First Hitting Time Analysis of the Independence Metropolis Sampler

Romeo Maciuca and Song-Chun Zhu

In this paper, we study a special case of the Metropolis algorithm, the Independence Metropolis Sampler (IMS), in the finite state space case. The IMS is often used in designing components of more complex Markov Chain Monte Carlo algorithms. We present new results related to the *first hitting time* of individual states for the IMS. These results are expressed mostly in terms of the eigenvalues of the transition kernel. We derive a simple form formula for the mean first hitting time and we show tight lower and upper bounds on the mean first hitting time with the upper bound being the product of two factors: a "local" factor corresponding to the target state and a "global" factor, common to all the states, which is expressed in terms of the total variation distance between the target and the proposal probabilities. We also briefly discuss properties of the distribution of the first hitting time for the IMS and analyze its variance. We conclude by showing how non-independence Metropolis-Hastings algorithms can perform better than the IMS.

Keywords: Eigenanalysis, Expectation, Fundamental Matrix, Metropolized Gibbs Sampler, Variance.

The authors are with the Department of Statistics, 8130 Math. Science Bldg, Box 951554, University of California, Los Angeles, CA 90095.

emails: {rmaciuca,sczhu}@stat.ucla.edu

1 Introduction

Along with the surge of Markov chain Monte Carlo methods in the scientific community, the Metropolis algorithm with its variations soon became one of the most popular MCMC techniques, in fact it appears on the top of the list of 10 most popular algorithms in a recent review (Sullivan, 2000). In this paper, we study a special case of the Metropolis algorithm – the Independence Metropolis Sampler (IMS), for finite state spaces. The IMS is often used in designing components of more complex Markov Chain Monte Carlo algorithms. Using an acceptance-rejection mechanism described in section 3, the IMS simulates a Markov chain with target probability $p = (p_1, p_2, \ldots, p_n)$, by drawing samples from a more tractable probability $q = (q_1, q_2, \ldots, q_n)$.

In the last two decades a considerable number of papers have been devoted to studying properties of the IMS. Without trying to be comprehensive, we shall briefly review some of the results that were of interest to us. For finite state spaces, Diaconis (1992) and Liu (1996) proved various upper bounds for the total variation distance between updated and target distributions for the IMS. They showed that the convergence rate of the Markov chain is upper bounded by a quantity that depends on the second largest eigenvalue:

$$\lambda_{slem} = 1 - \min_i \{\frac{q_i}{p_i}\}.$$

A complete eigenanalysis of the IMS kernel was performed by Liu (1996). Smith and Tierney (1996) have extended Liu's results to obtain exact m-step transition probabilities for any m for both discrete and continuous state spaces. In the continuous case, if denoting

$$r^* = 1 - \inf_x \{\frac{q(x)}{p(x)}\},$$

they showed that if r^* is strictly less than 1, the chain has a geometric rate of convergence, while if r^* is equal to 1, the convergence is not geometric anymore. Similar results were obtained by Mengersen and Tweedie (1994). These results show that the convergence rate of the Markov chain for the IMS is subject to a *worst-case* scenario. For the finite case, the state corresponding to the least probability ratio q_i/p_i is determining the rate of convergence, that is just one state from a potentially huge state space decides the rate of convergence of the Markov chain. A similar situation occurs in continuous spaces. To illustrate it let us consider the following simple example.

Example: Let q and p be two Gaussians having equal variances and means slightly shifted. Then q, as proposal distribution, will approximate the target p very well. However, it is easy

to see that $inf_x\{q(x)/p(x)\} = 0$ and therefore the IMS algorithm will not have a geometric rate of convergence. This dismal behavior motivated our interest for studying the mean first hitting time as a measure of "speed" for Markov chains. It is particularly appropriate when dealing with stochastic search and optimization algorithms, when the focus could be on finding individual states rather than on the global convergence of the chain.

The concept of first hitting times (f.h.t) has been widely used in various areas stretching from search problems in artificial intelligence (Pearl, 1995) or sensitivity analysis (Cho, 1999) to finance problems (Sodal, 2001).

In this paper, we present new results related to the f.h.t for the IMS. These results are expressed mostly in terms of the eigenvalues of the transition kernel.

- We start with reviewing some formulas for first hitting times. Then, we derive a formula for the mean f.h.t for ergodic kernels in terms of its eigen-elements and show that when the starting distribution of the chain is equal to one of the rows of the transition kernel, the mean f.h.t will have a particularly simple form.

Using this result together with the eigen-analysis of the IMS kernel (briefly reviewed in section 3), we prove the main result, which gives an analytical formula for the mean f.h.t of individual states, as well as bounds.

- We show that, if in running an IMS chain the starting distribution is the same as the proposal distribution q, then after ordering the states according to their probability ratio, and if denoting by λ_i the i^{th} eigenvalue of the transition kernel, we have:

  $i)$ $E[\tau(i)] = \dfrac{1}{p_i(1-\lambda_i)}$

  $ii)$ $\dfrac{1}{\min\{q_i, p_i\}} \leq E[\tau(i)] \leq \dfrac{1}{\min\{q_i, p_i\}} \dfrac{1}{1 - \|p-q\|_{TV}}$,

  where $\tau(i)$ stands for the f.h.t of i, and $\|p-q\|_{TV}$ denotes the total variation distance between the proposal and target distributions.

The result can be extended from individual sets to some subsets of state space, as we shall see in section 3. We then illustrate these findings through a simple example.

- We conclude the section by proving that when starting from $j \neq i$, the mean f.h.t of i are decreasing, with the smallest being equal to the mean f.h.t of i when starting from q:

If $q_1/p_1 \leq q_2/p_2 \leq \ldots \leq q_n/p_n$ then :

$$E_1[\tau(i)] \geq E_2[\tau(i)] \geq \ldots \geq E_{i-1}[\tau(i)] \geq E_{i+1}[\tau(i)] = \ldots = E_n[\tau(i)] = E[\tau(i)], \forall i.$$

Section 3.5 is devoted to studying the tail distribution and the variance of the f.h.t for the IMS.

- We first give an exponential upper bound on its tail distribution: $P(\tau(i) > m) \leq \exp\{-m(p_i w_1)\}$, $\forall m > 0$.

- Then, we analyze the variance of the f.h.t using similar techniques as for the expectation. We find that, if Z denotes the fundamental matrix associated with the IMS kernel then:

$$Var[\tau(i)] = \frac{2Z_{ii}(1-\lambda_i) - 3p_i(1-\lambda_i) + 2p_i - 1}{p_i^2(1-\lambda_i)^2}, \forall i.$$

  We also prove various bounds on the variance.

Finally, in section 4 we show how a special class of Metropolis-Hastings algorithms can outperform the IMS in terms of mean first hitting times.

- We prove that if Q is a stochastic proposal matrix satisfying $Q_{ji}/p_i \geq 1, Q_{ij}/p_j \geq 1, \forall i, \forall j \neq i$, and R is the corresponding Metropolis-Hastings kernel then, for any initial distribution q,

$$E_q^Q[\tau(i)] \leq 1 + \frac{1-q_i}{p_i},$$

  and as a corollary,

$$E_q^Q[\tau(i)] \leq \frac{1}{\min\{q_i, p_i\}} \leq E_q^{IMS}[\tau(i)] \ \ \forall i,$$

  where we denoted by $E_q^{IMS}[\tau(i)]$ the mean f.h.t of the IMS kernel associated to q and p.

2 General f.h.t for finite spaces

Consider an ergodic Markov chain $\{X_m\}_m$ on the finite space $\Omega = \{1, 2, \ldots, n\}$. Let $\mathbf{K}$ be the transition kernel, p its unique stationary probability, and q the starting distribution. For each state $i \in \Omega$, the *first hitting time* is defined below.

Definition 2.1 *The* first hitting time *for a state i is the number of steps for reaching i for the first time in the Markov chain sequence, $\tau(i) = \min\{m \geq 1 : \ X_m = i\}$.*

$E[\tau(i)]$ is the expected first hitting time of i for the Markov chain governed by $\mathbf{K}$.

Let $\{\lambda_j\}_{0\leq j\leq n-1}$ be the eigenvalues of $\mathbf{K}$ with the corresponding right and left eigenvectors v_j, u_j such that $U'V = \mathbf{I}$, where $U' = \{u_k\}_k, V = \{v'_k\}_k$. As $\mathbf{K}$ is a stochastic matrix with stationary probability p, we have $\lambda_0 = 1$ and we can consider $v_0 = \mathbf{1}$ and $u_0 = p$ respectively. Moreover, all the eigenvalues have real values and $|\lambda_j| < 1, \forall j > 0$.

There are two ways of looking at the mean and variance of the f.h.t.

2.1 Inverse matrix approach

Let i be the target state and let us denote by $\mathbf{K}_{-i}$ the $(n-1)\times(n-1)$ matrix obtained from $\mathbf{K}$ by deleting the i-th column and row, that is, $\mathbf{K}_{-i}(k,j) = \mathbf{K}(k,j), \forall k \neq i,\ j \neq i$. Also let $q_{-i} = (q_1, ..., q_{i-1}, q_{i+1}, ..., q_n)$. Then it is facile to show that $P(\tau(i) > m) = q_{-i}\mathbf{K}_{-i}^{m-1}\mathbf{1}$, where $\mathbf{1} = (1, 1, ..., 1)'$ is a vector of ones. This leads to the following formula for the expectation:

$$E_q[\tau(i)] = 1 + q_{-i}(\mathbf{I} - \mathbf{K}_{-i})^{-1}\mathbf{1}, \tag{2.1}$$

where $\mathbf{I}$ denotes the identity matrix. The existence of the inverse of $\mathbf{I} - \mathbf{K}_{-i}$ is assured by the substochasticity of $\mathbf{K}_{-i}$ and the irreducibility of $\mathbf{K}$ (Bremaud, 1999).

