sum_{k=1}^{N} p_k^\alpha\right) \qquad \alpha > 0 \text{ and } \alpha \neq 1 \tag{24}$$

Rényi's $\alpha-$entropy equation is therefore a general expression for entropy and comprises of a family of entropies for different values of the parameter α. Shannon's entropy is a special case of Rényi's entropy in the limit as $\alpha \to 1$. The argument of the logarithm function in (Eq. (24)) is called the Information Potential. The α-Information Potential is expressed as:

$$V_\alpha(X) = \sum_{k=1}^{N} p_k^\alpha \tag{25}$$

Substituting (Eq. (25)) in (Eq. (24)), we get the following expression for Rényi's entropy in terms of the Information Potential:

$$H_\alpha(X) = \frac{1}{(1-\alpha)} \log_2(V_\alpha(X)) \tag{26}$$

The Information Potential in (Eq. (25)) can be written as the expected value of the PDF of the sample distribution raised to $\alpha - 1$:

$$V_\alpha(X) = \sum_{k=1}^{N} p_k^\alpha = \sum_{k=1}^{N} p_k p_k^{\alpha-1} = E\left[p_k^{\alpha-1}\right] \tag{27}$$

For $\alpha = 2$ in (Eq. (24)), we get Rényi's quadratic entropy, which has the useful property that it allows us to compute the entropy directly from the samples. The equations for Rényi's Quadratic Entropy (QE) and Quadratic Information Potential (QIP) are obtained by substituting $\alpha = 2$ in (Eqs. (26) and (27)):

$$\begin{aligned} H_2(X) &= \frac{1}{(1-2)} \log_2(V_2(X)) = -\log_2(V_2(X)) \\ \text{where } V_2(X) &= E\left[p_k^{2-1}\right] = E\left[p_k\right] \end{aligned} \tag{28}$$

The QIP is therefore the expected value of the PDF of the given data samples.

4.2 Rényi's quadratic information potential (QIP) estimator

From (Eq. (28)), it is evident that to compute the QIP we need to know the PDF of the given data samples. In practical applications an analytical expression of the PDF is rarely available. Therefore, the QIP computation involves a non-parametric estimator of the PDF directly from the samples [14]. The Parzen-Rosenblatt window estimator [15, 16] is a non-parametric way to estimate the PDF of a random variable from its sample values. This estimator places a kernel function with its center at each of the samples. The resulting output values are averaged over all the samples to estimate the PDF. The laws governing the interaction of the Information Theoretic particles is defined by the shape of the kernel. We use a Gaussian kernel, since this kernel when placed over the samples, behaves like an Information Theoretic field whose strength decays with increasing distance between the samples. Just like a charge in space creates an electric field, the samples of a probability distribution behave like Information Particles with unit charge. Information particles exert Information Forces on other particles through this Information Theoretic field.

For scalar samples $x_1, x_2, \ldots x_N$, the Parzen window PDF estimator with a Gaussian kernel is expressed as:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} G_\sigma(x - x_i) \tag{29}$$

where $G_\sigma(u)$ is the following standard univariate Gaussian kernel:

$$G_\sigma(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{u}{\sigma}\right)^2\right] \tag{30}$$

σ is the kernel bandwidth of the estimator and it must be carefully chosen to obtain an accurate and unbiased estimate of the PDF. The Parzen window estimator of a multivariate PDF for vector samples $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N$ of dimension d is expressed as:

$$\hat{p}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} G_{\mathbf{C}}(\mathbf{x} - \mathbf{x}_i) \tag{31}$$

where $G_{\mathbf{C}}(u)$ is the following standard multivariate Gaussian kernel:

$$G_{\mathbf{C}}(\mathbf{u}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{C}|}} \exp\left[-\frac{1}{2}\mathbf{u}^T \mathbf{C}^{-1} \mathbf{u}\right] \tag{32}$$

d is the dimension of the input vector $\mathbf{u}$, $\mathbf{C}$ is the $d \times d$ covariance matrix and $|\mathbf{C}|$ is the determinant of the covariance matrix. For the multivariate PDF case, the kernel bandwidth $\mathbf{C}$ must be carefully chosen to obtain an accurate and unbiased estimate of the PDF.

Rényi's quadratic entropy for a continuous random variable is expressed as:

$$H_2(X) = -\log \int_{-\infty}^{\infty} p^2(x)dx \tag{33}$$

Substituting $\hat{p}(x)$ from (Eq. (29)) for $p(x)$ in the above equation as described in [5], we get the following equation for the QE estimator:

$$\hat{H}_2(X) = -\log\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i)\right] \tag{34}$$

where:

$$G_{\sigma\sqrt{2}}(u) = \frac{1}{\sqrt{2\pi(\sigma\sqrt{2})^2}} \exp\left[-\frac{1}{2}\left(\frac{u}{\sigma\sqrt{2}}\right)^2\right] \tag{35}$$

