Information Geometry of Gibbs Sampler

Kazuya Takabatake
Neuroscience Research Institute
AIST Central 2, Umezono 1-1-1, Tsukuba
JAPAN 305-8568
k.takabatake@aist.go.jp

Abstract: - This paper shows some information geometrical properties of Gibbs sampler which is one of Markov chain Monte Carlo(MCMC) methods. The Gibbs sampler belongs to the class of the single-component-update MCMC, in which two or more components are never updated simultaneously. When a component is updated in the single-component-update MCMC, the chain's distribution moves along a m-flat manifold. In cases of the Gibbs sampler, the distribution moves to the point which minimizes the KL divergence to the target distribution on the m-flat manifold. From this viewpoint, the Gibbs sampler is interpreted as a greedy algorithm which minimizes the KL divergence in each update.

Key-Words: - Gibbs sampler, Markov chain Monte Carlo, information geometry, KL-divergence, greedy algorithms, convergence

1 Introduction

The most straightforward way to generate random numbers according to the given target distribution is using the table of the target distribution which consists of probabilities for all values the random variable takes. However the size of this table is proportional to the exponential of the dimension of the random variable, therefore this straightforward way is not practical for generating large-dimensional random variables.

Markov chain Monte Carlo(MCMC)[7] is a class of random number generators which uses Markov chains whose distribution converges to the target distribution.

Let $\{X^{(t)}\}(t = 0, 1, ...)$ be a Markov chain, V be the range of value $X^{(t)}$ takes [1], $p^{(t)}$ be the distribution of $X^{(t)}$, D_V be the set of distributions on V. Any distribution $q \in D_V$ can be represented in a vector form [2]

$$q = (q(x))(x \in V). \tag{1}$$

Let $W^{(t)}$ be the transition [3] from $X^{(t-1)}$ to $X^{(t)}(p^{(t-1)}$ to $p^{(t)})$. $W^{(t)}$ can be represented in a matrix form

$$W^{(t)} = (W^{(t)}_{xy})(x, y \in V) \tag{2}$$

$$W^{(t)}_{xy} := \Pr(X^{(t)} = y | X^{(t-1)} = x). \tag{3}$$

Then

$$p^{(t)} = p^{(t-1)} W^{(t)} \tag{4}$$

therefore

$$p^{(t)} = p^{(0)} W^{(1)} ... W^{(t)} \tag{5}$$

holds.

The mission of MCMC is generating $X^{(t)}$ for the given target variable $X^{(\infty)}$ and its distribution π. In designing MCMC, the sequence $\{W^{(t)}\}(t = 1, 2, ...)$ is designed as $p^{(t)}$ converges to π when $t \to \infty$. Under some conditions, we can make such $\{W^{(t)}\}$ without knowing the complete table of the target distribution π.

In this paper, we review the Metropolis-Hastings algorithm[7] and its special case of the Gibbs sampler[3, 7] at first. Then we show some information geometrical properties of the single-component-update MCMC and its special case of the Gibbs sampler.

[1] For the sake of simplicity, all random variables in this paper take finite discrete values.

[2] In this paper, symbols for distributions or transitions also denote their vector or matrix form.

[3] Any transition is a linear operator $D_V \to D_V$.

2 Metropolis-Hastings Algorithm and Gibbs Sampler

2.1 Metropolis-Hastings algorithm

Metropolis-Hastings algorithm[7] is the following algorithm.

step0 Prepare an arbitrary sequence of conditional distribution $\{q^{(t)}(X'|X)\}(t = 1, 2, ...)$, which is called proposal distribution, where X' is a candidate variable for next time. Set an arbitrary value to x. Set $t = 0$.

step1 Generate a random number x' according to $q^{(t)}(x'|x)$.

step2 Set $x = x'$ with probability

$$\alpha^{(t)}(x, x') = \min\left(1, \frac{\pi(x')q^{(t)}(x|x')}{\pi(x)q^{(t)}(x'|x)}\right), \quad (6)$$

which is called acceptance probability, otherwise keep x as it is.

step3 Set $t = t + 1$ and go to step1.

