Approximate Zero-Variance Simulation in Rare Event Settings

Pierre L'Ecuyer
Université de Montréal, Canada

partly based on joint work with

Jose H. Blanchet, Columbia University, New York, USA
Zdravko Botev, Université de Montréal, Canada
Peter W. Glynn, Stanford University, USA
Bruno Tuffin, INRIA, Rennes, France

IMACS MCM 2011, Borovets, Bulgaria, August 2011

Monte Carlo integration: basic setting

We want to estimate $\mu_0 = \mathbb{E}[X]$ where X is the output of a stochastic simulation. Basic Monte Carlo (MC) method:

- Generate n independent replicates of X, say $X_1, \ldots, X_n$;
- estimate μ_0 by $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Monte Carlo integration: basic setting

We want to estimate $\mu_0 = \mathbb{E}[X]$ where X is the output of a stochastic simulation. Basic Monte Carlo (MC) method:

- Generate n independent replicates of X, say $X_1, \ldots, X_n$;
- estimate μ_0 by $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Strong law of large numbers: $\bar{X}_n \to \mu_0$ with probability 1 when $n \to \infty$.
For confidence interval on μ_0, can use central limit theorem:

$$\mathbb{P}\left[\mu_0 \in \left(\bar{X}_n - \frac{c_\alpha S_n}{\sqrt{n}},\ \ \bar{X}_n + \frac{c_\alpha S_n}{\sqrt{n}}\right)\right] \approx 1 - \alpha$$

where S_n^2 is an estimator of $\sigma^2 = \mathrm{Var}[X]$.

Accuracy of estimator X can be measured by half-width $c_\alpha n^{-1/2}\sigma$ or relative half-width $c_\alpha n^{-1/2}\sigma/\mu_0$ of confidence interval.

Monte Carlo integration: basic setting

We want to estimate $\mu_0 = \mathbb{E}[X]$ where X is the output of a stochastic simulation. Basic Monte Carlo (MC) method:

- Generate n independent replicates of X, say $X_1, \ldots, X_n$;
- estimate μ_0 by $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Strong law of large numbers: $\bar{X}_n \to \mu_0$ with probability 1 when $n \to \infty$.
For confidence interval on μ_0, can use central limit theorem:

$$\mathbb{P}\left[\mu_0 \in \left(\bar{X}_n - \frac{c_\alpha S_n}{\sqrt{n}},\ \ \bar{X}_n + \frac{c_\alpha S_n}{\sqrt{n}}\right)\right] \approx 1 - \alpha$$

where S_n^2 is an estimator of $\sigma^2 = \mathrm{Var}[X]$.

Accuracy of estimator X can be measured by half-width $c_\alpha n^{-1/2}\sigma$ or relative half-width $c_\alpha n^{-1/2}\sigma/\mu_0$ of confidence interval.

Other types of situations: estimate a quantile, optimization, etc.

Example: a static network reliability problem

The system has d components, in state 1 (failed) or 0 (operating).

System state: $\mathbf{B} = (B_1, \ldots, B_d)$.

Complementary structure function: $\Phi : \{0,1\}^d \to \{0,1\}$.

System failed iff $X \stackrel{\text{def}}{=} \Phi(\mathbf{B}) = 1$.

Unreliability: $u = \mathbb{P}[\Phi(\mathbf{B}) = 1]$.

Example: a static network reliability problem

The system has d components, in state 1 (failed) or 0 (operating).
System state: $\mathbf{B} = (B_1, \ldots, B_d)$.
Complementary structure function: $\Phi : \{0,1\}^d \to \{0,1\}$.

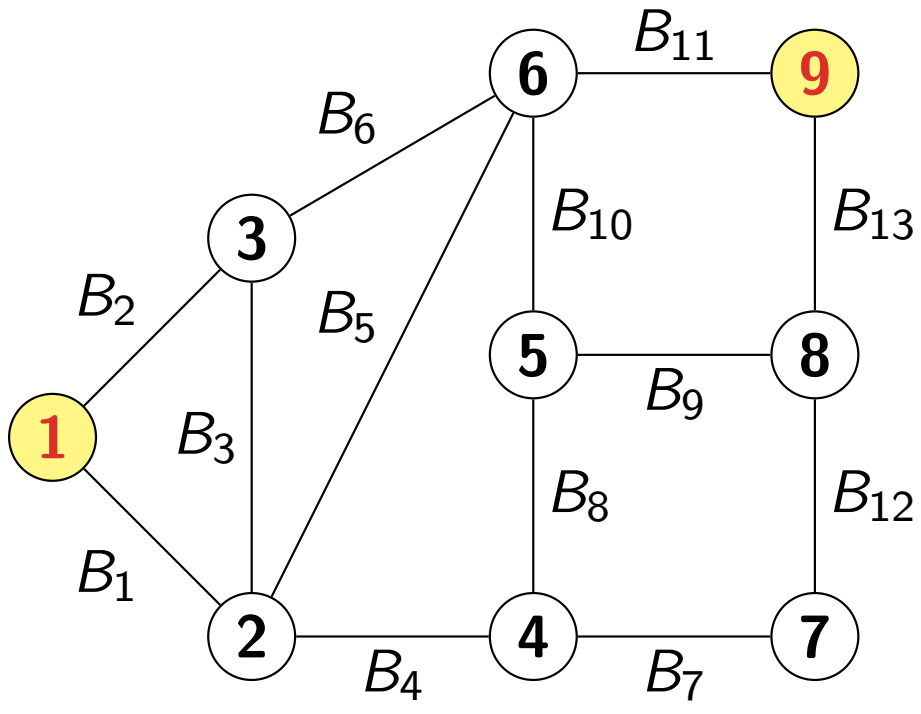
System failed iff $X \stackrel{\text{def}}{=} \Phi(\mathbf{B}) = 1$.

Unreliability: $u = \mathbb{P}[\Phi(\mathbf{B}) = 1]$.

Monte Carlo: Generate n i.i.d. realizations of $\mathbf{B}$, say $\mathbf{B}_1, \ldots, \mathbf{B}_n$, compute $X_i = \Phi(\mathbf{B}_i)$ for each i, and estimate u by $\bar{X}_n = (X_1 + \cdots + X_n)/n$.

For example, easy if the components are independent and $\mathbb{P}[B_j = 1] = q_j$.

In our examples, Φ is defined via a graph (or network).
Link j "works" iff $B_j = 0$.
The system works if all nodes in a given set $\mathcal{K}$ are connected.



Given the B_j's, $X = \Phi(\mathbf{B})$ is easy to evaluate by graph algorithms (e.g., minimal spanning tree).

Here, u is very close to 0 (failure is a rare event). For example, if $u = 10^{-10}$, the system will fail only once per 10 billion runs on average.

Here, u is very close to 0 (failure is a rare event). For example, if $u = 10^{-10}$, the system will fail only once per 10 billion runs on average.

In fact, X is Bernoulli with $\mathbb{E}[X] = u$, $\mathrm{Var}[X] = u(1-u)$, and

$$\mathrm{MSE}[\bar{X}_n] \overset{\text{here}}{=} \mathrm{Var}[\bar{X}_n] = \frac{u(1-u)}{n} \approx \frac{u}{n}.$$

We want at least to beat the trivial estimator $Y = 0$, for which $\mathrm{MSE}[Y] = \mathrm{bias}^2[Y] = u^2$.

Here, u is very close to 0 (failure is a rare event). For example, if $u = 10^{-10}$, the system will fail only once per 10 billion runs on average.

In fact, X is Bernoulli with $\mathbb{E}[X] = u$, $\mathrm{Var}[X] = u(1-u)$, and

$$\mathrm{MSE}[\bar{X}_n] \overset{\text{here}}{=} \mathrm{Var}[\bar{X}_n] = \frac{u(1-u)}{n} \approx \frac{u}{n}.$$

We want at least to beat the trivial estimator $Y = 0$, for which $\mathrm{MSE}[Y] = \mathrm{bias}^2[Y] = u^2$.

When u is small, a relevant quality measure is the relative error

$$\mathrm{RE}[\bar{X}_n] \overset{\text{def}}{=} \frac{\sqrt{\mathrm{MSE}[\bar{X}_n]}}{u} \overset{\text{here}}{=} \frac{\sqrt{1-u}}{\sqrt{nu}} \to \infty \quad \text{when } u \to 0.$$

For example, if $u \approx 10^{-10}$, to have $\mathrm{RE}[\bar{X}_n] \leq 10\%$ we need $n \approx 10^{12}$. Requires much more efficient methods than crude MC!

Examples of situations involving rare events

- Probability that the completion time of a large project exceeds a given threshold.
- Probability of buffer overflow, or mean time to overflow, in a queueing system.
- Proportion of packets lost in a communication system.
- Probability of a large loss from an investment portfolio.
- Ruin probability for an insurance firm.
- Value-at-risk (quantile estimation).
- Expected amount of radiation that crosses a given protection shield.
- Air traffic control.
- Mean time to failure or other reliability or availability measure for a highly reliable system (e.g., fault-tolerant computers, safety systems).
- Artificial settings: counting problems, combinatorial optimization, etc.

A framework for asymptotic analysis

We estimate a small quantity $\mu_0 = \mu_0(\varepsilon) > 0$, where $\mu_0(\varepsilon) \to 0$ when the rarity parameter $\varepsilon \to 0$, by an unbiased estimator $X = X(\varepsilon) \geq 0$.

A framework for asymptotic analysis

We estimate a small quantity $\mu_0 = \mu_0(\varepsilon) > 0$, where $\mu_0(\varepsilon) \to 0$ when the rarity parameter $\varepsilon \to 0$, by an unbiased estimator $X = X(\varepsilon) \geq 0$.

In a queueing system with buffer size B and s servers, we can take $\varepsilon = 1/B$ if we are interested in very large values of B, and $\varepsilon = 1/s$ if we are interested in what happens when there is a large number of servers.

In a reliability model, the failure probabilities or failure rates may be taken as polynomial functions of ε.

A framework for asymptotic analysis

We estimate a small quantity $\mu_0 = \mu_0(\varepsilon) > 0$, where $\mu_0(\varepsilon) \to 0$ when the rarity parameter $\varepsilon \to 0$, by an unbiased estimator $X = X(\varepsilon) \geq 0$.

In a queueing system with buffer size B and s servers, we can take $\varepsilon = 1/B$ if we are interested in very large values of B, and $\varepsilon = 1/s$ if we are interested in what happens when there is a large number of servers.

In a reliability model, the failure probabilities or failure rates may be taken as polynomial functions of ε.

