(To appear in the Journal for Theoretical Probability)

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 some non-independence Metropolis-Hastings algorithms can perform better than the IMS and deriving general lower and upper bounds for the mean first hitting times of a Metropolis-Hastings algorithm.

Keywords: Eigenanalysis, Expectation, First Hitting Time, Independence Metropolis Sampler, Metropolis-Hastings.

The authors are with the Department of Statistics, 8125 Math. Science Bldg, Box 951554, University of California, Los Angeles, CA 90095.

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

1 Introduction

In this paper, we study a special case of the celebrated 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 $p = (p_1, p_2, \dots, p_n)$, by drawing samples from a more tractable probability $q = (q_1, q_2, \dots, q_n)$.

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 and Hanlon [3] and Liu [9] 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). He also compared the IMS with other two well known sampling techniques, rejection sampling and importance sampling. By making use of Liu's results, Smith and Tierney [12] obtained exact m-step transition probabilities for IMS, for both discrete and continuous state spaces.

In the continuous case, if denoting by

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

Mengersen and Tweedie [10] showed that if r^* is strictly less than 1, the chain is uniformly ergodic, while if r^* is equal to 1, the convergence is not even geometric anymore. Similar results were obtained by Smith and Tierney. 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 the 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 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. This is particularly appropriate when dealing with stochastic search algorithms, when the focus could be on finding individual states rather than on the global convergence of the chain. For instance, in computer vision problems, one is often searching for the most probable interpretation of a scene and, to this end, various Metropolis-Hastings type algorithms can be employed. See Tu and Zhu [13] for examples and discussions. In such a context, knowledge of the behavior of the first hitting time of some states, like the modes of the posterior distribution of a scene given the input images, is of interest.

The IMS was thoroughly studied, but most of the analysis focused on its convergence properties. Here, we analyze its first hitting time and derive formulas for its expectation and variance. These formulas are expressed mostly in terms of the eigenvalues of the transition kernel.

- We first review some general 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) $\dfrac{1}{\min\{q_i, p_i\}} \le E[\tau(i)] \le \dfrac{1}{\min\{q_i, p_i\}} \dfrac{1}{1 - \|p-q\|_{TV}}$,

  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 results by 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 :

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

Next, in section 3.5 and section 3.6 we focus on the tail distribution and the variance of the f.h.t for the IMS and we determine:

- An exponential upper bound on the tail: $P(\tau(i) > m) \leq \exp\{-m(p_i w_1)\}$, $\forall m > 0$.
- If Z denotes the fundamental matrix associated with the IMS kernel then:

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

  Various bounds on the variance are also presented.

Further, 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.

- We conclude by presenting lower and upper bounds on the hitting times for general Metropolis-Hastings kernels.

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 mean first hitting time of i for the Markov chain governed by* $\mathbf{K}$.

For any i, 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 immediate that $P(\tau(i) > m) = q_{-i}\mathbf{K}_{-i}^{m-1}\mathbf{1}$, where $\mathbf{1} := (1, 1, ..., 1)'$. 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 implied by the sub-stochasticity of $\mathbf{K}_{-i}$ and the irreducibility of $\mathbf{K}$ (Bremaud [2]).

More generally, the mean f.h.t of a subset A of Ω is given by

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

A different route is to consider the first hitting times if starting from a fixed $j \neq i$. Here, we should define a different stopping time, by not counting starting from j as an initial step, but for simplicity, we will use the same notation and refer to $E_j[\tau(i)]$ as the mean f.h.t of i when starting from state j. Then, for all $j \neq i$, one has $E_j[\tau(i)] = (Z_{ii} - Z_{ji})/p_i$. Z denotes the fundamental matrix, which, we recall, is defined to be $Z = (\mathbf{I} - \mathbf{K} + P)^{-1}$, where P denotes the matrix having all rows equal to p. When starting from q instead from a fixed state j, one has:

$$E_q[\tau(i)] = 1 + \sum_{j \neq i} q_j E_j[\tau(i)] = 1 + \frac{1}{p_i} \sum_{j \neq i} q_j (Z_{ii} - Z_{ji}). \tag{2.3}$$

For the rest of the paper we shall drop the subscript q whenever this will not create any notation confusion.

