Approximation of normalizing constants using random cooling schedules

Mark Huber[1] and Sarah Schott[2]

[1]Department of Mathematics and Computer Science, Claremont McKenna College

[2]Department of Mathematics, Duke University

17 Apr, 2010

The art of numerical integration

In every block of marble I see a statue as plain as though it stood before me, shaped and perfect in attitude and action. I have only to hew away the rough walls that imprison the lovely apparition to reveal it to the other eyes as mine see it.
-Michelangelo

The integration problem

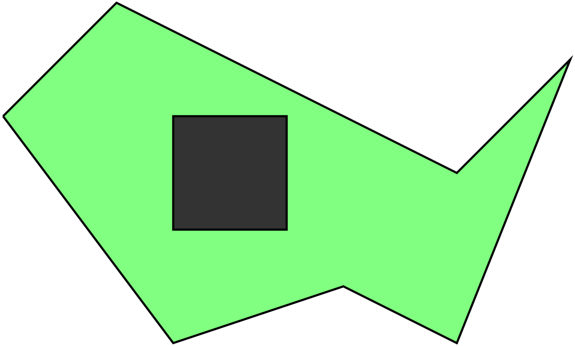
What is:

Volume of 

Volume of 

My version of the block and sculpture

What is the measure of a set?



Green area = B

Black area = B'

$$\mu(B) = \mu(B')\frac{\mu(B)}{\mu(B')}$$

Example applications

Integration:

$$\mu(B) = \int_{\vec{x} \in B} f(\vec{x})\, d\vec{x}, \text{ where } f(\vec{x}) \geq 0$$

- Volume of a convex set
- Normalizing constant for unnormalized density

Summation:

$$\mu(B) = \sum_{i \in A} w(i), \text{ where } w(i) \geq 0$$

- Normalizing constant for Ising model
- Permanent of a matrix with nonnegative entries

Bayesian statistical applications

$$\text{posterior density} \propto \text{prior density} \times \text{likelihood}$$

Normalizing posterior

- B =parameter space, μ proportional to posterior
- $\mu(B)$ is integration likelihood/evidence
- Appears in Bayes Factors for model selection

Posterior mean of nonnegative parameter θ

- A =parameter space, μ has density θ against posterior
- $\mu(B) = \mathbb{E}[\theta]$

Spatial statistics

- For likelihoods like Ising model, need normalizing constant before can build posterior
- Often called doubly intractable

Variance dependent methods

Classical approach: first write $\mu(B) = \mathbb{E}[X]$

- Draw $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} X$
- $\mu(B) \approx \bar{X} = \frac{1}{n} \sum_i X_i$
- $\mathbb{E}[\bar{X}] = \mu(B)$, $\mathbb{V}(\bar{X}) = (1/n)\mathbb{V}(X_i)$ could be huge

Today

- Variance free estimation
- No need to calculate or estimate a variance

Variance Free Monte Carlo

Randomized Approximation Algorithms

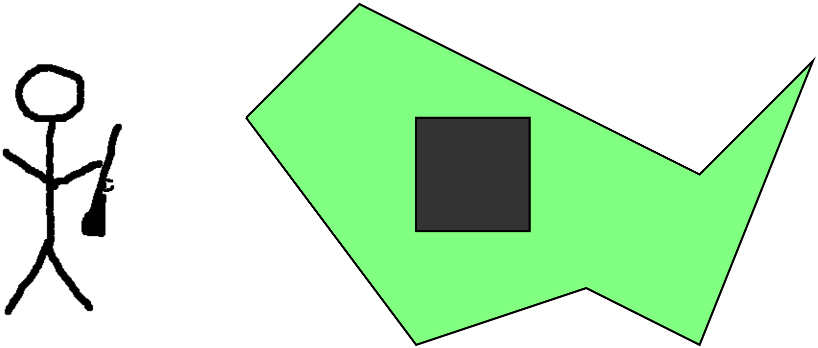
Definition

Let A^* be the output of an algorithm when A is the true answer. Then the algorithm is an (ϵ, δ)-randomized approximation algorithm if

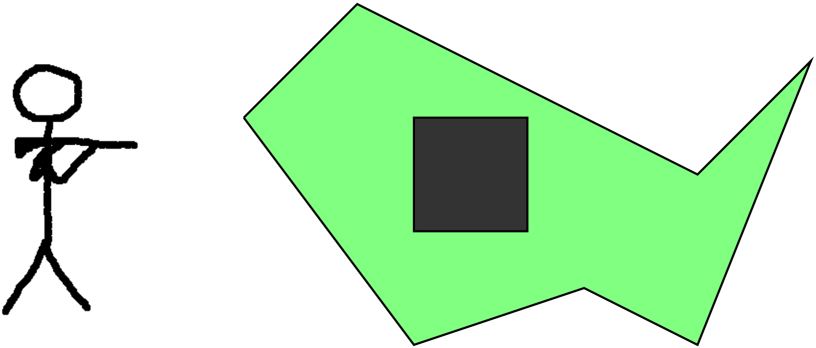
$$\mathbb{P}((1+\epsilon)^{-1} \leq A^*/A \leq 1+\epsilon) \geq 1-\delta.$$

