Analog Optimization with Wong's Stochastic Hopfield Network*

George Kesidis, *Member, IEEE*

E&CE Dept, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

IEEE Trans. Neural Net., Vol. 6, No. 1, pp. 258-260, Jan. 1995

Abstract

We describe E. Wong's stochastic Hopfield network and show that it can be used, in principle, to perform analog optimization. The optimization dynamics are analogous to those of simulated annealing. To show this, we use the theory developed in Holley and Stroock [5] for the continuous-time simulated annealing process.

1 Introduction

The optimization of a function $V(x)$, $x \epsilon D$, when the dimension of the space D is large and multiple local minima exist, is a computationally difficult problem. A class of stochastic algorithms, known as simulated annealing, has been developed for the case where D is countable [5]. These algorithms have recently been extended to the case where D is the n-cube $(-1,1)^n$. This method, known as the diffusion machine, can be viewed as a continuous-state, continuous-time Hopfield network [6]. Noise is injected at each node of the network in such a way as to ensure a Gibbs stationary distribution for the configuration of states. Simulated annealing can then be achieved by a suitable cooling schedule.

Simulated annealing algorithms involve the transition of a state that is a continuous-time Markov chain x_t on D. The generator of x_t depends on

*supported by NSERC of Canada

a time-varying parameter T_t called the temperature. If T_t is constant, the time-homogeneous Markov chain will have a stationary distribution of the form:

$$\mu_{1/T}(\cdot) = \frac{\exp(-V(\cdot)/T)}{Z_{1/T}}$$

This is known as the Gibbs distribution and $Z_{1/T}$ is a normalizing constant known as the partition function. Note that as T goes to zero, the Gibbs distribution becomes a singular function entirely concentrated on the subset, D^*, of D consisting of points that globally minimize V over D. Therefore, the goal of the time-inhomogeneous Markov chain is to converge (in probability) to points in D^* only. To achieve this, the T_t must approach zero slowly enough to maintain "quasi-equilibrium." In other words, when T_t changes slowly enough, the distribution of x_t can track μ_{1/T_t}.

Another simulated annealing algorithm, called the Langevin equation, also addresses the problem of global minimization over $[-1,1]^n$ [3],[4]. Boundary conditions on V are required, however, to extend the definition of $V(x)$ to $x \epsilon \mathbf{R}^n$. Also, the state of the Langevin equation is not bounded. On the other hand, the diffusion machine is stable and requires no such boundary conditions on V.

A Hopfield network [6] with state $x_t \;\epsilon\; [-1,1]^n$ and inputs $b_t \;\epsilon\; \mathbf{R}^n$ is described by the following equations:

$$\dot{u}_t = -\nabla V(x_t) + b_t$$

$$x_t^k = \tanh(u_t^k/w)$$

Where $k = 1, 2, ..., n$ and w is a fixed positive constant. The diffusion machine is a Hopfield network with the input taken to be modulated white noise. The modulating terms are chosen in a special way so that the stationary distribution of the state of the network is Gibbs [8]. These special modulating terms are functions of the temperature parameter *and* x_t.

In this note, we prove that the state of the diffusion machine will converge so as to globally minimize its Hamiltonian when the white noise modulating terms tend to zero ($T_t \downarrow 0$) slowly enough. The methods employed by Holley and Stroock [5] to the convergence of the simulated annealing algorithm are applied. As in previous simulated annealing algorithms, a logarithmic cooling schedule will guarantee convergence to the global minimum. We conclude with a lemma which indicates that a logarithmic cooling schedule is necessary.

Because the diffusion machine is formulated as a Hopfield network, an analog realization is possible (hence the term "machine"). Of course, in an analog system, annealing would not be "simulated." An analog realization of the diffusion machine would also be faster than a simulated version and would not involve a discrete time approximation of a diffusion. An attempt to realize the diffusion machine using MOSFETs operating in their subthreshold region is described in [7]; a simulation of this circuit was conducted using SPICE.

2 Convergence of the Diffusion Machine

Consider a system described by equations

$$\begin{aligned} du_t &= -\nabla V(x_t)\, dt + \Sigma(t, x_t)\, dB_t \\ x_t &= G(u_t) \end{aligned} \tag{1}$$

where the first equation is a stochastic differential equation of the Ito type,

∇ is gradient with respect to the x variables,
$u_t \epsilon \mathbf{R}^n$, $u_t = (u_t^1, ..., u_t^n)$,
$x_t \epsilon (-1,1)^n$, $x_t = (x_t^1, ..., x_t^n)$,
B_t is n dimensional Brownian motion,
$V : [-1,1]^n \to \mathbf{R}$, $V \epsilon C^\infty$, $\sup_{[-1,1]^n} V - \inf_{[-1,1]^n} V < \infty$,
and $G : \mathbf{R}^n \to (-1,1)^n$.

