Monte Carlo methods for high dimensional integration

Mark Huber
Fletcher Jones Foundation Associate Professor of Mathematics and Statistics and George R. Roberts Fellow
Chair of the Department of Mathematical Sciences
Claremont McKenna College

3 June, 2017

What are Monte Carlo methods?

Monte Carlo for high dimensional integration

Simulation

How do I draw a vector $(X_1, \ldots, X_n)$ of nonidentical, highly dependent random variables when n is large?

Estimation

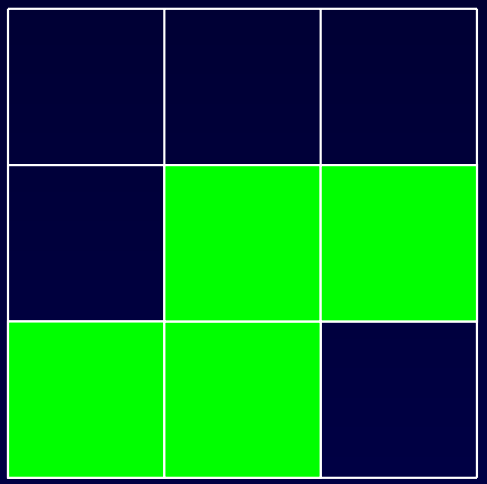
What is the best way to use these samples to estimate high dimensonal integrals and sums?

$$\int_{\mathbb{R}^n} f(\vec{x})\ d\mathbb{R}^n, \quad \sum_{\mathbf{Z}^n} w(\vec{x})\ d\mathbb{R}^n.$$


Motivation

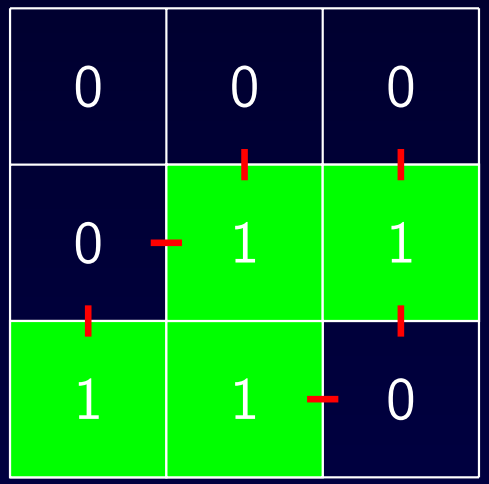
Spatial interaction and the statistical analysis of lattice systems (with Discussion)
J. Besag
JRRS Ser B Stat Methodology, *36:192–236, 1974*

Consider a 3 by 3 set of plots of land
Each plot either has good soil or bad soil
Can represent the state space by $\{0, 1\}$



Assigning a probability measure

- Label the good plots with a 1 and bad with a 0
- So state space is $\{0,1\}^9$.
- There are $2^9 = 512$ different states
- Let $c(x)$ be the # of different squares with the same label



$$c(x) = 6$$

- Mathematically, $c(x)$ is just the size of the cut between nodes labeled 1 and nodes labeled 0.

Assigning a probability measure, part 2

- For $x \in \{0,1\}^9$ and $\gamma > 0$, let

$$w(x) = \gamma^{c(x)},$$
$$Z_\gamma = \sum_{x \in \{0,1\}} w(x),$$
$$\mathbb{P}_\gamma(X = x) = \frac{w(x)}{Z_\gamma}.$$

- Call Z_γ the *normalizing constant* or *partition function*
- When $\gamma \in [0,1]$, this is the *ferromagnetic Ising model*

Assigning a probability measure, part 3

0	0	0
0	1	1
1	1	0

$c(x) = 6$

$w(x) = \gamma^6$

$\mathbb{P}_\gamma(X = x) = \gamma^6 / Z_\lambda$

Same model comes up in different ways

1. Statistical physics
2. Computer science notions of problem complexity
3. Frequentist statistical inference
4. Bayesian statistical inference

Statistical physics

Beitrag zur Theorie des Ferromagnetismus
E. Ising
Zeitschrift für Physik, *31:253–258, 1925*

- Model of magnetism, each domain either spin up or spin down
- Lenz gave problem to his student Ising
- Ising model in 2 dimensions exhibits a phase transition
- As the model passes critical γ, it goes from isolated connected components to giant component
- Notation: $\gamma = \exp(-1/T) = \exp(-\beta)$, where β called the *inverse temperature*

Computational complexity

Definition (Polynomial checkable set)

Say that a set S is *polynomially checkable* if there exists a polynomial running time Turing machine that for all x can verify if $x \in S$ or $x \notin S$.

Definition (Class NP of problems)

A problem is in NP if it asks if $S = \emptyset$ or $S \neq \emptyset$, where S is polynomially checkable.

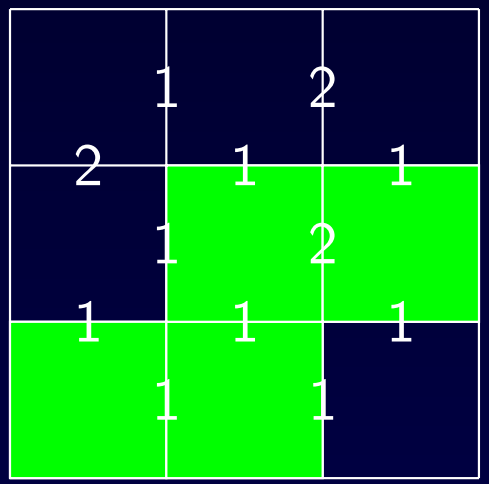
Definition (Class #P of problems)

A problem is in #P (number P) if it asks what is $\#(S)$ (the number of elements of S) where S is polynomially checkable.

