PERFECT

Simulation

Lecture 4

Perfect simulation IV

Perfect Integration

Mark Huber
Claremont McKenna College

July, 2018

Supported by NSF grant DMS 1418495

Why Monte Carlo?

The purpose of Monte Carlo or Quasi Monte Carlo is to approximate the value of high dimensional integrals

$$I = \int_{\vec{x} \in A} g(x)\ dA$$

Sample average approach

Write integral as

$$I = \int_{\vec{x} \in A} f(x) f_X(x)\ d\vec{x}$$

where f_X is a probability density. Then draw

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_X,$$

use

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

There must be a better way!

Seen many approaches to improvement this last week

- SMC concentrate particles where $g(x) = f(x)f_X(x)$ large
- Multilevel Monte Carlo: lots of samples to get initial estimate, small number of samples to get correction
- QMC and variance reduction to get better spread of random variables

Perfect integration

- Learn from our samples about the space
- End with guaranteed relative error bounds on $\hat{I}$
- Need relative error since I usually exponentially large/small in dimension

Want user set relative error and chance of success

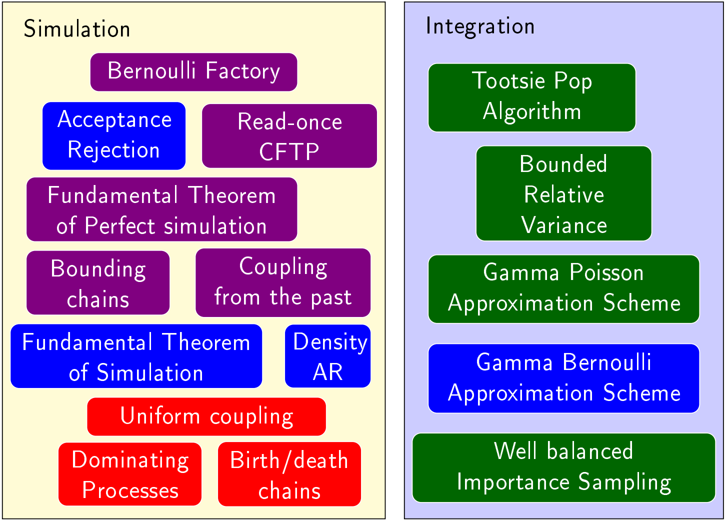
Definition

An algorithm is an (ϵ, δ)-**randomized approximation scheme** if for all $\epsilon, \delta > 0$, the output $\hat{a}$ of the algorithm satisfies (with respect to true answer a)

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

Our Goal

Develop (ϵ, δ)-ras for I where the number of samples grows polynomially (linearly if possible) in the dimension of the problem



Simulation

Bernoulli Factory

Acceptance Rejection

Read-once CFTP

Fundamental Theorem of Perfect simulation

Bounding chains

Coupling from the past

Fundamental Theorem of Simulation

Density AR

Uniform coupling

Dominating Processes

Birth/death chains

Integration

Tootsie Pop Algorithm

Bounded Relative Variance

Gamma Poisson Approximation Scheme

Gamma Bernoulli Approximation Scheme

Well balanced Importance Sampling

Can I use perfect integration methods wtih imperfect samples?

Absolutely!

Bad

Imperfect samples with imperfect integration

Good

Imperfect samples with perfect integration

Great!

Perfect samples with perfect integration

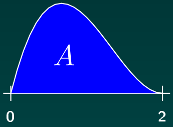
Gamma Bernoulli Approximation Scheme

All integrals are area/volume problems

For f a nonnegative function:

$$\int_{(x_1,\ldots,x_n)\in A} f(x_1,\ldots,x_n)\ d\mathbb{R}^n =$$

$$\mathsf{vol}(\{x_1,\ldots,x_n,y\}|(x_1,\ldots,x_n)\in A, y\in[0,f(x_1,\ldots,x_n)])$$

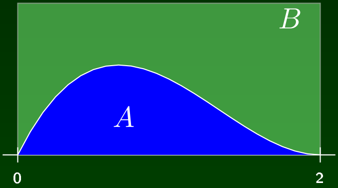


$\mathsf{Area}(A) = \int_0^2 4x^3 - 4x^2 + 4x\ dx$

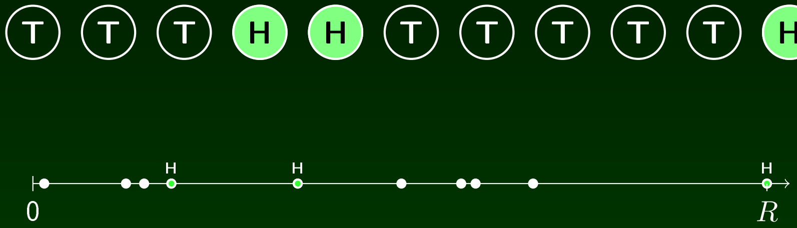
Convert area problem to mean of Bernoulli problem

Embed A in larger region B

1. Draw $X \leftarrow \text{Unif}(B)$ where $m(B)$ known
2. Then $\mathbb{P}(X \in A) \sim \text{Bern}(m(A)/m(B))$



Gamma Bernoulli Approximation Scheme outline



1. Use your Bern(p) to thin a 1D Poisson point process of rate 1
2. Result is a Poisson point process of rate p
3. Let R be location of kth unthinned point (then $R \sim \text{Gamma}(k, p)$)
4. Let $\hat{p} = (k-1)/R$ (so $p/\hat{p} \sim \text{Gamma}(k, k-1)$)

