Illinois Journal of Mathematics
Volume 50, Number 1, Spring 2006, Pages 67–91
S 0019-2082

# EXAMPLES COMPARING IMPORTANCE SAMPLING AND THE METROPOLIS ALGORITHM

FEDERICO BASSETTI AND PERSI DIACONIS

Abstract. Importance sampling, particularly sequential and adaptive importance sampling, have emerged as competitive simulation techniques to Markov-chain Monte-Carlo techniques. We compare importance sampling and the Metropolis algorithm as two ways of changing the output of a Markov chain to get a different stationary distribution.

## 1. Introduction

Let $\mathcal{X}$ be a finite set and $\pi(x)$ be a probability on $\mathcal{X}$. For $f : \mathcal{X} \to \mathbb{R}$, we want to approximate

$$\mu = \sum_x f(x)\pi(x). \tag{1.1}$$

Suppose we have available a reversible Markov chain $K(x, y)$ on $\mathcal{X}$ with stationary distribution $\sigma(x) > 0$ for all $x$ in $\mathcal{X}$. Two classical procedures are available.

*Metropolis.* Change the output of the $K$ chain to have stationary distribution $\pi$ by constructing

$$M(x,y) = \begin{cases} K(x,y)A(x,y), \quad A(x,y) := \min\left(\frac{\pi(y)K(y,x)}{\pi(x)K(x,y)}, 1\right), & x \neq y, \\ K(x,x) + \sum_{z\neq x} K(x,z)(1 - A(x,z)), & x = y. \end{cases}$$

Generate $Y_1$ from $\pi$ and then $Y_2, \ldots, Y_N$ from $M(x, y)$. It follows that

$$\hat{\mu}_M = \frac{1}{N}\sum_{i=1}^{N} f(Y_i) \tag{1.2}$$

is an unbiased estimator of $\mu$, the *Metropolis estimate*.

Received June 2, 2005; received in final form January 10, 2006.
2000 *Mathematics Subject Classification.* Primary 60J10. Secondary 82C80, 62E25, 65C05.

*Importance sampling.* Generate $X_1$ from $\sigma$ and then $X_2, \dots, X_N$ from $K(x,y)$. Then

$$\hat{\mu}_I = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(X_i)}{\sigma(X_i)} f(X_i) \tag{1.3}$$

is an unbiased estimate of $\mu$, the *importance sampling estimate*. Often one uses

$$\tilde{\mu}_I = \frac{1}{\sum_{i=1}^{N} \frac{\pi(X_i)}{\sigma(X_i)}} \sum_{i=1}^{N} \frac{\pi(X_i)}{\sigma(X_i)} f(X_i)$$

instead of $\hat{\mu}_I$. The advantage for choosing $\tilde{\mu}_I$ instead of $\hat{\mu}_I$ is that the importance sampling ratios only need to be evaluated up to an unknown constant.

The Metropolis algorithm was introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller in [16] and later generalized by Hastings [8] and Peskun [17], [18]. There exists a large body of literature on the Metropolis algorithm, the interested reader is referred to [24], [14], [19] and references therein. For this introduction, we have started both Markov chains out in their stationary distribution. For the study of the rate of convergence of the Metropolis algorithm see the survey [3]. The systematic development of importance sampling began in the 1950's with works by Khan [9], [10]. See also [22], [7], [15]. More recent references can be found in [14]. Importance sampling has seen many extensions and adaptations in recent years. For sequential importance sampling see [1], for particle filtering see [5], for adaptive importance sampling see [20]. All of these developments seem worthy of further mathematical study.

Both $\hat{\mu}_M$ and $\hat{\mu}_I$ take the output of the Markov chain $K$ and reweight to get an unbiased estimate of $\mu$. The work involved is comparable and it is natural to ask which estimate is better.

In this note we address this question through examples; a random walk on binary $d$–tuples with $K$ based on changing a random coordinate (Section 3) and the independence proposal chain (Section 4). Moreover, in Section 5 we discuss a problem of Knuth on non-self-intersecting paths and develop the theory for monotone paths in fairly complete detail. In most of our examples the Metropolis algorithm is either comparable or else dominates, sometimes by an exponential amount. The proofs are based on explicit spectral decompositions which give exact expressions for variances as determined in the following section.

## 2. Variance computation

Let $P(x,y)$ be a reversible Markov chain on the finite set $\mathcal{X}$ with stationary distribution $p(x)$. Thus, $p(x)P(x,y) = p(y)P(y,x)$. Throughout we assume all Markov chains are ergodic so $p$ is the unique stationary distribution for $P$. Let $L^2(p) = \{f : \mathcal{X} \to \mathbb{R}\}$ with $\langle f, g \rangle_p = E_p(fg) = \sum_x f(x)g(x)p(x)$.

Reversibility is equivalent to $P : L^2 \to L^2$ being self-adjoint. Here $Pf(x) = \sum_y f(y)P(x,y)$. The spectral theorem implies that $P$ has real eigenvalues $1 = \beta_0 > \beta_1 \geq \beta_2 \geq \cdots \geq \beta_{|\mathcal{X}|-1} > -1$ with an orthonormal basis of eigenfunctions $\psi_i : \mathcal{X} \to \mathbb{R}$ ($P\psi_i(x) = \beta_i\psi_i(x)$, $\langle \psi_i, \psi_j \rangle_p = \delta_{ij}$).

PROPOSITION 2.1. *Let $f \in L^2(p)$ have $\sum_x f(x)p(x) = 0$, expand $f(x) = \sum_{i\geq 1} a_i\psi_i(x)$ (with $a_i = \langle f, \psi_i \rangle_p$). Let $Z$ be chosen from $p$ and $Z_1, \ldots, Z_N$ be a realization of the $P(x,y)$ chain. Then*

$$\hat{\mu}_P = \frac{1}{N}\sum_{i=1}^{N} f(Z_i)$$

*has variance*

$$\mathrm{Var}_p(\hat{\mu}_P) = \frac{1}{N^2}\sum_{k\geq 1} |a_k|^2 W_N(k) \tag{2.1}$$

*with*

$$W_N(k) = \frac{N + 2\beta_k - N\beta_k^2 + 2\beta_k^{N+1}}{(1-\beta_k)^2}. \tag{2.2}$$

*Proof.* Because $\hat{\mu}$ has mean zero,

$$\mathrm{Var}_p(\hat{\mu}_P) = E_p(\hat{\mu}_p^2) = \frac{1}{N^2}\sum_{i,j} Ef(Y_i)f(Y_j).$$

For $i \leq j$,

$$\begin{aligned}
Ef(Y_i)f(Y_j) &= E_p\left\{\left(\sum_k a_k\psi_k(Y_i)\right)\left(\sum_l a_l\psi_l(Y_l)\right)\right\} \\
&= E_p\left\{\left(\sum_k a_k\psi_k(Y_i)\right) E_p\left(\sum_l a_l\psi_l(Y_l)|Y_i\right)\right\} \\
&= E_p\left\{\left(\sum_k a_k\psi_k(Y_i)\right)\left(\sum_l a_l\beta_l^{j-i}\psi_l(Y_i)\right)\right\} \\
&= \sum_k a_k^2\beta_k^{j-i}.
\end{aligned}$$

The last equality uses the orthonormality of $\psi_j$. The next to last equality uses $E_p(\psi_l(Y_j)|Y_i) = \beta_l^{j-i}\psi_l(Y_i)$. Summing over $i, j$, using the identity

$$\sum_{1\leq i<j\leq N} x^{j-i} = \left\{(N-1)x - Nx^2 + x^{N+1}\right\}/(1-x)^2,$$

one gets

$$E_p(\hat{\mu}_P^2) = \frac{1}{N^2}\sum_k a_k^2 \left\{ N + 2 \sum_{1\le i<j\le N} \beta_k^{j-i} \right\}$$
$$= \frac{1}{N^2}\sum_k a_k^2 W_N(k). \qquad \square$$

REMARK 2.2. (a) If $\beta_k = 1 - h_k$ for $N$ large and $h_k$ is small, $W_N(k) \sim 2N/h_k$ and $\mathrm{Var}_p(\hat{\mu}_P) \sim \frac{2}{N}\sum_{k\ge1} a_k^2/h_k$. Of course this last relation is just heuristic. We will see that it is accurate in examples.

More formally,

$$\sigma^2_\infty(\hat{\mu}_P) := \lim_{N\to+\infty} N\,\mathrm{Var}_p(\hat{\mu}_P) = \sum_{k\ge1} |a_k|^2 \frac{1+\beta_k}{1-\beta_k} \tag{2.3}$$
$$\le \frac{2}{1-\beta_1}\|f\|^2_{2,p}.$$

This last inequality is classical and it is the usual way of relating spectral gaps to asymptotic variance. It is used to compare proposal chains [17] and as a standard bound or rough estimate of the actual variance of the estimator. For small state spaces and long runs, this is reasonable. However, for large state spaces and runs a few multiples of the relaxation time, it can be badly off.

