Monte Carlo methods

outline

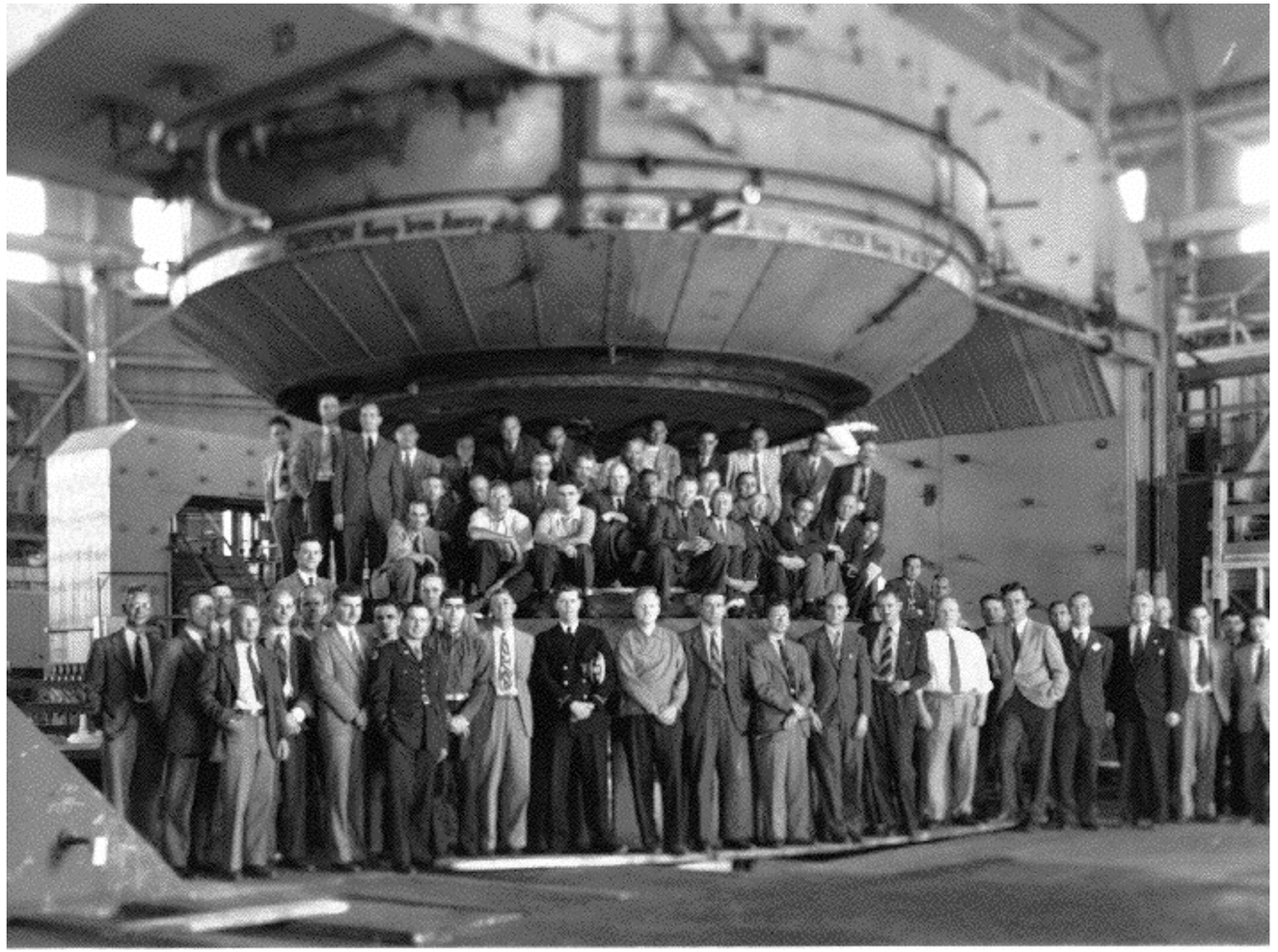
Stochastic Processes in Science Inference

- History of Monte Carlo Methods
- Application of MC to probabilistic inference
- A simple MC simulation
- MC simulations applications in Urban Science - Traffic flow, Resque
- Rejection & Importance Sampling

