BULLETIN OF THE
INSTITUTE OF MATHEMATICS
ACADEMIA SINICA
Volume 20, Number 1, March 1992

ON OCCUPATION TIMES OF ANNEALING PROCESSES

BY

YUNSHYONG CHOW (周雲雄) AND JUNE HSIEH (謝 仲)

Abstract. Simulated annealing is a probabilistic optimization algorithm for finding the global minimum (or minima) of a certain function on a finite set S. It is known that the algorithm converges weakly to a certain distribution $(\beta_i)_{i\in S}$ concentrated on the global minima set $\underline{S}$. In this paper we show that the fraction of the time spent at each state i converges a.e. to β_i.

1. Introduction. Many combinatorial optimization problems can be formulated as finding the maximum of a certain cost function U over a finite state space S, say $S = \{1, 2, \ldots, N\}$. The popular gradient method converges fast. But the limiting state depends on the initial state and is usualy only a local minimum. Based on ideas from statistical physics, Kirkpatrick et al [13] and Černy [3] proposed a probabilistic optimization algorithm: simulated annealing. It can be described as a (continuous or discrete) time inhomogeneous Markov process $\{X(t) : t \in R^+ \text{or } I^+\}$ with stat space S and transition rates $q_{ij}(t)$ $(= \lim_{\Delta t \downarrow 0}\{P(X(t+\Delta t) = j | X(t) = i) - \delta_{ij}\}/\Delta t$ for $t \in R^+$; $= P(X(t+1) = j | X(t) = i)$ for $t \in I^+)$ of the following type

$$q_{ij}(t) = \begin{cases} p(i,j)\exp[-(U(j) - U(i))^+/T(t)], & \text{for } j \neq i \\ 1_{\{discretetime\}} - \sum_{k\neq i} q_{ik}(t), & \text{for } j = i, \end{cases} \tag{1.1}$$

where $T(t)$ is the "temperature" function converging to 0 as time t goes

Received by the editors April 16, 1991 and in revised form Novermber 18, 1991.
Subject Classification: Primary 60J27, Secondary 60F15, 90C15.
Key Words: Simulated annealing, inhomogeneous Markov process, strong convergence, occupation times.

to infinity, and $P = (p(i,j))$ is a nonnegative matrix which determines, in particular, the neighborhood structure on S. For example, $p(i,j) = 0$ iff $j \notin N(i)$, i.e., j is not a neighbor of i.

The difference between the gradient method and simulated annealing lies in the definition of $q_{ij}(t)$ for $j \in N(i)$ and $U(j) > U(i)$. In the former case, $q_{ij}(t) = 0$ so that the process is in fact time–homogeneous and all (local or global) minima are essentially absorbing states. In the latter case, $q_{ij}(t) = p(i,j)[\exp -1/T(t)]^{U(j)-U(i)}$ so that "up-hill" movements are allowed, though not encouraged. If the accumulated effect is sufficiently strong, one might be able to escape from the local minima and thus, reach the global minima. The crucial issue is how to choose the cooling schedule $T(t)$.

Applying simulated annealing to image restoration problems, Geman and Geman [8] first showed, in the case they considered, that if $T(t) = c/\log(t+1)$ and the constant c is sufficiently large, then $\lim_{t\to\infty} P(X(t) \in \underline{S}) = 1$, where $\underline{S} = \{i \in S : U(i) = \min U\}$ is the global minima set. Since then, there has been a wide interest in the applications and convergence of simulated annealing. Some of these works are listed in the reference of this paper. The reader can find more from the books cited there.

Throughout the paper, $\{X(t)\}$ is assumed to be irreducible and weakly reversible. The readers are referred to Chiang and Chow [4,5] for all the terminologies used in the paper. In particular, one can find there the definitions and physical interpretations as active energy of two depth constants d_H and d_V. Note that by definition $d_H \le d_V$ and equality holds if $|\underline{S}| = 1$.

Let $\lambda(t) = \exp(-1/T(t))$. Besides tending to 0, $\lambda(t)$ is required to satisfy some regularity conditions. In the following E is a nonnegative parameter.