We study the asymptotic behavior when $\varepsilon \to 0$ to understand what happens when ε is very small.

With standard MC, we often have $\mathrm{RE}[X(\varepsilon)] \to \infty$ when $\varepsilon \to 0$.

Classical robustness properties in this context

Commonly-used characterizations of $X(\varepsilon)$ in rare-event setting:

- It has bounded relative error (BRE) (bounded relative variance) if

$$\lim_{\varepsilon \to 0} \frac{\mathrm{Var}[X(\varepsilon)]}{\mu_0^2(\varepsilon)} < \infty.$$

Classical robustness properties in this context

Commonly-used characterizations of $X(\varepsilon)$ in rare-event setting:

- It has bounded relative error (BRE) (bounded relative variance) if

$$\lim_{\varepsilon \to 0} \frac{\mathrm{Var}[X(\varepsilon)]}{\mu_0^2(\varepsilon)} < \infty.$$

- It is logarithmically efficient (LE) or asymptotically optimal if

$$\lim_{\varepsilon \to 0} \frac{\ln \mathbb{E}[X^2(\varepsilon)]}{2 \ln \mu_0(\varepsilon)} = 1.$$

  Means (roughly) that if $\mu_0(\varepsilon) \to 0$ at an exponential rate, then the standard deviation converges at least at the same exponential rate.

- BRE is stronger than LE, and can be more difficult to reach.

Generalization: BRM-k and LE-k

L., Blanchet, Glynn, Tuffin (2010)

An estimator $X(\varepsilon)$ with mean $\mu_0(\varepsilon)$ has bounded relative moment of order k (BRM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[X^k(\varepsilon)]}{\mu_0^k(\varepsilon)} < \infty.$$

Generalization: BRM-k and LE-k

L., Blanchet, Glynn, Tuffin (2010)

An estimator $X(\varepsilon)$ with mean $\mu_0(\varepsilon)$ has bounded relative moment of order k (BRM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[X^k(\varepsilon)]}{\mu_0^k(\varepsilon)} < \infty.$$

It has logarithmic efficiency of order k (LE-k) if

$$\lim_{\varepsilon \to 0} \frac{\ln \mathbb{E}[X^k(\varepsilon)]}{k \ln \mu_0(\varepsilon)} = 1.$$

Interesting and relevant for situations where we need estimators of the variance or of other moments higher than the mean.

Relevant for the validity of Berry-Esseen bound, for example.

Generalization: BRM-k and LE-k

L., Blanchet, Glynn, Tuffin (2010)

An estimator $X(\varepsilon)$ with mean $\mu_0(\varepsilon)$ has bounded relative moment of order k (BRM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[X^k(\varepsilon)]}{\mu_0^k(\varepsilon)} < \infty.$$

It has logarithmic efficiency of order k (LE-k) if

$$\lim_{\varepsilon \to 0} \frac{\ln \mathbb{E}[X^k(\varepsilon)]}{k \ln \mu_0(\varepsilon)} = 1.$$

Interesting and relevant for situations where we need estimators of the variance or of other moments higher than the mean.

Relevant for the validity of Berry-Esseen bound, for example.

Is BRM-k the best that we can hope for?

Vanishing relative moments

$X(\varepsilon)$ has vanishing relative centered moment of order k (VRCM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[|X(\varepsilon) - \mu_0(\varepsilon)|^k]}{\mu_0^k(\varepsilon)} = 0.$$

Vanishing relative moments

$X(\varepsilon)$ has vanishing relative centered moment of order k (VRCM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[|X(\varepsilon) - \mu_0(\varepsilon)|^k]}{\mu_0^k(\varepsilon)} = 0.$$

True if and only if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[X^k(\varepsilon)]}{\mu_0^k(\varepsilon)} = 1.$$

Vanishing relative moments

$X(\varepsilon)$ has vanishing relative centered moment of order k (VRCM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[|X(\varepsilon) - \mu_0(\varepsilon)|^k]}{\mu_0^k(\varepsilon)} = 0.$$

True if and only if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[X^k(\varepsilon)]}{\mu_0^k(\varepsilon)} = 1.$$

It has vanishing relative variance or relative error (VRE), if

$$\limsup_{\varepsilon \to 0} \frac{\sigma(\varepsilon)}{\mu_0(\varepsilon)} = 0.$$

When VRCM occurs, the rare event difficulty is reversed! May seem strange and perhaps unachievable at first sight, but does happen.

Challenge in rare-event simul.: build estimators with these properties.

Vanishing relative moments

$X(\varepsilon)$ has vanishing relative centered moment of order k (VRCM-k) if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[|X(\varepsilon) - \mu_0(\varepsilon)|^k]}{\mu_0^k(\varepsilon)} = 0.$$

True if and only if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}[X^k(\varepsilon)]}{\mu_0^k(\varepsilon)} = 1.$$

It has vanishing relative variance or relative error (VRE), if

$$\limsup_{\varepsilon \to 0} \frac{\sigma(\varepsilon)}{\mu_0(\varepsilon)} = 0.$$

When VRCM occurs, the rare event difficulty is reversed! May seem strange and perhaps unachievable at first sight, but does happen.

Challenge in rare-event simul.: build estimators with these properties.

Is VRCM the best we can dream of?

Ultimate dream: a zero-variance estimator

Can be achieved in principle via importance sampling (IS) or via control variates (CV), as we will see later in the talk.

Exact implementation of this is impractical, since it would require the knowledge of μ_0 (and usually much more) in the first place.

Ultimate dream: a zero-variance estimator

Can be achieved in principle via importance sampling (IS) or via control variates (CV), as we will see later in the talk.

Exact implementation of this is impractical, since it would require the knowledge of μ_0 (and usually much more) in the first place.

But by plugging crude approximations of the unknown quantities in place of the exact ones in the zero-variance sampling strategies, we may reduce the variance tremendously, and sometimes the convergence rate as well.

This is what we call **approximate zero-variance simulation**.

Has been studied for IS by Booth (1985, 1987), Kollman et al. (1999), Baggerly et al. (2000), and for CV by Henderson and Glynn 2002, Gobet and Maire (2006), and Kim and Henderson (2006, 2007), among others.

Importance Sampling (IS)

Suppose $X = h(Y)$ where Y is a random vector with density f.
Instead of generating Y from f, we can generate it from another density $\tilde{f}$, such that $\tilde{f}(y) > 0$ whenever $h(y)f(y) \neq 0$. We have

$$\mathbb{E}[X] = \int h(y)f(y)dy = \int \left[\frac{h(y)f(y)}{\tilde{f}(y)}\right] \tilde{f}(y)dy = \tilde{\mathbb{E}}\left[\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right],$$

where $\tilde{\mathbb{E}}$ is the expectation under the new density $\tilde{f}$.

Importance Sampling (IS)

Suppose $X = h(Y)$ where Y is a random vector with density f.
Instead of generating Y from f, we can generate it from another density $\tilde{f}$, such that $\tilde{f}(y) > 0$ whenever $h(y)f(y) \neq 0$. We have

$$\mathbb{E}[X] = \int h(y)f(y)dy = \int \left[\frac{h(y)f(y)}{\tilde{f}(y)}\right] \tilde{f}(y)dy = \tilde{\mathbb{E}}\left[\frac{h(Y)f(Y)}{\tilde{f}(Y)}\right],$$

where $\tilde{\mathbb{E}}$ is the expectation under the new density $\tilde{f}$.

To estimate $\mu_0 = \mathbb{E}[X]$ with IS, we generate $Y_1, \ldots, Y_n$ i.i.d. from density $\tilde{f}$ and compute

$$\bar{X}_{\mathrm{is},n} = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\frac{f(Y_i)}{\tilde{f}(Y_i)}.$$

We want to select $\tilde{f}$ so that $\mathrm{Var}[h(Y)f(Y)/\tilde{f}(Y)]$ is small under $\tilde{f}$.

Theorem: If $h \geq 0$, then taking $\tilde{f}(y)$ proportional to $h(y)f(y)$ gives zero variance:

$$X_{\text{is}} = h(Y)\frac{f(Y)}{\tilde{f}(Y)}$$

is a constant, which must equal μ_0.

Hard to implement, but can sometimes be approximated.

Theorem: If $h \geq 0$, then taking $\tilde{f}(y)$ proportional to $h(y)f(y)$ gives zero variance:

$$X_{\text{is}} = h(Y)\frac{f(Y)}{\tilde{f}(Y)}$$

is a constant, which must equal μ_0.

Hard to implement, but can sometimes be approximated.

Also works if Y has a discrete distribution: $\int$ is replaced by $\sum$.

Also, Y can be a stochastic process or another type of random object.

VRCM implies convergence to a zero-variance sampling measure

Suppose

$$\mu_0(\varepsilon) = \mathbb{E}_{\mathbb{P}_\varepsilon}[X(\varepsilon)] = \int_\Omega X(\varepsilon, \omega) d\mathbb{P}_\epsilon(\omega) \to 0$$

when $\varepsilon \to 0$. The zero-variance measure $\mathbb{P}^*_\varepsilon$ here satisfies

$$\frac{d\mathbb{P}^*_\varepsilon(\omega)}{d\mathbb{P}_\varepsilon(\omega)} = \frac{X(\varepsilon, \omega)}{\mu_0(\varepsilon)}.$$

Theorem (L., Blanchet, Glynn, Tuffin 2010).
If $X(\varepsilon)$ has VRCM-$(1+\delta)$ for some $\delta > 0$, then

$$\lim_{\varepsilon \to 0} \sup_{A \in \mathcal{F}} |\mathbb{P}_\varepsilon(A) - P^*_\varepsilon(A)| = 0.$$

That is, the sampling distribution must converge in total variation to the zero-variance measure associated with $X(\varepsilon)$, regardless of what sampling strategy we use (IS or not).

