Some Properties of Nonlinear MAP Estimation by Simulated Annealing

A. Kehagias
Division of Applied Mathematics
Brown University
Providence, RI 02912
E-Mail Address: kehagias@brown.csc.edu

October 11, 2001

Abstract

A Simulated Annealing method is presented for the solution of nonlinear time series estimation problems, by maximization of the a Posteriori Likelihood function. Homogeneous temperature annealing is proposed for smoothing problems and inhomogeneous temperature annealing for filtering problems. Both methods of annealing guarantee convergence to the Maximum A Posteriori Likelihood (MAP) estimate. Entropy change with temperature provides a heuristic evaluation of speed of convergence.

1 Introduction

The problem addressed in this paper is the following: Consider a discrete time stochastic dynamic system with state vector x_n taking values in R^d, obeying the equation

$$x_n = f(x_{n-1}) + u_n, \tag{1}$$

and observation vector y_n taking values in R^c

$$y_n = g(x_n) + v_n. \tag{2}$$

Here, $u_n, v_n, \;\; n = 1, 2, ...$ are zero mean, white and independent of each other stochastic processes, with probability laws $p_{u_n}(u), \; p_{v_n}(v)$ respectively. Furthermore, we assume a random initial state x_0, with probability law $p_{x_0}(x)$, and independent of u_n, v_n.

Let $X^N = [x_0, ..., x_N], \;\; Y^N = [y_1, ..., y_N]$. The Maximum A Posteriori Estimation problem is: given Y^N, find vectors $Z \in R^{d \cdot N}, z \in R^d$ such that for given Y^N,

$$p_{X^N}(Z \mid Y_N), \;\; p_{x_N}(z \mid Y_N)$$

are maximized. (Here the vertical bars indicate conditional probabilities.) Maximization of the first probability is the smoothing problem, maximization of the second probability is the filtering problem.

2 Notation - Conventions

In general lowercase subscripted letters refer to vectors in R^d, e.g. x_n. The same letters in uppercase refer to vectors made up by a sequence of R^d-valued vectors. E.g. $X^n = [x_1, ..., x_n]$. I use the following naming convention: an **estimator** is a stochastic process (**s.p.**, denoted by a doubly subscripted letter, e.g. $X^{t,n}$, $x_{t,n}$) and an **estimate** is a random variable (**r.v.**, e.g. $\hat{X}^{m,n}, \hat{x}_{m,n}$). For example, we could have the **estimator** converging to an **estimate**: $\lim_{t\to t_m} X^{t,n} = \hat{X}^{m,n}$. This is denoted by using the same letter for both, with a hat on top for the r.v. We have two levels of time here: time indicated by $m, n, n+1, ...$ is the time in which the system (1)-(2) evolves; time indicated by t is the time in which the estimation algorithm runs. I assume that all probability distributions in this paper, except possibly the one on the global minima of H_n (see, the following sections), do not concentrate on only one point. In other words, $p(X) > 0$ for at least two distinct values of X. I also assume that the log-likelihood function H^n is nonnegative.

3 Smoothing - Preliminaries

Consider a fixed number of observations, n. Note that

$$p(X^n \mid Y^n) = \frac{p(Y^n \mid X^n)p(X^n)}{p(Y^n)}, \tag{3}$$

$$p(X^n) = p(x_n \mid x_{n-1})p(x_{n-1} \mid x_{n-2})...p(x_1 \mid x_0)p(x_0) =$$
$$p_{u_n}(x_n - f(x_{n-1}))p_{u_{n-1}}(x_{n-1} - f(x_{n-2}))...p_{u_1}(x_1 - f(x_0))p_{x_0}(x_0), \tag{4}$$

$$p(Y^n \mid X^n) = p_{v_n, v_{n-1}, ..., v_1}(y_n - g(x_n), ..., y_1 - g(x_1)) =$$
$$p_{v_n}(y_n - g(x_n))p_{v_{n-1}}(y_{n-1} - g(x_{n-1}))...p_{v_1}(y_1 - g(x_1)). \tag{5}$$

Hence, for given Y^n, maximization of $p(X^n \mid Y^n)$ is equivalent to maximization of $H^{*,n}(X^n) \doteq p(Y^n \mid X^n)p(X^n)$. But, we have

$$H^{*,n}(X^n) = p_{u_n}(x_n - f(x_{n-1}))p_{u_{n-1}}(x_{n-1} - f(x_{n-2}))...p_{x_0}(x_0) \cdot$$
$$p_{v_n}(y_n - g(x_n))...p_{v_1}(y_1 - g(x_1)). \tag{6}$$

