Using TPA for Bayesian inference

Mark Huber[1] and Sarah Schott[2]

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

[2]Department of Mathematics, Duke University

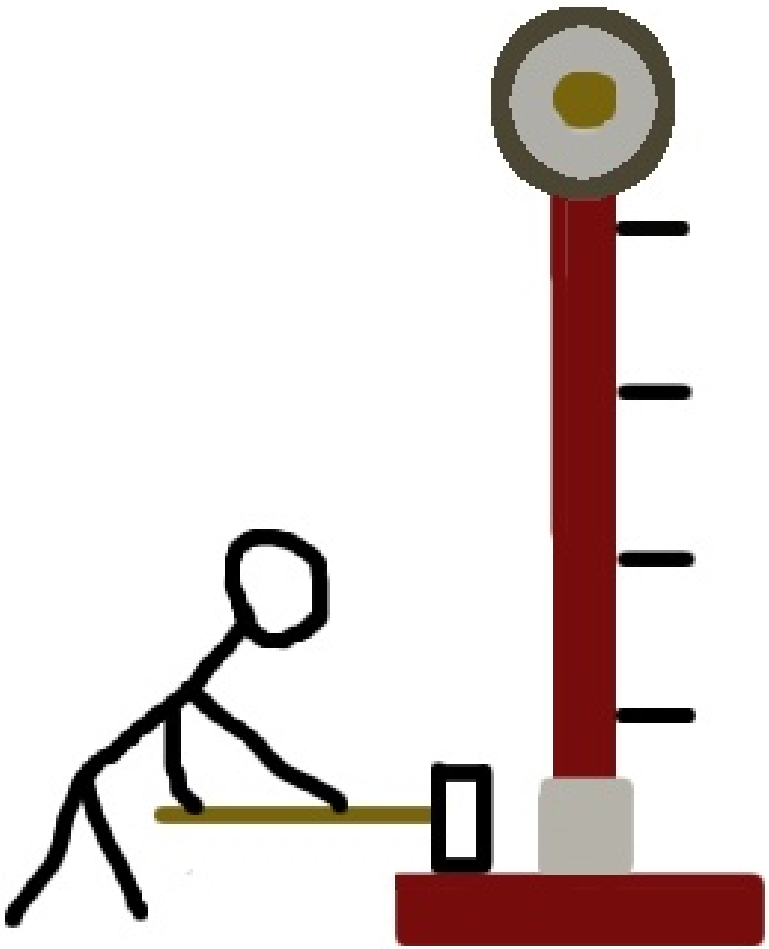
5 June, 2010

Supported by: SAMSI Population Monte Carlo group and NSF CAREER grant DMS-05-48153

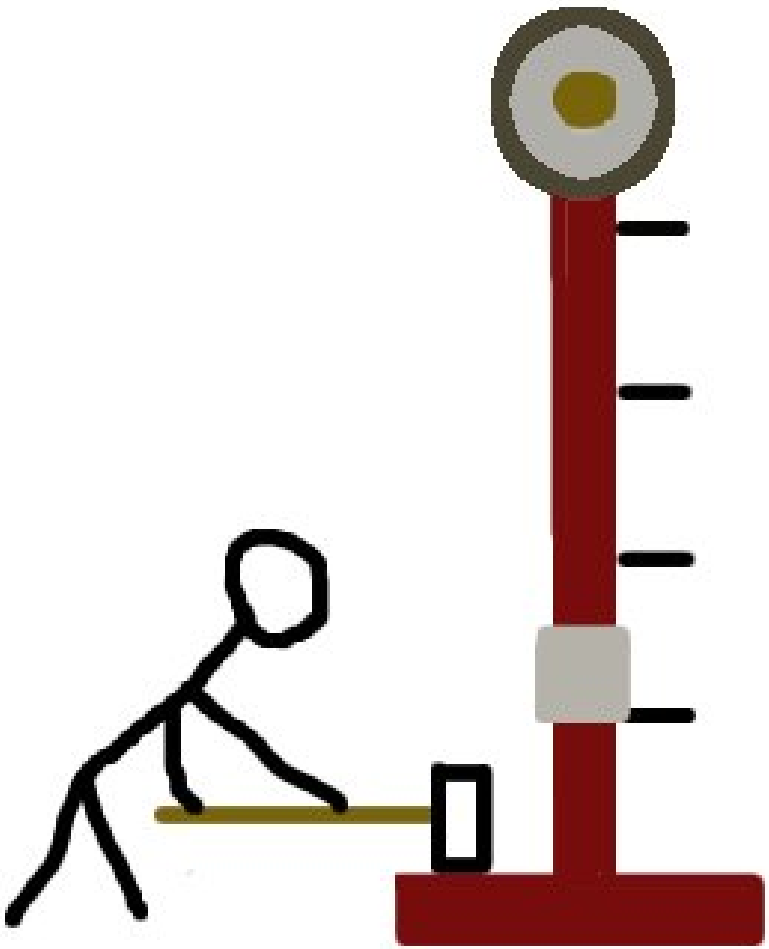
Monte Carlo Integration


ICH DIEN
100
90
80
70
60
50
40
30
20
10
A Prize Every Time
3 Strikes £1.50 Serve
Test Your Strength
GAMES FOR ALL AGES
STAND CLEAR
The Mighty Striker
EVERY ONE A WINNER
STAND CLEAR

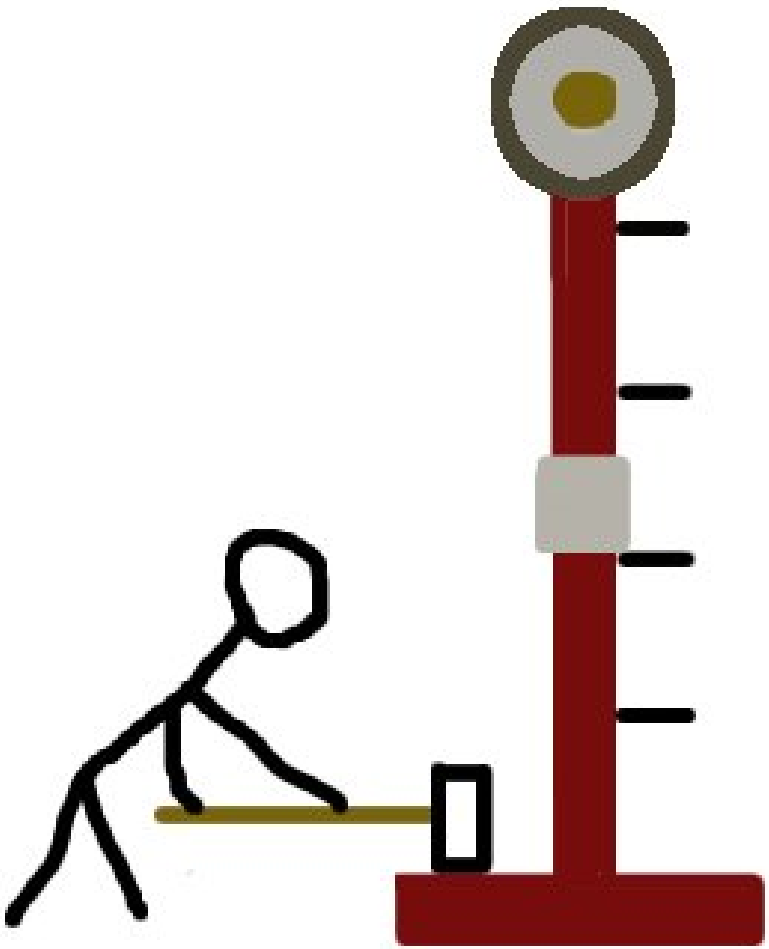
Have to ring bell in one shot...