The variance of the f.h.t can also be derived from the fundamental matrix Z. It is known that the second moment of $\tau(i)$, when starting from j, is determined by:

$$E_j[\tau(i)]^2 = \frac{2}{p_i}(Z^2_{ii} - Z^2_{ji}) - \frac{1}{p_i}(Z_{ii} - Z_{ji}) + \frac{2}{p_i^2} Z_{ii}(Z_{ii} - Z_{ji}), \forall j \neq i, \tag{2.4}$$

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

$$E[\tau(i)^2] = 1 + \frac{2}{p_i}\sum_j q_j(Z^2_{ii} - Z^2_{ji}) - \frac{1}{p_i}\sum_j q_j(Z_{ii} - Z_{ji}) + \frac{2Z_{ii}}{p_i^2}\sum_j q_j(Z_{ii} - Z_{ji}), \tag{2.5}$$

which readily leads to a formula for the variance. For more on the properties of the fundamental matrix Z and its connections to hitting times refer to Kemeny and Snell [6].

Next, we will show how knowing the eigen-structure of the transition matrix allows the direct computation of the mean f.h.t.

Let $\{\lambda_j\}_{0 \leq j \leq n-1}$ be the eigenvalues of $\mathbf{K}$ and let $v_k = \{v_{kl}\}_{0 \leq l \leq n-1}$, and $u_k = \{u_{kl}\}_{0 \leq l \leq n-1}$ be their corresponding right and left eigenvectors, 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 fix $v_0 = \mathbf{1}$ and $u_0 = p$ respectively. Moreover, all the eigenvalues have real values and $|\lambda_j| < 1, \forall j > 0$.

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: Using (2.3),

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{j \neq i} q_j(Z_{ii} - Z_{ji}). \tag{2.6}$$

Let us recall that Z and $\mathbf{K}$ share the same system of eigenvectors, while the eigenvalues of Z are $\beta_0 = 1, \beta_j = 1/(1-\lambda_j), \forall 1 \leq j \leq n-1$. Therefore, we can apply the spectral decomposition theorem (see Bremaud [2]), 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,

$$Z_{ii} - Z_{ji} = \sum_{k=1}^{n-1} \frac{1}{1-\lambda_k}(v_{ki} - v_{kj})u_{ki}. \tag{2.7}$$

From (2.7) and (2.6), 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

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki} \sum_{j\neq i} q_j(v_{ki} - v_{kj}). \tag{2.8}$$

Noting that $\sum_{j\neq i} q_j(v_{ki} - v_{kj})$ can be rewritten as $v_{ki} - \sum_l q_l v_{kl}$, the first part of the proof is completed. For the second part, 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 for λ_k, we get $\sum_l q_l v_{kl} = \lambda_k v_{kj}$ and by plugging this into the general formula just proved,

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

Or

$$E[\tau(i)] = 1 + \frac{1}{p_i}\sum_{k=1}^{n-1} \frac{1}{1-\lambda_k} u_{ki}(v_{ki} - v_{kj}) + \frac{1}{p_i}\sum_{k=1}^{n-1} u_{ki} v_{kj}. \tag{2.9}$$

We have to consider two cases:

i) $j = i$. In this case, $\sum_{k=1}^{n-1} u_{ki} v_{kj} = \sum_{k=0}^{n-1} u_{ki} v_{ki} - p_i = 1 - p_i$ since $\sum_{k=0}^{n-1} u_{ki} v_{ki} = 1$. Therefore, from (2.9) 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 (2.9)

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

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 [9]). 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 $q = (q_1, q_2, \dots, q_n)$ 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. In section 3.3, we show 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 want 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* 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: $i_{\max} = \arg\max_{i\in\Omega}\{\ w_i\ \}$, $i_{\min} = \arg\min_{i\in\Omega}\{\ w_i\ \}$.*

Liu [9] noticed that the transition kernel can be written in a simpler form by reordering the states increasingly according to their informedness. Since 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. 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 $\lambda_k = T_k - w_k \cdot S_k, \forall\ 1 \leq k \leq n-1, \lambda_0 = 1$, and they are decreasing as $\lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} \geq 0$. Moreover, the right eigenvector corresponding to $\lambda_k, k > 0$ is $v_k = (0, \cdots, 0, S_{k+1}, -p_k, \cdots, -p_k)$, with the first $k-1$ entries being 0 and $v_0 = (1, 1, \ldots, 1)'$.*

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 [12] 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 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 eigen-structure of the IMS kernel as well as of Proposition 2.1.

