
The Theory and Algorithm of Ergodic Inference

Yichuan Zhang¹

Abstract

Approximate inference algorithm is one of the fundamental research fields in machine learning. The two dominant theoretical inference frameworks in machine learning are variational inference (VI) and Markov chain Monte Carlo (MCMC). However, because of the fundamental limitation in the theory, it is very challenging to improve existing VI and MCMC methods on both the computational scalability and statistical efficiency. To overcome this obstacle, we propose a new theoretical inference framework called ergodic Inference based on the fundamental property of ergodic transformations. The key contribution of this work is to establish the theoretical foundation of ergodic inference for the development of practical algorithms in future work.

1. Introduction

Statistical inference is the cornerstone of probabilistic modelling in machine learning. The research on inference algorithms always attracts a great attention in the research community, because it is the fundamentally important in the computation of Bayesian inference, deep generative models. The majority of research is focused on algorithmic development in two theoretical frameworks: variational inference (VI) and Markov chain Monte Carlo (MCMC). These two methods are significantly different. VI is an optimisation-based approach, in particular, which fits a simple distribution to a given target. In contrast, MCMC is a simulation-based approach, which sequentially generates asymptotically unbiased samples of arbitrary target.

Unfortunately, both VI and MCMC suffer from fundamental limitations. VI methods are in general biased because the density function of approximate distribution must be in closed-form. MCMC methods are also biased in practice because the Markov property limits the sample simulation in a local sample space close to previous samples. However,

¹Department of Engineering, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Yichuan Zhang <yichuan.zhang@eng.cam.ac.uk>.

VI is in general more scalable in computation. Optimising variational distribution and simulating samples in VI are computationally efficient and can be accelerated by parallelization on GPU. In contrast, simulating Markov chains is computationally inefficient and, more importantly, asynchronous parallel simulation of multiple Markov chains has no effect on reducing sample correlations but multiplies the computation.

Ergodic Measure preserving flow (EMPF), introduced by (Zhang et al., 2018), is a recent novel optimisation-based inference method that overcomes the limitations of both MCMC and VI. However, there is no theoretical proof of the validity of EMPF. In this work, we will generalize EMPF to a novel inference framework called ergodic inference. In particular, the purpose of this work is to establish the theoretical foundation of ergodic inference. We list the key contribution of this work as following

- The mathematical foundation of ergodic inference. (Section 3 and 4)
- A tractable loss of ergodic inference and the proof of the validity of the loss. (Section 5)
- An ergodic inference model: deep ergodic inference networks (Section 6)
- Clarification of differences between ergodic inference, MCMC and VI (Section 6)

2. The background

Convergence of probability measures is the foundation of statistical inference. Distance metric between probability measures is critical in the study of convergence. We will review the basics of distance metrics between probability measures and connect these metrics to theoretical foundation of inference methods.

2.1. Distance Metric of Probability Measures

Total variation distance is fundamentally important in probability theory, because it defines the strongest convergence of probability measure. Let (Ω, \mathcal{F}) be a measure space, where Ω denotes the sample space and \mathcal{F} denotes the collection of measurable subsets of Ω . Given two probability

measure P and Q defined on (Ω, \mathcal{F}) , the TV distance between Q and P is defined as

$$D_{\text{TV}}(Q, P) = \sup_{A \in \mathcal{F}} |Q(A) - P(A)|. \quad (1)$$

Convergence in TV, that is $D_{\text{TV}}(Q, P) = 0$, means Q and P cannot be distinguished on any measurable set.

The Kullback-Leibler (KL) divergence is an important measure of difference between probability measures in statistical methods. For a continuous sample space Ω , the KL divergence is defined as

$$D_{\text{KL}}(Q||P) = \int_{\Omega} dQ \log \frac{dQ}{dP}, \quad (2)$$

where dP denote the density of probability measure.

2.2. Approximate Monte Carlo Inference

Monte Carlo method is the most popular simulation based inference technique in probabilistic modelling. For example, to fit a probabilistic model π by maximum likelihood estimation, it is essential to compute the gradient of the partition function $Z(\theta) = \int \pi^*(z) dz$. Given the unnormalised density function $\log \pi^*(z)$, computing the gradient becomes a problem of expectation estimation

$$\partial_{\theta} Z(\theta) = \mathbf{E}_{\pi(z)}[\partial_{\theta} \log \pi^*(z)].$$

Monte Carlo methods allow us to construct unbiased estimator of expectation as

$$\mathbf{E}_{\pi(z)}[f(z)] = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N f(z_i),$$

where z_i denotes samples from π . Unfortunately, it is intractable to generate samples from complex distributions, like the posterior distributions in model parameters or latent variables. Because of this challenge, approximate Monte Carlo Inference is fundamentally important. We will review the theoretical foundation of two important inference methods: variational Inference (VI) and Markov chain Monte Carlo (MCMC) in the next two sections.

2.3. Variational Inference

The theoretic foundation of VI is Pinsker's inequality. Pinsker's inequality states that the KL divergence is a upper bound of TV distance

$$D_{\text{TV}}(Q, P) \leq D_{\text{KL}}(Q||P). \quad (3)$$

Given a parametric distribution Q and the target distribution π , minimising the KL divergence $D_{\text{KL}}(Q||\pi)$ implies the less TV distance $D_{\text{TV}}(Q, \pi)$. The key challenge of VI is how to construct the parametric family \mathcal{Q} so that

the estimation of the KL divergence is tractable and family \mathcal{Q} is expressive to approximate complex target. This forces most VI methods to choose Q with closed-form density function. Otherwise, the estimation of entropy term $H(Q) = - \int Q(dz) \log q(z)$ becomes challenging. In practice, the approximation family \mathcal{Q} in most VI methods are rather simple, like Gaussian distribution, so the approximation bias due to oversimplified Q is the key issue of VI.