An example of a poly checkable set

An example S

- Start with a labeling x of the nodes from $\{0, 1\}$
- For each edge $\{i, j\} \in E$
  - If $x(i) = x(j)$ then let $y(\{i, j\}) \in \{1, 2\}$
  - else $x(i) \neq x(j)$, and must set $y(\{i, j\}) = 1$.



Given $(x, y) \in \{0, 1\}^V \times \{1, 2\}^E$, easy to check if $(x, y) \in S$ or not

Relation of example to Ising

- Note that each edge with like labels has twice the number of labelings as an edge whose labels are different
- Equivalent to giving unlike labeled edge $1/2$ the weight of like labeled edges

Fact

Suppose $(X, Y) \sim$ *Unif*(S)*. Then* X *has the Ising distribution with* $\gamma = 1/2$*.*

Hardness of the problem

Definition

Say that a problem with set S is #P complete if when there exists a polynomial time algorithm for finding $\#(S)$, there exists a polynomial time algorithm for finding $\#(S')$ for any polynomial time checkable set.

So if a poly time algorithm exists for the problem, then **all** problems in #P can be solved in polynomial time!

A side note

The number sign # has

- Two lines at or almost vertical
- Two lines that are exactly horizontal

The sharp sign ♯ (from music notation) has

- Two lines that are exactly vertical
- Two lines with slightly positive slope (so they can be read on top of the staff lines with music on them)

Hardness of Ising

Polynomial-time approximation algorithms for the Ising model
M. Jerrum and A. Sinclair
SIAM J. Comput., *22:1087–1116, 1993*

Theorem (Jerrum Sinclair 1993)

The Ising model with $\gamma = 1/2$ is $\#P$ complete.

- So if there was some poly time method, then we have a poly time method for all S that are poly checkable
- Since #P harder than NP, also gives poly time method for factorization and traveling salesman problem
- Try for approximating Z_γ rather than getting exact value

Frequentist inference

- In data $c(x) = 5$
- Likelihood of getting x with $c(x) = 5$ is γ^5 / Z_γ
- Maximum likelihood estimator is

$$\arg\max_{\gamma \in [0,1]} \frac{\gamma^5}{Z_\gamma}$$

Bayesian inference

- Put prior density f_{prior} on γ
- Take data, find $c(x) = 5$
- Use Bayes rule to find posterior density for γ:

$$\begin{aligned} f_{\text{posterior}}(g) &\propto f_{\text{prior}}(g) g^5 / Z_g \\ &= \frac{f_{\text{prior}}(g) g^5 / Z_g}{\int_{g' \in [0,1]} f_{\text{prior}}(g') (g')^5 / Z_{g'}} \end{aligned}$$

Statistical inference

- To do either frequentist or Bayesian inference for γ, need to know Z_γ for $\gamma \in [0, 1]$.
- Dimension of sum/integral to computer Z typically equals the number of data points gathered
  - For our 3 by 3 Ising model, each of nine plots are 0 or 1, that is nine data points, so nine dimensional problem

Why generate samples from high dim distributions?

Computer science

- #P-complete problems count number of discrete structures with a given property
- Example: count independent sets of a graph or number of matchings in a graph
- Randomized approximation scheme for these problems

Frequentist statistical models

- Dimension of problem typically equal to number of data points
- Samples needed for exact p-values, exact confidence intervals
- Example: Spatial models

Bayesian posterior

- Dimension of problem typically equal to number of data points
- Dimensions weakly dependent through prior
- Example: mixture models, hierarchical models


ALGORITHM

My work: algorithms for simulation and estimation

Simulation

- Bounding chains (Huber 2006)
- Randomness recycler (H. & Fill 2001)
- Birth-death-swap chains (H. 2012)
- Bernoulli factories (H. 2016, 2017)

Estimation

- Tootsie Pop Algorithm (TPA) (H. & Schott 2014)
- Gamma Bernoulli Approximation Scheme (H. to appear)
- Approximating normalizing constants for Gibbs distributions (H. 2015)

Bernoulli random variables

Estimating the size of a set using draws from a larger set

Bernoulli

If a random variable is 0 or 1, call it a Bernoulli random variable. Named after Jakob Bernoulli (1654–1705).

Procedure

For $S \subseteq S'$:

1. Draw $A \sim \text{Unif}(S')$
2. If $A \in S$ set $B = 1$, otherwise $B = 0$
3. Return B

- Can also just use indicator function notation:

$$\text{output} = \mathbb{1}(A \in S), \text{ where } A \sim \text{Unif}(S')$$

- Bernoulli random variables also called indicator r.v.'s
- Note $\mathbb{P}(B = 1) = \#(S)/\#(S')$

For Ising with $\gamma = 1/2$

1. For each node i, choose $X(i) \sim \text{Unif}(\{0, 1\})$
2. For each edge e, choose $Y(e) \sim \text{Unif}(\{1, 2\})$
3. If $\forall$ edges $\{i, j\}$ with $X(i) \neq X(j)$, $Y(\{i, j\}) = 1$, return 1,
4. Otherwise return 0.

The $\mathbb{P}(\text{output} = 1) = \#(S)/2^{\#V + \#E}$.

Also

$$Z_{1/2} = \frac{\#(S)}{2^{\#E}} = 2^{\#V}\mathbb{P}(\text{output} = 1)$$

For general ferromagnetic Ising with $\gamma \in [0, 1]$

1. For each node i, choose $X(i) \sim \text{Unif}(\{0, 1\})$
2. For each edge e, choose $Y(e) \sim \text{Unif}([0, 1])$
3. If $\forall$ edges $\{i, j\}$ with $X(i) \neq X(j)$, $Y(\{i, j\}) \leq \gamma$, return 1,
4. Otherwise return 0.