The equation for the QE estimator shows that we can compute the QE estimate directly from the samples of a distribution without knowing its PDF, by applying the Parzen-Rosenblatt kernel on these samples. From (Eqs. (28) and (34)), the QIP estimator can be expressed as:

$$\hat{V}_2(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i) \tag{36}$$

4.3 Information potential energy and the information force of information theoretic particles

The total QIP energy estimate of the system is given by (Eq. (36)). The QIP energy estimate of sample x_j due to the Information Potential field of a single sample x_i is:

$$\hat{V}_2(x_j; x_i) = G_{\sigma\sqrt{2}}(x_j - x_i) \tag{37}$$

The Quadratic Information Potential energy estimate of scalar sample x_j in the Information Field created by all the samples $x_i \in \mathbb{R}$, for $i = 1, 2, \dots N$ is defined as the average of $\hat{V}_2(x_j; x_i)$ taken over all the samples x_i:

$$\begin{aligned} \hat{V}_2(x_j) &= \frac{1}{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i) \\ &= \frac{1}{N}\frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})}\sum_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right] \end{aligned} \tag{38}$$

If the samples are d dimensional vectors, then the Quadratic Information Potential energy estimate of vector sample $\mathbf{x}_j$ in the Information Potential Field created by all vector samples $\mathbf{x}_i \in \mathbb{R}^d$, for $i = 1, 2, \ldots N$ is defined as:

$$\hat{V}_2(\mathbf{x}_j) = \frac{1}{N}\sum_{i=1}^{N} G_{2\mathrm{C}}(\mathbf{x}_j - \mathbf{x}_i) \tag{39}$$

where:

$$G_{2\mathrm{C}} = \frac{1}{\sqrt{(2\pi)^d |\mathbf{C}| (2^d)}} \sum_{i=1}^{N} \exp\left[-\frac{1}{2}(\mathbf{x}_j - \mathbf{x}_i)^T (2\mathbf{C})^{-1} (\mathbf{x}_j - \mathbf{x}_i)\right] \tag{40}$$

From (Eqs. (39) and (40)) we can re-write the QIP energy estimate for vector samples of d dimensions as:

$$\hat{V}_2(\mathbf{x}_j) = \frac{1}{N} \frac{1}{\sqrt{(2\pi)^d |\mathbf{C}| (2^d)}} \sum_{i=1}^{N} \exp\left[-\frac{1}{2}(\mathbf{x}_j - \mathbf{x}_i)^T (2\mathbf{C})^{-1} (\mathbf{x}_j - \mathbf{x}_i)\right] \tag{41}$$

To obtain the Quadratic Information Force estimate on scalar sample x_j due to the Information Potential field of sample x_i, we take the derivative of the Quadratic Information Potential energy estimate:

$$\begin{aligned}
\hat{F}_2(x_j; x_i) &= \frac{\partial}{\partial x_j}\hat{V}_2(x_j; x_i) = \frac{\partial}{\partial x_j} G_{\sigma\sqrt{2}}(x_j - x_i) \\
&= \frac{\partial}{\partial x_j}\left[\frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right]\right] \\
&= \frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right]\left(-\frac{1}{2(2\sigma^2)}\right)\frac{\partial}{\partial x_j}(x_j - x_i)^2 \\
&= \frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right]\left(-\frac{1}{2(2\sigma^2)}\right)[2(x_j - x_i)] \\
\hat{F}_2(x_j; x_i) &= \left(\frac{1}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right][(-1)(x_j - x_i)] \\
&= \left(\frac{1}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right](x_i - x_j) \qquad (42) \\
&= \left(\frac{1}{2\sigma^2}\right) G_{\sigma\sqrt{2}}(x_j - x_i)(x_i - x_j)
\end{aligned}$$

The Quadratic Information Force on scalar sample x_j in the Information Potential Field created by all the samples $x_i \in \mathbb{R}$, for $i = 1, 2, \ldots N$ is defined as the average of $\hat{F}_2(x_j; x_i)$ taken over all the samples x_i:

$$\begin{aligned}\hat{F}_2(x_j) &= \frac{1}{N(2\sigma^2)} \sum_{i=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i)(x_i - x_j) \\ &= \frac{1}{(2N\sigma^2)} \frac{1}{\sqrt{2\pi}(\sigma\sqrt{2})} \sum_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{x_j - x_i}{\sigma\sqrt{2}}\right)^2\right](x_i - x_j)\end{aligned} \tag{43}$$

If the samples are d dimensional vectors, then the Quadratic Information Force on vector sample $\mathbf{x}_j$ in the Information Potential Field created by all samples $\mathbf{x}_i \in \mathbb{R}^d$, for $i = 1, 2, \dots N$ is defined as:

$$\begin{aligned}\hat{F}_2(\mathbf{x}_j) &= \frac{1}{(2^d N|\mathbf{C}|)} \sum_{i=1}^{N} G_{2\mathbf{C}}(\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j) \\ &= \frac{1}{(2^d N|\mathbf{C}|)} \frac{1}{\sqrt{(2\pi)^d |\mathbf{C}|(2^d)}} \\ &\times \sum_{i=1}^{N} \exp\left[-\frac{1}{2}(\mathbf{x}_j - \mathbf{x}_i)^T (2\mathbf{C})^{-1}(\mathbf{x}_j - \mathbf{x}_i)\right](\mathbf{x}_i - \mathbf{x}_j)\end{aligned} \tag{44}$$

5. Hamiltonian MCMC with information theoretic particles

The expression for the potential energy in the Hamiltonian function (Eq. (11)) is similar to the expression for Rényi's quadratic entropy (Eq. (28)). This is consistent with the principles of statistical mechanics where the entropy is related to the dissipation of the potential energy of the system. Based on this intuition from statistical mechanics, we replace the PDF of the position vector $\mathbf{q}$ in (Eq. (11)) with the QIP energy estimator as follows:

$$U\left(\mathbf{q}_j\right) = -\log\left[P\left(\mathbf{q}_j\right)\right] = -\log\left[\hat{V}_2\left(\mathbf{q}_j\right)\right] \tag{45}$$