(b) Laurent Saloff-Coste suggests that the asymptotic variance can also be bounded by

$$\sigma^2_\infty(\hat{\mu}_P) \le 2|a_*|^2 \sum_{k\ge1} \frac{1}{1-\beta_k} \tag{2.4}$$

with $a_* := \max_{i\ge1} |a_i|$.

The following examples show that both bounds (2.3) and (2.4) are useful.

EXAMPLE 2.3. Let $\mathcal{X} = \mathbb{Z}_n$, the integers modulo $n$, with $n = 2m-1$ an odd number. Let

$$P(x,y) = \begin{cases} 1/2 & \text{if} \quad |x-y| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

be the transition matrix for the simple random walk. This has stationary distribution $p(x) = 1/n$, and the eigenvalues and orthonormal eigenfunctions are well known:

$$\beta_0 = 1, \psi_0 = 1,$$
$$\beta_j = \cos(2\pi j/n), \quad \psi_j^c(h) = \sqrt{2}\cos(2\pi jh/n),$$
$$\psi_j^s(h) = \sqrt{2}\sin(2\pi jh/n), \quad 1 \le j \le (n-1)/2.$$

(The non-unit eigenvalues have multiplicity two.) If $f(h) = \delta_0(h) - \frac{1}{n}$, then $a_j^c = \sqrt{2}/n$, $a_j^s = 0$. The asymptotic variance is

$$\sigma_\infty^2 = \sum_{j=1}^{\frac{n-1}{2}} \frac{2}{n^2} \frac{1+\cos(2\pi j/n)}{1-\cos(2\pi j/n)} \sim \frac{\pi^2}{3}.$$

The bound (2.3) is of order $n^2(1-1/n)/n \sim n$. Thus here the easiest bound is off while the bound (2.4) is bounded in $n$. Similar results hold for $f(h) = \delta_{[-a,a]}(h)$. Saloff-Coste has suggested that the $a_*$ bound (2.4) will be better for one and two dimensional random walk problems, but will not be an improvement more generally.

EXAMPLE 2.4. Let $\mathcal{X} = \mathbb{Z}_2^d$ be the hypercube. Let

$$P(x,y) = \begin{cases} 1/d & \text{if} \quad |x-y| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

be the transition matrix for the nearest neighbor random walk. This has stationary distribution $p(x) = 1/2^d$. The eigenvalues and othonormal eigenfunctions are well known. It is convenient to index them by $x \in \mathbb{Z}_2^d$:

$$\beta_x = 1 - \frac{2|x|}{d}, \quad \psi_x(y) = (-1)^{x \cdot y}.$$

Here $|x| = \sum_{i=1}^d x_i$ is the Hamming-weight. Note that $1-2j/d$ has multiplicity $\binom{d}{j}$, $0 \le j \le d$. If $f(x) = \delta_0(x) - 1/2^d$, the Fourier coefficients at $x \neq 0$ are $a_x = 1/2^d$. The asymptotic variance is

$$\sigma_\infty^2 = \frac{1}{2^{2d}} \sum_{j=1}^d \binom{d}{j} \frac{2+2j/d}{2j/d} = \frac{1}{2^{2d}} \sum_{j=1}^d \left[ \frac{d}{j} \binom{d}{j} + 2^d \right] \le \frac{4}{2^d}.$$

On the other hand, $\|f\|_2^2 = \frac{1}{2^d}(1 - \frac{1}{2^d})$. The crude upper bound from (2.3) is $2d\frac{1}{2^d}(1-\frac{1}{2^d})$, which is off by a factor of $d$. The $a_*$ bound from (2.4) for $\sigma_\infty^2$ is

$$\frac{2}{2^{2d}} \sum_{j=1}^{d-1} \binom{d}{j} \frac{d}{2j} \le \frac{C}{2^d}.$$

This has the right order.

If $f(x) = \sum_{j=1}^d (-1)^{x_j} = d - 2|x|$, $a_x = 0$ if $|x| > 1$ and $a_x = 1$ if $|x| = 1$. Thus

$$\sigma_\infty^2 = d \frac{2 - 2/d}{2/d} = d(d-1).$$

Here, $\|f\|_2^2 = d$. The crude bound from Remark 2.2 is $2d/(2/d) = d^2$. The $a_*$ bound is

$$\sum_{j=1}^{d} \binom{d}{j} \frac{d}{2j} \sim c2^d$$

for some $c$. This is wildly off.

There is another approach to bounding the asymptotic variance $\sigma_\infty^2(\hat{\mu}_P)$ defined by (2.3). This uses the basic Poincaré inequality ([4], [21])

$$\|f\|_{2,p}^2 \leq \frac{1}{1-\beta_1}\mathcal{E}(f,f)$$

with $\mathcal{E}(f,f)$ the Dirichlet form

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))^2 p(x)P(x,y).$$

Using this in (2.3) gives

$$\sigma_\infty^2(\hat{\mu}_P) \leq \frac{2}{(1-\beta_1)^2}\mathcal{E}(f,f).$$

The point is that sometimes $\mathcal{E}(f,f)$ can be usefully bounded since it only involves knowing how $f$ varies under a local change. In Example 2.4 of this section with $f(x) = d-2|x|$, $(f(x)-f(y))^2 = 4$ when $P(x,y) > 0$, so $\mathcal{E}(f,f) = 2$. Thus our bound gives $\sigma_\infty^2 \leq d^2$. This is essential sharp. In Example 2.3, with $f(x) = \delta_0 - 1/n$, $(f(x)-f(y))^2$ equals 1 when $(x,y) = (0,\pm 1), (\pm 1, 0)$, and 0 otherwise. It follows that $\mathcal{E}(f,f) = 2/n$. Thus our bound gives $\sigma_\infty^2 \leq 4/(n(1-\cos(2\pi/n))^2)$. Here the bound is off. Of course, these are simple examples. The power of these Poincaré arguments will only show in more complex problems where little is known about the stationary distribution or the eigenfunctions.

## 3. The hypercube

Let $\mathcal{X} = \mathbb{Z}_2^d$ be the set of binary $d$–tuples. Let

$$\pi(x) = \theta^{|x|}(1-\theta)^{d-|x|}$$

with $1/2 \leq \theta \leq 1$ and $|x|$ the number of ones in the $d$–tuple $x$. Let the base chain $K(x,y)$ be given by "from $x$, pick a coordinate at random and change it to one or zero with probability $p$ or $1-p$" $(1/2 < p \leq 1)$. The $K$–chain has stationary distribution

$$\sigma(x) = p^{|x|}(1-p)^{d-|x|}.$$

This example models a high-dimensional problem where the desired distribution $\pi(x)$ is concentrated in a small part of the space. We have available a sampling procedure—run the chain $K(x,y)$—where the stationary distribution is roughly right (if $p$ is close to $\theta$) but not spot on.

Let us begin by diagonalizing the Metropolis chain. This may be presented as

$$M = \frac{1}{d}\sum_{i=1}^{d} M_i \tag{3.1}$$

with $M_i$ the Metropolis construction operating on the $i$–th coordinate. The transition matrix restricted to the $i^{th}$ coordinate is

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} \bar{p} & p \\ p\,\bar{\theta}/\theta & 1 - p\,\bar{\theta}/\theta \end{pmatrix} \end{array} \qquad \text{with} \quad \bar{p} = 1-p,\ \bar{\theta} = 1-\theta.$$

This matrix has stationary distribution $(\bar{\theta}, \theta)$ on $\{0,1\}$. The eigenvalues are $\beta_0 = 1$, $\beta_1 = 1 - p/\theta$, with normalized eigenvectors

$$\psi_0(0) = \psi_0(1) = 1, \qquad \psi_1(0) = \sqrt{\theta/\bar{\theta}}, \quad \psi_1(1) = -\sqrt{\bar{\theta}/\theta}. \tag{3.2}$$

PROPOSITION 3.1. *The Metropolis chain* (3.1) *on* $\mathbb{Z}_2^d$ *has* $2^d$ *eigenvalues and eigenvectors which will be indexed by* $\zeta \in \mathbb{Z}_2^d$. *These are*

$$\beta_\zeta = 1 - \frac{|\zeta| p}{d\theta}, \tag{3.3}$$

$$\psi_\zeta(x) = \prod_{i=1}^{d} \psi_{\zeta_i}(x_i) = \prod_{i=1}^{d} \left(\sqrt{\frac{\theta}{\bar{\theta}}}\right)^{\zeta_i(1-x_i)} \left(-\sqrt{\frac{\bar{\theta}}{\theta}}\right)^{\zeta_i x_i}, \tag{3.4}$$

*with* $\psi_i$ *defined in* (3.2). *The eigenvectors are orthonormal in* $L^2(\pi)$.

*Proof.* This is a straightforward verification from (3.2) and the basic structure of the product chains. For more details, see [2, Sec. 5]. □

Using these tools we may compute the variance of the Metropolis algorithm for a variety of functions $f$. We take $f$ to be the number of ones normalized to have mean zero.