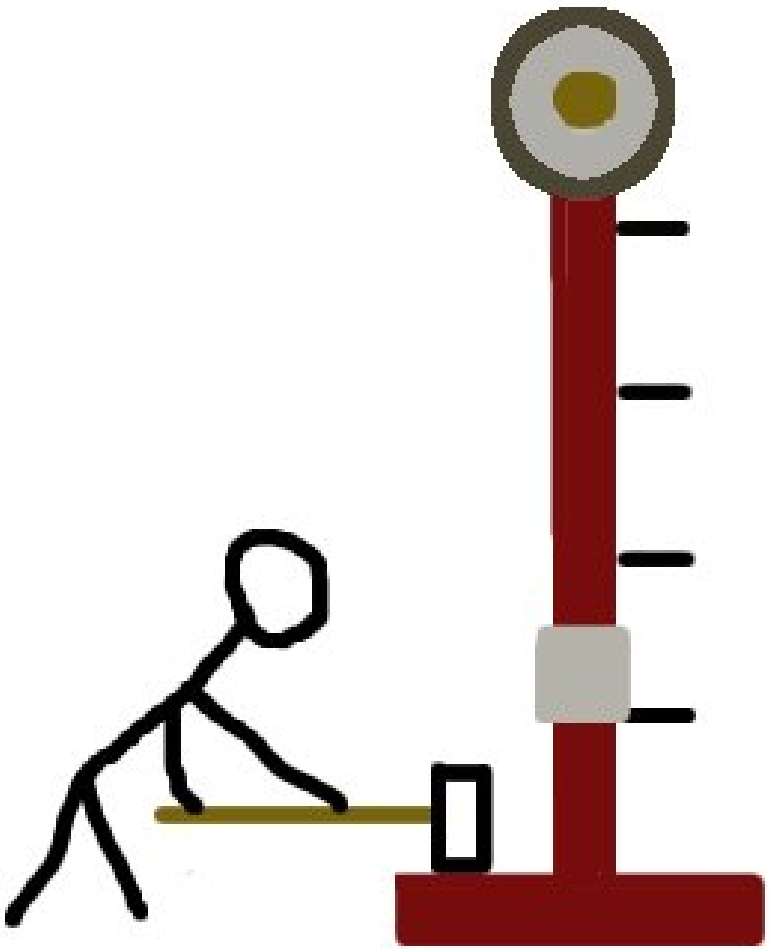
Have to ring bell in one shot...



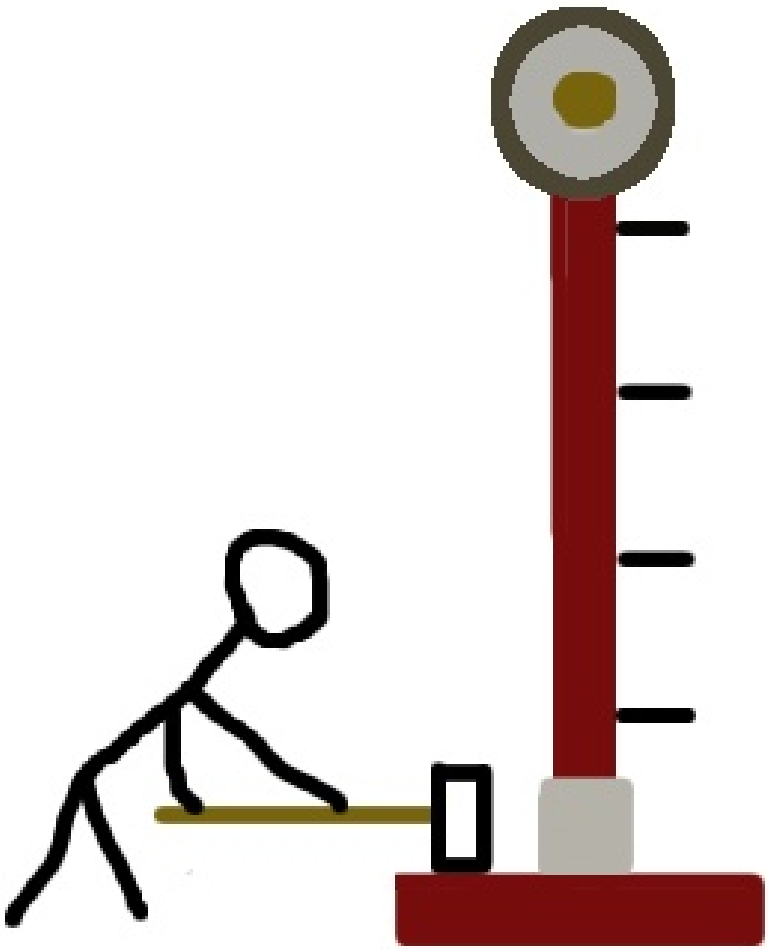
Have to ring bell in one shot...



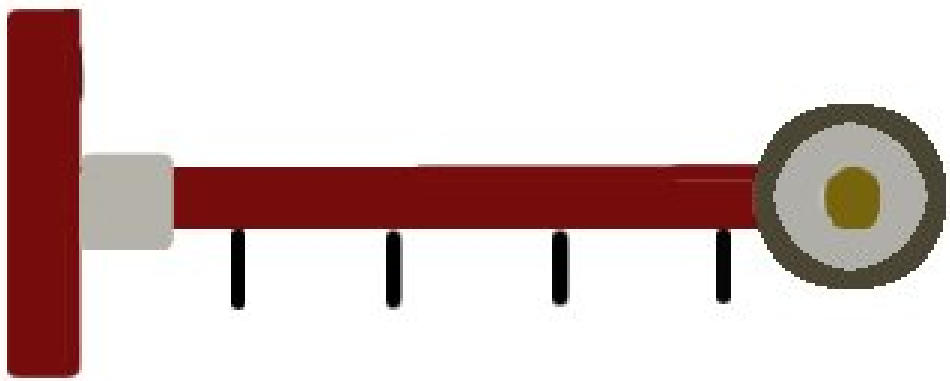
Have to ring bell in one shot...



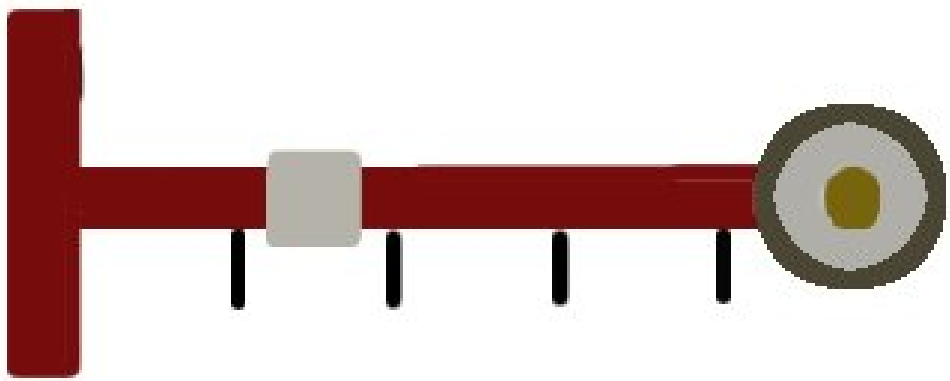
Have to ring bell in one shot...



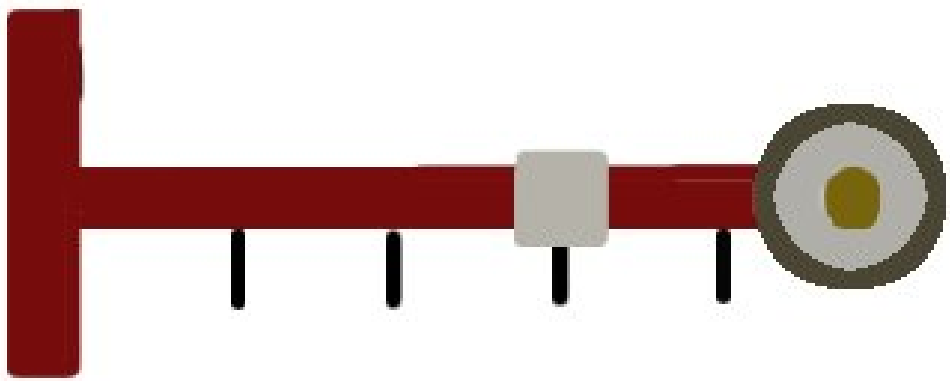
Easier if you could tip over...



