The Paired Product Estimator for normalizing constants of Gibbs distribution

Mark Huber[1] and Sarah Schott[2]

[1]Department of Mathematical Sciences, Claremont McKenna College

[2]Department of Mathematics, Duke University

2 June, 2011

The importance of keeping variance small

Daedalus warned Icarus not to fly too close to the sun, as it would melt his wings, and not too close to the sea, as it would dampen them and make it hard to fly.

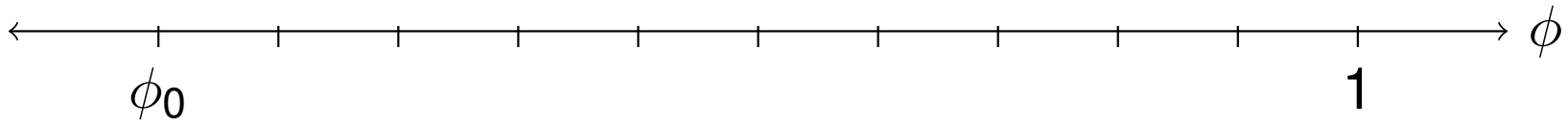
Previous talks

Relation to Gareth's talk

Gareth's Example 2 yesterday is Ising model

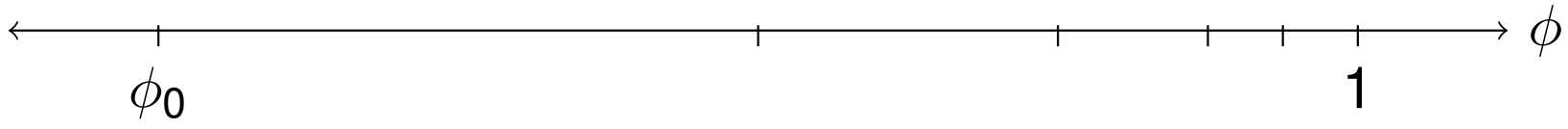
Relation to Alex's talk

Bridging (unnormalized) densities between Π and Π^{ϕ_0}



Equal spacing in ϕ

Today I will also use bridging densities
Let $Z(\phi)$ be normalizing constant for Π^{ϕ}:



Equal spacing in $\ln(Z(\phi))$

The Objective

Today

The Goal

Approximate the normalizing constant Z for a Gibbs distribution:

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

Previous work

Method	$\tilde{O}(\ln(Z))$	$\Omega(\ln(Z)^2)$	Easy to implement?
TPA		X	Yes
Nested Sampling		X	Yes
Štefankovič, et al.	X		No
Today	X		Yes

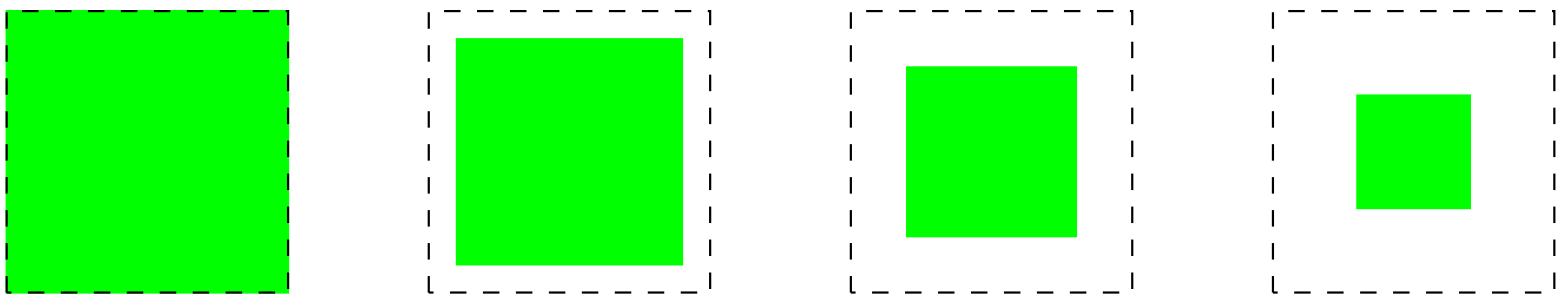
Cooling schedule

Definition

A *cooling schedule* is a sequence of distributions that interpolate between a starting and ending distribution.

