Small-sample Behavior for Importance Sampling
Rare-event Estimators

Peter W. Glynn
Stanford University

Joint work with Jihye Choi

SAMSI Rare Events Workshop,
February, 2012
Rare event Simulation:

**Goal:** Compute $\alpha = P(A)$, where $A$ is "rare"
- e.g. buffer overflow
- large financial loss
- failure of distributed database

**Method:** $\alpha = \mathbb{E}_Q \mathbb{I}(A)L$

where

$$L(\omega) = \left[ \frac{dP}{dQ} \right](\omega)$$

$Q$ is the "importance distribution"

**Remark:** The estimator is zero-variance if we choose

$$Q^*(\cdot) = P(\cdot | A)$$

**Problem:** We often have only a very vague idea of what the conditional distribution looks like
e.g. $P(S_{10} > 20) = ?$

where $S_{10} = Z_1 + \cdots + Z_{10}; Z_i$’s iid

$P(Z_i \in dz) = e^{-z}dz$, so that $\mathbb{E}Z_i = 1$

Note that:

$$\mathbb{E} [Z_i \mid S_{10}] = \frac{S_{10}}{10} \approx 2$$

on $\{S_{10} > 20\}$

Choose $Q$ so that under $Q$,

$$Z_i \sim N(2, 1)$$
Figure: Sample mean of the IS estimator

Poor performance!
Perhaps if we compute confidence intervals, this will “diagnose” the poor performance.
Perhaps if we compute confidence intervals, this will “diagnose” the poor performance.

Figure: Sample mean of the IS estimator with 95\% confidence interval
The sample variance can be misleading when the importance
distribution is poorly chosen

Another possible diagnostic:
effective sample size

\[ n \frac{\mathbb{E}L^2}{\sum_{i=1}^{n} L_i} = \frac{(\sum_{i=1}^{n} L_i)^2}{\sum_{i=1}^{n} L_i^2}, \]

estimated via
Importance sampling is a “high-risk” variance reduction method, in the sense that a poorly chosen importance distribution can lead to disastrous increases in variance

[ unlike control variates, common random numbers, conditional Monte Carlo, etc ]

Building importance samplers that are provably good is hard

And diagnosing an importance sampler that is bad is also hard
In practice:

- Importance samplers will often be heuristically obtained, be “suboptimal”, and will come without provable guarantees.
- The diagnostics may fail precisely when we need them the most.

Question: What happens when one uses a sub-optimal importance distribution?

Focus of the remainder of the talk
To get some theoretical traction, we consider problems that can be naturally embedded in an asymptotic setting:

e.g. Compute $\alpha = P(S_{10} > 20)$

with $S_{10} = Z_1 + \cdots + Z_{10}$, with the $Z_i$’s iid having $EZ_1 = 1$.

Embed in the asymptotic regime:

Compute $\alpha_n = P(S_n > an)$, $(n = 10, a = 2)$

with $S_n = Z_1 + \cdots + Z_n$ (Zi’s iid) with $n \to \infty$, where $a > EZ_1$
Large deviations allows us to easily compute “rough asymptotics” for $\alpha_n$.

When the $Z_i$’s are “light-tailed” (i.e. $\mathbb{E}\exp(\theta Z_i) < \infty$ for $\theta$ in a neighborhood of 0),

$$\frac{1}{n} \log P(S_n > an) \to -\inf\{I(x) : x \in [a, \infty)\}$$

where $I(x) = \sup\{\theta x - \Lambda(\theta) : \theta \in \mathbb{R}\}$ and $\Lambda(\theta) = \log \mathbb{E}\exp(\theta Z_i)$.