Let us note that this is just a particular case of the more general formula for the mean f.h.t of a subset A of Ω. That is, more generally,

$$E_q[\tau(A)] = 1 + q_{-A}(\mathbf{I} - \mathbf{K}_{-A})^{-1}\mathbf{1}, \quad \forall A \subset \Omega. \tag{2.2}$$

To be thorough, we give the corresponding formula for the variance, with the mention that we shall not use it in our analysis.

$$Var[\tau(i)] = 2q_{-i}(\mathbf{I} - \mathbf{K}_{-i})^{-2}\mathbf{1} - q_{-i}(\mathbf{I} - \mathbf{K}_{-i})^{-1}\mathbf{1} - [q_{-i}(\mathbf{I} - \mathbf{K}_{-i})^{-1}\mathbf{1}]^2, \forall i \in \Omega. \tag{2.3}$$

2.2 The fundamental matrix approach

The fundamental matrix Z is defined to be $Z = (\mathbf{I} - \mathbf{K} + P)^{-1}$ or, equivalently, $Z = \mathbf{I} + \sum_{k\geq 1}(\mathbf{K}^k - P)$, where P denotes the matrix having all rows equal to p. We summarize below some of its properties:

i) $(\mathbf{I} - \mathbf{K})Z = Z(I - \mathbf{K}) = \mathbf{I} - P$

ii) $PZ = P, Z\mathbf{1} = \mathbf{1}$.

Also, it is noted that Z and $\mathbf{K}$ share the same system of eigenvectors, while the eigenvalues of Z are $\beta_0 = 1, \beta_j = 1/(1-\lambda_j), \forall 1 \leq j \leq n-1$. To prove this, we note that $(\mathbf{I} - \mathbf{K} + P)v_k =$

$v_k - \mathbf{K}v_k + Pv_k = (1-\lambda_k)v_k, \forall k > 0$, where we used $u_0' v_k = 0$, or $Pv_k = 0$ and also that v_k is a right eigenvector of $\mathbf{K}$ corresponding to the eigenvalue λ_k. Now, as $Z = (\mathbf{I} - \mathbf{K} + P)^{-1}$, it follows that $v_k = (1-\lambda_k)Zv_k$ or equivalently, $Zv_k = \beta_k v_k, k > 0$. For $k = 0$ we use $ii)$ from above. As the right eigenvectors for Z and $\mathbf{K}$ coincide then the left eigenvectors will also be the same as it can also be easily seen by repeating the above computations for u_k.

Mean f.h.t can be described using the fundamental matrix in the following way. Let us denote by $E_j[\tau(i)]$ the mean f.h.t of i when starting from state j. Then, for all $j \neq i$, one has

$$E_j[\tau(i)] = (Z_{ii} - Z_{ji})/p_i. \tag{2.4}$$

When we start from q instead from a fixed state j, we have:

$$E_q[\tau(i)] = 1 + \sum_{j \neq i} q_j E_j[\tau(i)] = 1 + \frac{1}{p_i}\sum_{j \neq i} q_j(Z_{ii} - Z_{ji}), \tag{2.5}$$

where the 1 corresponds to the first step of the chain. For the rest of the paper we shall drop the subscript q from the expectation whenever this will not create any notation confusion.

The variance of the f.h.t can also be derived from the fundamental matrix Z. As before, the formulas refer to the chain that starts from a fixed state j. It is known that the second moment of $\tau(i)$, when starting from j, is determined by:

$$E_j[\tau(i)]^2 = \frac{2}{p_i}(Z_{ii}^2 - Z_{ji}^2) - \frac{1}{p_i}(Z_{ii} - Z_{ji}) + \frac{2}{p_i^2}Z_{ii}(Z_{ii} - Z_{ji}), \forall j \neq i, \tag{2.6}$$

where the first term refers to the matrix Z^2. Knowing this, it is immediate that the second moment of the f.h.t when starting from q is just:

$$E[\tau(i)^2] = 1 + \frac{2}{p_i}\sum_j q_j(Z_{ii}^2 - Z_{ji}^2) - \frac{1}{p_i}\sum_j q_j(Z_{ii} - Z_{ji}) + \frac{2Z_{ii}}{p_i^2}\sum_j q_j(Z_{ii} - Z_{ji}). \tag{2.7}$$

For a detailed account on the properties of the fundamental matrix Z and its connections with hitting times refer to Kemeny (1976).

Remark: We need to note that in our notation $E_j[\tau(i)]$ is not the same as $E_\eta[\tau(i)]$ for $\eta = \{\delta_{jl}\}_l$. In $E_j[\tau(i)]$ we do not count starting from j as one step of the chain, while in $E_\eta[\tau(i)]$ we would have to count it for consistence. For our purposes we shall only use $E_j[\tau(i)]$ per se and not as a particular case of $E_q[\tau(i)]$.

It is worth showing briefly how formulas like (2.4) or (2.6) can be obtained. We shall illustrate the method for (2.4).

From (2.1) it follows that we need to compute $(\mathbf{I}-\mathbf{K}_{-i})^{-1}\mathbf{1}$. With the hindsight that the inverse will depend on Z let us compute $(\mathbf{I}-\mathbf{K})(aZ+b\mathbf{1}'\mathbf{1})$ where a, b will be determined latter on. Using $i)$ and $\mathbf{K1}=\mathbf{1}$, we get $(\mathbf{I}-\mathbf{K})(aZ+b\mathbf{1}'\mathbf{1})=a(\mathbf{I}-P)$. Now if we consider b such that $(aZ+b\mathbf{1}'\mathbf{1})_{ii}=0$, it will follow that

$$(\mathbf{I}-\mathbf{K}_{-i})(aZ+b\mathbf{1}'\mathbf{1})_{-i}=[(\mathbf{I}-\mathbf{K})(aZ+b\mathbf{1}'\mathbf{1})]_{-i}=a(\mathbf{I}-P)_{-i}.$$

Writing this equality only for column i we get $(\mathbf{I}-\mathbf{K}_{-i})(aZ_{\cdot i}+b\mathbf{1})_{-i}=a(\delta_i-p_i\mathbf{1})_{-i}$ or equivalently,

$$(\mathbf{I}-\mathbf{K}_{-i})^{-1}\mathbf{1}=-\frac{b}{ap_i}\mathbf{1}-\frac{Z_{\cdot i}}{p_i}.$$

But b is given by $(aZ+b\mathbf{1}'\mathbf{1})_{ii}=0$ which leads to $b=-aZ_{ii}$ so finally, $(\mathbf{I}-\mathbf{K}_{-i})^{-1}\mathbf{1}=Z_{ii}/p_i-Z_{\cdot i}/p_i$ which is exactly (2.4).

As a matter of notation we have denoted by $Z_{\cdot i}$ the i^{th} column of Z and we used the subscript $-i$ to indicate that we eliminated row and column i (for matrices) or just component i for vectors. For simplifying the notation we did not subscript $Z_{\cdot i}$ with $-i$ in the last two statements where it appears. We also used $\mathbf{I}$ to stand for the identity matrix for both order n and order $n-1$.

The second moment of the f.h.t could be determined in the same way, by computing $(\mathbf{I}-\mathbf{K})(aZ^2+bZ+c\mathbf{1}'\mathbf{1})$ for a, b, c such that $(aZ^2+bZ+c\mathbf{1}'\mathbf{1})_{ii}=0$, and then equating $(\mathbf{I}-\mathbf{K})_{-i}(aZ^2+bZ+c\mathbf{1}'\mathbf{1})_{-i}=(Z_{ii}-Z_{\cdot i})/p_i$ to get $(\mathbf{I}-\mathbf{K}_{-i})^{-2}\mathbf{1}$. The same method would work for any moment of the f.h.t, but we shall not go further on this route.

2.3 The mean f.h.t for the general case

Here, we shall derive a formula for the expectation of the f.h.t in terms of its eigenvalues and eigenvectors. Let us expand the eigenvectors by denoting $v_k=\{v_{kl}\}_{0\le l\le n-1}$ and $u_k=\{u_{kl}\}_{0\le l\le n-1}$.