This algorithm simulates the Markov chain whose transition matrix is

$$W_{xy}^{(t)} = \begin{cases} q^{(t)}(y|x)\alpha^{(t)}(x, y) & x \neq y \\ 1 - \sum_{x \neq y} q^{(t)}(y|x)\alpha^{(t)}(x, y) & x = y \end{cases}. \quad (7)$$

This transition matrix holds the following so-called detailed balance equation[7] for all x, y, t.

$$\pi_x W_{xy}^{(t)} = \pi_y W_{yx}^{(t)} \quad (8)$$

Summing up eq.(8) about x, we get

$$\pi W^{(t)} = \pi \quad (9)$$

i.e. the transition by $W^{(t)}$ does not move π. It is known that if the Markov chain is weakly ergodic[2, 5] then the distribution $p^{(t)}$ converges to π when $t \to \infty$.

To perform this algorithm, we do not need to know the complete table of the target distribution π but just the ratio of probability $\pi(x')/\pi(x)$ in eq.(6). This is the major merit of the Metropolis-Hastings algorithm.

2.2 Single-component Metropolis-Hastings

Single-component Metropolis-Hastings[7] is a special case of Metropolis-Hastings algorithm. It is used in cases that X is multi-dimensional i.e. $X = (X_0, ..., X_{N-1})$.

In the single-component Metropolis-Hastings, only one component is updated in each transition. Therefore candidate x' differ from x in one component. Assume it is X_i and let $X_{\bar{i}}$ be the joint variable of other components. Then for all $x'_{\bar{i}} \neq x_{\bar{i}}$

$$q^{(t)}(x'_i, x'_{\bar{i}}|x_i, x_{\bar{i}}) = 0. \quad (10)$$

There are several ways to select the component updated at time t[7]. Let $i(t)$ be the suffix of the component updated at time t. In this paper, we adopt sequential-update:

$$i(t) = t \mod N. \quad (11)$$

2.3 Gibbs sampler

Gibbs sampler[3, 7] is a special case of Single-component Metropolis-Hastings. Its proposal distribution is [4]

$$q^{(t)}(x'_i, x'_{\bar{i}}|x_i, x_{\bar{i}}) = \begin{cases} \pi(x'_i|x_{\bar{i}}) & x'_{\bar{i}} = x_{\bar{i}} \\ 0 & x'_{\bar{i}} \neq x_{\bar{i}} \end{cases} \quad (12)$$

therefore acceptance probability is

$$\begin{aligned} &\alpha^{(t)}((x_i, x_{\bar{i}}), (x'_i, x_{\bar{i}})) \\ &= \min\left(1, \frac{\pi(x'_i, x_{\bar{i}})\pi(x_i|x_{\bar{i}})}{\pi(x_i, x_{\bar{i}})\pi(x'_i|x_{\bar{i}})}\right) \\ &= \min\left(1, \frac{\pi(x'_i, x_{\bar{i}})\pi(x_i, x_{\bar{i}})\pi(x_{\bar{i}})}{\pi(x_i, x_{\bar{i}})\pi(x'_i, x_{\bar{i}})\pi(x_{\bar{i}})}\right) \\ &= 1. \end{aligned} \quad (13)$$

It shows that the candidate x' is always accepted in step2 in section 2.1. Substituting eq.(12),(13) into eq.(7), we get

$$W^{(t)}_{(x_i,x_{\bar{i}})(y_i,y_{\bar{i}})} = \begin{cases} \pi(y_i|x_{\bar{i}}) & x_{\bar{i}} = y_{\bar{i}} \\ 0 & x_{\bar{i}} \neq y_{\bar{i}} \end{cases} \quad (14)$$