Theorem 3.4 *Assume a Markov chain starting from* q *is simulated according to the IMS transition kernel having proposal* q *and target probability* p*. 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:

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

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 $E[\tau(i)] = \delta_{in}/p_i = 1/[p_n(1 - \lambda_n)]$, while for $i < n$ one has

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

Using the eigen-analysis for the IMS, we can write $u_{ii}(v_{ii} - v_{in}) = (S_{i+1} - (-p_i))/S_i = 1$, so the expectation becomes $E[\tau(i)] = 1/[p_i(1 - \lambda_i)]$, 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, we only need to show 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 follows quickly since for any $j < i$, $w_j \leq w_i \iff q_j \leq p_j w_i$.

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. Equality is attained for $p_j = q_j, \forall j \in [i, k]$, 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 $\lambda_i \leq T_{k+1} - S_{k+1}$. Because $\lambda_i \leq \lambda_{k+1}$, it suffices to prove that $\lambda_{k+1} \leq T_{k+1} - S_{k+1}$, or $T_{k+1} - w_{k+1}S_{k+1} \leq T_{k+1} - S_{k+1}$, which is 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. The following corollary holds true.

Corollary 3.5 *Let $A \subset \Omega$ of the form $A = \{i+1, i+2, \ldots, i+k\}$, with $w_1 \leq w_2 \leq \ldots \leq w_n$. Denote $p_A := p_{i+1} + p_{i+2} + \ldots + p_{i+k}, q_A := q_{i+1} + q_{i+2} + \ldots + q_{i+k}, w_A := q_A/p_A$ and $\lambda_A := (q_{i+1} + \ldots + q_n) - (p_{i+1} + \ldots + p_n)w_A$. Then i) and ii) from Theorem 3.4 hold ad-literam with i replaced by A.*

Proof: We will only prove part *i)* since the proof of *ii)* is analogous to the one in Theorem 3.4.

Let $A = \{i+1, i+2, \ldots, i+k\}$. We notice 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 mean f.h.t of A reduces to computing the mean 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 would be q restricted to 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 using Theorem 3.4 for Ω_A, $E_\Omega[\tau(A)] = E_{\Omega_A}[\tau(\{A\})] = 1/[p_A(1 - \lambda_A)]$. We used the subscripts Ω or Ω_A to indicate which space we are working on. □

In the introduction part we hinted at showing 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 that $w_1 \leq w_2 \leq \ldots \leq w_n$, the following inequalities hold true:*

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

Proof: We saw that

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

i) $j > i$. Then one has $u_{ki}(v_{ki} - v_{kj}) = u_{ki}(v_{ki} - v_{kn}), \forall k > 0$ and therefore,

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

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 too, 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 note that $(v_{j(j+1)} - v_{jj}) u_{ji} = (-p_j - S_{j+1})(-p_i/(S_j S_{j+1})) = p_i/S_{j+1}$. Similarly, $(v_{(j+1)(j+1)} - v_{(j+1)j}) u_{(j+1)i} = S_{j+2}(-p_i/(S_{j+1}S_{j+2})) = -p_i/S_{j+1}$. Hence,

$$E_j[\tau(i)] - E_{j+1}[\tau(i)] = \frac{1}{p_i} \Big(\frac{1}{1-\lambda_j} - \frac{1}{1-\lambda_{j+1}} \Big) \frac{p_i}{S_{j+1}} = \frac{1}{S_{j+1}} \Big(\frac{1}{1-\lambda_j} - \frac{1}{1-\lambda_{j+1}} \Big) \geq 0.$$

Equality case is obtained if $w_j = w_{j+1}$ which implies $\lambda_j = \lambda_{j+1}$. 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

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

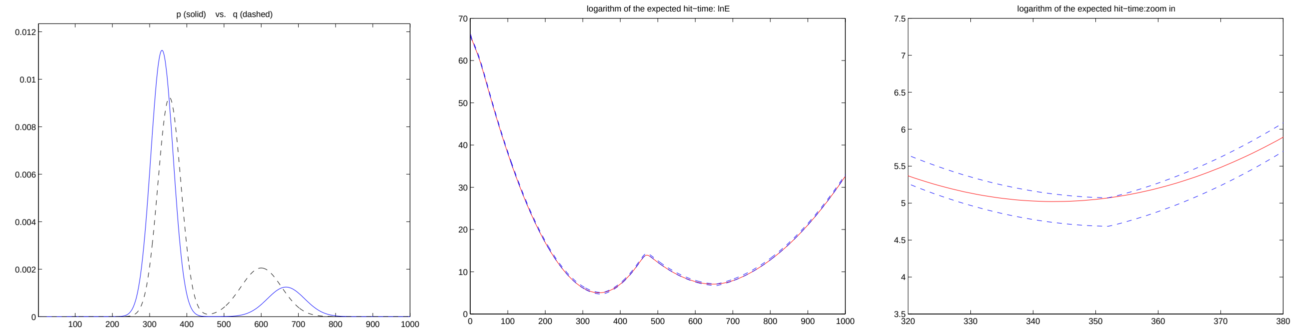
As before, all the terms cancel except for $k = i-1, i$ and analogously,

