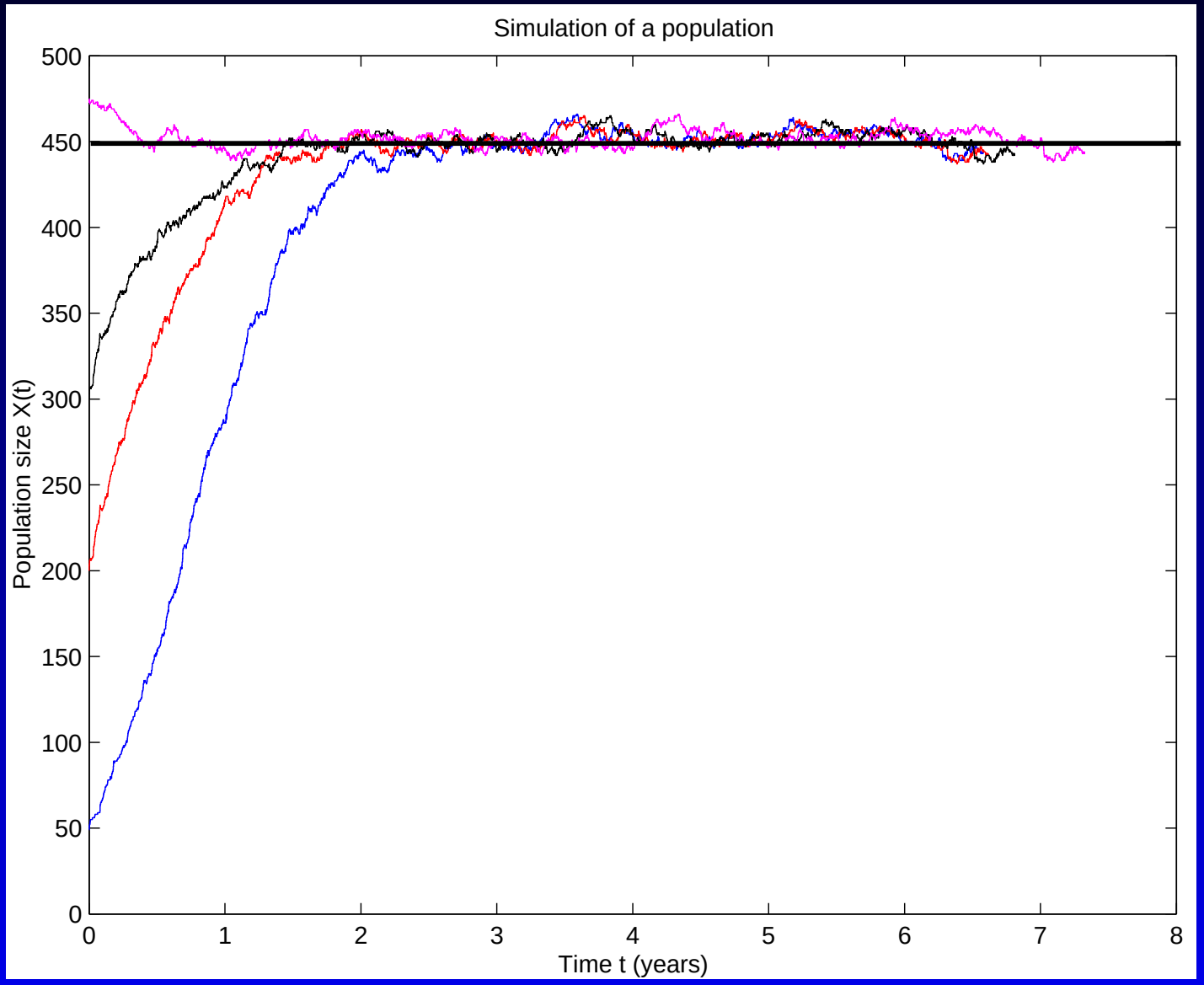
Rare Event Simulation using Importance Sampling and Cross-Entropy

Caitlin James

Supervisor: Phil Pollett


Simulation of a population
Population size X(t)
Time t (years)
0
50
100
150
200
250
300
350
400
450
500
0
1
2
3
4
5
6
7
8

Outline

- A rare event problem
- Stochastic SIS Logistic Epidemic (SIS)
- Crude Monte Carlo Method (CMCM)
- Importance Sampling (IS)
- Cross-Entropy Method (CE)
- Comparison of simulation estimates

A rare event problem

Denote $X(t)$ as the size of the population at time t.