The change in momentum of the j^{th} Information Theoretic particle in the dynamical system is equal to the negative potential energy gradient defined in (Eq. (16)). This can be expressed in terms of the QIP energy estimator as:

$$\frac{d\mathbf{p}}{dt} = -\frac{dU\left(\mathbf{q}_j\right)}{d\mathbf{q}_j} = -\frac{d\log\left[P\left(\mathbf{q}_j\right)\right]}{d\mathbf{q}_j} = -\frac{d\log\left[\hat{V}_2\left(\mathbf{q}_j\right)\right]}{d\mathbf{q}_j} \tag{46}$$

From the above expression, we derive the expression for the Hamiltonian system's negative potential gradient in terms of the Information Potential and the Information Force as follows:

$$-\frac{d\log\left[\hat{V}_2\left(\mathbf{q}_j\right)\right]}{d\mathbf{q}_j} = -\frac{d}{d\mathbf{q}_j}\log\left[\begin{array}{l}\dfrac{1}{N}\dfrac{1}{\sqrt{(2\pi)^d|\Sigma|(2^d)}}\times \\ \sum\limits_{i=1}^{N}\exp\left[-\dfrac{1}{2}\left(\mathbf{q}_j - \mathbf{q}_i\right)^T(2\Sigma)^{-1}\left(\mathbf{q}_j - \mathbf{q}_i\right)\right]\end{array}\right]$$

$$
\begin{aligned}
-\frac{d\log\left[\hat{V}_2\left(\mathbf{q}_j\right)\right]}{d\mathbf{q}_j} &= -\frac{d}{d\mathbf{q}_j}\log\left[\frac{1}{N}\frac{1}{\sqrt{(2\pi)^d|\Sigma|(2^d)}}\right] \\
&\quad -\frac{d}{d\mathbf{q}_j}\log\left[\sum_{i=1}^{N}\exp\left[-\frac{1}{2}\left(\mathbf{q}_j-\mathbf{q}_i\right)^T(2\Sigma)^{-1}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right]\right] \\
&= -\frac{1}{\sum_{i=1}^{N}\exp\left[-\frac{1}{2}\left(\mathbf{q}_j-\mathbf{q}_i\right)^T(2\Sigma)^{-1}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right]} \\
&\quad \times\sum_{i=1}^{N}\frac{d}{d\mathbf{q}_j}\exp\left[-\frac{1}{2}\left(\mathbf{q}_j-\mathbf{q}_i\right)^T(2\Sigma)^{-1}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right] \\
&= -\frac{1}{\sum_{i=1}^{N}\exp\left[-\frac{1}{2}\left(\mathbf{q}_j-\mathbf{q}_i\right)^T(2\Sigma)^{-1}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right]} \\
&\quad \times\frac{1}{\left(2^d|\Sigma|\right)}\sum_{i=1}^{N}\exp\left[-\frac{1}{2}\left(\mathbf{q}_j-\mathbf{q}_i\right)^T(2\Sigma)^{-1}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right]\left(\mathbf{q}_i-\mathbf{q}_j\right) \\
&= -\frac{\hat{F}_2\left(\mathbf{q}_j\right)}{\hat{V}_2\left(\mathbf{q}_j\right)}
\end{aligned}
\tag{47}
$$

This result shows that the gradient of the potential energy of the Hamiltonian system of Information Particles is just the Information Force estimate normalized by the Information Potential energy estimate. This also shows that the Information Force vector influences the trajectory of sample proposals in the HMC algorithm. This equation is one of the important contributions of our work. Our simulation of the HMC of a dynamical system of Information Theoretic particles uses this potential energy gradient equation to evolve the system over time.

5.1 Quality of the information potential energy estimator

As described in [5], the Information Potential energy estimator is a kernel estimator of the 2-norm of the underlying PDF of the Information Particles. Just like a PDF estimator, we can define metrics to describe the quality of the Information Potential energy estimator. The Mean Integrated Square Error (MISE) is an important metric used to assess the quality of an estimator. This is expressed as:

$$
\begin{aligned}
MISE\left[\hat{V}_2\left(q_j\right)\right] &= E\left[\int\left(\hat{V}_2\left(q_j\right)-V_2\left(q_j\right)\right)^2 dq\right] \\
&= \int E\left\{\hat{V}_2\left(q_j\right)-E\left[\hat{V}_2\left(q_j\right)\right]\right\}^2 dq+\int\left\{E\left[\hat{V}_2\left(q_j\right)\right]-V_2\left(q_j\right)\right\}^2 dq \\
&= \int Variance\left(\hat{V}_2\left(q_j\right)\right)dq+\int Bias^2\left(\hat{V}_2\left(q_j\right)\right)dq
\end{aligned}
\tag{48}
$$

The bias and variance of the Information Potential estimator can be derived as follows:

Bias:

$$\begin{aligned} E\left[\hat{V}_2\left(q_j\right)\right] - V_2\left(q_j\right) &= E\left[\frac{1}{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}\left(q_j - q_i\right)\right] - V_2\left(q_j\right) \\ &= E\left[G_{\sigma\sqrt{2}}\left(q_j - q_i\right)\right] - V_2\left(q_j\right) \end{aligned} \tag{49}$$