Goal is rand. approx. alg. for $\mu(A)$ for all positive ϵ and δ

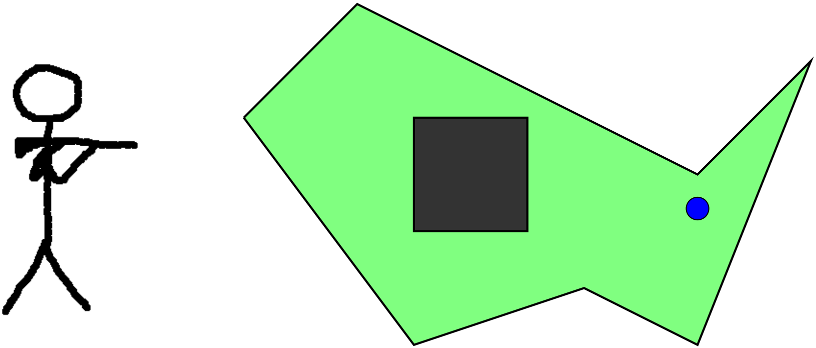
Acceptance/Rejection a.k.a "Shoot at it randomly"



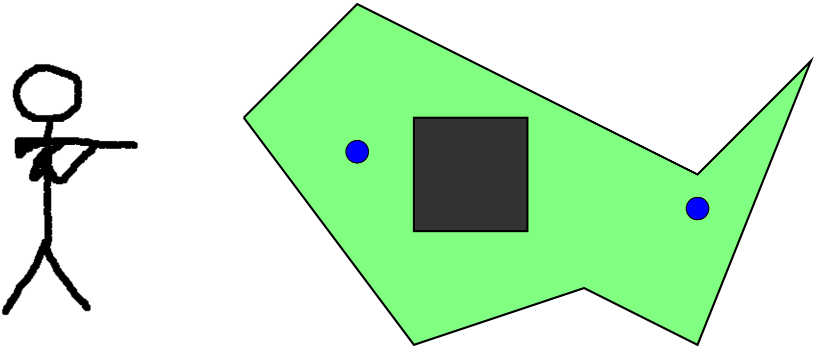
Acceptance/Rejection a.k.a “Shoot at it randomly”



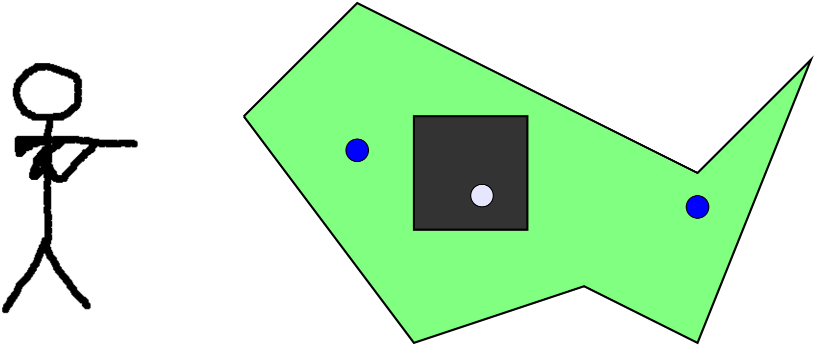
Acceptance/Rejection a.k.a “Shoot at it randomly”



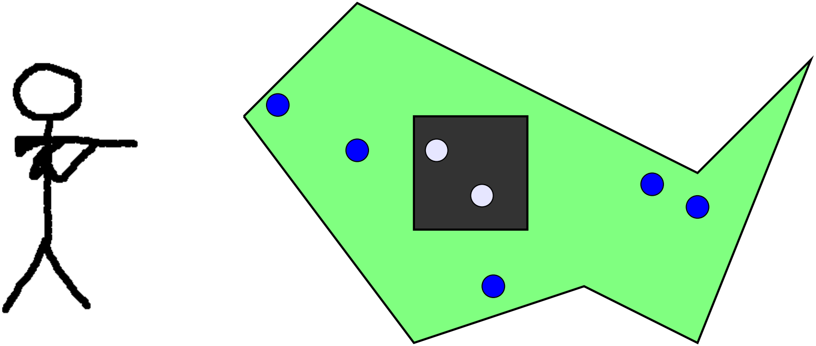
Acceptance/Rejection a.k.a "Shoot at it randomly"