Define $H^n(X^n) \doteq -log\ H^{*,n}(X^n)$: the negative log-likelihood function of X^n. Maximization of $H^{*,n}(X^n)$ is equivalent to minimization of $H^n(X^n)$. By defining functions $U_0(.) \doteq -log\ p_{x_0}(.), U_m(.) \doteq -log p_{u_m}(.), V_m(.) \doteq -log\ p_{v_m}(.)$, we can write

$$H^n(X^n) = U_0(x_0) + \sum_{m=1}^{n} [U_m(x_m - f(x_{m-1})) + V_m(y_m - g(x_m))]. \tag{7}$$

Now, take the usual aproximation of the continuous valued variables x_m by discrete variables x_m, taking values $x_m \in \Lambda = \{\lambda_1, ..., \lambda_L\}$. Assume the approximation is good enough that minimization of $H^n(X^n)$ over $R^{d \cdot n}$ is equivalent to minimization over $\Omega = \Lambda^n$. Define $J = \mid \Omega \mid$.

4 Smoothing by Simulated Annealing

The method of simulated annealing was introduced by Kirkpatrick [4] for combinatorial optimization problems. Geman and Geman [3] proved important convergence results and used the method for image restoration. The present work is very much in line with their paper.

The function $H^n(X^n)$ has a number of global minima $X_1^n, ..., X_K^n \in R^{d \cdot n}$. The smoothing problem is to find one of them. Simulated annealing solves the smoothing problem in the following way: we will consider a stochastic process $X^{t,n}$ which will be constructed in such a way as to ensure that the estimator $X^{t,n}$ will converge to one of the minima of H^n. More exactly, define the random variable $\hat{X}^n$ which takes values in the set $\{X_1^n, ..., X_K^n\}$ with a uniform probability distribution. Then we will construct $X^{t,n}$ in such a way that it converges in distribution to $\hat{X}^n$, as $t \to \infty$.

To achieve this convergence, perform the following algorithm:

Homogeneous Annealing Algorithm

Given a positive function of time, $T(t)$,
and a random variable $\hat{Z}$, taking values in Ω,

choose $X^{0,n}$ arbitrarily and, given $X^{t,n}$,
choose $X^{t+1,n}$ in the following way:

Pick $\hat{Z} \in \Omega$ randomly and, given $T(t)$, compute q:

$$q \doteq \frac{exp(-\frac{H^n(\hat{Z})}{T(t)})}{exp(-\frac{H^n(X^{t,n})}{T(t)})}. \tag{8}$$

If $q \geq 1$ $X^{t+1,n} = \hat{Z}$ with probability 1.
If $q < 1$ $X^{t+1,n} = \hat{Z}$ with probability q (else $X^{t+1,n} = X^{t,n}$).
Continue for the next t.

In this way we generate a stochastic process $X^{t,n}$. It is easy to prove (by similar arguments as the ones used in Geman and Geman [3] for the "Gibbs Sampler", a slightly different algorithm) that to have $X^{t,n} \to \hat{X}^n$, it is sufficient that two conditions are satisfied. The first condition has to do with the probability law of $\hat{Z}$ and further details can be found in Geman and Geman [3]. The second condition is on $T(t)$, the "temperature" parameter: $lim\, T(t)_{t\to\infty} = 0$ at a slow enough pace - specifically

$$T(t) \geq \frac{K}{logt}. \tag{9}$$

If the above conditions hold, then

$$\lim_{t\to\infty} X^{t,n} = \hat{X}^n$$

(the limit is in the distribution sense).

Here, let us consider only annealing schedules that are piecewise constant:

$$T(t) = \begin{cases} T_1 & t_0 \le t \le t_1 \\ T_2 & t_1 \le t \le t_2 \\ ... & ... \\ T_n & t_{n-1} \le t \le t_n \end{cases}$$

and respect the logarithmic bound of eq.(9).

We know [3] that for $t_1 - t_0,\ t_2 - t_1, ...$ big enough the stochastic process $X^{t,n} \to \hat{X}^{m,n}$ (for $t \to t_m$) where $\hat{X}^{m,n}$ is a random variable (the estimate) with

$$\pi_i^m \doteq p(X^m = X_i) = \frac{exp(-\frac{1}{T_m} H^n(X_i))}{\sum_{i=1}^J exp(-\frac{1}{T_m} H^n(X_i)}) \tag{10}$$

(The quantity $Z(T) \doteq \sum_{i=1}^J exp(-\frac{1}{T} H^n(X_i))$ is called the partition function.)

