数理解析研究所講究録 1240 巻 2001 年 58-63

Accelerating Diffusions and Related Problems

Chii-Ruey Hwang
Institute of Mathematics, Academia Sinica, Taipei,TAIWAN

ABSTRACT

Let $\pi(x)$ be a density function proportional to $exp - U(x)$ in R^d. The following diffusion $X(t)$ is often used to sample from $\pi(x)$.

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2}dW(t), X(0) = x_0.$$

To accelerate the convergence, one may consider the following family of diffusions with $\pi(x)$ as the equilibrium distribution.

$$dX(t) = (-\nabla U(X(t) + C(X(t)))dt + \sqrt{2}dW(t), X(0) = x_0.$$

Let L_C be the corresponding infinitesimal generator. The following criterion is used to describe the convergence rate.

$$\lambda(C) = \{ \text{ real part of } \mu : \mu \text{ is in the spectrum of } L_C, \mu \text{ is not zero } \}.$$

The smaller the $\lambda(C)$ is the better the convergence rate. Adding an extra drift C(x) will accelerate the convergence except in some situation that no improvement can be made. Related problems are also discussed.

1 Introduction

High dimensional probability distributions appear frequently in applications. To sample from these distributions directly is not feasible in practice. One has to resort to approximations. A Markov process with the underlying distribution as its equilibrium could be used for the approximation. How good the approximation is depends on the comparison criterion. And regarding a Markov process as some sort of conceptual algorithm is useful in the theoretical study as well as in applications.

The underlying distribution π is assumed to have a density proportional to $\exp -U(x)$ with some smooth condition on U. The following diffusion is used commomly for sampling from π,

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2}dW(t), \quad X(0) = x_0, \quad (1)$$

where $W(t)$ is the Brownian motion in R^d.

For using diffusions to sample underlying distributions in real applications, one may consult [Grenander and Miller 1994], [Miller, Srivastava and Grenander 1995], [Srivastava 1996] and references therein.

If a diffusion is regarded as a theoretical algorithm, then it is natural to consider a family of diffusions (algorithms). And within that family one may try to do the comparison and even tries to find an optimal one.

A family of diffusions of the following form is considered:

$$dX(t) = -\nabla U(X(t))dt + C(X(t))dt + \sqrt{2}dW(t), \quad X(0) = x_0. \quad (2)$$

Under suitable conditions on $C(x)$, π is their common equilibrium. An specific example for $C(X)$ is $\mathrm{div}(C(X)\exp -U(x)) = 0$ and there is no explosion. We are interested in the convergence rate of the process, with different choices of $C(X)$, to the equilibrium.

The Gaussian diffusion has some satisfactory answers [Hwang, Hwang-Ma and Sheu 1993]. Once beyond the Gaussian case, the situation is more complicated even for the very definition of convergence rate. And the related problems are very challenging.

Let L_C denote the infinitesimal generator of $X(t)$ and for $C = 0$, let $L = L_0$. Let $T(t) = e^{tL_C}$ denote the corresponding semigroup,

$$T(t)f(x) = E_x f(X(t)) = \int p(t,x,y)f(y)\,dy,$$

where $p(t,x,y)$ is the transition density. Note that the index C is suppressed from $T(t)$ for the sake of brevity.

Now we start to sketch the problem under consideration.

Since $E_x f(X(t)) \to \pi(f)$, it is reasonable to consider the average case formulation, i.e. averaging over the starting point x :

$$\int (E_x f(X(t)) - \pi(f))^2 \pi(x)\,dx = \| T(t)f - \pi(f) \|^2 \le c\| f \|^2 e^{2\lambda t}, \quad (3)$$

where $\|\ \ \|$ is the norm in $L^2(\pi)$.

Now consider the worst-case analysis over f, then the spectral radius of $T(1)$ in the space $\{f \in L^2(\pi), \pi(f) = 0\}$ is an indicator for the convergence rate of diffusions. Furthermore the weak spectral mapping theorem holds between L_C and e^{tL_C} [Nagel 1986]. Hence,

$$\lambda(C) = \{\text{real part of } \mu : \mu \text{ in the spectrum of } L_C, \mu \neq 0\} \quad (4)$$

is a good candidate to serve as a criterion for the comparison of the convergence rates.

The main result is $\lambda(C) \le \lambda(0)$ and the equality holds in some rare situation which is characterized completely. In other words by adding an extra drift will accelerate the convergence.

The results are stated in Section 2. The proofs may be found in the forthcoming paper "Accelerating diffusions" by Chii-Ruey Hwang, Shu-Yin Hwang-Ma, Shuenn-Jyi Sheu. Section 3 is for the related problems and discussions.