Then

$$Z_\gamma = 2^{\#V} \mathbb{P}(\text{output} = 1)$$

Relative error

- Because $\mathbb{P}(\text{output} = 1)$ is being multiplied by such a large term, interested in relative error
- For true value a and estimate $\hat{a}$, the relative error is

$$\frac{\hat{a}}{a} - 1$$

Randomized approximation schemes

Definition

An estimate $\hat{a}$ for a is an (ϵ, δ)-*randomized approximation scheme* if

$$\mathbb{P}\left(\left|\frac{\hat{a}}{a} - 1\right| > \epsilon\right) \leq \delta.$$

The Gamma Bernoulli Approximation Scheme

A new way to estimate the mean of data that has the Bernoulli distribution

Main results about GBAS

Consider

$$B_1, B_2, \ldots \overset{\text{iid}}{\sim} \text{Bern}(p)$$

GBAS gives a (ϵ, δ)-ras $\hat{a}$ for $\mathbb{E}[B_i]$.

1. Can make unbiased.
2. Biased version can use fewer samples than the Central Limit Theorem

Stronger result

For GBAS, the distribution of $\hat{a}/a$ is known completely. That is, it does not depend on a in any way!

First estimate for Bernoulli's that does this

- Takes advantage of nature of Monte Carlo data
- Uses a random number of flips of the coin
- Adds extra smoothing randomness to the mix

Main results about relative error

Theorem (H. 2013)

Given $B_1, B_2, \ldots$ *iid Bern*(p) *where* p *is unknown, there exists an estimate* $\hat{p}_k = \hat{p}_k(B_1, \ldots, B_T)$ *such that the distribution of* $p/\hat{p}_k \sim$ *Gamma*$(k, (k-1)c)$ *for user-specified* k *and* c*, and* $\mathbb{E}[T] = k/p$.

Theorem (H. 2016)

Given $A_1, A_2, \ldots$ *iid Pois*(μ) *where* μ *is unknown, there exists an estimate* $\hat{\mu}_k = \hat{\mu}_k(A_1, \ldots, A_T)$ *such that the distribution of* $\mu/\hat{\mu}_k \sim$ *Gamma*$(k, (k-1)c)$ *for user-specified* k *and* c*, and* $\mathbb{E}[T] \in [k/\mu, k/\mu + 1]$.

Main results with added bias

Theorem (Feng, H., Ruan 2016)

For both of the previous theorems, setting $c = 1$ *gives an unbiased estimator. Setting*

$$c = \frac{2\epsilon}{(1-\epsilon^2)[\ln(1+\epsilon) - \ln(1-\epsilon)]} = 1 + \frac{2}{3}\epsilon^2 + O(\epsilon^4)$$

makes

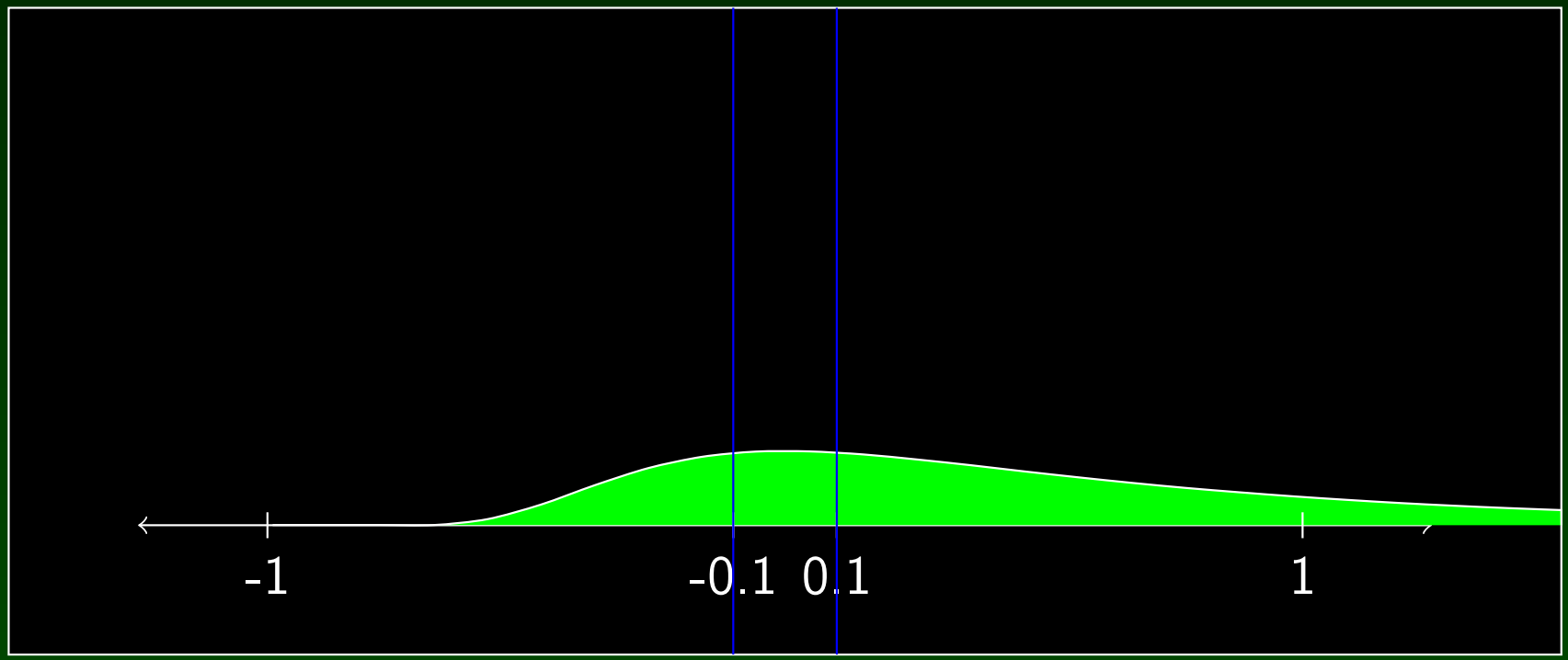
$$\mathbb{P}(|\textit{relative error}| > \epsilon) \leq c_1 \cdot c_2^k$$