G is such that $x_t^k = g(u_t^k)$ where g is a sigmoid threshold function commonly found in neural networks:

$$g(u_t^k) = \tanh(u_t^k / w) \text{ and } w > 0.$$

If $\Sigma \equiv 0$ then the two relations in (1) describe a continuous-time Hopfield network with no external inputs. E. Wong [8] showed that if

$$\Sigma(t, x_t) = \mathrm{diag}\left(\sqrt{\frac{2}{\beta_t f(x_t^1)}}, ..., \sqrt{\frac{2}{\beta_t f(x_t^n)}}\right)$$

where

$$f(y) = g'(g^{-1}(y)) = \frac{1}{w}(1 - y^2)$$

and β_t is constant $(\infty > \beta_t \equiv \beta > 0)$, then the stationary distribution of the (x_t) process is the Gibbs distribution:

$$\mu_\beta(x) = \frac{1}{Z_\beta} \exp(-\beta V(x))$$

where Z_β is the partition function.

With this Σ, the relations in (1) describe what Wong called a diffusion machine. By Ito's rule [9], the Fokker-Planck operator of the (x_t) process is:

$$L_\beta(p) \quad = \quad \operatorname{div}[A(\frac{1}{\beta}\nabla p + p\nabla V)] \tag{2}$$

where

$$A(x) = \operatorname{diag}(f(x^1), ..., f(x^n)).$$

The state of a Hopfield net $(\Sigma \equiv 0)$ will evolve so as to perform a gradient descent on V and settle in a local minimum of V [6]. As in simulated annealing, it turns out that if $\beta_t = B \log(2 + t)$, for small enough $B > 0$, then the state (x_t) of the diffusion machine will converge so as to globally minimize V. That is,

$$\lim_{t\to\infty} \text{Prob}(\ V(x_t) > \inf_{[-1,1]^n} V + \delta\ ) \quad = \quad 0, \quad \forall \delta > 0. \tag{3}$$

Let $\gamma(\beta)$ be the gap between 0 and the rest of the spectrum of L_β. Using arguments found in [5] one can show that

$$\lim_{t\to\infty} \frac{\gamma(\beta_t)}{\dot{\beta}_t} \quad = \quad \infty \tag{4}$$

is a sufficient condition for (3). See the Appendix for an outline of their approach.

For sufficiently small B, $\beta_t = B \log(2 + t)$ satisfies condition (4) by the following simple claim.

Claim

If $0 < \beta < \infty$ then there exists a $c > 0$ such that

$$\gamma(\beta) \geq \frac{c}{\beta} \exp(-2\beta M)$$

where $M = \sup V - \inf V < \infty$.
<u>Proof</u>

A straightforward computation yields [2]:

$$\gamma(\beta) \quad = \quad \inf_{\phi} \frac{\frac{2}{\beta} \int (\nabla \phi)^T A (\nabla \phi) \mu_\beta dx}{\int \int (\phi(x) - \phi(y))^2 \mu_\beta(x) \mu_\beta(y) dy dx} \tag{5}$$

subject to the constraint that ϕ is not constant.

Equivalently,

$$\gamma(\beta) = \inf_{\phi \not\equiv 0} \frac{1}{\beta} \int (\nabla \phi)^T A (\nabla \phi) \mu_\beta dx$$

subject to

$$\int \phi^2 \mu_\beta dx = 1 \text{ and } \int \phi \mu_\beta dx = 0.$$

With (2), it can be easily shown that μ_β is the *unique* stationary distribution of the (x_t) process. Thus $\gamma(\beta) > 0$.

From (5) we see that

$$\gamma(\beta) \geq \frac{c}{\beta} \frac{\inf \mu_\beta}{(\sup \mu_\beta)^2}$$

where

$$c = \inf_{\phi} \frac{2 \int (\nabla \phi)^T A (\nabla \phi) dx}{\int \int (\phi(x) - \phi(y))^2 dy dx}$$

subject to the constraint that ϕ is not constant. But,

$$\frac{\inf \mu_\beta}{(\sup \mu_\beta)^2} = \frac{Z_\beta \exp(-\beta \sup V)}{\exp(-2\beta \inf V)} = \exp(-\beta M) \int \exp(-\beta(V - \inf V)) dx \geq \exp(-2\beta M)$$

$$\Rightarrow \quad \gamma(\beta) \geq \frac{c}{\beta} \exp(-2\beta M).$$

Note that

$$\frac{c}{\beta} \frac{\sup \mu_\beta}{(\inf \mu_\beta)^2} \geq \gamma(\beta) > 0 \quad \Rightarrow \quad c > 0.$$

Q.E.D.