(A.1; E) $$\int_0^\infty (\lambda(t))^E dt \left[\text{or} \sum_{t=0}^{\infty} (\lambda(t))^E\right] = \infty.$$

(A.2; E) $$\lambda'(t) \left[\text{or } \lambda(t+1) - \lambda(t)\right] = o((\lambda(t))^{E+1}).$$

(A.3; E)

There are constants $u \leq v < 1$ such that $t^{-v} \leq (\lambda(t))^E \leq t^{-u}$ for t large.

$$\text{(A.4)} \qquad \int_0^\infty (\lambda(t))^p dt \left[\text{or } \sum_{t=0}^{\infty} (\lambda(t))^p\right] < \infty \text{ for some } p > 0.$$

Note that (A.3; E) implies (A.1; E) and (A.4). (A.4) is a reasonable condition in the sense that the rate of convergence (see Theorem 1 below) will not be too slow. On the other hand, in order to satisfy (A.1; E), $\lambda(t)$ cannot converge to 0 too fast.

By analyzing the Kolmogorov's forward equation for $P(X(t) = i|X(t_0))$, it can be shown that whatever the initial distribution $X(0)$ is, the following results hold.

Theorem 1. (i) *Assume* (A.1; d_H) *and* (A.2; d_H). *Then for* t *large* $P(X(t) \in \underline{S}|X(0)) = 1 + O(\lambda^a(t))$, *where* $a = \min_{i \notin \underline{S}} U(i) - \min U$.

(ii) *Assume* (A.1; d_V) *and* (A.2; d_H). *Then there are constats* $\beta_i > 0$ *such that for* $i \in S$ *and* t *large*

$$\text{(1.2)} \qquad P(X(t) = i|X(0)) = [\beta_i + o(1)](\lambda(t))^{U(i) - \min U}.$$

(iii) *Assume* (A.1; d_V) *and* (A.2; d_V). *Then for* $i \in S$ *and* t *large*

$$\text{(1.3)} \qquad P(X(t) = i|X(0)) = [\beta_i + O(\lambda^a(t))](\lambda(t))^{U(i) - \min U}.$$

(iv) *Assume* (A.1; d_V), (A.2; d_V) *and* (A.3; d_V). *Then there are constants* $\delta < 1$, C *and* t_1 *such that for* $t_2 \geq t_1$, $t \geq t_2 + t_2^\delta$ *and* $i \in S$

$$\text{(1.4)} \qquad |\{P(X(t) = i|X(t_2))/(\lambda(t))^{U(i) - \min U}\} - \beta_i| \leq C\lambda^a(t).$$

Theorem 1 (i) and (ii) can be found in Chiang and Chow [4, 5]. Theorem 1 (iii) and (iv) are new. Both give the exact error estimates of $\{X(t)\}$ to its limit distribution. However, Theorem 1 (iv) offers an estimate of the waiting–time to reach the limit distribution, which is needed later.

Theorem 1 (ii) implies that there is an ergodic distribution concentrated on $\underline{S}$. In the theory of time–homogeneous Markov processes, there are also almost sure covergence results. This motivates us to study the sample path

properties of simulated annealing, which, if they hold, would be helpful for running computer simulations in its applications.

Let $\gamma_i(t) = \int_0^t 1_{\{X(s)=i\}} ds$ (or $\sum_{s=0}^t 1_{\{X(s)=i\}}$ in the discrete time case) be the occupation time up to t at state i.

Theorem 2. (i) *Assume* (A.1; d_H), (A.2; d_H) *and* (A.4). *Then w.p.1,* $\lim_{t\to\infty}\{\sum_{i\notin \underline{S}} \gamma_i(t)\}/t = 0$. *Hence* $\lim_{t\to\infty}\{\sum_{i\in \underline{S}} \gamma_i(t)\}/t = 1$ *a.e..*
(ii) *Assume* (A.1; d_V), (A.2; d_V) *and* (A.3; d_V). *Then w.p.1,*

$$\lim_{t\to\infty} \gamma_i(t)/t = \beta_i \text{ for } i \in \underline{S}. \tag{1.5}$$

In particular, it is easy to check that Theorem 2 holds for the commonly used cooling schedule $T(t) = c/\log(t+1)$ with $c > d_V$. Note that in this case $\lambda(t) = (t+1)^{-1/c}$.