where c_2 *is as small as possible.*

With GBAS the user knows the relative error distribution

User set $k = 5$, $c = 1$

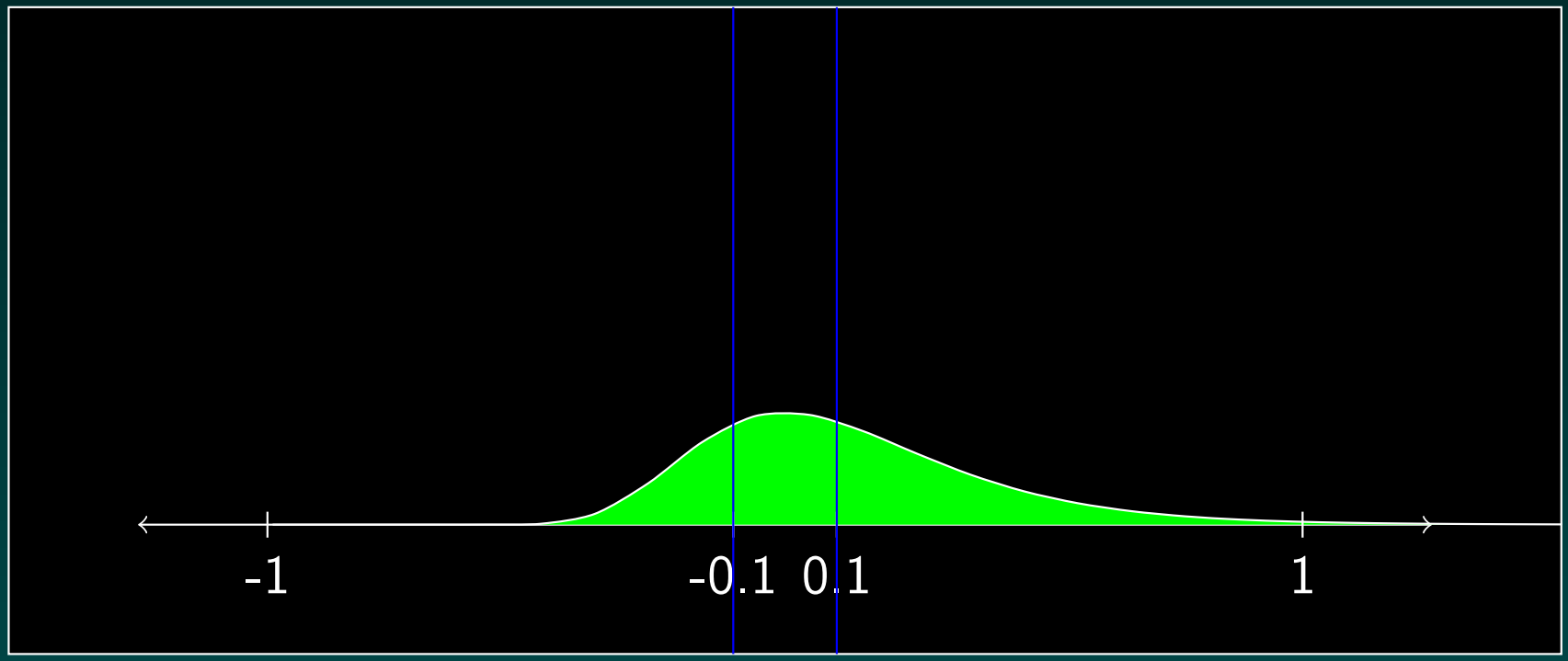
$\mathbb{P}(|\text{rel err}| > 0.1) \approx 92.6\%$