Let N denote the maximum population size.

We wish to estimate

$$\alpha = \mathbb{P}(X(t) \text{ hits 0 before N}).$$

Notation

- Suppose $(X(t) : t \geq 0)$ is a birth-death process on finite space $\mathcal{X} = \{0, 1, \ldots, N\}$.

Figure 1: Transition diagram of a finite-state birth-death process

Notation

- Suppose $(X(t) : t \geq 0)$ is a birth-death process on finite space $\mathcal{X} = \{0, 1, \ldots, N\}$.

- Let $(X_m, m = 0, 1, \ldots)$ denote the corresponding jump chain.

Notation

- Suppose $(X(t) : t \geq 0)$ is a birth-death process on finite space $\mathcal{X} = \{0, 1, \ldots, N\}$.
- Let $(X_m, m = 0, 1, \ldots)$ denote the corresponding jump chain.
- Let $\mathcal{A}$ be the collection of all the sample paths of (X_m) & let $\mathcal{A}_0$ be the collection of all the paths that hit 0 before N.

Notation

- Suppose $(X(t) : t \geq 0)$ is a birth-death process on finite space $\mathcal{X} = \{0, 1, \ldots, N\}$.
- Let $(X_m, m = 0, 1, \ldots)$ denote the corresponding jump chain.
- Let $\mathcal{A}$ be the collection of all the sample paths of (X_m) & let $\mathcal{A}_0$ be the collection of all the paths that hit 0 before N.
- Let $\boldsymbol{X} = (X_0, X_1, \ldots)$ be a random sample path of $\mathcal{A}$ and let A be the event $\{\boldsymbol{X} \in \mathcal{A}_0\}$.

Equation

Then we can write

$$\begin{aligned} \alpha &= \mathbb{P}(A) \\ &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) f(\boldsymbol{x}; u, P) \mu(d\boldsymbol{x}). \end{aligned}$$

where

- $H(\boldsymbol{X})$ is the indicator function of the rare event A defined by

$$H(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in \mathcal{A}_0 \\ 0 & \text{if } \boldsymbol{x} \notin \mathcal{A}_0. \end{cases}$$

Equation

Then we can write

$$\begin{aligned} \alpha &= \mathbb{P}(A) \\ &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) f(\boldsymbol{x}; u, P) \mu(d\boldsymbol{x}). \end{aligned}$$

and

- $f(\boldsymbol{x}; u, P) = u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1}).$

Equation

Then we can write

$$\begin{aligned} \alpha &= \mathbb{P}(A) \\ &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) f(\boldsymbol{x}; u, P) \mu(d\boldsymbol{x}). \end{aligned}$$

and

- $f(\boldsymbol{x}; u, P) = u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1})$.
- $u = (u(i) : i \in \mathcal{X})$ is the initial distribution.

Equation

Then we can write

$$\begin{aligned} \alpha &= \mathbb{P}(A) \\ &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) f(\boldsymbol{x}; u, P) \mu(d\boldsymbol{x}). \end{aligned}$$

and

- $f(\boldsymbol{x}; u, P) = u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1})$.
- $u = (u(i) : i \in \mathcal{X})$ is the initial distribution.
- $P = (P(i,j) : i, j \in \mathcal{X})$ is the one-step transition matrix of (X_m).

Stochastic SIS Logistic Epidemic

- The SIS model is a finite-state birth and death process for which the origin is an absorbing state.

Figure 2: Transition diagram of SIS model

Stochastic SIS Logistic Epidemic

- The SIS model is a finite-state birth and death process for which the origin is an absorbing state.
- The rate of infection per contact is denoted by λ and μ denotes the per-capita death rate.