So, in the limit $t \to t_m$, the estimate has a distribution $\pi^m = \{\pi_1, ..., \pi_j\}$ and, accordingly, an average log-likelihood

$$E(-H^n(\hat{X}^{m,n})) = G_m. \tag{11}$$

Define

$$G(t) \doteq G_m \quad for \quad t_{m-1} \le t \le t_m. \tag{14}$$

Now, Jaynes [1,2] proves that the equilibrium distribution π^m is the maximum entropy distribution of all distributions satisfying (11). Obviously, G_m in (11) is a measure of the likelihood of $\hat{X}^{m,n}$ (as an estimate of X^n). For a given temperature we pick the max entropy estimate subject to fixed average likelihood. The max entropy estimate is the most honest to choose (see [2]). However, we actually want an estimate with small entropy This is resolved, as temperature decreases. When $T \downarrow 0$, $E(-H^n) \uparrow$. This is true because

$$G(t) = \frac{d}{d(-1/T(t))} (log \sum_{i=1}^J exp(-\frac{1}{T(t)} H^n(X_i)) \tag{13}$$

$$\frac{dG}{dT} = \frac{1}{T^2} \cdot (E(H^n)^2 - E((H^n)^2)) \le 0. \tag{14}$$

And secondly, for the entropy, defined by $S \doteq -\sum_{i=1}^J \pi_i log\ \pi_i$, we have

$$S(t) = \sum_{i=1}^J \frac{1}{T(t)} H^n(X_i) \frac{exp(-\frac{1}{T(t)} H^n(X_i))}{Z(T(t))} \ - \ logZ(T(t)), \tag{15}$$

$$\frac{dS}{dT} = \frac{1}{T^3} \cdot (E((H^n)^2) - E(H^n)^2) \tag{16}$$

At every temperature step ($t_{m-1} \le t \le t_m$), we effectively fix a likelihood level G_m and pick the maximum entropy estimate with such likelihood. When going to a lower temperature, we increase the likelihood ($T_m \downarrow$, so $G_m \uparrow$) and decrease the maximum entropy. This is then a minimax procedure on the entropy, with a simultaneous increase of the likelihood of the estimator. In the limit of $T(t) \to 0$ we get a very likely estimate with a low entropy.

One might object that the average value of G_m does not yield much information about what the actual value of G_m is. However, we can extract more information using Markov's inequality. Set $p \doteq P(\hat{X}^{m,n} \mid Y^n)$. Then

$$P(-log\ p \geq \alpha) \leq \frac{1}{\alpha} E(-log\ p) = \frac{1}{\alpha} G_m \Rightarrow$$

$$P(e^{-log\ p} \geq e^{\alpha}) \leq \frac{1}{\alpha} G_m \Rightarrow$$

$$P(P(\hat{X}^{m,n} \mid Y^n) \leq e^{-\alpha}) \leq \frac{1}{\alpha} G_m. \tag{17}$$

5 Filtering

For filtering we would like to maximize $p_{x_n}(z \mid Y^n)$. But this task is not well suited to the Simulated Annealing method. The power of linear systems filtering (see Kalman [5]) is in that it can be done in a recursive way. But this is not true of Simulated Annealing. On the other hand SA is well suited to maximizing $p_{X^n}(Z \mid Y^n)$. This maximization yields more information, but is also computationally more intensive. Ideally, we would like to use the basic idea of recursive filtering, that is, to use some of the previous work in the next step, but retain the Simulated Annealing context, which makes treatment of nonlinear problems as easy as that of linear ones. Another complication is that for $m_n \doteq min\ H^n(X^{n+1})$, we will generally have $m_n \to \infty$, and no useful estimates of the form of eq.(17) can be found.