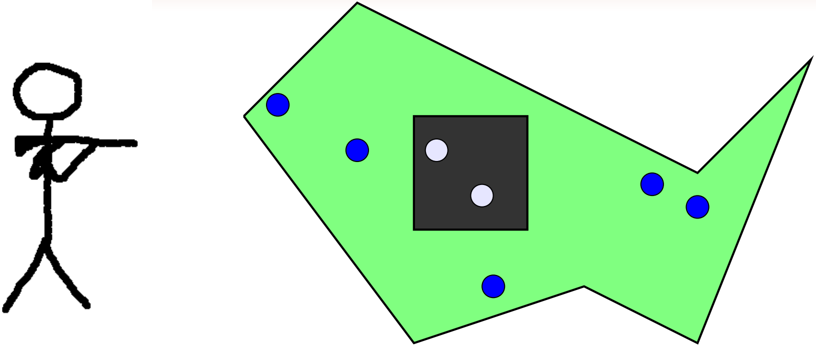
Acceptance/Rejection a.k.a "Shoot at it randomly"



Acceptance/Rejection a.k.a "Shoot at it randomly"



Acceptance/Rejection a.k.a "Shoot at it randomly"



Best estimate:

$$\hat{\mu}(B) = \frac{7}{2}\mu(B'), \quad B' = \text{black rectangle inside region}$$

Analysis of Accept/Reject

Algorithm

- Fire n times at box
- Say you hit H times

Analysis

- Let $p =$ chance hit center region B'
- $H \sim \text{Bin}(n, p)$

$$p = \frac{\mu(B')}{\mu(B)}$$

- Estimate

$$\hat{p} = \frac{H}{n}, \quad \hat{\mu}(B) = \frac{n}{H}\mu(B');$$

How well does it work?

True answer: 3.0933...

- After 10 iterations 5.04
- After 10^3 iterations 2.8656
- After 10^5 iterations 3.0870
- After 10^7 iterations 3.0935

About a factor of 100 per extra digit

How well does it work?

True answer: 3.0933...

- After 10 iterations 5.04
- After 10^3 iterations 2.8656
- After 10^5 iterations 3.0870
- After 10^7 iterations 3.0935

About a factor of 100 per extra digit

How well does it work?

True answer: 3.0933...

- After 10 iterations 5.04
- After 10^3 iterations 2.8656
- After 10^5 iterations 3.0870
- After 10^7 iterations 3.0935

About a factor of 100 per extra digit

How well does it work?

True answer: 3.0933...

- After 10 iterations 5.04
- After 10^3 iterations 2.8656
- After 10^5 iterations 3.0870
- After 10^7 iterations 3.0935

About a factor of 100 per extra digit

How well does it work?

True answer: 3.0933...

- After 10 iterations 5.04
- After 10^3 iterations 2.8656
- After 10^5 iterations 3.0870
- After 10^7 iterations 3.0935

About a factor of 100 per extra digit

Bounding error

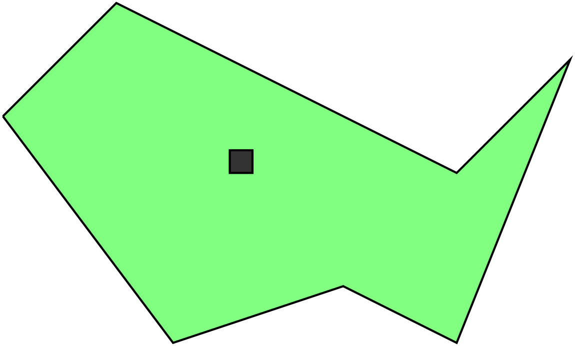
Get relative error below ϵ with probability at least $1-\delta$:

- Need to analyze tails of binomial distribution to show:

$$n \approx \frac{2(1+\epsilon)}{p} \cdot \frac{1}{\epsilon^2} \cdot \ln \frac{1}{\delta}$$

- The $(1/\epsilon^2)\ln(1/\delta)$ often called "Monte Carlo" error
- Best you can do in general
- So concentrate on improving $1/p$ part

When p small, runtime large



Usually p exponentially small in dimension of problem

Running times

Acceptance/Rejection:

$$2 \cdot \frac{1}{p}.$$

Product Estimator [1]:

$$192 \cdot \left[\log \frac{1}{p}\right]^2.$$

Goal for TPA:

- Get $[\log 1/p]^2$ performance
- With decent constant out in front

Nested Sampling [Skilling 2007]

Nested sampling another approach to these integrals

- Mix of product estimator-like algorithm and classical 1-D numerical integration
- Not quite approximation algorithm
- Roughly speaking also $[\log(1/p)]^2$
- Does introduce a nice idea
- Combination nice idea + product estimator = TPA

The New Algorithm

TPA

Idea

- Product estimator...
- ...plus idea from Nested sampling

Result

- Product estimator with random cooling schedule
- Output can be analyzed exactly (like A/R)

The Tootsie Pop Algorithm

What is a Tootsie Pop?