However, simple approximate family gives VI methods great computational advantage in practice. First, the main loss function in VI is known as the evidence lower bound (ELBO)

$$L_{\text{ELBO}} = \int_{\Omega} dQ \log \frac{d\pi^*}{dQ} \leq \log \int d\pi^*. \quad (4)$$

With analytic form of the entropy of Q , ELBO can be efficiently computed and optimized using standard gradient descent algorithm. Second, simulating i.i.d. samples from a simple variational family Q is straightforward and very efficient.

2.4. Markov Chain Monte Carlo

The theoretical foundation of Markov chain Monte Carlo (MCMC) is ergodic theorem. Ergodic theorem states that, given an ergodic Markov chain (Z_n) with a stationary distribution π , the average cross states of chain is equivalent to the average in state space of the chain, that is

$$\mathbf{E}_{\pi}[f] = \lim_{m \rightarrow \infty} \frac{1}{M} f(Z_{\infty}^m) = \lim_{n \rightarrow \infty} \frac{1}{N} f(Z_n),$$

where Z_{∞}^m denotes the sample of a well-mixed Markov chains after infinitely long transitions. Ergodic theorem implies that we can generate unbiased samples from every Markov transition without waiting forever for the chains to reach stationary state. Therefore, we can trade computational efficiency with a bias that may decrease in a long time. The key challenge of MCMC methods is to define ergodic Markov chains with any given stationary distribution π . This challenge was solved first by Metropolis-Hastings algorithm. We will discuss in detail in Section 4.2.

Ergodic Markov chains enjoy strong stability. Irrespective of the distribution of initial state $\mu(z_0)$ and the parameter of Markov kernel $K(\cdot, \cdot)$, the distribution of the state of the chain is guaranteed to converge to the stationary distribution in total variation after every transition. Formally, that means the reduce of TV distance to stationary for all $L \geq 0$

$$D_{\text{TV}}(Q_{L+1}, \pi) < D_{\text{TV}}(Q_L, \pi)$$

where q_L denotes the marginal distribution of the L -th state and

$$q_L(dz') = \int K(z, dz') q_L(dz).$$

As L increases, the distribution q_L converges to a unique stationary distribution π

$$\lim_{l \rightarrow \infty} D_{\text{TV}}(Q_l, \pi) = 0.$$

In spite of the theoretical convergence property, the convergence of MCMC chains is not guaranteed in practice. Because the burn-in stage cannot be infinite long, the samples from MCMC methods are often biased. The problem is that there is no reliable measurement of such a sampling bias related to TV distance or KL divergence. The iterative simulation of Markov chain is another limitation in computational efficiency. Each sample from MCMC methods requires one simulation of Markov transition and this can only be executed in a sequential manner due to the nature of Markov chain. Therefore, the sampling time of MCMC grows linearly with the number of samples.

3. Ergodic Inference Principle

In this section, we present the mathematical foundation of ergodic inference principle.

3.1. Motivation

First, we would like to propose the the following properties of ideal inference method:

- Parallelizable: the simulation of each sample is computationally independent;
- Statistically efficient: there is zero correlation between samples;
- Asymptotic unbiased: more computational power guarantees diminishing of simulation bias. The bias can be eliminated in theory with sufficient computation.

Both MCMC and VI fail to have all the properties above. For this reason, there are existing works on a hybrid methods that combine MCMC and VI, for example, accelerate the burn-in of MCMC using variational approximation in (Hoffman, 2017) or optimise ELBO based on tractable density function of MCMC kernel in (Salimans et al., 2015). To some extend, such algorithmic hybrid approach can be useful in practice. However, the limitation in theoretical foundation of MCMC and VI cannot be eliminated by algorithmic modification. To achieve an ideal inference method, it is necessary to have a new mathematical theoretical foundation.

3.2. The Theoretical Foundation

Different from Pinsker's inequality and ergodic theorem, the theoretical motivation of the proposed inference is the

characteristic property of ergodic Markov transition: *there is a unique invariant distribution for every ergodic Markov Kernel*. Formally, let K_π be an ergodic Markov transition kernel with an invariant distribution π . By construction of K_π , π is guaranteed to be the only distribution satisfies the condition $\pi(dz') = \int K_\pi(z, dz')\pi(dz)$.

Based on the property of ergodic Markov kernel, we construct the following criteria to verify if a distribution is equivalent to the stationary distribution of the kernel. Given a distribution q , the distribution of q after one Markov transition by K is given by

$$q_1(z') = \int K_\pi(z, z')q(dz). \quad (5)$$

We say the distribution q is preserved by K_π if

$$D_{\text{TV}}(q_1, q) = 0. \quad (6)$$

By the uniqueness of the invariant distribution of ergodic kernel K_π , the preservation of q by K_π as (6) implies $D_{\text{TV}}(q, \pi) = 0$. This motivates the following loss function.

Definition 3.1. Given a Markov kernel $K_\pi(z, z')$ that is ergodic w.r.t. a distribution π , the ergodic loss of a distribution q is defined as

$$L^*(q, K_\pi) = D_{\text{TV}} \left(\int K_\pi(z, \cdot)q(dz), q(\cdot) \right).$$

As mentioned earlier, the loss $L^*(q, K_\pi)$ is equal to 0 if and only if $D_{\text{TV}}(q, \pi)$ is equal to 0.

Let π be the target distribution and q be the approximate distribution in a parametric family \mathcal{Q} . Given an ergodic Markov kernel K_π , the closest $q \in \mathcal{Q}$ to the target π can be identified by the parameter ϕ^* optimising the ergodic loss $L^*(\cdot, K_\pi)$

$$\phi^* = \arg \min_{\phi} L^*(q_\phi, K_\pi).$$

If the target distribution is in \mathcal{Q} , then the optimal parameter ϕ^* should have the loss

$$L^*(q_{\phi^*}, K_\pi) = 0,$$

otherwise the L^2 norm of the gradient of the loss should vanish

$$\|\partial_{\phi^*} L^*(q_{\phi^*}, K_\pi)\|_2^2 = 0.$$

3.3. Technical Challenges

There are two technical challenges of ergodic inference methods in practice. First, we need a tractable estimation of a loss function equivalent to $D_{\text{TV}}(q_1, q)$. The estimation of the gradient of the loss should also be tractable for the optimisation of the parameter ϕ . Second, we need a general

parametric family \mathcal{Q} that can approximate any target distribution up to a certain amount of error. More specific, the error can be controlled and even eliminated by increase the complexity of approximation family of \mathcal{Q} , i.e. the number of parameters of \mathcal{Q} is unlimited. The computational cost of optimisation is associated with the complexity of \mathcal{Q} .