For $\beta_t = B\log(2+t)$ with $0 < B < 1/(2M)$:

$$\lim_{t\to\infty}\frac{\gamma(\beta_t)}{\dot{\beta}_t} \geq \lim_{t\to\infty}\frac{c}{\beta_t\dot{\beta}_t}\exp(-2\beta_t M) = \lim_{t\to\infty}\frac{c}{B^2}\frac{(2+t)^{1-2BM}}{\log(2+t)} = \infty.$$

Thus condition (4) holds, and so this logarithmic cooling schedule will guarantee convergence to a global minimum.

An exponential upper bound for $\gamma(\beta)$ would suggest that the logarithmic cooling schedule is the fastest possible to guarantee convergence to the global minimum ((4) is only sufficient). Such an upper bound was found for the case $n = 1$ in [7]. The proof of the upper bound was motivated by Holley and Stroock [5] as well.

3 Conclusions

We have shown that E. Wong's stochastic Hopfield network (or "diffusion machine") can be used, in principle, to perform analog optimization. The optimization dynamics are analogous to those of simulated annealing and require a logarithmic cooling schedule. A logarithmic cooling schedule may be impractically slow. A "quantum" simulated annealing algorithm exists that allows for a faster than logarithmic cooling schedule [1]–a quantum version of the diffusion machine is currently under study.

Appendix: A Convergence Theorem of Holley and Stroock

An outline of the proof of a convergence theorem in [5] follows. We now write $x(t) = x_t$ and $\beta(t) = \beta_t$.

Theorem

If there exists a positive, differentiable function $\beta(t)$ such that $\lim_{t\to\infty}\beta(t) = \infty$ and

$$\lim_{t\to\infty}\frac{\gamma(\beta(t))}{\dot{\beta}(t)} = \infty, \qquad (6)$$

then

$$\lim_{t\to\infty} \text{Prob}(\ V(x(t)) > \inf_{[-1,1]^n} V + \delta\ ) \ \ = \ \ 0, \ \ \forall\delta > 0. \qquad (7)$$

<u>Outline of Proof</u>

We begin by defining: $U_t(S) = \int_S \mu_{\beta(t)}(x)dx$ is the Gibbs measure, m_t is the probability density function of $x(t)$ ($\dot{m}_t = L_{\beta(t)} m_t$), $M = \sup V - \inf V$, and

$$z_t \ \ = \ \ \int_{[0,1]^n} \frac{m_t^2(x)}{\mu_{\beta(t)}(x)} dx.$$

1. Write γ in terms of the Dirichlet form of the process $x(t)$ (equation (5)).
2. By direct differentiation, show that

$$\dot{z}_t \ \ \leq \ \ (-2\gamma(\beta(t)) + M\beta(t))\, z_t + 2\gamma(\beta(t)).$$

3. By integrating, we get that the hypothesis of the theorem implies that

$$\lim_{t\to\infty} z_t \ \ \leq \ \ 1.$$

Thus, by the continuity of z_t, there exists a positive constant $C < \infty$ such that $z_t \leq C$ for all $t \geq 0$.
4. By the Cauchy-Schwarz inequality, $z_t \leq C$ implies that

$$\text{Prob}(\ V(x(t)) > \inf V + \delta\ ) \ \ \leq \ \ \sqrt{C\ U_t(x\ :\ V(x) > \inf V + \delta)}.$$

The right hand side of this inequality converges to zero when $t \to \infty$ as desired. ♠

References

[1] B. Appoloni, C. Carvalho, and D. deFalco. Quantum stochastic optimization. *Stochastic Processes and their Applications*, Vol. 33, No. 2: pp. 233-244, 1989.

[2] H. Brand, A. Schenzle, and G. Schroder. Lower and upper bounds for the eigenvalues of the Fokker-Planck operator in detailed balance. *Physical Review A*, Vol. 25, No. 4: pp. 2324-2338, April 1982.

[3] S. Geman and C.Hwang. Diffusions for global optimization. *SIAM J. Control and Optimization,* Vol. 24, No. 5: pp. 1031-1043, Sept. 1986.

[4] B.Gidas. Global optimization via the Langevin equation. *Proc. of the 24th Conf. on Decision and Control, Ft. Lauderdale, Fl.,* pp. 774-778, Dec 1985,

[5] R.Holley and D.Stroock. Simulated Annealing via Sobolev inequalities. *Commun. in Math. Phys.*, No. 115: pp. 553-569, 1988.

[6] J.J.Hopfield. Neural networks and physical systems with emergent computational abilities. *Proc. Natl Acad. of Sci. USA,* Vol. 79: pp. 2554-2558, April 1982.

[7] G.Kesidis. Topics in Stochastic Neural Networks: Circuit Simulation and Cooling Schedule. Masters Report, EECS Dept, University of California at Berkeley, 1990.

[8] E.Wong. Stochastic Neural Networks. ERL Memorandum No.UCB/ERL M89/9, University of California at Berkeley, 1989.

[9] E.Wong and B.Hajek. *Stochastic Processes in Engineering Systems.* Springer-Verlag, New York, 1985.