Nonreversible Perturbations Accelerate Convergence*

Chii-Ruey Hwang
Institute of Mathematics, Academia Sinica, Taipei, Taiwan 11529

ABSTRACT

To sample from distributions in high dimensional spaces or finite large sets directly is not feasible in practice, especially when the corresponding densities are known up to normalizing constants only. One has to resort to approximations. A Markov process with the underlying distribution as its equilibrium is often used to generate an approximation ("MCMC"). How good the approximation is depends on the approximating Markov process and on the specific criterion used for comparison. One may investigate the convergence properties of some particular Monte Carlo Markov processes, or compare the convergence rate within a family of Markov processes (with the same equilibrium) w.r.t. different criteria, or even try to find optimal solutions in that family. Mathematical problems arising from this approach are challenging. By simply adding a weighted divergence-free drift to a reversible diffusion, the convergence to equilibrium is accelerated. In other words, from an algorithmic point of view the nonreversible algorithm performs better. Note that two criteria are considered. The analysis is related to the study of antisymmetric perturbations of self-adjoint infinitesimal generators. The optimal solution is still open. However on torus the rate could be pushed to infinity. As for finite sample space nonreversible perturbations reduce variance.

1 Introduction

In this talk we consider nonreversible (weighted antisymmetric) perturbations of reversible diffusions and finite state Markov chains. By simply adding a weighted

*Partially Supported by NSC GRANT 94-2115-M-001-001

AMS 2000 subject classifications. Primary 60J60, 47D07;secondary 65B99, 35P05.

Key words and phrases. Diffusion, convergence rate, acceleration, spectral gap, spectrum, variational norm, ergodicity, MCMC, Monte Carlo Markov process, variance reduction, antisymmetric perturbation, nonreversible perturbation.

divergence-free drift to a reversible diffusion, the convergence to equilibrium is accelerated. In other words, from an algorithmic point of view the nonreversible algorithm performs better. The analysis is related to the study of antisymmetric perturbations of self-adjoint infinitesimal generators. As in the Markov chain case, antisymmetric perturbations of reversible dynamic Monte Carlo methods have variance reduction effect.

To sample directly from distributions in high dimensional spaces or large finite sets is not feasible in practice, especially when the corresponding densities are known up to normalizing constants only. One has to resort to approximations. A Markov process with the underlying distribution as its equilibrium is often used to generate an approximation ("MCMC"). How good the approximation is depends on the approximating Markov process and on the specific criterion used for comparison. One may investigate the convergence properties of some particular Monte Carlo Markov processes, or compare the convergence rate within a family of Markov processes (with the same equilibrium) w.r.t. different criteria, or even try to find optimal solutions in that family. Mathematical problems arising from this approach are challenging. Related works may be found in Amit (1991), Amit and Grenander (1991), Frigessi, Hwang and Younes (1992), Frigessi, Hwang, Sheu and di Stefano (1993), Hwang, Hwang-Ma and Sheu (1993), Amit (1996), Athreya, Doss and Sethuraman (1996), Gilks and Roberts (1996), Mengersen and Tweedie (1996), Stramer and Tweedie (1997), Chang and Hwang (1998), Hwang and Sheu (1998, 2000), Roberts and Rosenthal (2004).

We formulate the diffusion case first. Let U be a given real-valued function defined in R^d satisfying some smoothness conditions. The underlying distribution π is assumed to have a density proportional to $\exp -U(x)$. The following diffusion is commonly used for sampling from its equilibrium 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 convenience, π will be used to denote the underlying probability measure as well as its probability density.

If a diffusion is regarded as a useful approach to sampling, then it is natural to consider a family of diffusions with π as their common equilibrium :

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

under suitable conditions on $C(x)$. Roughly speaking, the conditions are that $\mathrm{div}(C(x)\exp -U(x)) = 0$ and there is no explosion in (2), i.e. $\mid X(t) \mid$ does not tend to infinity in a finite time. A strict definition of explosion can be found on p.172 of Ikeda and Watanabe (1989). It is easy to pick such a C. For example $C(x) = S(\nabla U(x))$, for any skew symmetric matrix S. We are interested in how $C(x)$ influences the convergence of the diffusion (2) to equilibrium.