Since the Gaussian kernel is symmetric under the expectation operation:

$$G_{\sigma\sqrt{2}}\left(q_j - q_i\right) = G_{\sigma\sqrt{2}}\left(q_i - q_j\right) \tag{50}$$

Substituting this in (Eq. (49)) and using the definition of $G_{\sigma\sqrt{2}}$ from (Eq. (35)):

$$\begin{aligned} E\left[\hat{V}_2\left(q_j\right)\right] - V_2\left(q_j\right) &= = E\left[G_{\sigma\sqrt{2}}\left(q_i - q_j\right)\right] - V_2\left(q_j\right) \\ &= \frac{1}{\sigma\sqrt{2}} E\left[G\left(\frac{q_i - q_j}{\sigma\sqrt{2}}\right)\right] - V_2\left(q_j\right) \\ &= \frac{1}{\sigma\sqrt{2}} \int G\left(\frac{s - q_j}{\sigma\sqrt{2}}\right) V_2(s) ds - V_2\left(q_j\right) \end{aligned} \tag{51}$$

In the above equation $s$is the dummy variable of integration. Let $y = \frac{s-q_j}{\sigma\sqrt{2}}$.This implies that $dy = \frac{ds}{\sigma\sqrt{2}}$. Substituting this in (Eq. (51)), we get:

$$E\left[\hat{V}_2\left(q_j\right)\right] - V_2\left(q_j\right) = \int G(y) V_2\left(q_j + \sigma\sqrt{2}y\right) dy - V_2\left(q_j\right) \tag{52}$$

When $\sigma\sqrt{2}$ is small, we can write the Taylor series expansion of $V_2\left(q_j + \sigma\sqrt{2}y\right)$ as:

$$V_2\left(q_j + \sigma\sqrt{2}y\right) = V_2\left(q_j\right) + \sigma\sqrt{2}yV_2'\left(q_j\right) + \frac{1}{2}2\sigma^2 y^2 V_2''\left(q_j\right) + o\left(\sigma^2\right) \tag{53}$$

Substituting this in (Eq. (52)), we get:

$$\begin{aligned} &E\left[\hat{V}_2\left(q_j\right)\right] - V_2\left(q_j\right) \\ &= \int G(y)\left[V_2\left(q_j\right) + \sigma\sqrt{2}yV_2'\left(q_j\right) + \sigma^2 y^2 V_2''\left(q_j\right) + o\left(\sigma^2\right)\right] dy - V_2\left(q_j\right) \\ &= V_2\left(q_j\right)\int G(y)dy + \sigma\sqrt{2}V_2'\left(q_j\right)\int yG(y)dy + \sigma^2 V_2''\left(q_j\right)\int y^2 G(y)dy + o\left(\sigma^2\right) - V_2\left(q_j\right) \\ &= V_2\left(q_j\right)(1) + \sigma\sqrt{2}V_2'\left(q_j\right)(0) + \sigma^2 V_2''\left(q_j\right)\int y^2 G(y)dy + o\left(\sigma^2\right) - V_2\left(q_j\right) \\ &= \sigma^2 V_2''\left(q_j\right)\int y^2 G(y)dy + o\left(\sigma^2\right) \end{aligned} \tag{54}$$

This result implies that as the kernel bandwidth $\sigma \rightarrow 0$ the bias of the Information Potential energy estimator for sample $q_j$reduces at the rate of $O\left(\sigma^2\right)$. From the above equation it is also evident that the main reason for the bias is the second derivative of the true Information Potential energy (i.e., the rate of curvature of the true PDF of the samples). In other words, if the true PDF of the samples has a sharp spike, the bias of the Information Potential energy estimator will increase. The Information Potential energy estimator tends to smooth out sharp curvatures or spikes in the PDF which increases bias. The amount of smoothness is governed by the bandwidth parameter σ.

Variance:

$$\begin{aligned} E\left\{\left[\hat{V}_2\left(q_j\right)\right]^2\right\} - \left\{E\left[\hat{V}_2\left(q_j\right)\right]\right\}^2 &= E\left\{\left[\hat{V}_2\left(q_j\right)\right]^2\right\} - \frac{1}{N}\left(V_2\left(q_j\right) + Bias\right)^2 \\ &= E\left\{\left[\hat{V}_2\left(q_j\right)\right]^2\right\} + O\left(N^{-1}\right) \\ &= E\left\{\left[\frac{1}{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}\left(q_j - q_i\right)\right]^2\right\} + O\left(N^{-1}\right) \\ &= E\left\{\left[\frac{1}{N}\sum_{i=1}^{N} G_{\sigma\sqrt{2}}\left(q_i - q_j\right)\right]^2\right\} + O\left(N^{-1}\right) \\ &= \frac{1}{N} E\left\{\left[G_{\sigma\sqrt{2}}\left(q_i - q_j\right)\right]^2\right\} + O\left(N^{-1}\right) \\ &= \frac{1}{2N\sigma^2}\int G^2\left(\frac{s - q_j}{\sigma\sqrt{2}}\right) V_2(s)ds + O\left(N^{-1}\right) \end{aligned} \tag{55}$$

Let $y = \frac{s-q_j}{\sigma\sqrt{2}}$ .This implies that $dy = \frac{ds}{\sigma\sqrt{2}}$. Substituting this in (Eq. (55)), we get:

$$\begin{gathered} E\left\{\left[\hat{V}_2\left(q_j\right)\right]^2\right\} - \left\{E\left[\hat{V}_2\left(q_j\right)\right]\right\}^2 = \\ \frac{1}{N\sigma\sqrt{2}}\int G^2(y) V_2\left(q_j + \sigma\sqrt{2}y\right) dy + O\left(N^{-1}\right) \end{gathered} \tag{56}$$