Example

Uniform distribution over the following sets:

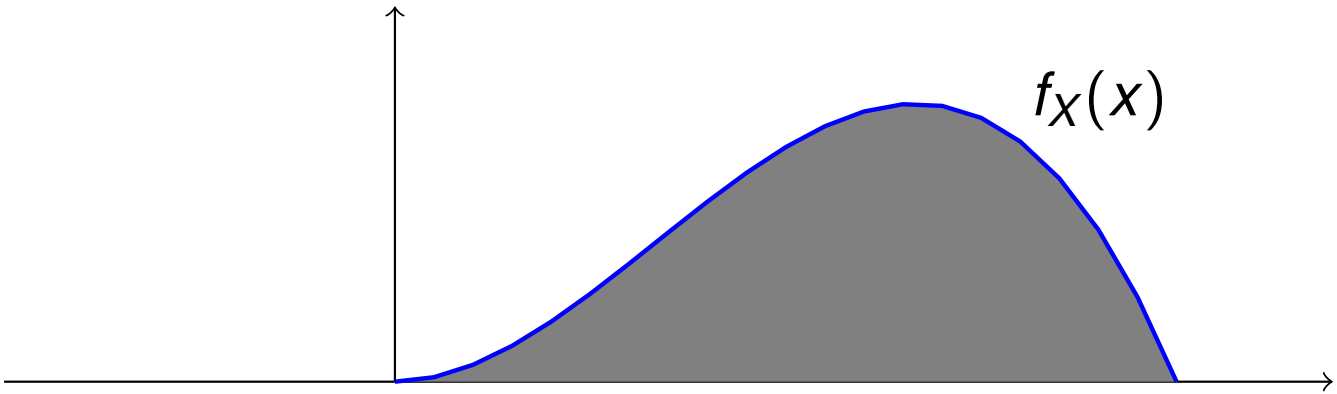


Fundamental Theorem of Simulation

Any distribution can be made uniform with auxiliary variables.

Simple example of Fun Thm of Simulation

To sample from density $f_X(x)$



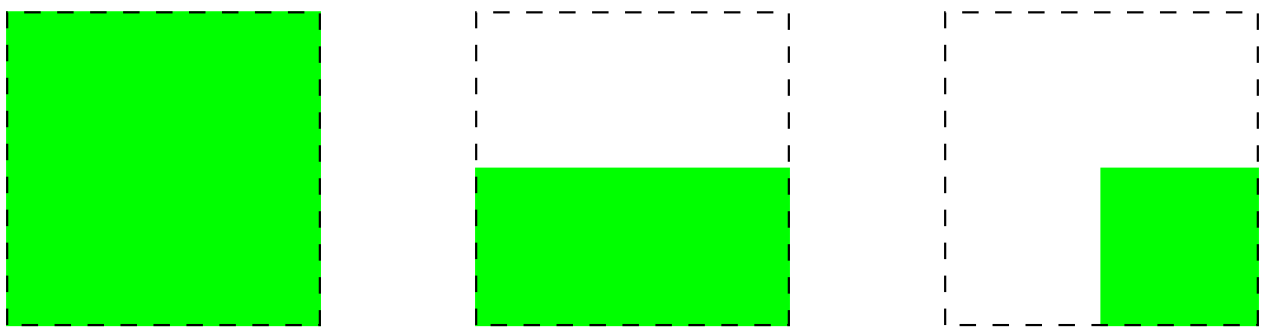
Sample (X, Y) uniformly from gray region, keep X
[Here Y is auxiliary variable]

Well balanced cooling schedule

Definition

A schedule is *well balanced* if the ratio of the normalizing constants of successive sets lies in $[1/e, 2/e] \approx [37\%, 74\%]$.