Proof (uses Jensen's inequality):

$$
\begin{aligned}
\sup_{A\in\mathcal{F}} |\mathbb{P}^*_\varepsilon(A) - \mathbb{P}_\varepsilon(A)| &\leq \sup_{A\in\mathcal{F}} |\mathbb{E}_{\mathbb{P}_\varepsilon}\left[(d\mathbb{P}^*_\varepsilon/d\mathbb{P}_\varepsilon)\,\mathbb{I}(A)\right] - \mathbb{E}_{\mathbb{P}_\varepsilon}[\mathbb{I}(A)]| \\
&\leq \mathbb{E}_{\mathbb{P}_\varepsilon}\left|d\mathbb{P}^*_\varepsilon/d\mathbb{P}_\varepsilon - 1\right| \\
&\leq \mathbb{E}^{1/(1+\delta)}_{\mathbb{P}_\varepsilon}\left[\left|d\mathbb{P}^*_\varepsilon/d\mathbb{P}_\varepsilon - 1\right|^{(1+\delta)}\right] \\
&\leq \mathbb{E}^{1/(1+\delta)}_{\mathbb{P}_\varepsilon}\left[\left|X(\varepsilon)/\mu_0(\varepsilon) - 1\right|^{(1+\delta)}\right] \\
&= [c_{1+\delta}(\varepsilon)]^{1/(1+\delta)} \\
&\overset{\varepsilon\to 0}{\longrightarrow} 0.
\end{aligned}
$$

A discrete-time Markov chain (DTMC) framework

A Markov chain $\{Y_j, j \geq 0\}$ with (large) state space $\mathcal{Y}$, and a set of absorbing states $\Delta \subset \mathcal{Y}$.

Stopping time: $\tau = \inf\{j : Y_j \in \Delta\}$.

Transition kernel: $P(C \mid y) = \mathbb{P}[Y_j \in C \mid Y_{j-1} = y)$.

One-step cost $c(y, y')$ for each transition $y \to y'$.

Total cost: $X = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j)$.

Expected cost-to-go from state y: $\mu(y) = \mathbb{E}[X \mid Y_0 = y]$.

We assume that $\mathbb{E}[\tau \mid Y_0 = y] < \infty$ and $\mu(y) < \infty$ for all $y \in \mathcal{Y}$.

Want to estimate $\mu_0 = \mu(y_0)$ for some initial state y_0.

This covers a wide range of situations, including a finite time horizon.

Recurrence equation for μ

The function $\mu : \mathcal{Y} \to \mathbb{R}$ satisfies the recurrence (Poisson) equation

$$\mu(y) = \mathbb{E}[c(y, Y_1) + \mu(Y_1) \mid Y_0 = y] = \int_{\mathcal{Y}} [c(y, z) + \mu(z)] \, \mathrm{d}P[dz \mid y]$$

for $y \notin \Delta$, and $\mu(y) = 0$ for $y \in \Delta$.

Recurrence equation for μ

The function $\mu : \mathcal{Y} \to \mathbb{R}$ satisfies the recurrence (Poisson) equation

$$\mu(y) = \mathbb{E}[c(y, Y_1) + \mu(Y_1) \mid Y_0 = y] = \int_{\mathcal{Y}} [c(y, z) + \mu(z)] \, \mathrm{d}P[dz \mid y]$$

for $y \notin \Delta$, and $\mu(y) = 0$ for $y \in \Delta$.

If $\mathcal{Y}$ is finite, this becomes a linear system of equations. But can be huge!

If $\mathcal{Y}$ is continuous, one may approximate μ by a linear combination of basis functions, or more generally by tuning the parameters of a parameterized function, $v(y; \theta)$.

These techniques are used in machine learning and approximate dynamic programming.

Limitation: the error can be large and difficult to estimate, or the approximation can be too costly to compute. Then, we may use simulation.

Dynamic IS and its interpretation as a Markov decision process (MDP)

At each step of the Markov chain, we can change the transition kernel P for another kernel $\tilde{P}$. That is, $\tilde{P}(C \mid y) = \tilde{\mathbb{P}}[Y_j \in C \mid Y_{j-1} = y)$.

Want to select the transition kernels dynamically in a way that minimizes the variance. The decision (selection) at each step may depend on past and current history.

An optimal (selection) policy gives zero variance. It takes

$$\mathrm{d}\tilde{P}(y_1 \mid y) \quad \text{proportional to} \quad \mathrm{d}P(y_1 \mid y)[c(y, y_1) + \mu(y_1)],$$

with proportionality constant $1/\mu(y)$.

We can approximate it by using an approximation v of μ.
Often, a crude approximation of μ can be computed cheaply.

IS for a discrete-time Markov chain

We change P to $\tilde{P}$ such that $\tilde{\mathbb{E}}[\tau] < \infty$ and $\tilde{P}(C \mid y) > 0$ whenever $\int_C [c(y, y_1) + \mu(y_1)] \mathrm{d}P(y_1 \mid y) > 0$.

The estimator X is replaced by

$$X_{\mathrm{is}} = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j) \prod_{i=1}^{j} (\mathrm{d}P/\mathrm{d}\tilde{P})(Y_i \mid Y_{i-1}).$$

IS for a discrete-time Markov chain

We change P to $\tilde{P}$ such that $\tilde{\mathbb{E}}[\tau] < \infty$ and $\tilde{P}(C \mid y) > 0$ whenever $\int_C [c(y, y_1) + \mu(y_1)] \mathrm{d}P(y_1 \mid y) > 0$.

The estimator X is replaced by

$$X_{\mathrm{is}} = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j) \prod_{i=1}^{j} (\mathrm{d}P/\mathrm{d}\tilde{P})(Y_i \mid Y_{i-1}).$$

Theorem. If we choose $\tilde{P}$ so that

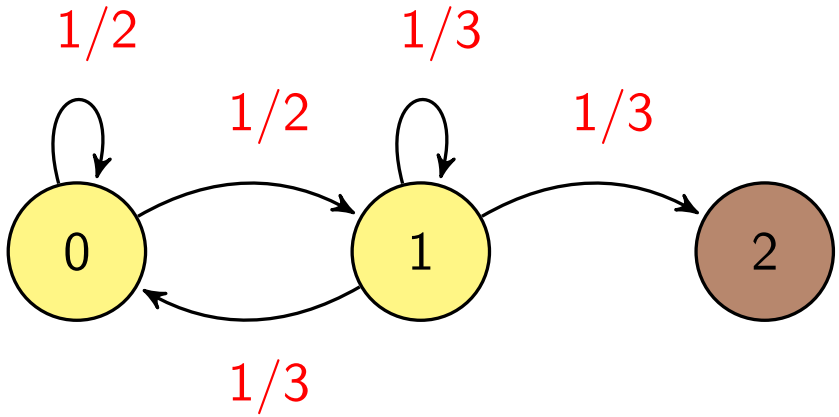
$$(\mathrm{d}\tilde{P}/\mathrm{d}P)(y_1 \mid y) = \begin{cases} \dfrac{c(y, y_1) + \mu(y_1)}{\mu(y)} & \text{if } \mu(y) > 0, \\ 1 & \text{if } \mu(y) = 0, \end{cases}$$

then X_{is} has zero variance.

Proof: E.g., by backward induction on j.

An example where zero-variance gives $\tilde{\mathbb{E}}[\tau] = \infty$

Is zero variance always the perfect thing?

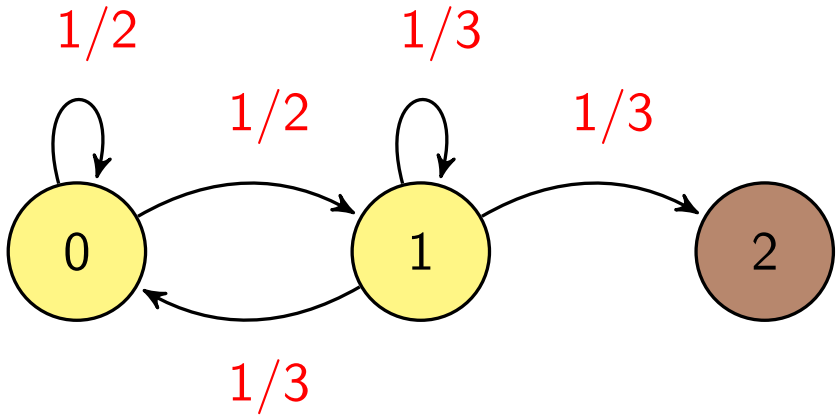


Cost $c(y) = 1$ for $y = 1, 2$, and $c(2) = 0$.
Here, $\mu(y)$ is the expected number of transitions before reaching state 2, given that we are in state y. We have $\mu(2) = 0$.

Zero-variance IS gives $\tilde{p}(2|y) = 0$ for $y = 0, 1$, so the chain will never reach the stopping time τ under these new probabilities!
We have zero variance but infinite computing cost.

An example where zero-variance gives $\tilde{\mathbb{E}}[\tau] = \infty$

Is zero variance always the perfect thing?

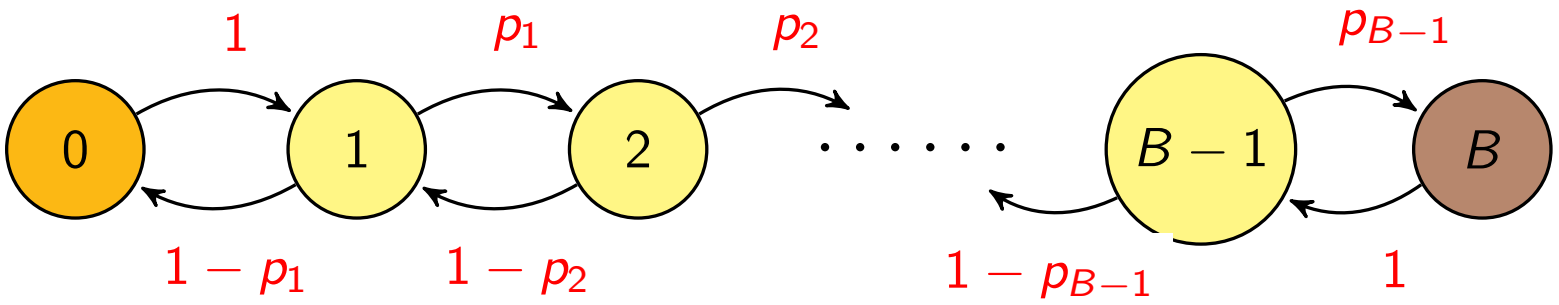


Cost $c(y) = 1$ for $y = 1, 2$, and $c(2) = 0$.
Here, $\mu(y)$ is the expected number of transitions before reaching state 2, given that we are in state y. We have $\mu(2) = 0$.

Zero-variance IS gives $\tilde{p}(2|y) = 0$ for $y = 0, 1$, so the chain will never reach the stopping time τ under these new probabilities!
We have zero variance but infinite computing cost.

