Fast approximation algorithms for partition functions of Gibbs distributions

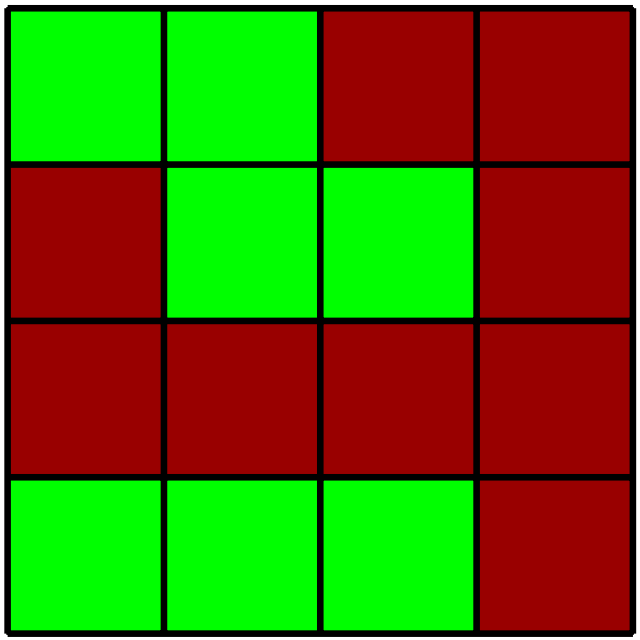
Mark Huber
Department of Mathematical Sciences
Claremont McKenna College

24 May, 2012

The Problem

Autonormal (Ising) model

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



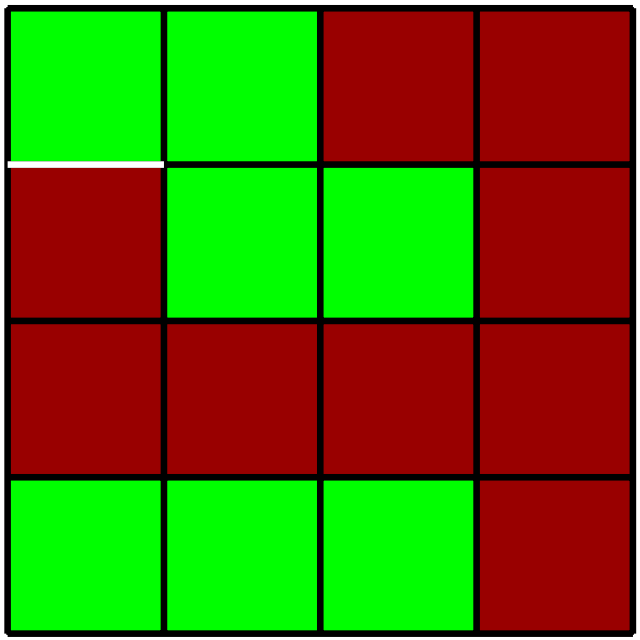
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



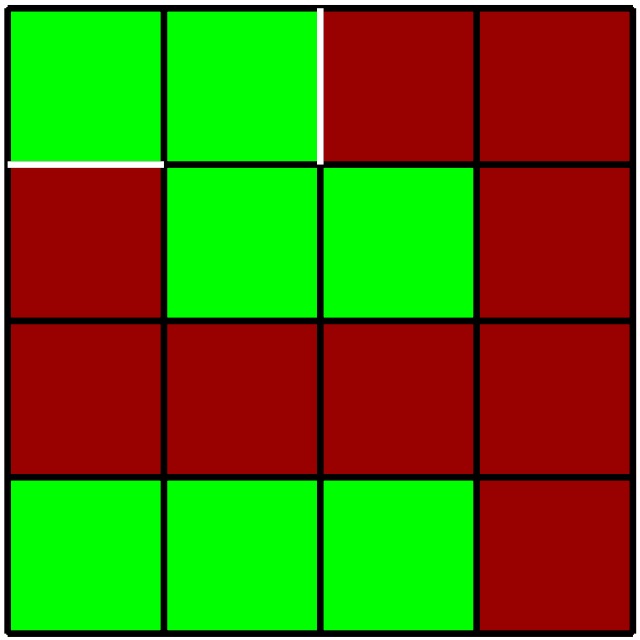
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



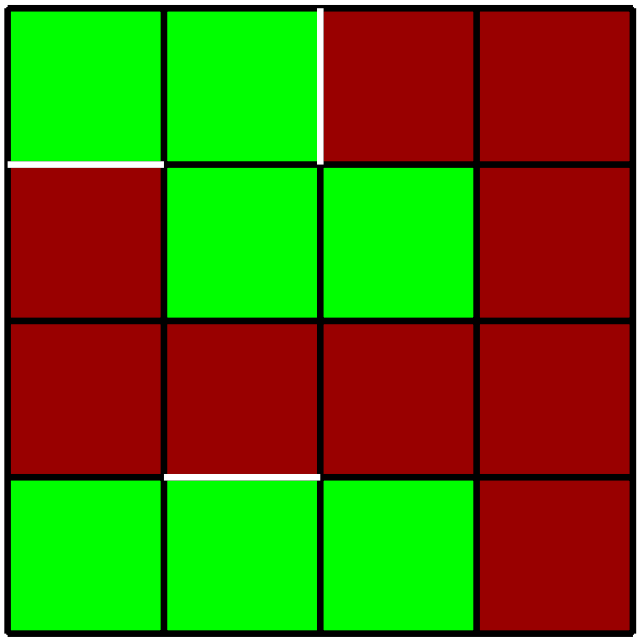
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



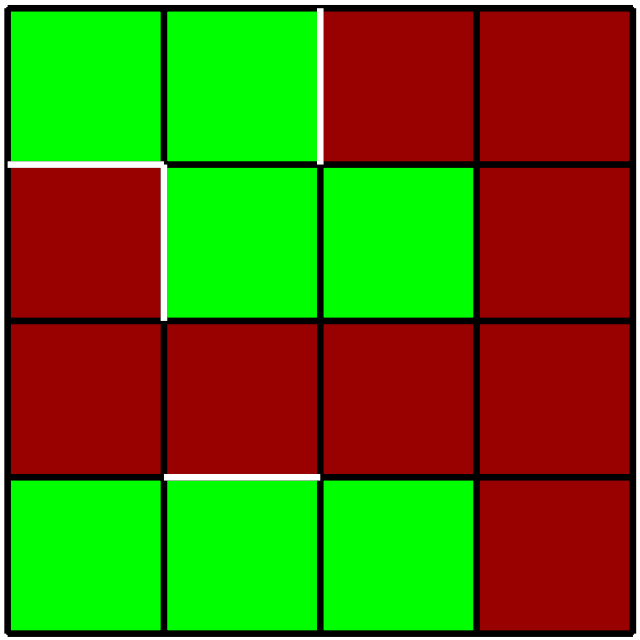
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