Figure 2: Transition diagram of SIS model

Stochastic SIS Logistic Epidemic

The non-zero transition rates are defined as

$$\lambda_i = \lambda \frac{1}{N} i(N-i) \quad \text{and} \quad \mu_i = \mu i.$$

Stochastic SIS Logistic Epidemic

The non-zero transition rates are defined as

$$\lambda_i = \lambda \frac{1}{N} i(N-i) \quad \text{and} \quad \mu_i = \mu i.$$

The jump chain (X_m) has jump probabilities

$$p_i = \frac{\lambda(N-i)}{\mu N + \lambda(N-i)}$$

$$\text{and} \quad q_i = (1-p_i) = \frac{\mu N}{\mu N + \lambda(N-i)}.$$

Crude Monte Carlo Method

We can estimate our probability using CMCM. This involves simulating n replicates, $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$, from $f(\cdot\,; u, P)$ and setting

$$\hat{\alpha} = \frac{1}{n} \sum_{k=1}^{n} H(\boldsymbol{X}^k).$$

We simulate our model until the process hits N or 0. We count 1 if the process hits 0.

Crude Monte Carlo Method

We can estimate our probability using CMCM. This involves simulating n replicates, $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n$, from $f(\,\cdot\,; u, P)$ and setting

$$\hat{\alpha} = \frac{1}{n} \sum_{k=1}^{n} H(\boldsymbol{X}^k).$$

The number of trials needed to get one successful trial has a geometric distribution with the expected value being $1/p$, where p is the probability of a successful trial.

CMCM example

If we start in State 8 and we wish to estimate the probability of reaching 0 before 10, then we would require $1/\alpha$ runs to see one successful path hit 0 before 10. (Failure is hitting 10 before 0.)

CMCM example

If we start in State 8 and we wish to estimate the probability of reaching 0 before 10, then we would require $1/\alpha$ runs to see one successful path hit 0 before 10. (Failure is hitting 10 before 0.)

The exact value of α starting in state 8 is 1.18E-05. Thus we need around $1/1.18$E-$05 \approx 85,000$ runs to see our first path hit 0 before 10.

CMCM example

If we start in State 8 and we wish to estimate the probability of reaching 0 before 10, then we would require $1/\alpha$ runs to see one successful path hit 0 before 10. (Failure is hitting 10 before 0.)

The exact value of α starting in state 8 is $1.18\text{E-}05$. Thus we need around $1/1.18\text{E-}05 \approx 85,000$ runs to see our first path hit 0 before 10.

If we use CMCM, then we would require many more than $85,000$ runs to obtain a reasonable estimate for α.

Importance Sampling

- $\tilde{u}$ to be an alternative initial distribution

Importance Sampling

- $\tilde{u}$ to be an alternative initial distribution
- $\tilde{P}$ to be an alternative transition matrix

Importance Sampling

- $\tilde{u}$ to be an alternative initial distribution
- $\tilde{P}$ to be an alternative transition matrix
- $g(\boldsymbol{x}; \tilde{u}, \tilde{P}) = \tilde{u}(x_0) \prod_{k=0}^{m-1} \tilde{P}(x_k, x_{k+1})$

Importance Sampling

- $\tilde{u}$ to be an alternative initial distribution
- $\tilde{P}$ to be an alternative transition matrix
- $g(\boldsymbol{x}; \tilde{u}, \tilde{P}) = \tilde{u}(x_0) \prod_{k=0}^{m-1} \tilde{P}(x_k, x_{k+1})$

The following conditions must also hold:

$$u(i) > 0 \quad \text{implies} \quad \tilde{u}(i) > 0$$

$$\text{and} \quad P(i,j) > 0 \quad \text{implies} \quad \tilde{P}(i,j) > 0.$$

Importance Sampling

Under the alternative measure g we can estimate α by

$$\begin{aligned}\alpha &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) \frac{f(\boldsymbol{x}; u, P)}{g(\boldsymbol{x}; \tilde{u}, \tilde{P})} g(\boldsymbol{x}; \tilde{u}, \tilde{P}) \mu(d\boldsymbol{x}) \\ &= \mathbb{E}_g \left(H(\boldsymbol{X}) L_T(\boldsymbol{X}; u, P, \tilde{u}, \tilde{P}); T < \infty \right),\end{aligned}$$