Proposition 2.1 *Using the same notations as before, for any ergodic kernel* $\mathbf{K}$ *and any initial distribution* q*, the mean first hitting time of* $i\in\Omega$ *is*

$$E[\tau(i)]=1+\frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki}-\sum_l q_l v_{kl}).$$

In particular, if q *is chosen to be row* j^{th} *of* $\mathbf{K}$ *for arbitrary* $j\in\Omega$*, then*

$$E[\tau(i)]=\frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki}-v_{kj})+\frac{\delta_{ij}}{p_i}.$$

Proof: We use (2.5) which gives

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{j\neq i} q_j(Z_{ii} - Z_{ji}). \tag{2.8}$$

Knowing the complete eigenstructure of Z, we can apply the spectral decomposition theorem, to get:

$$Z_{li} = \sum_{k=0}^{n-1} \beta_k v_{kl} u_{ki} = v_{0l}u_{0i} + \sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} v_{kl}u_{ki} = p_i + \sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} v_{kl}u_{ki}, \forall l, i.$$

Therefore, we can compute $Z_{ii} - Z_{ji}$ in terms of eigen-elements of $\mathbf{K}$:

$$Z_{ii} - Z_{ji} = \sum_{k=1}^{n-1} \frac{1}{1-\lambda_k}(v_{ki} - v_{kj})u_{ki}. \tag{2.9}$$

Combining (2.9) and (2.8), we get

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{j\neq i} q_j(Z_{ii} - Z_{ji}) = 1 + \frac{1}{p_i}\sum_{j\neq i} q_j \sum_{k=1}^{n-1} \frac{1}{1-\lambda_k}(v_{ki} - v_{kj})u_{ki},$$

which, by changing the summation order, turns into

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki} \sum_{j\neq i} q_j(v_{ki} - v_{kj}). \tag{2.10}$$

We note that we can rewrite $\sum_{j\neq i} q_j(v_{ki} - v_{kj})$ as $v_{ki} - \sum_l q_l v_{kl}$. Hence, from (2.10), we get the desired form for $E[\tau(i)]$, that is:

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki}(v_{ki} - \sum_l q_l v_{kl}).$$

Now, assume that $q = \mathbf{K}_{j\cdot}$. This implies that $\sum_l q_l v_{kl} = \sum_l K_{jl} v_{kl} = (Kv_k)_j$. But as v_k is a right eigenvector associated with the eigenvalue λ_k, we get $\sum_l q_l v_{kl} = \lambda_k v_{kj}$ and by plugging this into the above expression of the expectation one will get

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki}(v_{ki} - \lambda_k v_{kj}) = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki}(v_{ki} - v_{kj} + (1-\lambda_k)v_{kj}).$$

Splitting the above in two parts

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki}(v_{ki} - v_{kj}) + \frac{1}{p_i}\sum_{k=1}^{n-1} u_{ki}v_{kj}. \tag{2.11}$$

We have to consider two cases:

i) $j = i$.

In this case, $\sum_{k=1}^{n-1} u_{ki}v_{kj} = \sum_{k=0}^{n-1} u_{ki}v_{ki} - p_i = 1 - p_i$, as $\sum_{k=0}^{n-1} u_{ki}v_{ki} = 1$ from $U'V = \mathbf{I}$. Therefore, from (2.11) it follows that $E[\tau(i)] = 1/p_i$, the first sum cancelling for $j = i$.

ii) $j \neq i$.

Then, again, $\sum_{k=1}^{n-1} u_{ki}v_{kj} = \sum_{k=0}^{n-1} u_{ki}v_{kj} - p_i = \delta_{ij} - p_i = -p_i$. Now, using (2.11)

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki} - v_{kj}) - 1 = \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki} - v_{kj}). \quad \square$$

3 Hitting time analysis for the IMS

Here, we shall capitalize on the previous result to prove our main theorem. But first, let us set the stage by briefly introducing the IMS.

3.1 The Independence Metropolis Sampler

The IMS is a Metropolis-Hastings type algorithm with the proposal independent of the current state of the chain. It has also been called Metropolized Independent Sampling (Liu, 1996). Let $\Omega = \{1, 2, \ldots, n\}$ be the state space. As for any MCMC algorithm, the goal is to simulate a Markov chain $\{X_m\}_{m\geq 0}$ taking values in Ω and having stationary distribution p (the target probability) . To do this, at each step a new state $j \in \Omega$ is sampled from the proposal probability $q = (q_1, q_2, \ldots, q_n)$ according to $j \sim q_j$, which is then accepted with probability

$$\alpha(i, j) = \min\{1, \frac{q_i}{p_i}\frac{p_j}{q_j}\}.$$

Therefore, the transition from X_m to X_{m+1} is decided by the transition kernel having the form

$$\mathbf{K}(i, j) = \begin{cases} q_j\alpha(i, j) & j \neq i, \\ 1 - \sum_{k\neq i}\mathbf{K}(i, k) & j = i. \end{cases}$$

The initial state could be either fixed or generated from a distribution whose natural choice in this case is q. We shall see later, in section 3.3, why it is more efficient to generate the initial state from q instead of choosing it deterministically.

It is easy to show that p is the invariant (stationary) distribution of the chain. In other words, $p\,\mathbf{K} = p$. Since from $q > 0$ it follows that $\mathbf{K}$ is ergodic, then p is also the equilibrium distribution of the chain. Therefore, the marginal distribution of the chain at step m, for m large enough, is approximately p.

However, instead of trying to sample from the target distribution p, one may be interested in searching for a state i^* with maximum probability: $i^* = \arg\max_{i\in\Omega} p_i$. Here is where the mean f.h.t can come into play. $E[\tau(i)]$ is a good measure for the speed of search in general. As a special case we may need to know $E[\tau(i^*)]$ for the optimal state.

As it shall become clear later, a key quantity to the analysis is the probability ratio $w_i = q_i/p_i$. It measures how much knowledge the heuristic q_i has about p_i, or in other words how *informed* is q about p for state i. Therefore we define the following concepts.

Definition 3.1 *A state i is said to be* over-informed *if $q_i > p_i$ and i is* under-informed *if $q_i < p_i$.*

There are three special states defined below.

Definition 3.2 *A state i is* exactly-informed *if $q_i = p_i$. A state i is* most-informed *(or* least-informed*) if it has the highest (or lowest) ratio w_i: $i_{\max} = \arg\max_{i\in\Omega}\{w_i\}$, $i_{\min} = \arg\min_{i\in\Omega}\{w_i\}$.*

Let us observe that because of its special form, the transition kernel can be written in a simpler form by reordering the states increasingly according to their informedness. Noticing that for $i \neq j$, $\mathbf{K}_{ij} = q_j \min\{1, w_i/w_j\}$, if $w_1 \leq w_2 \leq \ldots \leq w_n$ it follows that

$$\mathbf{K}_{ij} = \begin{cases} w_i p_j & i < j, \\ 1 - \sum_{k<i} q_k - w_i \sum_{k>i} p_k & i = j, \\ q_j = w_j p_j & i > j. \end{cases}$$

Without loss of generality, we shall assume for the rest of the paper that the states are indexed such that $w_1 \leq w_2 \leq \ldots \leq w_n$, to allow for this more tractable form of the transition kernel.

Proposition 2.1 can be used to compute mean first hitting times whenever an eigen-analysis for the transition kernel is available. In practice, this situation is quite rare though. However, such an eigen-analysis is available for the IMS when the state space is finite. We review these results below and then proceed with our results.

3.2 The eigenstructure of the IMS

A first result concerns the eigenvalues and right eigenvectors of the IMS kernel.

Theorem 3.1 (J. Liu, 1996) *Let $T_k = \sum_{i\geq k} q_i$ and $S_k = \sum_{i\geq k} p_i$. Then the eigenvalues of the transition matrix $\mathbf{K}$ are $\lambda_k = T_k - w_k \cdot S_k, \forall\, 1 \leq k \leq n-1$, and they are decreasing as $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} \geq 0$. Moreover, the right eigenvector corresponding to λ_k is $v_k = (0, \cdots, 0, S_{k+1}, -p_k, \cdots, -p_k)$, where for $k > 0$ the first $k-1$ entries are 0.*

Obviously, $v_0 = (1, 1, \ldots, 1)'$.

Remark: It is easy to see now that the eigenvalues of $\mathbf{K}$ are "incorporated" in the diagonal terms of $\mathbf{K}$ through the equality $K_{ii} = \lambda_i + q_i$, which will be often used later on.

Smith and Tierney (1996) computed the exact k-step transition probabilities for the IMS. One of their results reveals in fact the very structure of the left eigenvectors.

Suppose δ_k is the unit vector with 1 in the k'th position ($1 \leq k \leq n$) and 0 everywhere else. They showed that:

Proposition 3.2 (Smith and Tierney, 1996) *For* $1 \leq k \leq n-1$,

$$\delta_k = p_k v_0 + \frac{1}{S_k} v_k - p_k \sum_{j=1}^{k-1} \frac{v_j}{S_j S_{j+1}}$$

while for $k = n$,

$$\delta_n = p_n v_0 - p_n \sum_{j=1}^{n-1} \frac{v_j}{S_j S_{j+1}}.$$

As a corollary, the left eigenvectors of $\mathbf{K}$ are given by:

Corollary 3.3

$$u_0 = p, u_k = (0, 0, \ldots, 0, \frac{1}{S_k}, -\frac{p_{k+1}}{S_k S_{k+1}}, \ldots, -\frac{p_n}{S_k S_{k+1}})^T, 1 \leq k \leq n-1,$$

where for $k > 0$ *the first* $k-1$ *entries are 0.*

3.3 Our Main Result

We are now able to compute the mean f.h.t for the IMS and provide bounds for it, by making use of the eigenstructure of the IMS kernel as well as of Proposition 2.1.