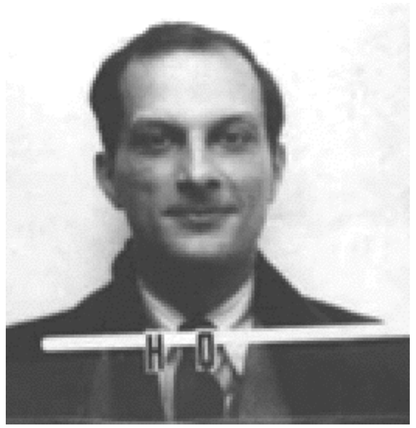
Markov Chain Monte Carlo

- Markovian Processes and Markov chains
- Bayes theorem and the posterior distribution
- Metropolis-Hastings (and Gibbs sampling) MCMC
- Affine Invariant MCMC
- convergence criteria

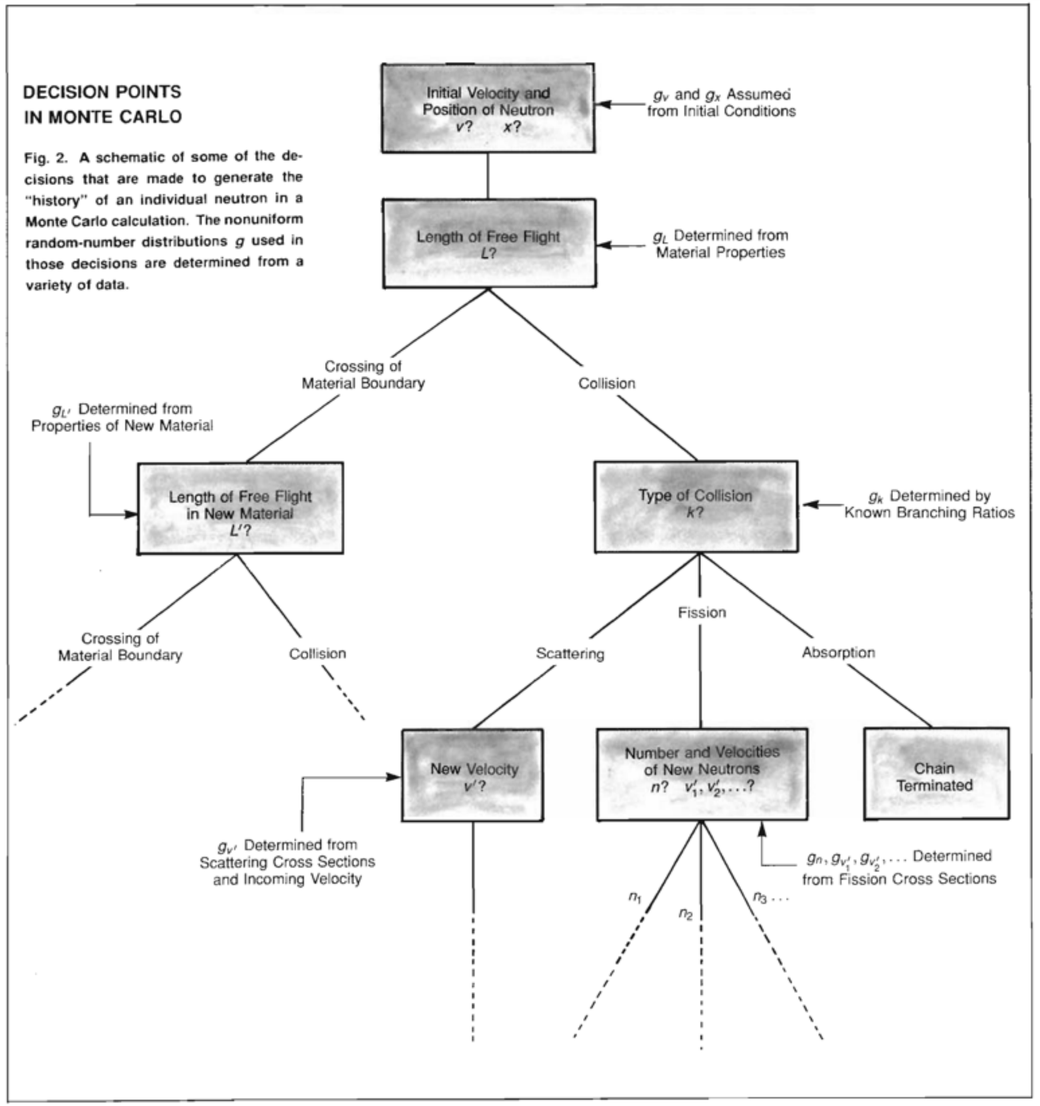
The Manhattan Project



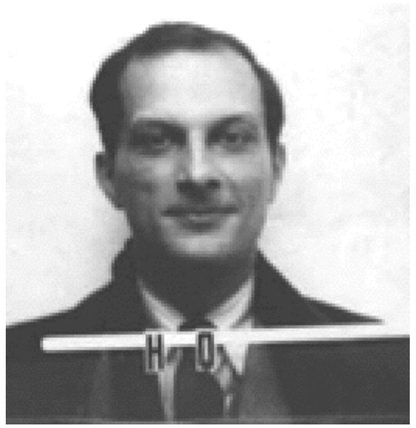