Gamma Bernoulli Approximation Scheme

```
Gamma_Bernoulli_Approximation_Scheme
Input: $k$    Output: $\hat{p}_k$

$S \leftarrow 0$, $R \leftarrow 0$.
Repeat
    $X \leftarrow \text{Bern}(p)$, $A \leftarrow \text{Exp}(1)$
    $S \leftarrow S + X$, $R \leftarrow R + A$
Until $S = k$
$\hat{p}_k \leftarrow (k - 1)/R$
```

M. Huber, A Bernoulli mean estimate with known relative error distribution, *Random Struc. & Alg., arXiv:1309.5413, 50:173–182, 2017*

Performance of GBAS

Recall that an (ϵ, δ)-ras satisfies

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

Average number of Bernoulli draws for (ϵ, δ)-ras at most

$$\lceil 2\epsilon^{-2} \ln(\delta^{-1}) \rceil / p$$

Bad for small p!

Theoretical Computer Science

How hard is optimization and integration?

How hard is integration?

Optimization $\leq$ Integration

NP problems

If you can evaluate $f(x)$ in polynomial time, the problem of deciding if $\exists x$ such that $f(x) \geq a$ for some a is in the NP class of problems.

Notable NP hard problems

1. Traveling Salesman Problem
2. Integer Programming
3. Max-cut in a graph

#P problems

If you can evaluate $f(x)$ in polynomial time, the problem of finding the measure of the set of x such that $f(x) \geq a$ for some a is in the #P (read as "Number-P") class of problems.

Notable #P hard problems

1. Find the number of Hamiltonian cycles in a graph
2. Volume of a convex body
3. How many cuts of a certain size are there in a graph?

#P harder than NP

- If you can solve the #P problem of measure, NP problem of existence is easy
- NP $\approx$ optimization, #P $\approx$ integration
- Sampling gives existence, so harder than NP
- Integration gives conditional marginals, so harder than sampling

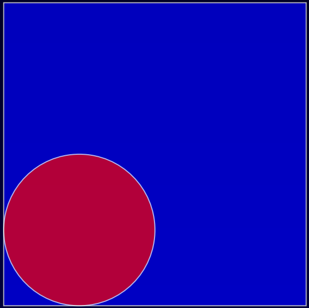
$$\text{optimization} \leq \text{sampling} \leq \text{integration}$$

- Goal here is to show that good sampling gives rise to good approximate integration

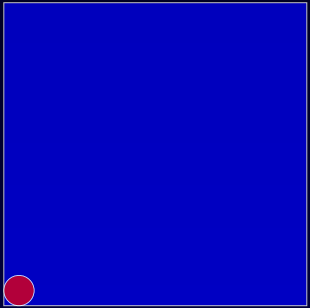
$$\text{sampling} \geq \text{approximate integration}$$

The Tootsie Pop Algorithm

A tale of two circles



$p = 0.1963$



$p = 0.007853$

Need 25 times as many samples for same relative error for circle that is 25 times as small

Handling small p

For statistical applications,

- p typically exponentially small in dimension of problem
- Want running time to grow as $\ln(1/p)$, not $1/p$
- Have to adapt our samples as we take steps

Tootsie Pop Algorithm

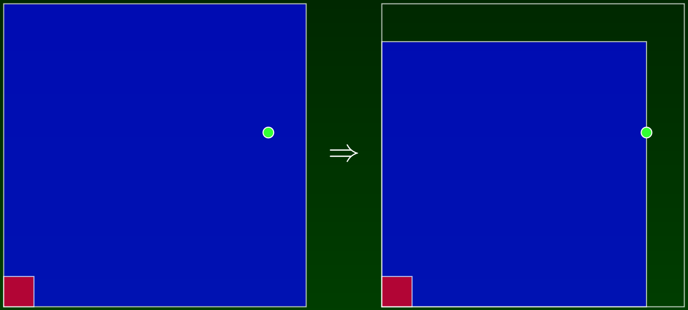
Acceptance-rejection

- Draw from large region, only accept if make it to small region in a single step

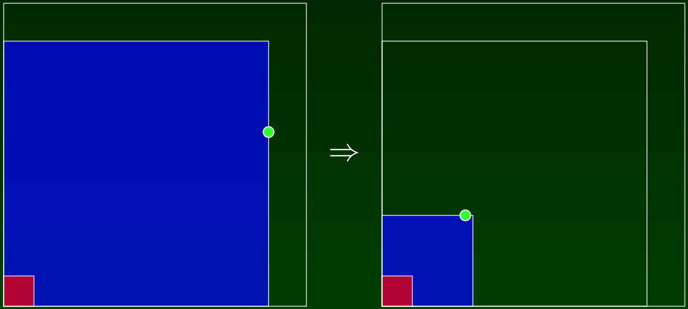
TPA

- Suppose draw from large region does not reach all the way to the small region.
- Remove everything in large region farther away from small region than your draw
- On average, remove half of large region at each step