[4] In this paper, readers are expected to interpret symbols for distributions with x_i or $x_{\bar{i}}$ as appropriate conditional or marginal distributions. For example

$$\pi(x_i|x_{\bar{i}}) = \Pr(X_i^{(\infty)} = x_i|X_{\bar{i}}^{(\infty)} = x_{\bar{i}}).$$

and substituting this into eq.(4), we get

$$\begin{aligned} p^{(t)}(y) &= p^{(t)}(y_i, y_{\bar{i}}) \\ &= \sum_{x_i, x_{\bar{i}}} p^{(t-1)}(x_i, x_{\bar{i}}) W^{(t)}_{(x_i,x_{\bar{i}})(y_i,y_{\bar{i}})} \\ &= \sum_{x_i} p^{(t-1)}(x_i, y_{\bar{i}}) \pi(y_i|y_{\bar{i}}) \\ &= p^{(t-1)}(y_{\bar{i}}) \pi(y_i|y_{\bar{i}}) \\ &= p^{(t)}(y_{\bar{i}}) \pi(y_i|y_{\bar{i}}). \end{aligned} \tag{15}$$

The information about the target distribution π required to perform the Gibbs sampler is the full conditional distribution[7] $\pi(x_i|x_{\bar{i}})$ in eq.(12).

3 Information Geometry of MCMC

As we described in the introduction, we can treat a distribution as a point in a vector space. $\{p^{(t)}\}$ is the series of points which converges to π. In this section we show some information geometrical properties of the MCMC which updates only one component in each transition, and show some special properties the Gibbs sampler has.

3.1 Single-component-update MCMC

When i-th component is updated, other components keep their value therefore the marginal distribution about $X_{\bar{i}}$ is succeeded:

$$\forall x_{\bar{i}} \quad p^{(t)}(x_{\bar{i}}) = p^{(t-1)}(x_{\bar{i}}). \tag{16}$$

Let $M(p, i)$ be the manifold of distributions defined by

$$M(p, i) := \{q \in D_V | \forall x_{\bar{i}} \quad q(x_{\bar{i}}) = p(x_{\bar{i}})\}. \tag{17}$$

This manifold is m-flat(see appendix).

The important point is that updating i-th component can move p only along the manifold $M(p, i)$.

3.2 Gibbs sampler

For quick convergence of $p^{(t)} \to \pi$, it is a natural idea that we choose the closest point to π on $M(p^{(t-1)}, i)$. Some distance measure is required to determine the meaning of "closest". We adopt the KL-divergence $KL(p||\pi)$ as the distance measure.

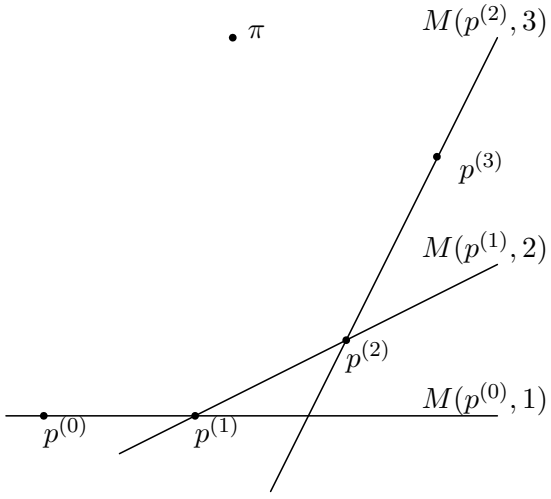

Fig. 1: This figure illustrates the movement of some $p^{(t)}$ in the distribution space D_V in the case of a single-component-update MCMC. Dots represent distributions and lines represent m-flat manifolds

In any MCMC, if we choose $W^{(t)}$ as it satisfies eq.(9), we get

$$\begin{aligned} KL(p^{(t)}||\pi) &= KL(p^{(t-1)}W^{(t)}||\pi W^{(t)}) \\ &\leq KL(p^{(t-1)}||\pi) \end{aligned} \tag{18}$$

from the data processing inequality(see appendix). It shows that $KL(p^{(t)}||\pi)$ decreases [5] as time goes.