MC - history



Stanislav Ulam



MC - history



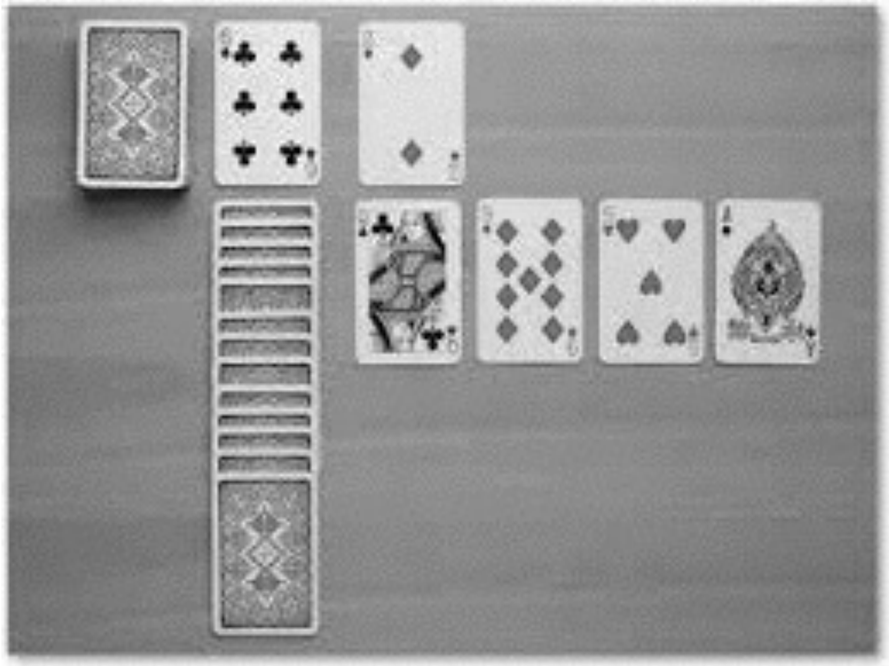
Stanislav Ulam

What are the chances that a Canfield solitaire laid out with 52 cards will come out successfully?

The number of different games is

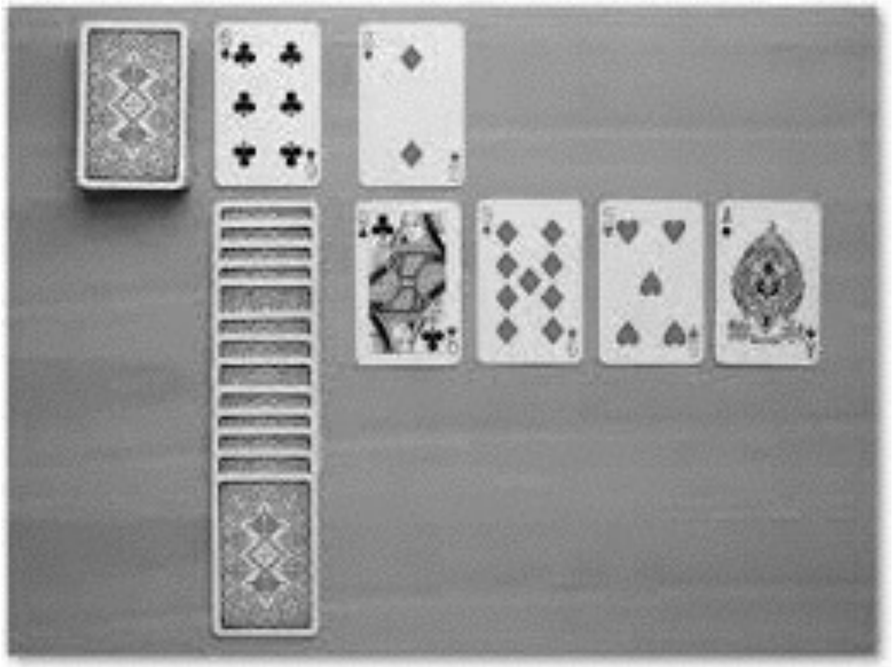
$52! = 52x51x52\ldots x3x2x1 \sim 8x10^{67}$