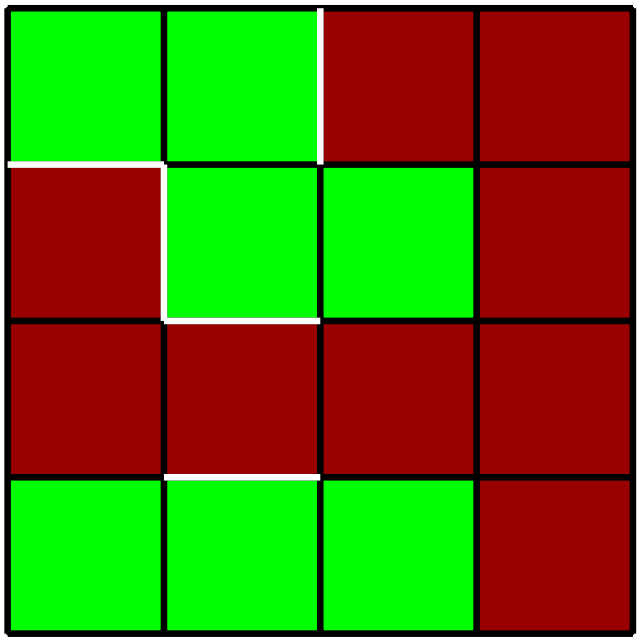
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



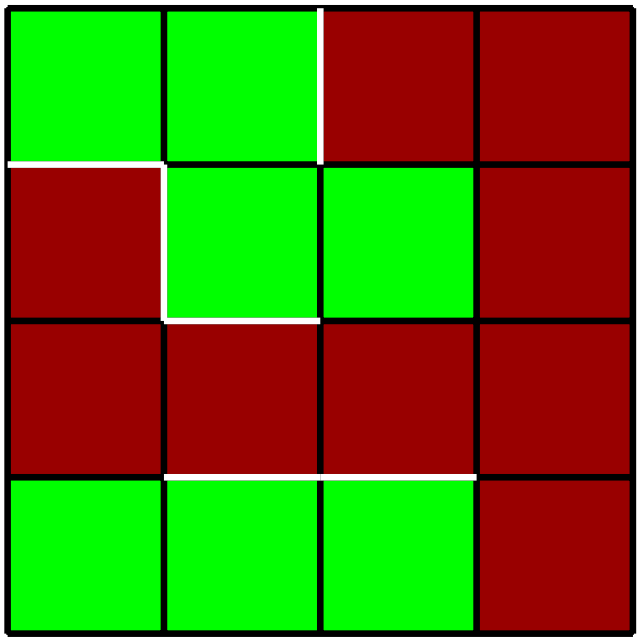
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



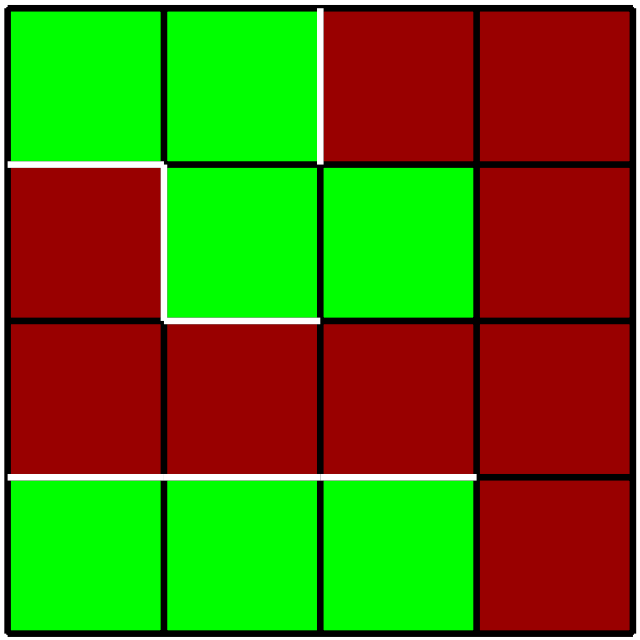
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



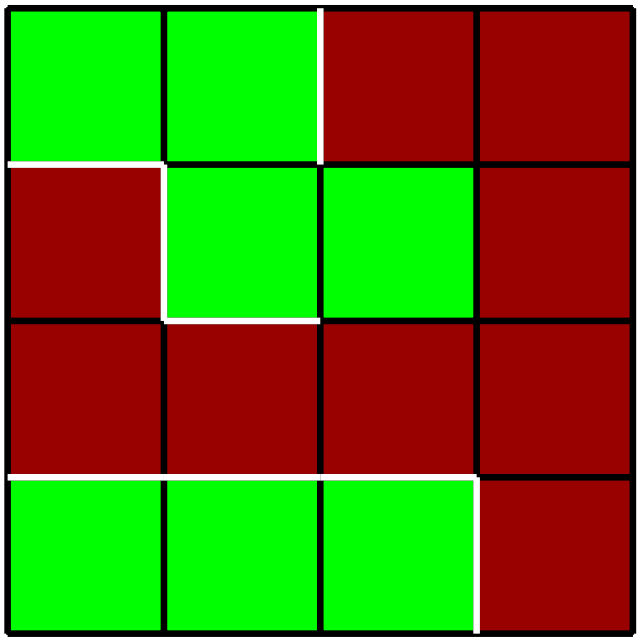
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



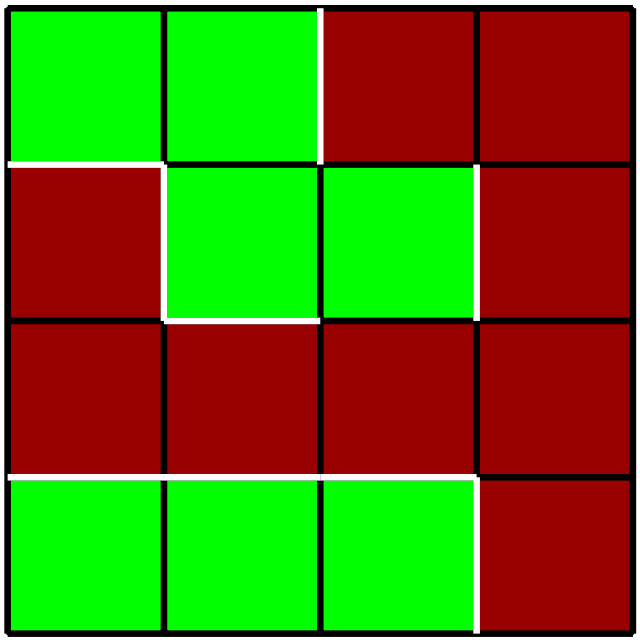
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



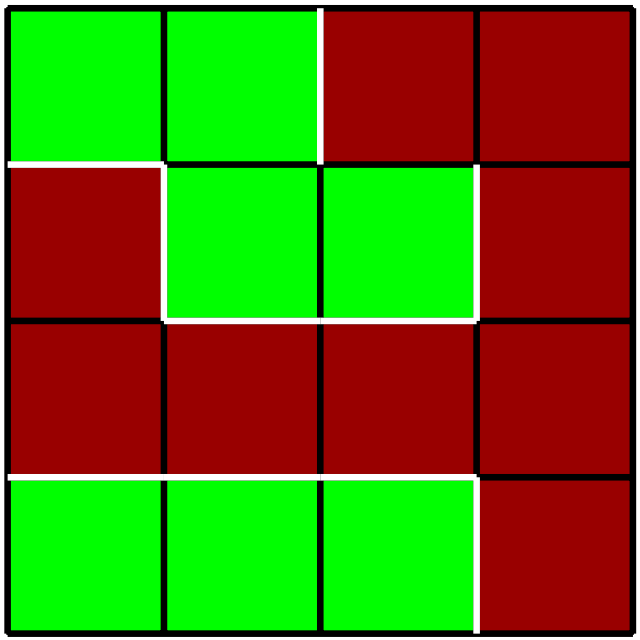
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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