$$E_{i-1}[\tau(i)] - E_{i+1}[\tau(i)] = \frac{1}{S_i}\Big(\frac{1}{1-\lambda_{i-1}} - \frac{1}{1-\lambda_i}\Big) \geq 0.$$

As $E_{i+1}[\tau(i)] = E_j[\tau(i)] = E[\tau(i)], \forall j > i$, 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. 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.



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

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

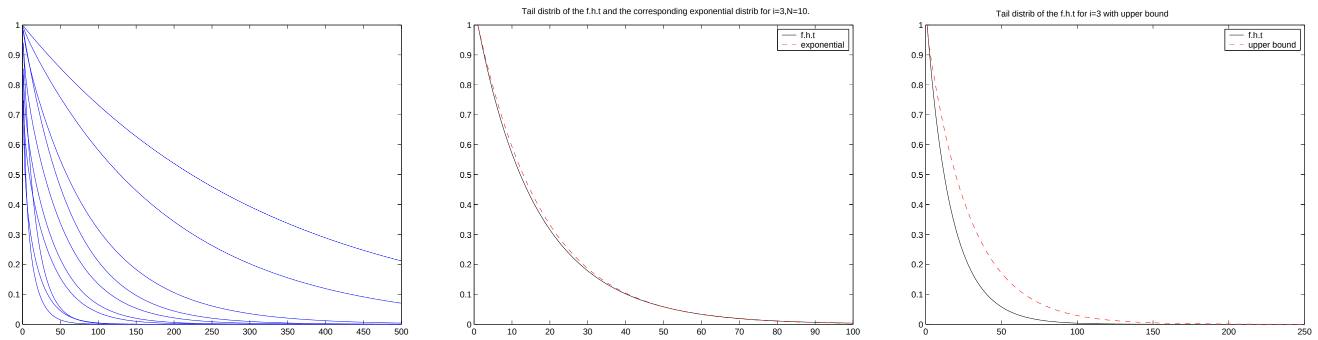
(c) zoom in for (b)

Figure 1: Mean f.h.t and bounds

3.5 Tail Distribution

It is known (Abadi [1]) that the distribution of first hitting times is generally well approximated by an exponential distribution. This can be illustrated on a small example.

Consider a state space with $N = 10$ states, with p and q being discretized mixtures of Gaussians as before. Fig. 2a plots the tail distributions of the f.h.t for all the states of the space. It is apparent that their shapes generally resemble exponential tails. In Fig. 2b, we plotted both the tail distribution of the f.h.t (in solid) and the corresponding exponential distribution (dashed) for an arbitrary state ($i = 3$). Even though we are not able to quantify the approximation error in general, we can give an exponential upper bound on the tail distribution of the f.h.t. The bound is shown in Fig. 2c.



(a) Tail distributions (b) F.h.t and the corresponding exponential distribution (c) Upper bound for the tail

Figure 2: Tails, exponential approximation and the upper bound

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, $\mathbf{K}^l_{-i}\mathbf{1} \leq (1 - p_i w_1)^l \mathbf{1}$, $\forall l$.

We recall that $P(\tau(i) > m) = q_{-i}\mathbf{K}^{m-1}_{-i}\mathbf{1}$. 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, $P(\tau(i) > m) \leq (1 - q_i)(1 - p_i w_1)^{m-1}$.

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. □

Remarks: 1) 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, 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 $1/E[\tau(i)] \geq p_i w_1$. Then, obviously, $P(\mu(i) > m) = \exp\{-m/E[\tau(i)]\} \leq \exp\{-m(p_i w_1)\}$.