Theorem 3.4 *If* $\mathbf{K}$ *corresponds to the IMS kernel and the initial distribution of the chain is the same as the proposal probability* q*, then, using previous notations:*

i) $E[\tau(i)] = \dfrac{1}{p_i(1-\lambda_i)}, \forall i \in \Omega,$

ii) $\dfrac{1}{\min\{q_i, p_i\}} \leq E[\tau(i)] \leq \dfrac{1}{\min\{q_i, p_i\}} \dfrac{1}{1 - \|p-q\|_{TV}},$

where we define λ_n *to be equal to zero and* $\|p-q\|_{TV}$ *denotes the total variation distance between* p *and* q*. Equality is attained for the three special states from Definition 3.2.*

Proof: i) Let us first note that we are in the situation from the second part of Proposition 2.1. That is, after reordering the states according to their probability ratios, our initial distribution q is equal to the n^{th} row of $\mathbf{K}$ as it can easily be seen.

Then, from Proposition 2.1, one has:

$$E[\tau(i)] = \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki}-v_{kn}) + \frac{\delta_{in}}{p_i}. \tag{3.1}$$

Let us note that from Theorem 3.1, $v_{ki} = v_{kn}, \forall k < i$ while from Corollary 3.3, $u_{ki} = 0$ for $k > i$, hence $u_{ki}(v_{ki} - v_{kn}) = 0, \forall k \neq i$. If $i = n$ then the only term left in the above expression of the expectation is $\delta_{in}/p_i = 1/p_n = 1/[p_n(1-\lambda_n)]$, meanwhile for $i < n$ one has

$$E[\tau(i)] = \frac{u_{ii}(v_{ii}-v_{in})}{p_i(1-\lambda_i)}.$$

But, using the eigenanalysis for the IMS, $u_{ii}(v_{ii}-v_{in})$ is nothing more than $(S_{i+1}-(-p_i))/S_i = S_i/S_i = 1$, so the average f.h.t becomes

$$E[\tau(i)] = \frac{1}{p_i(1-\lambda_i)},$$

and the proof of i) is completed.

ii) By using i) it is obvious that $E[\tau(i)] \geq 1/p_i$ since $0 \leq \lambda_i < 1$. Therefore, the proof of the lower bound reduces to showing that $1 - \lambda_i \leq w_i$ which would imply that $E[\tau(i)] \geq 1/q_i$. Noting that $\lambda_i = q_i + q_{i+1} + \ldots + q_n - (p_i + p_{i+1} + \ldots + p_n)w_i$, we need to prove that

$$w_i = \frac{q_i}{p_i} \geq \frac{q_1 + q_2 + \ldots + q_{i-1}}{p_1 + p_2 + \ldots + p_{i-1}}.$$

This is quite obvious since for any $j < i$, $w_j \leq w_i \Longleftrightarrow q_j \leq p_j w_i$. By summing the last inequality with j from 1 to $i-1$ we get the desired result.

To prove the upper bound, let us first get a more tractable form for $\|p-q\|_{TV}$. We partition the state space into two sets: under-informed and over-informed with the exactly-informed states in either set: $\Omega = \Omega_{\text{under}} \cup \Omega_{\text{over}}$. As the states are sorted, let $k \leq n$ be their dividing point

$$\Omega_{\text{under}} = \{i \leq k : q_i \leq p_i\}, \quad \Omega_{\text{over}} = \{i > k : q_i > p_i\},$$

where Ω_{over} can be the empty set if $q = p$. By definition, $\|p-q\|_{TV} = \frac{1}{2}\sum_i |p_i - q_i|$. Since $\sum_{i\in\Omega}(p_i - q_i) = 0$, we have

$$\begin{aligned}\|p-q\|_{TV} &= \frac{1}{2}\sum_{i\in\Omega}|p_i-q_i| = \frac{1}{2}\sum_{i\in\Omega_{\text{under}}}(p_i-q_i) + \frac{1}{2}\sum_{i\in\Omega_{\text{over}}}(q_i-p_i) \\ &= \sum_{i\in\Omega_{\text{over}}}(q_i-p_i) = T_{k+1} - S_{k+1},\end{aligned} \tag{3.2}$$

where we define $T_{n+1} = S_{n+1} = 0$. We prove the upper bound for the under-informed and over-informed states respectively.

Case I. upper bound for under-informed states $i \leq k$.

For under-informed states, $q_i = \min\{p_i, q_i\}$. As $\lambda_i = T_i - w_i S_i$, it follows that:

$$p_i(1-\lambda_i) = p_i(1-T_i) + q_i S_i = p_i(1-T_{i+1}) - p_i q_i + q_i S_{i+1} + q_i p_i = p_i(1-T_{i+1}) + q_i S_{i+1}.$$

Therefore, $p_i(1-\lambda_i) \geq q_i(1-T_{i+1}+S_{i+1})$. By using (3.2), we get $\min\{p_i, q_i\}(1-\|p-q\|_{TV}) = q_i(1-T_{k+1}+S_{k+1})$. Thus, we only need to show that $S_{i+1} - S_{k+1} \geq T_{i+1} - T_{k+1}$. By definition, this is equivalent to $p_{i+1} + p_{i+2} + \ldots + p_k \geq q_{i+1} + q_{i+2} + \ldots q_k$, which is obviously true because states $i+1, \ldots, k$ are under-informed.

The equality is attained, as noticed from the proof, when $p_j = q_j, \forall j \in [i,k]$, which is at the exactly-informed states.

Case II. upper bound for over-informed states $i > k$.

As $\min\{p_i, q_i\} = p_i$, it suffices to show that $p_i(1-\lambda_i) \geq p_i(1-T_{k+1}+S_{k+1})$, or $\lambda_i \leq T_{k+1} - S_{k+1}$. As $\lambda_i \leq \lambda_{k+1}$, it is enough to prove that

$$\begin{aligned} & \lambda_{k+1} \leq T_{k+1} - S_{k+1}, \\ \text{or} \quad & T_{k+1} - w_{k+1} S_{k+1} \leq T_{k+1} - S_{k+1}, \\ \text{or} \quad & S_{k+1}(1-w_{k+1}) \leq 0. \end{aligned}$$

The last step becomes trivial since $w_{k+1} \geq 1$ for over-informed states.

Equality in this case is obtained if $\lambda_i = \lambda_{i-1} = \ldots = \lambda_{k+1}$ and $w_{k+1} = 1$ which is equivalent to $w_{k+1} = w_{k+2} = \ldots = w_i = 1$.

Theorem 3.4 can be extended by considering the first hitting time of some particular sets. We give the following corollary:

Corollary 3.5 *Let $A \subset \Omega$ of the form $A = \{i+1, i+2, \ldots, i+k\}$, with $w_1 \leq w_2 \leq \ldots \leq w_n$. Also, let us define $p_A = p_{i+1} + p_{i+2} + \ldots + p_{i+k}$, $q_A = q_{i+1} + q_{i+2} + \ldots + q_{i+k}$ and $w_A = q_A/p_A$. We denote $\lambda_A = (q_{i+1} + \ldots + q_n) - (p_{i+1} + \ldots + p_n) w_A$. Then,*

i) $E[\tau(A)] = \dfrac{1}{p_A(1-\lambda_A)}$,

ii) $\dfrac{1}{\min\{q_A, p_A\}} \leq E[\tau(A)] \leq \dfrac{1}{\min\{q_A, p_A\}} \dfrac{1}{1-\|p-q\|_{TV}}$.

Proof: In the following, we will only prove part i) since for ii) the proof is similar to the one we made for part ii) of Theorem 3.4.

Let $A = \{i+1, i+2, \ldots, i+k\}$.

First, we observe that $w_1 \leq w_2 \leq \ldots \leq w_i \leq w_A \leq w_{i+k+1} \leq \ldots \leq w_n$. Therefore, if we consider A to be a singleton, the problem of computing the average first hitting time of A reduces to computing the average f.h.t of the singleton A in the "reduced" space $\Omega_A := \{1, 2, \ldots, i, \{A\}, i+k+1, \ldots, n\}$. The new proposal (respectively target) probability will be q restricted on the space Ω_A by putting mass q_A on the state $\{A\}$ (similarly for p).

It is easy to check that $\mathbf{K}_{-A} = \mathbf{K}_{-\{A\}}$, where the last matrix is obtained if we consider A to be a singleton (it is essential that the ordering of the states according to the probability ratios is the same in Ω as in Ω_A).

Now, we can apply (2.2) to obtain

$$E_\Omega[\tau(A)] = 1 + q_{-A}(\mathbf{I} - \mathbf{K}_{-A})^{-1}\mathbf{1}' = 1 + q_{-\{A\}}(\mathbf{I} - \mathbf{K}_{-\{A\}})^{-1}\mathbf{1},$$

and by applying Theorem 3.4 for Ω_A

$$E_\Omega[\tau(A)] = E_{\Omega_A}[\tau(\{A\})] = \frac{1}{p_A(1-\lambda_A)}.$$

We used subscripts Ω or Ω_A to indicate which space we are working on. $\square$

In the introduction we have hinted at motivating why generating the initial state from q is preferable to starting from a fixed state $j \neq i$. The following result attempts to clarify this issue.

Proposition 3.6 *Assuming the states are ordered as $w_1 \leq w_2 \leq \ldots \leq w_n$, the following holds:*

$$E_1[\tau(i)] \geq E_2[\tau(i)] \geq \ldots \geq E_{i-1}[\tau(i)] \geq E_{i+1}[\tau(i)] = \ldots = E_n[\tau(i)] = E[\tau(i)], \forall i \in \Omega.$$

Proof: We have seen that

$$E_j[\tau(i)] = \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki} - v_{kj}),$$

for $j \neq i$.