To overcome the above problems, we will use a trick that is usual in filtering theory: Define $W_m(x_m, x_{m-1}) \doteq U_m(x_m, x_{m-1}) + V_m(x_m, x_{m-1})$ if $m > 0$ and $W_m(x_m, x_{m-1}) \doteq U_m(x_m)$ for $m = 0$ (where U_m, V_m refer to eq.(..)). Define

$$H^n(X^{t,n}) \doteq \sum_{l=1}^{n} W_l(x_{t,l}, x_{t,l-1}). \tag{18}$$

(This is the H^n function of the previous sections with $X^{t,n}$ as argument.) Define

$$\mathcal{H}^n_{\lambda(m)}(X^{t,n}) \doteq \sum_{l=1}^{n} \lambda^l(m) W_l(x_{t,l}, x_{t,l-1}) \tag{19}.$$

Define the functions $I^{n,q}, J^{n,q}, \mathcal{I}^{n,q}_{\lambda(m)}, \mathcal{J}^{n,q}_{\lambda(m)}$ as follows:

$$I^{n,q} \doteq \sum_{l=1}^{q} W_l, \quad J^{n,q} \doteq \sum_{l=q+1}^{n} W_l \quad \mathcal{I}^{n,q}_{\lambda(m)} \doteq \sum_{l=1}^{q} \lambda(m)^l W_l, \quad \mathcal{J}^{n,q}_{\lambda(m)} \doteq \sum_{l=q+1}^{n} \lambda(m)^{l-q} W_l. \tag{20}$$

Also we will make the following assumptions about these functions:

Assumptions:

1. $\exists\ m_1,\ m_2\ \ s.t.\ \ \forall Z \in \Omega,\ \forall\ l\ \ m_1 \leq W_l(Z) \leq m_2$
2. $\exists\ k_1,\ q\ \lambda(m)\ \ \ s.t.\ \ \forall\ n\ \ E(\mathcal{I}^{n,q}_{\lambda(m)}) \geq n \cdot k_1 \cdot E(\mathcal{J}^{n,q}_{\lambda(m)})$.

3. $\exists\, k_2,\ q,\ \lambda(m)\quad s.t.\ \forall\, n\ \ E((\mathcal{I}^{n,q}_{\lambda(m)})^2) \geq n \cdot k_2 \cdot E((\mathcal{J}^{n,q}_{\lambda(m)})^2)$.
4. $\exists\, \epsilon\ \ s.t.\ \forall\, n\ q\ \lambda(m)\ \ E((\mathcal{I}^{n,q}_{\lambda(m)})^2) \geq \epsilon \cdot E(\mathcal{I}^{n,q}_{\lambda(m)})^2$

(The above assumptions will certainly hold for any practical case and are in fact easily derivable from the boundedness of W_l, the finiteness of the state space Ω and the nondegeneracy of the probability distributions. They are listed explicitly only for convenience.)

It is easy to see that $\forall$ fixed $m,\ n,\ \lambda(m) < 1$ (by Assumption 1) $\mathcal{H}^n_{\lambda(m)} < M < \infty\ \ \forall\ n,\ m$. Also, it is easy to check that $\{\mathcal{H}^n_{\lambda(m)}(X^{t,n})\}_{t=1}^{t_m}$ is a **submartingale** with respect to the σ-algebra generated by $\{X^{t,n}\}_{t=1}^{t_m}$. Hence it is a bounded submartingale (for fixed n) and so converges w.p. 1 to some r.v., call it $\hat{M}_n$ [1] . However, $lim\ _{t\to t_m}\mathcal{H}^n_{\lambda(m)}(X^{t,n}) = \mathcal{H}^n_{\lambda(m)}(\hat{X}^{m,n})$ weakly. So it follows that $lim\ _{t\to t_m}\mathcal{H}^n_{\lambda(m)}(X^{t,n}) = \mathcal{H}^n_{\lambda(m)}(\hat{X}^{m,n})$ w.p. 1.

Why bother to change the likelihood function? Apart from the convenience of introducing a submartingale, there is a heuristic justification and computational advantages to be gained. Also, it will be proven that, by taking an appropriate limiting procedure, we can recover the solution to the original problem.

For the purpose of this discussion, use the following terminology: At time q we have an estimate $\hat{X}^q$ and at time n an estimate $\hat{X}^n = [\bar{X}^n_q, \tilde{X}^n_q]$. We will call $\bar{X}^n_q$ (which is a $q \cdot d$ vector) the "first part" of $\hat{X}^n$ and $\tilde{X}^n_q$ the "last part". We will use "first" and "last" in the same sense for the estimator $X^{t,n+1}$.