We will present the solution to the first challenge in Section 5 and the solution to the second challenge in Section 6.

4. Ergodic Transformations

The key of solving the technical challenges in ergodic inference is the reparameterization of the ergodic Markov kernel. This is important in both algorithmic development and theoretical analysis.

4.1. Ergodic Transformations and Markov Kernels

Ergodic Markov kernels are essentially conditional distributions, which can be reparameterized by deterministic transformations known as measure preserving transformations (MPTs). Given a probability measure μ , a deterministic transformation T preserves μ if for any measurable subset of sample space A , $\mu(T^{-1}(A)) = \mu(A)$. The shear transformation $T(x, y) = (x + y, x)$, which preserves Lebesgue measure, is a classic example of MPT (Billingsley, 1986). The following conditions are often used in the literature MCMC theory for verification of ergodic property:

1. Irreducibility: $T(A) \neq A, \forall A \in \mathcal{F}$ except \emptyset and Ω .
2. Density preservation: $\pi(T(\mathbf{z})) = \pi(\mathbf{z})$.
3. Lebesgue preservation: the determinant of the Jacobian of T is equal to 1.

Formally, we define the reparameterisation of Ergodic Markov chains as following.

Definition 4.1. (Ergodic Reparameterisation of MCMC) Given a target distribution $\pi(\mathbf{z})$, a MCMC kernel $K(\mathbf{z}, \mathbf{z}')$ with invariant π can be reformed as two steps:

1. Simulate an auxiliary variable \mathbf{r} with distribution $\mu(\mathbf{r})$
2. Deterministic transformation $(\mathbf{z}', \mathbf{r}') = T_{\pi\mu}(\mathbf{z}, \mathbf{r})$,

where $T_{\pi\mu}$ is an ergodic transformation that preserves the probability measure $\pi(\mathbf{z})\mu(\mathbf{r})$.

Remark. The transformation $T_{\pi\mu}$ in ergodic reparameterisation is fundamentally different from volume preserving transformation $V(\mathbf{z})$ in the sample space of \mathbf{z} for two reasons.

- $T_{\pi\mu}(\mathbf{z}, \mathbf{r})$ does not preserve the volume/entropy in the sample space of \mathbf{z} , but $V(\mathbf{z})$ must preserves the volume/entropy in the space of \mathbf{z} .

- $T_{\pi\mu}(\mathbf{z}, \mathbf{r})$ preserves the probability measure $\pi(\mathbf{z})$, but $V(\mathbf{z})$ does not preserve $\pi(\mathbf{z})$ in general.

Ergodic transformations also allow us to form the expectation under Markov transition as composition of functions, that is not used in classic MCMC literature. Formally, this is given by the following proposition.

Proposition 1. Given an ergodic transformation T_π w.r.t. π , the expectation is preserved by the transformation, which means, for any function f

$$\int_{\Omega} f(\mathbf{z})\pi(d\mathbf{z}) = \int_{\Omega} f \circ T_\pi(\mathbf{z})\pi(d\mathbf{z}) = \int_{\Omega'} f(\mathbf{z}')T_{\pi*}\pi(d\mathbf{z}'),$$

where Ω' is the image of Ω under T_π and $T_{\pi*}\pi(\cdot)$ denotes the pushforward probability measure of π under T_π . Because T_π preserves π , $\Omega' = \Omega$ and $D_{TV}(\pi, T_{\pi*}\pi) = 0$.

In the next two sections, we will demonstrate the ergodic reparameterization with two well-known MCMC kernels.

4.2. Metropolis-Hastings Transformations

Metropolis-Hastings (MH) algorithm is the first and most well-known MCMC methods. We will show that it is straightforward to form the MH transition kernel as an ergodic transformation. Given a target distribution $\pi(\mathbf{z})$ and a transition proposal distribution $q(\mathbf{r}|\mathbf{z})$, MH kernel in most text books is described as following two steps:

1. Sample \mathbf{r} from $q(\cdot|\mathbf{z})$.
2. Return the new state of the chain as \mathbf{r} with probability

$$p_{MH} = \min \left\{ 1, \frac{\pi(\mathbf{r})q(\mathbf{z}|\mathbf{r})}{\pi(\mathbf{z})q(\mathbf{r}|\mathbf{z})} \right\}, \quad (7)$$

otherwise the state remains as \mathbf{z} .

It is straightforward to verify that MH transition kernel preserves the density function as

$$\begin{aligned} & \pi(\mathbf{z}) \left[q(\mathbf{r}|\mathbf{z}) \min \left\{ 1, \frac{\pi(\mathbf{r})q(\mathbf{z}|\mathbf{r})}{\pi(\mathbf{z})q(\mathbf{r}|\mathbf{z})} \right\} \right] \\ &= \min \{ \pi(\mathbf{z})q(\mathbf{r}|\mathbf{z}), \pi(\mathbf{r})q(\mathbf{z}|\mathbf{r}) \} \\ &= \pi(\mathbf{r}) \left[q(\mathbf{z}|\mathbf{r}) \min \left\{ 1, \frac{\pi(\mathbf{z})q(\mathbf{r}|\mathbf{z})}{\pi(\mathbf{r})q(\mathbf{z}|\mathbf{r})} \right\} \right], \end{aligned}$$

where the MH transition kernel $K_{MH}(\cdot, \cdot)$ is in squared brackets. This verification of stationary distribution is known as detailed balance. It is important because it proves the existence of stationary distribution.