where

Importance Sampling

Under the alternative measure g we can estimate α by

$$\begin{aligned}\alpha &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) \frac{f(\boldsymbol{x}; u, P)}{g(\boldsymbol{x}; \tilde{u}, \tilde{P})} g(\boldsymbol{x}; \tilde{u}, \tilde{P}) \mu(d\boldsymbol{x}) \\ &= \mathbb{E}_g \left(H(\boldsymbol{X}) L_T(\boldsymbol{X}; u, P, \tilde{u}, \tilde{P}); T < \infty \right),\end{aligned}$$

where

$$L_m(\boldsymbol{x}; u, P, \tilde{u}, \tilde{P}) = \frac{f(\boldsymbol{x}; u, P)}{g(\boldsymbol{x}; \tilde{u}, \tilde{P})} = \frac{u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1})}{\tilde{u}(x_0) \prod_{k=0}^{m-1} \tilde{P}(x_k, x_{k+1})}.$$

Importance Sampling

Under the alternative measure g we can estimate α by

$$\begin{aligned}\alpha &= \mathbb{E}_f H(\boldsymbol{X}) = \int_{\mathcal{A}} H(\boldsymbol{x}) \frac{f(\boldsymbol{x}; u, P)}{g(\boldsymbol{x}; \tilde{u}, \tilde{P})} g(\boldsymbol{x}; \tilde{u}, \tilde{P}) \mu(d\boldsymbol{x}) \\ &= \mathbb{E}_g \left(H(\boldsymbol{X}) L_T(\boldsymbol{X}; u, P, \tilde{u}, \tilde{P}); T < \infty \right),\end{aligned}$$

where

$$L_m(\boldsymbol{x}; u, P, \tilde{u}, \tilde{P}) = \frac{f(\boldsymbol{x}; u, P)}{g(\boldsymbol{x}; \tilde{u}, \tilde{P})} = \frac{u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1})}{\tilde{u}(x_0) \prod_{k=0}^{m-1} \tilde{P}(x_k, x_{k+1})}.$$

with $T = \inf\{m : X_m = 0 \text{ or } X_m = N\}$ ("Stopping Time")

Importance Sampling estimator

We can estimate α using the *importance sampling* (IS) estimator, defined by

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} H(\boldsymbol{X}) L_m(\boldsymbol{X}; u, P, \tilde{u}, \tilde{P}),$$

where (recall)

$$L_m(\boldsymbol{x}; u, P, \tilde{u}, \tilde{P}) = \frac{f(\boldsymbol{x}; u, P)}{g(\boldsymbol{x}; \tilde{u}, \tilde{P})} = \frac{u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1})}{\tilde{u}(x_0) \prod_{k=0}^{m-1} \tilde{P}(x_k, x_{k+1})}.$$

Optimal IS density

Which change of measure gives the IS estimator with smallest variance?

Optimal IS density

Which change of measure gives the IS estimator with smallest variance?

The smallest variance is obtained when $g = g^*$, the *optimal importance sampling density*, given by

$$g^*(\boldsymbol{x}) = \frac{|H(\boldsymbol{x})| f(\boldsymbol{x}; u, P)}{\int_{\mathcal{A}} |H(\boldsymbol{x})| f(\boldsymbol{x}; u, P) \mu(d\boldsymbol{x})}.$$

Optimal IS density

Which change of measure gives the IS estimator with smallest variance?

The smallest variance is obtained when $g = g^*$, the *optimal importance sampling density*, given by

$$g^*(\boldsymbol{x}) = \frac{|H(\boldsymbol{x})| f(\boldsymbol{x}; u, P)}{\int_{\mathcal{A}} |H(\boldsymbol{x})| f(\boldsymbol{x}; u, P) \mu(d\boldsymbol{x})}.$$

If $H(\boldsymbol{x}) \geq 0$, then

$$g^*(\boldsymbol{x}) = \frac{H(\boldsymbol{x}) f(\boldsymbol{x}; u, P)}{\alpha}.$$

Cross-Entropy Method

The Kullback-Leibler Cross-Entropy (CE) measure defines a "distance" between two densities g and h

$$\begin{aligned} D(g,h) &= \int g(\boldsymbol{x}) \ln \frac{g(\boldsymbol{x})}{h(\boldsymbol{x})} \mu(d\boldsymbol{x}) \\ &= \int g(\boldsymbol{x}) \ln g(\boldsymbol{x}) \mu(d\boldsymbol{x}) - \int g(\boldsymbol{x}) \ln h(\boldsymbol{x}) \mu(d\boldsymbol{x}). \end{aligned}$$

The purpose of CE is to choose the IS density h such that the "distance" between the optimal IS density g^* and density h is as small as possible.