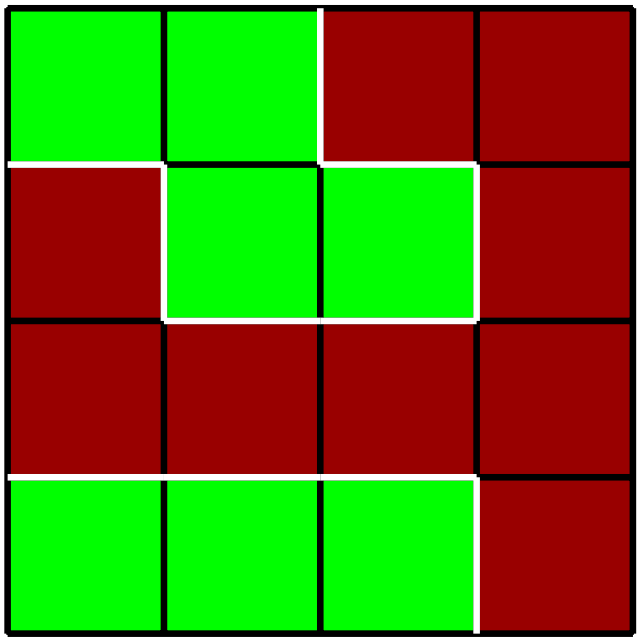
$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Autonormal (Ising) model

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

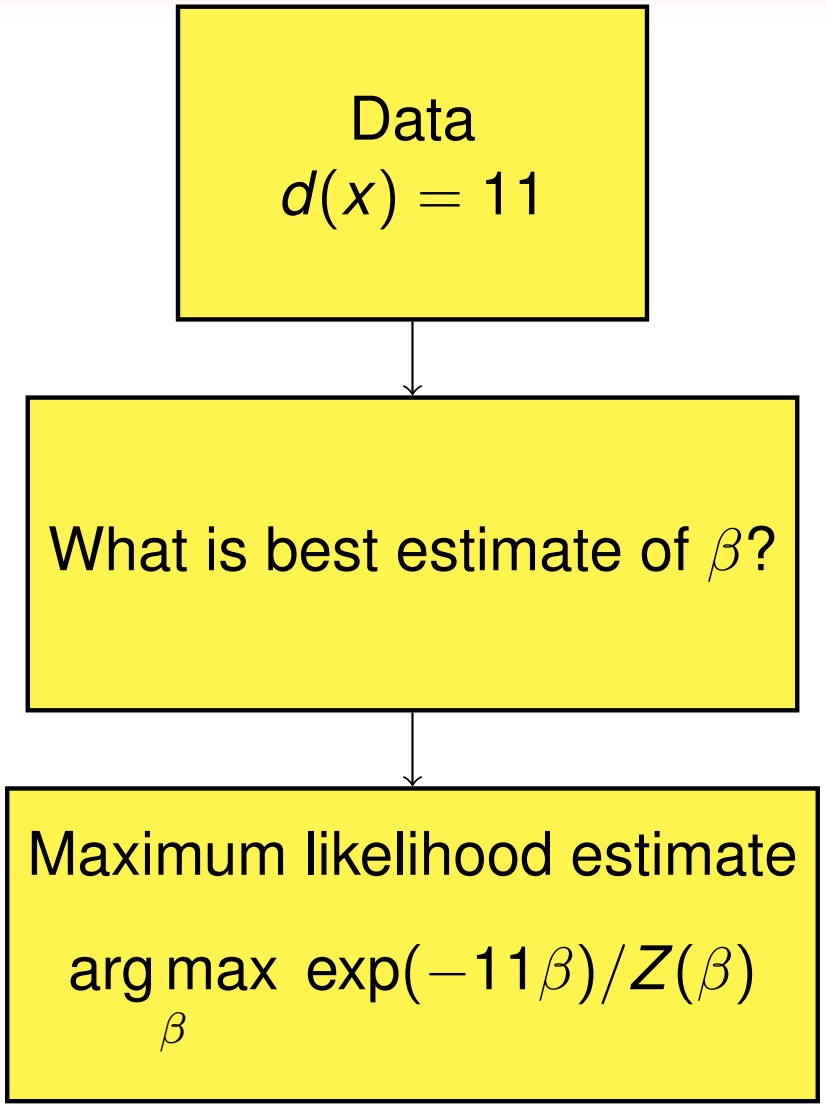


$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Basic inference



To use the MLE:

Need to know the function $Z(\beta)$ to find MLE
Typically hard to compute

Brute force computation $Z(\beta)$
Requires $2^{\#\text{ of vertices}}$ steps for autonormal model

Bayesian model selection also need $Z(\beta)$
$Z(\beta)$ appears in Bayes' Factor

$Z(\beta)$: The Partition Function

Partition function of Gibbs distribution

Ingredients:

State space $\Omega \subseteq \mathbb{R}^n$
Nonnegative function $H(x)$
Parameter β

Then the goal is to approximate:

$$Z(\beta) = \int_{x \in \Omega} \exp(-\beta H(x)) \, dx$$

Discrete version

Ingredients:

State space $\Omega \subseteq \{1, 2, \ldots, c\}^n$
Nonnegative function $H(x)$
Parameter β

Then the goal is to approximate:

$$Z(\beta) = \sum_{x \in \Omega} \exp(-\beta H(x))$$

Terminology

Definition (Gibbs distribution)

X has a (discrete) *Gibbs distribution* π_β if for all $x \in \Omega$,

$$\mathbb{P}(X = x) = \frac{1}{Z(\beta)} \exp(-\beta H(x)).$$

Definition (Partition function)

The *partition function* of a Gibbs distribution is

$$Z(\beta) = \sum_{x \in \Omega} \exp(-\beta H(x)).$$

Today's result

An approximation algorithm where

- Given $\epsilon > 0$, $\alpha > 0$
- Given the ability to sample from $\pi_{\beta'}$ for $\beta' \in [0, \beta]$
- Outputs $\hat{Z}(\beta)$ so that

$$\mathbb{P}\left(\frac{1}{1+\epsilon} \leq \frac{\hat{Z}(\beta)}{Z(\beta)} \leq 1+\epsilon\right) \geq 1-\alpha.$$

- Requires a number of samples

$$O(\ln(Z(0)/Z(\beta)))\ln(H_m)\epsilon^{-2}\ln(\alpha^{-1})],$$

$$H_m = \text{ median value of } H(X) \text{ where } X \sim \pi_0.$$

Is that good?

Why is problem hard?