Now we consider the following minimization problem

$$\min_{p^{(t)} \in M(p^{(t-1)}, i)} KL(p^{(t)}||\pi) \tag{19}$$

The minimizer $p^{(t)}$ is the e-projection(see appendix) of π onto m-flat manifold $M(p^{(t-1)}, i)$. Using the chain rule of KL-divergence(see appendix) we get

$$\begin{aligned} & KL(p^{(t)}||\pi) \\ &= KL(X_i^{(t)} X_{\bar{i}}^{(t)} || X_i^{(\infty)} X_{\bar{i}}^{(\infty)}) \\ &= KL(X_i^{(t)} X_{\bar{i}}^{(t-1)} || X_i^{(\infty)} X_{\bar{i}}^{(\infty)}) \\ &= KL(X_{\bar{i}}^{(t-1)} || X_{\bar{i}}^{(\infty)}) \\ &\quad + KL(X_i^{(t)} || X_i^{(\infty)} | X_{\bar{i}}^{(t-1)}). \end{aligned} \tag{20}$$

The minimization of eq.(19) is equivalent to

$$\min_{p^{(t)} \in M(p^{(t-1)}, i)} KL(X_i^{(t)} || X_i^{(\infty)} | X_{\bar{i}}^{(t)}) \tag{21}$$

[5] Here "decreases" means "at least never increase".

because the transition by $W^{(t)}$ does not move $X_{\bar{i}}^{(t)}$. From eq.(34), it is clear that the minimization is achieved when and only when

$$\forall x_{\bar{i}} \quad KL(X_i^{(t)}||X_i^{(\infty)}|x_{\bar{i}}) = 0. \tag{22}$$

It is equivalent to

$$\forall x_i, x_{\bar{i}} \quad p^{(t)}(x_i|x_{\bar{i}}) = \pi(x_i|x_{\bar{i}}). \tag{23}$$

Multiplying $p^{(t)}(x_{\bar{i}})$, we get the same equation as eq.(15). It implies that the Gibbs samplers transition matrix $W^{(t)}$ moves $p^{(t-1)}$ to $p^{(t)}$ which is the e-projection of π onto $M(p^{(t-1)}, i)$. The important point is that we need no information about $p^{(t)}$ to design the transition matrix $W^{(t)}$. In other words, by using Gibbs sampler's transition matrix, we can move any distribution p towards the closest point(e-projection) to π without knowing where p is.

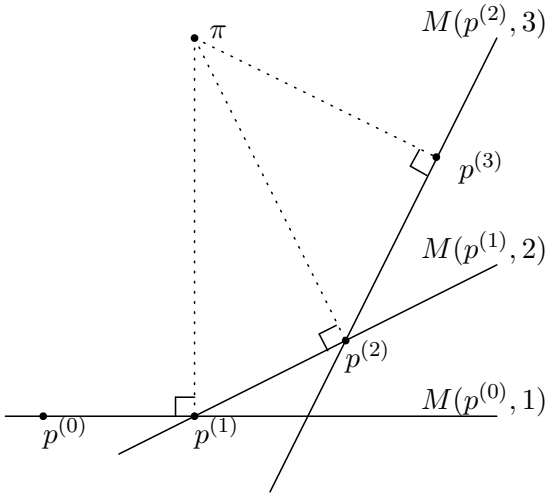

Fig. 2: This figure illustrates the movement of some $p^{(t)}$ in the distribution space D_V in the case of a Gibbs sampler. Dots represent distributions, lines represent m-flat manifolds and dashed lines represent e-geodesics

From the property discribed above, we can interpret the Gibbs sampler as the following algorithm.

step0 Set an arbitrary distribution to initial p.

step1 Move p to the e-projection of π onto $M(p, i)$.

step2 Go to step1.

This algorithm is a kind of greedy algorithm, because p is moved to the minimizer of the cost $KL(p||\pi)$ in *each* step1. In other words, p is not moved to the optimal point in two or more movements but in *each* movement. Imagine the simplified example shown in Fig.3. There are a point p and a target point π on a two dimensional plane. We can move p towards the east or the west for the first movement and towards the north-east or the south-west for the second movement. In this example lines from the west to the east and lines from the south-west th the north-east correspond to the manifold $M(p, i)$. If we move p to the closest point to π for the first movement, we can not make p reach π for the second movement. It is clear that the path written dashed lines is the optimal path to approach π.