Properties of CE

1. $D(\cdot\,,\cdot)$ is non-symmetric ie: $D(g,h) \neq D(h,g)$, thus $D(g,h)$ is not a true distance between g and h in a formal sense, although

2. $D(g,h) \geq 0$.

3. $D(g,g) = 0$.

Restrict the IS density

If we restrict the density to belong to some family $\mathcal{F}$ which contains the original density

$$f(\boldsymbol{x}; u, P) = u(x_0) \prod_{k=0}^{m-1} P(x_k, x_{k+1})$$

and the alternative density

$$f(\boldsymbol{x}; \tilde{u}, \tilde{P}) = \tilde{u}(x_0) \prod_{k=0}^{m-1} \tilde{P}(x_k, x_{k+1})$$

CE optimisation problem

Then the CE method aims to solve the parametric optimisation problem

$$\min_{(\tilde{u},\tilde{P})} D(g^*, f(\,\cdot\,;\tilde{u},\tilde{P})),$$

where (recall)

$$g^*(\boldsymbol{x}) = \frac{H(\boldsymbol{x}) f(\,\cdot\,; u, P\,)}{\alpha}.$$

CE reference parameter

Since $f(\,\cdot\,; u, P)$ does not depend on $(\tilde{u}, \tilde{P})$, minimising the CE distance between g^* and $f(\,\cdot\,; \tilde{u}, \tilde{P})$ is equivalent to *maximising*, with respect to $(\tilde{u}, \tilde{P})$,

$$\int |H(\boldsymbol{x})| f(\boldsymbol{x}; u, P) \ln f(\boldsymbol{x}; \tilde{u}, \tilde{P}) \mu(d\boldsymbol{x})$$

$$= \mathbb{E}_{(u,P)} |H(\boldsymbol{X})| \ln f(\boldsymbol{X}; \tilde{u}, \tilde{P}).$$

CE reference parameter

Since $f(\,\cdot\,;u,P)$ does not depend on $(\tilde{u},\tilde{P})$, minimising the CE distance between g^* and $f(\,\cdot\,;\tilde{u},\tilde{P})$ is equivalent to *maximising*, with respect to $(\tilde{u},\tilde{P})$,

$$\int |H(\boldsymbol{x})| f(\boldsymbol{x};u,P) \ln f(\boldsymbol{x};\tilde{u},\tilde{P}) \mu(d\boldsymbol{x})$$

$$= \mathbb{E}_{(u,P)} |H(\boldsymbol{X})| \ln f(\boldsymbol{X};\tilde{u},\tilde{P}).$$

Assuming $H(\boldsymbol{X}) \geq 0$, the optimal $(\tilde{u},\tilde{P})$ (with respect to CE) is the solution to

$$(\tilde{u}^*,\tilde{P}^*) = \arg\max_{(\tilde{u},\tilde{P})} \mathbb{E}_{(u,P)} H(\boldsymbol{X}) \ln f(\boldsymbol{X};\tilde{u},\tilde{P}).$$

CE reference parameter

Since $f(\,\cdot\,;u,P)$ does not depend on $(\tilde{u},\tilde{P})$, minimising the CE distance between g^* and $f(\,\cdot\,;\tilde{u},\tilde{P})$ is equivalent to *maximising*, with respect to $(\tilde{u},\tilde{P})$,

$$\int |H(\boldsymbol{x})| f(\boldsymbol{x};u,P) \ln f(\boldsymbol{x};\tilde{u},\tilde{P}) \mu(d\boldsymbol{x})$$

$$= \mathbb{E}_{(u,P)} |H(\boldsymbol{X})| \ln f(\boldsymbol{X};\tilde{u},\tilde{P}).$$

Assuming $H(\boldsymbol{X}) \geq 0$, the optimal $\tilde{P}$ (with respect to CE) is the solution to

$$\tilde{P}^* = \arg\max_{\tilde{P}} \mathbb{E}_{(u,P)} H(\boldsymbol{X}) \ln f(\boldsymbol{X};u,\tilde{P}).$$