2 Results

We assume that

$$C \text{ and } \nabla U \text{ are in } L^2_{loc}(\pi).\ C \text{ is in } L^1(\pi).\ \text{For } f \in C_0^\infty, \int (C \cdot \nabla f)\pi = 0.$$

Then there is no explosion and π is the equilibrium distribution of (2) (Stannat 1999). Note that for $f \in C_0^\infty$,

$$Lf = -\nabla U \cdot \nabla f + \Delta f \quad \text{and} \quad L_C f = Lf + C \cdot \nabla f.$$

Intuitively L_C is a perturbation of a symmetric operator L by an anti-symmetric operator $C \cdot \nabla$. And we are interested in how the spectrum changes. The spaces considered are real vector spaces of real functions. However for spectral analysis, one has to consider the complex vector spaces. We will make the distinction when it is necessary.

Moreover we assume that

$$1/2|\ \nabla U(x)\ |^2 - \Delta U(x) \longrightarrow \infty \text{ as } |\ x\ | \rightarrow \infty.$$

Lemma1.(Reed and Simmon 1978,p.249)
Under the above assumptions, L has compact resolvents and moreover it has purely discrete spectrum and a complete orthonormal base consisting of eigenfunctions.

Define

$$\epsilon^0(f,g) = \int (\nabla f \cdot \nabla g)\pi, \quad f, g \in C_0^\infty.$$

ϵ^0 is closable in $L^2(\pi)$. Let $D(\cdot)$ denotes "the domain of".

Lemma2. (Stannat 1999,p.124)
Let $f \in D(L_C)$, then

$$f \in D(\epsilon^0) \text{ and } \epsilon^0(f,f) \leq -\int (L_C f) f \pi.$$

Lemma3.
L_C does not have continuous spectrum.

Lemma4.
If there exists a sequence of eigenvalues $\{a_n + ib_n\}$ of L_C such that

$$a_n < \lambda(0), \quad a_n \longrightarrow \lambda(0), \quad b_n \longrightarrow \infty \quad (or \quad b_n \longrightarrow -\infty),$$

then $\lambda(0)$ is the real part of an eigenvalue of L_C. The same assertion holds for the residual spectrum.

Lemma5.
If λ is the real part of an eigenvalue (or an element in the residual spectrum) of L_C, then $\lambda \leq \lambda(0)$. Equality holds if and only if $C \cdot \nabla$ maps a nonzero subspace of the eigenspace with eigenvalue $\lambda(0)$ of L into itself.

Theorem.
The diffusion (1) with the gradient drift has the worst convergence rate. By adding an extra drift C, the diffusion (2) does accelerate the convergence to the equilibrium π except in the case decsribed in Lemma5 that no improvement can be made.

3 Discussions and Related Problems

The main theorem gives only a general and qualitative answer. It does not shed any light on how the rate depends on C. Each C corresponds to a diffusion (algorithm). Then what is the best algorithm within a certain family? E.g. let **G** and **S** denote the family of C satisfying general conditions described in the previous section and $C = S(\nabla U)$ for any antisymmetric matrix S respectively. And we are interested in
1) $\inf_C \lambda(C)$, $C \in \mathbf{G}$.
2) $\inf_C \lambda(C)$, $C \in \mathbf{S}$.

For 1) we even do not know if $\inf_C \lambda(C)$ equals $-\infty$. 2) is open in general. For the Gaussian case 2) has a satisfactory answer, the optimal structure is known [Hwang, Hwang-Ma, Sheu 1993]. If an independent diffusion is added to the original one to form a $d+1$ dimensional diffusion and consider the original d dimensional projection of the corresponding optimal $d+1$ dimensional diffusion, then the projection is still Gaussian and the convergence rate can be speeded up to infinity theoretically. This is closely related to the Swensen-and-Wang algorithm in spirit. A lot is still unknown.

Basically the above formulation is to find the best "spectral gap" in a family of differential operators. On compact Riemannian manifolds, we may consider similar problems. Here is a generic case: on two dimensional torus consider
3) $\inf_C \lambda(C)$, $L_C = \Delta + c \cdot \nabla$, C is divergence free.
Again, is this quantity $-\infty$?

From a more probabilistic point of view, we consider

$$\int | p(t,x,y) - \pi(y) | \, dy \leq M(x) \ e^{\rho t}.$$

Note that

$$\int \mid p(t,x,y) - \pi(y) \mid dy = 2 \ \sup_A (P(t,x,A) - \pi(A)),$$

which is twice the variational norm between the transition probability and the equilibrium. This is the worse-case analysis of the difference of those two probabilities on sets. $M(x) \in L^1(\pi)$ means averaging over the initials. Let $\rho(C)$ denote the infimum over all ρ's. How do we compare $\rho(C)$ and $\lambda(C)$? Is there any inconsistency for these two comparison criteria?

One may consider the dependence of the convergence rate on the initial point. In the Ornstein-Uhlenbeck processes, i.e.2), if the mean is known, say mean 0, then the process starting at 0 has a much faster convergence. Now an essential question is: how to use the known information to choose the process? E. g. the mean and covariance or some eigenfunctions of the corresponding operators are known.