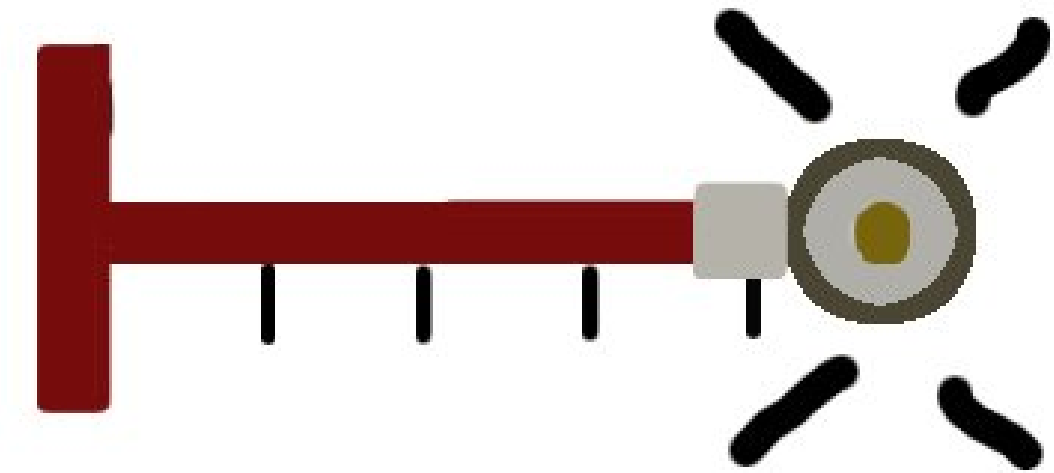
Easier if you could tip over...



Easier if you could tip over...



Easier if you could tip over...



Two methods of Monte Carlo Integration

Hit it in one shot = Acceptance/Rejection

Multiple hits = TPA

Classical Monte Carlo Integration

Suppose want integrated likelihood

$$Z = \int_{\text{parameter space}} \text{likelihood}\, d\mu_{\text{prior}}$$

Find random variable X such that

$$\mathbb{E}[X] = Z$$

Approximate Z with sample average of n iid draws from X

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad X_1, \ldots, X_n \overset{\text{iid}}{\sim} X$$