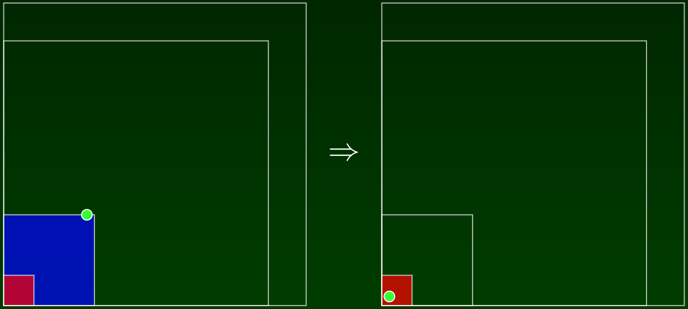
First step TPA



Second step TPA



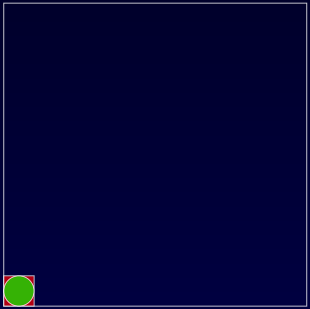
Third step TPA



It took three 3 steps to get to the red region

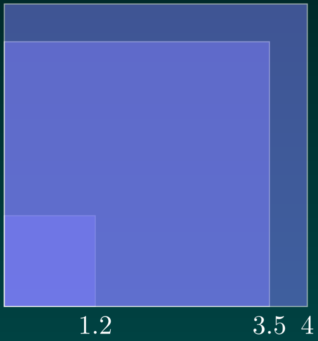
Getting to the small circle

Now that we know how to get to the small square, easy to get to the small circle with standard AR



Indexing the regions

Note that each blue region can be indexed by the length of the side of the square - call this index β



$\Rightarrow$

Measuring the regions

Let B_β be region indexed by β, and Z_β the area of B_β

$$\begin{aligned} Z_4 &= 16 \\ Z_{3.5} &= 12.25 \\ Z_{1.2} &= 1.44 \end{aligned}$$

One step of TPA

1. Draw $X \leftarrow \text{Unif}(B_\beta)$
2. Set $\beta \leftarrow \inf\{b : X \in B_b\}$

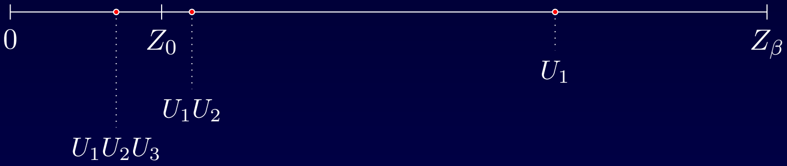
Main result about TPA

Theorem

Suppose $X \sim$ Unif(B_β), $\beta' = \inf\{b : X \in B_b\}$. If Z_β is continuous in β, then

$$\frac{Z_\beta}{Z_{\beta'}} \sim \textit{Unif}([0,1]).$$

So on average the size of the large region is cut in half in a step of TPA. It's like a randomized Zeno's walk!



Moving to log-space

Fun fact, if $U \sim \mathsf{Unif}([0,1])$, then

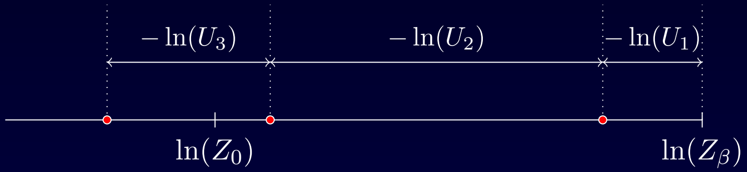
$$-\ln(U) \sim \mathsf{Exp}(1)$$

Reminder

$$\ln(U_1 U_2 \cdots U_k) = \sum_{i=1}^{k} \ln(U_i)$$

In log-space, the product of uniforms is the sum of exponentials

Poisson point process



How this gives us an integration algorithm

Corollary

The number of steps taken by TPA before landing in the small region is Poisson distributed with mean equal to the natural logarithm of the ratio between the small region and the original large region. So the expected number of steps taken by TPA is

$$\ln(1/p) + 1.$$

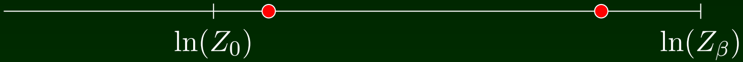
M. Huber and S. Schott, Random construction of interpolating sets for high dimensional integration, *Journal of Applied Probability*, arXiv:1112.3692, 51(1):92–105, 2014

Why Tootsie Pop Algorithm?

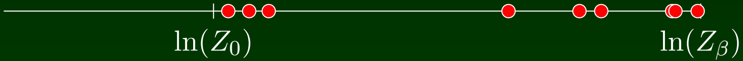
- A Tootsie Pop is a candy with a chocolately center surrounded by a candy shell
- An old ad campaign has Mr. Owl being asked "How many licks does it take to get to the center of the Tootsie Pop"
- For us, if it takes N, then $p = \exp(-\mathbb{E}[N] + 1)$

Easy to parallelize

One run of TPA gives Poisson point process of rate 1



Combine k runs of TPA to give Poisson point process of rate k

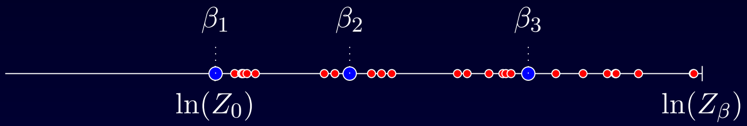


Number of points is then

$$\mathsf{Pois}(k[\ln(Z_\beta) - \ln(Z_0)])$$

Well balanced schedule

Suppose $k = 10$, and we consider every 10th point



- Distance between points (in log-space) is about 1
- Each β_i value has

$$\frac{Z_{\beta_{i+1}}}{Z_{\beta_i}} \approx e,$$

call such a schedule **well balanced**

What are well balanced schedules used for?