$\hat{X}^q,\ \bar{X}^n_q$ are estimates of the same quantity. It is intuitively obvious, that the new observations $y_{q+1}, ..., y_n$ will not completely invalidate our old estimate, and, in fact, we can be fairly confident that $\hat{X}^n \simeq \bar{X}^n_q$. At any rate, it is a good idea to start the new annealing algorithm with the old estimate as initial value for the first part of $X^{t,n}$.

On the other hand, the choice of the initial temperature, as indicated in the previous section, is related to the confidence we have in the estimate. If we consider a piecewise constant cooling schedule, as we go to lower and lower temperatures, we consider classes of more and more likely estimates. Furthermore our state of higher confidence in the first part of the estimate must be reflected in the annealing scheme we will use. Finally, entropy is often used as an indication of the distance from equilibrium (see **[6]**).

I will now show that an annealing scheme that uses two temperatures, T_1 (low) for the first part of X^n and T_2 (high) for the second part of X^n, reduces entropy and reflects the differing degrees of confidence in the two parts of our estimate. I call this new scheme **inhomogeneous annealing**

In place of one constraint (eq.(11)) on the likelihood of $\hat{X}^{m,n+1}$, we will use the two constraints:

$$E(-\mathcal{I}^{n,q}_{\lambda(m)}(\hat{X}^{m,n})) \doteq G_{m,1}, \quad E(-\mathcal{J}^{n,q}_{\lambda(m)}(\hat{X}^{m,n})) \doteq G_{m,2} \tag{21}$$

Annealing according to a law $\sim exp(-\frac{\mathcal{I}}{T_1} - \frac{\mathcal{J}}{T_2})$ will yield the max-entropy distribution that respects the above constraints. Now, dropping the indices for brevity, it is clear that $\mathcal{H} = \mathcal{I} + \lambda(m)^q\mathcal{J}$. Also, $\mathcal{H}^n_{\lambda(m)} = H^n$ when $\lambda(m) = 1$. Define $\bar{\mathcal{H}} \doteq \mathcal{I} + \alpha\mathcal{J}$. Annealing $\bar{\mathcal{H}}$ **homogeneously** at temperature T_1 (that is with law $\sim exp(\frac{\bar{\mathcal{H}}}{T_1})$) is equivalent to

[1] Strictly speaking this is not true, since t_m is finite. Assume t_m to be large enough that convergence is "almost achieved".

annealing $\mathcal{H}$ **inhomogeneously** at temperatures T_1, T_2 (that is with law $\sim exp(-\frac{\mathcal{I}}{T_1}-\frac{\mathcal{J}}{T_2})$). with $T_2 = \frac{T_1}{\alpha}$. Clearly, the choice of α determines whether we perform homogeneous or inhomogeneous annealing of H^n. For $\alpha = 1$ we have homogeneous annealing. For $\alpha = \lambda^q(m) < 1$ we have inhomogeneous annealing. It will be shown that inhomogeneous annealing has lower entropy, hence smaller computational load, and reflects our higher confidence in the likelihood of the first part of the estimator during the initial stages of the computation. As we proceed, we would like the difference in the likelihood of the two parts of the estimator to decrease. This implies that we want $\lambda(m) \to 1$, which has the effect of gradually equalizing the confidence we have in the two parts of the estimate. To prove the above claims we need the following formulas (which are easy to prove):

$$\frac{dG_{m,1}}{dT_1} = \frac{1}{T_1^2} \cdot \{ - E(\mathcal{I}^2) + E(\mathcal{I})^2\} \tag{22}$$

$$\frac{dG_{m,2}}{dT_1} = \frac{1}{T_1^2} \cdot \{ - E(\mathcal{I}\mathcal{J}) + E(\mathcal{I})E(\mathcal{J})\} \tag{23}$$

$$\frac{dS}{dT_1} = \frac{1}{T_1^2} \cdot \{E(\frac{\mathcal{I}}{T_1} + \frac{\mathcal{J}}{T_2}\mathcal{I}) - E(\frac{\mathcal{I}}{T_1} + \frac{\mathcal{J}}{T_2})E(\mathcal{I})\} \tag{24}$$

Using the above we can also prove that:

1. For fixed T_2, entropy decreases as T_1 decreases (equivalently, as $\alpha \leq 1$). This follows from the fact that

$$\frac{dS}{dT_1} \geq \frac{1}{T_1^3} \cdot [E(\mathcal{I}^2) - \frac{n \cdot k_1 \cdot T_2 + T_1}{n \cdot k_1 \cdot T_1 \cdot T_2} E(\mathcal{J})^2)] \geq 0 \ \ for \ \ large \ \ n. \tag{25}$$

(Where we used Assumptions 2 and 4.)