2) We can always use Markov's inequality to upper bound the tail distribution with respect to the expectation of the f.h.t of some i. That is, for any k positive integer, one has:

$$P(\tau(i) > kE[\tau(i)]) \leq \frac{1}{k}.$$

If we were to avoid the use of eigenvalues, whose values might be too difficult to compute, we could rely on the upper bound from Theorem 3.4 which combined with the above gives:

$$P(\tau(i) \geq k\lceil \frac{1}{\min\{p_i, q_i\}(1 - ||p - q||)} \rceil) \leq \frac{1}{k}.$$

3.6 Variance

In this section, we derive a formula for the variance of the f.h.t for the IMS. We show that:

Theorem 3.8 *If Z is the fundamental matrix associated to the IMS kernel and using the same notations as before, the variance of the first hitting time of i is given by:*

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

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 (2.5):

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

We can rewrite the above as:

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

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, $\sum_j q_j Z^2_{ji} = \sum_j \mathbf{K}_{nj} Z^2_{ji} = (KZ^2)_{ni}$. From (3.5) it also follows that $KZ^2 = Z^2 + P - Z$, hence $\sum_j q_j Z^2_{ji} = Z^2_{ni} + p_i - Z_{ni}$. By transforming (3.4) we get

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

which further reduces to

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

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 $\lambda_n = 0$.

If $i < n$, let us rewrite (3.6), for clarity:

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

We use the spectral decomposition theorem for Z^2 and obtain

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

Therefore, we have

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

which, just as before, leads to

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

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

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

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

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

3.6.1 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 us denote* $E_i := E[\tau(i)]$, *for any* $i \in \Omega$. *Then,*

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

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

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

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

Proof of the lemma: We use again the identity

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

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

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

Note that $1/(1-\lambda_i) \leq 1/(1-\lambda_j) \leq 1/(1-\lambda_1), \forall\, 1 \leq j \leq i$, which gives

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

Since $\sum_{k=1}^{i}\lambda_k v_{ki}u_{ki} = \sum_{k=1}^{n-1}\lambda_k v_{ki}u_{ki} = \mathbf{K}_{ii} - p_i = q_i + \lambda_i - p_i$, it follows that $1+(q_i+\lambda_i - p_i)/(1-\lambda_i) \leq Z_{ii} \leq 1 + (q_i + \lambda_i - p_i)/(1-\lambda_1)$, or further, $(1+q_i-p_i)/(1-\lambda_i) \leq Z_{ii} \leq (1+q_i-p_i+\lambda_i-\lambda_1)/(1-\lambda_1)$.

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 $p_i \leq p_i(1-\lambda_i)/w_1$. This combined with the upper bound for Z_{ii} will show that

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

Since from Theorem 3.8, $Var[\tau(i)] = [2Z_{ii}(1-\lambda_i) - 3p_i(1-\lambda_i) + 2p_i - 1]/[p_i^2(1-\lambda_i)^2]$, it follows that $Var[\tau(i)] \leq [2(1+q_i)(1-\lambda_i)/w_1 - 3p_i(1-\lambda_i) - 1]/[p_i^2(1-\lambda_i)^2]$, 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 $Var[\tau(i)] \leq E_i(a-E_i)$. Consequently, $a \geq E_i > 0$, and since the maximum value of function $f(x) := x(a-x)$ on $(0,a)$ is obtained for $x = a/2$, we conclude that $f(E_i) \leq a^2/4$, which is the upper bound.

4 IMS vs. a special class of 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 can be written as $\mathbf{R}_{ij} = Q_{ij}\min\{1, Q_{ji}p_j/(Q_{ij}p_i)\}$, for any $i \neq j$ (Hastings [5]).

Theorem 4.1 *Let Q be a stochastic proposal matrix satisfying the condition $Q_{ji} \geq p_i, \forall j \neq i$.*

Then, for any initial distribution q, the Metropolis-Hastings Markov chain that uses Q as a proposal matrix has the property that

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

with equality for Q equal to the stationary matrix.

Proof: Let R be the Metropolis-Hastings kernel associated to the proposal Q and the target probability p. Let $i \in \Omega$. As $Q_{ji} \geq p_i$ and $Q_{ij} \geq p_j$, 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 $R_{j\cdot}\mathbf{1}_{n-1} \leq 1 - p_i$. As the previous inequality holds true for all $j \neq i$, we get that $R_{-i}\mathbf{1} \leq (1-p_i)\mathbf{1}$ or equivalently $(\mathbf{I} - R_{-i})\mathbf{1} \geq p_i 1$.