Typically $Z(\beta)$ is exponential in n, the input size of problem

Many methods for lowering variance

- Multistage Sampling [Valleau Card 1972]
- Bridge Sampling [Meng Wong 1996]
- Nested Sampling [Skilling 2006]

The above are not approximation algorithms

No *guarantee* on quality of estimate obtained
(But they could be faster in practice)

History of Approximation Algorithms

Let $q = \ln(Z(0)/Z(\beta))$ and ignore ϵ, α factors

Self-reducibility [Jerrum, Valiant, Vazirani 1986]
$O[q^2]$ time under best conditions, best constant ≈ 10

TPA [H., Schott 2010]
$O[q^2]$, and much simpler

SVV [Štefankovič, Vempala, Vigoda 2009]
$O[q \ln(M)^5]$
(where $H(x) \leq M$ for all x and with 10^{10} constant out in front)

Today [H. 2012]
$O[q + 1 + \beta \ln(H_m)] +$ preproccessing
(H_m median of $H(X)$ for $X \sim \pi_0$)

Applications

- Exact p values [Frequentist Statistics]
- Model selection [Bayesian Statistics]
- Phase Transitions [Statistical Physics]
- Approximation Algorithms for #P complete problems [Theoretical Computer Science]

The algorithm

New algorithm

Uses some old ideas

- Ratios of $Z(\beta)$ values at different values of β
- Importance sampling
- Product estimator/Self-reducibility

Also some new ideas

- Paired estimator
- Well balanced cooling schedule

Ratios of partition function values

$Z(\beta)$ **usually easy for some values of** β

- In discrete models $Z(0) = \#\Omega$
- For Ising model $Z(0) = 2^{\#\text{ of nodes in graph}}$
- To find $Z(\beta)$, instead estimate:

$$r = \frac{Z(\beta)}{Z(0)}$$

then multiply by $Z(0)$

Importance sampling

Suppose can sample X from π_0, let

$$W := \exp(-\beta H(X))$$

which makes

$$\mathbb{E}[W] = \frac{\sum_{x \in \Omega} \exp(-\beta H(X))}{Z(0)} = \frac{Z(\beta)}{Z(0)}$$

so it has the right mean!

Spread

Second moment

$$\begin{aligned}\mathbb{E}[W^2] &= \frac{\sum_{x\in\Omega}(\exp(-\beta H(X)))^2}{Z(0)} \\ &= \frac{\sum_{x\in\Omega}\exp(-2\beta H(X))}{Z(0)} \\ &= \frac{Z(2\beta)}{Z(0)}\end{aligned}$$

Relative variance (square of coefficient of variation)

$$\mathbb{V}_{\text{rel}}(W) = \frac{\mathbb{V}(W)}{\mathbb{E}[W]^2} = \frac{Z(2\beta)/Z(0)}{[Z(\beta)/Z(0)]^2} - 1 = \frac{Z(2\beta)Z(0)}{Z(\beta)^2} - 1$$

Standard Monte Carlo

Relative variance (square of coefficient of variation)

1. Let n be $\mathbb{V}_{\text{rel}}(\exp(-\beta H(X)))$, where $X \sim \pi_0$
2. Draw $X_1, \ldots, X_n$ iid from π_0
3. Let $\hat{r} = (1/n) \sum_{i=1}^{k} \exp(-\beta H(X_i))$

Properties

$$\mathbb{E}[\hat{r}] = Z(\beta)/Z(0), \quad \mathbb{V}(\hat{r}) = 1$$

So number of samples needed is on order of

$$\frac{Z(2\beta)Z(0)}{Z(\beta)^2} - 1$$

Understanding the relative variance

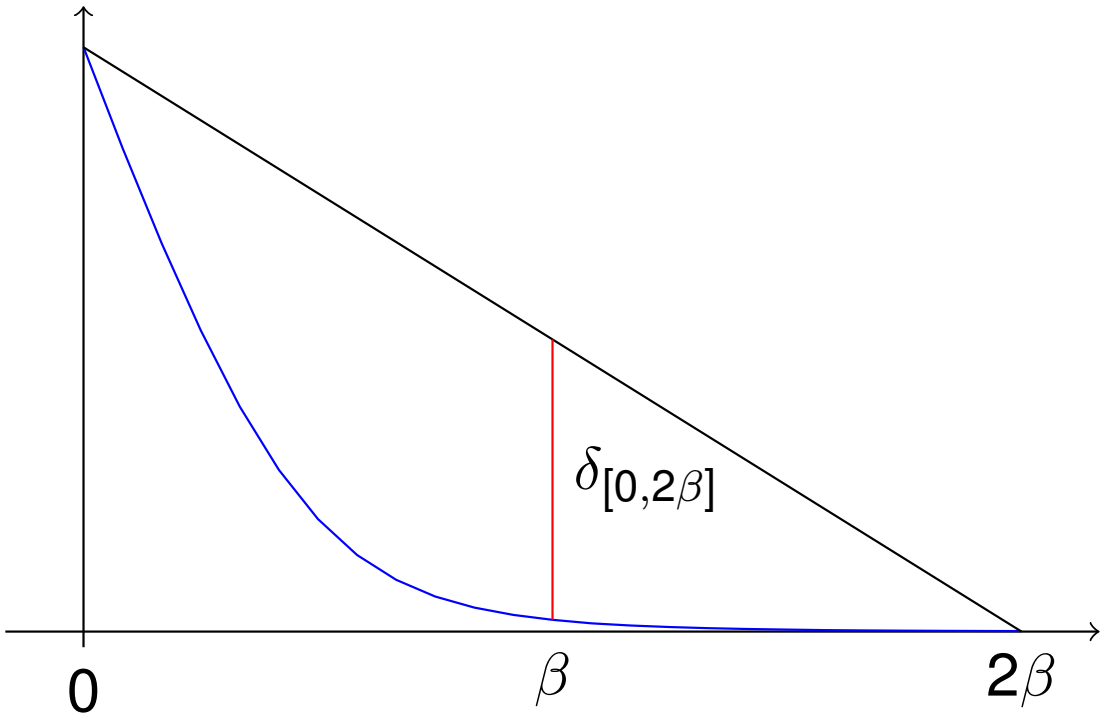
Easier to work in log space

$$z(\beta) := \ln(Z(\beta))$$

Then z is convex

$$\begin{aligned}
\ln\left(\frac{Z(2\beta)Z(0)}{Z(\beta)^2}\right) &= z(2\beta) + z(0) - 2z(\beta) \\
&= 2\left(\frac{z(2\beta) + z(0)}{2} - z(\beta)\right) \\
&= 2\delta_{[0,2\beta]}
\end{aligned}$$