CE reference parameter

The optimal transition probability matrix is given by

$$\tilde{P}^*(i,j) = \frac{\mathbb{E}_{(u,P)} H(\boldsymbol{X}) \sum_{k:X_k=i} 1_{(X_{k+1}=j)}}{\mathbb{E}_{(u,P)} H(\boldsymbol{X}) \sum_{k:X_k=i} 1}.$$

We can approximate this optimal transition probability matrix by implementing IS to obtain:

CE reference parameter

The optimal transition probability matrix is given by

$$\tilde{P}^*(i,j) = \frac{\mathbb{E}_{(u,P)} H(\boldsymbol{X}) \sum_{k:X_k=i} 1_{(X_{k+1}=j)}}{\mathbb{E}_{(u,P)} H(\boldsymbol{X}) \sum_{k:X_k=i} 1}.$$

We can approximate this optimal transition probability matrix by implementing IS to obtain:

$$\tilde{P}^*_{l+1}(i,j) = \frac{\mathbb{E}_{(u,\tilde{P}_l)} H(\boldsymbol{X}) L_m(\boldsymbol{X}; u, P, u, \tilde{P}_l) \sum_{k:X_k=i} 1_{(X_{k+1}=j)}}{\mathbb{E}_{(u,\tilde{P}_l)} H(\boldsymbol{X}) L_m(\boldsymbol{X}; u, P, u, \tilde{P}_l) \sum_{k:X_k=i} 1}$$

CE reference parameter

The optimal transition probability matrix is given by

$$\tilde{P}^*(i,j) = \frac{\mathbb{E}_{(u,P)} H(\boldsymbol{X}) \sum_{k:X_k=i} 1_{(X_{k+1}=j)}}{\mathbb{E}_{(u,P)} H(\boldsymbol{X}) \sum_{k:X_k=i} 1}.$$

We can approximate this optimal transition probability matrix by implementing IS to obtain:

$$\tilde{P}^*_{l+1}(i,j) \approx \frac{\sum_{\boldsymbol{X}=\boldsymbol{X}^1}^{\boldsymbol{X}^n} H(\boldsymbol{X}) L_m(\boldsymbol{X}; u, P, u, \tilde{P}_l) \sum_{k:X_k=i} 1_{(X_{k+1}=j)}}{\sum_{\boldsymbol{X}=\boldsymbol{X}^1}^{\boldsymbol{X}^n} H(\boldsymbol{X}) L_m(\boldsymbol{X}; u, P, u, \tilde{P}_l) \sum_{k:X_k=i} 1},$$