The inverse of $\mathbf{I} - R_{-i}$ exists and it is equal to $\sum_m R_{-i}^m$ and therefore $(\mathbf{I} - R_{-i})^{-1} \geq 0$. This said, we can multiply the inequality $(\mathbf{I} - R_{-i})\mathbf{1} \geq p_i\mathbf{1}$ by $q_{-i}(\mathbf{I} - R_{-i})^{-1}$ and get $q_{-i}(\mathbf{I} - R_{-i})^{-1}\mathbf{1} \leq (1-q_i)/p_i$ or finally, $E_q^Q[\tau(i)] \leq 1 + (1-q_i)/p_i$, 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 mean f.h.t of the IMS kernel associated to q and p.

Proof: 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. □

The above corollary thus gives a simple way to construct Metropolis-Hastings samplers that would perform better than a corresponding IMS sampler in terms of first hitting times.

It is worth noting that there are known examples of samplers that satisfy the condition in Theorem 4.1. Such a sampler is the *"Metropolized Gibbs Sampler"* (Liu [8]) or simply MGS. A recent application of this sampler is described in Tu and Zhu [13].

For the MGS, the proposal matrix Q is defined as: $Q_{ij} = p_j/(1 - p_i)$, $\forall i \neq j$ which satisfies the above mentioned condition.

Interestingly, the MGS can also 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 assume that $p_1 \leq p_2 \leq \ldots \leq p_n$. Now, if we denote by $q_i := p_i/(1 - p_i), \forall i < n$ and $q_n := 1 - \sum_{i<n} q_i$, we note that $\mathbf{R}$ has the same form as the IMS transition 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.

Remark: The MGS is a modified Gibbs sampler, the main difference being that it will never propose the current state. Thus, it travels through the state space in a more efficient manner. However, a M-H acceptance probability needs to be introduced to maintain the correct invariant distribution. Rejections could still cause the sampler to stay in the same state. Nevertheless, Liu [8] showed that the MGS is more efficient than the ordinary Gibbs sampler in the sense that the asymptotic variance of the estimators based on the Markov chain samples is smaller for the MGS than for the Gibbs sampler. Thus, the expected gain in efficiency would justify *metropolizing* the stationary probability p.

A recent review of various types of "efficiency" definitions for MCMC samplers as well as theoretical results linking these types of efficiency notions can be found in Mira [11]. Our approach is to consider the mean first hitting time as an indicator of efficiency when the focus is on searching for a few states through a finite state space.

5 General bounds for expected f.h.ts for Metropolis-Hastings kernels

It turns out that one can get lower and upper bounds on the expected f.h.t for any Metropolis-Hastings kernel by a reasoning similar to the one in Theorem 4.1 as shown with the result below.

Theorem 5.1 *Let p and Q be the target probability and the proposal matrix respectively for a Metropolis-Hasting sampler. Let $M = \max_{i,j} Q_{ij}/p_j$ and $m = \min_{i,j} Q_{ij}/p_j$. We assume $m > 0$. Then for any initial distribution q, the expected f.h.ts are bounded by*

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

Equality is attained if $Q_{ij} = p_j, \forall i, j$.

Proof: The proof is similar to the one for Theorem 4.1 so it will only be sketched. Firstly one shows that $mp_i \leq K_{ji} \leq Mp_i$ which then leads to $(mp_i)\mathbf{1} \leq (\mathbf{I} - \mathbf{K}_{-i})\mathbf{1} \leq (Mp_i)\mathbf{1}$, which in turn, by an argument analogous to the one in Theorem 4.1, gives $(1-q_i)/Mp_i \leq q_{-i}(\mathbf{I} - \mathbf{K}_{-i})^{-1}\mathbf{1} \leq (1-q_i)/mp_i$. Now using the corresponding identity for expected f.h.t, $E_q^Q[\tau(i)] = 1 + q_{-i}(\mathbf{I} - \mathbf{K}_{-i})^{-1}\mathbf{1}$, one immediately gets the result stated in Theorem 5.1.□

For some particular choices of q, the bounds can be made more "concrete". Two particular choices seem both intuitive and convenient:

i) $q_i = \sum_j \alpha_j Q_{ji}, \forall i, (\sum_j \alpha_j = 1)$

ii) $q_i = \dfrac{\max_j Q_{ji}}{\sum_k \max_j Q_{jk}}, \forall i.$

It is immediate to check that both *i)* and *ii)* are valid probability distributions. The first one is just a linear combination of the elements on the ith column, which in the particular case when the proposal does not depend on the current state (the IMS), reduces to the proposal distribution for the IMS. The second distribution described above would use the maximum value of the proposal matrix on each column as an initial step, after normalization.