Example



The importance of well balanced cooling schedule

Reasons to use well balanced schedules

- Make tempering chains mix well
  [Woodard, Schmidler, H 2009, 2010]
- Today: getting low variance estimates of Z

Gibbs distributions

Definition

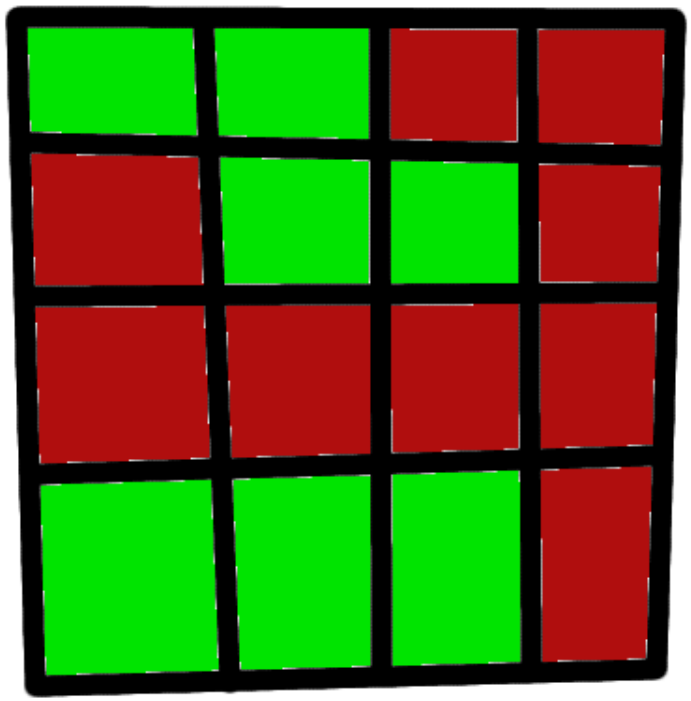
A *Gibbs distribution* is a family of distributions $\{\pi(\beta)\}$ indexed by the *inverse temperature* β, where

$$\pi(\beta)(\{x\}) = \exp(-\beta H(x))/Z(\beta).$$

Here $Z(\beta) = \sum_{x \in \Omega} \exp(-\beta H(x))$, is called the *partition function* and $H(x) \geq 0$ is the *Hamiltonian*.

Example: The Ising model

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



$H(x) = 2 \cdot \#\text{adj red-green plots}$

Here $H(x) = 22$

$$\pi(x) = \frac{\exp(-\beta H(x))}{Z(\beta)}$$

β penalizes dissimilarity

Ratios of $Z(\beta)$ easier

A technical note

Usually $Z(0)$ or $Z(\infty)$ easy to find, so suffices to gives a method for estimating

$$\frac{Z(\beta_{\text{start}})}{Z(\beta_{\text{end}})}$$

for some values of β_{start} and β_{end}

Example

For the Ising model

$$Z(0) = 2^{\#\text{of plots}}, \quad Z(\infty) = 2$$

Product Estimators

Product estimators

Definition

A *product estimator* of $a = a_1 a_2 \cdots a_\ell$ is an estimator of the form

$$\hat{a} = \hat{a}_1 \hat{a}_2 \cdot \hat{a}_\ell,$$

where each $\hat{a}_i$ is an independent estimator of a_i.

Product estimator for Gibbs distributions

Example

To determine $Z(\beta)$ given $Z(0)$, use cooling schedule indexed by

$$\beta = \beta_0 > \beta_1 > \beta_2 > \cdots > \beta_{\ell-2} > \beta_{\ell-1} = 0.$$

and

$$Z(\beta) = \frac{Z(\beta_0)}{Z(\beta_1)} \frac{Z(\beta_1)}{Z(\beta_2)} \cdots \frac{Z(\beta_{\ell-2})}{Z(\beta_{\ell-1})} Z(0)$$