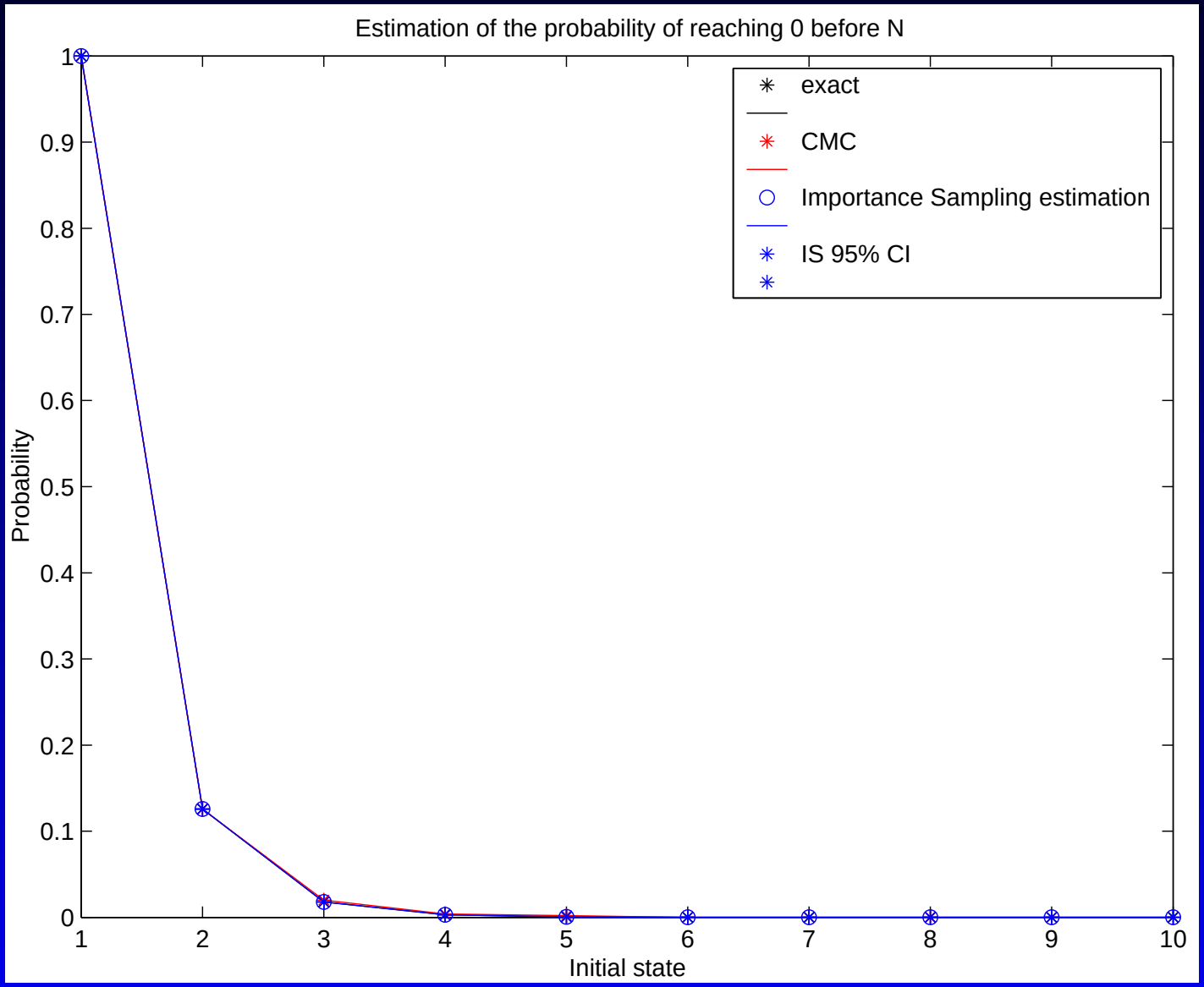
Initial CE parameter

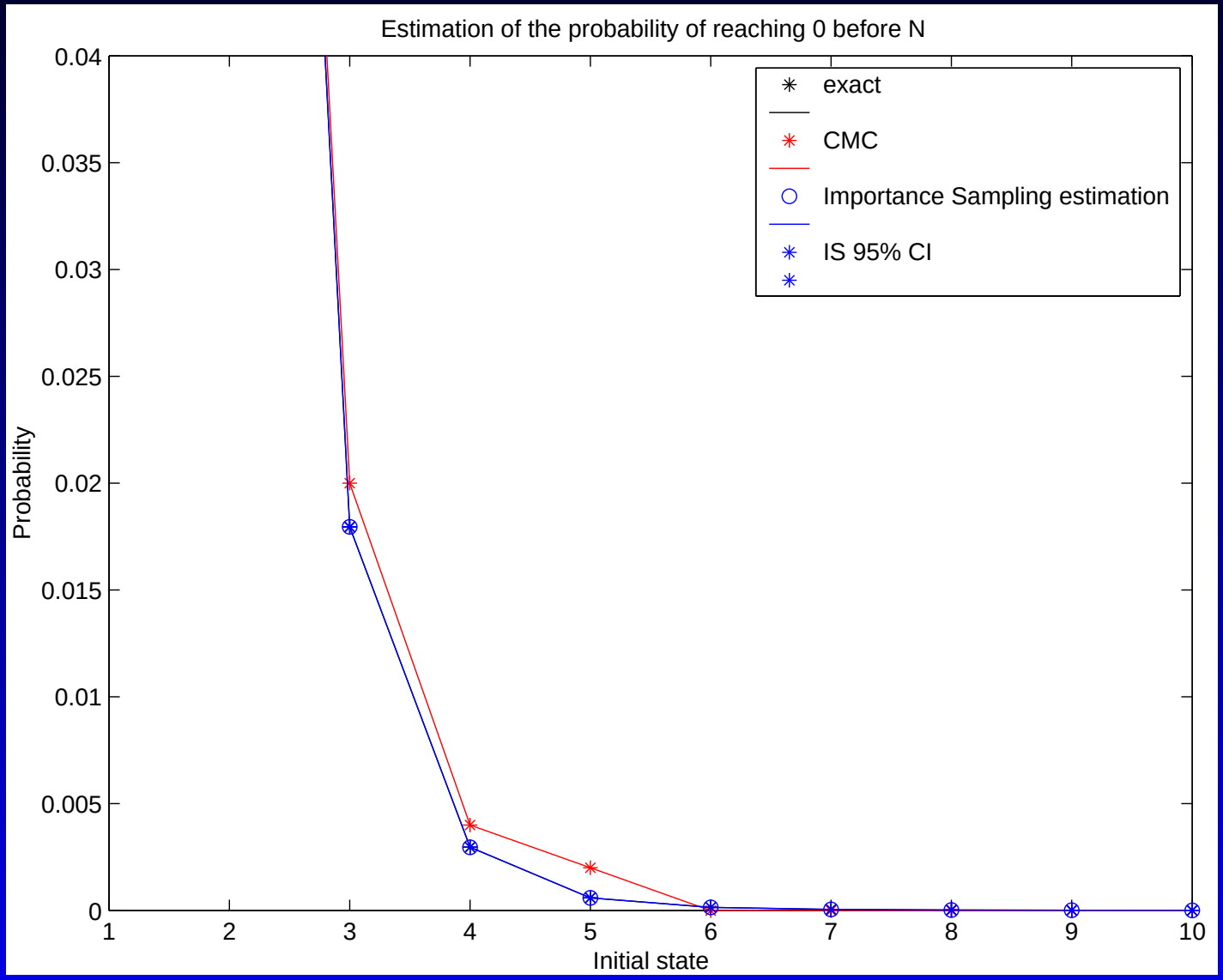
Original parameters:

$$p_i = \frac{\lambda(N-i)}{\mu N + \lambda(N-i)} \quad \text{and} \quad q_i = (1-p_i) = \frac{\mu N}{\mu N + \lambda(N-i)}.$$

Initial change of measures:

$$\tilde{p}_i = \frac{\mu(N-i)}{\lambda N + \mu(N-i)} \quad \text{and} \quad \tilde{q}_i = (1-\tilde{p}_i) = \frac{\lambda N}{\lambda N + \mu(N-i)}.$$


Estimation of the probability of reaching 0 before N
exact
CMC
Importance Sampling estimation
IS 95% CI
Probability
Initial state
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
2
3
4
5
6
7
8
9
10


Estimation of the probability of reaching 0 before N
exact
CMC
Importance Sampling estimation
IS 95% CI
Probability
Initial state