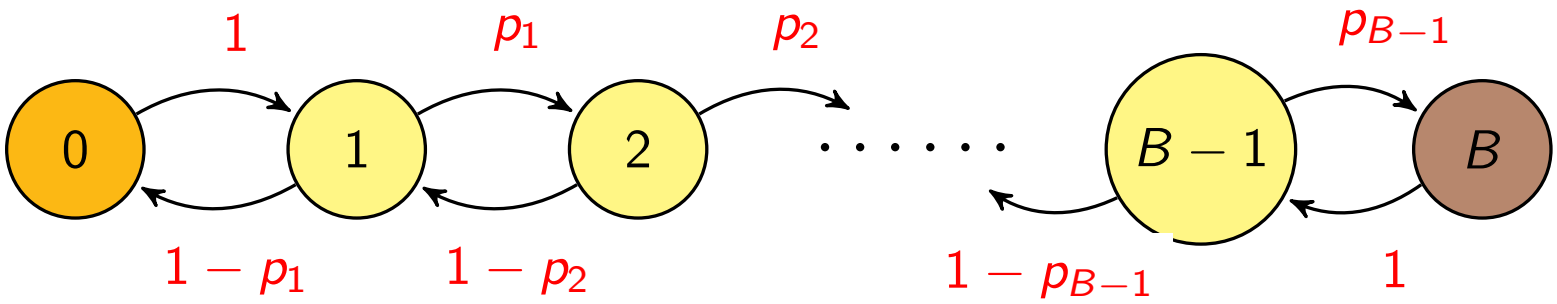
Trick to resolve this: add a cost $\delta > 0$ to any transition that enters Δ.
Afterwards, subtract δ to the final (zero-variance) estimator.

Example 2: A birth-and-death process



Let $\tau = \inf\{j > 0 : Y_j \in \{0, B\}\}$ and define $\mu(y) = \mathbb{P}[Y_\tau = B \mid Y_0 = y]$. We want to estimate $\mu_0 = \mu(1)$, the probability of reaching B before coming back to 0.

Example 2: A birth-and-death process



Let $\tau = \inf\{j > 0 : Y_j \in \{0, B\}\}$ and define $\mu(y) = \mathbb{P}[Y_\tau = B \mid Y_0 = y]$. We want to estimate $\mu_0 = \mu(1)$, the probability of reaching B before coming back to 0. We have the recurrence:

$$\mu(y) = p_y \mu(y+1) + (1 - p_y)\mu(y-1)$$

for $y = 1, \ldots, B-1$, with $\mu(0) = 0$ and $\mu(B) = 1$. Zero-variance change of measure gives

$$\tilde{p}_y = p_y \mu(y+1)/\mu(y) \qquad \text{for } y \geq 1.$$

Because $\mu(0) = 0$, we also see that $\tilde{p}_1 = 1$ and that no sample path will ever return to 0 under zero-variance IS.

Example 2 with $p_y = p$ for $1 \leq y \leq B-1$

For $\rho = p/(1-p) \neq 1/2$, it is known that $\mu(y) = (1-\rho^{-y})/(1-\rho^{-B})$, so

$$\tilde{p}_y = \frac{1-\rho^{-y-1}}{1-\rho^{-y}}\, p = \frac{(1-\rho^{y+1})}{(1-\rho^{y})}\, \frac{1}{\rho}\, p.$$

Those probabilities do not depend on B, but depend on y.
The cycles do not contribute to the likelihood ratio:

$$\frac{p_{y-1}}{\tilde{p}_{y-1}} \frac{(1-p_y)}{(1-\tilde{p}_y)} = 1.$$

Example 2 with $p_y = p$ for $1 \leq y \leq B-1$

For $\rho = p/(1-p) \neq 1/2$, it is known that $\mu(y) = (1-\rho^{-y})/(1-\rho^{-B})$, so

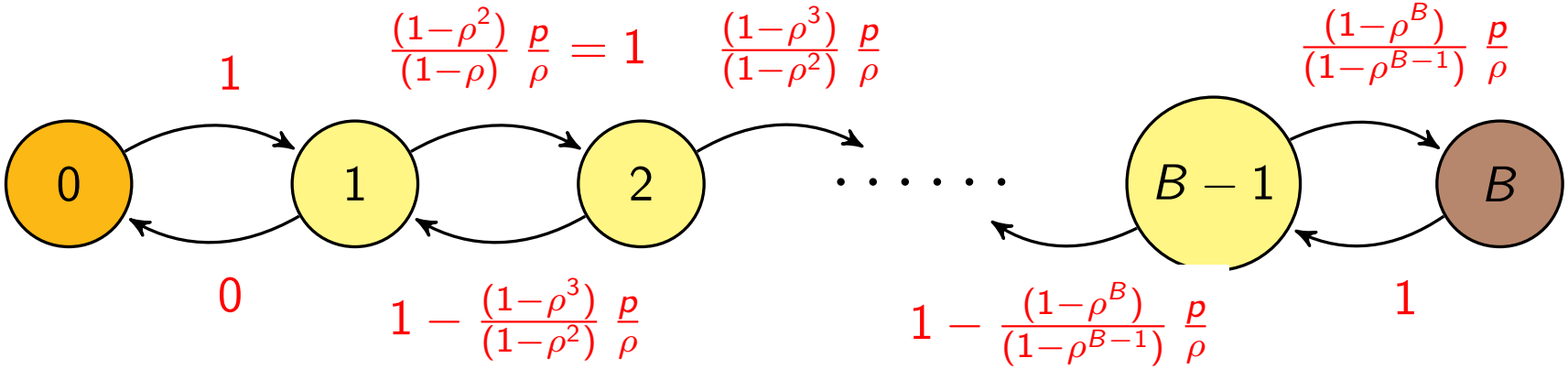
$$\tilde{p}_y = \frac{1-\rho^{-y-1}}{1-\rho^{-y}}\, p = \frac{(1-\rho^{y+1})}{(1-\rho^{y})}\,\frac{1}{\rho}\, p.$$

Those probabilities do not depend on B, but depend on y.
The cycles do not contribute to the likelihood ratio:

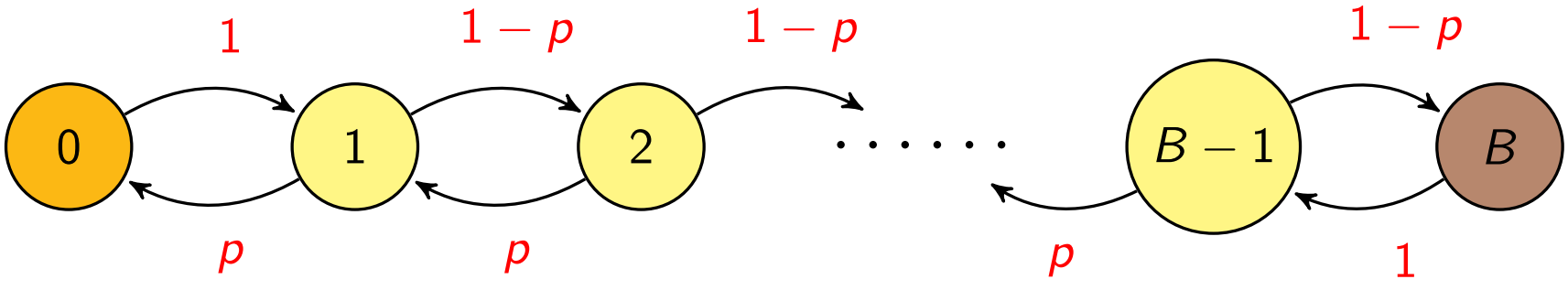
$$\frac{p_{y-1}}{\tilde{p}_{y-1}}\frac{(1-p_y)}{(1-\tilde{p}_y)} = 1.$$

For large B, $\mu(y) = (\rho^{B-y} - \rho^B)/(1-\rho^B) \approx \rho^{B-y}$.
This approximation leads to VRCM-k if $p \to 0$ and fixed B,
because then $\rho^{B-y}/\mu(y) \to 1$.
If $B \to \infty$ for fixed $p < 1/2$, it gives (only) BRM-k.

Case where $p_y = p$, $\rho = p/(1-p)$.



Approximation with $\mu(y) \approx \rho^{B-y}$: note that $p/\rho = 1-p$ and $1 - p/\rho = p$.



First-Passage Probability in a Markov Chain

$A, B \subset \mathcal{Y}$, $A \cap B = \emptyset$, $\Delta = A \cup B$.

$\mu(y) = \mathbb{P}[\text{hitting } B \text{ before } A]$. Want to estimate $\mu(y_0)$.

We have $\mu(y) = 1$ for $y \in B$ and $\mu(y) = 0$ for $y \in A$.

Here, P, A, and B may depend on ε.

First-Passage Probability in a Markov Chain

$A, B \subset \mathcal{Y}$, $A \cap B = \emptyset$, $\Delta = A \cup B$.

$\mu(y) = \mathbb{P}[\text{hitting } B \text{ before } A]$. Want to estimate $\mu(y_0)$.

We have $\mu(y) = 1$ for $y \in B$ and $\mu(y) = 0$ for $y \in A$.

Here, P, A, and B may depend on ε.

Zero variance IS:

$$\tilde{P}\left(\mathrm{d}y_1 \mid y\right) = P\left(\mathrm{d}y_1 \mid y\right) \frac{\mu(y_1)}{\mu(y)}.$$

Approximation:

$$P_v\left(\mathrm{d}y_1 \mid y\right) = P\left(\mathrm{d}y_1 \mid y\right) \frac{v\left(y_1\right)}{w\left(y\right)},$$

where $v : \mathcal{Y} \to [0, \infty)$ is a good approximation of $\mu(\cdot)$ and

$$w\left(y\right) = \int_{y_1 \in \mathcal{Y}} P\left(\mathrm{d}y_1 \mid y\right) v\left(y_1\right).$$

IS estimator of $\mu(y_0)$:

$$X = X(\varepsilon) = \mathbb{I}[\text{hit } B \text{ before } A]\, L,$$

where

$$L = \prod_{k=1}^{\tau} \frac{w(Y_{k-1})}{v(Y_k)} = \frac{w(Y_0)}{v(Y_\tau)} \prod_{k=1}^{\tau-1} \frac{w(Y_k)}{v(Y_k)}.$$

Can take $v(y) = 1$ for $y \in B$ and $v(y) = 0$ for $y \in A$.

IS estimator of $\mu(y_0)$:

$$X = X(\varepsilon) = \mathbb{I}[\text{hit } B \text{ before } A]\, L,$$

where

$$L = \prod_{k=1}^{\tau} \frac{w(Y_{k-1})}{v(Y_k)} = \frac{w(Y_0)}{v(Y_\tau)} \prod_{k=1}^{\tau-1} \frac{w(Y_k)}{v(Y_k)}.$$

Can take $v(y) = 1$ for $y \in B$ and $v(y) = 0$ for $y \in A$.