$i)$ $j > i$. Then, $u_{ki}(v_{ki} - v_{kj}) = u_{ki}(v_{ki} - v_{kn}), \forall k > 0$ since for $k > i$, $u_{ki} = 0$ and for $k \leq i, v_{kj} = -p_k = v_{kn}$. Therefore,

$$E_j[\tau(i)] = \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}u_{ki}(v_{ki} - v_{kn}) = E[\tau(i)],$$

as the last equality has been already proven in (3.1).

ii) $j < i$. Let us compute the difference $E_j[\tau(i)] - E_{j+1}[\tau(i)]$ for arbitrary j.

$$\begin{aligned} E_j[\tau(i)] - E_{j+1}[\tau(i)] &= \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}(v_{ki} - v_{kj})u_{ki} - \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}(v_{ki} - v_{k(j+1)})u_{ki} = \\ &= \frac{1}{p_i}\sum_{k=1}^{n-1}\frac{1}{1-\lambda_k}(v_{k(j+1)} - v_{kj})u_{ki} \end{aligned} \tag{3.3}$$

If $j < i-1$ then for $k < j$ we have $v_{k(j+1)} = 0 = v_{kj}$ while for $j+1 < k < i$, $v_{k(j+1)} = -p_k = v_{kj}$, so in both cases the difference is zero, which cancels the corresponding terms in (3.3). The terms for $k > i$ cancel also because $u_{ki} = 0$. The only remaining terms are those for $k = j, j+1$. Therefore,

$$E_j[\tau(i)] - E_{j+1}[\tau(i)] = \frac{1}{p_i}\Big[\frac{1}{1-\lambda_j}(v_{j(j+1)} - v_{jj})u_{ji} + \frac{1}{1-\lambda_{j+1}}(v_{(j+1)(j+1)} - v_{(j+1)j})u_{(j+1)i}\Big]$$

We now note that, according to the eigen-analysis,

$$(v_{j(j+1)} - v_{jj})u_{ji} = (-p_j - S_{j+1})(-\frac{p_i}{S_jS_{j+1}}) = \frac{p_i}{S_{j+1}}.$$

And similarly

$$(v_{(j+1)(j+1)} - v_{(j+1)j})u_{(j+1)i} = (S_{j+2} - 0)(-\frac{p_i}{S_{j+1}S_{j+2}}) = -\frac{p_i}{S_{j+1}}.$$

Hence,

$$E_j[\tau(i)] - E_{j+1}[\tau(i)] = \frac{1}{p_i}\Big(\frac{1}{1-\lambda_j} - \frac{1}{1-\lambda_{j+1}}\Big)\frac{p_i}{S_{j+1}} = \frac{1}{S_{j+1}}\Big(\frac{1}{1-\lambda_j} - \frac{1}{1-\lambda_{j+1}}\Big).$$

This is obviously a positive quantity since $\lambda_j \geq \lambda_{j+1}$. The equality case is obtained if $w_j = w_{j+1}$ which would imply that $\lambda_j = \lambda_{j+1}$. Therefore, if states j and $j+1$ have the same informedness, it would make no difference from which one of them the sampler would start.

The only thing left to prove is that $E_{i-1}[\tau(i)] \geq E[\tau(i)]$. To do this, we note that one can write (3.3) with $i-1$ in the place of j and $i+1$ instead of $j+1$. This gives

$$E_{i-1}[\tau(i)] - E_{i+1}[\tau(i)] = \frac{1}{p_i}\sum_{k=1}^{i}\frac{1}{1-\lambda_k}(v_{k(i+1)} - v_{k(i-1)})u_{ki}.$$

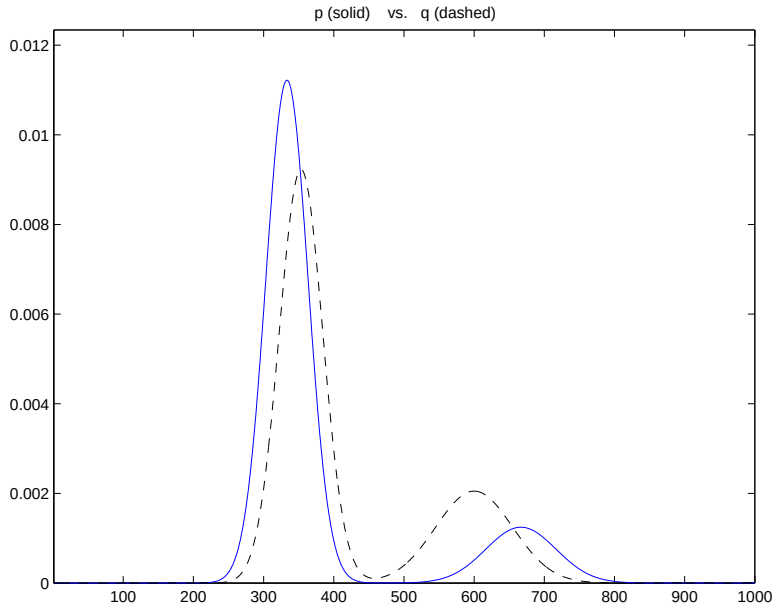
As before, all the terms cancel except for $k = i-1, i$ and then, the same way as above, after making the calculations we end up with

$$E_{i-1}[\tau(i)] - E_{i+1}[\tau(i)] = \frac{1}{S_i}\Big(\frac{1}{1-\lambda_{i-1}} - \frac{1}{1-\lambda_i}\Big),$$

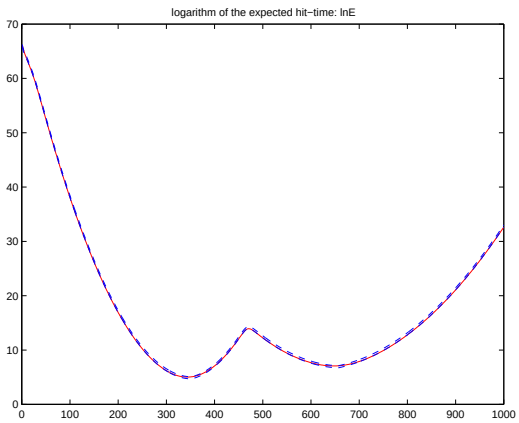
which is again positive since $\lambda_{i-1} \geq \lambda_i$. As we have already proved that $E_{i+1}[\tau(i)] = E_j[\tau(i)] = E[\tau(i)], \forall j > i$, the proof of Proposition 3.6 is completed.□

3.4 Example

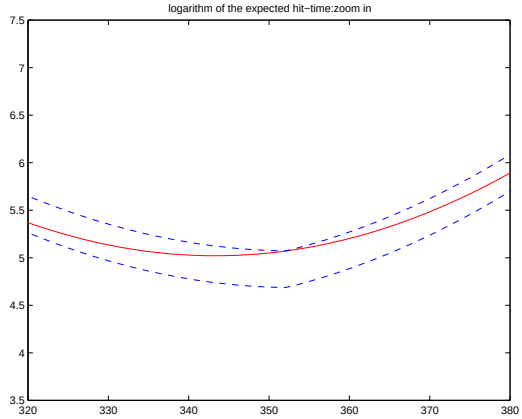
We can illustrate the main results in Theorem 3.4 through a simple example. We consider a space with $n = 1000$ states. Let p and q be mixtures of two discretized Gaussians with tails truncated and then normalized to one. They are plotted as solid (p), dashed (q) curves in Fig.1a .



(a) p (solid) vs q (dashed)



(b) $\ln E[\tau(i)]$ (solid) and bounds (dashed)



(c) $\ln E[\tau(i)]$ (solid) and bounds (dashed)-zoomed in

Figure 1: Mean f.h.t and bounds

Fig.1b plots the logarithm of the expected first hitting-time $\ln E[\tau(i)]$. The lower and upper bounds from Theorem 3.4 are plotted in logarithm scale as dashed curves which almost coincide with the hitting-time plot. For better resolution we focused on a portion of the plot around the mode, the three curves becoming more distinguishable in Fig.1c. We can see that the mode $x^* = 333$ has $p(x^*) \approx 0.012$ and it is hit in $E[\tau_{x^*}] \approx 162$ times on average for q. This is much smaller than $n/2 = 500$ which would be the average time for exhaustive search. In comparison, for an uninformed (i.e uniform) proposal the result is $E[\tau_{x^*}] = 1000$. Thus, it becomes visible how a "good" proposal q can influence the speed of such a stochastic sampler.

3.5 Other properties of the f.h.t for the IMS

In this section we discuss additional properties of the f.h.t for the IMS. Two items will be of interest here: the distribution and the variance of the f.h.t.

3.5.1 The Tail Distribution

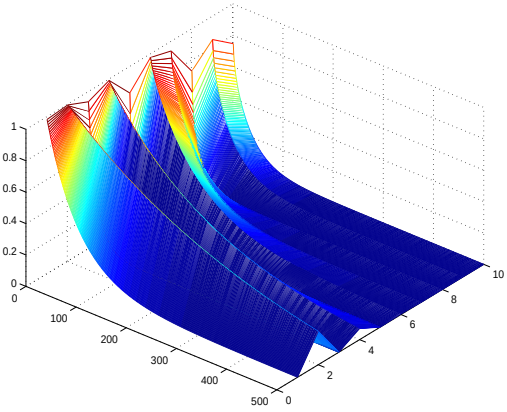
It is known (Abadi, 2001) that the distribution of first hitting times is generally well approximated by an exponential distribution after some waiting time. Our simulations showed that for the IMS this is indeed the case and moreover, the approximation seems to be good at all times. We illustrate this assertion in Figure 2 for a state space with $N = 10$ states, with p and q being discretized mixtures of Gaussians as before.