Mean and spread of product estimator

Product estimators with a_i independent

If each $\hat{a}_i$ is unbiased for a_i,

$$\mathbb{E}[\hat{a}] = \prod_i \mathbb{E}[\hat{a}_i] = \prod_i a_i = a,$$

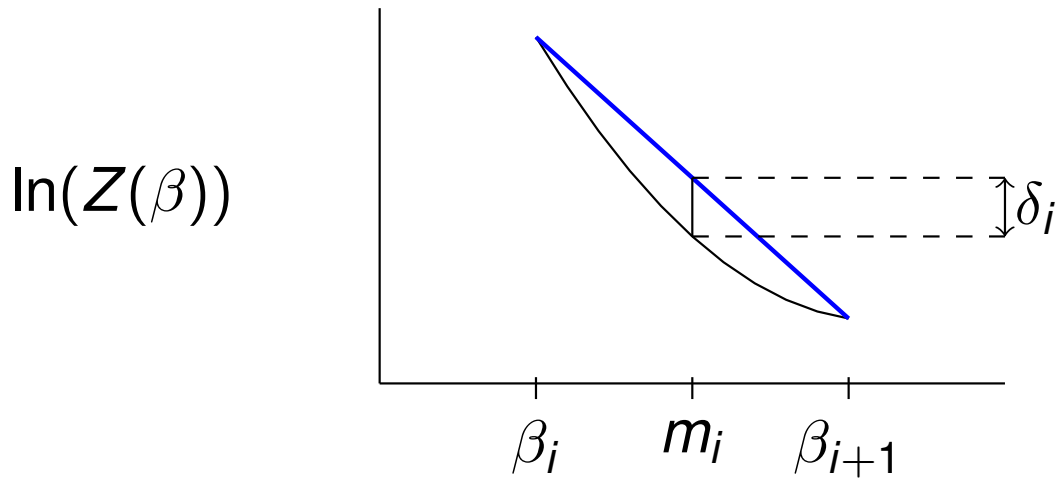
and

$$\frac{\mathbb{E}[\hat{a}^2]}{\mathbb{E}[\hat{a}]^2} = \prod_i \frac{\mathbb{E}[a_i^2]}{\mathbb{E}[a_i]^2}.$$

A geometric view of spread for Gibbs distributions

Draw $X_i \sim \pi(\beta_i)$, then

$$W_i = \exp(H(X_i)(\beta_i - m_i)).$$

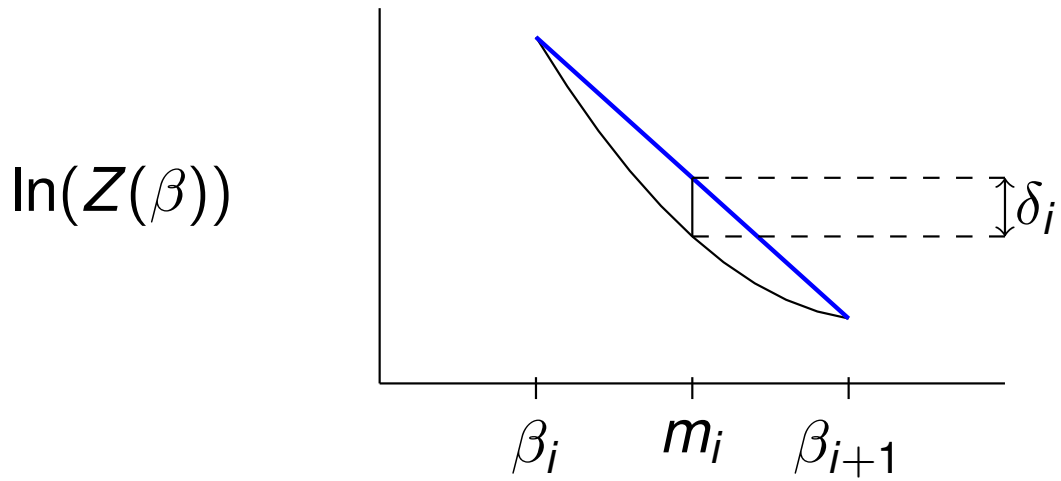


$$\mathbb{E}[W_i] = \frac{Z(m_i)}{Z(\beta_i)}, \quad \frac{\mathbb{E}[W_i^2]}{\mathbb{E}[W_i]^2} = \exp(2\delta_i).$$