2. For fixed T_2, G_1 increases as T_1 decreases. This is a direct consequence of eq.(22).

3. For fixed T_2, $G_1 - G_2$ increases as T_1 decreases. This follows from the fact that

$$\frac{d(G_1 - G_2)}{dT_1} = \frac{1}{T_1^2} \cdot [-E(\mathcal{I}\mathcal{J}) + E(\mathcal{I})E(\mathcal{J}) + E(\mathcal{I}^2) - E(\mathcal{I})^2)] \geq$$

$$\frac{1}{T_1^2}[E(\mathcal{I}^2) - (1 + \frac{k_2}{n})E(\mathcal{I})^2] \geq 0 \tag{26}.$$

(Where we have made use of Assumptions 3 and 4.)

Now, we have three facts:

1. The entropy decrease justifes the claim about computational savings.

2. Since $\{\mathcal{H}^n_{\lambda(m)})$ is a submartingale we can use the Markov inequality for martingales and reason about the value of the log-likelihood function from its expectation:

$$P(max_{t \leq t_m} \ P(X^{t,n} \mid Y^n) \leq e^{-\alpha}) \leq \frac{1}{\alpha}(G_{m,1} + G_{m,2}) \tag{27}.$$

This justifies, along with eq.(21), the confidence equalization argument.

3. As mentioned, $\lambda(m) \to 1$; but this implies that we are approaching the minimizer of H^n. Taking $\lambda(m) \uparrow 1$ and using the Dominated Convergence Theorem, we get that

$$lim_{\lambda(m) \to 1} \mathcal{H}_n^{\lambda(m)}(X^{m,n}) = H_n(X^{m,n}) \quad w.p.1. \tag{28}$$

So, in the limit $t_m \to \infty$, $\mathcal{H}^n_{\lambda(m)}(.)$ and $H^n(.)$ have the same minima. This fact supports the following conjecture (also supported by numerical experimentation):

Conjecture:

For any n, exists m large enough that we have:

$$lim_{\bar{n}\to\infty}[argmin\ \mathcal{H}^n_{\lambda(m)}(.) - argmin\ H^{n+\bar{n}}(.)]_i = 0 \quad (for\ \ 1 \leq i \leq n\ ) \tag{29}$$

(where $[Z]_i$ indicates the i-th component of a vector).

The significance of this result is that, as we get longer and longer time series and obtain optimal estimates, the estimate $\hat{x}_{m,n}$ tends to a limiting value. This value (by the convergence of $\mathcal{H}^n$ to H^n) has to be the minimizing value for H^n.

In short, annealing with an inhomogeneous temperature law yields reduced computation (by the empirical entropy argument), reflects the difference of confidence levels in parts of the estimate, introduces a submartingale process that allows use of the Markov inequality and solves the problem of unbounded log-likelihood functions. Hence we have a viable solution to the filtering problem.

6 Conclusion

We have looked at two problems: MAP smoothing and MAP filtering of observations of nonlinear dynamical systems. These are both hard problems for which no satisfactory and general solution exists to date. Smoothing can be done as a straightforward extension of the Homogeneous Annealing Algorithm. The ease of ipmplementation of this algorithm for any kind of nonlinear system and the guaranteed convergence make it very attractive. Filtering is a harder problem, but the Inhomogeneous Annealing Algorithm makes a more efficient computation possible and retains simplicity of implementation and guaranteed convergence.

7 References

1. E.T. Jaynes, "On the Rationale of Maximum-Entropy Methods", Proc. IEEE, Vol.70, No.9, Sept. 1982.
2. E.T. Jaynes, "Prior Probabilities", IEEE SSC-4, No.3, Sep. 1968.
3. S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", IEEE PAMI-6, No.6, Nov. 1984.
4. S. Kirkpatrick, C.D. Gellatt Jr., M.P. Vecchi, "Optimization by Simulated Annealing", IBM Thomas Watson Research Center, NY 1982.
5. R. Kalman, "New Results in Filtering Theory", ASME J. of Basic Eng., March 1960.
6. J.L. Lutton, E. Bonomi, "The N-City Travelling Salesman Problem: Statistical Mechanics and the Metropolis Algorithm" , SIAM Review, 26, 1984.