Hwang, Hwang-Ma and Sheu (1993) focused on a special case, the study of a family of Gaussian diffusions where $2U(x) = (-Dx) \cdot x, \quad -\nabla U(x) = Dx, \quad C(x) =$

SDx, and where D is a strictly negative-definite real matrix and S is any skew symmetric real matrix. In this case, $\pi(x)$ is Gaussian with mean 0 and covariance matrix $-D^{-1}$ and $X(t)$ is an Ornstein-Uhlenbeck process with drift $(D + SD)x$. Using the rate of convergence of the covariance of $X(t)$ (or together with $EX(t)$) as the criterion, the reversible diffusion with drift Dx (i.e. $C = 0$) is the worst choice and the optimal solution is obtained in this setup.

If $C(x)$ is not zero, then the corresponding diffusion, regarded as a Markov process, is nonreversible. In general it is difficult to analyze nonreversible processes. We just cite some related works in different settings. In Geman and Geman (1984), Amit and Grenander (1991), Hwang and Sheu (1998) the convergence properties of some nonreversible Gibb samplers is studied. The ergodicity of systematic sweep in stochastic relaxation, again nonreversible, is investigated in Hwang and Sheu (1992).

Two comparison criteria are considered here. Basic questions such as the acceleration of convergence and the consistency of the comparison w.r.t. these two criteria are answered.

Let L_C denote the infinitesimal generator of the diffusion $X(t)$ from (2) 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 if it exists. Note that the index C is suppressed from $T(t)$ and $p(t, x, y)$ for the sake of brevity.

We define now the spectral gap of L_C in $L^2(\pi)$ as the first comparison criterion. Since $E_x f(X(t)) \to \pi(f)$ for any starting point x, one may consider the average case formulation by averaging the difference $(E_x f(X(t)) - \pi(f))^2$ over the starting point w.r.t. π:

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

for some λ less than or equal to 0, where $\pi(f)$ means integration of f w.r.t. π. Now consider the worst-case analysis over f, then $\| T(t) - \pi \| \leq$ constant $e^{\lambda t}$. The infimum over such λ's indicates the convergence rate. This shows that the spectral radius of $T(1)$ in the space $\{f \in L^2(\pi), \pi(f) = 0\}$ is a measure of convergence rate of diffusions to equilibrium. Furthermore the weak spectral mapping theorem holds between L_C and e^{tL_C} (Nagel (1986) p.91). Hence, the spectral gap of L_C in $L^2(\pi)$ defined by

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

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

The constant in (3) may depend on C. If instead we reformulate the inequality in (3) without the constant term,

$$\| T(t)f - \pi(f) \| \leq \| f - \pi(f) \| e^{\lambda t},$$

for some λ, then the inequality depends only on the behavior of the process around time 0 and the rate will be the same regardless of perturbations (Chen (1992) p.312). Our interest here is instead in the large-time behavior.

We will always assume that there is no explosion for the diffusions under consideration. For simplicity we just assume that the following assumption holds throughout this paper,

$$\textbf{(A1)}: C \text{ and } \nabla U \text{ are in } L^1(\pi) \cap L^l_{loc}(\pi), \text{ for some } l > d;$$

$$\text{for } f \in C_0^\infty, \int (C \cdot \nabla f)\pi = 0.$$

Under **(A1)** there is no explosion in the diffusion (2) and the transition density exists with π as its equilibrium distribution (Stannat (1999), Bogachev, Krylov and Röckner (2001)). For $f \in C_0^\infty$, $\int (C \cdot \nabla f)\pi = 0$ means that C is weakly weighted divergence free. This is essential for π to be an invariant measure.