The accuracy problem

The standard deviation of estimate typically unknown

$$\mathrm{SD}(\bar{X}) = \frac{1}{\sqrt{n}}\,\mathrm{SD}(X)$$

Tails could be heavy

Goal

Develop approximation with bounded tails

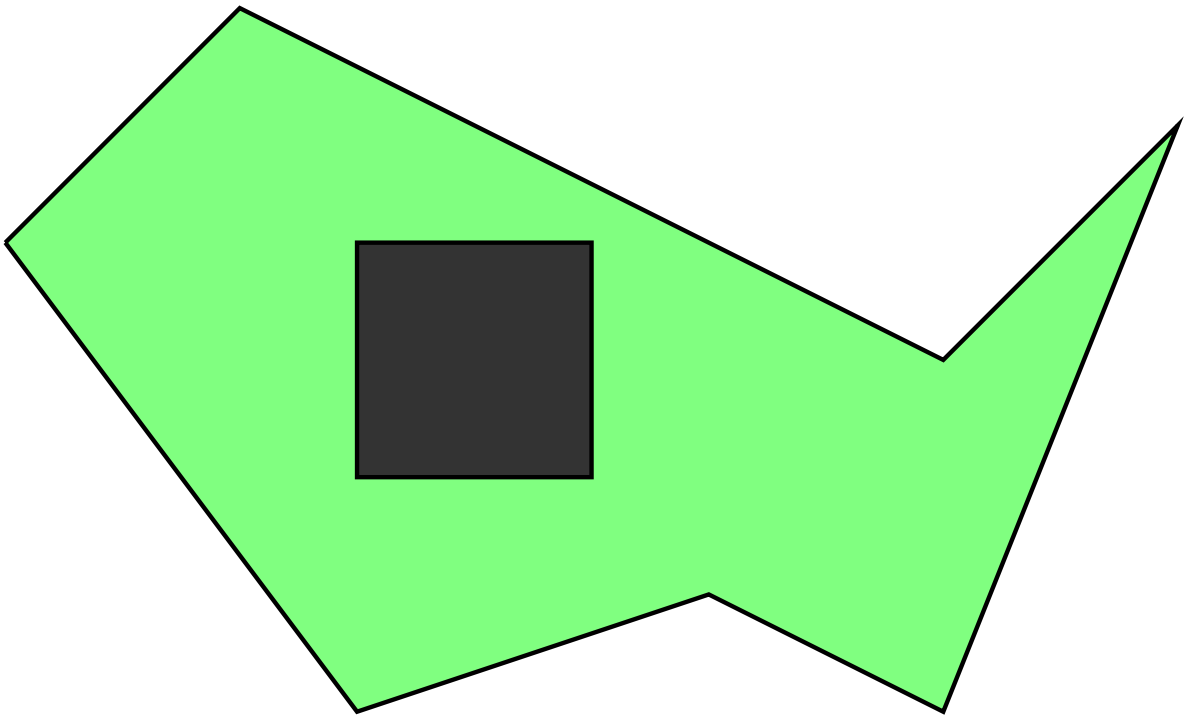
- Given $\epsilon, \delta > 0$ want

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

- No need to worry about unknown variance

Abstract problem: find the measure of a set

Find easy set inside hard set

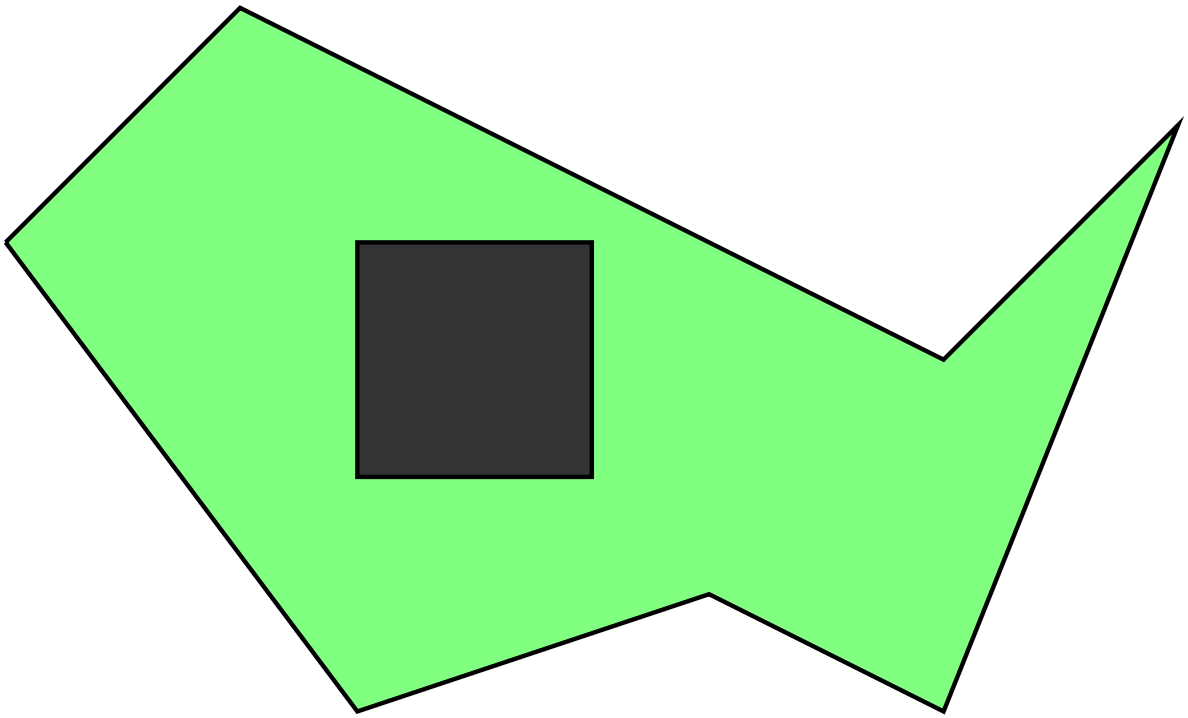


Green area = B

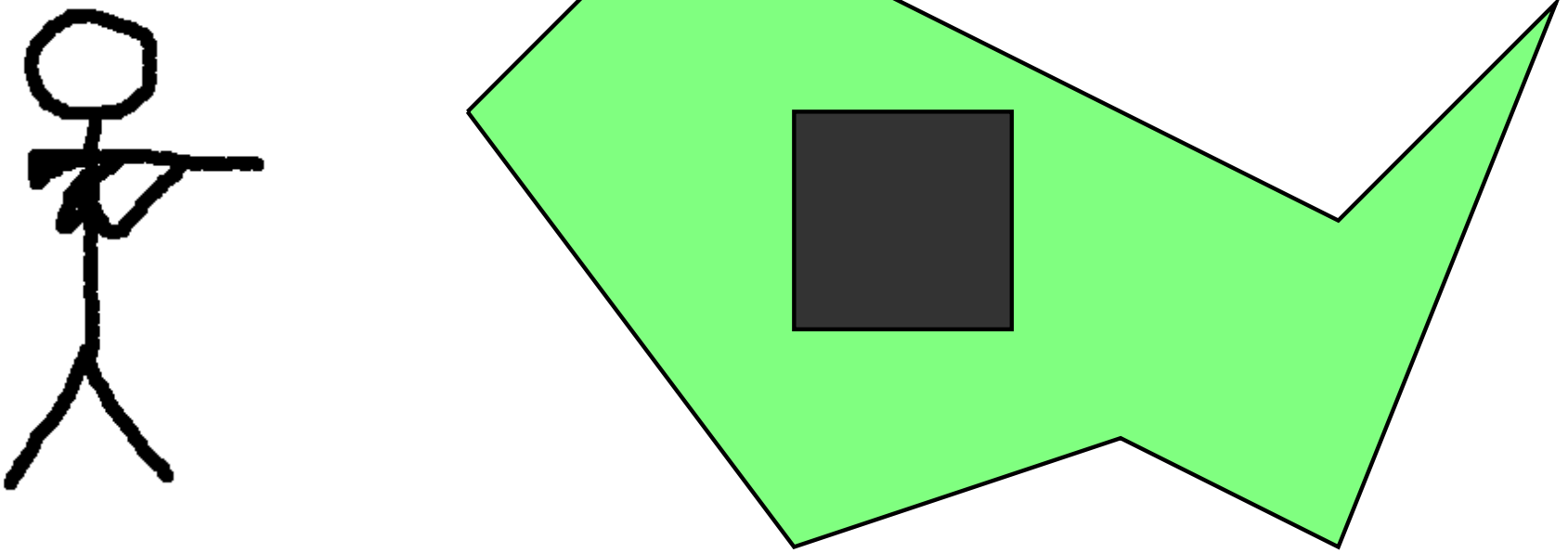
Black area = B'

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

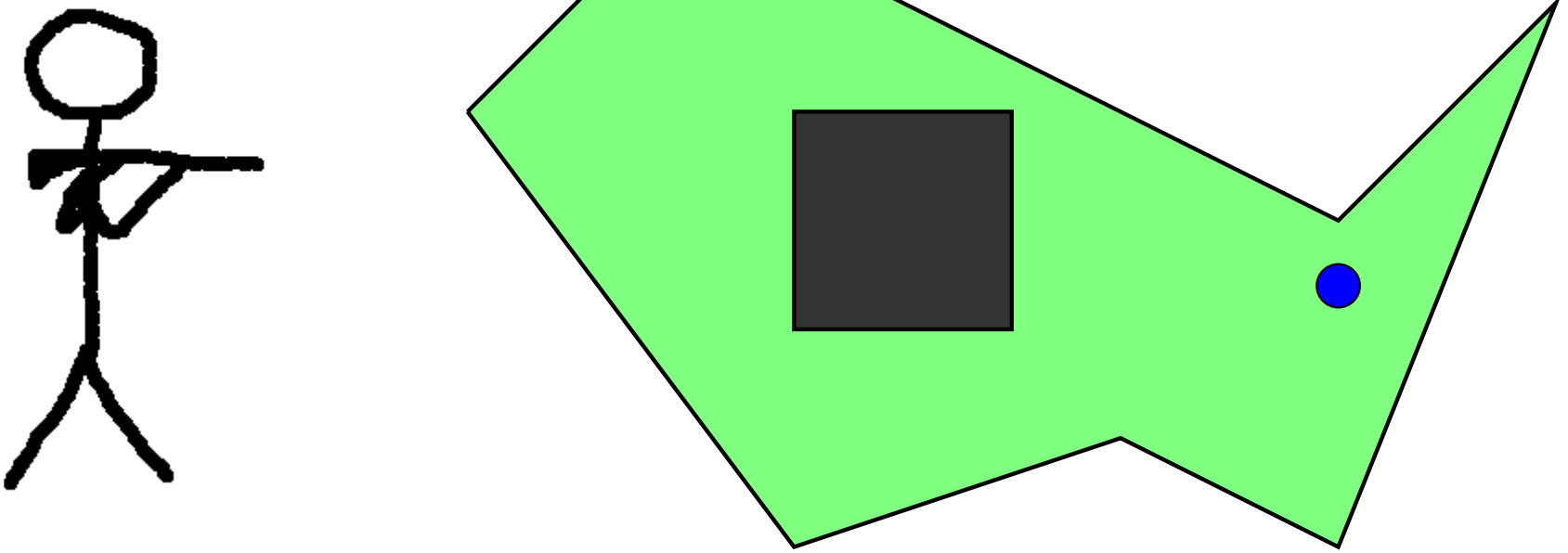
Acceptance/Rejection a.k.a “Shoot at it randomly”



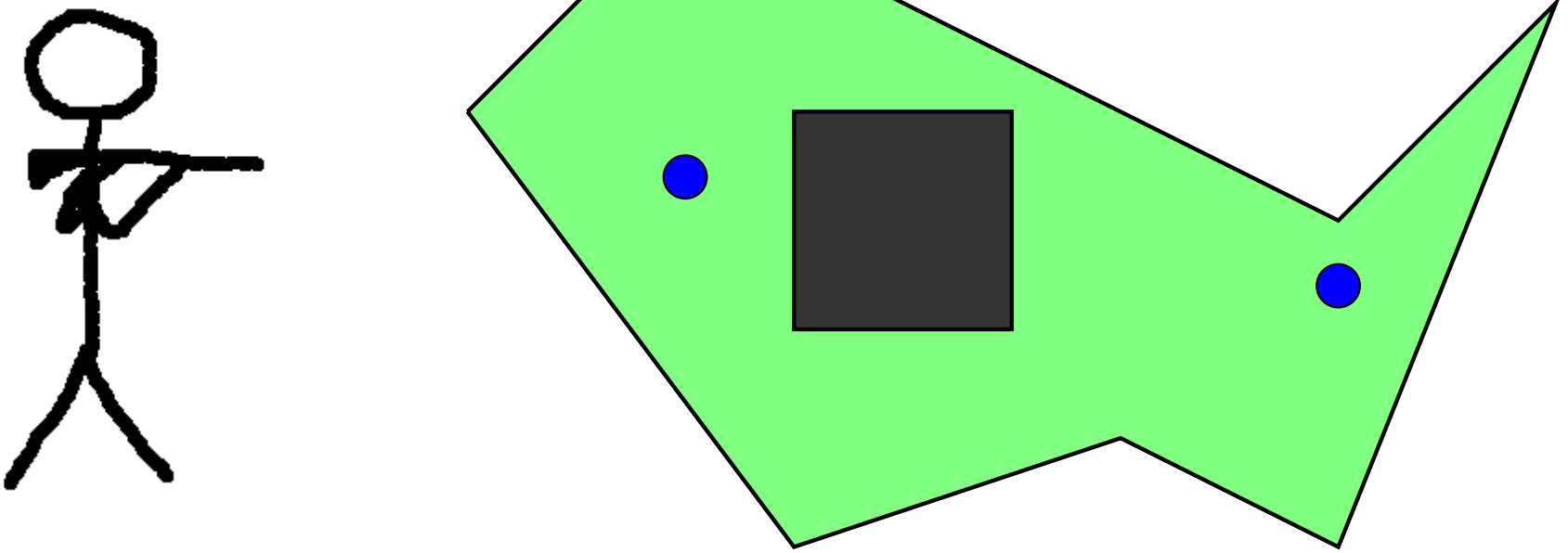
Acceptance/Rejection a.k.a “Shoot at it randomly”