- Hard candy lollipops with a tootsie roll (chewy chocolate) at the center

In 1970, Mr. Owl was asked the question:

- How many licks does it take to get to the center of a Tootsie Pop?

List of ingredients of TPA

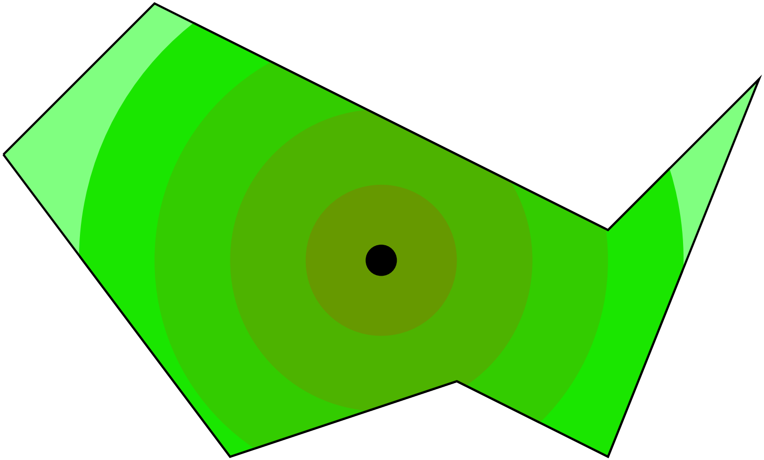
(a) A measure space $(\Omega, \mathcal{F}, \mu)$

(b) Two measurable sets: the *center* B' and the *shell* B with $B' \subset B$

(c) A family of sets $\{A(\beta)\}$ where
  1. $\beta' < \beta$ implies $A(\beta') \subseteq A(\beta)$,
  2. $\mu(A(\beta))$ is continuous in β

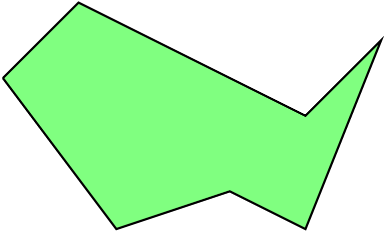
(d) Two special values β_B and $\beta_{B'}$ with $A(\beta_B) = B$ and $A(\beta_{B'}) = B'$.