$I(\cdot)$ is called the rate function.
Furthermore,

\[ P(Z_1 \in dz_1, \ldots, Z_k \in dz_k \mid S_n > an) \Rightarrow \prod_{i=1}^{k} Q^*(Z_i \in dz_i) \]

where

\[ Q^*(Z_i \in dz) = \exp \left( \theta^*(a)z - \hat{\psi}(\theta^*) \right) P(Z_i \in dz) \]

and \( Q^* \) is such that \( E_{Q^*}Z_1 = a. \)
I(x)

“Dominating point”
Large deviations for random walk: $\mathbb{R}^2$

$S_n = \sum_{i=1}^{n} Z_i \in \mathbb{R}^2$, $Z_i$ are iid with distribution $p$ and mean 0.

As before, as $n \to \infty$,

$$P(S_n/n \in dx) \approx e^{-I(x)n},$$

and

$$P(S_n/n \in A) \approx \sup_{x \in A} e^{-I(x)n} \triangleq e^{-I(x^*)n}.$$
If crude Monte Carlo is used, the required sample size needed to compute $\alpha_n$ to relative precision $\epsilon$ is of order

$$\frac{1}{\epsilon^2} \frac{1}{P(S_n \in nA)} \approx \frac{1}{\epsilon^2} \exp(nI(x^*))$$

If we use the distribution $Q^*$ under which the $Z_i$’s are iid with

$$Q^*(Z_i \in dz) \propto \exp(\theta^* z)P(Z_i \in dz)$$

where $Q^*$ is chosen so that $E_{Q^*} Z_1 = “dominating point”$ for the rate function $I(x)$ over $A$, then the required sample size needed is of order

$$\frac{1}{\epsilon^2} \exp(o(n))$$

i.e. sub-exponential (“asymptotically efficient”)
Suppose $\theta$ is not chosen optimally

We sample from $Q$:

$$Q(Z_i \in dz) = \exp(\theta z - \Lambda(\theta))P(Z_i \in dz), \ i \geq 1$$

Importance sampling estimator involves generating $m$ copies of

$$L_n I(S_n \in nA)$$

where

$$L_n = \exp(-\theta S_n + n\Lambda(\theta))$$

What does the distribution of this estimator look like?
Let $I^\theta(\cdot)$ be the rate function associated with $Q$

Then,

$$P_Q(S_n \in nA) \approx \exp\left(-n \inf_{x \in A} I^\theta(x)\right)$$

Let the sample size

$$m = \exp(rn + o(n))$$

With sample size $m$, we will rarely see samples of $S_n$ outside $nA_r$, where

$$A_r = \left\{ x \in A : I^\theta(x) \leq r \right\}$$

Basically, for any $\delta > 0$,

$$P_Q \left( \frac{S_n}{n} \in A^C_{r+\delta} \right) \to 0$$

exponentially fast
Cover $A_r$ with small balls: Each ball has exponentially many samples within it; likelihood ratio is constant (at exponential scale) within each ball

Importance estimator is mixture of the contributions from the balls

The estimator inherits the behavior of the largest importance ratios seen within $A_r$

$$\frac{1}{n} \log(\hat{\alpha}_n) \overset{P}{\to} - \inf_{x \in A_r} I(x)$$

The algorithm is solving a “modified rare-event simulation problem”
Dominating point $x^*$

Mean under $P$

Mean under $Q$
Dominating point $x^*$

Mean under $P$

Mean under $Q$
Dominating point $x^*$

Mean under $P$

Mean under $Q$
Dominating point $x^*$

Mean under $P$

Mean under $Q$
Dominating point $x^*$

Mean under $P$

Mean under $Q$

$A_{x^*}$
Dominating point $x^*$

Mean under $P$

Mean under $Q$

$A_{f_2}$

$A_N$
Dominating point $x^*$

Mean under $P$

Mean under $Q$
As a function of $r$, 

$$
\frac{1}{n} \log (\hat{\alpha}_n^r) \xrightarrow{P} - \inf_{x \in A_r} I(x)
$$

uniformly on compact sets.

There is a critical $r^*$ at which the estimator begins to estimate $\alpha_n$ correctly (basically, when the algorithm “covers” the dominating point for $I(\cdot)$ over $A$).
Note that in logarithmic scale, importance sampling estimator increases monotonically to the correct answer.