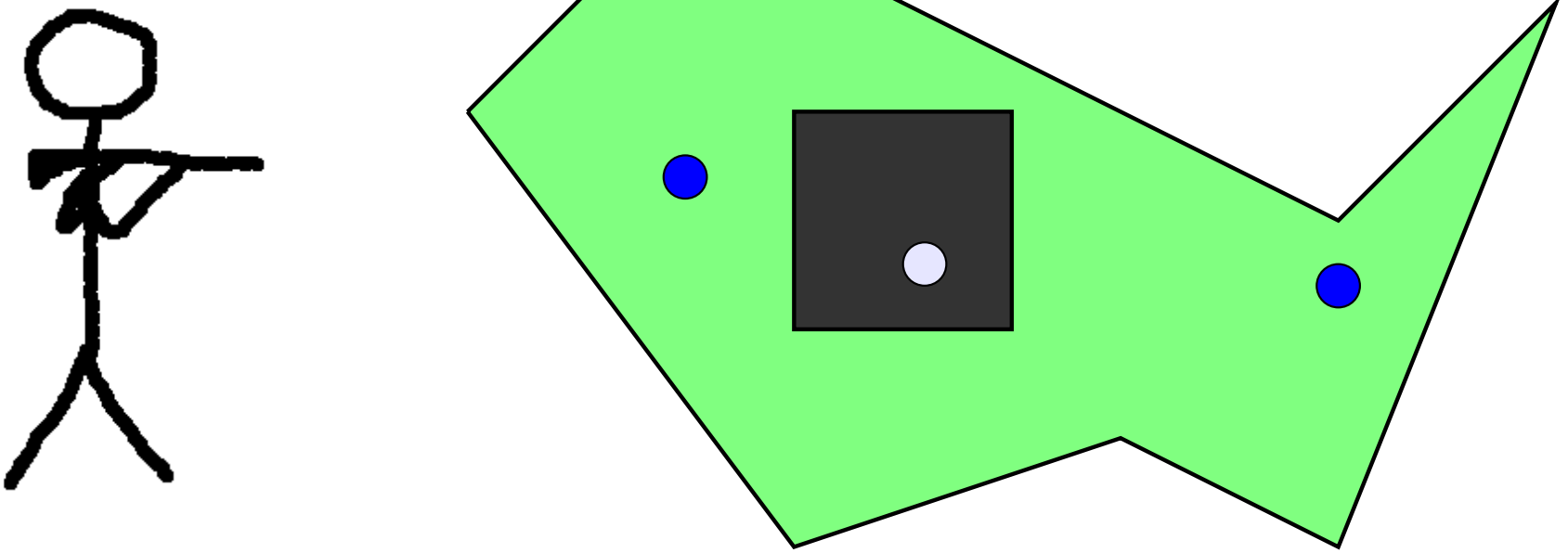
Acceptance/Rejection a.k.a “Shoot at it randomly”



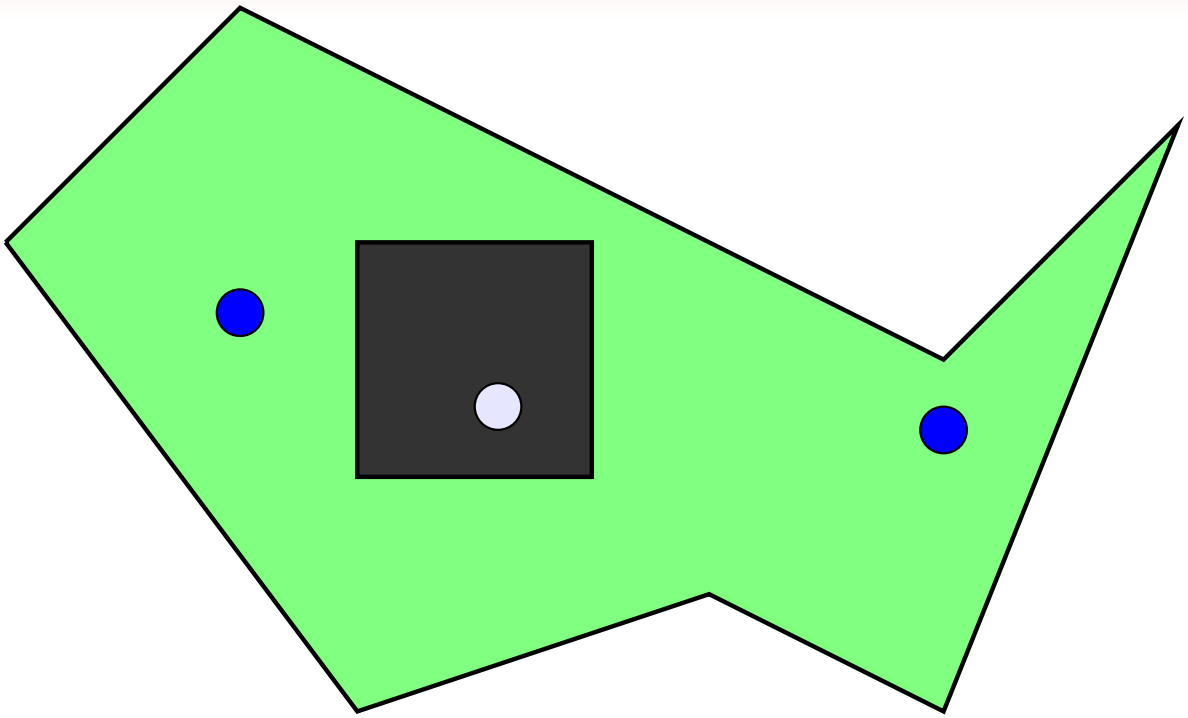
Acceptance/Rejection a.k.a “Shoot at it randomly”



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



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



Best estimate:

$$\hat{\mu}(B) = 3\mu(B'), \quad B' = \text{black rectangle inside region}$$

How many samples need to be taken?

Let

$$p = \frac{\mu(B')}{\mu(B)} \quad \text{(probability draw from } B \text{ lands in } B')$$

Number of samples used by algorithm

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

Acceptance/Rejection:

$$f_{\mathrm{AR}}(p) = \frac{1}{p}$$

Running times

Acceptance/Rejection:

$$\frac{1}{p}$$

Product Estimator [1]:

$$81 \cdot \left[\log \frac{1}{p}\right]^2$$

TPA:

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

The New Algorithm: TPA

Nested Sampling [Skilling 2006]

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 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?