Example of nested sets



$A(\beta) =$ all points within distance β of center

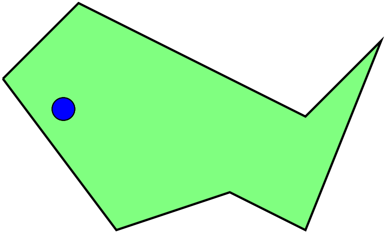
Idea behind TPA



Step	β
0	∞
1	
2	
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

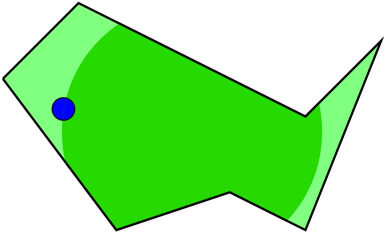
Idea behind TPA



Step	β
0	∞
1	
2	
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

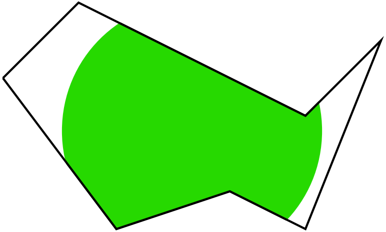
Idea behind TPA



Step	β
0	∞
1	1.72
2	
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

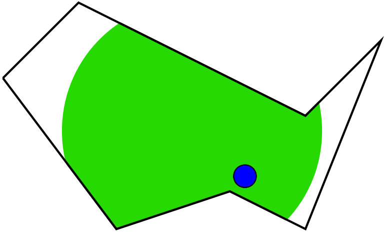
Idea behind TPA



Step	β
0	∞
1	1.72
2	
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

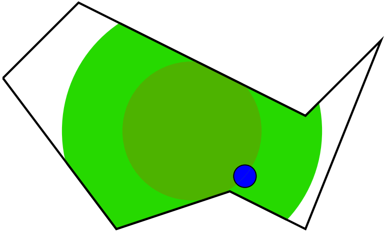
Idea behind TPA



Step	β
0	∞
1	1.72
2	
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

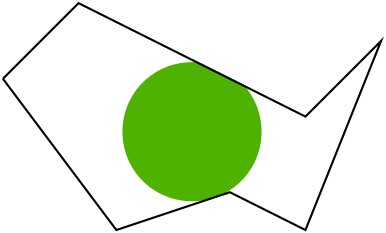
Idea behind TPA



Step	β
0	∞
1	1.72
2	0.92
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

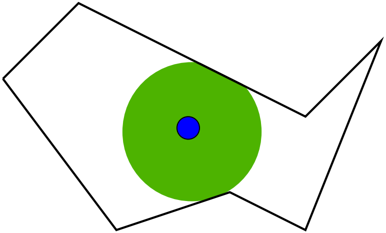
Idea behind TPA



Step	β
0	∞
1	1.72
2	0.92
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

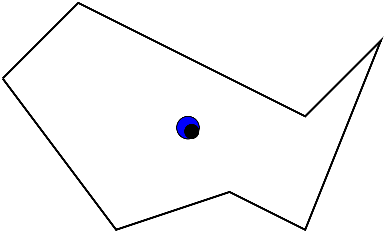
Idea behind TPA



Step	β
0	∞
1	1.72
2	0.92
3	

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

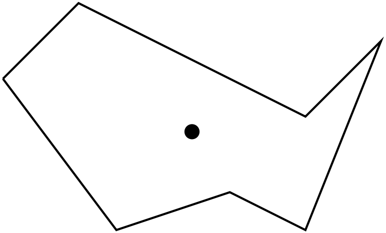
Idea behind TPA



Step	β
0	∞
1	1.72
2	0.92
3	0.09

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

Idea behind TPA



Step	β
0	∞
1	1.72
2	0.92
3	0.09

1. $\beta \leftarrow \beta_B$
2. Repeat
3. Draw $X \leftarrow \mu(A(\beta))$
4. $\beta \leftarrow \inf\{\beta' : X \in A(\beta')\}$
5. Until $\beta \leq \beta_{B'}$

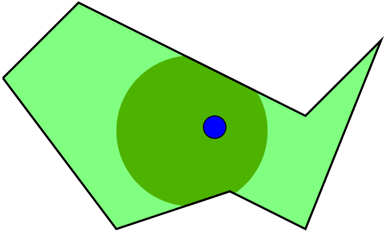
How much is shaved off at each step?

Notation: $Z(\beta) := \mu(A(\beta))$

Lemma

Say $X \sim \mu(A(\beta))$ *and* $\beta' = \min\{\beta' : X \in A(\beta')\}$. *Then*

$$\frac{Z(\beta')}{Z(\beta)} \sim \text{Unif}([0,1])$$



Proof by picture:

Let b satisfy $Z(b)/Z(\beta) = 1/3$

Then

$$\mathbb{P}\left(\frac{Z(\beta')}{Z(\beta)} \leq 1/3\right) = \mathbb{P}(X \in A(b))$$

Each step removes on average 1/2 the measure

Continue until reach center

- Original measure $\mu(B) = Z(\beta_B)$
- After k steps measure $Z(\beta_k) = Z(\beta_B) r_1 r_2 \cdots r_k$, where $r_i \overset{\text{iid}}{\sim} \text{Unif}([0,1])$
- Recall $\beta_{B'}$ index of center
- Let ℓ be number of steps until hit center

$$\ell := \min\{k : Z(\beta_B) r_1 \cdots r_k < Z(\beta_{B'})\} - 1$$

Question: what is the distribution of ℓ?

Logarithms

Recall if $U \sim \text{Unif}([0,1])$**,**

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

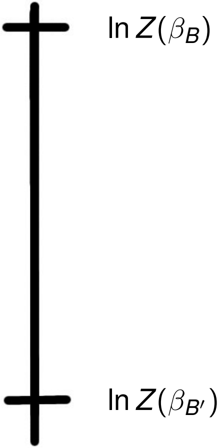
Since

$$\frac{Z(\beta_k)}{Z(\beta_B)} \sim r_1 r_2 \cdots r_k, \text{ where } r_i \overset{\text{iid}}{\sim} \text{Unif}([0,1]),$$

Consider the points

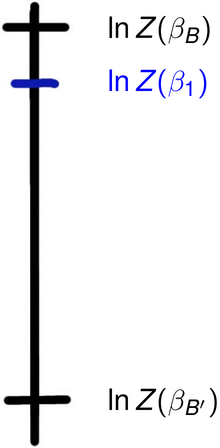
$$P_i = -\ln\left(\frac{Z(\beta_k)}{Z(\beta_B)}\right) \sim e_1 + e_2 + \cdots + e_k, \text{ where } e_i \overset{\text{iid}}{\sim} \text{Exp}(1)$$

The Poisson point process



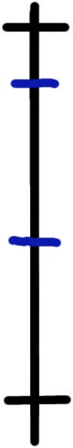
- Better to work in $\ln Z(\beta_i)$ space
- Recall: $U \sim \text{Unif}([0,1]) \Rightarrow -\ln U \sim \text{Exp}(1)$
- In log space, each step moves down $\text{Exp}(1)$
- Result: a Poisson point process from $\ln Z(\beta_B)$ to $\ln Z(\beta_{B'})$