New algorithms know relative error distribution

User set $k = 20$, $c = 1$

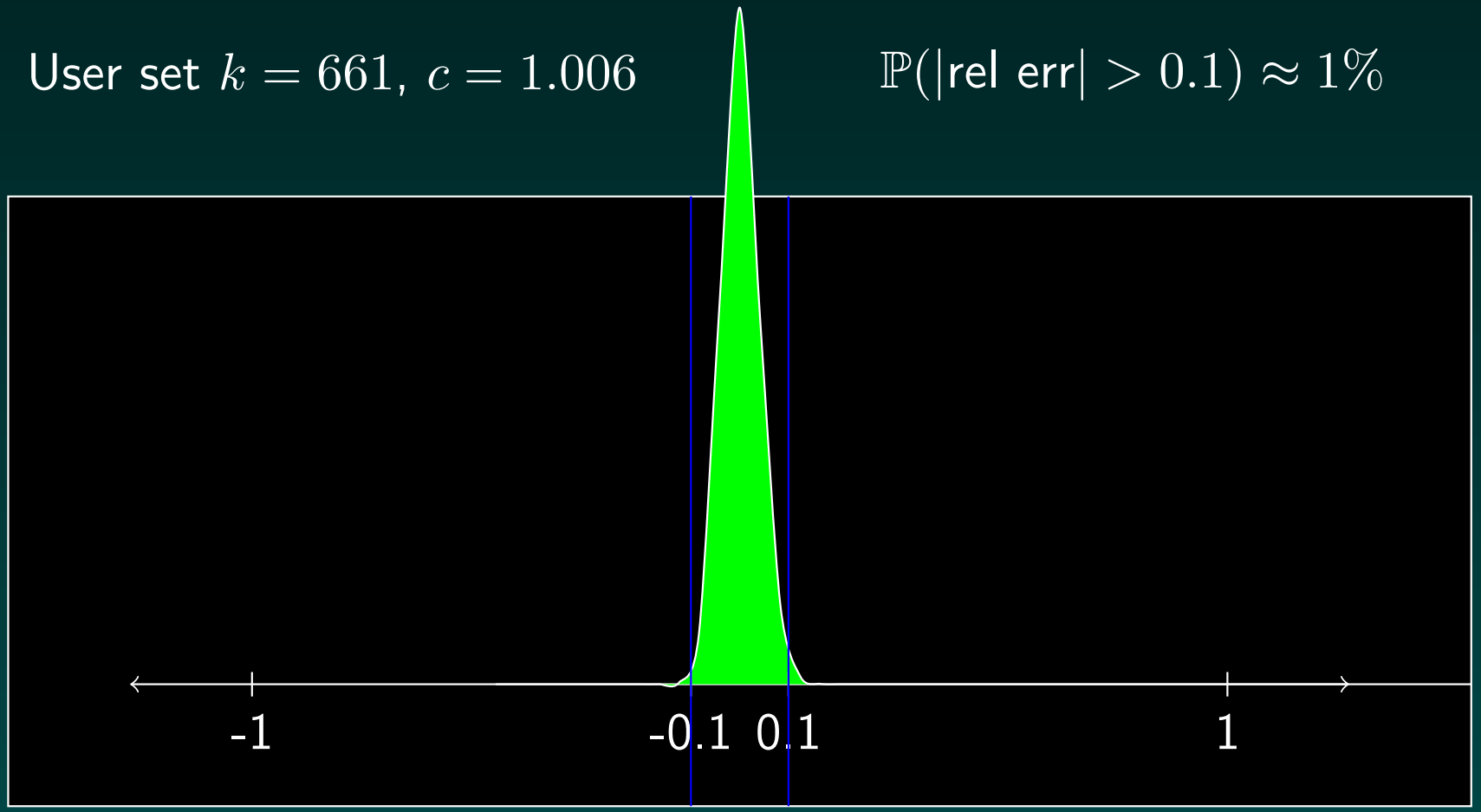
$\mathbb{P}(|\text{rel err}| > 0.1) \approx 66.1\%$



New algorithm knows relative error distribution

User set $k = 661$, $c = 1.006$

$\mathbb{P}(|\text{rel err}| > 0.1) \approx 1\%$



Running time

Lemma

Given k, and $X_1, X_2, \ldots$ iid Bern(p), the expected number of draws used by the algorithm is

$$\frac{k}{\mathbb{E}[X_i]}.$$

Setting the relative error

Suppose want relative error of 10%

- GBAS user decides k
- Example: if $k = 661$, then

$$\mathbb{P}(\text{relative error} > 0.1) < 0.01$$

- If CLT holds perfectly: data $X_1, X_2, \ldots$ iid $\mathsf{N}(\mu, \mu)$ then

$$k > 663.48$$

Method of moments estimator

Definition

The *basic estimate* for $\mathbb{E}[B_i]$ given $B_1, B_2, \ldots$ is

$$\hat{p}_{\mathsf{BE}} = \frac{B_1 + \cdots + B_n}{n}.$$

$$(1)(1)(0)(1)(0)(0)(1)(1)(1)(0) \to \hat{p}_{\mathsf{BE}} = \frac{6}{10}$$

Properties of the basic estimate

Pros

- Unbiased
- Easy to implement

Cons

- Not an (ϵ, δ)-ras

Basic estimate relative error

The values that relative error takes on depends on p

For $n = 5$

$$B_1 + B_2 + \cdots + B_5 \in \{0, 1, 2, 3, 4, 5\}$$

so

$$\text{rel err} \in \left\{ \frac{0}{5p} - 1, \frac{1}{5p} - 1, \ldots, \frac{5}{5p} - 1 \right\}$$

Measuring error in estimate

- Random error in estimate measured with standard deviation
- Say $\mathbb{E}[X] = \mu$, $\mathrm{SD}[X] = \sigma$
- Then for $\hat{\mu}_n = (X_1 + \cdots + X_n)/n$,

$$\mathbb{E}[\hat{\mu}_n] = \mu, \ \mathrm{SD}[\hat{\mu}] = \frac{\sigma}{\sqrt{n}}$$

- To make $\mathrm{SD}[\hat{\mu}] \leq \epsilon\mu$, need

$$n = \epsilon^{-2}\frac{\sigma^2}{\mu^2}$$

samples

How many samples needed for Bernoulli?

For basic estimate:

$$\hat{p}_{BE} = \frac{B_1 + \cdots + B_n}{n},$$

- $\mathbb{E}[B_i] = p$, and $\mathrm{SD}(B_i) = \sqrt{p(1-p)}$.
- Need
$$n \approx \epsilon^{-2} \frac{1-p}{p}.$$
- But we don't know p!

DKLR

An optimal algorithm for Monte Carlo estimation
P. Dagum, R. Karp, M. Luby, and S. Ross
SIAM J. Comput., *Vol 29, No 5, pp. 1484–1496, 2000*

- Idea: Use $\{B_i\}$ to form $\{G_i\}$, where $G_1, G_2, \ldots$ iid Geo(p)
- $\mathbb{E}[G_i] = 1/p$, $\mathrm{SD}(G_i) = (1/p)\sqrt{1-p}$
- Use

$$\hat{p} = \left[\frac{G_1 + \cdots G_k}{k}\right]^{-1}$$

where $k \approx \epsilon^{-2}(1-p)$ to get relative error below ϵ

Two problems with DKLR

- Biased estimate (in general $\mathbb{E}[1/X] \neq 1/\mathbb{E}[X]$)
- Hard to get correct constant and lower order terms for k

Gamma Bernoulli Approximation Scheme (GBAS)

A Bernoulli mean estimate with known relative error distribution
M. Huber
Random Structures & Algorithms, *arXiv:1309.5413, to appear.*