When $\sigma\sqrt{2}$ is small, we can write the Taylor series expansion of $V_2\left(q_j + \sigma\sqrt{2}y\right)$ as:

$$V_2\left(q_j + \sigma\sqrt{2}y\right) = V_2\left(q_j\right) + \sigma\sqrt{2}yV_2'\left(q_j\right) + o(\sigma) \tag{57}$$

Substituting this in (Eq. (56)), we get:

$$\begin{aligned} &E\left\{\left[\hat{V}_2\left(q_j\right)\right]^2\right\} - \left\{E\left[\hat{V}_2\left(q_j\right)\right]\right\}^2 \\ &= \frac{1}{N\sigma\sqrt{2}}\int G^2(y)\left[V_2\left(q_j\right) + \sigma\sqrt{2}yV_2'\left(q_j\right) + o(\sigma)\right] ds + O\left(N^{-1}\right) \\ &= \frac{1}{N\sigma\sqrt{2}} V_2\left(q_j\right)\int G^2(y) + o\left(\frac{1}{N\sigma\sqrt{2}}\right) \end{aligned} \tag{58}$$

This result shows that as the number of samples $N \to \infty$ and kernel bandwidth $\sigma \to \infty$, the variance of the Information Potential energy estimator for the sample

q_j reduces at the rate of $O\left(\frac{1}{N\sigma\sqrt{2}}\right)$. However, as $\sigma \to 0$, the variance of the estimator increases. The result also shows that the variance of the estimator is large where the value of the Information Potential energy $V_2\left(q_j\right)$ (i.e., true probability of the sample) is also large. This happens when there are many Information Particles closer together.

5.2 The Kernel bandwidth parameter and the information potential energy estimator bias-variance trade-off

We have shown that the Gaussian kernel bandwidth σ directly influences the bias and variance of the Information Potential energy estimator. This in turn affects the sample distribution of the PDF estimate generated by the Hamiltonian MCMC. From (Eq. (54)) it is evident that the bias of the estimator reduces when we decrease the kernel bandwidth σ. However, (Eq. (58)) clearly shows that the decreasing σ increases the variance of the estimator. Therefore, we must choose an optimum bandwidth which minimizes both the systematic error (bias) and the random error (variance) of the Information Potential energy estimator. An iterative algorithm to converge to the optimum kernel bandwidth is described in the following section.

5.3 Computational complexity of the information potential energy estimator

From (Eqs. (38) and (41)) it may appear that the complexity of computing the Information Potential is $O\left(N^2\right)$. However, as described in [5], the Information Potential can be written as a symmetric positive Gramm Matrix which can be approximated using the incomplete Cholesky decomposition (ICD) as an $N \times D$ matrix where $D \ll N$. Using this technique, the time complexity for computing the Information Potential reduces to $O\left(ND^2\right)$ and the space complexity reduces to $O(ND)$.

6. Maximum-likelihood iterative algorithm to adapt the kernel bandwidth of the information potential energy estimator

There are many iterative kernel bandwidth adaptation techniques available in the literature. We present a simple iterative technique to illustrate how MCMC with Hamiltonian of Information Theoretic Particles can be used to adjust the bandwidth parameter of the iterative PDF estimator. Here, we have chosen to minimize the Kullback–Leibler (K-L) divergence between the samples of the estimated PDF and the target sample distribution as the criteria for adapting the kernel bandwidth of the Information Potential energy and Information Force estimator. As described in [17], this is equivalent to maximizing the likelihood that the estimated PDF samples output by the MCMC, has the same distribution as the target samples.

The ML estimate of the optimum kernel bandwidth $\mathbf{C}_{ML}$ for vector Information Particle samples $\mathbf{q}_j$ is the solution to the following log-likelihood maximization problem:

$$\mathbf{C}_{ML} = \arg\max_{\mathbf{C}} \sum_{j=1}^{N} \log\left[\hat{V}\left(\mathbf{q}_j|\mathbf{C}\right)\right] \tag{59}$$

Using (Eq. (41)) in the summation of the above equation, we get:

$$\sum_{j=1}^{N}\log\left[V\left(\mathbf{q}_j|\mathbf{C}\right)\right]=\sum_{j=1}^{N}\log\left[\frac{1}{N-1}\sum_{\substack{i=1\\i\neq j}}^{N}G_{2\mathrm{C}}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right]$$

$$=\sum_{j=1}^{N}\log\left[\begin{array}{l}\dfrac{1}{(N-1)}\dfrac{1}{\sqrt{(2\pi)^d\left(2^d\right)|\mathbf{C}|}}\times\\ \sum\limits_{\substack{i=1\\i\neq j}}^{N}\exp\left[-\dfrac{1}{2}\left(\mathbf{q}_j-\mathbf{q}_i\right)^T(2\mathbf{C})^{-1}\left(\mathbf{q}_j-\mathbf{q}_i\right)\right]\end{array}\right] \tag{60}$$