The formulation of a fixed self-adjoint operator perturbed by a family of anti-symmetric operators, can be considered as a sort of dual problem of stability of linear system [Bellman 1960, Kransosel'skij et al 1989]. The continuous time approach for sampling shares some similar spirit with the works of [Chu 1992, 1995, Chu and Driessel 1990] in numerical analysis. All of these deserve further investigation.

One may consider other criteria. We remark that in formula (3), if the constant is taken to be one, then the rate depends only on the behavior around time 0 and remains the same for different C's [Chen 1992].

REFERENCES

Amit, Y. (1996). Convergence properties of the Gibbs sampler for perturbations of Gaussians, Ann. Stat. 24, 122-140.

Athreya, K. A. , H. Doss and J. Sethuraman (1996). On the convergence of the Markov Chain simulation method. Ann. Stat. 24, 69-100.

Bellman, R. (1970). Methods of Nonlinear Analysis. Math. Sci. Engrg. 61-I. Academic, New York.

Chang, H. -C. ,C. -R. Hwang (1998) On the average-case analysis of dynamic Monte Carlo schemes.

Chen, C. N. , C. I. Chou, C. -R. Hwang, J. Kang, T. K. Lee and S. P. Li (1999). Monte Carlo dynamics in global optimization, Physical Review E, 60, 2388-2393.

Chen, M. F. (1992). From Markov Chains to Non-equilibrium Particle Systems, World Scientific.

Chu, M. T. (1995). Constructing a Hermitian matrix from its diagonal entries and eigenvalues, SIAM J. Matrix Anal. Appl. ,16, 207-217.

Chu, M. T. (1992). Numerical methods for inverse singular value problems. SIAM J. Numer. Anal. , 29, 885-903.

Chu, M. T. and K. R. Driessel (1990). The projected gradient method for least squares matrix approximations with spectral constraint, SIAM J. Numer. Anal. , 27, 1050-1060.

Frigessi, A. , Hwang, C. -R. and Younes, L. (1992). Optimal spectral structures of reversible stachastic matrices, Monte Carlo methods and the simulation of Markov

random fields. Ann. Appl. Probab. 2, 610-628.

Frigessi, A. , Hwang, C. -R. , Sheu, S. -J. and di Stefano, P. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm, and other single-site updating dynamics. J. Roy. Statist. Soc. Ser. B 55, 205-219.

Goodman, J. and Sokal, A. D. (1989). Multigrid Monte Carlo method. Conceptual foundations. Phys. Rev. D 40, 2035-2071.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems. JRSS, B, 56, 549-603.

Hwang, C. -R. , S. -Y. Hwang-Ma and S. -J. Sheu (1993). Accelerating Gaussian diffusions. Ann. Appl. Probab. , 3, 897-913.

Hwang, C. -R. , S. -Y. Hwang-Ma and S. -J. Sheu (2001). Accelerating diffusions, in preparation.

Hwang, C. -R. and Sheu, S. -J. (1990). Large-time behavior of perturbed diffusion Markov processes with applications to the second eigenvalue problem for Fokker-Planck operators and simulated annealing. Acta Appl. Math. 19, 253-295.

Hwang, C. -R. and S. -J. Sheu (1998). On the geometric convergence of Gibbs sampler in R^d, J. Multi. Analy. ,66,22-37

Hwang, C. -R. and S. -J. Sheu (2000). On some quadratic perturbation of Ornstein-Uhlenbeck processes, Soochow J. of Math. ,26,205-244.

Kato, T. (1995). Perturbation Theory for Linear Operators. Classics in Mathematics, Springer.

Krasnosel'skij, M. A. , Je. A. Lifshits and A. V. Sobolev (1989). Position Linear Systems: The Method of Positive Operators. Sigma Series in Applied Math. 5, Heldermann Verlag.

Nagel, R. (1986). One-parameter Semigroup of Positive Operators. LN 1184, Springer.

Miller, M. I. , A. Srivastava and U. Grenander (1995). Conditional-mean estimation via jump-diffusion processes in multiple target tracking/recognition. IEEE Transac. Sig. Processing, 43, 2678-2690.

Reed, M. and Simon, B. (1978). Methods of Modern Mathematical Physics 4, Academic, New York.

Srivastava, A. (1996). Inferences on transformation groups generating patterns on rigid motions. Ph. D. thesis, Dept. E. E. , Washington University.

Stannat, W. (1999). (Nonsymmetric) Dirichlet operators on L^1:existence, uniqueness and associated Markov process,Ann. Scola Norm. Sup. Pisa Cl. Sci., XXVIII, 99-140.

Varadhan, S. R. S. (1980). Lectures on Diffusion Problems and Partial Differential Equations. Springer, New York.

Yosida, K. (1980). Functional Analysis, 6th Ed. , Springer.