Starting with sequence of Bernoulli random variables with mean p

1. Construct a sequence of exponential random variables with mean $1/p$
2. Estimate the mean of these exponential random variables
3. Take reciprocal to get estimate for p

Converting Bernoulli to Geometric

This is straightforward:

- Start with $B_1, B_2, \ldots \sim \text{Bern}(p)$
- Let $G = \inf\{t : B_t = 1\}$
- Then say G is geometric with mean $1/p$, write $G \sim \text{Geo}(1/p)$

$$\mathbb{E}[G] = 1/p, \ \text{SD}(G) = \sqrt{1-p}/p.$$

So the standard deviation already same magnitude as mean

- This is what DKLR used: $\Theta(\epsilon^{-2} \ln(1/\delta))$ iid draws from G

Poisson point process

Definition (Poisson point process)

Say $P \subset \mathbb{R}$ is a Poisson point process (PPP) of rate λ if

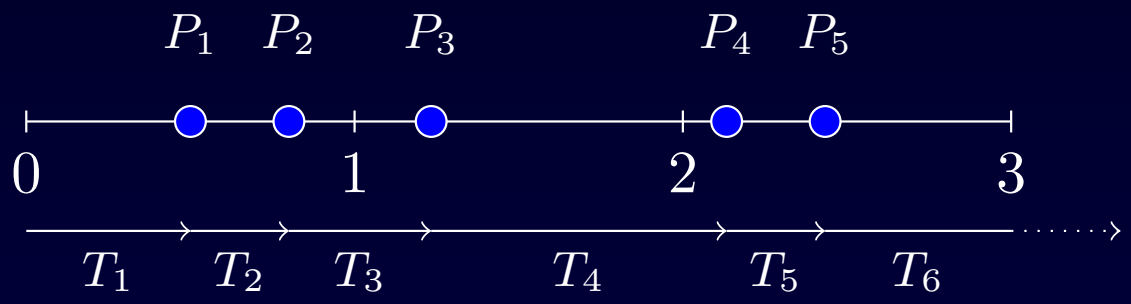
- Mean points in interval For all $a < b$,

$$\mathbb{E}[\#P \cap [a, b]] = \lambda \cdot (b - a).$$

- Independent intervals For all $a < b < c < d$, $P \cap [a, b]$ and $P \cap [c, d]$ are independent.

How to turn Bernoullis into exponentials, part 1

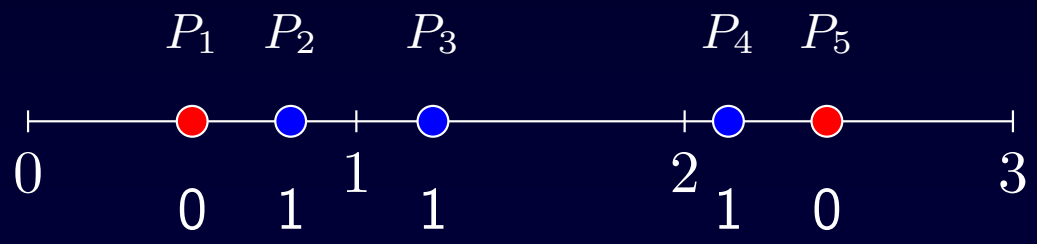
Start with a Poisson point process of rate 1



Fact about PPP: $T_1, T_2, \ldots$ are iid Exp(1)

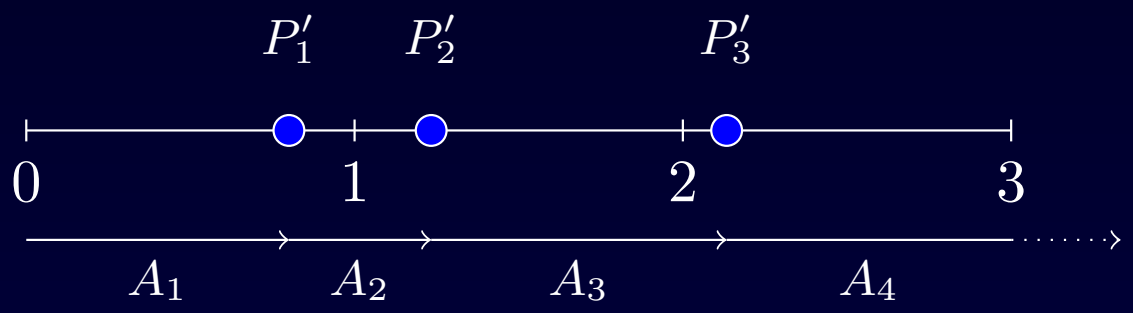
How to turn Bernoullis into exponentials, part 2

For each point of the process, flip a Bern(p) coin



How to turn Bernoullis into exponentials, part 3

Only keep those with Bernoulli draw 1



Result is a Poisson point process of rate p

So $A_1, A_2, \ldots$ iid Exp(p)

Thinning Poisson processes

- Suppose buses arrive at a stop at rate 1 per hour
- Chance a bus is red is 0.3
- Rate at which red buses arrive is 0.3 per hour
- Multiply rate by chance of a coin flip being heads

Geometric sum of exponentials is exponential

The formal proof uses the following two facts.

Fact

If $B_1, B_2, \ldots \sim$ Bern(p) are iid, then $G = \min\{i : B_i = 1\}$ has a geometric distribution with parameter p. Write $G \sim$ Geo(p).

Fact

Let $G \sim$ Geo(p), and $[R|G] \sim$ Gamma$(G, 1)$. Then $R \sim$ Exp(p).

How big should k be?

Bounding tails of gamma distributions for applications
J. Feng, M. Huber, S. Ruan, Y. Zhang
preprint, 2016

Short answer: at most $k = 2\epsilon^{-2}\ln(1/\delta) + 1$.