To maximize the above equation, we take the derivative and equate it to 0. This gives us the following update equation for scalar Information Theoretic particles:

$$\sigma_{t+1}^2=\left[\frac{1}{2N(N-1)}\sum_{j=1}^{N}\frac{1}{\hat{V}\left(q_j\right)}\sum_{\substack{i=1\\i\neq j}}^{N}G_{\sigma\sqrt{2}}\left(q_j-q_i\right)\left(q_j-q_i\right)^2\right]_t \tag{61}$$

In the above equation, σ_{t+1} is the kernel bandwidth at iteration $t+1$. It is updated with the result of the right-hand side of the equation obtained at time t. This kernel bandwidth update equation (Eq. (61)) is in the form of a fixed-point (or invariant point) equation. This equation is like the equation in [18] except for the factor of 1/2. For vector Information Theoretic particles, the kernel bandwidth update equation is:

$$\mathbf{C}_{t+1}=\left\{\frac{1}{2N(N-1)}\sum_{j=1}^{N}\frac{1}{\hat{V}\left(\mathbf{q}_j\right)}\sum_{\substack{i=1\\i\neq j}}^{N}G_{2C}\left(\mathbf{q}_j-\mathbf{q}_i\right)\left[\left(\mathbf{q}_j-\mathbf{q}_i\right)\left(\mathbf{q}_j-\mathbf{q}_i\right)^T\right]\right\}_t \tag{62}$$

In this equation, **C** is the kernel bandwidth matrix and can have unequal elements along its diagonal or non-zero off-diagonal elements. If the kernel bandwidth matrix is constrained to an identity matrix multiplied by a scaling factor, the kernel bandwidth matrix update equation can be expressed as:

$$\mathbf{C}_{t+1}=\left\{\frac{1}{2N(N-1)}\sum_{j=1}^{N}\frac{1}{V\left(\mathbf{q}_j\right)}\sum_{\substack{i=1\\i\neq j}}^{N}G_{2C}\left(\mathbf{q}_j-\mathbf{q}_i\right)\left\|\left(\mathbf{q}_j-\mathbf{q}_i\right)\right\|^2\right\}_t \tag{63}$$

From the fixed- or invariant-point theorem, the range over which the fixed-point bandwidth update equations will converge to a unique solution is:

$$\left[\overline{\frac{\min\left(q_j-q_i\right)^2}{2}},Trace\left\{E\left[\mathbf{q}\mathbf{q}^T\right]\right\}\right] \tag{64}$$

In the above equation, $q_{i,}q_{j}$ are information particles from the target sample distribution and $\mathbf{q}$ is the column vector of all the target information particles. From the fixed-point theorem, this fixed-point equation will converge to a unique solution if $\left| f'(\sigma^2) \right| < 1$.

7. Simulation results

The potential energy surface, which is the plot of (Eq. (41)), of a Hamiltonian system of Information Theoretic particles for a bivariate Gaussian distribution is shown in **Figure 1**. From this figure it is evident that the potential energy surface of the Hamiltonian system has larger values when the Information Theoretic particles are sparse and is lowest at the bottom of the bowl-shaped surface where the particles have the highest density.

The momentum variable of the HMC algorithm occasionally moves the "probe" particle to a higher energy level but the Hamiltonian system has the tendency to fall back to its lowest energy level along the bowl-shaped surface. As a result, the HMC tends to sample the given target distribution more often where the density of the Information Theoretic particles is the largest.

Figure 2 shows the potential energy gradient of the same bivariate Gaussian distribution. This is the plot of (Eq. (47)) for this distribution. Each surface in this figure is one component of the potential energy gradient. Each surface tilts towards the corresponding mean value $\mu = [-5, 6]$ of the bivariate Gaussian distribution. The figure shows that the potential energy gradient of the Hamiltonian system is lowest near the mean of the distribution and is highest further away from the mean. The time evolution trajectory of the Hamiltonian system lies on this surface.

The iterative PDF estimate of a bivariate Gaussian distribution with $\mu = [-5, 6], \Sigma = [3, 0; 0, 4]$ using MCMC with 3 different kernel bandwidths is shown in **Figure 3**.

From **Figure 3**, it is evident that the MCMC algorithm based on the Hamiltonian of Information Theoretic particles accurately estimates the PDF of the target distribution. The sample points generated by the HMC algorithm covers most of the target samples in this figure. This figure shows that our intuition of comparing the Entropy to the system's potential and also using the Information Potential in the derivation of the potential gradient (Eq. (47)) of the Hamiltonian system of Information Theoretic particles, was correct.

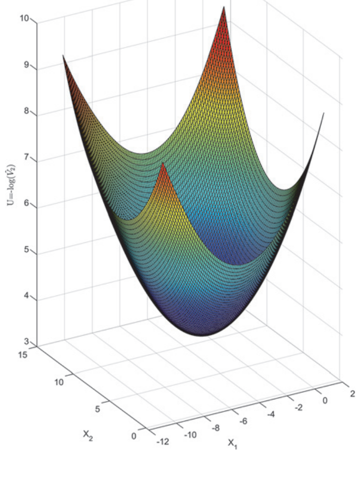

Figure 1.
Potential energy surface of the Hamiltonian system of a bivariate Gaussian $(\mu = [-5, 6], \Sigma = [3, 0; 0, 4])$ *distribution of Information Theoretic particles.*