Certain Markov chains

- Simulated annealing
- Simulated tempering

Why well balanced temperatures

- Temp levels roughly reduce size of space by constant factor
- Necessary condition for rapid mixing of the chain [1] [2]

[1] D. B. Woodward, S. C. Schmidler, and M. Huber, Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions, *Annals of Applied Probability* 19(2):617–640, 2012

[2] D. B. Woodward, S. C. Schmidler, and M. Huber, Sufficient conditions for torpid mixing of parallel and simulated tempering, *Electronic Journal of Probability* 14:780–804, Article 29, 2009

TPA with Markov chains

- Suppose that I have a set of N particles
- Use min β that covers all particles
- Rate of Poisson point process now N instead of 1
    - If using Markov chains, duplicate one point, let points wander for a while
    - If using perfect simulation, draw one new point to replace

About Poisson random variables

Variance of a Poisson with mean $\ln(1/p)$ is

$$\ln(1/p)$$

Need to run TPA $\Theta(\ln(1/p))$ times for (ϵ, δ)-ras. Total steps

$$\lceil 2\epsilon^{-2}\ln(\delta^{-1})\ln(1/p)\rceil(\ln(1/p)+1) = \Theta((\ln(1/p))^2)$$

Can we get user specified relative error for Poisson (as Bernoulli)?

Gamma Poisson Approximation Scheme

Gamma Poisson Approximation Scheme

Outline

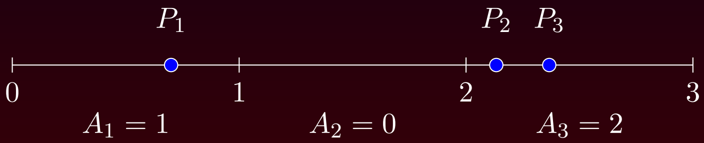
- Can turn stream of Pois(μ) rv's into stream of Exp(μ) rv's.
- Then proceed as with GBAS.

How Poisson to Exponential?

- Use relationship between Poisson process of rate μ and Exponential random variables of rate μ

How to turn Poissons into Exponentials

- Use $A_1, A_2, \ldots$ iid Pois(μ) for Poisson point process rate μ
- Use $A_i \sim$ Pois(μ) to determine how many points in $[i-1, i]$
- Generate points uniformly in interval



- $P_1, P_2 - P_1, P_3 - P_2, \ldots$ iid Exp(μ)

Pseudocode for GPAS

```
Gamma_Poisson_Approximation_Scheme
Input: $k$, $c$    Output: $\hat{\mu}_k$

$A \leftarrow 0, i \leftarrow 0$
While $A < k$                    [Draw $k$ points.]
    $T \leftarrow \text{Pois}(\mu)$
    If $A + T \geq k$            [Then have $k$ points.]
        $T' \leftarrow i + \text{Beta}(k - A, T - (k - A) + 1)$
    $A \leftarrow A + T$, $i \leftarrow i + 1$
$\hat{\mu}_k \leftarrow (k - 1)c/T'$
```

Our integration story so far

Direct Acceptance/Rejection

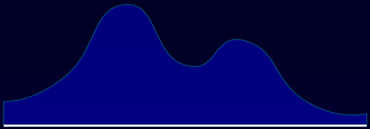
$$F \cdot 1/p$$

TPA

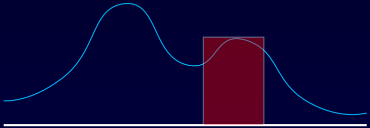
$$F \cdot [\ln(1/p)]^2$$

TPA for general integration using three sets approach

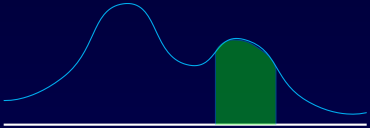
Using TPA for integration, three sets approach



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^*

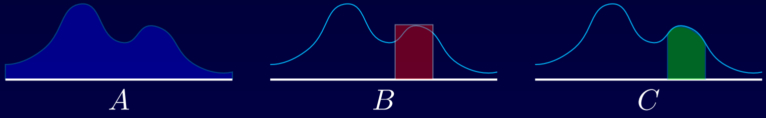


$C = A \cap B$

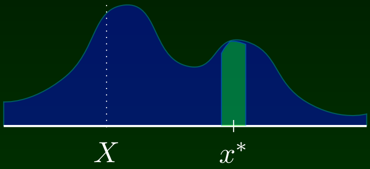
Using the three sets

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

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



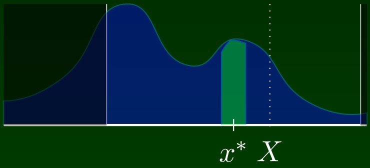
Take several draws from area under A



Set $\beta \leftarrow \infty$

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

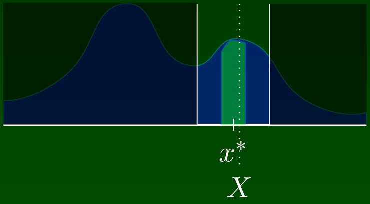
Set $\beta \leftarrow \mathsf{dist}(X, x^*)$



Repeat

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

Set $\beta \leftarrow \mathsf{dist}(X, x^*)$



Until X falls in C

Gibbs distributions

Gibbs distributions

Definition

Say that X has a **Gibbs distribution** if

$$\mathbb{P}(X = x) = \frac{\exp(\beta h(x))}{Z_\beta},$$

where $h(x)$ is called a **Hamiltonian function** and β is the **inverse temperature**.