PROPOSITION 3.2. *On* $\mathbb{Z}_2^d$, *let*

$$f(x) = \sum_{i=1}^{d} (x_i - \theta). \tag{3.5}$$

*Under the Metropolis chain* (3.1), *with* $\hat{\mu}_M$ *defined by* (1.2), *we have* $\mu = 0$ *and*

$$\mathrm{Var}_\pi(\hat{\mu}_M) = \frac{2d^2\bar{\theta}\theta^2}{Np} - \frac{d\theta\bar{\theta}}{N} + \frac{2d^3\theta^3\bar{\theta}}{N^2p^2}\left(1 - \frac{p}{d\theta}\right)^{N+1} + \frac{2d^3\bar{\theta}\theta^3}{N^2p^2}\left(1 - \frac{p}{d\theta}\right).$$

*Here*

$$\sigma^2_\infty(\hat{\mu}_M) \sim \frac{2d^2\bar{\theta}\theta^2}{p} \qquad (d \to +\infty).$$

*Proof.* To use Proposition 2.1, we must compute the expansion of $f$ in (3.5) with respect to the eigen basis $\psi_\zeta$ of (3.4). For $\zeta \neq 0$, by the orthogonality of $\psi_\zeta$ with $\psi_0 \equiv 1$, $\sum_x \psi_\zeta(x) f(x)\pi(x) = \sum_x (\sum_i x_i)\psi_\zeta(x)\pi(x)$. For fixed $i$, write $x^i$ for $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$. If $\zeta_i = 0$,

$$\sum_x x_i \psi_\zeta(x)\pi(x) = 0.$$

If $\zeta_i = 1$,

$$\begin{aligned}\sum_x x_i\psi_\zeta(x)\pi(x) &= \sum_{x^i} \pi(x^i)\psi_{\zeta^i}(x^i)(0\bar{\theta}\psi_{\zeta_i}(0) - \theta\psi_{z_i}(1)) \\ &= \begin{cases} -\sqrt{\theta\bar{\theta}} & \text{if } \zeta^i = 0, \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

Hence, $a_\zeta = -\sqrt{\theta\bar{\theta}}$ if $|\zeta| = 1$ and $a_\zeta = 0$ otherwise. Now, Proposition 2.1 and (3.3) give

$$\mathrm{Var}_\pi(\hat{\mu}_M) = \frac{d\bar{\theta}\theta}{N^2}\left(\frac{2Nh - Nh^2 + 2(1-h) + 2(1-h)^{N+1}}{h^2}\right)$$

with $h = (\theta d)^{-1}p$. □

Consider next the importance sampling chain with the same $K$ and $\sigma$ considered above. Represent $K$ as

$$K = \frac{1}{d}\sum_{i=1}^d K_i \tag{3.6}$$

with $K_i$ having matrix (restricted to the $i^{th}$ coordinate)

$$\begin{array}{c} \\ 0 \\ 1 \end{array}\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} \bar{p} & p \\ \bar{p} & p \end{pmatrix} \end{array} \qquad \text{with} \quad \bar{p} = 1 - p.$$

This matrix has stationary distribution $(\bar{p}, p)$ on $\{0, 1\}$. The eigenvalues are $\beta^*_0 = 1$, $\beta^*_1 = 0$ with normalized eigenvectors $\psi^*_0(0) = \psi^*_1(1) = 1$, $\psi^*_1(0) = \sqrt{p/\bar{p}}$, $\psi^*_1(1) = -\sqrt{\bar{p}/p}$. Arguing as for Proposition 3.2 we have the following:

PROPOSITION 3.3. *The Markov chain* (3.6) *on* $\mathbb{Z}_2^d$ *has* $2^d$ *eigenvalues and eigenvectors indexed by* $\zeta$ *in* $\mathbb{Z}_2^d$. *These are*

$$\beta^*_\zeta = 1 - \frac{|\zeta|}{d}, \tag{3.7}$$

$$\psi_z^*(x) = \prod_{i=1}^d \psi_{\zeta_i}^*(x_i) = \prod_{i=1}^d \left(\sqrt{\frac{p}{\bar{p}}}\right)^{\zeta_i(1-x_i)} \left(-\sqrt{\frac{\bar{p}}{p}}\right)^{\zeta_i x_i}. \tag{3.8}$$

*The eigenvectors are orthonormal in $L^2(\sigma)$, where $\sigma(x) = p^{|x|}(1-p)^{d-|x|}$.*

The importance sampling estimate is

$$\hat{\mu}_I = \frac{1}{N}\sum_{i=1}^N \frac{\pi(X_i)}{\sigma(X_i)} f(X_i) \tag{3.9}$$

with

$$\frac{\pi(x)}{\sigma(x)} = ab^{|x|}$$

for $a = ((1-\theta)/(1-p))^d$, $b = (\theta(1-p)/(p(1-\theta)))^d$.

To use the machinery above, we need to compute the spectral coefficients.

PROPOSITION 3.4. *Let $\psi_\zeta^*$ be defined as in* (3.8) *and*

$$g(x) = \left(\sum_{i=1}^d (x_i - \theta)\right) \frac{\pi(x)}{\sigma(x)}.$$

*Then $\langle g, \psi_\zeta^* \rangle_\pi = -\alpha|\zeta|\beta^{|\zeta|}$ with $\alpha := (\theta\bar{\theta}\sqrt{p/\bar{p}} + \bar{\theta}\theta\sqrt{\bar{p}/p})/(\bar{\theta}\sqrt{p/\bar{p}} - \theta\sqrt{\bar{p}/p})$ and $\beta := \bar{\theta}\sqrt{p/\bar{p}} - \theta\sqrt{\bar{p}/p}$.*

*Proof.* Write

$$\langle g, \psi_\zeta^* \rangle_\pi = \sum_{i=1}^d E_\sigma \frac{\pi(x)}{\sigma(x)} \psi_\zeta^*(x)(x_i - \theta).$$

Under $\sigma$, the coordinates are independent, taking values one or zero with probability $p$ and $1-p$. The integrand is a product and we may compute it componentwise. Consider the $i^{th}$ term in the sum. For $j \neq i$, the expectation of the $j^{th}$ term in the product is 1 if $\zeta_j = 0$, and $\bar{\theta}\sqrt{p/\bar{p}} - \theta\sqrt{\bar{p}/p}$ if $\zeta_j = 1$. For $j = i$, the expectation of the $i^{th}$ term in the product is 0 if $\zeta_i = 0$ and $-\theta\bar{\theta}\sqrt{p/\bar{p}} - \theta\bar{\theta}\sqrt{\bar{p}/p}$ if $\zeta_i = 1$. Computing the product and summing in $i$ gives the stated result. □

Combining these results gives a formula for the variance.

PROPOSITION 3.5. *For the Markov chain $K$ of* (3.6) *and $f(x) = \sum_{i=1}^d (x_i - \theta)$, the importance sampling estimate $\hat{\mu}_I$ of* (3.9) *has mean zero and variance*

$$\mathrm{Var}_\sigma(\hat{\mu}_I) = \frac{\alpha^2}{N^2} \sum_{i=1}^d \binom{d}{i} i^2 \beta^{2i} W_N^*(i),$$

*with $\alpha$ and $\beta$ from Proposition 3.4 and*

$$W_n^*(i) = \frac{2Nd}{i} - N + \frac{2d^2}{i^2}(1 - i/d) + \frac{2d^2}{i^2}(1 - i/d)^{N+1}.$$

REMARK 3.6. (a) The lead term in $\mathrm{Var}_\sigma(\hat{\mu}_I)$ is

$$\frac{\sigma^2}{N^2}\sum_{i=1}^{d}\binom{d}{i}i^2\beta^{2i}\frac{2Nd}{i} = \frac{2\alpha^2 d}{N}\sum_{i=1}^{d}\binom{d}{i}i\beta^{2i} = 2\beta^2\frac{\alpha^2 d^2}{N}(1+\beta^2)^{d-1}.$$

The last equality used $d(1+x)^{d-1}x = \sum_{i=0}^{d}\binom{d}{i}ix^i$. For our running example, $\theta = 7/8$, $p = 3/4$, $\beta = -0.2887$, $1 + \beta^2 = 1.0833$; for large $d$ and $N$, $\mathrm{Var}_\sigma(\hat{\mu}_I)$ is exponentially worse (in $d$) than the Metropolis variance.

(b) The next term is

$$-\frac{\alpha^2}{N^2}\sum_{i=1}^{d}\binom{d}{i}i^2\beta^{2i}N = -\frac{\alpha^2}{N}(1+d\beta^2)s\beta^2(1+\beta^2)^{d-2};$$

the sum of these two lead terms is

$$\frac{\alpha^2 d\beta^2}{N}(1+\beta^2)^{d-2}\{2(1+\beta^2)d - (1+d\beta^2)\}.$$

(c) For the next term we need

$$\begin{aligned}\sum_{i=1}^{d}\binom{d}{i}i^2\beta^{2i}\frac{d^2}{i^2}\left(1-\frac{i}{d}\right) &= d^2((1+\beta^2)^d - 1) - d^2(1+\beta^2)^{d-1}\beta^2 \\ &= d^2((1+\beta^2)^{d-1} - 1).\end{aligned}$$

From this, the third term is $2d^2\alpha^2((1+\beta^2)^{d-1} - 1)/N^2$.