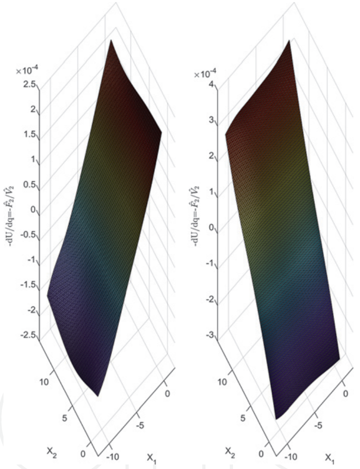

Figure 2.
Vector components of the potential energy gradient of the Hamiltonian system of a bivariate Gaussian $(\mu = [-5, 6], \Sigma = [3, 0; 0, 4])$ *distribution of Information Theoretic particles.*

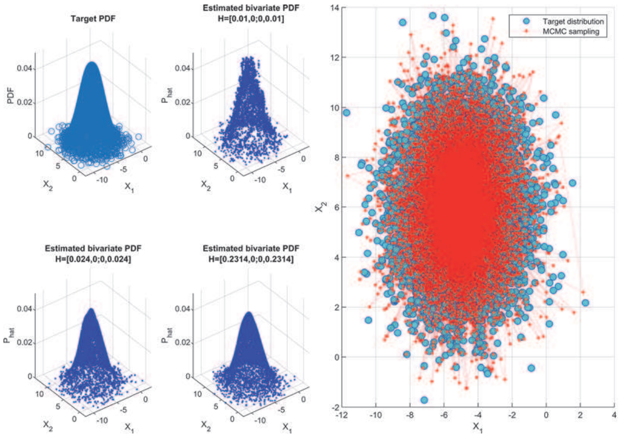

Figure 3.
The left-hand side figure shows the iterative PDF estimate of a bivariate Gaussian distribution $(\mu = [-5, 6], \Sigma = [3, 0; 0, 4])$ *with the MCMC method using the Hamiltonian of Information Theoretic particles. The right-hand side figure shows that the samples generated by the HMC method mostly overlaps the samples of the target distribution.*

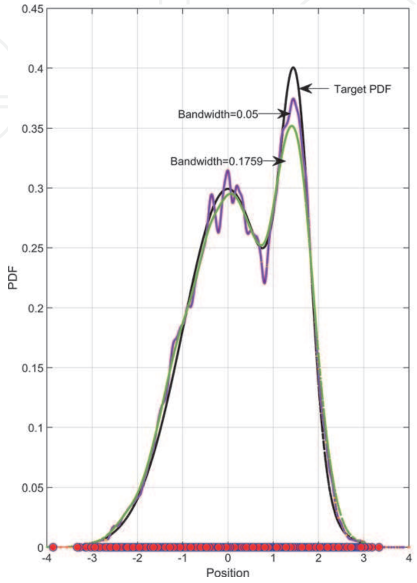

Figure 4.
Iterative estimation of the PDF of a bivariate Gaussian mixture distribution with the MCMC method using the Hamiltonian of Information Theoretic particles.

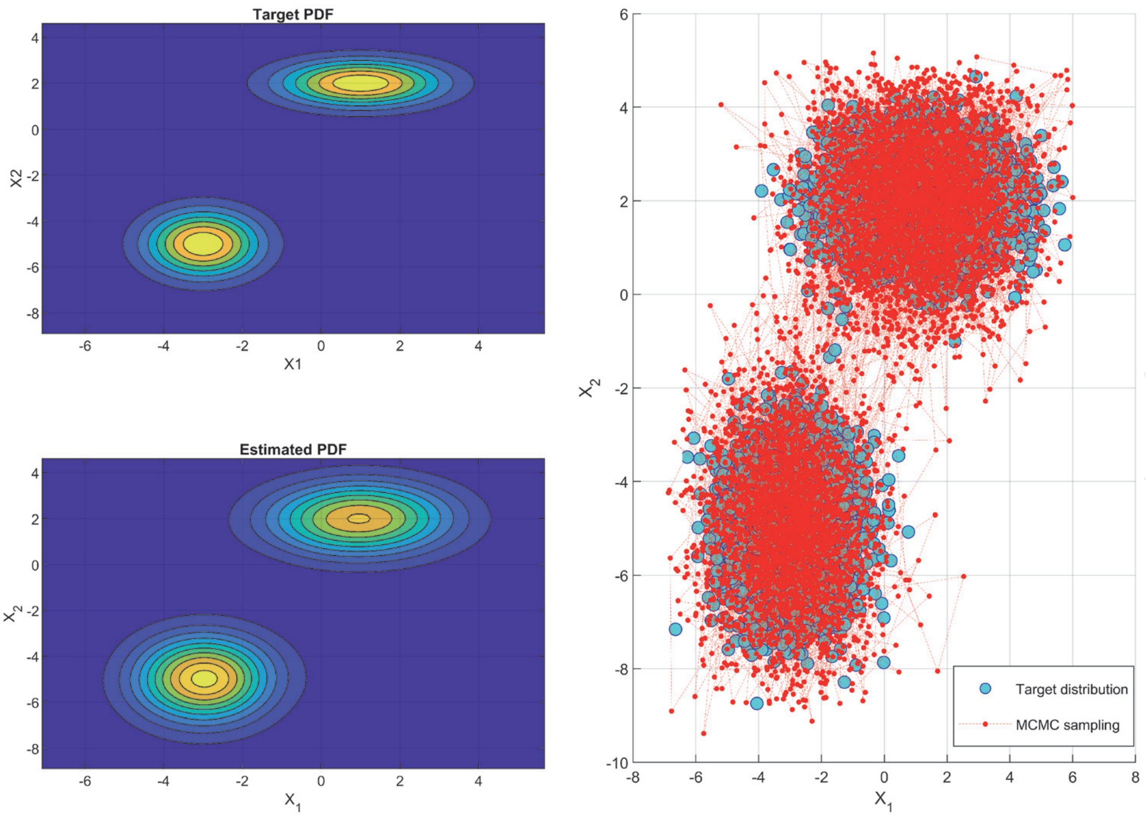

Figure 5.
Contour plots of the target PDF and the estimated PDF of the bivariate Gaussian mixture distribution. Samples generated by the MCMC algorithm using the Hamiltonian of Information Theoretic particles.