On other hand, crude Monte Carlo gives an upper bound on the probability $P(A)$:

$n$ samples until no realizations of $A$

\[ \downarrow \]

probability is of order $1/n$ or smaller (Bayesian posterior)

Potential basis for new stopping rules.
Theorem (Exponential c.o.m., \( k \)'th moment)

Let the sample size \( m_n = \exp(rn + o(n)) \). Then

\[
\frac{1}{n} \log \left( \frac{1}{m_n} \sum_{j=1}^{m_n} (L_n^{(j)})^{(j)} \right)^k \xrightarrow{P} \sup_{x \in A_r} \left\{ k(\theta x - \Lambda(\theta)) - I^\theta(x) \right\}
\]

where \( A_r = \{ x : x \in A, I^\theta(x) < r \} \)

If \( k = 2 \), this theorem describes the small sample behavior of the sample variance.
Estimate $P(S_n/n \in A)$

\[ L_n = \left( \prod_{i=1}^{n} \frac{p(Z_i)}{q(Z_i)} \right) I(S_n/n \in A) \]

\[ = \exp \left( \sum_{i=1}^{n} \log \left( \frac{p(Z_i)}{q(Z_i)} \right) \right) I(S_n/n \in A) \]

Now $L_n$ depends on not just $S_n$, but the entire empirical distribution of $Z_i$'s.

We need the large deviation theorem for the empirical distribution of $Z_i$'s.
Sanov’s theorem: LDP for empirical distributions

Empirical distribution $\beta_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

**Example** Suppose that we roll a die 100 times.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td># appears</td>
<td>17</td>
<td>22</td>
<td>10</td>
<td>14</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>$\beta_n$</td>
<td>0.17</td>
<td>0.22</td>
<td>0.10</td>
<td>0.14</td>
<td>0.21</td>
<td>0.16</td>
</tr>
</tbody>
</table>

$\beta_n = (0.17, 0.22, 0.10, 0.14, 0.21, 0.16)$
As \( n \to \infty \) the empirical distribution \( \beta_n \) will get close to \( \left( \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right) \), the distribution \((\mu)\) from which \( X \) sampled. i.e.,

\[
\beta_n \xrightarrow{P} \mu
\]

How fast do rare event probabilities converge to zero?
Sanov’s theorem

As \( n \to \infty \) the empirical distribution \( \beta_n \) will get close to \( \left( \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right) \), the distribution \((\mu)\) from which \( X \) sampled. i.e.,

\[
\beta_n \xrightarrow{P} \mu
\]

How fast do rare event probabilities converge to zero?

**Theorem (Sanov’s theorem)**

\[
- \inf_{\nu \in \Gamma^o} H(\nu | \mu) \leq \lim_{n \to \infty} \frac{1}{n} \log P_\mu(\beta_n \in \Gamma) \leq - \inf_{\nu \in \bar{\Gamma}} H(\nu | \mu)
\]

where \( H(\nu | \mu) \triangleq \int \log \frac{\nu}{\mu} d\nu \) (Relative entropy, Kullback–Leibler divergence).

\( H(\nu | \mu) \) measures the distance between two measures \( \nu \) and \( \mu \).
Small sample behavior: General IS distribution

Estimate $P(S_n/n \in A)$

$$Z_n = \left( \prod_{i=1}^{n} \frac{p(X_i)}{q(X_i)} \right) I(S_n/n \in A)$$

$$= \exp \left[ n \int_{\mathbb{R}^d} \log \left( \frac{p(y)}{q(y)} \right) \beta_n(dy) \right] I(S_n/n \in A)$$

where $\beta_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ (empirical distribution)
Virtually all importance sampling (IS) theory to date has focused on use of variance calculations.