(d) For the last term we need

$$\begin{aligned}2d^2\sum_{i=1}^{d}\binom{d}{i}\beta^{2i}\left(1-\frac{i}{d}\right)^{N+1} &\le 2d^2\sum_{i=1}^{d}\binom{d}{i}\beta^{2i}e^{-i(N+1)/d} \\ &\le 2d^2\left(\left(1+\beta^2 e^{-(N+1)/d}\right)^d - 1\right).\end{aligned}$$

From this the last term is positive and bounded above by

$$2d^2\alpha^2\left(\left(1+\beta^2 e^{-(N+1)/d}\right)^d - 1\right)/N^2.$$

(e) The bottomline is

$$\mathrm{Var}_\sigma(\hat{\mu}_I) \sim \frac{2\alpha^2 d^2\beta^2}{N}(1+\beta^2)^{d-1}.$$

This is exponentially worse (in $d$) than the estimate

$$\mathrm{Var}_\pi(\hat{\mu}_M) \sim \frac{2d^2\theta^2\bar{\theta}}{Np}$$

from Proposition 3.2.

(f) From Remark 3.6 (a), the variance of the importance sampling estimator blows up as $d^2(1+\beta^2)^d$, with $\beta^2 = \bar{\theta}^2 p/\bar{p} + \bar{p}/p\theta^2 - 2\theta\bar{\theta}$. It is natural to ask how close $p$ must be to $\theta$ so that this doesn't blow up. If $p = \theta + \epsilon$, a straightforward calculation gives

$$\beta^2 = \frac{\epsilon^2}{\theta\bar{\theta}} + O(\epsilon^2).$$

It follows that $\epsilon$ of order $1/\sqrt{d}$ is required to keep the variance from exponential explosion.

For another example we take $f(x) = \delta_d(|x|) - \theta^d$. Again we compute the spectral coefficients.

PROPOSITION 3.7. *Let $\psi_\zeta$ and $\psi^*_\zeta$ be defined as in* (3.3) *and* (3.8). *Let $f(x) = \delta_d(|x|) - \theta^d$ and $g(x) = f(x)\pi(x)/\sigma(x)$. Then $\mu = E_\pi(f) = 0$ and $\mathrm{Var}_\pi(f) = \theta^d(1-\theta^d)$. Moreover,*

$$a_\zeta := \langle f, \psi_\zeta\rangle_\pi = (-1)^d \left(\sqrt{\frac{\bar{\theta}}{\theta}}\right)^{|\zeta|} \theta^d,$$

$$a^*_\zeta := \langle f, \psi^*_\zeta\rangle_\pi = (-1)^d \left(\sqrt{\frac{\bar{\theta}}{\theta}}\right)^{|\zeta|} \theta^d \left\{1 - \left(\frac{\theta - p}{1-p}\right)^{|\zeta|}\right\}$$

*holds true for every $\zeta$ with $|\zeta| \neq 0$, and $a_0 = a^*_0 = 0$.*

*Proof.* This is a straightforward verification. Indeed, by orthogonality, $\langle f, \psi_\zeta\rangle_\pi = \langle \delta_d(|x|), \psi_\zeta\rangle_\pi$, and then $\langle f, \psi_\zeta\rangle_\pi = \theta^d \psi_\zeta(\mathbf{1})$, where $\mathbf{1} = (1, 1, \ldots, 1)$. Moreover, arguing as in the proof of Proposition 3.4, we get

$$\begin{aligned} a^*_\zeta &= \langle \delta_d(|x|), \psi^*_\zeta\rangle_\pi - \theta^d \langle 1, \psi^*_z\rangle_\pi \\ &= (-1)^{|\zeta|} \left(\sqrt{\frac{\bar{p}}{p}}\right)^{|\zeta|} \theta^d - \theta^d \left(\bar{\theta}\sqrt{\frac{p}{\bar{p}}} - \theta\sqrt{\frac{\bar{p}}{p}}\right)^{|\zeta|} \\ &= (-1)^{|\zeta|} \left(\sqrt{\frac{\bar{p}}{p}}\right)^{|\zeta|} \theta^d \left(1 - \left(\frac{\theta - p}{1-p}\right)^{|\zeta|}\right). \end{aligned}$$

□

Combining the previous results we get:

PROPOSITION 3.8. *On $\mathbb{Z}_2^d$, let $f(x) = \delta_d(|x|) - \theta^d$. Under the Metropolis chain* (3.1), *with $\hat{\mu}_M$ defined by* (1.2), *we have $\mu = 0$ and*

$$\mathrm{Var}_\pi(\hat{\mu}_M) = \frac{\theta^{2d}}{N^2} \sum_{i=1}^{d} \binom{d}{i} \left(\frac{\bar{\theta}}{\theta}\right)^i W^*_N(ip/\theta),$$

*with*

$$W_n^*(i) = \frac{2Nd}{i} - N + \frac{2d^2}{i^2}(1 - i/d) + \frac{2d^2}{i^2}(1 - i/d)^{N+1}.$$

*Moreover, for the Markov chain $K$ of* (3.6) *the importance sampling estimate $\hat{\mu}_I$ of* (3.9) *has mean zero and variance*

$$\mathrm{Var}_\sigma(\hat{\mu}_I) = \frac{\theta^{2d}}{N^2} \sum_{i=1}^{d} \binom{d}{i} \left(\frac{\bar{p}}{p}\right)^i \left(1 - \left(\frac{\theta - p}{1-p}\right)^i\right)^2 W_N^*(i).$$

REMARK 3.9. (a) The lead term in $\mathrm{Var}_\pi(\hat{\mu}_M)$ is

$$\frac{\theta^{2d}}{N} \sum_{i=1}^{d} \binom{d}{i} \left(\frac{\bar{\theta}}{\theta}\right)^i \left(\frac{2d\theta}{p}\frac{1}{i} - 1\right) = \frac{2\theta^{2d}\theta}{Np} d A_d(\bar{\theta}/\theta) - \theta^d(1-\theta^d)/N,$$

with

$$A_d(x) := \sum_{i=1}^{d} \binom{d}{i} x^i/i.$$

Since

$$A_d(x) = (x(d+1))^{-1} \sum_{i=2}^{d+1} \binom{d+1}{i} x^i i/(i-1),$$

for every positive $x$, we can write

$$\frac{[(1+x)^{d+1} - 1 - x(d+1)]}{x(d+1)} \le A_d(x) \le 2\frac{[(1+x)^{d+1} - 1 - x(d+1)]}{x(d+1)}. \tag{3.10}$$

Hence, the lead term in $\mathrm{Var}_\pi(\hat{\mu}_M)$ can be written as

$$N^{-1}(2\theta^{d+1} d B_d(\bar{\theta}/\theta)/(p\bar{\theta}(d+1)) - \mathrm{Var}_\pi(f))$$

with $B_d(\bar{\theta}/\theta)$ bounded in $d$. More exactly, (3.10) gives

$$\left[1 - \theta^{d+1} - \bar{\theta}\theta^d(d+1)\right] \le B_d(\bar{\theta}/\theta) \le 2\left[1 - \theta^{d+1} - \bar{\theta}\theta^d(d+1)\right].$$

This suggests that $N\,\mathrm{Var}_\pi(\hat{\mu}_M) = O(\mathrm{Var}_\pi(f))$ for $d \to +\infty$. This will be formalized in the next proposition.

(b) The first term in $\mathrm{Var}_\sigma(\hat{\mu}_I)$ can be written as

$$\begin{aligned}
&\frac{2\theta^{2d}d}{N} \sum_{i=1}^{d} \binom{d}{i} \left(\frac{\bar{p}}{p}\right)^i \frac{1}{i} \left[1 + \left(\frac{\theta - p}{1-p}\right)^{2i} - 2\left(\frac{\theta - p}{1-p}\right)^i\right] \\
&\quad = \frac{2\theta^{2d}d}{N} \left[A_d(\bar{p}/p) + A_d\left(\frac{(\theta-p)^2}{p(1-p)}\right) - 2A_d\left(\frac{\theta - p}{p}\right)\right] \\
&\quad = \frac{2\theta^d d}{N(d+1)} \left(\frac{\theta}{p}\right)^d B_d^*(p,\theta)
\end{aligned}$$

with

$$B_d^*(p,\theta) = (d+1)p^d \left[ A_d(\bar{p}/p) + A_d\left(\frac{(\theta-p)^2}{p(1-p)}\right) - 2A_d\left(\frac{\theta-p}{p}\right)\right].$$

Again using (3.10), it is easy to see that $d \mapsto B_d^*(p,\theta)$ is a bounded function. The second term in $\mathrm{Var}_\sigma(\hat{\mu}_I)$ is

$$\begin{aligned} &-\frac{\theta^d}{N}\left(\frac{\theta}{p}\right)^d \left[1 - 2\theta^d + \left(\frac{p\bar{p} + (\theta-p)^2}{\bar{p}}\right)^d\right] \\ &\quad = -\frac{\theta^d}{N}\left(\frac{\theta}{p}\right)^d \left[1 - 2\theta^d + \left(\frac{p + \theta^2 - 2\theta p}{\bar{p}}\right)^d\right] \\ &\quad = -\frac{\theta^d}{N}\left(\frac{\theta}{p}\right)^d C_d^*(p,\theta) \end{aligned}$$

with

$$C_d^*(p,\theta) = \left[1 - 2\theta^d + \left(\frac{p + \theta^2 - 2\theta p}{\bar{p}}\right)^d\right].$$

If $1/2 < p < \theta < 1$, then $(p + \theta^2 - 2\theta p)/\bar{p} < 1$. Hence, $C_d^*(p,\theta) = 1 + o(1)$ for $d \to +\infty$, while some simple computations show that $\theta(p + \theta^2 - 2\theta p)/(p(1-p)) < 1$ if $1/2 < \theta < p < 1$.