The right half of interval

To estimate the "right half" of the interval:

$$Y_i \sim \pi_{\beta_{i+1}}, \quad V_i = \exp(H(Y_i)(\beta_{i+1} - m_i))$$



$$\mathbb{E}[V_i] = \frac{Z(m_i)}{Z(\beta_{i+1})}, \quad \frac{\mathbb{E}[V_i^2]}{\mathbb{E}[V_i]^2} = \exp(2\delta_i)$$

The Paired Product Estimator

Definition

$$W = \prod_i W_i, \text{ so } \mathbb{E}[W] = \frac{Z(m_0)Z(m_1)\cdots Z(m_{\ell-1})}{Z(\beta_0)Z(\beta_1)\cdots Z(\beta_{\ell-1})}$$

$$V = \prod_i V_i, \text{ so } \mathbb{E}[V] = \frac{Z(m_0)Z(m_1)\cdots Z(m_{\ell-1})}{Z(\beta_1)Z(\beta_2)\cdots Z(\beta_{\ell})}$$

Then (W, V) is the *paired product estimator* with the property that

$$\frac{\mathbb{E}[W]}{\mathbb{E}[V]} = \frac{Z(\beta_\ell)}{Z(\beta_0)}$$

Better estimators with well balanced schedules

Theorem (H, Schott 2011)

Let $\bar{W}_i$ be the sample average of r draws of W_i, and $\hat{W} = \prod \bar{W}_i$. Then for a well balanced schedule

$$\frac{\mathbb{E}[\hat{W}^2]}{\mathbb{E}[\hat{W}]^2} \leq \left(\frac{\mathbb{E}[H(X)]}{\mathbb{E}[H(Y)]} \right)^{(e-1)/r}$$

where $X \sim \pi(\beta_{start})$ and $Y \sim \pi(\beta_{end})$.

Using the theorem

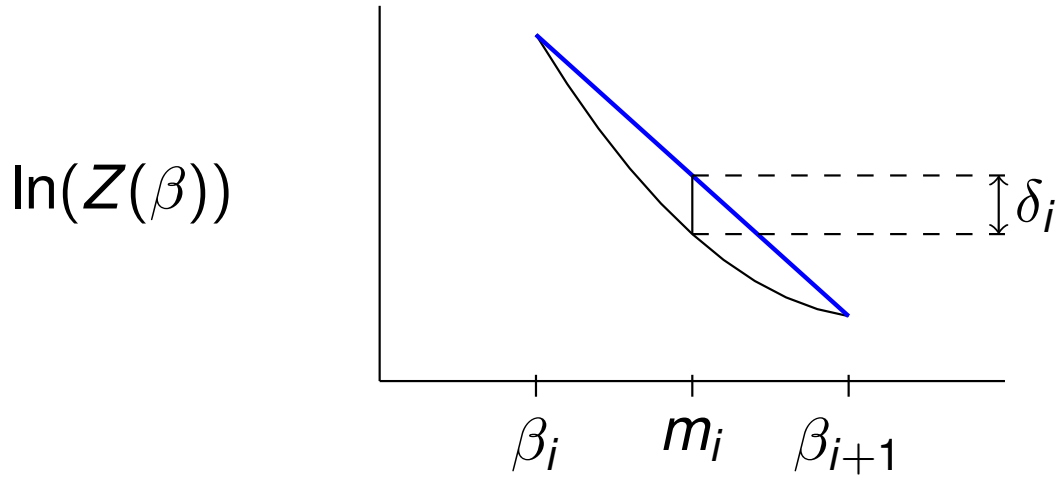
Number of samples

$$\Theta\left(\ln\left(\frac{\mathbb{E}[H(X)]}{\mathbb{E}[H(Y)]}\right)\ln(Z)\right)$$

samples are necessary to get small variance.

Usually $\mathbb{E}[H(X)]/\mathbb{E}[H(Y)]$ grows as the dimension.

Proof idea



- For δ_i to be large, $\ln(Z(\beta))$ must curve
- Changes in the slope accumulate
- Slope of $\ln(Z(\beta))$ is mean of $H(X)$ where $X \sim \pi(\beta)$

The Algorithm

The chicken and the egg

The conundrum

- To find Z quickly, need a well balanced schedule
- To get well balanced schedule, need to know Z!