To establish robustness properties such as LE-k, BRM-k, and VRCM-k, we need an asymptotic bound on $\mathbb{E}[X^k(\varepsilon)]/\mu^k(y_0, \varepsilon)$.

Proposition: Bounds via Lyapunov inequalities.
Suppose there are positive constants κ_1 and κ_2 and a function $h_k : \mathcal{Y} \to [0, \infty)$ such that $v(y) \geq \kappa_1$ and $h_k(y) \geq \kappa_2$ for each $y \in B$, and

$$\mathbb{E}\left[\left(\frac{w(y)}{v(y)}\right)^k h_k(Y_1) \mid Y_0 = y\right] \leq h_k(y)$$

for all $y \notin \Delta$. Then, for all $y \notin \Delta$,

$$\mathbb{E}[X^k \mid Y_0 = y] \leq \frac{v^k(y) h_k(y)}{\kappa_1^k \kappa_2},$$

and therefore

$$\frac{\mathbb{E}[X^k]}{\mu^k(y_0)} \leq \frac{[v(y_0)/\mu(y_0)]^k h_k(y_0)}{\kappa_1^k \kappa_2}.$$

Corollary.
Under the proposition's conditions:

(i) If

$$\lim_{\varepsilon \to 0} \frac{\ln v(y_0, \varepsilon) + k^{-1} \ln h_k(y_0, \varepsilon)}{\ln \mu(y_0, \varepsilon)} = 1,$$

then $X(\varepsilon)$ is LE-k.

(ii) If

$$\lim_{\varepsilon \to 0} [v(y_0, \varepsilon)/\mu(y_0, \varepsilon)]^k h_k(y_0, \varepsilon) < \infty,$$

then $X(\varepsilon)$ is BRM-k.

(iii) If

$$\lim_{\varepsilon \to 0} \frac{[v(y_0, \varepsilon)/\mu(y_0, \varepsilon)]^k h_k(y_0, \varepsilon)}{\kappa_1^k \kappa_2} = 1,$$

then $X(\varepsilon)$ is VRCM-k.

Example: a random walk on the real line

Let $Y_j = Y_0 + D_1 + \cdots + D_j$, the D_j's are i.i.d., $\mathbb{E}[D_j] < 0$, $B = [0, \infty)$,

$$\mu(y) = \mathbb{P}\left[\sup_{j \geq 0} Y_j \geq 0 \mid Y_0 = y \leq 0\right],$$

and $\mu_0 = \mu(-1/\varepsilon)$. Several applications. May represent the ruin probability for an insurance company with initial reserve $-1/\varepsilon$, or the probability that a steady-state delay in a single-server queue exceeds $-1/\varepsilon$.

Example: a random walk on the real line

Let $Y_j = Y_0 + D_1 + \cdots + D_j$, the D_j's are i.i.d., $\mathbb{E}[D_j] < 0$, $B = [0, \infty)$,

$$\mu(y) = \mathbb{P}\left[\sup_{j \geq 0} Y_j \geq 0 \mid Y_0 = y \leq 0\right],$$

and $\mu_0 = \mu(-1/\varepsilon)$. Several applications. May represent the ruin probability for an insurance company with initial reserve $-1/\varepsilon$, or the probability that a steady-state delay in a single-server queue exceeds $-1/\varepsilon$.

If D_j has a light-tailed distribution, $\theta^* > 0$, $\mathbb{E}[\exp(\theta^* D_j)] = 1$ and $\mathbb{E}[D_j \exp(\theta^* D_j)] < \infty$, then taking $v(y) = \exp(\theta^* y)$ gives BRE.

Example: a random walk on the real line

Let $Y_j = Y_0 + D_1 + \cdots + D_j$, the D_j's are i.i.d., $\mathbb{E}[D_j] < 0$, $B = [0, \infty)$,

$$\mu(y) = \mathbb{P}\left[\sup_{j \geq 0} Y_j \geq 0 \mid Y_0 = y \leq 0\right],$$

and $\mu_0 = \mu(-1/\varepsilon)$. Several applications. May represent the ruin probability for an insurance company with initial reserve $-1/\varepsilon$, or the probability that a steady-state delay in a single-server queue exceeds $-1/\varepsilon$.

If D_j has a light-tailed distribution, $\theta^* > 0$, $\mathbb{E}[\exp(\theta^* D_j)] = 1$ and $\mathbb{E}[D_j \exp(\theta^* D_j)] < \infty$, then taking $v(y) = \exp(\theta^* y)$ gives BRE.

If D_j has a regularly varying tail: for each $b > 0$,

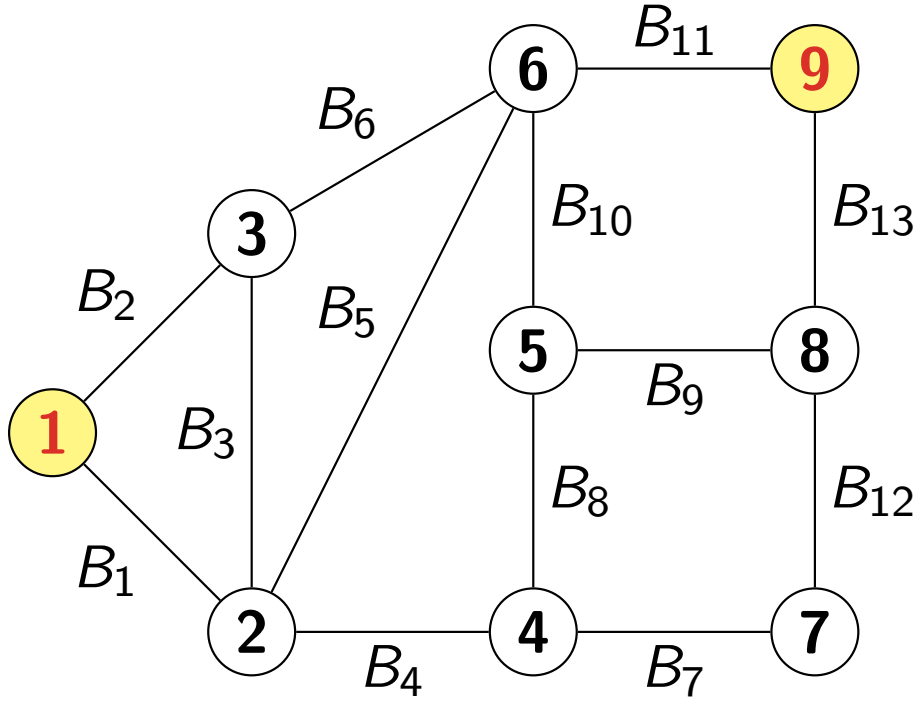
$$\lim_{t \to \infty} \mathbb{P}\left(D_j > bt\right) / \mathbb{P}\left(D_j > t\right) = b^{-\alpha}$$

for some $\alpha > 1$, then taking

$$v(y) = \min\left(1, \frac{-1}{\mathbb{E}[D_j]} \int_{a_k^* - y}^{\infty} \mathbb{P}[D_j > s] \mathrm{d}s\right).$$

gives BRM-k and even VRCM-k, for some $a_k^* > 0$ that may depend on k.

Static Network Reliability Problem



Suppose the B_j's are independent and $\mathbb{P}[B_j = 1] = q_j$.
We generate $B_1, B_2, \ldots, B_{13}$ in this order (could renumber before).

Can be seen as a Markov chain with state $Y_j = (B_1, \ldots, B_j)$ at step j.

$$\begin{aligned} \mu_j(b_1, \ldots, b_{j-1}) &= \mathbb{E}[\phi(\mathbf{B}) \mid B_1 = b_1, \ldots, B_{j-1} = b_{j-1}] \\ &= q_j \; \mu_{j+1}(b_1, \ldots, b_{j-1}, 1) + (1 - q_j) \; \mu_{j+1}(b_1, \ldots, b_{j-1}, 0). \end{aligned}$$

Zero-variance scheme:

$$\tilde{q}_j = q_j \, \frac{\mu_{j+1}(b_1, \ldots, b_{j-1}, 1)}{\mu_j(b_1, \ldots, b_{j-1})}.$$

The (unknown) μ_j can be replaced by easily-computable approximations.

Can be seen as a Markov chain with state $Y_j = (B_1, \ldots, B_j)$ at step j.

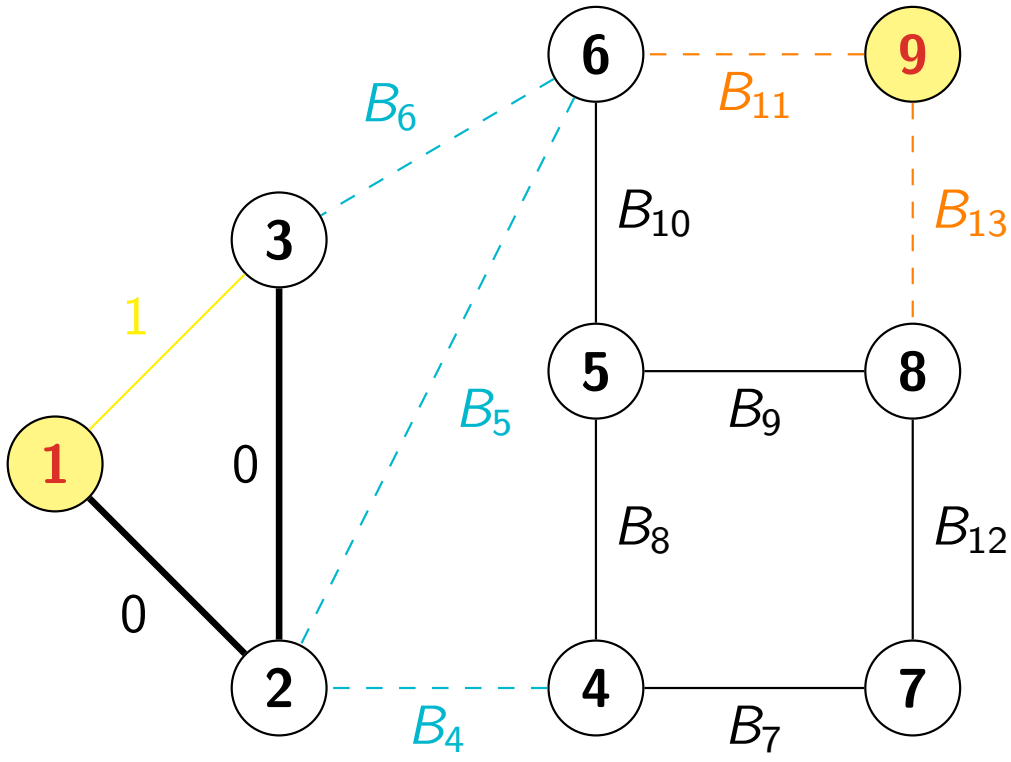
$$\begin{aligned} \mu_j(b_1, \ldots, b_{j-1}) &= \mathbb{E}[\phi(\mathbf{B}) \mid B_1 = b_1, \ldots, B_{j-1} = b_{j-1}] \\ &= q_j \, \mu_{j+1}(b_1, \ldots, b_{j-1}, 1) + (1 - q_j) \, \mu_{j+1}(b_1, \ldots, b_{j-1}, 0). \end{aligned}$$

Zero-variance scheme:

$$\tilde{q}_j = q_j \, \frac{\mu_{j+1}(b_1, \ldots, b_{j-1}, 1)}{\mu_j(b_1, \ldots, b_{j-1})}.$$

The (unknown) μ_j can be replaced by easily-computable approximations.