Intuitively L_C is a perturbation of a self-adjoint operator L by an antisymmetric operator $C \cdot \nabla$ in $L^2(\pi)$. We are interested in how the spectrum changes. Note that in general this perturbation is neither small nor relatively compact. For general references, refer to Kato (1995), Yosida (1980). L_C is not self-adjoint for nonzero C. The spaces considered are real vector spaces of real functions. However for spectral analysis, one has to consider complex vector spaces. Let C_+ denote $L_C - L$ and C_- denote $L_{-C} - L$.

We assume that the reversible diffusion (1) w.r.t. π has an exponential convergence rate. Equivalently L has a spectral gap in $L^2(\pi)$, i.e.

(A2): $\lambda(0) < 0$.

Note that the exponential convergence rate assumption is imposed only on the reversible diffusion. As a consequence of Theorem 1 the perturbed diffusion (2) has a better exponential convergence rate. In other words, adding an extra drift accelerates convergence.

For the nonexplosion of (1), **(A2)**, and $\lambda(0)$ in the discrete spectrum of L to all hold, the following is a sufficient condition (Reed and Simon (1978)):

$$1/2 \mid \nabla U(x) \mid^2 - \Delta U(x) \longrightarrow \infty \text{ as } \mid x \mid \rightarrow \infty. \quad (5)$$

From a probabilistic point of view, one may consider the rate of convergence of $p(t,x,y)$ to π in variational norm as a comparison criterion. The variational norm of two probability measures is defined as the supremum of the difference between the two probabilities over all events. This may be regarded as some kind of worst case analysis. Note that the variational norm equals one half of the $L^1(dy)$ distance between the two corresponding densities. Hence, $\rho(C)$ defined below is used as a comparison criterion,

$$\rho(C) = Inf\{\rho : \int \mid p(t,x,y) - \pi(y) \mid dy \leq g(x)e^{\rho t}\}. \quad (6)$$

$g(x)$ may depend on C. Usually g is assumed to be essentially locally bounded or locally integrable w.r.t. π. It needs further study for unrestricted g. Theorem 2 and Theorem 3 show that $\rho(C) \leq \lambda(C)$ and equality holds for the reversible case. Again using $\rho(C)$ as the comparison criterion, adding an antisymmetric perturbation does help. This result is consistent with the previous one.

It is not clear how the perturbations affect $\rho(C)$ directly. We compare $\rho(C)$ and $\rho(0)$ via $\lambda(C)$ and $\lambda(0)$.

Results for the above two criteria are given in section 2. Details, proofs, and discussion for other criteria can be found in Hwang, Hwang-Ma, Sheu (2005).

One may ask for the existence of an optimal C or whether $Inf\lambda(C) = -\infty$. We have answer only for torus case, it is still open even for R^d or S^d. We consider antisymmetric perturbations of Laplacian on n dimensional torus, namely, the Brownian motion is perturbed by a divergence free drift C to accelerate the convergence to its equilibrium, the uniform distribution on torus.

The approach goes as follows. Some specific form of C's are chosen. For each picked C, $Inf_k\lambda(kC)$ is characterized explicitly. From the characterzation, one can show that $Inf\lambda(C) = -\infty$, the infimun is taken over general C, Hwang, Pai (2005). This result is Theorem 4.

Now we turn to the dynamic Monte Carlo method. Assume that S is a finite set. We are interested in the evaluation of $\sum_{s\in S} f(s)\pi(s)$, denoted by $\pi(f)$. Let $X_0, X_1, ..., X_n, ...$ be a Markov chain with transition matrix P and invariant probability π. Under suitable condition on P, it is known that $\frac{1}{n}\sum_{k=0}^{n-1} f(X_k)$ converges to $\pi(f)$ and the corresponding asymptotic variance $v(f, P)$ depends only on f and P.