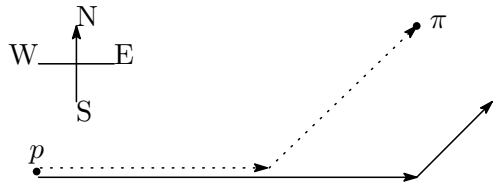

Fig. 3: This figure illustrates the movement of p. The path written in solid lines represents the movement of the greedy algorithm and the path written in dashed lines is optimal movement to approach to π

Another interesting property of the Gibbs sampler is

$$KL(p^{(t-1)}||\pi) = KL(p^{(t-1)}||p^{(t)}) + KL(p^{(t)}||\pi), \tag{24}$$

which is derived from Pythagorean theorem(see appendix). It means that the divergence which $p^{(t)}$ moves is equal to the divergence which $p^{(t)}$ approaches to π in each transition. Let $TD^{(n)}$ be the travelling divergence defined by

$$TD^{(n)} := \sum_{t=1}^{n} KL(p^{(t-1)}||p^{(t)}). \tag{25}$$

Then we get

$$TD^{(n)} + KL(p^{(n)}||\pi) = KL(p^{(0)}||\pi) \tag{26}$$

It implies that $TD^{(n)}$ has upper bound $KL(p^{(0)}||\pi)$ and if $p^{(t)}$ converges to π when $t \to \infty$, $TD^{(t)}$ converges to $KL(p^{(0)}||\pi)$.

4 Conclusion

Markov chain Monte Carlo(MCMC) is a class of random number generators which uses a Markov chain whose distribution $p^{(t)}$ converges to the target distribution π when $t \to \infty$.

In single-component-update MCMC, updating i-th component of the multi-dimensional variable X moves X's distribution p along the m-flat manifold $M(p, i)$.

Gibbs sampler is one of single-component-update MCMC. From the viewpoint of information geometry, the Gibbs sampler is interpreted as the following algorithm.

step0 Set an arbitrary distribution to initial p.

step1 Move p to the e-projection of π onto $M(p, i)$.

step2 Go to step1.

This algorithm is a kind of greedy algorithm, because p is moved to the minimizer of the cost $KL(p||\pi)$ in each step1.

In the Gibbs sampler, $p^{(t)}$ does not travels infinite divergence in the distribution space. The divergence $p^{(t)}$ travels is less or equal to $KL(p^{(0)}||\pi)$. If $p^{(t)}$ converges to π when $t \to \infty$, the traveling divergence converges to $KL(p^{(0)}||\pi)$.

References:

[1] Csiszár, I., Körner, J., *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981

[2] Seneta, E., *Non-negative Matrices and Markov Chains, Second Edition*, Springer-Verlag, 1981

[3] Geman, S., Geman, D., Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 6, 1984, pp.721-741

[4] Amari, S., *Differential-Geometrical Methods in Statistics*, Lecture Notes in Statistics, Vol. 28. Springer-Verlag, 1985

[5] van Laarhoven, P. J. M., Aarts, E. H. L., *Simulated Annealing: Theory and Applications*, Kluwer Academic Publishers, 1987

[6] Amari, S., Information Geometry of the EM and em Algorithms for Neural Networks, *Neural Networks*, Vol. 8, No. 9, 1995, pp.1379-1408

[7] Gilks, W. R., Richardson, S., Spiegelhalter, D. J., Introducing Markov chain Monte Carlo, In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J.(ed.), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, 1996, pp.1-19

[8] Gilks, W. R.: Full conditional distributions, In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J.(ed.), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, 1996, pp.1-19