Now we consider an alternative representation of MH kernel. In particular, we define a stationary distribution as the joint distribution of all random variables involved in

the target π and MH kernel K_{MH} , that is $\pi(\mathbf{z}, \mathbf{r}, u) = \pi(\mathbf{z})q(\mathbf{r}|\mathbf{z})\nu(u)$, where $\nu(u)$ denotes uniform distribution between $[0, 1]$. Following the ergodic reparameterization (Definition 4.1), we can rewrite the MH algorithm as

1. Resample \mathbf{r} from $q(\cdot|\mathbf{z})$ and u from $\nu(\cdot)$.
2. Return the next state $(\mathbf{z}', \mathbf{r}', u') = T_{MH}(\mathbf{z}, \mathbf{r}, u)$ defined as

$$T_{MH}(\mathbf{z}, \mathbf{r}, u) = (\mathbf{z}, \mathbf{r}, u)\delta(u > p_{MH}) + (\mathbf{r}, \mathbf{z}, u)\delta(u < p_{MH}), \quad (8)$$

where $\delta(\cdot)$ denotes indicator function.

Notice that the transformation $T_{MH}(\mathbf{z}, \mathbf{r}, u)$ above is a deterministic function. It is obvious that resampling \mathbf{r} and u from their conditional distribution leaves $\pi(\mathbf{z}, \mathbf{r}, u)$ invariant. Then, it is straightforward to show the preservation of density function

$$\pi(\mathbf{s})\delta(\mathbf{s}' = T_{MH}(\mathbf{s})) = \pi(\mathbf{s}')\delta(\mathbf{s} = T_{MH}(\mathbf{s}')),$$

where \mathbf{s} denote the triple $(\mathbf{z}, \mathbf{r}, u)$. It is also easy to verify that the determinate of Jacobian of $\partial_{(\mathbf{z}, \mathbf{r}, u)}T_{MH}(\mathbf{z}, \mathbf{r}, u)$ is always equal to 1.

4.3. Hamiltonian Measure Preserving Transformations

Hamiltonian Monte Carlo (HMC), originally known as Hybrid Monte Carlo, is an important MCMC method. Originally, HMC is considered as a hybrid method, because its combines both deterministic and stochastic simulation. The deterministic simulation in HMC essentially refers to any dynamics that generalize the classic Hamiltonian dynamics in physics.

Hamiltonian system in physics is a system of moving particles in an energy field and the energy of the system is constant over time. Given n particles, the state of Hamiltonian system is defined by the position $\mathbf{z} \in \mathbb{R}^n$ and the momenta $\mathbf{r} \in \mathbb{R}^n$. The position \mathbf{z} is associated with potential energy $U : \mathbb{R}^n \rightarrow \mathbb{R}$ and the momentum \mathbf{r} is associated with kinetic energy $K : \mathbb{R}^n \rightarrow \mathbb{R}$. The state (\mathbf{z}, \mathbf{r}) evolves over time t , according to Hamiltons equations:

$$\dot{\mathbf{z}}(t) = \partial_{\mathbf{r}}K(\mathbf{r}); \dot{\mathbf{r}}(t) = -\partial_{\mathbf{z}}U(\mathbf{z}), \quad (9)$$

where $\dot{\mathbf{z}}$ denotes the derivative of \mathbf{z} w.r.t. time t . It is straightforward to verify that the total energy $H = U + K$ does not change over time

$$\dot{H}(\mathbf{z}, \mathbf{r}) = (\partial_{\mathbf{r}}U(\mathbf{z}))^T \partial_{\mathbf{r}}K(\mathbf{r}) - (\partial_{\mathbf{z}}U(\mathbf{z}))^T \partial_{\mathbf{z}}K(\mathbf{r}) = 0.$$

Given an initial condition (\mathbf{z}, \mathbf{r}) , the solution of Hamiltonian dynamics is a function of time t

$$(\mathbf{z}(t), \mathbf{r}(t)) = T_H(t, \mathbf{z}, \mathbf{r}).$$

Given a fixed time t , the solution T_H becomes a map $T_{H,t} : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ between two states (\mathbf{z}, \mathbf{r}) and $(\mathbf{z}', \mathbf{r}')$ with the same total energy H . Intuitively, $\mathbf{z}(t)$ forms a trajectory of particle traversing in a n -dimensional space and the velocity of the particle is given by $\dot{\mathbf{z}}(t) = \partial_{\mathbf{r}}K(\mathbf{r}(t))$.

It is well-known in MCMC literature that $T_{H,t}$ is essentially a family of measure preserving transformations with any parameter $t \in \mathbb{R} \neq 0$. It is clear that $T_{H,t}$ is irreducible if $t \neq 0$ and density preserving w.r.t. $\exp(-H)$. The volume preservation property of Hamiltonian dynamics in the state space (\mathbf{z}, \mathbf{r}) is a well-known result of Liouville's Theorem. Therefore, we know that $T_{H,t}(\mathbf{z}, \mathbf{r})$ with any $t \neq 0$ is an ergodic transformation w.r.t. the distribution $\pi(\mathbf{z})\mu(\mathbf{r}) \propto \exp(-H(\mathbf{z}, \mathbf{r}))$. This implies $T_{H,t}$ also preserves $\pi \propto \exp(-U)$ by the definition of marginal distribution.

In practice, Hamiltonian dynamics do not have closed-form solutions. Fortunately, there is a rich literature on the numeric simulation of Hamiltonian dynamics. The most known approximate approach in HMC is Leapfrog algorithm, which is constructed as a sequential of shear transformations. Leapfrog algorithm enjoys strong stability and good approximation error is around squared discretized step size. See more detailed analysis in (Neal, 2010; Leimkuhler & Reich, 2004).