For any two real-valued functions f and g defined on S, the weighted inner product $< f, g >$ is defined by

$$< f, g > \equiv \sum_{s\in S} f(s)g(s)\pi(s).$$

The following result is well-known, see e. g. Theorem 4.8 in Iosifescu 1980. Under suitable condition on P, for any initial distribution μ_0,

$$\lim_{n\to\infty} nE_{\mu_0}[\frac{1}{n}\sum_{0}^{n-1} f(X_k) - \pi(f)]^2 = v(f, P),$$

where the asymptotic variance $v(f, P)$ equals

$$< (I - P)^{-1}(I + P)(f - \pi(f)), f - \pi(f) > .$$

The inverse is taken in the space $\mathcal{N} = \{f : \pi(f) = 0\}$ and the $\pi(f)$ appearing in $f - \pi(f)$ means constant function with value $\pi(f)$.

Regarding P as a theoretical algorithm and without exploiting any prior knowledge on any specific f, the worst case analysis for reversible P was investigated in Frigessi, Hwang, Younes (1992) where the optimal P is characterized.

For any fixed reversible P, we consider the perturbation of P by adding an antisymmetric (w.r.t. the above mentioned inner product) Q with row sums of Q

being zero and entries of $P + Q$ being positive. The perturbation reduces variance, this is Theorem 5. Details can be found in Hwang, Hwang-Ma (2005).

2 Results

If $\lambda(0)$ is in the discrete spectrum of L in $L^2(\pi)$, then by definition its corresponding eigenspace, denoted by $\mathbf{M}$, is finite dimensional. In this section our analysis assumes $\pi(f) = 0, f \in L^2(\pi)$.

Theorem 1. If **(A1)** **(A2)** hold, then $\lambda(C) \leq \lambda(0)$. Furthermore if $\lambda(0)$ is in the discrete spectrum of L, then equality holds if and only if C_+ or C_- leaves a nonzero subspace of $\mathbf{M}$ invariant.

Remark. It seems that a stronger result should hold: if $\lambda(C) = \lambda(0)$, then $\lambda(0)$ is the real part of an eigenvalue of L_C. If this is the case, Theorem 1 has a stronger form: the equality holds iff C_+ leaves a nonzero subspace of $\mathbf{M}$ invariant. If (5) holds, then $(L-a)^{-1}$ is compact for a in the resolvent of L (Reed and Simon (1978)). And the stronger statements hold.

Remark. As mentioned in the Introduction, the existence of the transition density is not needed here. A weaker assumption than **(A1)** suffices, e.g. C and ∇U are in $L^1(\pi) \cap L^2_{loc}(\pi)$ (Stannat (1999)).

Under **(A1)** the transition density $p(t,x,y)$ exists. Let $p_t(x,y)$ denote $p(t,x,y)/\pi(y)$; $p_t(x,y)$ is locally Hölder [Bogachev, Krylov, Röckner 2001].

Theorem 2. In addition to **(A1)** **(A2)** if ∇U and C are locally bounded, then there exists a locally bounded function g such that

$$\int \mid p_t(x,y) - 1 \mid \pi(y)dy \leq g(x)e^{\rho(c)t}.$$

Moreover, $\rho(C) \leq \lambda(C)$.

Remark. The local boundedness assumption in Theorem 4 is not needed for the reversible case, since $\int p_1^2(x,y)\pi(y)dy = p_2(x,x)$ is locally bounded.

The following theorem implies that for the reversible case, $\rho(0) = \lambda(0)$.

Theorem 3. For the reversible case, if there exists some g in $L^1_{loc}(\pi)$ such that

$$\int \mid p_t(x,y) - 1 \mid \pi(y)dy \leq g(x)e^{\rho t},$$

$$\text{then} \quad \| T(t) - \pi \| \leq e^{\rho t}.$$

We are going to consider the perturbation of Laplacian by divergence free vector field on n-torus.

Assume $C(x) = p\cos(q\cdot x)$, for $x = (x_1, ..., x_n) \in R^n$, $p, q \in Z^n$ satisfying $p\cdot q = 0$. It is not difficult to check that C is divergence free.