Theorems 1, 2 will be proved in Sections 2, 3 respectively. In the following we give some remarks.

Remark 1. In the theorey of time–homogeneous Markov processes Theorem 2(ii) is proved through the use of interarrival times, which form an i.i.d. sequence. When the time–homogeneous process undergoes a small perturbation, the same method should also work. But it seems difficult to be applied to simulated annealing. Because simulated annealing is sort of a singular perturbation in the sense that the limit of $(q_{ij}(t))$ in (1.1) is not the generator of an ergodic Markov process.

Remark 2. Theorem 2 has a similar result when the exponent $(U(j) - U(i))^+$ in (1.1) is replaced by a general cost function $U(i,j)$ which is defined from $S \times S$ to $[0,\infty]$. This is because (1.2) holds in this case. See Chiang and Chow [5].

Remark 3. When the state space S is a continuum, say $[0,1]$ or $\mathcal{R}^n$, a stochastic differential equation $dX(t) = -\bigtriangledown U(X(t))dt + \sqrt{2T(t)}dW_t$ is proposed to find the global minima. See Geman and Hwang [9], Hwang and Sheu [12] for details.

2. Proof of Theorem 1. Parts (i) and (ii) can be found in Chiang and Chow [4, 5]. The estimate (1.2) for each state cannot be obtained in one step, but through successive improvements of order $O(\lambda(t))$, $O(\lambda^2(t)), \ldots,$ interwoven with the merging procedures. In each step the following lemma is used one or several times with $f(t)$ as a linear combination of some $P(X(t) = i|X(0))$.

Lemma 2.1 *Let $f(t)$ be a complex-valued function and α a complex number with* Re $\alpha > 0$. *Suppose*

$$f'(t)[\text{or } f(t+1) - f(t)] = -\alpha\lambda^E(t)f(t) + \Delta\lambda^F(t), \tag{2.1}$$

where $\Delta = o(1)$ or $O(\lambda(t))$. Then $f(t) = \Delta\lambda^{F-E}(t)$ as $t \to \infty$, if

$$\begin{cases} (i) & \text{(A.1; } E\text{) and (A.2; } E\text{) hold, or} \\ (ii) & \Delta = o(1),\ F = E \text{ and (A.1; } E\text{) holds.} \end{cases} \tag{2.2}$$

Note that only in the last step (2.2)(ii) is applied with $E = d_H$ to obtain the final constant $\beta_i + o(1)$ in (1.2). The proof of part (iii) is the same as that in part (ii), except that in the last step we apply instead the stronger assumptions (A.1; d_V) and (A.2; d_V) to Lemma 2.1 to obtain the desired better estimate in (1.3). We remark that under (A.1; d_V) and (A.2; d_V), Lemma 2.1 is used for $E \le d_V$ and through (2.2) (ii) only.

To prove part (iv) we first observe that starting from a different time t_0 does not affect the sequence of order improvements nor the merging procedures. In other words, Lemma 2.1 is used in the same order as starting from $t = 0$. Therefore, we need to examine more carefully its proof.

For brevity we consider only $t \in R^+$. The discrete time case can be treated similarly. Since (2.2)(i) will never be applied under the present assumptions, $\Delta = O(\lambda)$ always. Denote the error term $\Delta\lambda^F(t)$ in (2.1) by $b(t)$. Suppose we know

$$|b(t)| \le C\lambda^{F+1}(t). \tag{2.3}$$

Here and after, C will be a generic constant depending only on $(p(i,j))$ and $U(\cdot)$. Clearly,

$$(2.4)\qquad f(t) = \left[\exp - \int_{t_2}^{t} \alpha\lambda^E(s)ds\right]\left\{f(t_2) + \int_{t_2}^{t} b(s)\left[\exp\int_{t_2}^{s} \alpha\lambda^E(\omega)d\omega\right]ds\right\}$$
$$= I_1 + I_2.$$