Key Idea:

Acceptance/Rejection

- Try to get to center in one step
- If fail, start over

Tootsie Pop Algorithm

- Try to get to center
- If fail, start from current position

List of ingredients of TPA

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

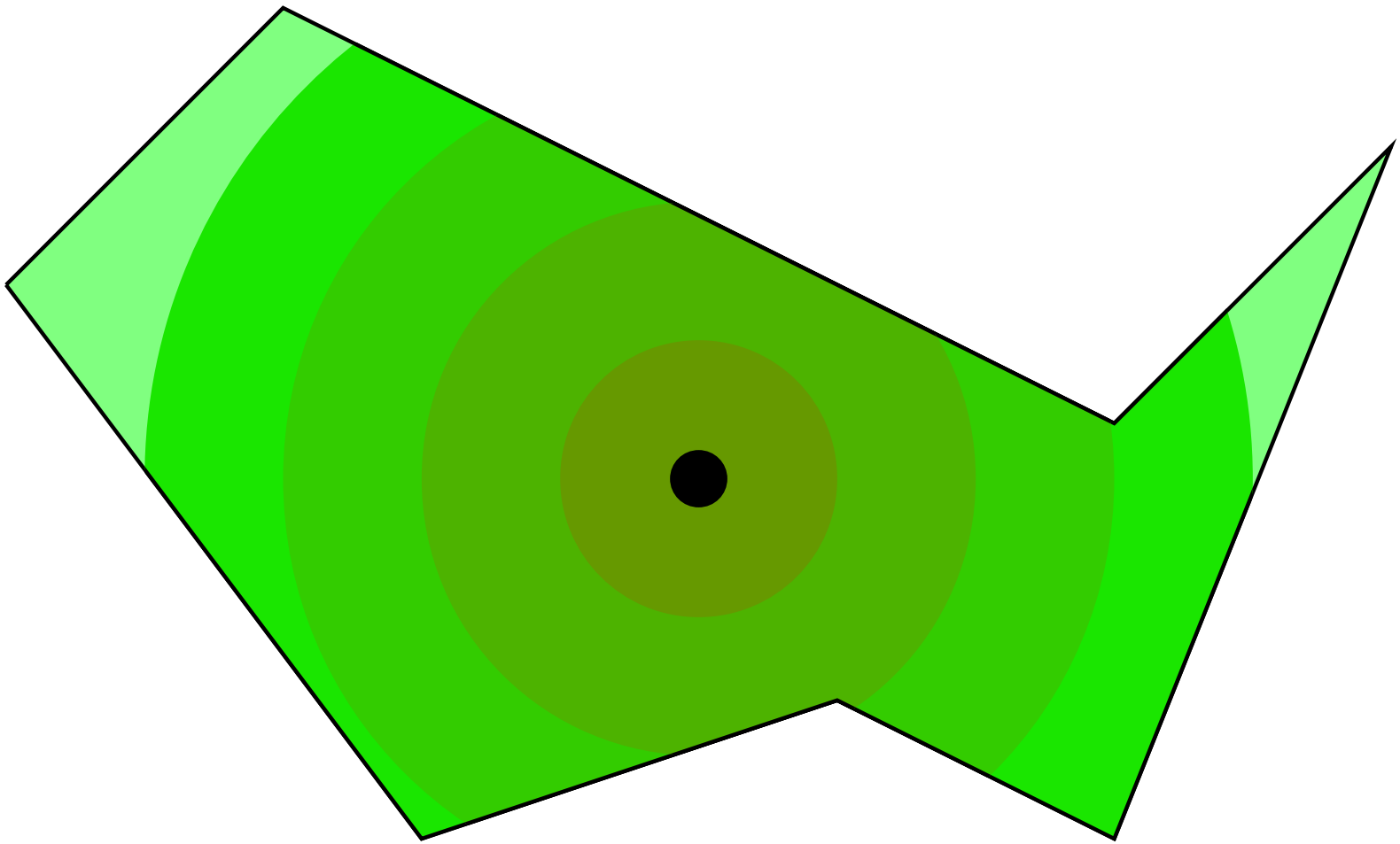
(b) Two measurable sets: the *center* B' and the *shell* B with $B' \subset B$

(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

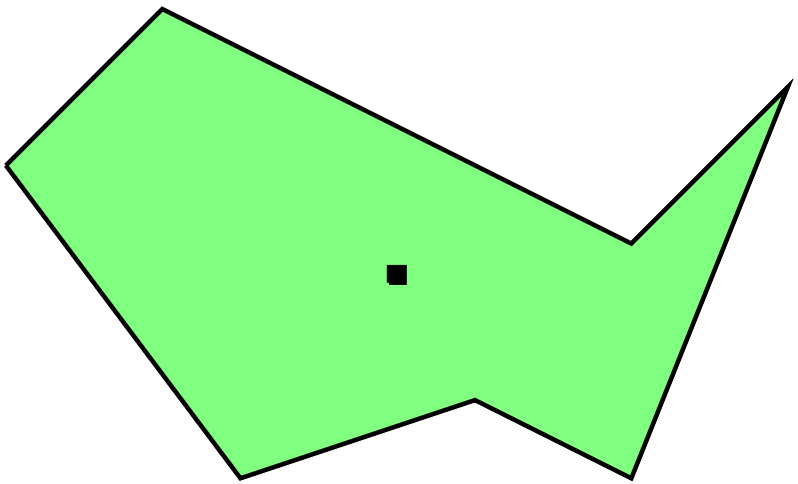
(d) Two special values β_B and $\beta_{B'}$ with $A(\beta_B) = B$ and $A(\beta_{B'}) = B'$.

Example of nested sets



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

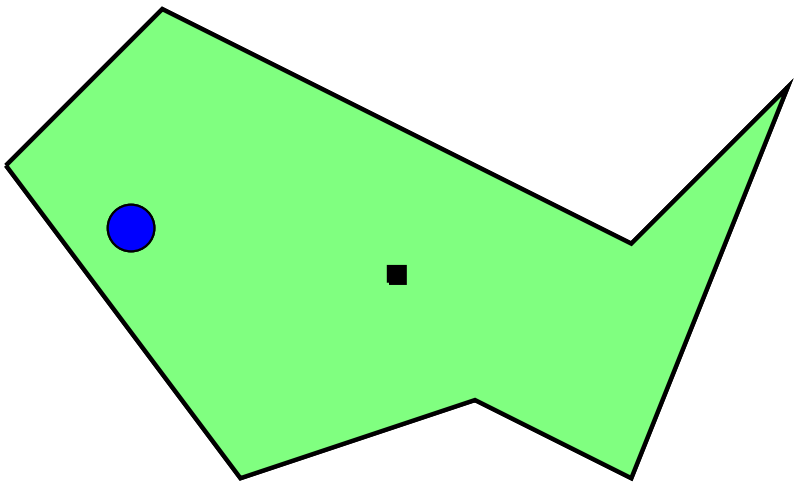
Idea behind TPA



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 $\beta \leq \beta_{B'}$

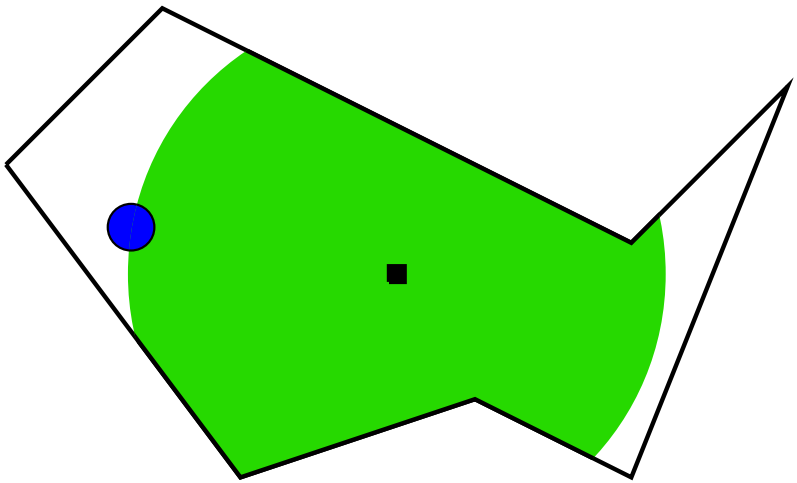
Idea behind TPA



Step	β
0	∞
1	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 $\beta \leq \beta_{B'}$