The relative variance picture



Why this is bad

Problems

- Only care about $[0, \beta]$ interval
- Variance involves Z function over $[0, 2\beta]$
- δ tends to grow linearly in problem size

$$\mathbb{V}_{\text{rel}}(W) = \exp(2\delta_{[0,2\beta]})$$

Paired Estimator

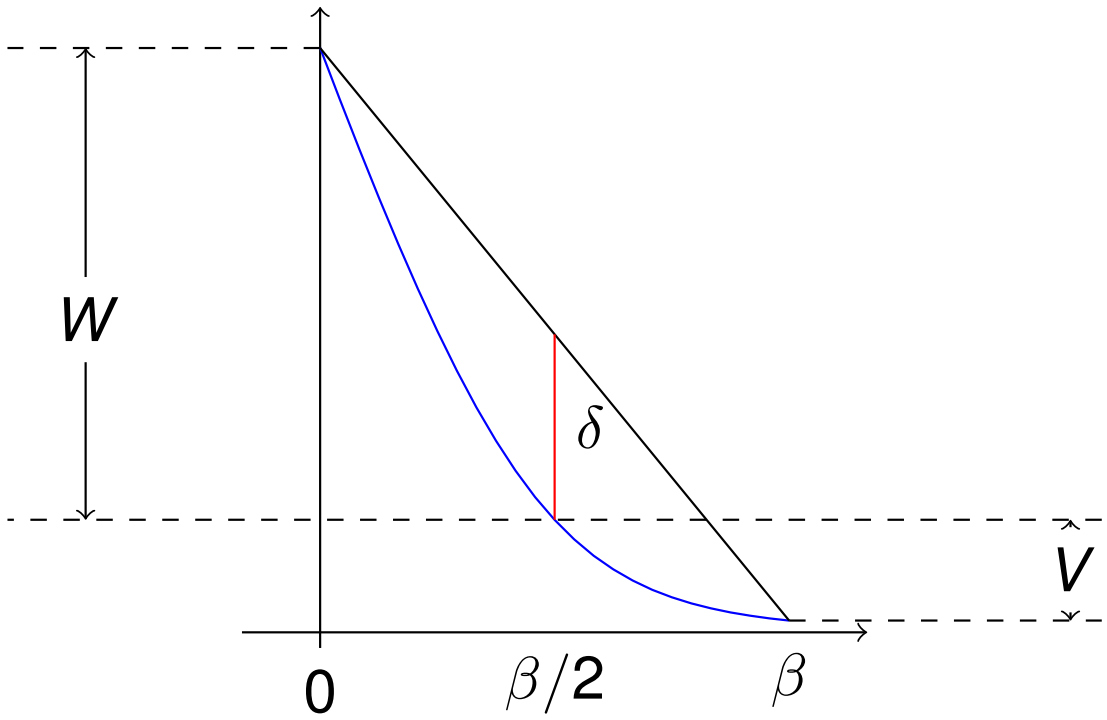
Use two estimators

$$X \sim \pi_0, \quad Y \sim \pi_\beta$$

$$W := \exp(-(\beta/2)H(X)), \quad V := \exp((\beta/2)H(Y))$$

$$W \text{ estimates } \frac{Z(\beta/2)}{Z(0)}, \quad V \text{ estimates } \frac{Z(\beta/2)}{Z(\beta)}$$

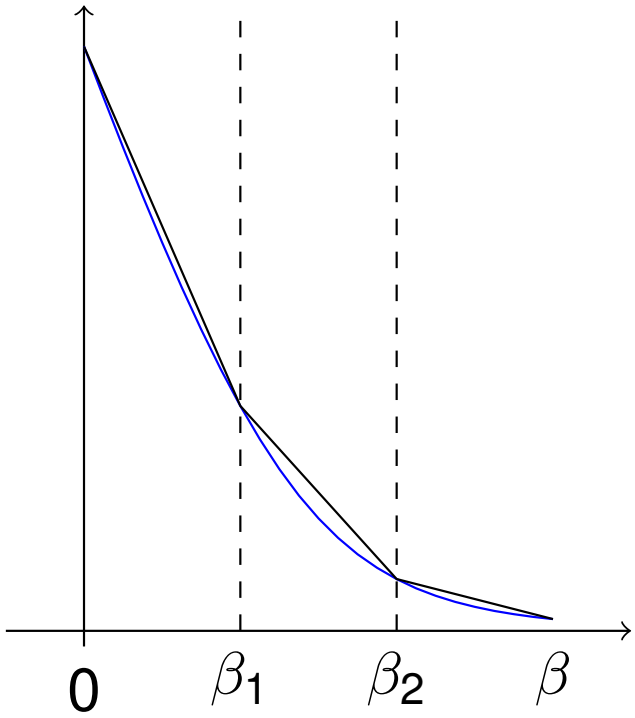
The relative variance for paired estimator



$$\mathbb{V}_{\mathrm{rel}}(W) = \mathbb{V}_{\mathrm{rel}}(V) = \exp(2\delta)$$

Breaking the interval

Multistage sampling [Valleau and Card 1972]



$$\frac{Z(\beta)}{Z(0)} = \frac{Z(\beta_1)}{Z(0)} \frac{Z(\beta_2)}{Z(\beta_1)} \frac{Z(\beta)}{Z(\beta_2)}$$

Paired Product Estimator

$$\frac{Z(\beta)}{Z(0)} = \frac{Z(\beta_1)}{Z(0)} \frac{Z(\beta_2)}{Z(\beta_1)} \frac{Z(\beta)}{Z(\beta_2)}$$

$$\frac{Z(\beta)}{Z(0)} \approx \frac{W_1}{V_1} \frac{W_2}{V_2} \frac{W_3}{V_3} = \frac{W}{V}$$

Analyzing Product Estimators

Variance of Product Estimator

Theorem

If $P = \prod P_i$ where the P_i are independent, then

$$\mathbb{E}(P) = \prod \mathbb{E}(P_i)$$

$$\mathbb{V}_{rel}(P) = -1 + \prod (1 + \mathbb{V}_{rel}(P_i))$$

For the products in our algorithm

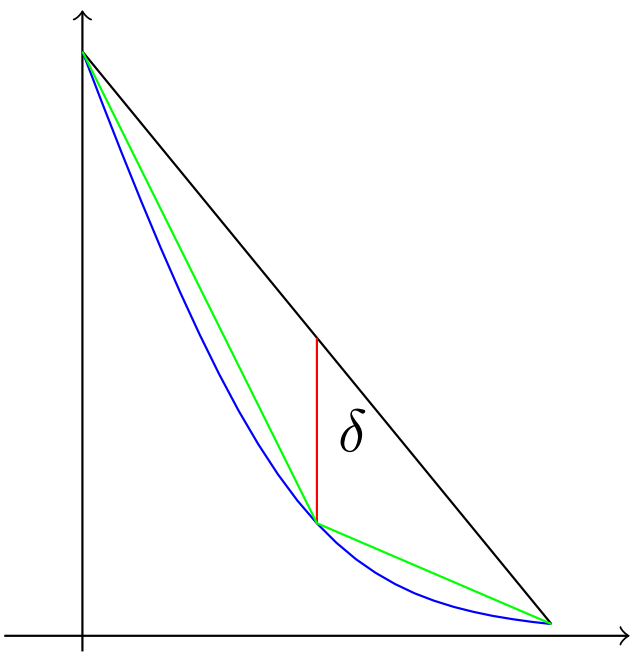
$$\mathbb{V}_{\text{rel}}(W_i) = \exp(2\delta_i) - 1 \Rightarrow \mathbb{V}_{\text{rel}}(W) = -1 + \exp\left(2\sum \delta_i\right)$$