Long answer

Theorem

For $\epsilon > 0$, $\delta > 0$, set

$$c = \frac{2\epsilon}{(1-\epsilon^2)\ln(1+2\epsilon/(1-\epsilon))},$$

and

$$k = 2\epsilon^{-2}\ln(1/\delta) + 1.$$

For $G \sim \mathsf{Gamma}(k, k-1)$,

$$\mathbb{P}\left(\frac{1}{cG} \in [1-\epsilon, 1+\epsilon]\right) > 1 - \delta\frac{1+\epsilon}{\sqrt{\pi\ln(1/\delta)}}.$$

The Tootsie Pop Algorithm

A general way of turning integrations into finding the mean of a Poisson random variable

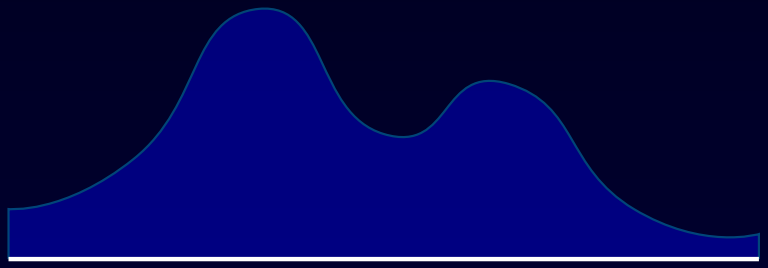
Numerical integration

Problem:

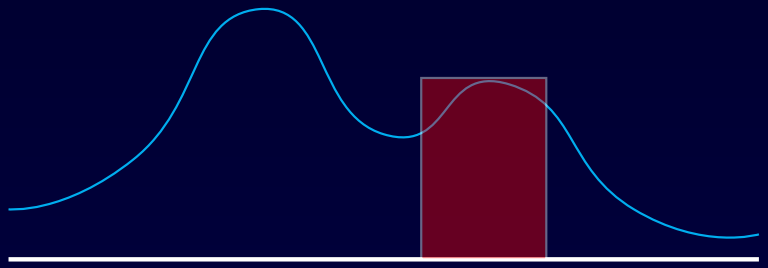
$$\int_{x \in \mathbb{R}^n} f(x)\ d\mathbb{R}^n,\ \ f(x) \geq 0$$

Suppose we can generate samples from f

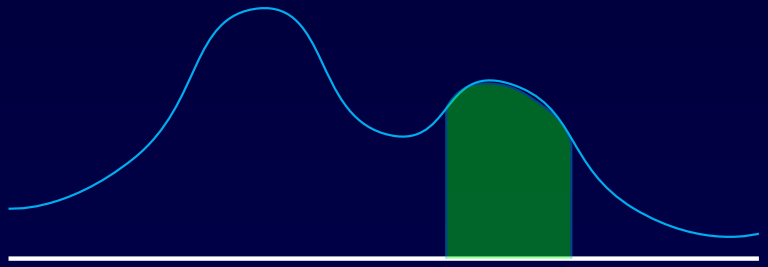
Three sets



A is the area under $f(x)$



Let $(x^*, f(x^*))$ be a local mode
B is area under $f(x^*)$ within distance α of x^*

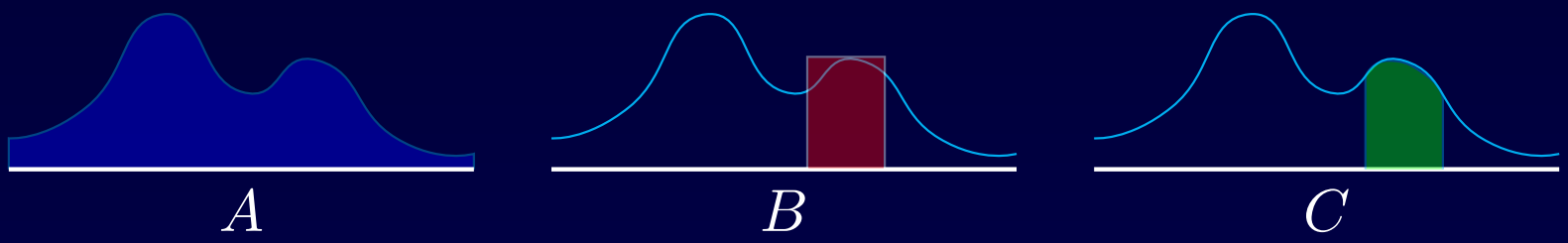


$C = A \cap B$

Using the three sets

- Know $\mu(B)$, want $\mu(A)$
- Estimate $\hat{p}_1 \approx \mu(C)/\mu(B)$
- Estimate $\hat{p}_2 \approx \mu(C)/\mu(A)$
- Then $\hat{\mu}(A) \approx \mu(A)$ where

$$\hat{\mu}(A) = \frac{\hat{p}_1}{\hat{p}_2}\mu(B)$$

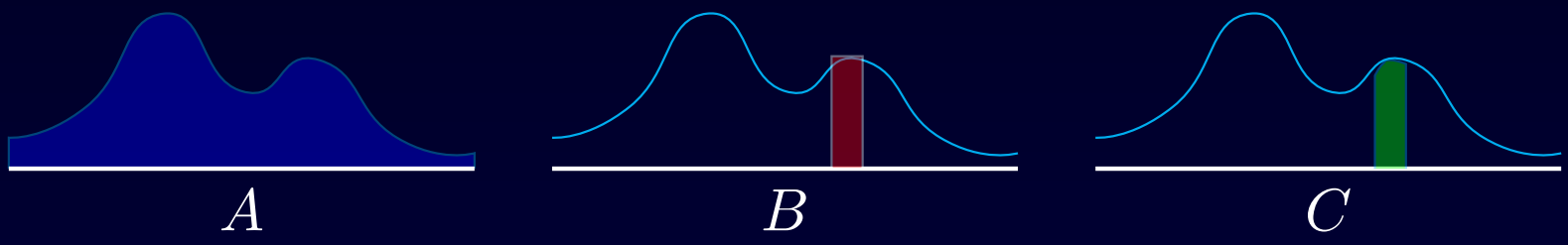


Again need relative error

If $\hat{p}_1$ and $\hat{p}_2$ have relative error at most ϵ,

$$\frac{1-\epsilon}{1+\epsilon}\mu(A) \leq \hat{\mu}(A) \leq \frac{1+\epsilon}{1-\epsilon}\mu(A)$$