Combining the previous remarks we get:

PROPOSITION 3.10. *For $f$ as in Proposition 3.8,*

$$\sigma_\infty^2(\hat{\mu}_M)/\mathrm{Var}_\pi(f) = \frac{2\theta}{p\bar{\theta}(1-\theta^d)}\frac{d}{d+1}B_d(\bar{\theta}/\theta) - 1 \sim K_1(\theta,p)$$

*with $d \mapsto B_d(\bar{\theta}/\theta)$ bounded, $K_1$ being a suitable constant. Moreover,*

$$\begin{aligned} \sigma_\infty^2(\hat{\mu}_I)/\mathrm{Var}_\pi(f) &= \frac{1}{1-\theta^d}\left(\frac{\theta}{p}\right)^d \left[\frac{2d}{d+1}B_d^*(p,\theta) - C_d^*(p,\theta)\right] \\ &\sim K_2(\theta,p)\left(\frac{\theta}{p}\right)^d \end{aligned}$$

*with $d \mapsto [\frac{2d}{d+1}B_d^*(p,\theta) - C^*(p,\theta)]$ bounded and strictly positive, for every $p$ and $\theta$ such that $1/2 < p < \theta < 1$, while*

$$\sigma_\infty^2(\hat{\mu}_I)/\mathrm{Var}_\pi(f) \sim K_3(\theta,p)\left[\left(\frac{\theta p + \theta^3 - 2\theta^2 p}{p(1-p)}\right)^d + \left(\frac{\theta}{p}\right)^d\right]$$

*with $\frac{\theta p + \theta^3 - 2\theta^2 p}{p(1-p)} < 1$ if $1/2 < \theta < p < 1$, $K_2, K_3$ being suitable constants.*

REMARK 3.11. The last proposition shows that, for $1/2 < p < \theta < 1$, the normalized asymptotic variance of $\hat{\mu}_I$ is exponentially worse (in $d$) than the normalized asymptotic variance of $\hat{\mu}_M$, while it is exponentially better if $1/2 < \theta < p < 1$. On the one hand, this fact agrees with the heuristic that the importance sampling estimator of $\pi(A)$ performs better than the Monte Carlo estimator, whenever the importance distribution $\sigma$ puts more weight on $A$ than $\pi$. On the other hand, the last example shows that even a small "loss of weight" in $A$ can cause an exponentially worse behavior.

## 4. Independence sampling

In this section the proposal chain is a sequence of independent and identically distributed variables with common probability density $\sigma$. Because of this, the structure of the state space does not matter. Throughout we take $\mathcal{X} = \{0, 1, \dots, m-1\}$ with $\sigma(i) > 0$ fixed and $\pi(i)$ the desired distribution. Without loss, suppose the states are numbered so that the importance weights $\pi(i)/\sigma(i)$ are decreasing, i.e.,

$$\frac{\pi(0)}{\sigma(0)} \geq \frac{\pi(1)}{\sigma(1)} \geq \cdots \geq \frac{\pi(m-1)}{\sigma(m-1)}. \tag{4.1}$$

This section makes use of an explicit diagonalization of the Metropolis chain due to Jun Liu [13].

*Metropolis.* For the proposal chain of independent picks from $\sigma$, the Metropolis chain starts with a pick from $\pi$. From state $i$ it proceeds by choosing $j$ from $\sigma$; if $j \leq i$ the chain moves to $j$. If $j > i$ the chain stays at $j$ with probability $(\pi(j)\sigma(i)/\pi(i)\sigma(j))$ and remains at $i$ with probability $(1-\pi(j)\sigma(i))/\pi(i)\sigma(j))$. Liu [13] proves that this chain has eigenvalue 1 and, for $1 \leq k \leq m-1$,

$$\beta_k = \sum_{i \geq k} \left( \sigma(i) - \pi(i) \frac{\sigma(k)}{\pi(k)} \right), \tag{4.2}$$

$$\psi_k = (\underbrace{0, \dots, 0}_{k-1}, S_\pi(k+1), -\pi(k), \dots, -\pi(k)), \quad S_\pi(k+1) := \sum_{j=k+1}^{m-1} \pi(j). \tag{4.3}$$

These eigenvectors are orthogonal in $L^2(\pi)$. From these facts and Proposition 2.1 we get the following:

PROPOSITION 4.1. *For the Metropolis algorithm based on independent proposals with distribution $\sigma(i)$ and stationary distribution $\pi(i)$, let $f : \mathcal{X} \to \mathbb{R}$ have representation $f(i) = \sum_{k=1}^{m-1} a_k \psi_k(i)$, with $\mu = \sum_{i=0}^{m-1} f(i)\pi(i) = 0$. Then the Metropolis estimator $\hat{\mu}_M$ of (1.2) satisfies*

$$\mathrm{Var}_\pi(\hat{\mu}_M) = \frac{1}{N^2} \sum_{k=1}^{m-1} b_k^2 W_N(k) \tag{4.4}$$

*with $W_N(k)$ given by* (2.2) *using the eigenvalues of* (4.2) *and*

$$b_k := \zeta_k \left( f(k) S_\pi(k+1)\pi(k) - \pi(k) \sum_{j \geq k} f(j)\pi(j) \right),$$

$$\zeta_k := (\pi(k) S_\pi(k) S_\pi(k+1))^{-1/2}.$$

The following lemma supplements Liu's results:

LEMMA 4.2. *With the notation as above, the eigenvalues $\beta_k$ in* (4.2) *satisfy*

$$1 - \frac{\sigma(0)}{\pi(0)} \geq \beta_1 \geq \beta_2 \geq \cdots \geq \beta_{m-1} = 0.$$

*Proof.* We have

$$\begin{aligned}
\beta_k &= \sigma(k) + \cdots + \sigma(m-1) - \frac{\sigma(k)}{\pi(k)}(\pi(k) + \cdots + \pi(m-1)) \\
&\geq \sigma(k+1) + \cdots + \sigma(m-1) - \frac{\sigma(k+1)}{\pi(k+1)}(\pi(k+1) + \cdots + \pi(m-1)) \\
&= \beta_{k+1}.
\end{aligned}$$

For $\beta_1$, note that

$$\beta_1 = 1 - \sigma(0) - \frac{\sigma(0)}{\pi(0)}(1 - \pi(0)) \leq 1 - \frac{\sigma(0)}{\pi(0)}.$$

For $\beta_{m-1}$ note that

$$\beta_{m-1} = \sigma(m-1) - \frac{\sigma(m-1)}{\pi(m-1)}\pi(m-1) = 0. \qquad \square$$

The variance of the importance sampling estimator is (for $\mu = 0$)

$$\mathrm{Var}_\sigma(\hat{\mu}_I) = \frac{1}{N^2} \sum_l \left( \frac{\pi(l)}{\sigma(l)} f(l) \right)^2 \sigma(l). \tag{4.5}$$

We record a different expression for this, similar to J. Liu's above.

LEMMA 4.3. *Let $P$ be the Markov chain on $\{0, 1, \ldots, m-1\}$ with all rows equal to $\sigma$. Then, $P$ has one eigenvalue $\beta_0 = 1$ with $\psi_0(i) \equiv 1$ and $m-1$ eigenvalues 0. An orthogonal basis for the zero eigenspace in $L^2(\sigma)$ is*

$$\psi_k = (0, \ldots, 0, S_\sigma(k+1), -\sigma(k), \ldots, -\sigma(k)), \qquad 1 \leq k \leq m-1. \tag{4.6}$$

REMARK 4.4. If $\sigma = \pi$, this basis agrees with the Metropolis basis of (4.3). This must be because then $P$ commutes with the Metropolis chain.