The Poisson point process



- Better to work in $\ln Z(\beta_i)$ space
- Recall: $U \sim \text{Unif}([0,1]) \Rightarrow -\ln U \sim \text{Exp}(1)$
- In log space, each step moves down Exp(1)
- Result: a Poisson point process from $\ln Z(\beta_B)$ to $\ln Z(\beta_{B'})$

The Poisson point process



$\ln Z(\beta_B)$

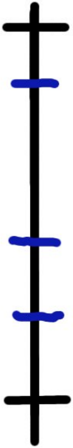
$\ln Z(\beta_1)$

$\ln Z(\beta_2)$

$\ln Z(\beta_{B'})$

- Better to work in $\ln Z(\beta_i)$ space
- Recall: $U \sim \text{Unif}([0,1]) \Rightarrow -\ln U \sim \text{Exp}(1)$
- In log space, each step moves down Exp(1)
- Result: a Poisson point process from $\ln Z(\beta_B)$ to $\ln Z(\beta_{B'})$

The Poisson point process



$\ln Z(\beta_B)$

$\ln Z(\beta_1)$

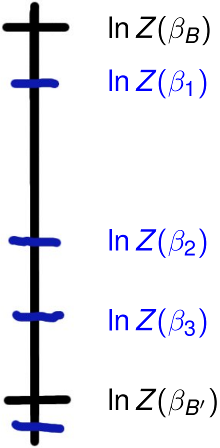
$\ln Z(\beta_2)$

$\ln Z(\beta_3)$

$\ln Z(\beta_{B'})$

- Better to work in $\ln Z(\beta_i)$ space
- Recall: $U \sim \text{Unif}([0, 1]) \Rightarrow -\ln U \sim \text{Exp}(1)$
- In log space, each step moves down Exp(1)
- Result: a Poisson point process from $\ln Z(\beta_B)$ to $\ln Z(\beta_{B'})$

The Poisson point process



- Better to work in $\ln Z(\beta_i)$ space
- Recall: $U \sim \text{Unif}([0, 1]) \Rightarrow -\ln U \sim \text{Exp}(1)$
- In log space, each step moves down Exp(1)
- Result: a Poisson point process from $\ln Z(\beta_B)$ to $\ln Z(\beta_{B'})$

The result

Output of TPA:

- $\ell \sim \text{Pois}(\ln(Z(\beta_B)/Z(\beta_{B'})))$

Output of A/R:

- $H \sim \text{Bin}(n, Z(\beta_{B'})/Z(\beta_B))$

Repeating the Poisson point process

Suppose run the Poisson point process twice

- Result also Poisson point process rate 2 instead of rate 1

Now run k times

- Result also Poisson point process rate k instead of rate 1
- Final answer $\text{Pois}(k \ln(Z(\beta_{B'})/Z(\beta_B)))$
- Divide by k, result close to $\ln[Z(\beta_{B'})/Z(\beta_B)]$
- Exponentiate, result close to $Z(\beta_{B'})/Z(\beta_B)$
- Can use Chernoff's Bound to choose k large enough

Bounding the tails

Theorem

Let $p = Z(\beta_{B'})/Z(\beta_B)$*. For* $p < \exp(-1)$ *and* $\epsilon < .3$*, after*

$$k = 2\left[\ln \frac{1}{p}\right]\left(\frac{3}{\epsilon} + \frac{1}{\epsilon^2}\right)\ln \frac{1}{2\delta}$$

runs, each of which uses on average $\ln(1/p)$ *samples, the output* $\hat{p}$ *satisfies:*

$$\mathbb{P}((1+\epsilon)^{-1} \leq \hat{p}/p \leq 1+\epsilon) > 1-\delta.$$

Bonus: Approximate for all parameters simultaneously

Can cut Poisson point process at any point:

Right half still Poisson point process
Yields omnithermal approximation

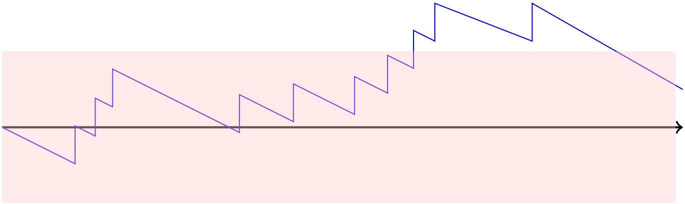
- Approximate $Z(\beta)/Z(\beta_{B'})$ for all $\beta \in [\beta_{B'}, \beta_B]$ at same time
- Number of runs still same:

$$k = 2 \left[\ln \frac{1}{p} \right] \left(\frac{3}{\epsilon} + \frac{1}{\epsilon^2} \right) \ln \frac{1}{2\delta}$$

Proof idea:

Poisson process

- Let $N(t)$ be rate r Poisson process
- $N(t) - rt$ is a right continuous martingale
- Omnithermal approximation valid means did not drift too far away from 0