Our HMC algorithm using Information Theoretic particles also works well for Gaussian mixture distributions. **Figure 4** shows that our MCMC algorithm using the Hamiltonian of Information Theoretic particles can be used to iteratively estimate the PDF of different multivariate distributions.

Figure 5 shows that the contour plot of the estimated PDF matches closely to the target PDF. The corresponding samples generated by the HMC algorithm traverses the two clusters of the bivariate Gaussian mixture distribution and covers most of the samples of the target distribution.

8. Conclusion

We have proposed a novel perspective on the MCMC method where we used it to iteratively estimate the PDF of a given target sample distribution. We have shown that the samples of a probability distribution can be viewed as Information Particles in an Information Field. These particles have Information Potential energy and are subject to Information Forces by virtue of their position in the field. The concept of Information Potential energy fits perfectly within the framework of the Hamiltonian of a dynamical system. We have derived an important result that the gradient of the potential energy of the Hamiltonian system of Information Particles is just the Information Force estimate normalized by the Information Potential energy estimate.

Our simulation results show that our intuition of comparing Rényi's Quadratic Entropy equation with the Hamiltonian potential energy equation to derive the equation for the potential gradient of a dynamical system of Information Theoretic particles was correct. Using this equation, we were able to accurately estimate univariate and multivariate PDFs. Based on the fixed- or invariant-point theorem, we also derived an equation to iteratively update the bandwidth parameter of the Information Potential and Information Force estimators.

In machine learning applications the dataset is sometimes resampled to the appropriate size before starting the learning operation. Our algorithm can be used to view the data samples as Information Theoretic particles and resample it using the HMC described in this chapter.

Author details

Tokunbo Ogunfunmi* and Manas Deb
Signal Processing Research Lab (SPRL), Department of Electrical and Computer Engineering, Santa Clara University, CA, USA

*Address all correspondence to: togunfunmi@scu.edu

References

[1] R. Neal, “Probabilistic Inference using Markov Chain Monte Carlo methods, Technical Report CRG-TR-93-1,” Department of Computer Science, University of Toronto, Toronto, 1993.

[2] R. Neal, “An improved acceptance procedure for the hybrid Monte Carlo algorithm, “*Journal of Computational Physics,* vol. 111, no. 1, pp. 194–203, 1994.

[3] M. Betancourt, “A Conceptual Introduction to Hamiltonian Monte Carlo, “arXiv: 1701.02434 [stat.ME], 2018.

[4] R. Neal, “MCMC using Hamiltonian Dynamics, “in *Handbook of Markov Chain Monte Carlo*, CRC Press, 2011, pp. 113–162.

[5] J. Principe, Information Theoretic Learning, New York: Springer, 2010.

[6] W. K. Hastings, “Monte Carlo Sampling Methods Using Markov Chains and Their Applications, “*Biometrika,* vol. 57, no. 1, pp. 97–109, 1970.

[7] A. Rényi, “On measures of entropy and information, “*Proceedings of the 4th Berkeley symposium on math, statistics and probability,* vol. 1, pp. 547–561, 1961.

[8] Z. Makó and Z. P. M. D. 7. 4. & Páles, “On the equality of generalized quasiarithmetic means.,” *Publicationes Mathematicae, Debrecen,* vol. 72, pp. 407–440, 2008.

[9] A. N. Kolmogorov, “Sur la notion de la moyenne,” *Atti Accad. Naz. Lincei. Rend. 12:9,* pp. 388–391, 1930.

[10] M. Nagumo, “Über eine klasse von mittelwerte,” *Japanese Journal of Mathematics,* vol. 7, pp. 71–79, 1930.

[11] G. H. Hardy, L. J. E. and P. G., Inequalities, Cambridge, 1934.

[12] R. V. L. Hartley, “Transmission of Information,” *Bell System Technical Journal,* vol. 7, p. 535, 1928.

[13] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal,* p. 535, 1928.

[14] M. Deb and T. Ogunfunmi, “Using Information Theoretic Learning techniques to train neural networks,” in *51st Asilomar conference on signals, systems and computers*, 2017.

[15] E. Parzen, “On Estimation of a Probability Density Function and Mode,” *Annals of Mathematical Statistics,* vol. 33, no. 3, pp. 1065–1076, 1962.

[16] M. Rosenblatt, “Remarks on some Nonparametric Estimates of a Density Function,” *Annals of Mathematical Statistics,* vol. 27, no. 3, pp. 832–837, 1956.

[17] T. Cover and J. Thomas, Elements of Information Theory, Wiley & Sons, 2012.

[18] J. M. Leiva-Murillo and A. Artés-Rodriguez, “Fixed point algorithm for finding the optimal covariance matrix in kernel density modeling,” *IEEE International Conference on Acoustics, Speech and Signal Procesing,* vol. 5, 2006.