On the next page, Fig. 2a and b plot the tail distributions of the first hitting times for all the states of the space. It is apparent that their shapes resemble exponential tails. In Fig. 2c, we have plotted both the tail distribution of the f.h.t and the corresponding exponential distribution for an arbitrary state (taken to be $i = 3$). One notes that the fit is quite good. Even though we were not able to quantify the approximation error, we can give an exponential upper bound on the tail distribution of the f.h.t. This was depicted in Fig. 2d.

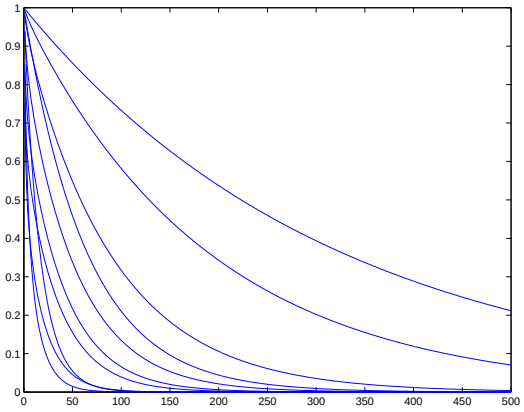
Proposition 3.7 *For all $i \in \Omega$, $P(\tau(i) > m) \leq (1 - q_i)(1 - p_i w_1)^m \leq \exp\{-m(p_i w_1)\}$, $\forall m > 0$.*

Proof: For all $j \neq i$ we can write $\mathbf{K}_{ji} = p_i \min\{w_i, w_j\}$. This shows that $\mathbf{K}_{ji} \geq p_i w_1, \forall j \neq i$. Or, equivalently, $1 - \mathbf{K}_{ji} \leq 1 - p_i w_1, \forall j \neq i$. By writing this set of inequalities in matrix form, one gets $\mathbf{K}_{-i}\mathbf{1} \leq (1 - p_i w_1)\mathbf{1}$. Now, we can iterate this inequality and therefore, $\mathbf{K}_{-i}^l \mathbf{1} \leq (1 - p_i w_1)^l \mathbf{1}$, $\forall l$.

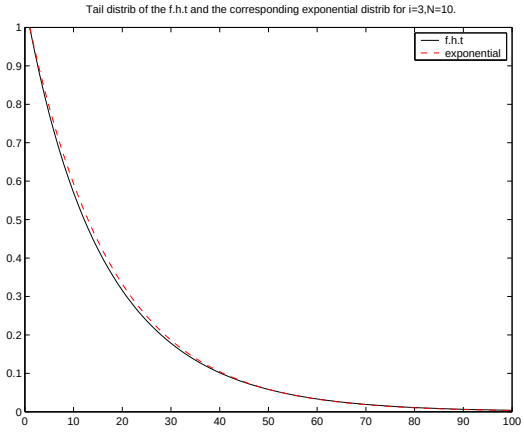
We recall that $P(\tau(i) > m) = q_{-i}\mathbf{K}_{-i}^{m-1}\mathbf{1}$. Hence, by taking $l = m - 1$ we shall obtain $P(\tau(i) > m) \leq (1 - p_i w_1)^{m-1} q_{-i}\mathbf{1}$, or finally, $P(\tau(i) > m) \leq (1 - q_i)(1 - p_i w_1)^{m-1}$.



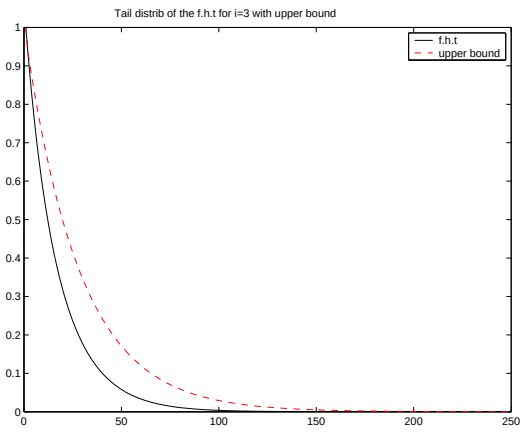
(a) Tail distributions of the f.h.t-mesh



(b) Tail distributions of the f.h.t



(c) F.h.t (solid) and corresponding exponential (dashed)



(d) Upper bound for the tail of the f.h.t

Figure 2: Tail distributions, exponential approximation and the upper bound

For the second of the inequalities we note that $w_i \geq w_1$, so $1 - q_i \leq 1 - p_i w_1$, which readily gives $P(\tau(i) > m) \leq (1 - p_i w_1)^m$. But $(1 - p_i w_1)^m \leq \exp\{-m(p_i w_1)\}$, since $\exp\{-x\} \geq 1 - x$, $\forall x$, and the proof is completed. $\square$

Remark: We note that the last of the inequalities in Proposition 3.7 holds also for the exponential distribution $\mu(i)$, having mean equal to $E[\tau(i)]$. That is, $P(\mu(i) > m) \leq \exp\{-m(p_i w_1)\}$, $\forall m > 0$. To see why it is so, note that $\lambda_i \leq \lambda_1 = 1 - w_1$, so $E[\tau(i)] = 1/[p_i(1 - \lambda_i)] \leq 1/(p_i w_1)$ or $1/E[\tau(i)] \geq p_i w_1$. Now, it suffices to say that $P(\mu(i) > m) = \exp\{-m/E[\tau(i)]\} \leq \exp\{-m(p_i w_1)\}$.

3.5.2 The Variance

In this section, we derive a formula for the variance of the f.h.t for the IMS which, as the mean, will be a function of p_i and λ_i and it will also depend on an extra term, that is Z_{ii}.

Theorem 3.8 *Let Z be the fundamental matrix associated to the IMS kernel $\mathbf{K}$. Then, the variance of the first hitting time takes the form:*

$$Var[\tau(i)] = \frac{2Z_{ii}(1 - \lambda_i) - 3p_i(1 - \lambda_i) + 2p_i - 1}{p_i^2(1 - \lambda_i)^2}, \forall i \in \Omega,$$

with $\lambda_n := 0$.

Proof: Already knowing the expectation of the f.h.t reduces the problem of computing the variance to finding $E(\tau(i)^2)$. This is given by (2.7):

$$E[\tau(i)^2] = 1 + \frac{2}{p_i}\sum_j q_j(Z_{ii}^2 - Z_{ji}^2) - \frac{1}{p_i}\sum_j q_j(Z_{ii} - Z_{ji}) + \frac{2Z_{ii}}{p_i^2}\sum_j q_j(Z_{ii} - Z_{ji}).$$

We can rewrite the above as:

$$E[\tau(i)^2] = 1 + \frac{2}{p_i}(Z_{ii}^2 - \sum_j q_j Z_{ji}^2) - \frac{1}{p_i}(Z_{ii} - \sum_j q_j Z_{ji}) + \frac{2Z_{ii}}{p_i^2}(Z_{ii} - \sum_j q_j Z_{ji}). \tag{3.4}$$

Let us note that $\sum_j q_j Z_{ji} = \sum_j \mathbf{K}_{nj} Z_{ji} = (KZ)_{ni}$. Also, recall that

$$KZ = Z + P - \mathbf{I}. \tag{3.5}$$

Thus, $\sum_j q_j Z_{ji} = Z_{ni} + p_i - \delta_{ni}$. Similarly, $\sum_j q_j Z_{ji}^2 = \sum_j \mathbf{K}_{nj} Z_{ji}^2 = (KZ^2)_{ni}$. Also, from (3.5) it follows that $KZ^2 = Z^2 + P - Z$ so $\sum_j q_j Z_{ji}^2 = Z_{ni}^2 + p_i - Z_{ni}$. Let us for now only consider the case $i < n$. Doing the necessary replacements in (3.4) we get

$$E[\tau(i)^2] = 1 + \frac{2}{p_i}(Z_{ii}^2 - Z_{ni}^2 - p_i + Z_{ni}) - \frac{1}{p_i}(Z_{ii} - Z_{ni} - p_i + \delta_{ni}) + \frac{2Z_{ii}}{p_i^2}(Z_{ii} - Z_{ni} - p_i + \delta_{ni})$$

By regrouping and cancellations one will further get

$$E[\tau(i)^2] = \frac{2}{p_i}(Z_{ii}^2 - Z_{ni}^2) - \frac{3}{p_i}(Z_{ii} - Z_{ni}) + \frac{2Z_{ii}}{p_i^2}(Z_{ii} - Z_{ni}) + \frac{\delta_{ni}}{p_i}(\frac{2Z_{ii}}{p_i} - 1) \tag{3.6}$$

For $i = n$ (3.6) becomes $E[\tau(n)^2] = (2Z_{nn} - p_n)/p_n^2$, so $Var[\tau(n)] = E[\tau(n)^2] - (E[\tau(n)])^2 = (2Z_{nn} - p_n)/p_n^2 - 1/p_n^2$ or finally,

$$Var[\tau(n)] = \frac{2Z_{nn} - p_n - 1}{p_n^2},$$

which is what I wanted since $\lambda_n = 0$.