Examples and behavior of Gibbs

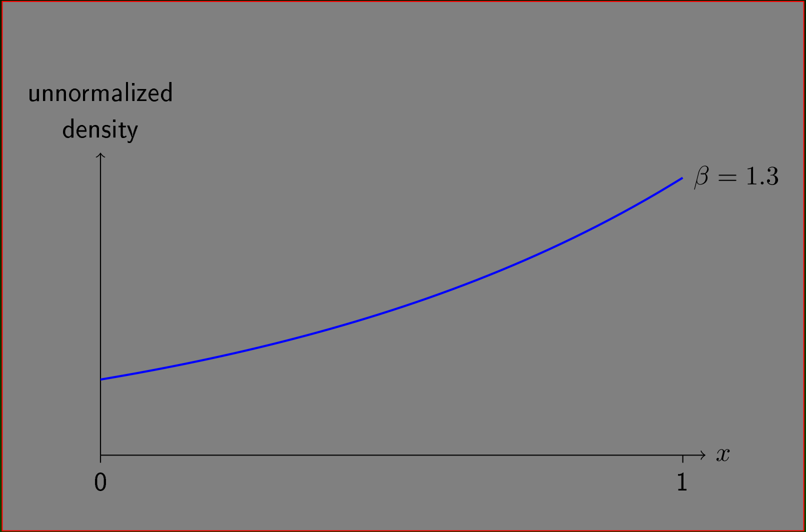
Examples

- Ising model
- Autonormal model
- Potts model

Typical behavior

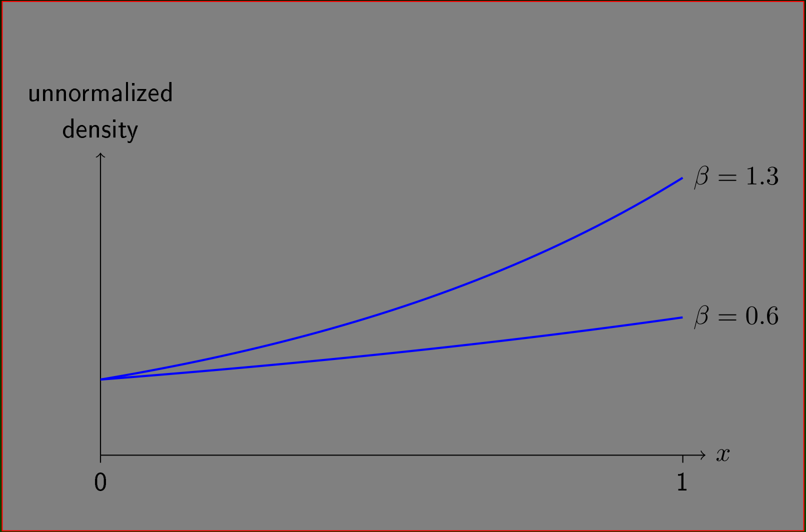
- Easy to draw from X when $\beta = 0$
- Easier to sample from for small values of β
- Can be set up so that $h(x) > 0$ for all x