Curving the z function

Theorem

In interval $[\beta_i, \beta_{i+1}]$, either δ_i is small, or the derivative of z changes rapidly. To be precise, if $\epsilon_i = z(\beta_i) - z(\beta_{i+1})$, then

$$\frac{z'(\beta_{i+1})}{z'(\beta_i)} \leq \exp(-2\delta_i/\epsilon_i)$$



Given δ:

Green line has smallest change in ratio of derivative

Telescoping the derivative

Corollary

This implies that

$$\frac{z'(\beta)}{z'(0)} = \frac{z'(\beta_1)}{z'(0)} \cdot \frac{z'(\beta_2)}{z'(\beta_1)} \cdots \frac{z'(\beta)}{z'(\beta_{\ell-1})} \leq \exp\left(-2\sum \frac{\delta_i}{\epsilon_i}\right)$$

Invert the inequality in corollary

$$\frac{z'(0)}{z'(\beta)} \geq \exp(2\sum \delta_i/\epsilon_i)$$

If $\epsilon \geq \epsilon_i$ for all i:

$$\left[\frac{z'(0)}{z'(\beta)}\right]^{\epsilon} \geq \exp(2\sum \delta_i)$$

Result is the following theorem

Theorem

If $z(\beta_i) - z(\beta_{i+1}) \leq \epsilon$ *for all* i*, then*

$$\mathbb{V}_{rel}(W) \leq \left[\frac{z'(0)}{z'(\beta)}\right]^{\epsilon}$$

Corollary

If $z(\beta_i) - z(\beta_{i+1}) \leq 1/[\ln(-z'(0)) - \ln(-z'(\beta))]$ *for all* i *then*

$$\mathbb{V}_{rel}(W) \leq e$$

Sidebar: z' as probabilistic object

Fact

If $X \sim \pi_\beta$, then

$$-z'(\beta) = \mathbb{E}_{\pi_\beta}[H(X)]$$

Proof.

$$\begin{aligned}
-z'(\beta) &= -(d/d\beta) \ln \left(\sum \exp(-\beta H(x)) \right) \\
&= -\frac{\sum (d/d\beta) \exp(-\beta H(x))}{\sum \exp(-\beta H(x))} \\
&= -\frac{\sum -H(x) \exp(-\beta H(x))}{Z(\beta)} \\
&= \mathbb{E}[H(X)]
\end{aligned}$$

□

Our algorithm so far

Ingredients:

Cooling schedule with $z(\beta_i) - z(\beta_{i+1}) \leq \epsilon$
Ability to sample from π_{β_i} for all i

Output

Estimates W and V with $\mathbb{E}[W]/\mathbb{E}[V] = r$
$\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V) \leq (z'(0)/z'(\beta))^{\epsilon}$

Natural questions

1. How long will the cooling schedule be?
2. How small should ϵ be?
3. What if $z'(0)$ is too large?
4. What if $z'(\beta)$ is too small?
5. How do we actually find the cooling schedule?

How long will the cooling schedule be?

For well balanced cooling schedule

- Suppose $z(\beta_i) - z(\beta_{i+1}) \approx \epsilon$ for all i
- Then length of cooling schedule about

$$\ell = [\ln(Z(0)) - \ln(Z(\beta))]/\epsilon$$

- Number of samples to get W and V is 2ℓ
- Make ϵ as large as possible: $1/\ln(z'(0)/z'(\beta))$ so

$$\mathbb{V}_{\text{rel}}(W) = \mathbb{V}_{\text{rel}}(V) \leq (z'(0)/z'(\beta))^{\epsilon} \leq e$$

What if $z'(0)$ is too big?

Making $z'(0)$ finite

- Technically, $z'(0)$ could be infinite!
- Trick: restrict state space using medians
- Let Ω_m be states with $H(x)$ value at most median of $H(X)$
- Only need two samples (on average) to get $X \in \Omega_m$
- For $X \in \Omega_m$, $\mathbb{E}[H(X)] \leq m$, always finite
- Estimating $Z(\Omega_m)/Z(\Omega)$ easy

Rarely needed in practice

- For Ising model, $z'(0)$ is half the edges in graph

What if $z'(\beta)$ is too small?

Easiest solution

- Form an adjusted H function:

$$H_{\text{adj}}(x) := 1 + H(x)$$

- Note π_β same with H_{adj} as with H
- $Z_{\text{adj}}(\beta) = \exp(-\beta)Z(\beta)$
- Which means

$$\ln(Z_{\text{adj}}(0)) - \ln(Z_{\text{adj}}(\beta)) = \beta + \ln(Z(0)) - \ln(Z(\beta))$$

- Always have $|z'_{\text{adj}}(\beta)| \geq 1$

How do we actually find the cooling schedule?

TPA algorithm [H.-Schott 2010]

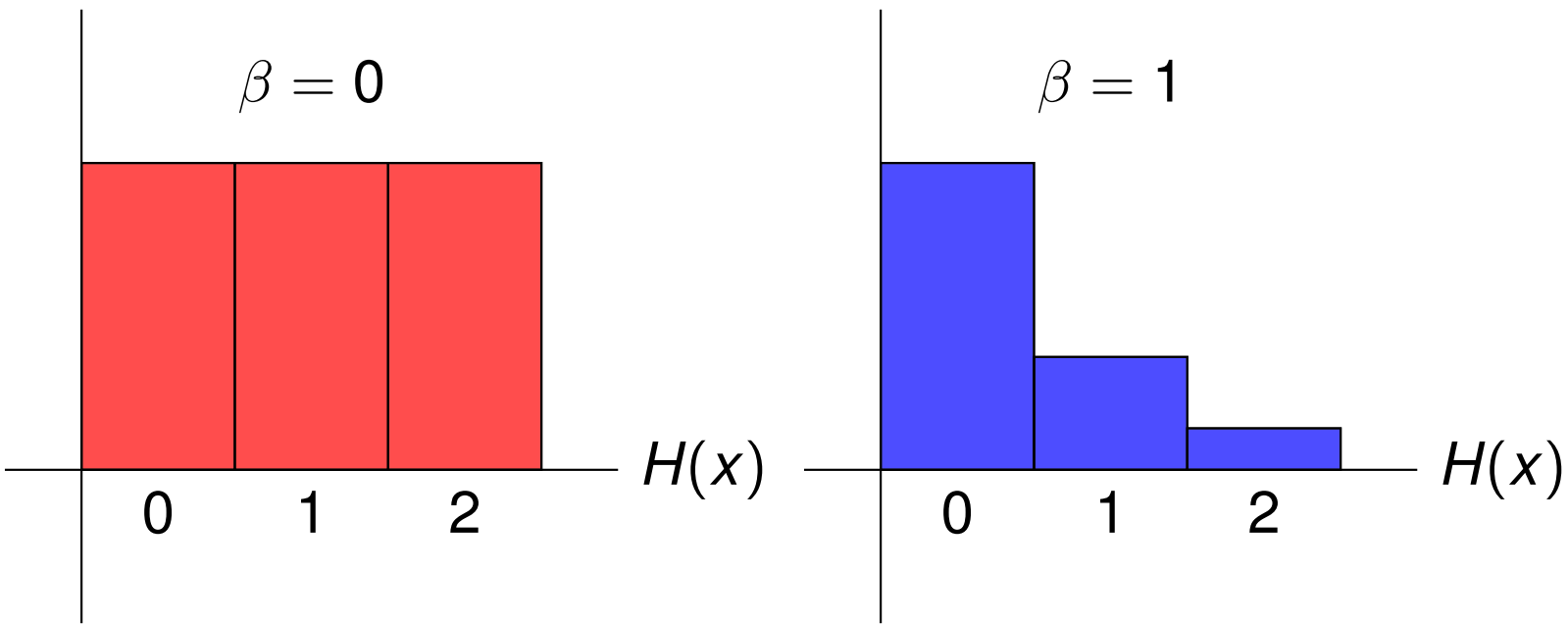
- Creates a well balanced schedule
- Number of samples used close to length of schedule