Therefore, we can now only consider $i < n$ and let us rewrite (3.6) again, in this case, for clarity:

$$E[\tau(i)^2] = \frac{2}{p_i}(Z_{ii}^2 - Z_{ni}^2) - \frac{3}{p_i}(Z_{ii} - Z_{ni}) + \frac{2Z_{ii}}{p_i^2}(Z_{ii} - Z_{ni}). \tag{3.7}$$

As in section 2, we shall use again the spectral decomposition theorem, for Z^2 this time:

$$Z_{li}^2 = p_i + \sum_{k=1}^{n-1} \frac{1}{(1-\lambda_k)^2} v_{kl} u_{ki}, \forall\, l, i.$$

Therefore, we have

$$Z_{ii}^2 - Z_{ni}^2 = \sum_{k=1}^{n-1} \frac{1}{(1-\lambda_k)^2} u_{ki}(v_{ki} - v_{kn}),$$

which, just as before, leads to

$$Z_{ii}^2 - Z_{ni}^2 = \frac{u_{ii}(v_{ii} - v_{in})}{(1-\lambda_i)^2} = \frac{1}{(1-\lambda_i)^2}. \tag{3.8}$$

It is also noted that $Z_{ii} - Z_{ni} = p_i E_n[\tau(i)]$. At the same time, from Proposition 3.6, $E_n[\tau(i)] = E[\tau(i)] = 1/[p_i(1-\lambda_i)]$. Therefore,

$$Z_{ii} - Z_{ni} = \frac{1}{1-\lambda_i}. \tag{3.9}$$

Now using (3.8) and (3.9) in (3.7) we obtain

$$E[\tau(i)^2] = \frac{2}{p_i(1-\lambda_i)^2} - \frac{3}{p_i(1-\lambda_i)} + \frac{2Z_{ii}}{p_i^2(1-\lambda_i)}.$$

Or

$$E[\tau(i)^2] = \frac{2Z_{ii}(1-\lambda_i) - 3p_i(1-\lambda_i) + 2p_i}{p_i^2(1-\lambda_i)^2}.$$

Now, as $Var[\tau(i)] = E[\tau(i)^2] - E[\tau(i)]^2$ and $E[\tau(i)] = 1/(p_i(1-\lambda_i))$, it is immediate that

$$Var[\tau(i)] = \frac{2Z_{ii}(1-\lambda_i) - 3p_i(1-\lambda_i) + 2p_i - 1}{p_i^2(1-\lambda_i)^2}. \quad \square$$

3.5.3 Bounds for the variance

Two corollaries of Theorem 3.8 will offer bounds on the variance of the f.h.t.

By bounding the term Z_{ii} in Theorem 3.8, we obtain Corollary 3.9, which gives bounds for the variance mainly in terms of the expectation of the f.h.t:

Corollary 3.9 *Let* $\mathbf{K}$ *be the IMS kernel and let us denote* $E_i := E[\tau(i)]$, *for any* $i \in \Omega$. *Then,*

$$E_i(E_i - 1) \leq E_i[(1 + 2q_i)E_i - 3] \leq Var[\tau(i)] \leq E_i[\frac{2(1 + q_i)}{w_1 p_i} - E_i - 3]$$

with equality if $w_i = w_{i-1} = \ldots = w_1$.

Proof: For the proof we first need to prove the following lemma:

Lemma 3.10

$$\frac{1 + q_i - p_i}{1 - \lambda_i} \leq Z_{ii} \leq \frac{1 + q_i - p_i}{w_1}$$

with equality if and only if $w_i = w_{i-1} = \ldots = w_1$.

Proof of the lemma: We recall that

$$Z_{ii} = p_i + \sum_{k=1}^{n-1} \frac{1}{1 - \lambda_k} v_{ki} u_{ki}.$$

As $1/(1 - \lambda_k) = 1 + \lambda_k/(1 - \lambda_k)$, we can rewrite Z_{ii} as

$$Z_{ii} = p_i + \sum_{j=1}^{n-1} v_{ki} u_{ki} + \sum_{k=1}^{n-1} \frac{\lambda_k}{1 - \lambda_k} v_{ki} u_{ki} = 1 + \sum_{k=1}^{n-1} \frac{\lambda_k}{1 - \lambda_k} v_{ki} u_{ki} = 1 + \sum_{k=1}^{i} \frac{\lambda_k}{1 - \lambda_k} v_{ki} u_{ki}.$$

As $1/(1 - \lambda_i) \leq 1/(1 - \lambda_j) \leq 1/(1 - \lambda_1), \forall\, 1 \leq j \leq i$, we get

$$1 + \frac{1}{1 - \lambda_i} \sum_{k=1}^{i} \lambda_k v_{ki} u_{ki} \leq Z_{ii} \leq 1 + \frac{1}{1 - \lambda_1} \sum_{k=1}^{i} \lambda_k v_{ki} u_{ki}.$$

But as we have already seen before,

$$\sum_{k=1}^{i} \lambda_k v_{ki} u_{ki} = \sum_{k=1}^{n-1} \lambda_k v_{ki} u_{ki} = \mathbf{K}_{ii} - p_i = q_i + \lambda_i - p_i,$$

and therefore,

$$1 + \frac{q_i + \lambda_i - p_i}{1 - \lambda_i} \leq Z_{ii} \leq 1 + \frac{q_i + \lambda_i - p_i}{1 - \lambda_1}$$

Or

$$\frac{1+q_i-p_i}{1-\lambda_i} \leq Z_{ii} \leq \frac{1+q_i-p_i+\lambda_i-\lambda_1}{1-\lambda_1}.$$

The lemma is proved if, for the right hand side term, we use $1-\lambda_1 = w_1$ and $\lambda_i \leq \lambda_1$. Clearly, equality on both sides is obtained if and only if $w_i = w_{i-1} = \ldots = w_1$.

Going back to the proof of Corollary 3.9, we note that, starting from the left side, the first inequality is trivial since $E_i \geq 1/q_i$. Also, proving that $E_i[(1+2q_i)E_i - 3] \leq Var[\tau(i)]$ is just a matter of applying Lemma 3.10 and regrouping the terms.

For the upper bound we notice that $w_1 = 1-\lambda_1 \leq 1-\lambda_i$ which gives $p_i \leq p_i(1-\lambda_i)/w_1$ which when combined with the upper bound for Z_{ii} will give

$$Z_{ii}(1-\lambda_i) + p_i \leq \frac{(1+q_i-p_i)(1-\lambda_i)}{w_1} + \frac{p_i(1-\lambda_i)}{w_1} = \frac{(1+q_i)(1-\lambda_i)}{w_1}.$$

But according to Theorem 3.8,

$$Var[\tau(i)] = \frac{2Z_{ii}(1-\lambda_i) - 3p_i(1-\lambda_i) + 2p_i - 1}{p_i^2(1-\lambda_i)^2}.$$

Hence

$$Var[\tau(i)] \leq \frac{2(1+q_i)(1-\lambda_i)/w_1 - 3p_i(1-\lambda_i) - 1}{p_i^2(1-\lambda_i)^2},$$

which easily turns into the pursued upper bound since, from Theorem 3.4, $E_i = 1/[p_i(1-\lambda_i)]$. The equality case shows up if $\lambda_i = \lambda_{i-1} = \ldots = \lambda_1$ which is equivalent to $w_i = w_{i-1} = \ldots = w_1$.

The bounds given by Corollary 3.9 can be further simplified, but weakened at the same time, if one uses the known lower bound for E_i on the left and maximizes the upper bound with respect to E_i. Thus, one gets:

Corollary 3.11 *If $M_i := 1/\min\{q_i, p_i\}$, for any $i \in \Omega$, then*

$$M_i(M_i-1) \leq M_i[M_i(1+2q_i)-3] \leq Var[\tau(i)] \leq \Big(\frac{1+q_i}{w_1 p_i} - \frac{3}{2}\Big)^2.$$

Proof: Obviously, for the lower bounds we apply inequality $E_i \geq M_i$ to the previous corollary. To prove the upper bound, we refer again to Corollary 3.9 and for simplicity, let us denote

$$2\frac{(1+q_i)}{w_1 p_i} - 3 := a.$$

Then, Corollary 3.9 gives $Var[\tau(i)] \leq E_i(a-E_i)$. One consequence of this is that $a \geq E_i > 0$, since the variance is a positive number. Hence, we could maximize function $f(x) := x(a-x)$ on $(0, a)$. As the maximum value of f is obtained for $x = a/2$, we conclude that $f(E_i) \leq a^2/4$, which is the upper bound.

4 Comparison between the IMS and general Metropolis-Hastings kernels

We have seen that for the IMS the mean f.h.t is always bounded below by $1/p_i$, for all proposal probabilities q. We shall prove that for more general Metropolis kernels, the mean f.h.t can be lower than $1/p_i$, and thus show formally, what was otherwise clear intuitively, that, because of its independence from the current state, the IMS kernel can be inferior to other samplers in terms of speed of hitting a certain state.

Firstly, we recall that a Metropolis-Hastings kernel R, induced by a proposal stochastic matrix Q is of the form

$$\mathbf{R}_{ij} = Q_{ij} \min\{1, \frac{Q_{ji}p_j}{Q_{ij}p_i}\},$$

for any $i \neq j$ (Hastings, 1970).

Theorem 4.1 *Let Q be a stochastic proposal matrix satisfying the condition*

$$\frac{Q_{ji}}{p_i} \geq 1, \frac{Q_{ij}}{p_j} \geq 1, \forall i, \forall j \neq i.$$

Let R be the Metropolis-Hastings kernel associated to the proposal Q and the target probability p. Then, for any initial distribution q,

$$E_q^Q[\tau(i)] \leq 1 + \frac{1 - q_i}{p_i}, \ \forall i \in \Omega,$$

with equality if Q is the stationary matrix.