PROPOSITION 4.5. *Let $g(i)$ satisfy $\sum_{i=0}^{m-1} g(i)\sigma(i) = 0$. Then*

$$\begin{aligned}\mathrm{Var}_\sigma(g) &= \sum_{i=0}^{m-1} g(i)^2\sigma(i) \\ &= \sum_{i=1}^{m-1} \left( g(i)S_\sigma(i+1) - \sigma(i)\sum_{j\geq i} g(j)\sigma(j) \right)^2 /(\sigma(i)S_\sigma(i)S_\sigma(i+1)).\end{aligned}$$

EXAMPLE 4.6. Let us compare the Metropolis and the importance estimator in the following case. For $\mathcal{X} = \{0, 1, \dots, m-1\}$, let $\pi(i) = a^i c(a)$ for fixed $a$ with $0 < a < 1$ with $c(a) = (1-a)/(1-a^m)$. For the proposal chain take $\sigma(i) = 1/m$, $0 \leq i \leq m-1$. Take $f(i) = (i-\mu)$ with $\mu = c(a)a(1 - a^{m-1}(a + (1-a)m))/(1-a)^2$. Thus $\mu = E_\pi(f) = 0$. For $a$ fixed and $m$ large $\mu \sim a/(1-a)$.

*Metropolis.* We must compute the expansion of $f$ in the basis $\psi_k$, $1 \leq k \leq m-1$. We have

$$a_k = \sum_j f(j)\psi_k(j)\pi(j) = (k-\mu)S_\pi(k+1)\pi(k) - \pi(k)\sum_{j=k+1}^{m-1}(j-\mu)\pi(j).$$

Here

$$S_\pi(k+1) = c(a)(1 - a^{m-k-1})a^{k+1}/(1-a),$$

and

$$\begin{aligned}&\sum_{j=k+1}^{m-1}(j-\mu)\pi(j) \\ &\qquad = c(a)(a^k((k+1)(1-a) - a) - a^{m-1}(m(1-a) - a))/(1-a)^2.\end{aligned}$$

It follows that

$$a_k = a^{2k} b(a, m, k)$$

with $b(a)$ bounded uniformly in $m$, $k$, $(1-a)^2$. It thus follow that $b_k^2$ in (4.4) equals $a^{2k}c(m,k)$ with $c(m,k)$ uniformly bounded. The eigenvalues $\beta_k$ of (4.2) become

$$\beta_k = 1 - \frac{k}{m} - \frac{1}{m}\left(\frac{1-a^{m-k}}{1-a}\right).$$

Plugging into (4.4) yields the following proposition.

PROPOSITION 4.7. *Fix $a$ with $0 < a < 1$, for independent proposal Metropolis sampling with uniform proposals on $\{0, 1, \dots, m-1\}$ for $\pi(i) = a^i c(a)$ the Metropolis algorithm has $\mathrm{Var}_\pi(\hat{\mu}_M(f)) = (m/N)A(a)$ with $A$ continuous and bounded uniformly in $m$ and $N$.*

*Importance sampling.* From (4.5),

$$\mathrm{Var}_\sigma(\hat{\mu}_I) = \frac{1}{N}\sum_{i=0}^{m-1}\left(\frac{\pi(i)}{\sigma(i)}(i-\mu)\right)^2\sigma(i) = \frac{m}{N}\sum_{i=0}^{m-1}\left(\pi(i)(i-\mu)\right)^2.$$

This expression is of the form

$$\frac{m}{N}B(a)$$

with $B(a)$ continuous in $a$ and bounded uniformly in $m$ for fixed $a \in (0,1)$. It follows that importance sampling and the Metropolis algorithm are roughly comparable for this example.

REMARK 4.8. The example above helps us to calibrate two obvious bounds. From (2.3), for independent Metropolis sampling applied to $f$ of mean zero,

$$\sigma^2_\infty(\hat{\mu}_M) \le 2\frac{\|f\|^2_{2,\pi}}{1-\beta_1} \le 2\frac{\pi(0)}{\sigma(0)}\|f\|^2_{2,\pi},$$

$$\begin{aligned}\mathrm{Var}_\sigma(\hat{\mu}_I) &= \frac{1}{N}\sum_i\left(\frac{\pi(i)}{\sigma(i)}f(i)\right)^2\sigma(i) = \frac{1}{N}\sum_i\frac{\pi(i)^2}{\sigma(i)}f(i)^2\\ &\le \frac{1}{N}\frac{\pi(0)}{\sigma(0)}\|f\|^2_{2,\pi}.\end{aligned}$$

## 5. Non-self-intersecting paths

Our interest in this area started with Donald Knuth's [12] study of non-self-intersecting paths in a grid. Knuth considered a $10\times 10$ grid. He wanted to estimate the number of non-self-intersecting lattice paths $\gamma$ that start at $(0,0)$ and end at $(10,10)$. He used the following sequential importance sampling (SIS) estimate: build a path $\gamma$ starting at $(0,0)$ sequentially, each time choosing one of the available nearest neighbors with equal probability. As the path grows, the past is recorded and only non-self-intersecting choices are considered. Thus, the first step may go up or to the right with probability $1/2$. Suppose it goes up. The next step can go up or to the right. Suppose it goes to the right. The third step has three possibilities (up, right, down) chosen with probability $1/3$, and so on. If the algorithm gets stuck, it simply starts again at $(0,0)$. Let $\sigma(\gamma)$ be the probability of a successful path. Thus $\sigma(\gamma) = \frac{1}{2}\frac{1}{2}\frac{1}{3}\cdots$ in the example. This is easily computed as the path is created. Let $X_i = 0$ if the $i^{th}$ trial fails and $X_i = 1/\sigma(\gamma)$ if the $i^{th}$ trial produces a legal path $\gamma$. Observe that

$$E(X_i) = \sum_\gamma \frac{1}{\sigma(\gamma)}\sigma(\gamma) = \text{number of paths.}$$

Thus if $X_1, X_2, \dots X_N$ is the result of $N$ trials,

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

is an unbiased estimate of the number of paths. Knuth used this sequential importance sampling algorithm to give the following estimates:

- The number of paths is $(1.6 \pm 0.3)10^{24}$.
- The average path length is $92 \pm 5$.
- The proportion of paths through $(5, 5)$ is $81 \pm 10$ percent.

In a personal communication he noted that usually $1/\sigma(\gamma)$ was between $10^{11}$ and $10^{17}$. A few of his sample values were much larger, accounting for the $10^{24}$. It is hard to bound or asses the variance of the sequential importance sampling estimator.

There is something to understand: D. Bressanini has found the exact answer to the Knuth problem in sequence A00764 in the online version of Sloane's Handbook of Integer Sequences. This contains further references [23]. The number of non-self-intersecting paths in a $10 \times 10$ grid equals

$$15687580304647500132141 06 = 1.5687 \times 10^{24}.$$

This is in good agreement with Knuth's estimate $(1.6 \pm 0.3)10^{24}$. Knuth reports further exact computations in an addendum to the reprinted version of [12]. He gives the exact number of paths going through $(5, 5)$ and the exact average length of a self-avoiding path (it is about 91.9). Again, the importance sampling estimates were quite accurate. More importantly, he has pointed us to a large class of practical problems (back-tracking algorithms) where such importance type estimates are routinely used and any theoretical justification is lacking. See [11].

**5.1. Monotone paths.** As a contribution to understanding Knuth's use of sequential importance sampling we consider an easier problem where all calculations can be carried out. Let $\mathcal{X}$ be the set of all monotone paths from $(0, 0)$ to $(n, n)$ in the usual lattice. Here, paths are only allowed to go up or to the right. Thus, if $n = 2$ there are 6 paths.

In general, $|\mathcal{X}| = \binom{2n}{n}$. Shown underneath the example is $\sigma(\gamma)$ for the sequential importance sampling applied to this setting: choose one of the two available next steps with probability $1/2$ until the walk hits the top or right side of the $n \times n$ "box" when the remainder of the walk is forced. If $T(\gamma)$ is the first time the walk hits the top or right side, $\sigma(\gamma) = 2^{-T(\gamma)}$. Both the uniform distribution $\pi(\gamma) = 1/\binom{2n}{n}$ and the distribution $\sigma(\gamma) = 2^{-T(\gamma)}$ have the property that, conditional on $T(\gamma) = j$, they are uniformly distributed. Thus things are determined by the behavior of the distribution of $T(\gamma)$. The following proposition determines this for $\pi$ and $\sigma$.