Of course, IS is consistent even when the variance is infinite.
Example (Glasserman and Wang, 1997)

- \( S_n = Z_1 + \cdots + Z_n; \) \( Z_i \)'s iid

- Compute \( P(S_n/n < -a \text{ or } S_n/n > b) \) where \( a, b > 0 \)

- If \( I(b) < I(-a) \), the optimal static importance distribution is same as for computing \( P(S_n/n > b) \)

- The variance of this estimator can be infinite
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$

![Graph showing Dynamic IS for $P(S_n > nb)$]
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$

![Graph showing $S_1$ and $nb$]
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$
There are a number of algorithmic fixes to deal with this ...

Dynamic IS for $P(S_n > nb)$

After each step, we adjust the IS distribution.
Dynamic IS is used for better variance, but dynamic IS can be hard to implement.

Depending on the relative accuracy to be achieved, the static IS estimator may be good enough.
Logarithmic relative accuracy: Random walk

Let $\alpha_n = P(S_n/n \in A)$, and let $\hat{\alpha}_n$ be the IS estimator with sample size $m_n$ (i.e., $\hat{\alpha}_n = \frac{1}{m_n} \sum_{j=1}^{m_n} L_n^j (S_n^j/n \in A)$) where $m_n = \exp(rn + o(n))$

We define the *logarithmic relative accuracy* (LRA) associated with sample size $\exp(rn)$

$$LRA(r) \triangleq \lim_{n \to \infty} -\frac{1}{n} \log \left| \frac{\hat{\alpha}_n}{\alpha_n} - 1 \right|$$

If $LRA(r) \approx \delta$, then

$$\hat{\alpha}_n = \alpha_n \left( 1 + \Theta(e^{-\delta n}) \right)$$
Second main result: Exponential c.o.m.

**Theorem (Exponential c.o.m.)**

$$-rac{1}{n} \log \left| \frac{\hat{\alpha}_n}{\alpha_n} - 1 \right| \xrightarrow{P} \inf_{x \in A} \left\{ I(x) + \frac{1}{2} \left( r - I^\theta(x) \right)^+ \right\} - \inf_{x \in A} I(x)$$
Second main result: Exponential c.o.m.

Theorem (Exponential c.o.m.)

\[-\frac{1}{n} \log \left| \frac{\hat{\alpha}_n}{\alpha_n} - 1 \right| \xrightarrow{P} \inf_{x \in A} \left\{ I(x) + \frac{1}{2} \left( r - I^\theta(x) \right)^+ \right\} - \inf_{x \in A} I(x)\]

Corollary $LRA(r) = \frac{1}{2} r$ for all $r \geq 0$ if and only if the estimator is asymptotically efficient
Logarithmic relative accuracy

\[ P \left( \frac{S_n}{n} < -a \text{ or } \frac{S_n}{n} > b \right), a = 1.3, b = 1, \text{ exp c.o.m. with mean } b. \]
Logarithmic relative accuracy

\[ P\left(\frac{S_n}{n} < -a \text{ or } \frac{S_n}{n} > b\right), a = 1.3, b = 1 \]
\[
P(S_n/n < -a \text{ or } S_n/n > b), \ a = 1.1, \ b = 1
\]
Second main result: General distribution

Theorem (General distribution)

\[-\frac{1}{n} \log \left| \frac{\hat{\alpha}_n}{\alpha_n} - 1 \right| \xrightarrow{P} \inf_{\nu \in \Gamma} \left\{ H(\nu|p) + \frac{1}{2} (r - H(\nu|q))^+ \right\} - \inf_{\nu \in \Gamma} H(\nu|p) \]

where \( \Gamma = \{ \nu : \int y\nu(y) \in A \} \).
It is worth bearing in mind that in typical rare-event simulations, one doesn’t care about getting many significant figures of accuracy.

The exponent of the rare-event probabilities is often enough.

So, low relative accuracy is often sufficient.

Perhaps presents new opportunities for algorithm design?
Conclusion

We have analyzed the small-sample behavior of IS estimators, provides insight into:

- Behavior of “suboptimal” importance distributions
- Possible error diagnosis for IS estimators
- Low relative accuracy estimators vs. high relative accuracy estimators
Thank you