Appendix: KL-divergence

Let $R(X)$ be the range of value X takes. For two random variables X, Y which have the same range, KL-divergence X to Y or KL-divergence their distributions p_X to p_Y is defined by

$$KL(X||Y) = KL(p_X||p_Y) := \sum_{x \in R(X)} p_X(x) \log \frac{p_X(x)}{p_Y(x)}. \quad (27)$$

KL-divergence is always non-negative and

$$KL(p_X||p_Y) = 0 \iff p_X = p_Y. \quad (28)$$

Let q be the distribution of a stochastic source, p be a data's distribution and N be the number of samples in the data. Log-likelihood of the data comes from the stochastic source is

$$L(p||q) = \sum_x Np(x) \log q(x) \quad (29)$$

and it takes maximum value $-NH(p)$ when and only when $p = q$, where $H(p)$ is Shannon's entropy:

$$H(p) = -\sum_x p(x) \log p(x) \quad (30)$$

KL-divergence $KL(p||q)$ is

$$KL(p||q) = -\frac{1}{N} L(p||q) - H(p) \quad (31)$$

Therefore the meaning of $KL(p||q)$ is "biased log-likelihood of data whose distribution is q comes out from distribution p. The bias is taken as $KL(p||q) = 0$ when $p = q$".

For any distribution vectors p, q and any transition matrix W,

$$KL(pW||qW) \leq KL(p||q) \tag{32}$$

holds. This inequality is called "data processing inequality".

Let X, Z be random variables which have the same range and Y, W be random variables which have the same range. The following equation holds for the joint variables XY and ZW.

$$KL(XY||ZW) = KL(X||Z) + KL(Y||W|X) \tag{33}$$

where

$$KL(Y||W|X) := \sum_{x \in R(X)} \Pr(X = x) KL(Y||W|x) \tag{34}$$

$$KL(Y||W|x) := \sum_{y \in R(Y)} \Pr(Y = y|X = x) \times \log \frac{\Pr(Y = y|X = x)}{\Pr(W = y|Z = x)}. \tag{35}$$

Eq.(33) is called "chain rule of KL-divergence" and the left side of eq.(34) is called "conditional KL-divergence".

Let q be a distribution and M be a manifold of distributions.

$$\arg\min_{p \in M} KL(p||q) \tag{36}$$

is called "e-projection of q onto M"[6]. If M has the following property

$$p, q \in M, 0 \leq \lambda \leq 1 \Rightarrow \lambda p + (1 - \lambda) q \in M, \tag{37}$$

we call "M is m-flat". It is known that if M is m-flat the e-projection of q onto M is unique for any distribution q.

The following curve is called e-geodesic from p_0 to p_1[6]:

$$\{q| \log q(x) = (1 - \lambda) \log p_0(x) + t \log p_1(x) - \log \phi(\lambda), \quad 0 \leq \lambda \leq 1\}. \tag{38}$$

where $\phi(t)$ is the term for the normalizing condition $\sum_x q(x) = 1$:

$$\phi(\lambda) = \sum_x p_0(x)^{1-\lambda} p_1(x)^{\lambda}. \tag{39}$$

Let q be a distribution, M be a manifold of distributions and p be the e-projection of q onto M. It is known that the e-geodesic form q to p and M are orthogonal at p. And for any distribution $r \in M$, the following equation holds.

$$KL(r||q) = KL(r||p) + KL(p||q) \tag{40}$$

It is called "Pythagorean theorem"[1].

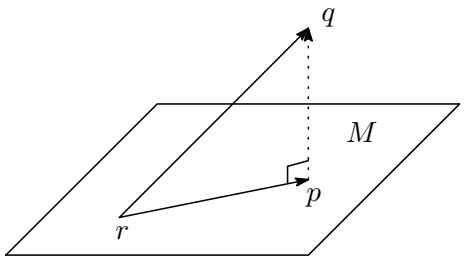

Fig. 4: This figure illustrates the Pythagorean theorem: $KL(r||q) = KL(r||p) + KL(p||q)$. M is a m-flat manifold of distributions and p is the e-projection of q onto M. Dashed line is the e-geodesic from q to p