5. Ergodic Loss

5.1. π -Ergodic Loss Function

By the definition of TV distance, we know that q is the stationary distribution of K if and only if for all function $f(\cdot)$ with $\mathbf{E}_\pi[f(\mathbf{z})] < \infty$,

$$\mathbf{E}_{q_1}[f(\mathbf{z})] = \mathbf{E}_q[f(\mathbf{z})]. \quad (10)$$

However, it is impossible to compare the expectation of all possible function f , but given specific function f it is possible to estimate

$$L_{K,f}(\phi) = |\mathbf{E}_{q_1}[f(\mathbf{z})] - \mathbf{E}_q[f(\mathbf{z})]|. \quad (11)$$

With the optimal choice of function f and certain condition, we can claim that $L_{K,f}(\phi) = 0$ implies $D_{TV}(q, \pi) = 0$. The log density function is an intuitive choice, because we can identify a distribution by its density function. Therefore, we define the following π -ergodic loss.

Definition 5.1. (Ergodic Loss Function)

$$L_{K,\pi}(\phi) = |\mathbf{E}_{q_1}[\log \pi(\mathbf{z})] - \mathbf{E}_q[\log \pi(\mathbf{z})]|. \quad (12)$$

Theorem 1. (Ergodic Loss Convergence Theorem) Given the ergodic Markov kernel K_π with invariant distribution π , the loss $L_{K,\pi}(\phi) = 0$ if and only if $\mathbf{E}_\pi[\log \pi(\mathbf{z})] = \mathbf{E}_q[\log \pi(\mathbf{z})]$.

Proof. The convergence of loss $L_{K,\pi}(\phi) = 0$ implies

$$\mathbf{E}_{q_1(\mathbf{z})\mu(\mathbf{r})}[\log \pi(\mathbf{z})] = \mathbf{E}_{q(\mathbf{z})\mu(\mathbf{r})}[\log \pi(\mathbf{z})], \quad (13)$$

where $q_1(\mathbf{z})$ is given by (5). Notice that q_1 is essentially the marginal of the pushforward of $q(\mathbf{z})\mu(\mathbf{r})$ under the measure preserving transformation $T_{\pi\mu}$. By Proposition 1, the expectations in (13) can be written as following

$$\mathbf{E}_{q_1(\mathbf{z})}[\log \pi(\mathbf{z})] \triangleq \int_{\Omega} \log \pi \circ T_{\pi\mu} d(q\mu) = \int_{\Omega} \log \pi d(q\mu), \quad (14)$$

where $d(q\mu)$ is the shorthand notations for $q(\mathbf{z})\mu(\mathbf{r})d\mathbf{z}d\mathbf{r}$. Replacing $q\mu$ on both sides in (14) with any distribution, the equality still holds. If we replace $q\mu$ in (14) with the pushforward probability measure of $q(\mathbf{z})\mu(\mathbf{r})$ under $T_{\pi\mu}$, denoted by $T_{\pi\mu*}(q\mu)$, we have

$$\int_{\Omega} \log \pi \circ T_{\pi\mu} \circ d(T_{\pi\mu*}(q\mu)) = \int_{\Omega} \log \pi \circ d(T_{\pi\mu*}(q\mu)),$$

which can be rewritten as

$$\int_{\Omega} \log \pi \circ T_{\pi\mu}^1 \circ T_{\pi\mu} d(q\mu^1) = \int_{\Omega} \log \pi \circ T_{\pi\mu} d(q\mu), \quad (15)$$

where $T_{\pi\mu}^1$ denotes $T_{\pi\mu} = (\mathbf{z}, \mathbf{r}_1)$ and $d\mu^1$ denotes $\mu(d\mathbf{r}_1)$. Notice that the LHS of (15) is an expectation under the distribution of \mathbf{z} after two ergodic Markov transitions from q , that is $\mathbf{E}_{q_2(\mathbf{z})}[\log \pi(\mathbf{z})]$. Therefore, by (14) and (15), we have

$$\begin{aligned} \mathbf{E}_{q_2(\mathbf{z})}[\log \pi(\mathbf{z})] &\triangleq \int_{\Omega} \log \pi \circ T_{\pi\mu}^1 \circ T_{\pi\mu} d(q\mu^1) \\ &= \int_{\Omega} \log \pi \circ T_{\pi\mu} d(q\mu) \\ &= \mathbf{E}_{q(\mathbf{z})}[\log \pi(\mathbf{z})]. \end{aligned} \quad (16)$$

By induction, we know the expectation of $\mathbf{E}_{q_n}[\log \pi]$ does not change after any number of measure preserving transformation $T_{\pi\mu}$, that gives

$$\mathbf{E}_{q_\infty(\mathbf{z})}[\log \pi(\mathbf{z})] = \mathbf{E}_{q(\mathbf{z})}[\log \pi(\mathbf{z})]. \quad (17)$$

By (17), we know if we simulate infinitely long ergodic Markov chain by kernel K_π , then the expectation $\mathbf{E}_{q_\infty(\mathbf{z})}[\log \pi(\mathbf{z})]$ is the same as the initial expectation $\mathbf{E}_{q(\mathbf{z})}[\log \pi(\mathbf{z})]$.

Because an ergodic Markov chain has unique invariant distribution, (17) implies

$$\mathbf{E}_{\pi(\mathbf{z})}[\log \pi(\mathbf{z})] = \mathbf{E}_{q(\mathbf{z})}[\log \pi(\mathbf{z})]. \quad (18)$$

Recall that the convergence of loss $L_{K,\pi^*}(\phi)$ cannot be sufficient for the convergence of the TV distance $D_{\text{TV}}(q, \pi) = 0$. Fortunately, under some reasonable condition, the loss $L_{K,\pi^*}(\phi) = 0$ implies the convergence in TV distance. Formally, this is given by the following theorem.