Rough idea

- Start with $b = \beta$
- Draw $X \sim \pi_b$, $Y \sim \text{Unif}([0, \exp(-\beta H(X))]$
- Let b be smallest value such that $Y \in [0, \exp(-bH(X))]$
- Repeat until $b = 0$.
- Set of β values is approximately 1-balanced schedule

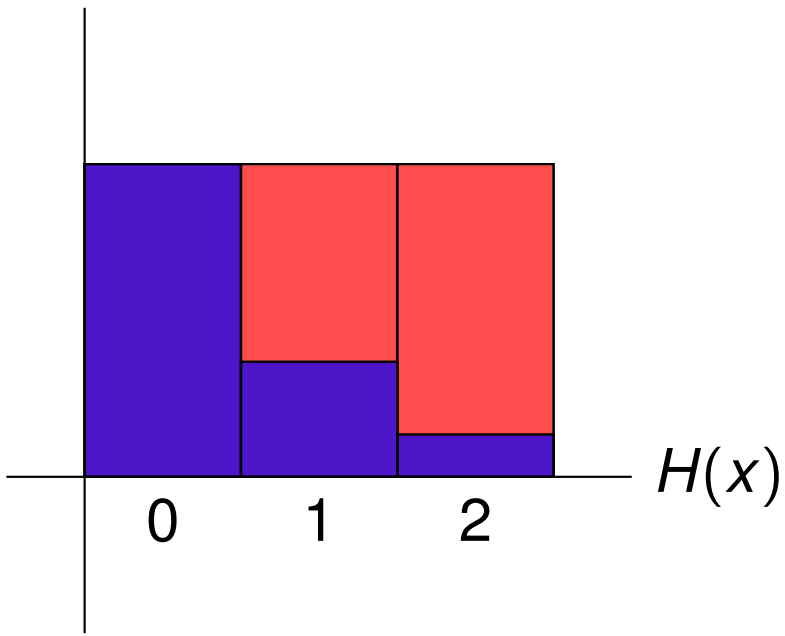
Geometric view

$$Z(\beta) = 1 + \exp(-\beta) + \exp(-2\beta)$$



Combining...

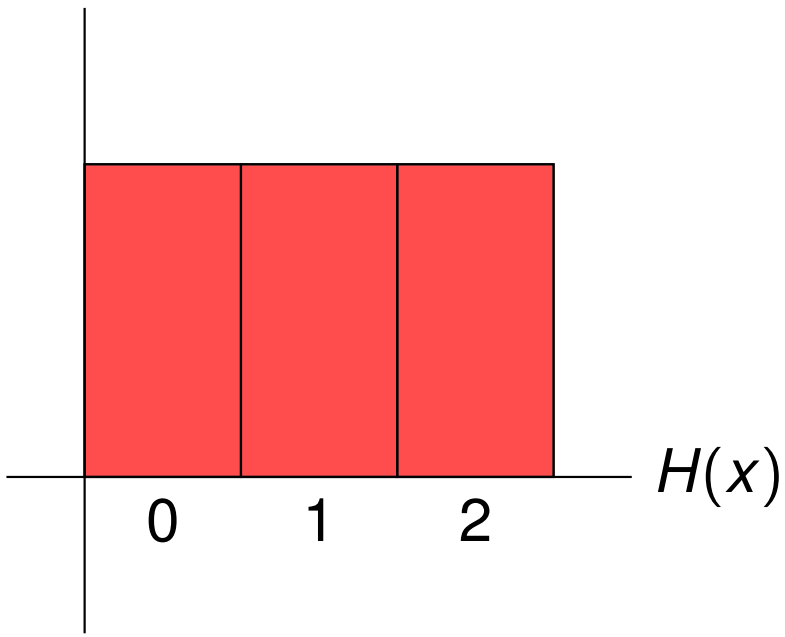
$$Z(\beta) = 1 + \exp(-\beta) + \exp(-2\beta)$$



TPA

- Draw (X, Y) uniformly from red region
- Find b so that point just touches blue region

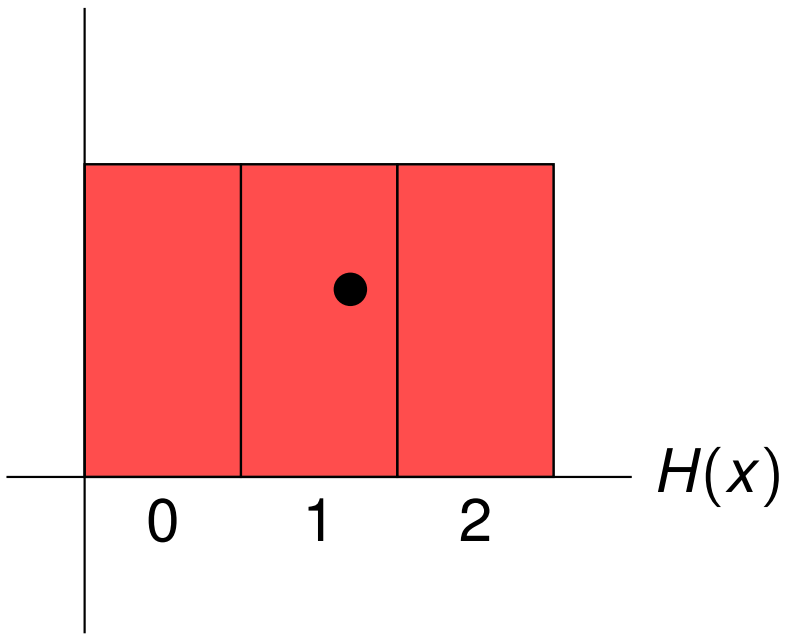
$$Z(\beta) = 1 + \exp(-\beta) + \exp(-2\beta)$$



TPA

- Draw (X, Y) uniformly from red region
- Find b so that point just touches blue region