Examples

Example 1: Mixture Gaussian Spikes

Multimodal toy example

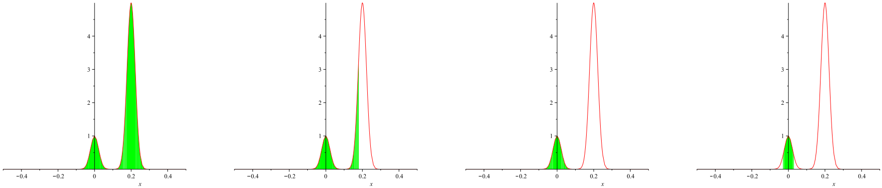
- Prior uniform over cube
- Likelihood mixture of two normals
- Small spike centered at $(0, 0, \ldots, 0)$
- Large spike centered at $(0.2, 0.2, \ldots, 0.2)$

$$p_\theta \sim \text{Unif}([-1/2, 1/2]^d)$$

$$L(\theta) = 100 \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi} u} \exp\left(-\frac{(\theta_i - 0.2)^2}{2u^2}\right) + \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi} v} \exp\left(-\frac{\theta_i^2}{2v^2}\right)$$

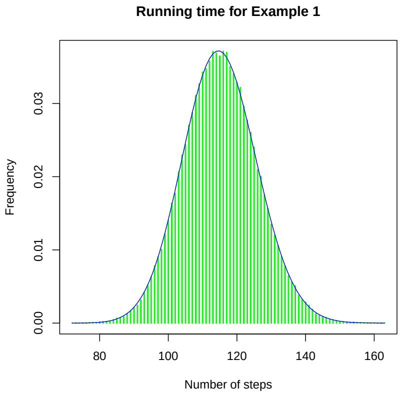
Parameter truncation

Create family by limiting distance to center of small spike



Running time results

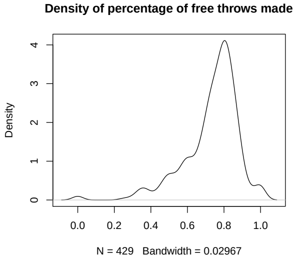
Problem: $d = 20$, $u = .01$, $v = .02$
True value: $\ln(1/p) = 115.0993$
Algorithm (10^5 runs): $\ln(1/p) \approx 115.10321$



Example 2: Beta-binomial model

Hierarchical model

- Data set: free throw numbers for 429 NBA players '08-'09
- Example data point: Kobe Bryant made 483 out of 564
- Model: number made by player i is $\text{Bin}(n_i, p_i)$
- n_i are known, $p_i \sim \text{Beta}(a, b)$
- Hyperparameters a and b, $a \sim 1 + \text{Exp}(1)$, $b \sim 1 + \text{Exp}(1)$



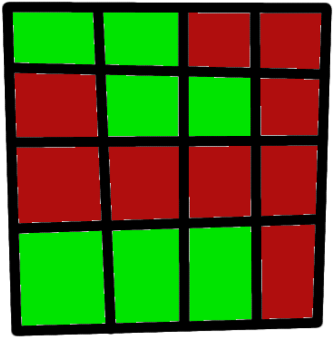
Again use parameter trunctation

Goal: find integrated likelihood

- Use β to limit distance from mode
- 2-D Unimodal problem so sampling easy
- True value (via numerical integration) -1577.250
- After 10^5 runs -1577.256

Example 3: Ising model

Besag[1974] modeled soil plots as good (green) or bad (red)



$h(x) = 13$ (# adj like colored plots)

$$\pi(x) = \frac{\exp(2\beta h(x))}{Z(\beta)}$$

parameter β is inv temp

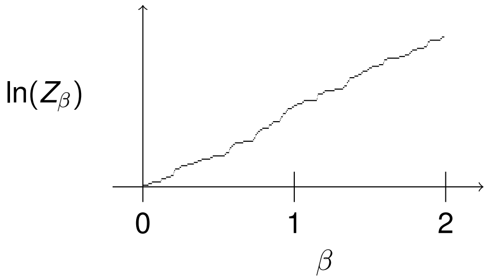
Integrated likelihood for Ising

Parameter space one dimensional

$$Z = \int_0^\infty p_\beta(b) \frac{\exp(2bh(x))}{Z(b)}\ db,$$

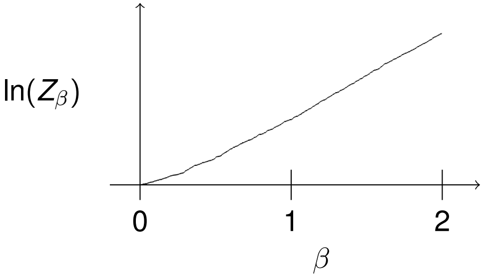
easy to do numerically if you know $Z(\beta)$ over $(0, \infty)$.