Theorem 2. (Ergodic Measure Convergence Theorem) Let K_π be an ergodic Markov kernel with invariant distribution π . Assume that the entropy of Q is not less than the entropy of π , that is $H(Q) \geq H(\pi)$, the loss $L_{K,\pi^*}(\phi) = 0$ if and only if $D_{\text{TV}}(q, \pi) = 0$.

Proof. By the definition of the KL divergence, we have

$$D_{\text{KL}}(q||\pi) = \mathbf{E}_q[\log q] - \mathbf{E}_q[\log \pi]. \quad (19)$$

By Theorem 1, we have

$$D_{\text{KL}}(q||\pi) = \mathbf{E}_q[\log q] - \mathbf{E}_\pi[\log \pi], \quad (20)$$

which is equivalent to

$$D_{\text{KL}}(q(\mathbf{z})||\pi) = H(\pi) - H(Q).$$

Because the KL divergence is never less than 0, we have

$$H(\pi) \geq H(Q).$$

Finally, by the assumption $H(\pi) \leq H(Q)$, we know $H(\pi) = H(Q)$, so we know $0 \leq D_{\text{TV}}(q, \pi) \leq D_{\text{KL}}(q||\pi) = 0$, which implies $D_{\text{TV}}(q, \pi) = 0$. \square

By the monotonic convergence in TV distance of ergodic Markov chain, it is straightforward to show that

Proposition 2. Given a smooth ergodic transformations w.r.t. the probability measure $\pi(\mathbf{z})$, if $\mathbf{E}_q[\log \pi] < \mathbf{E}_\pi[\log \pi]$, the loss

$$\mathbf{E}_q[\log \pi^*(\mathbf{z})] - \mathbf{E}_{q_1}[\log \pi^*(\mathbf{z})] > 0. \quad (21)$$

Assume that $\mathbf{E}_q[\log \pi] < \mathbf{E}_\pi[\log \pi]$, we have

$$L_{K,\pi^*}^*(\phi) = \mathbf{E}_q[\log \pi^*(\mathbf{z})] - \mathbf{E}_{q_1}[\log \pi^*(\mathbf{z})], \quad (22)$$

5.2. Optimising π^* -Ergodic Loss

Let $q_{01}(\mathbf{z}, \mathbf{z}_1)$ be the joint distribution $q(\mathbf{z})K(\mathbf{z}, \mathbf{z}_1)$. Then, we can rewrite (22) as

$$L_{K,\pi^*}^*(\phi) = \mathbf{E}_{q_{01}}[\log \pi^*(\mathbf{z}) - \log \pi^*(\mathbf{z}_1)], \quad (23)$$

which can be estimated by samples of $(\mathbf{z}, \mathbf{z}_1)$. To optimise the loss (23), we need to compute the gradient $\partial_\phi L_{K,\pi^*}^*(\phi)$. Notice that the \mathbf{z} and \mathbf{z}_1 are coupled by the kernel K and the density function of most MCMC kernels, which makes the computation of the gradient $\partial_\phi L_{K,\pi^*}^*(\phi)$ unstable. To avoid this, we reparameterize both $q(\cdot)$ and the ergodic

Markov kernel $K(\mathbf{z}, \cdot)$ by a transformation T_ϕ and a measure preserving transformation T_π respectively. This allows us to transform some simple random variable \mathbf{r} and \mathbf{r}_1 , that is independent of ϕ , into $(\mathbf{z}, \mathbf{z}_1)$ as

$$\mathbf{z} = T_\phi(\mathbf{r}), \quad \mathbf{z}_1 = T_\pi(\mathbf{z}, \mathbf{r}_1). \quad (24)$$

Therefore, we can compute the loss with following reformulation

$$L_{K, \pi^*}^*(\phi) = \mathbf{E}_{\mu(\mathbf{r})\mu_1(\mathbf{r}_1)}[L_{\pi^*, T_\phi, T_\pi}(\mathbf{r}, \mathbf{r}_1)], \quad (25)$$

where $L_{\pi^*, T_\phi, T_\pi} = \log \pi^*(\mathbf{z}) - \log \pi^*(\mathbf{z}_1)$ and $(\mathbf{z}, \mathbf{z}_1) = T_{\phi, \pi}$ as (24).

As discussed above, the only requirement of approximate family \mathcal{Q} in ergodic inference is the transformation T_ϕ is known and it is a measurable function. It is an important advantage over VI, where the density function of \mathcal{Q} must be in closed form.

6. Deep Ergodic Inference Model

Ergodic transformations are not only fundamentally important in the ergodic loss, they are also powerful tools for constructing flexible approximation family \mathcal{Q} . In this section, we will present how to construct and optimise the approximation family \mathcal{Q} by stacking multiple layers of ergodic transformations.

6.1. Definition

Let $\{K_1, K_2, \dots, K_N\}$ be N ergodic transition kernel with independent parameters $\{\phi_1, \phi_2, \dots, \phi_N\}$. Let q_0 be the distribution of initial state also has parameter ϕ_0 . By ergodic reparameterization, we reform each ergodic Markov kernel $K_n(\mathbf{z}, \mathbf{z}')$ as a transformation $\mathbf{z}_n = T_n(\mathbf{z}_{n-1}, \mathbf{r})$, where T_n is a deterministic function depends on the kernel parameter ϕ_n and \mathbf{r} is sampled from a standard distribution μ_n . We also reparameterize the initial distribution q from a simple distribution μ_0 by a transformation T_0 . Then, we can generate samples of \mathbf{z}_n by transforming samples of $(\mathbf{r}_0, \mathbf{r}_1, \dots, \mathbf{r}_{N-1})$ from $\mu(\cdot) = \prod_{i=0}^{N-1} \mu_i(\cdot)$ as

$$\mathbf{z}_n = T_{\mathbf{r}_{N-1}} \circ \dots \circ T_{\mathbf{r}_1} \circ T_0(\mathbf{r}_0), \quad (26)$$