Gibbs densities for various values of β



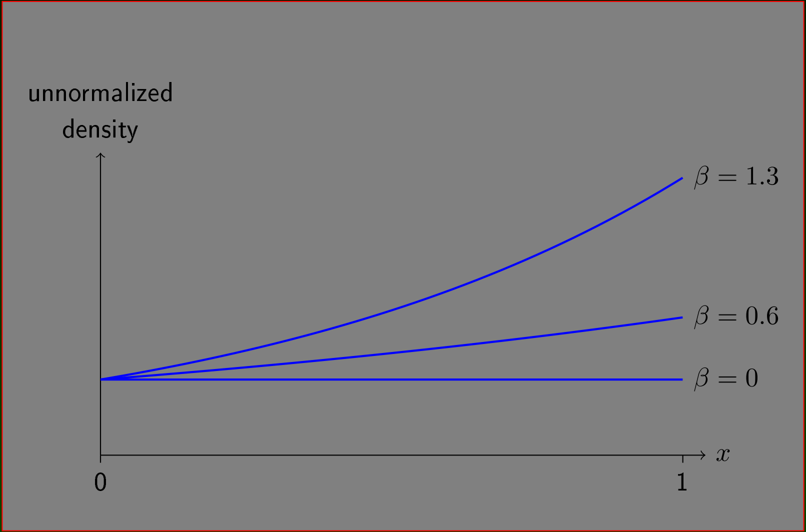
$$h(x) = x$$

Gibbs densities for various values of β



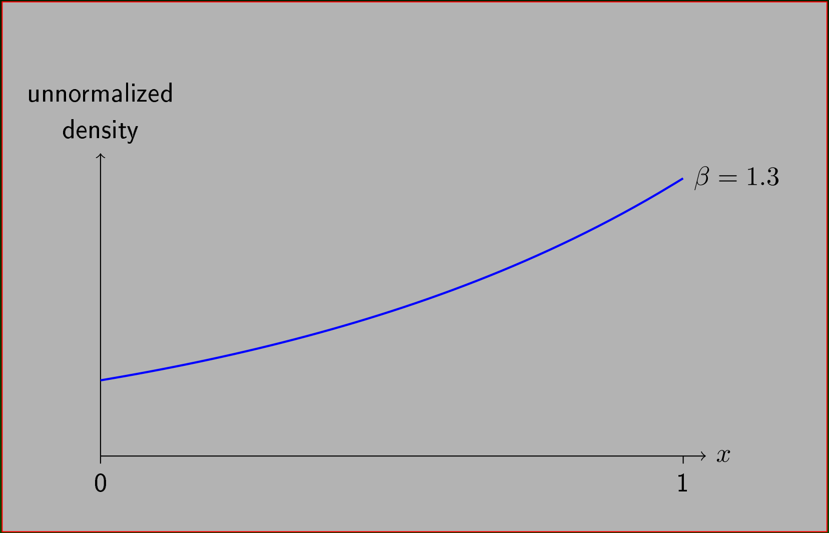
$h(x) = x$

Gibbs densities for various values of β



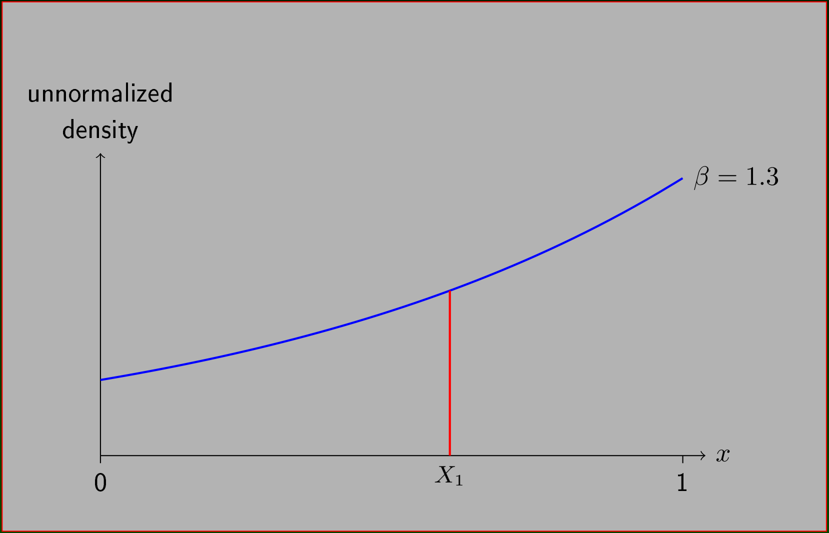
$h(x) = x$

TPA for Gibbs



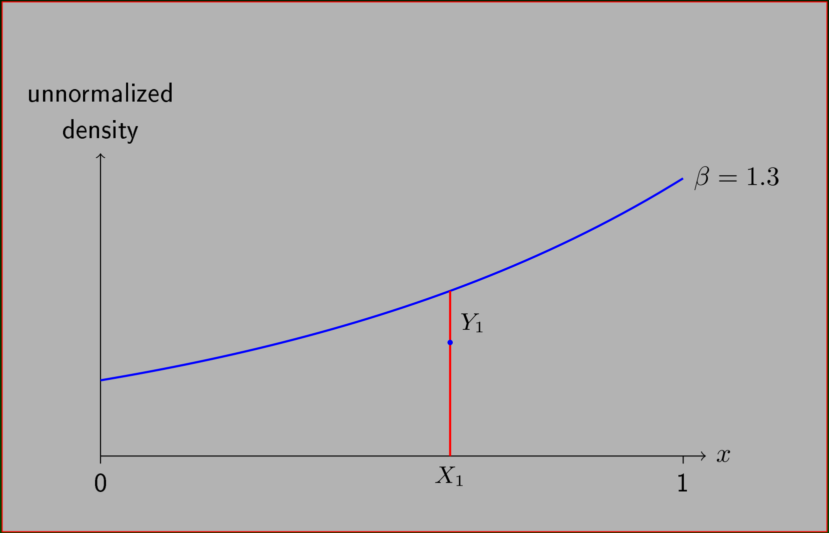
$$h(x) = x$$

TPA for Gibbs



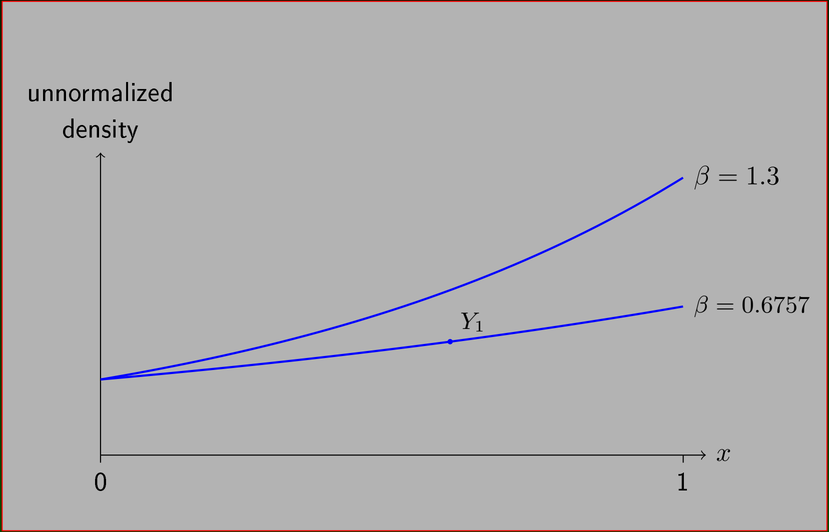
$$h(x) = x$$

TPA for Gibbs



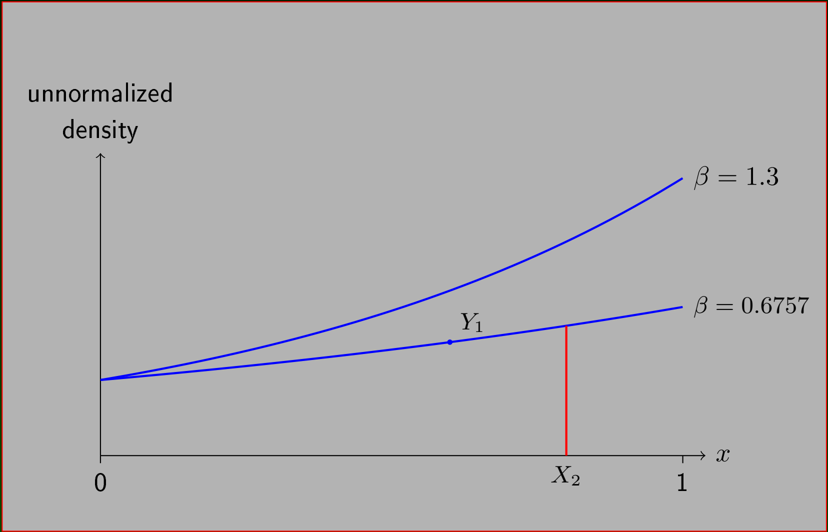
$$h(x) = x$$

TPA for Gibbs



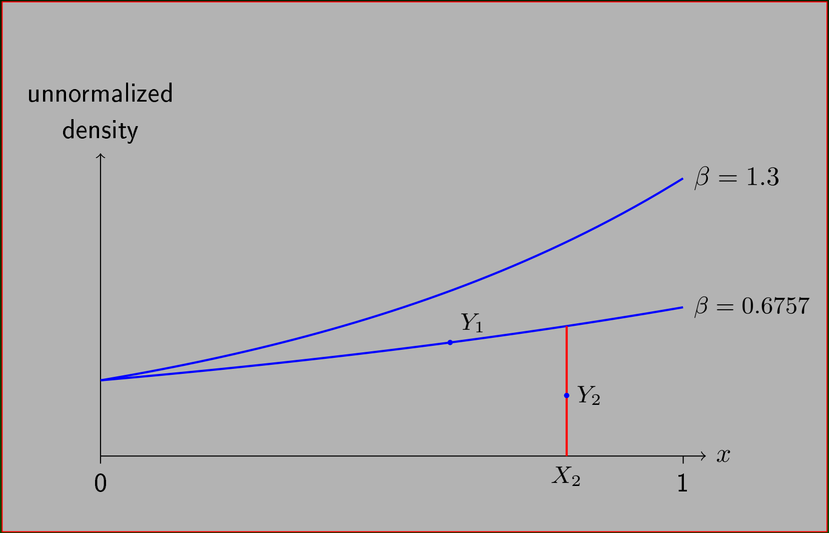
$$h(x) = x$$

TPA for Gibbs