The solution

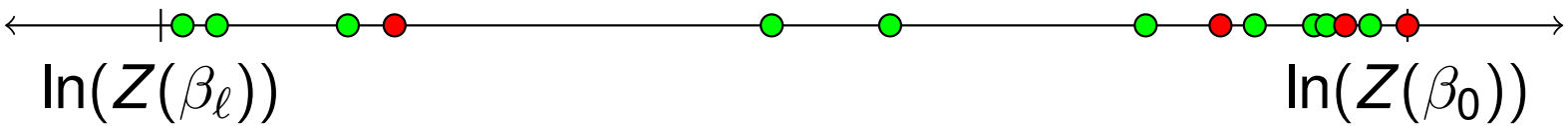
- Use the Tootsie Pop Algorithm (TPA)

TPA: The Tootsie Pop Algorithm

In $O(k \ln(Z))$ expected time

- Gives a Poisson point process of rate k in $\ln(Z(\beta))$ space

Example: $k = 4$



- Every kth points about 1 apart (Gamma$(1/k, k)$)
- For $k = \Theta(\ln(\ln(Z)))$, whole schedule balanced

Total running time

Theorem

The expected running time of the procedure is

$$\Theta(\ln(Z)[\ln(\ln(Z)) + \ln(\mathbb{E}[H(X)]/\mathbb{E}[H(Y]),$$

where X is a draw from the distribution indexed by β_{start} and Y is drawn from the distribution indexed by β_{end}

The algorithm (Part I: Well balanced schedule)

Using TPA to get well balanced schedule

Input: rate k

1. Set $P \leftarrow \emptyset$
2. Repeat k times
3. Set $\beta \leftarrow \beta_{\text{start}}$
4. While $\beta < \beta_{\text{end}}$
5. Draw $X \leftarrow \pi(\beta)$, draw $U \leftarrow \text{Unif}([0,1])$
6. Set $\beta \leftarrow \sup\{b : \exp(-b \cdot H(X)) = U \exp(-\beta H(X))\}$
7. If $\beta > \beta_{\text{end}}$ then $P \leftarrow P \cup \{\beta\}$
8. Sort P, return $P_{(k)}, P_{(2k)}, P_{(3k)}, \ldots$

The algorithm (Part II: Estimating Z)

Use a well balanced schedule to get $Z(\beta_{\text{start}})/Z(\beta_{\text{end}})$

Input: schedule $\beta_0, \ldots, \beta_{\ell-1}$, repetitions r

1. Set $W \leftarrow 1$, $V \leftarrow 1$
2. For i from 0 to $\ell - 1$
3. Draw $X_1, \ldots, X_r \leftarrow \pi(\beta_i)$, $Y_1, \ldots, Y_r \leftarrow \pi(\beta_{i+1})$
4. Set $W \leftarrow W \frac{1}{r} \sum_i \exp(H(X_i)(\beta_i - m_i))$
5. Set $V \leftarrow V \frac{1}{r} \sum_i \exp(H(Y_i)(\beta_{i+1} - m_i))$
6. Return W/V as estimate

General applications

Recommended course of action

1. Use TPA to find a well balanced cooling schedule
2. Use importance sampling to estimate pieces
3. Combine for overall estimate

Speeding things up

Improve efficiency by

- Use particle methods to get multiple samples at a temp
- Use parallel tempering/annealing to get multiple temps
- Reuse samples for W_i and V_{i-1} (positively correlated)

Conclusions

What have we done?

- First

$$O(\ln(Z)[\ln(\ln(Z)) + \ln(\mathbb{E}(Z))])$$

  algorithm for approximating Z for Gibbs distributions
- Straightforward to implement

Future work

- Remove nuisance $\ln(\ln(Z))$ term
- Remove nuisance $\ln(\mathbb{E}(H(X)))$ term
- Usually these are both about ln(dimension)

References

M. Huber and S. Schott
Using TPA for Bayesian inference
Bayesian Statistics **9**, 257–282, 2011

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