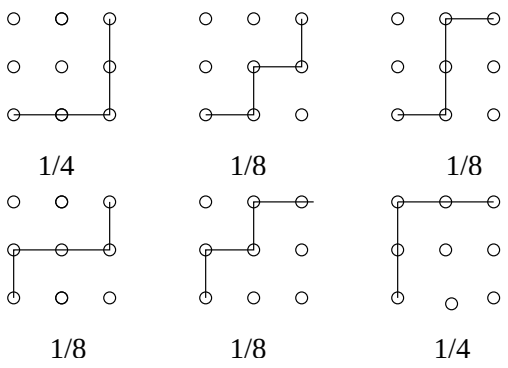


FIGURE 1

PROPOSITION 5.1. *For the monotone paths on an $n \times n$ grid we have:*

(a) *Under the uniform distribution $\pi$*

$$\pi\{T(\gamma) = j\} = 2\frac{\binom{j-1}{n-1}}{\binom{2n}{n}}, \qquad n \le j \le 2n-1.$$

(b) *For $n$ large and any fixed $k$*

$$\pi\{T(\gamma) = 2n-1-k\} \to \frac{1}{2^{k+1}}, \qquad 0 \le k < +\infty.$$

(c) *Under the importance sampling distribution $\sigma$*

$$\sigma\{T(\gamma) = j\} = 2^{1-j}\binom{j-1}{n-1}, \qquad n \le j \le 2n-1.$$

(d) *For $n$ large and any fixed positive $x$*

$$\sigma\left\{\frac{2n-1-T(\gamma)}{\sqrt{n}} \le x\right\} \to \frac{1}{\pi}\int_0^x e^{-y^2/4}dy.$$

*Proof.* Paths with $T(\gamma) = j$ may be coded as sequences of zeros and ones with $n$ zeros and $n$ ones having all zeros (or all ones) before $j$. This is twice the number with all ones before $j$. To count these, put zero at $j$ and the remaining $n-1$ zeros in the remaining $j-1$ places. This proves (a),(c). Part (b) is simple. For (d) we prove a local limit theorem. In (c), take $j = 2n-1-a$. Then,

$$\binom{j-1}{n-1} = \frac{(n-1)(n-2)\dots(n-1-a+1)}{(2n-2)(2n-3)\dots(2n-a+1)}\binom{2n-2}{n-1}.$$

Using Stirling's formula $\binom{2n-2}{n-1}2^{2n-2} \sim 1/\sqrt{\pi n}$. What is left is $2^{a-1}$ times

$$\frac{(n-1)(n-2)\dots(n-1-a+1)}{(2n-2)(2n-3)\dots(2n-a+1)} = \frac{(1-\frac{1}{n-1})\dots(1-\frac{a-1}{n-1})}{(1-\frac{1}{2(n-1)})\dots(1-\frac{a-1}{2(n-1)})}2^a.$$

Write the product in the numerator as

$$\prod_{i=1}^{a-1}\left(1-\frac{i}{n-1}\right)=\exp\left\{\sum_{i=1}^{a-1}\log\left(1-\frac{i}{n-1}\right)\right\}\sim e^{-\frac{1}{n}\binom{a}{2}}$$

where the asymptotics are valid for $a \ll n^{2/3}$. This may be justified and refined as in [6] Chapter 5. For the denominator,

$$\prod_{i=1}^{a-1}\left(1-\frac{i}{2(n-1)}\right)=\exp\left\{\sum_{i=1}^{a-1}\log\left(1-\frac{i}{2(n-1)}\right)\right\}\sim e^{-\frac{1}{2n}\binom{a}{2}}.$$

Putting things together, for $j = 2n-1-a$,

$$\sigma\{T(\gamma)=j\}\sim\frac{1}{\sqrt{\pi n}}e^{-\frac{1}{4n}a^2}.$$

For $a = x\sqrt{n}$, this last expression is $\frac{1}{\sqrt{\pi n}}e^{-x^2/4}$. Further details are omitted. □

REMARK 5.2. Thus, under the uniform distribution $\pi$, $T(\gamma)$ is close to its maximum $2n-1$. Under the importance sampling distribution $\sigma$, $T(\gamma)$ is usually $\sqrt{n}$ away from $2n-1$. We see below how our two algorithms deal with this. First, we treat the analog for estimating the size of the state space $|\mathcal{X}|$.

Let $\mu = |\mathcal{X}| = \binom{2n}{n}$. Generate paths $\gamma_i$, $1 \le i \le N$, independently from $\sigma(x)$ and set $\hat{\mu}_{SIS} = \frac{1}{N}\sum_{i=1}^{N} 1/\sigma(\gamma)$. As above, $E_\sigma(\hat{\mu}_{SIS}) = \mu$. The variance of the estimator $\hat{\mu}_{SIS}$ is given next.

PROPOSITION 5.3. *For $\mu$ the number of monotone paths in an $n \times n$ grid,*

$$\mathrm{Var}_\sigma(\hat{\mu}_{SIS})=\frac{1}{N}\frac{16^n}{4\sqrt{n}}\left(1+O\left(\frac{1}{n}\right)\right).$$

*Proof.* $\mathrm{Var}_\sigma(\hat{\mu}_{SIS}) = \frac{1}{N}\{E_\sigma(1/\sigma(\gamma)^2 - \mu^2\}$. We have

$$E_\sigma(1/\sigma(\gamma)^2)=\sum_\gamma 1/\sigma(\gamma)=\sum_{j=n}^{2n-1}2^{j+1}\binom{j-1}{n-1}.$$

This sum is dominated by its largest term. Thus

$$E(1/\sigma(\gamma)^2)=2^{2n}\binom{2n-2}{n-1}\left(1+O\left(\frac{1}{n}\right)\right).$$

On the other hand, $\mu^2 = \binom{2n}{n} \sim 16^n/(\pi n)$ is of lower order. □

REMARK 5.4. While the variance is exponentially large in $n$, the *relative* variance is

$$\mathrm{Var}_\sigma\left(\frac{\hat{\mu}_{SIS}}{\mu}\right)=\frac{1}{N}\sqrt{\pi n}\left(1+O\left(\frac{1}{n}\right)\right).$$

Thus a relatively small sample size $N$ suffices to get a useful relative error.

We next compare importance sampling and the Metropolis algorithm for estimating the mean of some simple functions of monotone paths. In our examples, the Metropolis algorithm dominates.

*Importance sampling.* Again $\mathcal{X}$ is the set of monotone paths from $(0,0)$ to $(n,n)$. Let $f : \mathcal{X} \to \mathbb{R}$ be any function. Let $\gamma_1, \ldots, \gamma_N$ be chosen independently using the sequential importance distribution $\sigma(\gamma) = 2^{-T(\gamma)}$. Then,

$$\hat{\mu}_{SIS}(f) := \frac{1}{N} \sum_{i=1}^{N} \frac{2^{T(\gamma_i)}}{\binom{2n}{n}} f(\gamma_i)$$

is an unbiased estimator of $\mu = \sum_\gamma f(\gamma)/\binom{2n}{n}$. We have

$$\mathrm{Var}_\sigma(\hat{\mu}_{SIS}(f)) = \frac{1}{N} \left\{ E\left( \frac{2^{T(\gamma)}}{\binom{2n}{n}} f(\gamma) \right)^2 - \mu^2 \right\}.$$

Using Proposition 5.3 above we may calculate this for simple functions $f$.

EXAMPLE 5.5. Let $f(\gamma) = T(\gamma)$ be the first hitting time of a uniformly chosen random path $\gamma$ to the top or right side of an $n \times n$ grid. Then

$$\mu = E_\pi(T) = \left(2 - \frac{2}{n+1}\right) n$$

and

$$\mathrm{Var}_\sigma(\hat{\mu}_{SIS}(T)) \sim \frac{\sqrt{\pi} n^{5/2}}{N}.$$

Indeed under the uniform distribution,

$$\begin{aligned} \mu &= \sum_{j=n}^{2n-1} 2j \binom{j-1}{n-1} \Big/ \binom{2n}{n} = 2n \sum_{j=n}^{2n-1} \binom{j}{n} \Big/ \binom{2n}{n} \\ &= \left(2 - \frac{2}{n+1}\right) n. \end{aligned}$$

For the variance,

$$E\left( \frac{2^{T(\gamma)}}{\binom{2n}{n}} T(\gamma) \right)^2 = \frac{1}{\binom{2n}{n}^2} \sum_\gamma 2^{2T(\gamma)} T^2(\gamma) 2^{-T(\gamma)} = \frac{2}{\binom{2n}{n}^2} \sum_{j=n}^{2n-1} 2^j j^2 \binom{j-1}{n-1}.$$

As before, the sum is dominated by its largest term. This is

$$2^{2n}(2n-1)^2 \binom{2n-2}{n-1} \Big/ \binom{2n}{n}^2 \sim \sqrt{\pi} n^{5/2}.$$

REMARK 5.6. Here, the expected squared relative error is

$$E\left(\frac{\hat{\mu}}{\mu}-1\right)^2 \sim \frac{\sqrt{\pi n}}{4N}.$$

Thus again, a sample size $N$ of order only larger than $\sqrt{n}$ suffices to give acceptable relative error.