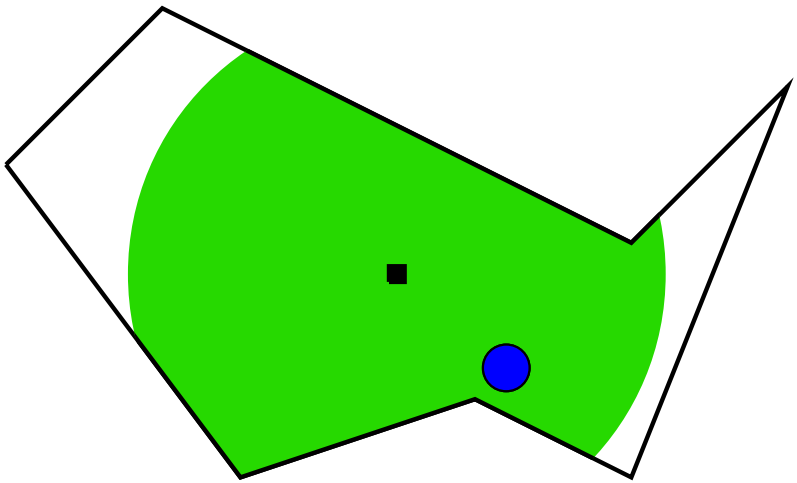
Idea behind TPA



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 $\beta \leq \beta_{B'}$

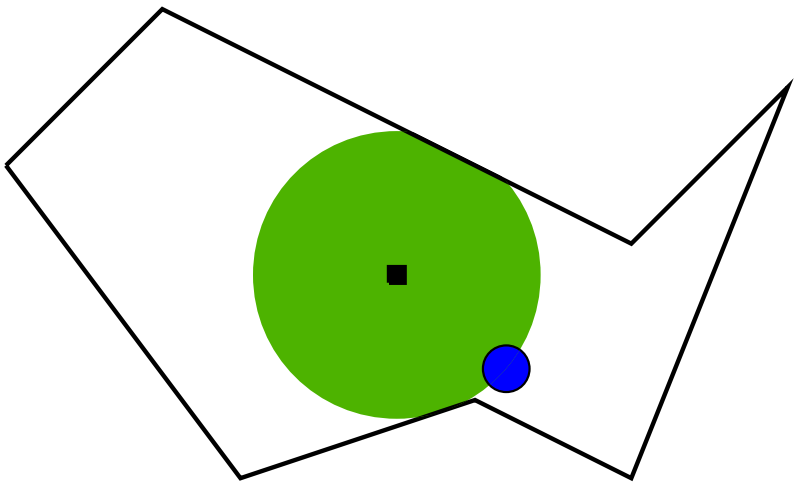
Idea behind TPA



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 $\beta \leq \beta_{B'}$

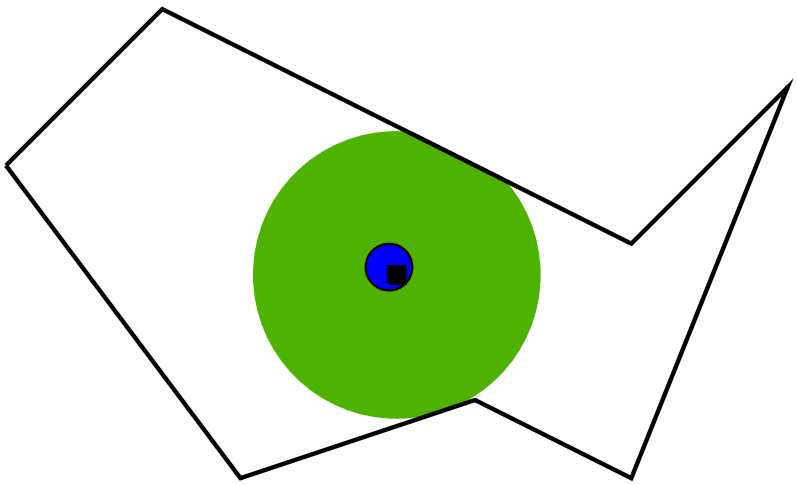
Idea behind TPA



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 $\beta \leq \beta_{B'}$

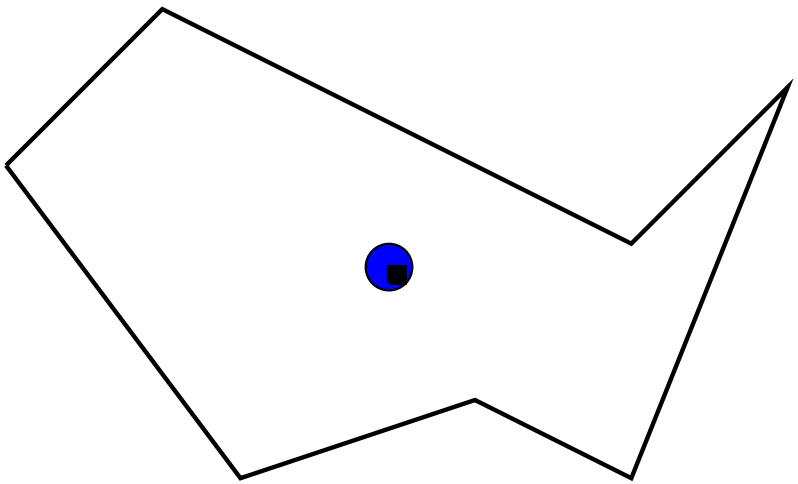
Idea behind TPA



Step	β
0	∞
1	1.72
2	0.92
3	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 $\beta \leq \beta_{B'}$

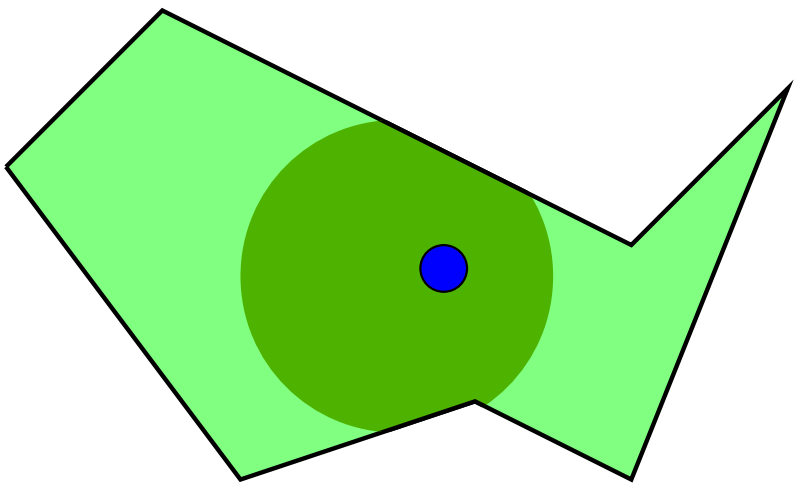
How much is shaved off at each step?

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

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

Product of uniforms

After k steps, measure is:

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

Taking the logarithm:

$$\ln Z(\beta_k) = \ln Z(\beta_B) + \ln r_1 + \ln r_2 + \cdots + \ln r_k, \text{ where}$$
$$-\ln r_1, -\ln r_2, \ldots, -\ln r_k \overset{\text{iid}}{\sim} \text{Exp}(1)$$

So $\{\ln(Z(\beta_i))\}_{i=1}^k$ forms a Poisson point process!

The result

Output of TPA:

- $\ell \sim \text{Poisson}(\ln(Z(\beta_B)/Z(\beta_{B'})))$
- $\mathbb{E}[\ell] = \ln(1/p)$, $\mathbb{V}(\ell) = \ln(1/p)$

Output of A/R:

- $H \sim \text{Geometric}(Z(\beta_{B'})/Z(\beta_B))$
- $\mathbb{E}[H] = 1/p$, $\mathbb{V}(H) = (1-p)/p^2$

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 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

Examples

Example 1: 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 Bin(n_i, p_i)
- n_i are known, $p_i \sim$ Beta(a, b)
- Hyperparameters a and b, $a \sim 1 +$ Exp(1), $b \sim 1 +$ Exp(1)

Natural question

Is this a good model?