Use omnithermal approximation



One run of TPA

Use omnithermal approximation



Sixteen runs of TPA

Connection to MCMC

Several sampling methods use temperatures

- Simulated annealing
- Simulated tempering

TPA easy for these problems

- Can speed up chain by giving well balanced cooling schedule

Another approach: Likelihood truncation

Works well when slice sampler works well
Let T bound the likelihood

$$Z(T) = \int \pi(\theta) \min\{T, L(\theta)\}\, d\theta.$$

Starting point

- $Z(\infty) = Z$, that is the starting point
- But where is the center?

Locate center separately

Draw k samples from π

- Get k different likelihood values
- Let m be median of these values

Now draw k' samples $\theta_1, \ldots, \theta_{k'}$ from π

- Accept θ_i with probability $\min\{1, L(\theta_i)/m\}$
- (Like Metropolis-Hastings)

Getting the center

What is probability of accepting?

$$\int_\theta \pi(\theta) \min\left\{1, \frac{L(\theta)}{m}\right\} \, d\theta = \frac{1}{m} \int_\theta \pi(\theta) \min\{m, L(\theta)\} \, d\theta = \frac{Z(m)}{m}$$

Idea:

- Use this A/R method to estimate $Z(m)$
- Since m was median, probability of acceptance at least about 1/2
- So can be used to get good approximation of $Z(m)$ quickly
- This way, $Z(m)$ becomes our center

Numerical example

Draw 11 random variates iid from π:

$$(\theta_1, \ldots, \theta_{11}) = (.37, .27, .72, .52, .33, .90, .07, .52, .05, .03, .60)$$

Plug into $L(\theta)$:

$$(L(\theta_1), \ldots, L(\theta_{11})) = (\boxed{.13}, .07, .52, .27, .11, .81, \ldots, .36)$$

Repeat 1000 times:

- Draw $\theta \sim \pi$, accept w/ prob. $\min\{1, L(\theta)/.13\}$

Suppose accept 650 times:

- $Z(.13) \approx \frac{650}{1000}(.13)$

Apples and Oranges

Comparisons

TPA draws ideas from...

- The Product Estimator
- Nested Sampling

Product Estimator

Cooling schedule fixed

$$\beta_{B'} = \beta_0 < \beta_1 < \beta_2 < \cdots < \beta_k = \beta_B$$

- Make sure $\mu(A(\beta_i))/\mu(A(\beta_{i+1}))$ is bounded away from 0
- Using basic A/R, find:

$$\hat{p}_i \approx \frac{\mu(A(\beta_i))}{\mu(A(\beta_{i+1}))}$$

- Take the product of individual estimates:

$$\hat{p} = \prod_{i=0}^{k-1} \hat{p}_i \approx \frac{\mu(A(\beta_0))}{\mu(A(\beta_k))}$$

Example product estimator

Suppose

$$\frac{\mu(A(\beta_0))}{\mu(A(\beta_1))} = .6, \ \frac{\mu(A(\beta_1))}{\mu(A(\beta_2))} = .4, \ \frac{\mu(A(\beta_2))}{\mu(A(\beta_3))} = .8,$$

Then

$$\frac{\mu(A(\beta_0))}{\mu(A(\beta_3))} = \frac{\mu(A(\beta_0))}{\mu(A(\beta_1))} \cdot \frac{\mu(A(\beta_1))}{\mu(A(\beta_2))} \cdot \frac{\mu(A(\beta_2))}{\mu(A(\beta_3))}$$

Output of product estimator

Final estimate

- Scaled product of binomial distributions
- Finding std. dev. easy, bounding tails is hard

Cooling schedule

- Need new schedule for every problem
- Works best when $\mu(A(\beta_i))/\mu(A(\beta_{i+1}))$ same for all i
- Difficult to do: if knew how to do that, wouldn't need the product estimator

Comparison to nested sampling

Why use TPA instead of nested sampling?

- Running time same order as nested sampling
- Output distribution known exactly
- Sets in nested sampling tend to exaggerate multimodality
- TPA tends to remove it
- Distributions usually easier (moving towards center) in later steps
- No unknown derivatives in error bounds

Conclusions

New algorithm: TPA

- Guaranteed performance bounds on Monte Carlo integration
- (No variance estimate or unknown derivatives appear)
- Speed: $2[\ln(1/p)]^2$ much better than product estimator

Future directions

- Convex sets and exponential families $\ln(1/p)$ algorithms exist
- Convex sets: pedestal method
- Exponential families: recursive adaptive scheduling
- Account for Rao-Blackwell-ization of estimate

References

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

M. Huber and S. Schott
Using TPA for Bayesian inference
Bayesian Statistics 9

J. Skilling,
Nested sampling for general Bayesian computation,
Bayesian Anal., **1**(4), 833–860, 2006