Proof: Let $i \in \Omega$. As $Q_{ji} \geq p_i$ and $Q_{ij} \geq p_i$, it follows that $R_{ji} = \min\{Q_{ji}, Q_{ij}p_i/p_j\} \geq p_i, \forall j \neq i$. This implies that $1 - R_{ji} \leq 1 - p_i$ or $R_{j.}\mathbf{1} \leq 1 - p_i$, since the sum of all elements on row j of R is 1. As the previous inequality holds true for all $j \neq i$, we get $R_{-i}1 \leq (1 - p_i)1$ or equivalently $(\mathbf{I} - R_{-i})1 \geq p_i 1$.

As seen before, in section 2, the inverse of $\mathbf{I} - R_{-i}$ always exists and it is equal to $\sum_m R_{-i}^m$ (see [3]), and therefore it is true that $(\mathbf{I} - R_{-i})^{-1} \geq 0$. This said, we can multiply the inequality $(\mathbf{I} - R_{-i})1 \geq p_i 1$ by $(\mathbf{I} - R_{-i})^{-1}$ and get $(\mathbf{I} - R_{-i})^{-1}\mathbf{1} \leq (1/p_i)\mathbf{1}$ which is immediately equivalent to $q_{-i}(\mathbf{I} - R_{-i})^{-1}\mathbf{1} \leq (1 - q_i)/p_i$ or finally,

$$E_q^Q[\tau(i)] \leq 1 + \frac{1 - q_i}{p_i},$$

where we have used formula (2.1) for the mean f.h.t when starting from q. We have equality if $R_{ji} = p_i, \forall j \neq i$, which is fulfilled if Q equals the stationary matrix. Naturally, there are also other Q's that accomplish equality, the condition being that either $Q_{ji} = p_i$ or $Q_{ij} = p_j$, $\forall j \neq i$. □

Combining Theorem 3.4 and Theorem 4.1, one gets:

Corollary 4.2 *For any initial distribution q and Q satisfying the assumption in Theorem 4.1,*

$$E_q^Q[\tau(i)] \leq \max\{\frac{1}{p_i}, \frac{1}{q_i}\} \leq E_q^{IMS}[\tau(i)],$$

where we denoted by $E_q^{IMS}[\tau(i)]$ the average f.h.t of the IMS kernel associated to q and p.

Proof: If using the two specified theorems the proof is immediate since, obviously, $1 + (1 - q_i)/p_i \leq \max\{1/p_i, 1/q_i\}$, with equality if and only if $q_i = p_i$ or, in other words, if i is an exactly-informed state for q. □

We would like to note that there are known examples of samplers that satisfy the condition in Theorem 4.1. Such a sampler is the *"Metropolized Gibbs Sampler"* (Liu, 2001) or simply MGS. One of the most recent applications of this sampler is described in Tu, Zhu(2003).

For the MGS, the proposal matrix Q is defined as : $Q_{ij} = p_j/(1-p_i)$, $\forall i \neq j$. Thus, obviously, Q satisfies the condition in Theorem 4.1. Naturally, in practice the sampler will have to use an approximative version of Q, since p is not usually known.

Interestingly, the MGS can be viewed as a particular case of the IMS. To see this, let us remark that after metropolizing Q through the usual acceptance-rejection mechanism, one gets the transition kernel having elements:

$$\mathbf{R}_{ij} = \begin{cases} \frac{p_j}{1-pi} & \text{if } i < j, \\ 1 - \sum_{k \neq i} \mathbf{R}_{ki} & \text{if } i = j, \\ \frac{p_j}{1-pj} & \text{if } i > j. \end{cases}$$

Without loss of generality, we assumed that $p_1 \leq p_2 \leq \ldots \leq p_n$. With this assumption and by denoting with

$$q_i = \frac{p_i}{1-p_i}, \forall i < n \text{ and } q_n = 1 - \sum_{i<n} q_i,$$

we note that $\mathbf{R}$ is equal to the IMS kernel matrix corresponding to p and q for $w_1 \leq w_2 \leq \ldots \leq w_n$. Therefore, if using as initial distribution the newly defined q, all the previous results pertaining to the IMS apply also to the MGS.

To conclude, we would like to point out the fact that, conceptually, there is a different way of arriving at the formulas of the expectation and variance that are based on the eigen-elements of **K**. To briefly describe the method, let **K** be an ergodic kernel and $\{\lambda_j\}_{0\leq j\leq n-1}$ be its eigenvalues with the corresponding right and left eigenvectors v_j, u_j such that $u_k^T v_l = \delta_{kl}, \forall k,l \in \Omega$. As **K** is a stochastic matrix with stationary probability p, one has $\lambda_0 = 1$ and $v_0 = \mathbf{1}, u_0 = p$ respectively.

Also, let us denote by $\{b_l\}_{l\geq 0}, b_l := (\mathbf{K}^l)_{ii}$, and let $\{a_l\}_{l\geq 0}$ be defined as $a_0 = b_0 = 1, a_l = b_{l-1} - b_l, \forall l > 0$.

The next step is to define the sequence $\{f_k\}_{k\geq 0}$ recursively, by the formula:

$$f_k = \sum_{l=0}^{k-1} a_{k-l} f_l, f_0 = 1.$$

Further more, we denote

$$F_{j,m} := \sum_{k=0}^{m} \lambda_j^{m-k} f_k, 0 \leq j \leq i, \forall m \geq 0.$$

Theorem 4.3 *Using the previous notations,* $\forall m \geq 0$ *and for all initial distributions* q*,*

$$i) \quad P(\tau(i) > m+2) = 1 - q_i - \sum_{k=0}^{n-1} \lambda_k u_{ki} F_{k,m} \sum_{l\neq i} q_l v_{kl}$$

$$ii) \quad \sum_{m=0}^{\infty} F_{k,m} = \frac{1}{p_i(1-\lambda_k)}, \forall k \geq 1$$

$$iii) \quad \sum_{m=0}^{\infty} (m+2) F_{k,m} = \frac{1}{p_i(1-\lambda_k)} \Big(1 + \frac{1}{1-\lambda_k} + \frac{1}{p_i} \sum_{j=1}^{n-1} \frac{1}{1-\lambda_j} v_{ji} u_{ji}\Big), \forall k \geq 1.$$

For proof and details refer to Maciuca(2004). This result allows the computation of the expectation and second moment in a straightforward way, and could also offer insights into properties of the tail distribution for the IMS which, as we saw, has a readily available eigen-analysis.

Acknowledgment

This work was supported by an NSF grant (IIS-0244763) and a Sloan fellowship.

References

[1] Abadi, M. and Galves, A. (2001), Inequalities for the occurrence times of rare events in mixing processes. The state of the art, Markov Proc. Relat. Fields, **7**, 97-112.

[2] Aldous, D. and Fill, J. Monograph on Markov Chain analysis, (chapters 2,3). Available from David Aldous's web page at UC Berkeley Statistics Department.

[3] Bremaud, P. (1999), Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues, Springer, New York.

[4] Cho, G.E. and Meyer, C.D. (1999), Markov Chain Sensitivity Measured by Mean First Passage Times, NCSU Technical Report.

[5] Diaconis, P. and Hanlon, P. (1992), Eigen Analysis for Some Examples of the Metropolis Algorithm, Contemporary Mathematics, **138**, 99-117.

[6] Diaconis, P. and Saloff-Coste, L. (1998), What Do We Know about the Metropolis Algorithm?, Journal of Computer and System Sciences, **57**, 20-36.

[7] Hastings, W.K. (1970), Monte Carlo Sampling Methods using Markov Chains and Their Applications, Biometrika, **57**, 97-109.

[8] Kemeny, J. G. and Snell, J. L. (1976), Finite Markov Chains, Springer Verlag.

[9] Liu, J.S. (2001), Monte Carlo Strategies in Scientific Computing, Springer Verlag.

[10] Liu, J.S. (1996), Metropolized Independence Sampling with Comparisons to Rejection Sampling and Importance Sampling, Statistics and Computing, **6**, 113-119.

[11] Mengersen, K.L. and Tweedie, R.L. (1994), Rates of Convergence of the Hastings and Metropolis Algorithms, Annals of Statistics, **24**, 101-121.

[12] Pearl, J. (1985), Heuristics: Intelligent Search Strategies for Computer Problem Searching, Addison Wesley.

[13] Maciuca, R. (2004), On the Tail Distribution of the First Hitting Time for the Independence Metropolis Sampler, Technical Report, Dept. of Statistics, UCLA.

[14] Smith, R.L. and Tierney, L. (1996), Exact Transition Probabilities for Metropolized Independence Sampling, Technical Report, Dept. of Statistics, Univ. of North Carolina.

[15] Sodal, S. (2001), Entry, Exit and Scrapping Decisions with Investment Lags: A Series of Investment Models Based on a New Approach, 5th Ann. Intl Conference on Real Options, UCLA.

[16] Sullivan, F. (2000), Great algorithms of 20th century scientific computing, Computing in Science and Engineering 2, Special Issue, 2000.

[17] Tu, Z.W. and Zhu, S.C. (2003), Parsing Images into Regions, Curves, and Curve Groups, submitted.