Let $\theta = \operatorname{Re}\alpha$ and $I(t) = \exp\int_{t_2}^{t} \theta\lambda^E(s)ds$. First taking the absolute value and then integrating by parts, $I(t)\cdot I_2$ is bounded by

$$\int_{t_2}^{t} C\lambda^{F+1}(s)\cdot I(s)$$
$$= C\theta^{-1}\left\{\lambda^{F+1-E}(s)I(s)\Big|_{t_2}^{t} - \int_{t_2}^{t}(F+1-E)\lambda^{F+1-E}(\lambda'/\lambda)I(s)ds\right\}.$$

Since Lemma 2.1 is used only for $E \le d_V$ and t_2 can be assumed large, we have by (A.2; d_V) that $\theta^{-1}(F+1-E)|\lambda'/\lambda| \le \lambda^E/2$ on $[t_2,\infty)$. The second term on the right-hand side of the equation above will be then of the same form as the left-hand side. A rearrangement shows

$$(2.5)\qquad |I_2| \le \quad 2C\lambda^{F+1-E}(t).$$

Let $g(t,t_2) = \lambda^{F+1-E}(t)\cdot I(t)$. By differentiating,

(2.6) $g(t,t_2)$ is increasing in t if t_2 is fixed and large enough.

Remember $E \le d_V$. If $1 > \delta > v$, a simple computation shows that under (A.3; d_V)

$$(2.7)\qquad g(t_2+t_2^{\delta},t_2) \ge (2t_2)^{-v(F+1-E)/d_V}[\exp(t_2^{\delta}\cdot\theta\cdot(2t_2)^{-v})] \ge C' > 0$$

if t_2 is away from 0. As a linear combination of some $P(X(t_2) = i|X(t_0))$, $f(t_2)$ is always bounded by some constant. Combining together with (2.6) and (2.7), we get from (2.4) that (2.5) holds for I_1 too, as soon as $t \ge t_2 + t_2^{\delta}$. Hence we have shown that for any $1 > \delta > v$, there exists t_1 such that the solution to (2.1) with an arbitrary initial $f(t_2)$ satisfies

$$(2.8)\qquad |f(t)| \le C\lambda^{F+1-E}(t) \text{ if } t_2 \ge t_1 \text{ and } t \ge t_2 + t_2^{\delta}.$$

It remains to check (2.3). When Lemma 2.1 is applied in the first step to those states with minimum out-going cost 0 but not in any 0th-order

cycle, $E = F = 0$ and $b(t)$ is of the form $\sum C_i \lambda^{u_i}(t) P(X(t) = i | X(t_0))$ with $u_i \geq 1$. Hence we have (2.3) automatically and then (2.8) holds if the initial time $t_0 \geq t_1$. The 0th–order cycles will be merged at the next step, in which $E = F = 0$ still, but $b(t)$ gets an extra contribution from the states just treated. By using (2.8) just obtained, (2.3) holds if $t_0 \geq t_1 + t_1^\delta$. As before (2.8) can be proved for this step, except that we need a new, larger triple (t_1, t_2, C). Repeating the previous procedure as many times as needed, we eventually have (2.3) and then (2.8) for each step. This completes the proof of part (iv).

3. Proof of Theorem 2. The basic idea of the proof is the same as in Chung [6; Theorem 5.1.2]. For brevity we consider $t \in R^+$ only. The discrete time case can be treated similarly. Let $J(t) = \int_0^t \lambda^a(s) ds$.