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
1
2
3
4
5
6
7
8
9
10

Comparison of simulation estimates

$N = 10$, $\lambda = 0.9$, $\mu = 0.1$, $\rho = 0.11$, $X_0 = 8$,

No. of CE runs$= 4$

Sample Size	Exact	IS	CMCM
50	1.18E-05	1.40E-05	0
100	1.18E-05	1.64E-05	0
1000	1.18E-05	1.18E-05	0
2000	1.18E-05	1.18E-05	0

Comparison of CE runs

$N = 10$, $\lambda = 0.9$, $\mu = 0.1$, $\rho = 0.11$, $X_0 = 8$,

Sample Size $= 1000$, Exact $= 1.18$E-05

CE run	IS	2 StD	$\tilde{p}_1$	$\tilde{p}_5$	$\tilde{p}_9$
0	1.47E-05	3.21E-06	0.091	0.053	0.011
1	1.16E-05	1.79E-06	0.091	0.212	0
2	1.18E-05	4.12E-07	0.127	0.256	0
3	1.18E-05	1.38E-07	0.118	0.277	0
4	1.18E-05	9.09E-08	0.125	0.282	0
5	1.18E-05	5.89E-08	0.134	0.270	0

Acknowledgements

- Phil Pollett
- Joshua Ross
- David Sirl
- Ben Gladwin