Mincut-maxprob approx. (L., Rubino, Saggadi, Tuffin 2010).
Given $b_1, \ldots, b_{j-1}$ fixed, take a set of disjoint minimal cuts made with the other edges, that disconnect $\mathcal{S}$ and have maximal probability.
Approximate μ_j by the sum of their probabilities.
The probability of a cut is the product of its q_j's.

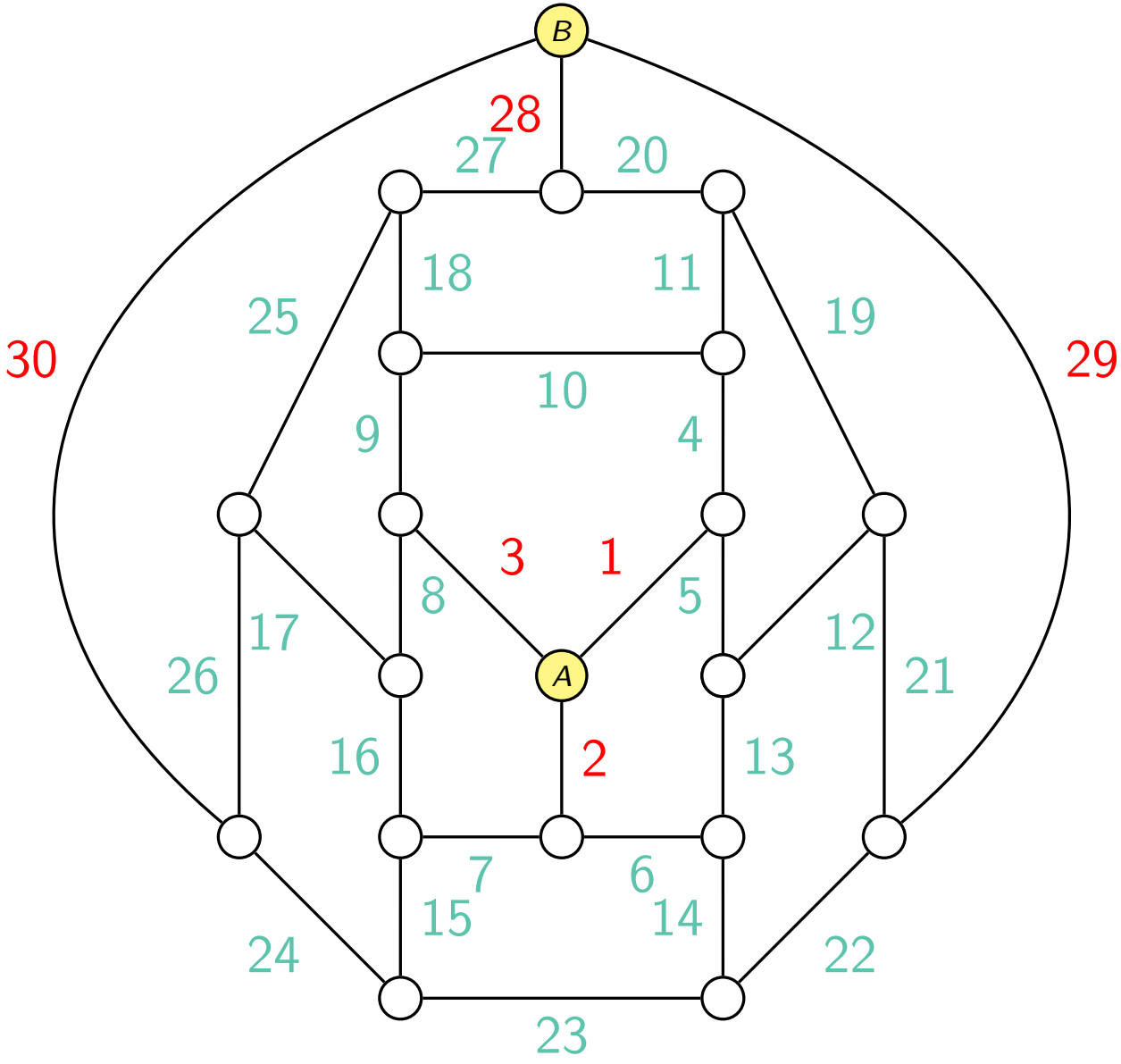


Two minimal cuts, in blue and orange.

Suppose $q_j = q_j(\varepsilon) = a_j \varepsilon^{b_j} + o(\varepsilon^{b_j})$.

Theorem: The mincut-maxprob approximation always gives BRE.
Under mild additional conditions, it also gives VRE.

A dodecahedron network



Results for dodecahedron network, with all $q_j = \epsilon$, for $n = 10^4$.

ϵ	estimate	standard dev.	relative error
10^{-1}	2.8960×10^{-3}	3.49×10^{-3}	1.2
10^{-2}	2.0678×10^{-6}	3.42×10^{-7}	0.17
10^{-3}	2.0076×10^{-9}	1.14×10^{-10}	0.057
10^{-4}	2.0007×10^{-12}	3.46×10^{-14}	0.017

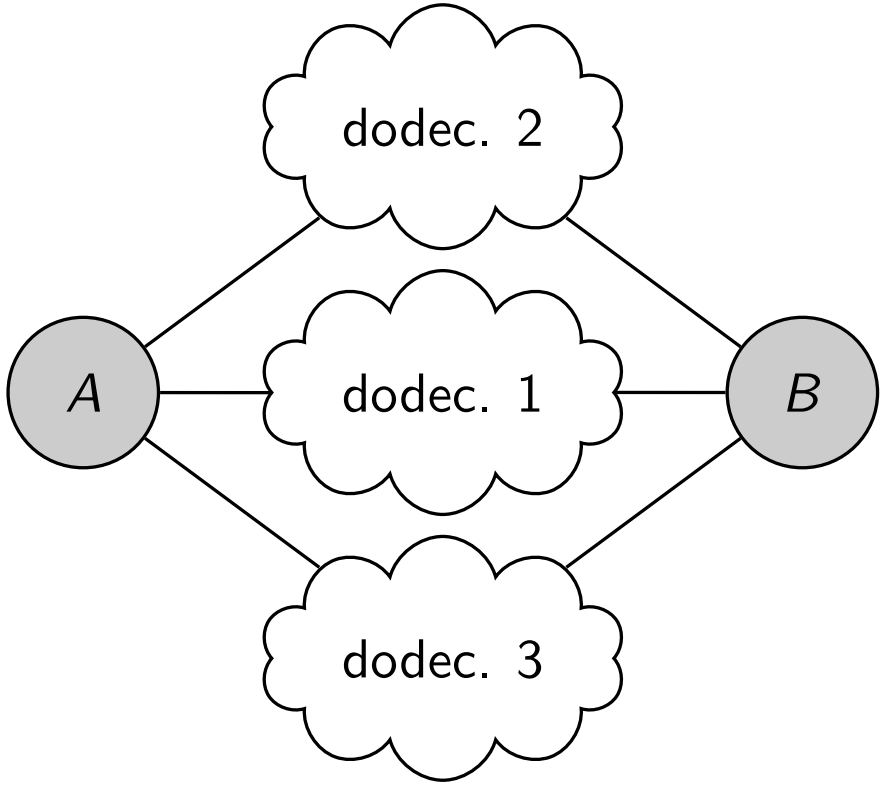
Results for dodecahedron network, with all $q_j = \epsilon$, for $n = 10^4$.

ϵ	estimate	standard dev.	relative error
10^{-1}	2.8960×10^{-3}	3.49×10^{-3}	1.2
10^{-2}	2.0678×10^{-6}	3.42×10^{-7}	0.17
10^{-3}	2.0076×10^{-9}	1.14×10^{-10}	0.057
10^{-4}	2.0007×10^{-12}	3.46×10^{-14}	0.017

Can combine the method with series-parallel reductions of the graph at each step (WSC 2011 paper).

Similar (dual) method based on disjoint paths to estimate μ_j.
Can combine the two estimates (based on cuts and on paths).

Three dodecahedrons in parallel. $q_j = \epsilon$ and for $n = 10^4$.



ϵ	estimate	standard dev.	relative error
0.10	2.3573×10^{-8}	5.49×10^{-8}	2.3
0.05	2.5732×10^{-11}	3.03×10^{-11}	1.2
0.01	8.7655×10^{-18}	2.60×10^{-18}	0.30

Dynamic Highly Reliable Markovian System

Similar to static network, but each component has a failure rate and a repair rate. Evolution is modeled by a continuous-time Markov chain.

State space partitioned into up states and down states.

Interesting reliability measures can be computed if we know the probability μ_0 that a chain starting in "all up" state reaches a down state before returning to all up state.

Goal: estimate this (very small) μ_0 by IS with zero-variance approx.

ZVA method based on paths (+ heuristic adjustments) proposed by L'Ecuyer and Tuffin (2010) for this problem.