Part (i). Let $\gamma(t) = \sum_{i \notin \underline{S}} \gamma_i(t)$ and $B = \int_0^\infty \lambda^p(s) ds$, where p is given in (A.4). By using Theorem 1(i) and the Hölder inequality,

$$E\gamma(t) \leq CJ(t) \leq \begin{cases} CB, & \text{if } a \geq p, \\ CB^{a/p} t^{1-a/p}, & \text{if } a < p. \end{cases}$$

Then $\sum E\{\gamma(t_n)/t_n\} < \infty$ for $t_n = n^{2p/a}$ if $a < p$, $= n^2$ if $a \geq p$. Hence $\sum \gamma(t_n)/t_n < \infty$ a.e. and in particular, $\lim \gamma(t_n)/t_n = 0$ a.e.. Because $\gamma(t_n) \leq \gamma(t) \leq \gamma(t_{n+1})$ for $t_n \leq t \leq t_{n+1}$ and $\lim t_{n+1}/t_n = 1$, the conclusion follows easily.

Part(ii). Apparently, we may assume $\gamma_i(t)$ counts from t_1 instead of 0, where t_1 is given in Theorem 1(iv). Write

$$\begin{aligned} E\{\gamma_i^2(t)\} &= E\left\{ \int_{t_1}^t 1_{\{X(s)=i\}} ds \right\}^2 \\ &= 2 \int_{t_1}^t ds \int_s^t P(X(s) = X(\omega) = i | X(t_1)) d\omega. \end{aligned} \tag{3.1}$$

The integrand above can be written as $P(X(s) = i | X(t_1)) \cdot P(X(\omega) = i | X(s) = i) \equiv h(s, \omega)$. It follows from Theorem 1(iii) and (iv) that $|h(s,\omega) - \beta_i^2| \leq C(\lambda^a(s) + \lambda^a(\omega))$ if $s + s^\delta < \omega$. Otherwise, use the trivial bound 1 for $h(s,\omega)$. A simple computation shows that, increasing the constant C,

$$|E(\gamma_i^2(t)) - (t-t_1)^2\beta_i^2| \le C[t^{1+\delta} + tJ(t)]. \tag{3.2}$$

Similarly $|E(\gamma_i(t)) - (t-t_1)\beta_i| \le CJ(t)$. Hence (increasing C)

$$E[(\gamma_i(t)/t) - \beta_i]^2 \le C\{t^{\delta-1} + t^{-1} + J(t)/t\}.$$

Since $\delta < 1$, (1.5) can be proved as did in part(i). The detail is omitted.

References

1. E. Arts and J. Korst, *Simulated Annealing and Boltzmann Machines*, Wiley, New York, 1989.

2. R. Azencott, *Simulated Annealing*, Seminaire Bourbaki, 1987–88, no. 697.

3. V. Černy, *Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm*, J. Opt. Theory Appl. **45** (1985), 41–51.

4. T. S. Chiang and Y. Chow, *On the convergence rate of annealing processes*, SIAM J. Control Optim. **26** (1988), 1455–1470.

5. T. S. Chiang and Y. Chow, *A limit theorem for a class of inhomogeneous Markov processes*, Ann. Probab. **17** (1989), 1483–1502.

6. K. L. Chung, *A Course in Probability Theory, 2nd Ed.*, Academic Press, New York, 1974.

7. L. Davis, *Genetic Algorithm and Simulated Annealing*, Pitman, London, 1987.

8. D. Geman and S. Geman, *Stochastic relaxation, Gibbs distribution and the Bayesion restoration of images*, IEEE PAMI **6** (1984), 721–741.

9. S. Geman and C. R. Hwang, *Diffusions for global optimization*, SIAM J. Control Optim. **24** (1986), 1031–1043.

10. B. Gidas, *Global optimization via the Langevin equation*, Proc. 24th IEEE Conf. Decision and Control, Fort Lauderdale, Florida, 1985, 774–778.

11. B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res. **13** (1988), 311–329.

12. C. R. Hwang and S. J. Sheu, *Large time behaviors of perturbed diffusion Markov processes with applications III: Simulated annealing*, preprint, 1986.

13. S. Kirkpatrick, C. Gelatt and M. Vecchi, *Optimizations by simulated annealing*, Science **220** (1983), 671–680.

14. P. J. M. van Laarhoven and E. Arts, *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht, 1987.

Institute of Mathematics, Academia Sinica, Taipei, Taiwan 115, R.O.C.