7. Volume estimates and Pagerank

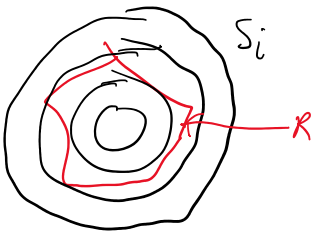
Wednesday, September 22, 2021 6:14 PM

Today:
- Estimating volume using MCMC
- Page Rank
- Overview of topics in data science

Estimating volume of convex sets

Rejection sampling fails in high dimensions vol of interior object compared to exterior goes to 0.

Instead, for a convex set R, choose concentric circles $S_1, \dots, S_k$ s.t. $S_1 \subseteq R \subseteq S_k$



Then $\mathrm{Vol}(R) = \mathrm{Vol}(S_k \cap R)$

$$= \frac{\mathrm{Vol}(S_k \cap R)}{\mathrm{Vol}(S_{k-1} \cap R)} \cdot \frac{\mathrm{Vol}(S_{k-1} \cap R)}{\mathrm{Vol}(S_{k-2} \cap R)} \cdots \frac{\mathrm{Vol}(S_2 \cap R)}{\mathrm{Vol}(S_1 \cap R)} \cdot \mathrm{Vol}(S_1)$$

If the radius of S_i is $1 + \frac{1}{d}$ times the radius of S_{i-1}, then $\mathrm{Vol}(S_i) = \left(1 + \frac{1}{d}\right)^d \mathrm{Vol}(S_{i-1})$

$$\Rightarrow 1 \le \frac{\mathrm{Vol}(S_i \cap R)}{\mathrm{Vol}(S_{i-1} \cap R)} \le \left(1 + \frac{1}{d}\right)^d < e$$

the larger circle contains more of R.

$$\frac{\mathrm{Vol}(S_i \cap R)}{\mathrm{Vol}(S_i)} \le \frac{\mathrm{Vol}(S_{i-1} \cap R)}{\mathrm{Vol}(S_{i-1})}$$

by convexity, for each ray from from the origin, the fraction of that ray occupied cannot increase

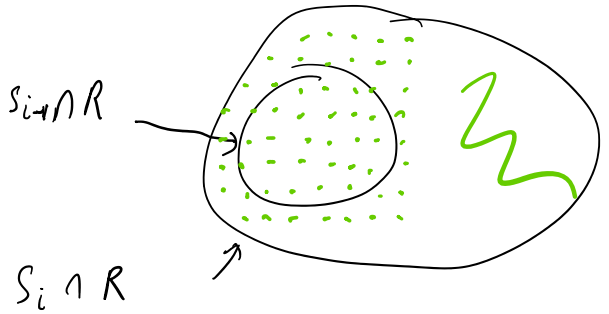
Thus, if we can sample from $S_i \cap R$, then we can approximate $\frac{\mathrm{Vol}(S_i \cap R)}{\mathrm{Vol}(S_{i-1} \cap R)}$ by rejecting sampling since accepted pts are $\ge \frac{1}{e}$ fraction

The number of spheres needed is $k \le 2 \log_{1+\frac{1}{d}} r$, where $r = \frac{\text{radius}(S_k)}{\text{radius}(S_1)}$.

$$\Rightarrow k \le \frac{2 \ln r}{\ln\left(1 + \frac{1}{d}\right)} \le \frac{2 \ln r}{\frac{1}{2d}} = 4d \ln r$$

So, if each ratio is estimated to error $1 \pm \frac{\varepsilon}{4ed \ln r}$, we can estimate

So, if each ratio is estimated to error $1 \pm \frac{\varepsilon}{4edlnr}$, we can estimate overall volume to error $1 \pm \varepsilon$.



How to estimate ratios?

Use Metropolis-Hastings on rectangular grid

The stationary dist. that is uniform on all pts inside $S_i \cap R$,

So we just need a random walk along the undirected graph with nodes λ $S_i \cap R$.

Fast convergence due to grid structure.

Pagerank

Consider the WWW (world wide web) of "webpages"

If you are a search engine, you two tasks:

(1) Find webpages containing a search term

(2) rank by importance

Model a "websurfer" as a random walk on the hyperlinks

Not strongly connected, so add random restarts to give stationary prob. dist.

Stationary probabilities are "Pagerank", and correspond to the frequency with which a page will be randomly visited over a period of time.