Zero-Variance via control variates (CV)

Same DTMC model. Still want to estimate $\mu_0 = \mu(y_0) = \mathbb{E}[X]$ where $X = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j)$. We can write

$$\mu(y_0) = X - M_\tau$$

where

$$\begin{aligned} M_\tau &= \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + \mu(Y_j) - \mu(Y_{j-1})] \\ &= \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + \mu(Y_j) - \mathbb{E}[c(Y_{j-1}, Y_j) + \mu(Y_j) \mid Y_{j-1}]] \,. \end{aligned}$$

Zero-Variance via control variates (CV)

Same DTMC model. Still want to estimate $\mu_0 = \mu(y_0) = \mathbb{E}[X]$ where $X = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j)$. We can write

$$\mu(y_0) = X - M_\tau$$

where

$$\begin{aligned} M_\tau &= \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + \mu(Y_j) - \mu(Y_{j-1})] \\ &= \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + \mu(Y_j) - \mathbb{E}[c(Y_{j-1}, Y_j) + \mu(Y_j) \mid Y_{j-1}]] \,. \end{aligned}$$

So if we could compute and subtract M_τ (as a CV) we would have zero-variance. But of course, μ is unknown.

Approximate zero variance via control variates

Replace μ in M_τ by an approximation v such that $v(y) = 0$ for $y \in \Delta$:

$$M_\tau = \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + v(Y_j) - \mathbb{E}[c(Y_{j-1}, Y_j) + v(Y_j) \mid Y_{j-1}]],$$

and define the CV estimator $X_{\mathrm{cv}} = X - M_\tau$.

We have $\mathbb{E}[M_\tau] = 0$, and thus $\mathbb{E}[X_{\mathrm{cv}}] = \mathbb{E}[X]$ (unbiased) regardless of v.

Variance can be reduced significantly if v is a good approximation of μ.

Approximate zero variance via control variates

Replace μ in M_τ by an approximation v such that $v(y) = 0$ for $y \in \Delta$:

$$M_\tau = \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + v(Y_j) - \mathbb{E}[c(Y_{j-1}, Y_j) + v(Y_j) \mid Y_{j-1}]],$$

and define the CV estimator $X_{\mathrm{cv}} = X - M_\tau$.

We have $\mathbb{E}[M_\tau] = 0$, and thus $\mathbb{E}[X_{\mathrm{cv}}] = \mathbb{E}[X]$ (unbiased) regardless of v.

Variance can be reduced significantly if v is a good approximation of μ.

However, this is not the right tool for rare-event simulation, because it does not make the rare events more frequent.

Approximate zero variance via control variates

Replace μ in M_τ by an approximation v such that $v(y) = 0$ for $y \in \Delta$:

$$M_\tau = \sum_{j=1}^{\tau} [c(Y_{j-1}, Y_j) + v(Y_j) - \mathbb{E}[c(Y_{j-1}, Y_j) + v(Y_j) \mid Y_{j-1}]],$$

and define the CV estimator $X_{\text{cv}} = X - M_\tau$.

We have $\mathbb{E}[M_\tau] = 0$, and thus $\mathbb{E}[X_{\text{cv}}] = \mathbb{E}[X]$ (unbiased) regardless of v.

Variance can be reduced significantly if v is a good approximation of μ.

However, this is not the right tool for rare-event simulation, because it does not make the rare events more frequent.

Extensions to regenerative simulation (Kim & Henderson 07), and to infinite-horizon models with discounting and stochastic differential equations (Henderson & Glynn 02).

Multilevel Splitting

Markov chain $\{Y_j, j \geq 0\}$. $\Delta = A \cup B \subset \mathcal{Y}$, where $A \cap B = \emptyset$.

$$\mu(y) = \mathbb{P}[\text{hit } B \text{ before } A \mid Y_0 = y], \quad \text{for } y \in \mathcal{Y}.$$

Want to estimate $\mu_0 = \mu(y_0)$ for some fixed initial state y_0.

Multilevel Splitting

Markov chain $\{Y_j, j \geq 0\}$. $\Delta = A \cup B \subset \mathcal{Y}$, where $A \cap B = \emptyset$.

$$\mu(y) = \mathbb{P}[\text{hit } B \text{ before } A \mid Y_0 = y], \quad \text{for } y \in \mathcal{Y}.$$

Want to estimate $\mu_0 = \mu(y_0)$ for some fixed initial state y_0.

Select an importance function $h : \mathcal{X} \to \mathbb{R}$ such that

$$A = \{x \in \mathcal{X} : h(x) \leq 0\} \text{ and } B = \{x \in \mathcal{X} : h(x) \geq \ell\} \text{ for some } \ell > 0.$$

Partition $[0, \ell)$ in m intervals with boundaries $0 = \ell_0 < \ell_1 < \cdots < \ell_m = \ell$.

For stage $k = 1, \ldots, m$: Clone (split) the R_{k-1} chains that have reached ℓ_{k-1} to get $N_{k-1} \geq R_{k-1}$ chains and simulate them independently until their $h(X_j)$ reaches ℓ_k or 0.

Estimate μ_0 by $\prod_{k=1}^{m} R_k / N_{k-1}$. Unbiased.

Several variants: fixed splitting vs fixed effort, etc.

- Choice of h is the most important (and difficult) issue.
- Not really a "zero-variance approximation" scheme by itself.
- Does not provide BRE or VRE.

Generalized splitting (GS) (Botev and Kroese 2010)

Random vector $\mathbf{Y}$ with known distribution, with density f. Suppose we want to estimate $\mu = \mathbb{P}[\mathbf{Y} \in B]$ (special case).

Zero-variance IS: sample $\mathbf{Y}$ from $\mathbb{P}[\mathbf{Y} \in \cdot \mid \mathbf{Y} \in B]$ (density f_m).

Generalized splitting (GS) (Botev and Kroese 2010)

Random vector $\mathbf{Y}$ with known distribution, with density f. Suppose we want to estimate $\mu = \mathbb{P}[\mathbf{Y} \in B]$ (special case).
Zero-variance IS: sample $\mathbf{Y}$ from $\mathbb{P}[\mathbf{Y} \in \cdot \mid \mathbf{Y} \in B]$ (density f_m).

At each level k, we construct a Markov chain $\{\mathbf{Y}_{k,j}, j \geq 0\}$ with transition density $\kappa_{k-1}(\cdot \mid \cdot)$, and whose stationary density is the density f_{k-1} of $\mathbf{Y}$ conditional on $h(\mathbf{Y}) > \ell_{k-1}$. **GS algorithm:**

Generate a $\mathbf{Y}$ from its unconditional density f.
if $h(\mathbf{Y}) > \ell_1$ **then** $\mathcal{X}_1 \leftarrow \{\mathbf{Y}\}$ **else return** $\mathcal{X}_\tau = \emptyset$ and $M = 0$.
for $k = 2$ **to** m **do**
    $\mathcal{X}_k \leftarrow \emptyset$ // list of states that have reached the level ℓ_k
    for all $\mathbf{Y}_{k,0} \in \mathcal{X}_{k-1}$ **do**
        for $j = 1$ **to** s **do**
            sample $\mathbf{Y}_{k,j}$ from the density $\kappa_{k-1}(\cdot \mid \mathbf{Y}_{k,j-1})$
            if $h(\mathbf{Y}_{k,j}) > \ell_k$ **then** add $\mathbf{Y}_{k,j}$ to $\mathcal{X}_k$
return the list $\mathcal{X}_m$ and its cardinality $M = |\mathcal{X}_m|$.

Some properties of GS

There are s^{m-1} possible copies of the chain that can reach ℓ_m.
For each, if it reaches ℓ_m, the conditional density of its last state is f_m.

Then, $\mathcal{X}_m$ contains a random number M of states having density f_m.
But those states and their number M and not independent.
Picking one of them at random does not give a state with density f_m.

Some properties of GS

There are s^{m-1} possible copies of the chain that can reach ℓ_m.
For each, if it reaches ℓ_m, the conditional density of its last state is f_m.

Then, $\mathcal{X}_m$ contains a random number M of states having density f_m.
But those states and their number M and not independent.
Picking one of them at random does not give a state with density f_m.

However, if we run GS n times independently, and pick a state at random from the union $\mathcal{X}_n^*$ of the n realizations of $\mathcal{X}_m$, the distribution of this state converges to that with density f_m when $n \to \infty$.

Thus, the empirical distribution of the states $\mathcal{X}_n^*$ could be taken as an approximation of the conditional distribution given B.

MCIS: use a one-step look-ahead density. For each state $\mathbf{Y} \in \mathcal{X}_n^*$, consider the conditional density $\kappa_m(\cdot \mid \mathcal{Y})$, and take a mixture of those densities, with equal weights, as an approximate zero-variance IS density.
Talk of Z. Botev later.

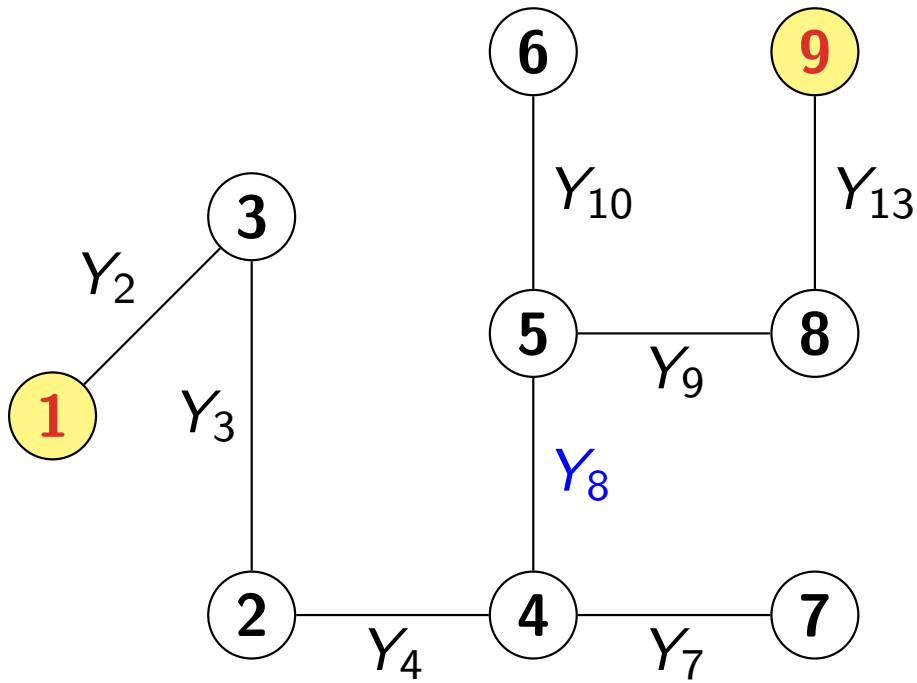
Example: static network with auxiliary variables

Suppose each link j is initially failed and gets repaired at time $Y_j \sim \text{Expon}(\lambda_j)$ where $\lambda_j = -\ln(q_j)$.
Then $\mathbb{P}[Y_j \leq 1] = q_j$. State: $\mathbf{Y} = (Y_1, \ldots, Y_d)$.
Importance function: $h(\mathbf{Y})$ = network repair time.