where $T_{\mathbf{r}_n}(\cdot)$ denotes $T_n(\cdot, \mathbf{r}_n)$. We call this multiple layer ergodic transformation $T_{\mathbf{r}_{N-1}} \circ T_{\mathbf{r}_1} \circ T_0(\cdot)$ deep ergodic inference network (DEIN). The expectation of q_N can be reformed as

$$\mathbf{E}_{q_N}[f(\mathbf{z}_N)] = \mathbf{E}_\mu[f \circ T_{\mathbf{r}_{N-1}} \circ \dots \circ T_{\mathbf{r}_1} \circ T_0(\mathbf{r}_0)],$$

which allows us to estimate the gradient of any function by Monte Carlo method

$$\partial_\phi \mathbf{E}_{q_N}[f(\mathbf{z}_N)] \approx \frac{1}{M} \sum_{i=1}^M \partial_\phi f \circ T_{\mathbf{r}_{N-1}^i} \circ \dots \circ T_{\mathbf{r}_1^i} \circ T_0(\mathbf{r}_0^i).$$

6.2. Optimisation and Convergence of DEINs

This is a non-parametric model because the number of parameters of this model grows with the number of transformations. Different from deep neural networks, DEIN has strong stability by the nature of ergodicity. In particular, DEINs can be arbitrarily deep and the stability and simulation quality is guaranteed to improve with the depth.

First, we define a loss (12) for each transition K_n as

$$L^n(\phi_n) = \mathbf{E}_{q_n}[\log \pi^*(\mathbf{z})] - \mathbf{E}_{q_{n-1}}[\log \pi^*(\mathbf{z})],$$

where q_N denotes the marginal of the last state

$$q_n(\mathbf{z}; \phi_{0:n}) = \int K(\mathbf{z}_{n-1}, \mathbf{z}_n) q_{n-1}(\mathbf{z}_{n-1}; \phi_{0:n-1}). \quad (27)$$

Proposition 3. Assume that $\mathbf{E}_{q_0}[\log \pi(\mathbf{z})] < \mathbf{E}_\pi[\log \pi(\mathbf{z})]$, minimizing the ergodic loss L_{K, π^*}^* in (22) with q_N of deep ergodic Inference network is equivalent to maximizing the total ergodic loss $\sum_{n=1}^N L^n(\phi_n)$

$$L_N(\phi) = \mathbf{E}_{q_N}[\log \pi^*(\mathbf{z})] - \mathbf{E}_{q_0}[\log \pi^*(\mathbf{z})]. \quad (28)$$

which is equivalent to

$$L_N(\phi; \phi_0) = \mathbf{E}_{q_N}[\log \pi^*(\mathbf{z})]. \quad (29)$$

when the parameter of q_0 is fixed.

The total loss (29) is consistent with the loss proposed by (Zhang et al., 2018) in ergodic measure preserving flows.

By Proposition 2, it is straightforward to show that DEINs enjoy incremental improvement as the depth grows.

Theorem 3. (Incremental Convergence of DEIN) Given a N -layer DEIN defined as (26), the optimal total ergodic loss $L_N(\phi^*) = \max_\phi L_N(\phi)$ increases monotonically as N increases.

Similar to the convergence of ergodic Markov chains, we have the asymptotic unbiased convergence of DEINs as following.

Theorem 4. (Asymptotic Unbiased Convergence of DEINs) For arbitrarily small $\epsilon > 0$, there always exists a DEIN with finite number of layer N , so that with the optimal distribution q_N^* has the ergodic loss $L_{K, \pi}^* = D_{TV}(\int K(\mathbf{z}, \cdot) dq_N^*, q_N^*(\cdot)) \leq \epsilon$.

6.3. Comparison with Auto-Tuning MCMC

From an algorithmic perspective, auto-tuning MCMC (AMCMC) and DEIN are very similar, because both methods simulate ergodic Markov chains and optimise the parameters of the kernel w.r.t. a loss. This may give a false impression of that AMCMC and DEIN share the same theoretical foundation.

To clear this impression, we will discuss the fundamental difference between DEINs and AMCMC. First of all, AMCMC is essentially a class of MCMC methods with auto-tuning strategy of kernel parameters. In particular, the purpose of auto-tuning is to boost the statistical power of samples from MCMC by encouraging distant jump between states in Euclidean space, which is inspired by the work of (Pasarica & Gelman, 2010) on reducing sample correlation of MCMC. In contrast, as a parametric family in ergodic inference methods. The parameters in DEINs is optimised w.r.t. the ergodic loss, which is based on the ergodic inference principle in Section 3.2.

The fundamental difference have two important effects in practice. The first effect is on the sample correlation. By the nature of Markov property, optimising the auto-tuning loss can never eliminate the correlation of samples from MCMC. In contrast, the samples from DEINs are generated by deterministic transformation of i.i.d. samples from initial distribution, which is still i.i.d. samples. The second consequence is on the MH-correction. In particular, MH correction is optional for DEINs for three reasons. First, DEIN is a parametric approximate family \mathcal{Q} rather than unbiased simulation procedure. Second, by optimising the ergodic loss, DEINs guarantee the convergence towards the target in TV distance. Finally, even with approximate ergodic transformations, the existence of a stationary distribution (not necessarily the target) is guaranteed by measure preserving property, in particularly with the depth of DEIN is always finite. In contrast, the convergence of AMCMC chains is only guaranteed with MH correction. In particular, without MH correction, the existence of a stationary distribution of MCMC chains becomes questionable. With unlimited number of recurrent Markov transitions, Markov chains are not guaranteed to converge to any distribution. The existence of stationary distribution is the necessary condition of ergodic theorem (Robert & Casella, 2005). Therefore, without MH-correction (implicitly proved by detailed balance condition), the bias of samples from MCMC may not be bounded. This is particularly true when the Markov kernel parameter is tuned to maximize the jumping distance between states.