Theorem 4.

$$\lim_{k\to\infty} \lambda(kC) = \sup\{- \mid \hat{m} \mid^2 \colon \hat{m}\cdot p = 0, \hat{m} \in Z^n - \{0\}\}.$$

Hence by properly chosen C, the spectral gap can be pushed to $-\infty$.

For the finite sample space, we are interested in the variance reduction for dynamic Monte Carlo evaluation of expection.

Theorem 5.
Let P be a stochastic matrix reversible w.r.t. π. If P has a cycle of length larger than two, then there exist an antisymmetric Q such that $P+Q$ is a stochatic matrix and $v(f, P+Q) \leq v(f, P)$. Equality holds for rare situations.

Acknowledgments. I would like to thank Professor Shigeyoshi Ogawa for the invitation and the travelling support to the workshop.

REFERENCES

Amit, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions, J. Multivariate Anal. 38, 82-99.

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

Amit, Y. and U. Grenander (1991). Comparing sweeping strategies for stochastic relaxation, J. Multivariate Anal. 37, 197-222.

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

Bogachev, V. I. , N. V. Krylov,M. Röckner (2001). On regularity of transition probabilities and invariant measures of singular diffusions under minimal conditions, Comm. Part. Diff. Equ. ,26, 2037-2080.

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

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

Chen, M. F. (2002). Ergodic convergence rates of Markov processes-eigenvalues, inequalities and ergodic theory, Vol. III, ICM 2002.

Frigessi, A. , Hwang, C. -R. and Younes, L. (1992). Optimal spectral structures of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. Ann. Appl. Probab. 2, 610-628.

Frigessi, A. , C. -R. Hwang, S. -J. Sheu, and P. di Stefano (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.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distribution, and the Baysian restoration of images, IEEE Trans. Pattn. Anal. Mach. Intell., 6, 721-741.

Gilks, W. R. and G. O. Roberts (1996). Strategies for improving MCMC, Markov Chain Monte Carlo in Practice, edited by Gilks, Richardson and Spiegelhalter, 89-114, Chapman & Hall.

Grenander, U. and M. I. Miller (1994). Representations of knowledge in complex systems, J. Roy. Statist. Soc. Ser. B, 56, 549-603.

Hwang, C. -R. and S. -J. Sheu (1992). A remark on the ergodicity of systematic sweep in stochastic relaxation, Lect. Notes Statist., 74, 199-202.

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

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.

Hwang, C. -R. , S. -Y. Hwang-Ma and S. -J. Sheu (2005). Accelerating diffusions. Ann. Appl. Probab. , 15, 1433-1444.

Hwang, C. -R. , H. -M. Pai (2005). Blowing up the spectral gap of Laplacian in T^n by antisymmetric perturbations, preprint.

Hwang, C. -R. , S. -Y. Hwang-Ma (2005). Variance reduction by antisymmetric perturbations, preprint.

Ikeda, N. and S. Watanaba (1989). Stochastic Differential Equations and Diffusion Processes, 2nd Edition, North-Holland.

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

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

Mergersen, K. L. and R. L. Tweedie (1996) Rates of convergence of the Hastings and Metropolis algorithms, Ann. Statist., 24, 101-121.

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

Roberts, G. O. and J. S. Rosenthal (2004). General state space Markov chains and MCMC algorithms, manuscript.

Röckner, M. and F. Y. Wang (2001). Weak Poincaré inequalities and L^2-convergence rates of Markov semigroups, J. Funct. Anal., 2001, 564-603.

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.

Sramer, O. and R. L. Tweedie (1997). Geometric and subgeometric convergence of diffusions with given distributions, and their discretizations, manuscript.

Trudinger, N. S. (1968). Pointwise estimates and quasilinear parabolic equations, Comm. Pure Appl. Math., 21, 205-226.

Wang, F. Y. (1999). Existence of spectral gap for elliptic operators, Ark. Mat., 37, 395-407.

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