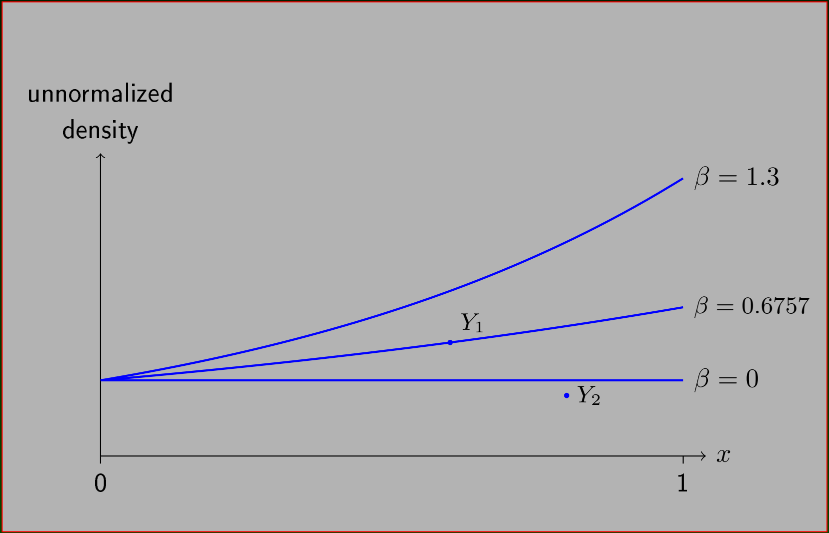
$$h(x) = x$$

TPA for Gibbs



$$h(x) = x$$

TPA for Gibbs



$$h(x) = x$$

TPA for Gibbs distribution

For X drawn from Gibbs, create auxilliary variable

$$[Y|X] \sim \mathsf{Unif}([0, \exp(\beta h(X))])$$

Let $\Omega_\beta = \{(x, y) : y \in [0, \exp(\beta h(x))]\}$, use with TPA

1. Draw X from Gibbs with parameter β
2. Draw Y uniformly from $[0, \exp(\beta h(X))]$
3. Let $\beta' = \inf\{b \geq 0 : (X, Y) \in \Omega_b\} = \ln(\max\{Y, 1\})/h(X)$

Behavior from earlier

For

$$p = Z_0/Z_\beta,$$

needed a number of draws that was about

$$\ln(1/p)^2$$

How big is that?