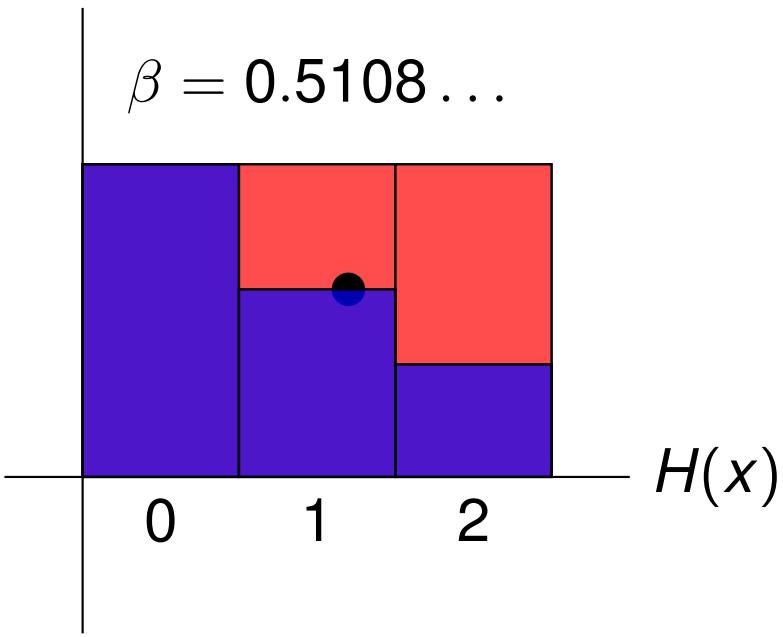
$$Z(\beta) = 1 + \exp(-\beta) + \exp(-2\beta)$$



TPA

- Draw (X, Y) uniformly from red region
- Find b so that point just touches blue region

$$Z(\beta) = 1 + \exp(-\beta) + \exp(-2\beta)$$



The result

Example run:

$$(1, .5108, 0.1342, 0)$$

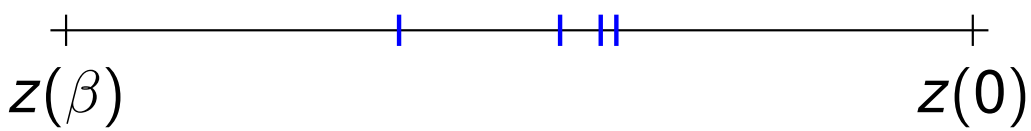
Fact (H.-Schott 2010)

If the values found in a run of TPA are $\beta = b_0, b_1, \ldots, b_k = 0$*, then* $z(b_1), z(b_2), \ldots, z(b_k)$ *form a Poisson point process on the interval* $[z(\beta), z(0)]$*.*

Why Poisson point process (PPP)?

Works as follows

- At each step, ratio of blue area to red area is random
- Random variable is uniform over $[0, 1]$
- Take $-\ln(U)$, get exponential distribution
- When differences exponential, get PPP



Poisson process

Fact (H.-Schott 2010)

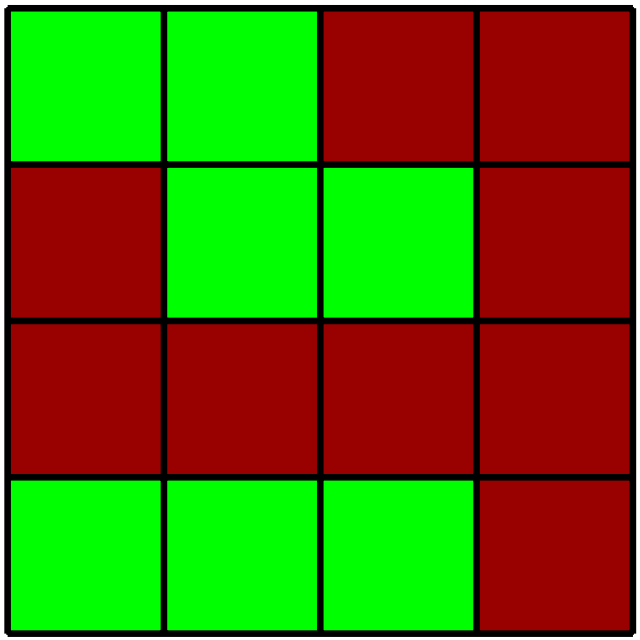
If the values found in a run of TPA are $\beta = b_0, b_1, \ldots, b_k = 0$*, then* $z(b_1), z(b_2), \ldots, z(b_k)$ *form a Poisson point process on the interval* $[z(\beta), z(0)]$*.*

- Run TPA several times to build large set of β values
- Points roughly evenly spaced (uniformly distributed)

Back to the Autonormal model

Autonormal (Ising) model

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



$d(x) = 11$
(number of adj nodes that disagree)

Model:

$$\mathbb{P}(X = x) = \frac{\exp(-\beta d(x))}{Z(\beta)}$$

Initial changes

What we want

- Cooling schedule for $Z(\beta)$ for all $\beta \in [0, \infty)$
- Let $h(x) = d(x) + 1$
- Since $d(x) \in [0, m]$, this means $h(x) \in [1, m + 1]$ (where $m =$ the number of edges)
- So $\ln(z'(0)/z'(\infty)) \leq \ln(m + 1)$
- Also $Z(0) = 2^n$ ($n =$ number of nodes)
- So length of schedule about $\ln(m + 1) \cdot n \cdot \ln(2)$

Procedure

1. Use TPA [H.-Schott 2010] to find a good cooling schedule
2. Generate W and V, use W/V as estimate

For better accuracy

- Generate several copies of W and V
- Find the sample average $\bar{W}$ and $\bar{V}$
- The estimate is then $\bar{W}/\bar{V}$

Initial changes

Improvement for 2D lattices

- Here $m \leq 4n$, β constant (below phase transition)
- Cooling schedule for $Z(\beta)$ for all $\beta \in [0, \infty)$
- Let $h(x) = d(x) + m$
- Since $d(x) \in [0, m]$, this means $h(x) \in [m, 2m]$ (where $m =$ the number of edges)
- So $\ln(z'(0)/z'(\infty)) \leq \ln(2)$
- Also $Z(0) = 2^n$ ($n =$ number of nodes)
- So length of schedule about $\ln(2) \cdot n \cdot \ln(2) + \beta m$
- Result: method linear in m

Future work

Running method in practice

Obtaining samples

- Simulated tempering runs all β values simultaneously
- Omnithermal sampling
- Lose analysis based on independence
- Gain from many more samples

Preprocessing TPA output

- Some temperatures will be too close together
- Use some fast samples to eliminate bunched β values

Ideas:

1. Is the $\ln(H_m)$ part really necessary? (Computer experiments indicate no)
2. Take advantage of the form:

$$Z(\beta) = \sum_{i=1}^{n} a_i \exp(-i\beta)$$

3. Flattening $Z(\beta)$ function by substracting median $H(x)$

Summary

Most imporant thing

First practical approx alg (nearly) linear in dimension

Uses some old ideas

Importance Sampling, Multistage Sampling, Product Estimator

Add some new ideas

Paired Product Estimator, Well-balanced schedule

The resulting approximation algorithm

Relatively simple to implement, guaranteed quality of estimate

References

M. L. Huber and S. Schott.
Using TPA for Bayesian inference.
Bayesian Statistics 9, pages 257–282, 2010.

D. Štefankovič, 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.

J.P. Valleau and D.N. Card.
Monte Carlo estimation of the free entergy by multistage sampling.
J. Chem. Phys., 57(12):5457–5462, 1972.