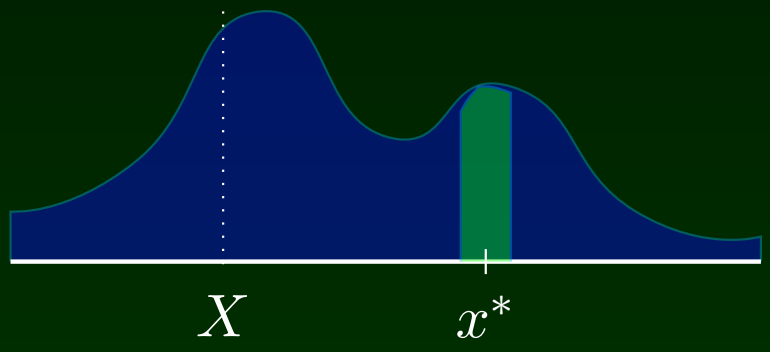
Small α



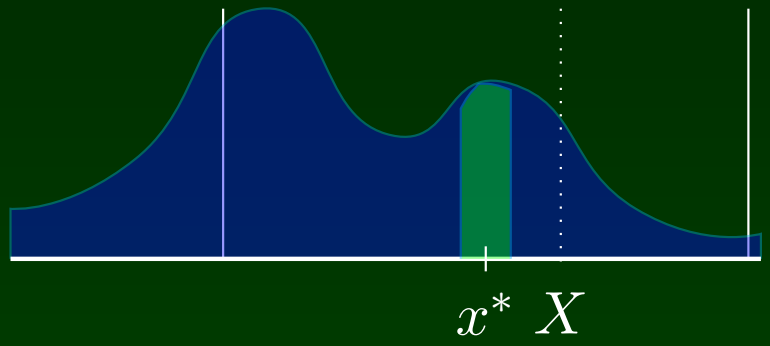
Makes

$$\frac{\mu(C)}{\mu(A)} \text{ small and } \frac{\mu(C)}{\mu(B)} \text{ large}$$

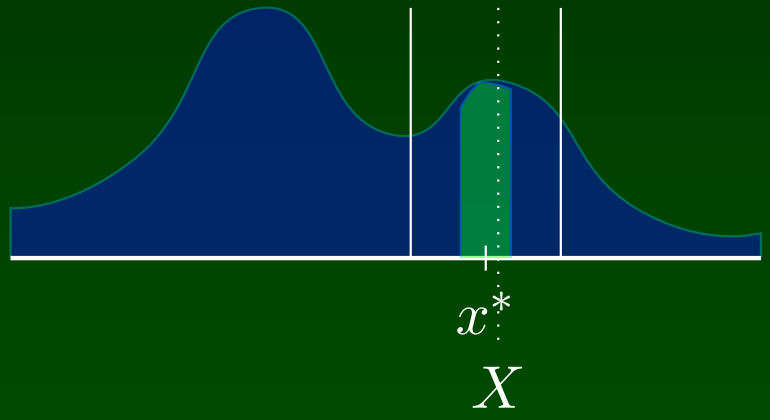
TPA in context of A and B



Set $\beta \leftarrow \infty$
Draw $X \sim f|\mathsf{dist}(X, x^*) \leq \beta$
Set $\beta \leftarrow \mathsf{dist}(X, x^*)$



Repeat
Draw $X \sim f|\mathsf{dist}(X, x^*) \leq \beta$
Set $\beta \leftarrow \mathsf{dist}(X, x^*)$



Until X falls in C

Poisson random variables

Definition (Poisson random variable)

Let P be a Poisson point process of rate μ. Then say that $N = \#(P \cap [0,1])$ is a Poisson random variable of mean μ. Write $N \sim \text{Pois}(\mu)$

Fact

For $N \sim$ Pois(μ), $\mathbb{V}(N) = \mu$.

What this gives us

Random construction of interpolating sets for high dimensional integration
M. Huber and S. Schott
J. of Applied Probability, arXiv:1112.3692. Vol 51, No 1, pp. 92–105, 2014.

Fact

Let N be the # of draws of X before X falls in C. Then

$$N \sim \textit{Pois}\left(\ln\left(\frac{\mu(A)}{\mu(C)}\right)\right).$$

Gamma Poisson Approximation Scheme

An estimator for Poisson means whose relative error distribution is known
M. Huber
arXiv:1605.09445, 2016

Algorithm similar to GBAS

- Gives an estimate $\hat{\mu}$ for μ such that $\mu/\hat{\mu} \sim \text{Gamma}(k, k-1)$
- Requires k/μ draws on average

Consequence

Change from needing

$$2\frac{\mu(A)}{\mu(C)}\epsilon^{-2}\ln(\delta^{-1})$$

samples for AR to

$$2\ln\left(\frac{\mu(A)}{\mu(C)}\right)\epsilon^{-2}\ln(\delta^{-1})$$

using Poisson with TPA

Monte Carlo methods

They work!

- GBAS, GPAS, & TPA are useful for approximating high dimensional integrals and sums
- Algorithms looks very different from traditional approaches
- Estimation also different from other statistical contexts
- Taking advantage of random number of choices can make for big improvements over classical statistical estimators

Want to know more about perfect simulation?