- Use Bayes' Factor...
- Need *evidence* aka *integrated likelihood*

$$Z = \int_{\text{parameter space}} \text{likelihood } d\mu_{\text{prior}}$$

- So we have an integration problem

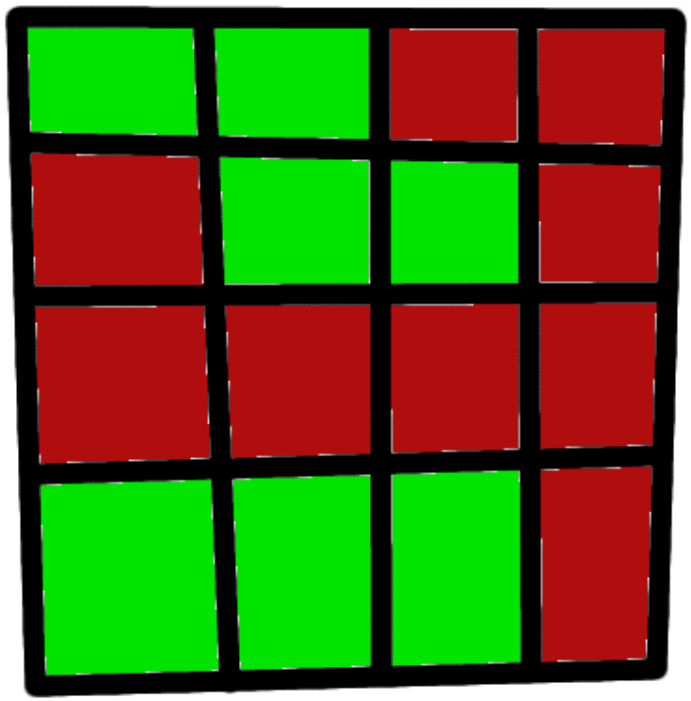
Parameter trunctation

Goal: find integrated likelihood

- Pick a point (a_0, b_0) in (a, b) space
- Let $A(\beta)$ be (a, b) values within distance β of (a_0, b_0)
- 2-D Unimodal problem so sampling from posterior easy
- True value $\ln Z$ (via numerical integration) -1577.250
- After 10^5 runs $\ln \hat{Z} = -1577.256$

Example 2: 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

Putting Ising into framework

Add an auxilliary variable

$$[Y|X] \sim \text{Unif}([0, \exp(2\beta h(X))])$$

With Y:

- (X, Y) uniform over weird shaped space
- As β grows, more allowable values for Y
- When $\beta = 0$, $Y \sim \text{Unif}([0, 1])$ independent of X
- So $Z(0) = 2^V$ (number of configurations of X)

Method works for any exponential family

Integrated likelihood for Ising

Often called "doubly-intractable"

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

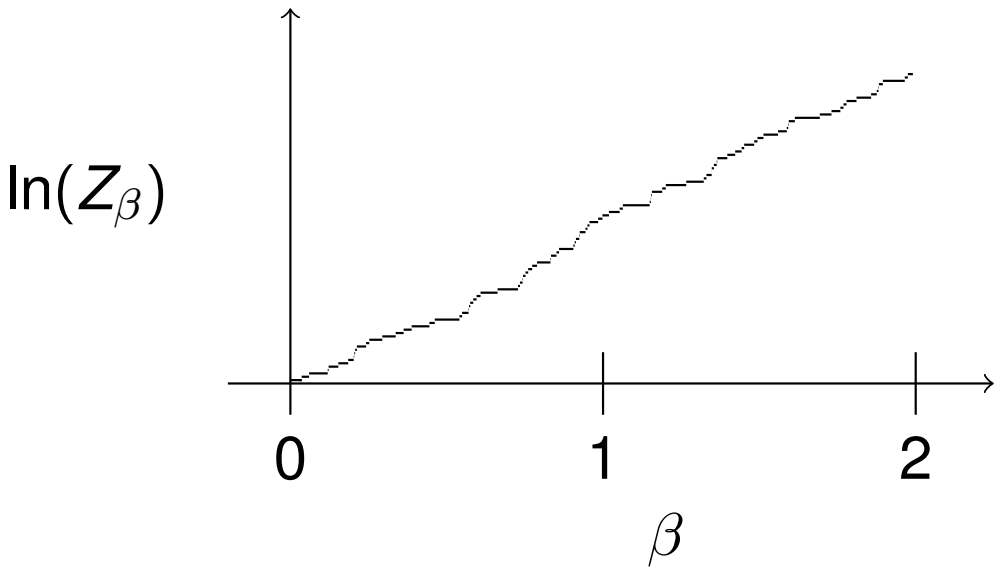
Intractable part is function $Z(b)$

- Parameter only one dimensional
- Easy to do numerically if you know $Z(\beta)$ over $(0, \infty)$

Poisson point process runs from $\ln(Z(\infty))$ **to** $\ln(Z(0))$

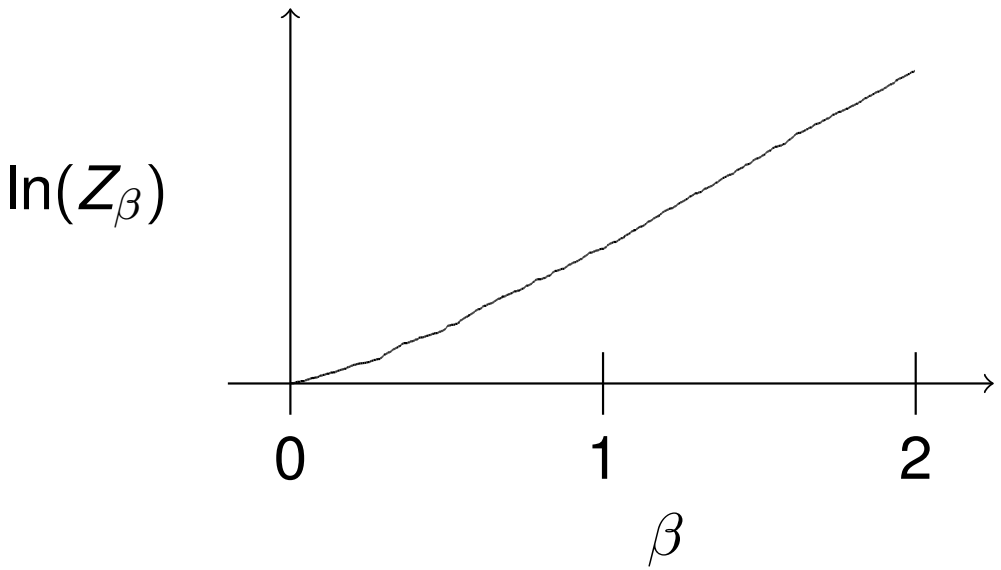
- Yields entire function $Z(b)$ for b from 0 to ∞
- Called *omnithermal* approximation

Use omnithermal approximation



One run of TPA

Use omnithermal approximation



Sixteen runs of TPA

Final thoughts

This is not nested sampling

Key differences

- Running time TPA same order as nested sampling
- Distribtution of output of TPA known exactly, nested sampling asymptotically
- Nested sets in nested sampling tend to exaggerate multimodality
- Nested sets in TPA tends to remove multimodality
- Nested sampling Monte Carlo/numerical analysis hybrid
- TPA pure Monte Carlo (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: $[\ln(1/p)]^2$ much better than product estimator

What else is in paper?

- Artificial multimodal example
- Using TPA with slice sampling
- Using TPA to build balanced cooling schedule

Next step

- $10^9 \ln(1/p)$ method for exponential families
- Remove constant with recursive adaptive scheduling

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