Example: static network with auxiliary variables

Suppose each link j is initially failed and gets repaired at time $Y_j \sim \text{Expon}(\lambda_j)$ where $\lambda_j = -\ln(q_j)$.
Then $\mathbb{P}[Y_j \leq 1] = q_j$. State: $\mathbf{Y} = (Y_1, \ldots, Y_d)$.
Importance function: $h(\mathbf{Y}) =$ network repair time.

Example: If $Y_3 < Y_2 < Y_{10} < Y_4 < Y_7 < Y_9 < Y_{13} < Y_8 < \cdots$:



We can define κ_{k-1} for GS via Gibbs sampling:

Require: $\mathbf{Y}$ for which $h(\mathbf{Y}) > \ell_{k-1}$ and a permutation π of $\{1, \ldots, d\}$
  for $j = 1$ **to** d **do**
    $i \leftarrow \pi(k)$
    if $S(Y_1, \ldots, Y_{i-1}, 0, Y_{i+1}, \ldots, Y_d) < \ell_{k-1}$ **then**
      // adding link i would connect $\mathcal{K}$
      resample Y_i from its density truncated to (ℓ_{k-1}, ∞)
    else
      resample Y_i from its original density
  return $\mathbf{Y}$ as the resampled vector.

We can define κ_{k-1} for GS via Gibbs sampling:

Require: $\mathbf{Y}$ for which $h(\mathbf{Y}) > \ell_{k-1}$ and a permutation π of $\{1, \ldots, d\}$
  for $j = 1$ **to** d **do**
    $i \leftarrow \pi(k)$
    if $S(Y_1, \ldots, Y_{i-1}, 0, Y_{i+1}, \ldots, Y_d) < \ell_{k-1}$ **then**
      // adding link i would connect $\mathcal{K}$
      resample Y_i from its density truncated to (ℓ_{k-1}, ∞)
    else
      resample Y_i from its original density
  return $\mathbf{Y}$ as the resampled vector.

Appropriate levels ℓ_k can be estimated in a pilot phase.

Here we take $s = 2$.

GS for the dodecahedron: $n = 10^6$, $\mathcal{K} = \{1, 20\}$

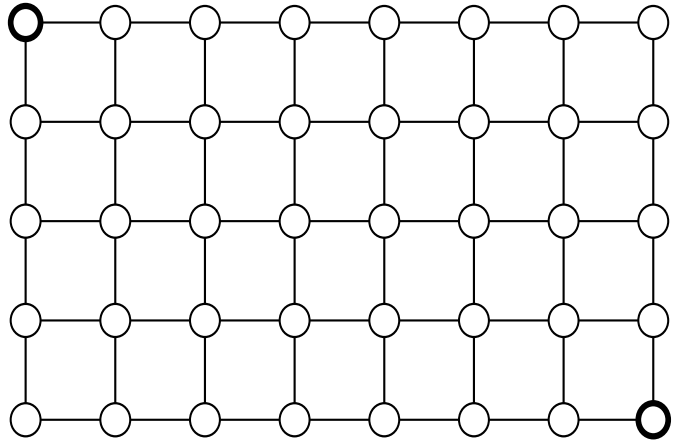
$q_j = \epsilon$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
m	9	19	29	39	49	59
$\bar{W}_n$	0.002877	2.054e-6	2.022e-9	2.01e-12	1.987e-15	1.969e-18
RE$[\bar{W}_n]$	0.0040	0.0062	0.0077	0.0089	0.0099	0.0112
T (sec)	93	167	224	278	334	376

GS for the three dodecahedrons in parallel: $n = 10^6$, $\mathcal{K} = \{1, 20\}$

$q_j = \epsilon$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
m	26	57	87	117	147	176
$\bar{W}_n$	2.377e-8	8.874e-18	8.182e-27	8.088e-36	8.237e-45	7.931e-54
RE$[\bar{W}_n]$	0.0071	0.0109	0.0137	0.0158	0.0185	0.0208
T (sec)	1202	2015	2362	2820	3041	3287

Lomonosov et al. conditional expectation method gives BRE.
GS does not, but works better in practice for large networks.

A lattice graph



GS for a 50×50 lattice graph, with 2500 nodes, 4900 edges, $n = 10^4$.

$q_j = \epsilon$	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
m	13	19	26	33	39
$\bar{W}_n$	2.148e-4	2.085e-6	2.179e-8	2.156e-10	1.932e-12
RE$[\bar{W}_n]$	0.0466	0.0604	0.0678	0.0785	0.0909
T (sec)	19818	19283	18413	17967	17851

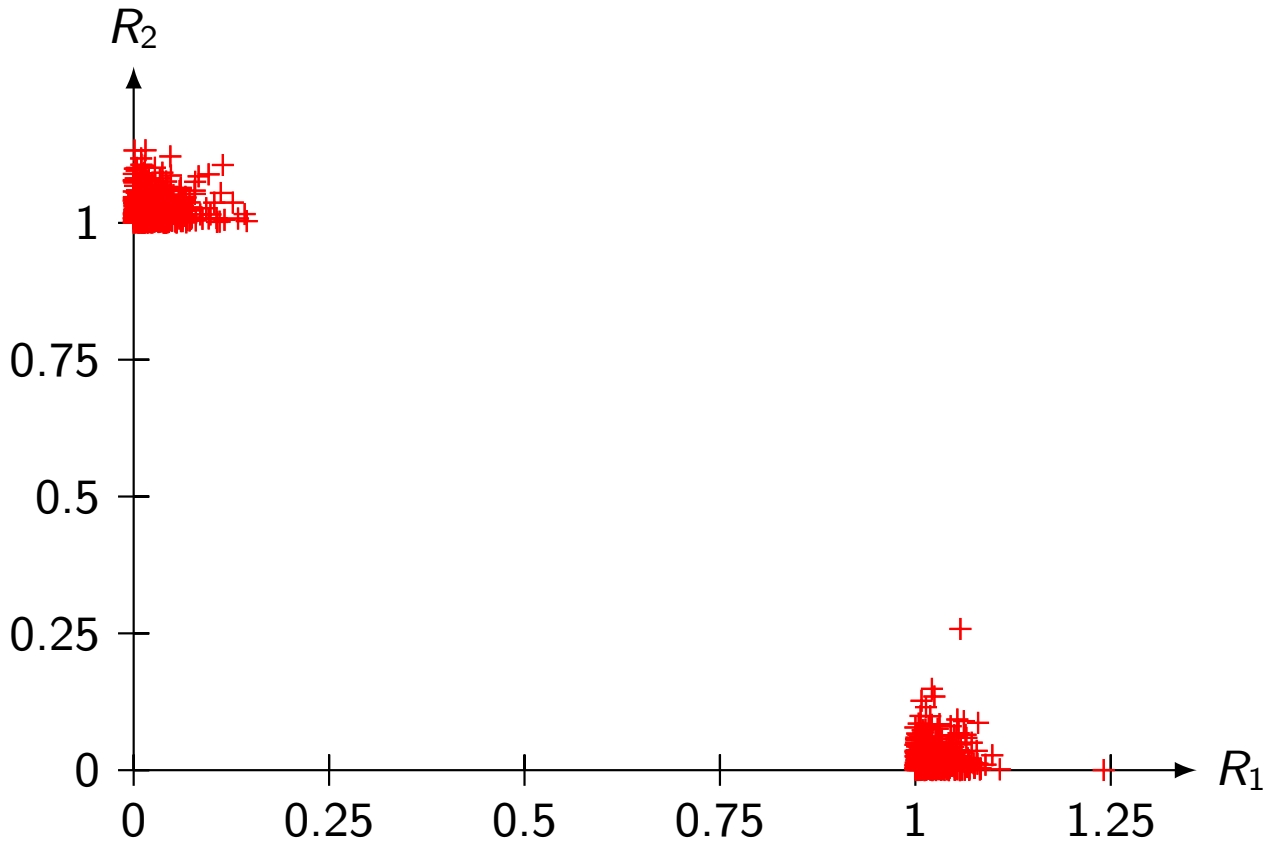
Dodecahedron: distribution of states at last level

The states $\mathbf{Y} \in \mathcal{X}_m$ (at the last level) have density f_m, which is the zero-variance IS density for $\mathbf{Y}$.

Even though their number is random and they are not independent, for large n their empirical distribution is close to this conditional distribution.

For the dodecahedron network with $q_j = 10^{-6}$, under f_m, we find that the mean repair time Y_j is near 0.57 for links $1, 2, 3, 28, 29, 30$, and is near 0.0724 (unchanged) for other links.

Out of the first 4000 runs, 2092 states reach the last level.
Scatter plot of $R_2 = \min(Y_{28}, Y_{29}, Y_{30})$ vs $R_1 = \min(Y_1, Y_2, Y_3)$:



For the dodecahedron network, for ϵ ranging from 10^{-1} to 10^{-12}, MCIS with $n = 100$ and $r = 1000$ replications gave a RE of less than 1%. Smaller variance than GS by a factor of roughly 1000.

For the dodecahedron network, for ϵ ranging from 10^{-1} to 10^{-12}, MCIS with $n = 100$ and $r = 1000$ replications gave a RE of less than 1%. Smaller variance than GS by a factor of roughly 1000.

Dependent links.

The vector $\mathbf{Y}$ can have a multivariate normal or multivariate Student distribution, for example.

In that setting, to resample $\mathbf{Y}$ at each step, we use hit-and-run instead of Gibbs sampling.

Conclusion

- Both IS and CV can achieve zero-variance in theory.
- Zero-variance can only be approximated, usually via a good approximation of the value function μ. Can provide large (unbounded) efficiency improvements in practice.
- Can approximate μ in a parametric class of functions $\mathcal{V} = \{v(\cdot;\theta) : \mathcal{Y} \to \mathbb{R}, \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$, and $\theta = (\theta_1, \ldots, \theta_d)$ is a vector of parameters that we try to optimize so that $v = v(\cdot;\theta)$ is close to μ in some sense. Can use a linear combination of a fixed set of basis functions. Difficulty: choice of those basis functions.
- Important hurdle for IS: approximation must be constructed to allow efficient random variate generation. Sampling under IS must remain simple!
- Splitting methods are also quite useful for rare events, although the required work increases and is unbounded when rarity parameter $\varepsilon \to 0$.