Typically p exponentially small in dimension of problem

- For Ising/autonormal/Strauss model, number of nodes in graph
- So $\ln(1/p)^2 = \Omega(n^2)$
- Too large to be effective in this case
- Use importance sampling to speed things up

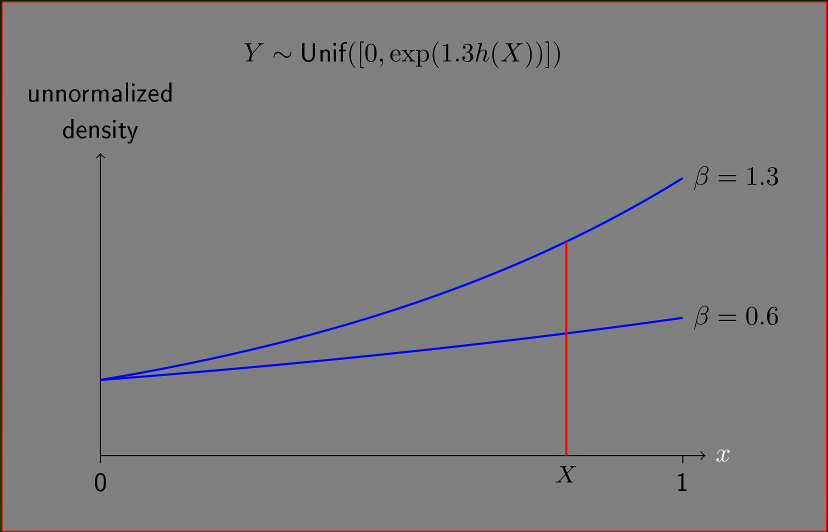
Importance sampling

Replacing random choices with means

Rao-Blackwellization

- In last step of TPA for Gibbs, draw Y uniformly over $[0, \exp(\beta h(X))]$, then choose new β
- Instead, start with new β, find probability that Y falls below new β

Illustration of importance sampling for Gibbs



$$\mathbb{P}(Y \leq \exp(0.6h(X)) = \exp(0.6h(X))/\exp(1.3h(X))$$

Moving from Z_0 to Z_β

Cooling schedule

- Connect Z_β to Z_0 using different values of β

$$0 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_k = \beta$$

- Then can multiply ratios to get target ratios

$$\frac{Z_\beta}{Z_0} = \frac{Z_{\beta_k}}{Z_{\beta_{k-1}}} \cdot \frac{Z_{\beta_{k-1}}}{Z_{\beta_{k-2}}} \cdot \frac{Z_{\beta_{k-2}}}{Z_{\beta_{k-3}}} \cdots \frac{Z_{\beta_1}}{Z_{\beta_0}}.$$

Product estimator

- Estimate each $r_k = Z_{\beta_k}/Z_{\beta_{k-1}}$ by $\hat{r}_k$ [3]
- Final estimtae is

$$\hat{r} = \hat{r}_1 \hat{r}_2 \cdots \hat{r}_k$$

- Called **product estimator** by Fishman [4]

[3] M. Jerrum and L. Valiant and V. Vazirani, Random generation of combinatorial structures from a uniform distribution, *Theoret. Comput. Sci.*, 43:169–188, 1986

[4] G. S. Fishman, *Monte Carlo: concepts, algorithms, and applications*, Springer-Verlag, New York, 1996

Using for Gibbs

- For Gibbs, using a well balanced cooling schedule makes variance in product estimator small
- Each $r_i \approx c \in [0, 1]$
- Basic approach difficult to analyze, but possible [5]
- Variant approach is the Paired Product Estimator [6]

[5] D. Štefankovič and S. Vempala and E. Vigoda, Adaptive Simulated Annealing: A Near-Optimal Connection between Sampling and Counting, *J. of the ACM*, 56(3):1–36,2009

[6] M. Huber, Approximation algorithms for the normalizing constant of Gibbs distributions, *Ann. Appl. Probab.*, arXiv:1206.2689, 51(1):92–105, 2015

Paired Product Estimator running time

Theorem

By employing the Paired Product Estimator we get an (ϵ, δ)-ras using about

$$2\epsilon^{-2}\ln(\delta^{-1})\ln(1/p)$$

samples on average.

Basic TPA approach to Gibbs

$$2\epsilon^{-2}\ln(\delta^{-1})[\ln(1/p)]^2$$

samples on average

Can I use inexact samples with these methods?

POBODY’S
NERFECT

What if I do not have a perfect simulator?

- Can still use AR, TPA, or PPE
- Normally two sources of mistakes with Monte Carlo integration:
  1. Samples are inexact
  2. Integration method has unknown variance
- With AR, TPA, or PPE only one source of error
  1. Samples are inexact

Summary

- AR integration with GBAS

$$F \cdot (1/p)$$

- TPA integration with GPAS

$$F \cdot [\ln(1/p)]^2$$

- Paired Product Estimator/IS for Gibbs

$$F \cdot \ln(1/p)$$