Using them as initial distributions for the Markov chain and applying Theorem 5.1 one can derive the corollary below:

Corollary 5.2 *Within the setup from Theorem 5.1, the following hold:*

1) If the initial distribution is given by i) then $1/M \leq p_i E_q^Q[\tau(i)] \leq 1/m, \forall i.$

2) If the initial distribution is given by ii), then $\max\{1/n, 1/M\} \leq p_i E_q^Q[\tau(i)] < 1 + 1/m, \forall i.$

Proof: To prove 1), let us first notice that from the way M and m where defined it follows that $mp_i \leq q_i \leq Mp_i$. Therefore

$$p_i + \frac{1-q_i}{M} \geq p_i + \frac{1-Mp_i}{M} = \frac{1}{M}$$

and analogously

$$p_i + \frac{1-q_i}{m} \leq p_i + \frac{1-mp_i}{m} = \frac{1}{m}$$

which proves 1) by means of Theorem 5.1.

For 2), we first need to show that $mp_i/n < q_i \leq Mp_i, \forall i$. We note that

$$1 = \sum_k Q_{1k} \leq \sum_k \max_j Q_{jk} < \sum_k 1 = n$$

Therefore,

$$\frac{\max_j Q_{ji}}{n} < q_i = \frac{\max_j Q_{ji}}{\sum_k \max_j Q_{jk}} \leq \max_j Q_{ji}$$

Hence $mp_i/n < q_i \leq Mp_i, \forall i$. Now, as for 2), one will get

$$\frac{1}{M} \leq p_i E_q^Q[\tau(i)] < p_i + \frac{1 - p_i m/n}{m} < 1 + \frac{1}{m}$$

The only thing left to prove is that $p_i E_q^Q[\tau(i)] \geq 1/n$. In order to show this, we employ the second basic identity for the kernel $\mathbf{K}$, which is $p\mathbf{K} = p$. If we denote by u the $n-1$ dimensional vector obtained from row i of $\mathbf{K}$ after deleting component K_{ii}, we can write $p_{-i}\mathbf{K}_{-i} + p_i u = p_{-i}$ or $p_{-i}(\mathbf{I} - \mathbf{K}_{-i}) = p_i u$. We note that $\mathbf{K}_{ij} \leq Q_{ij} \leq \max_i Q_{ij} < nq_j, \forall j$.

Therefore, $u < nq_{-i}$ so $p_{-i}(\mathbf{I}-\mathbf{K}_{-i}) < (np_i)q_{-i}$. As before, it follows that $p_{-i}\mathbf{1} < (np_i)q_{-i}(\mathbf{I}-\mathbf{K}_{-i})^{-1}\mathbf{1}$ or $1 - p_i < (np_i)(E_q^Q[\tau(i)] - 1)$. Hence $p_iE_q^Q[\tau(i)] > p_i + (1 - p_i)/n$, implying that $p_iE_q^Q[\tau(i)] > 1/n$ which concludes the proof of the corollary.□

Naturally, because of their generality the bounds developed in this section are quite weak in general, as m and M can take very extreme values in practice, rendering the bounds useless for such cases.

6 Conclusion

We were able to perform a detailed first hitting analysis of one special type of Metropolis-Hastings sampler, the Independence Metropolis Sampler. More practical general non-independence Metropolis-Hastings samplers seem to be too complex to allow for a detailed analysis. In the spirit of this paper, such an analysis could only be done if the eigenstructure of the kernel matrix would be available. This is typically not the case for most of the practical applications. However, even when the eigenstructure is unknown, insights into the behavior of mean first hitting time are possible, as seen in Theorem 4.1. Also, we have derived lower and upper bounds for the expected first hitting time for general Metropolis-Hastings algorithms. We are hoping that future work will allow obtaining better bounds in the case of the tail distribution for the IMS and further for more general cases. This would make our results more useful and amenable for using them in practice.

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

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

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

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

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

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

[8] Liu, J.S. (1996). Metropolized Gibbs sampler: an improvement. Technical report, Dept. Statistics, Stanford Univ.

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

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

[11] Mira, A. (2002). Ordering and improving the performance of Monte Carlo Markov Chains. Statistical Science 16, 340-350.

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

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