*Metropolis sampling.* Using independence Metropolis sampling as in Section 4 with $\pi(\gamma) = 1/\binom{2n}{n}$ and proposal distribution $\sigma(\gamma) = 2^{-T(\gamma)}$, we may use Liu's result. To proceed, we must order the state space of paths with decreasing importance weights $\pi(\gamma)/\sigma(\gamma)$ and thus by largest values of $T(\gamma)$. Using the binary coding of paths introduced in the proof of Proposition 5.3 we may order via $\gamma \leq \gamma'$ if $T(\gamma) \geq T(\gamma')$. If $T(\gamma) = T(\gamma')$, use the lexicographical order. We break paths into groups by $T(\gamma)$:

$$\begin{aligned}
&\text{group one with } T(\gamma) = 2n-1 \text{ of size } A_1 := 2\binom{2n-2}{n-1},\\
&\text{group two with } T(\gamma) = 2n-2 \text{ of size } A_2 := 2\binom{2n-3}{n-1},\\
&\dots\\
&\text{group } i \text{ with } T(\gamma) = 2n-i \text{ of size } A_i := 2\binom{2n-1-i}{n-1},\\
&\text{group } n \text{ with } T(\gamma) = n \text{ of size } A_n := 2.
\end{aligned}$$

PROPOSITION 5.7. *The independence proposal Markov chain on monotone paths with proposal distribution $\sigma(\gamma) = 2^{-T(\gamma)}$ and stationary distribution $\pi(\gamma) = 1/\binom{2n}{n}$ has $n$ distinct eigenvalues on each of the $n$ groups above with multiplicity the size of the $i^{th}$–group. If $s(i) = 2^{-(2n-1-i)}$, the eigenvalues are*

$$\begin{aligned}
&\beta_1 = 1 - s(0)\binom{2n}{n}, \quad \textit{multiplicity} \quad A_1 - 1,\\
&\beta_2 = 1 - s(1)\binom{2n}{n} + A_0(s(1)-s(0)), \quad \textit{multiplicity} \quad A_2,\\
&\dots\\
&\beta_i = 1 - s(i-1)\binom{2n}{n} + A_0(s(i-1)-s(0)) + \cdots\\
&\qquad + A_{i-2}(s(i-1)-s(i-2)), \quad \textit{multiplicity} \quad A_i,\\
&\dots
\end{aligned}$$

*Proof.* This follows from equation (4.2) by elementary manipulations. □

REMARK 5.8. (a) Using Stirling's formula, for fixed $j$

$$\beta_j = 1 - \frac{2^j}{\sqrt{\pi n}} + O_j\left(\frac{1}{n}\right)$$

From Lemma 4.2, $\beta_1$ is the second largest eigenvalue.

(b) The eigenvectors of this chain are simple to write down from (4.3). We do not do this explicitly here. Using the eigenvalues, eigenvectors and convergence results from [21], we have proved that the relaxation time of the Metropolis algorithm is of order $n^{3/2}$ and there is a sharp cut-off.

To conclude this section we bound the asymptotic variance of the Metropolis estimator of the function $T(\gamma)$ and show that it improves on the importance sampling estimator of Example 5.5. Recall from (2.3) that $\sigma^2_\infty(\hat{\mu}) = \lim_N N \operatorname{Var}(\hat{\mu})$.

EXAMPLE 5.9. Let $f(\gamma) = T(\gamma)$ be the first hitting time of a uniformly chosen random path to the top or right side of an $n \times n$ grid. Then,

$$\sigma^2_\infty(\hat{\mu}_{Met}(T)) \le \{8\sqrt{\pi} n^{3/2} + O(n).\}$$

To prove the last claim, use (2.3) to write

$$\sigma^2_\infty(\hat{\mu}_{Met}) \le \frac{2}{1-\beta_1} \operatorname{Var}_\pi(T).$$

We have shown above that $1/(1-\beta_0) \le \sqrt{\pi n} + o(1)$. Now,

$$\operatorname{Var}_\pi(T) = E_\pi(T^2) - \left(2 - \frac{2}{n+1}\right)^2 n^2,$$

$$E_\pi(T^2) = \frac{2n}{\binom{2n}{n}} \sum_{i=0}^{n-1} (n+i) \binom{n+1}{n}.$$

By elementary calculations

$$4n^2 \left(1 - \frac{1}{n}\right) \left(\frac{n+\frac{1}{2}}{n+1}\right) \le E_\pi(T^2) \le \frac{4n^3}{n+1}.$$

Hence, $\operatorname{Var}_\pi(T) = 4n + O(1)$. Combing bounds gives the result.

REMARK 5.10. Roughly $\operatorname{Var}_\pi(\hat{\mu}_{SIS}) \asymp \frac{n^{5/2}}{N}$ while $\operatorname{Var}_\pi(\hat{\mu}_{MET}) \asymp \frac{n^{3/2}}{N}$.

## Acknowledgments

We thank Joseph Blitzstein, D. Bressanini, Don Knuth, Neal Madras, Ron Peled, Mauro Piccioni and Laurent Saloff-Coste for their help with this paper.

## References


[1] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu, *Sequential Monte Carlo methods for statistical analysis of tables*, J. Amer. Statist. Assoc. **100** (2005), 109–120. MR 2156822 (2006f:62062)

[2] P. Diaconis and L. Saloff-Coste, *Comparison techniques for random walk on finite groups*, Ann. Probab. **21** (1993), 2131–2156. MR 1245303 (95a:60009)

[3] P. Diaconis and L. Saloff-Coste, *What do we know about the Metropolis algorithm?*, J. Comput. System Sci. **57** (1998), 20–36. MR 1649805 (2000b:68094)

[4] P. Diaconis and D. Stroock, *Geometric bounds for eigenvalues of Markov chains*, Ann. Appl. Probab. **1** (1991), 36–61. MR 1097463 (92h:60103)

[5] *Sequential Monte Carlo methods in practice*, Statistics for Engineering and Information Science, Springer-Verlag, New York, 2001. MR 1847783 (2003h:65007)

[6] W. Feller, *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons Inc., New York, 1968. MR 0228020 (37 #3604)

[7] J. M. Hammersley and D. C. Handscomb, *Monte Carlo methods*, Methuen & Co. Ltd., London, 1965. MR 0223065 (36 #6114)

[8] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their application*, Biometrika **57** (1970), 97–109.

[9] H. Kahn, *Modification of the Monte Carlo method*, Proceedings, Seminar on Scientific Computation, November, 1949, International Business Machines Corp., New York, N. Y., 1950, pp. 20–27. MR 0044896 (13,495c)

[10] ———, *Use of different Monte Carlo sampling techniques*, Symposium on Monte Carlo methods, University of Florida, 1954, John Wiley and Sons, Inc., New York, 1956, pp. 146–190. MR 0079819 (18,151c)

[11] D. E. Knuth, *Estimating the efficiency of backtrack programs*, Math. Comp. **29** (1975), 122–136. MR 0373371 (51 #9571)

[12] ———, *Mathematics and computer science: coping with finiteness*, Science **194** (1976), 1235–1242. MR 534161 (84i:68002a)

[13] J. Liu, *Metropolized independent sampling with comparisons to rejection sampling and importance sampling*, Statistics and Computing **6** (1996), 113–119.

[14] ———, *Monte Carlo strategies in scientific computing*, Springer Series in Statistics, Springer-Verlag, New York, 2001. MR 1842342 (2002i:65006)

[15] N. Madras and M. Piccioni, *Importance sampling for families of distributions*, Ann. Appl. Probab. **9** (1999), 1202–1225. MR 1728560 (2001e:60139)

[16] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys. **21** (1953), 1087–1092.

[17] P. H. Peskun, *Optimum Monte-Carlo sampling using Markov chains*, Biometrika **60** (1973), 607–612. MR 0362823 (50 #15261)

[18] ———, *Guidelines for choosing the transition matrix in Monte Carlo methods using Markov chains*, J. Comput. Phys. **40** (1981), 327–344. MR 617102 (84c:65015)

[19] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Texts in Statistics, Springer-Verlag, New York, 2004. MR 2080278 (2005d:62006)

[20] R. Y. Rubinstein, *Simulation and the Monte Carlo method*, John Wiley & Sons Inc., New York, 1981. MR 624270 (83k:68111)

[21] L. Saloff-Coste, *Lectures on finite Markov chains*, Lectures on probability theory and statistics (Saint-Flour, 1996), Lecture Notes in Math., vol. 1665, Springer, Berlin, 1997, pp. 301–413. MR 1490046 (99b:60119)

[22] D. Siegmund, *Importance sampling in the Monte Carlo study of sequential tests*, Ann. Statist. **4** (1976), 673–684. MR 0418369 (54 #6410)

[23] N. J. A. Sloane, *The online encyclopedia of integer sequences*, `http://www.research.att.com/~njas/sequences`.

[24] A. Sokal, *Monte Carlo methods in statistical mechanics: foundations and new algorithms*, Functional integration (Cargèse, 1996), NATO Adv. Sci. Inst. Ser. B Phys., vol. 361, Plenum, New York, 1997, pp. 131–192. MR 1477456 (98k:82101)

Federico Bassetti, Università degli Studi di Pavia, Dipartimento di Matematica, via Ferrata 1, 27100 Pavia, Italy

*E-mail address*: federico.bassetti@unipv.it

Persi Diaconis, Department of Mathematics and Statistics, Stanford University, Stanford, CA 94305, USA