6.4. Comparison with Normalising Flows

Normalizing Flow (NF), introduced by (Rezende & Mohamed, 2015), is a recent variational inference framework, where the variational parametric distribution is defined in an iterative procedure. The fundamental idea of NF is to define an expressive parametric family by a sequence of deterministic transformations with closed-form Jacobian. Let \mathbf{z}_0 be a random variable from a simple distribution μ , like Gaussian, and $f_1 \dots, f_M$ be M deterministic functions from \mathbb{R}^n to \mathbb{R}^n . We define a

sequence of random variable $\mathbf{z}_1 \dots \mathbf{z}_M$ as

$$\mathbf{z}_M = f_M \circ \dots \circ f_1(\mathbf{z}_0).$$

By the rule of changing variables, the density function of \mathbf{z}_M is given by

$$\log p(d\mathbf{z}_M) = \log q(d\mathbf{z}_0) - \sum_{i=1} \log |\det \partial_{\mathbf{z}_i} f_i(\mathbf{z}_i)|.$$

There are three important difference between DEINs and NFs. First, without manually engineering ergodic transformations, DEINs have theoretical guarantee of better performance with more transformations (Theorem 4). In contrast, the transformations f_i in NFs is predefined based on heuristics and experimental evidence. Second, ergodic transformations T_π has no closed form solutions, but the transformations f_i in NFs is limited to simple functions with tractable Jacobian. Finally, the distribution of DEINs is very expressive, which may not even have a closed form as (27). More importantly, there is no need to compute the density for optimising the parameters. It is the opposite for NFs. In particular, the transformations in NFs are often restricted to simple functions to have closed-form Jacobian. The computation of the Jacobian is also one of computational bottlenecks in optimisation.

6.5. Comparison Overview

The key difference between ergodic inference, AMCMC and VI is highlighted in the following table.

Method	TV-Loss	Implicit Simulation Density	Independent samples
VI	Yes	No	Yes
AMCMC	No	Yes	No
EI	Yes	Yes	Yes

- TV-Loss: Optimising the loss function leads to the convergence in TV distance.
- Independent samples: computationally and statistically independent sample simulation.
- Implicit Simulation Density: no closed-form density function of simulation distribution is required in training.

7. Related Works

Hamiltonian variational inference (HVI), introduced by (Salimans et al., 2015), is an interesting variational framework using MCMC kernel as variational parametric distribution. The motivation of HVI is that the joint density function of all the states of HMC chains is tractable to compute. Unfortunately, the variational lower bound is still

intractable to compute, because the reverse probability of HMC chain given the last state is intractable. To overcome this problem, they propose to approximate the reverse density function using neural network. Although HVI shows improvement in performance over VAEs, the additional approximation limits the potential of this method. However, optimising the HMC kernel parameters w.r.t. ELBO is still an attractive feature of HVI.

Hoffman (Hoffman, 2017) proposed another hybrid method based on VI and HMC without auxiliary approximation. The idea is to use a Monte Carlo estimation of the marginal likelihood by averaging over samples from HMC chains, that are initialized by variational distribution. In (Han et al., 2017) a very similar framework is proposed using Metropolis-adjusted Langevin dynamics. This idea is very similar to contrastive divergence in (Hinton, 2002). The main disadvantage of this methods is that the HMC parameters are manually pretuned. Especially, As mentioned by (Hoffman, 2017), No-U-turn Sampler (NUTS), an adaptive HMC, is not applicable due to engineering difficulties. (Neal, 2010) pointed out that HMC is very sensitive to the choice of Leapfrog step size and number of leaps.

Stein Variational Gradient Descent (SVGD) is a recent particle based dynamical inference method proposed by (Liu, 2017). In SVGD, the approximation distribution is a set point mass q generated by transforming a set of points sampled from a distribution μ using a perturbation function $T(x) = x + \phi(x)$, where ϕ is in a function space with boundary norm. With this setup, the optimisation of T w.r.t. the KL divergence between q and the target π is transformed into a stochastic optimisation in the kernel space of ϕ . The theoretical foundation of convergence of SVDG is sound and appealing. However, this method faces two practical challenges. First, the optimisation complexity grows quadratically with the number of particles. Second, it is very difficult to approximate high dimensional distribution well with a limited number of point mass approximation.

8. Summary

I proposed a new generic inference method based on optimization and ergodic deterministic transformations. This work provides us the very foundation of ergodic inference including: the fundamental ergodic inference principle; tractable estimation of ergodic loss and the its gradient; a generic construction of approximation family.

References

Billingsley, P. *Probability and Measure*. John Wiley and Sons, third edition, 1986.

Han, T., Lu, Y., Zhu, S.-C., and Wu, Y. N. Alternating Back-

Propagation for Generator Network. In *AAAI*, volume 3, pp. 13, 2017.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2014/09/08 2002. doi: 10.1162/089976602760128018. URL <http://dx.doi.org/10.1162/089976602760128018>.

Hoffman, M. D. Learning Deep Latent Gaussian Models with Markov Chain Monte Carlo. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1510–1519. PMLR, 2017.

Leimkuhler, B. and Reich, S. *Simulating Hamiltonian Dynamics*, volume 14. Cambridge university press, 2004.

Liu, Q. Stein variational gradient descent as gradient flow. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3115–3123. Curran Associates, Inc., 2017.

Neal, R. M. *MCMC using Hamiltonian Dynamics*. 2010.

Pasarica, C. and Gelman, A. Adaptively scaling the metropolis algorithm using expected squared jumped distance. *Statistica Sinica*, pp. 343–364, 2010.

Rezende, D. J. and Mohamed, S. Variational Inference with Normalizing Flows. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML’15, pp. 1530–1538, 2015.

Robert, C. P. and Casella, G. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387212396.

Salimans, T., Kingma, D., and Welling, M. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In *International Conference on Machine Learning*, pp. 1218–1226, 2015.

Zhang, Y., Hernández-Lobato, J. M., and Ghahramani, Z. Ergodic measure preserving flows. *CoRR*, abs/1805.10377, 2018. URL <http://arxiv.org/abs/1805.10377>.