Canfield Solitaire



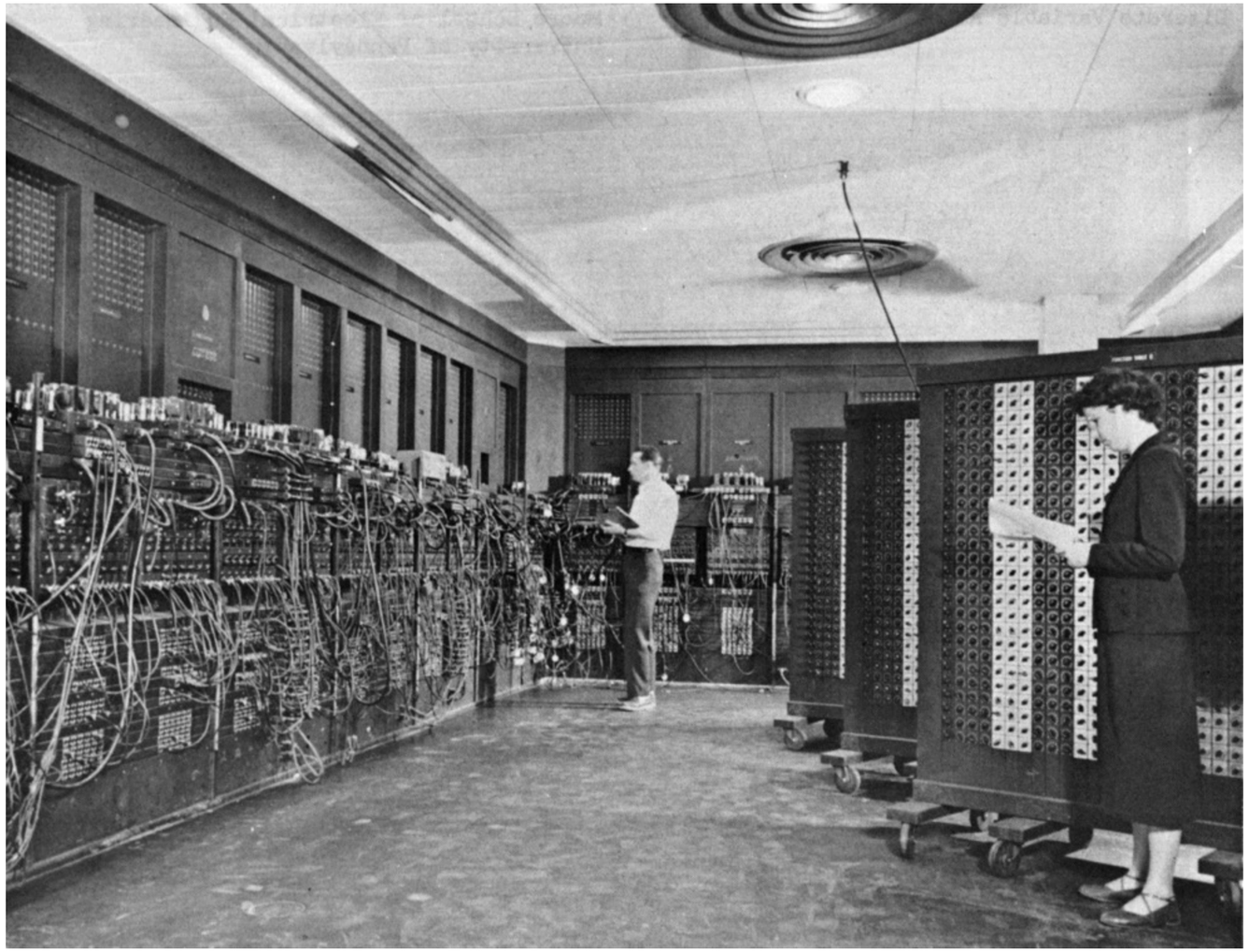
MC - history

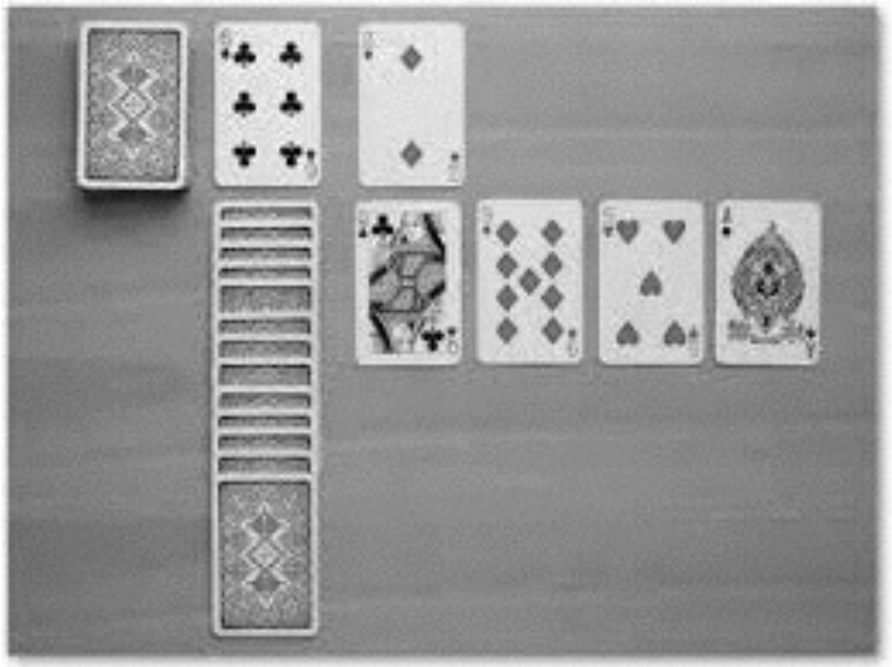
"What are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether **a more practical method than *abstract thinking* might not be to lay it out say one hundred times and simply observe and count the number of successful play"**



http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-88-9068

MC - history





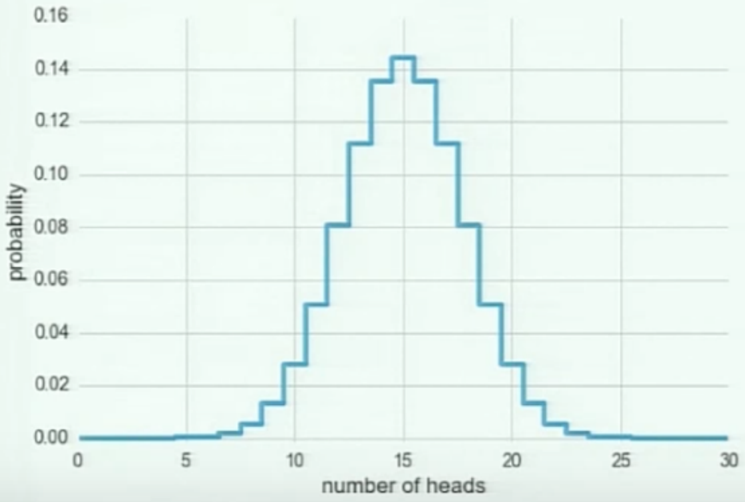
ENIAC It weighed more than 30 short tons (27 t), was roughly 2.4 m × 0.9 m × 30 m (8 × 3 × 100 feet) in size, occupied 167 m^2 (1,800 ft^2), consumed 150 kW of electricity.
500FLOPS vs today's Macbook pro ~1TeraFLOP

MC - history

Classic Method:

$N_H = 22,\ N_T = 8$

$$P(N_H, N_T) = \binom{N}{N_H}\left(\frac{1}{2}\right)^{N_H}\left(1-\frac{1}{2}\right)^{N_T}$$



Easier Method:

Just simulate it!

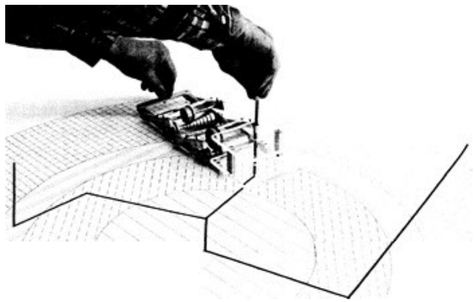
```
M = 0
for i in range(10000):
    trials = randint(2, size=30)
    if (trials.sum() >= 22):
        M += 1
p = M / 10000 # 0.008149
```

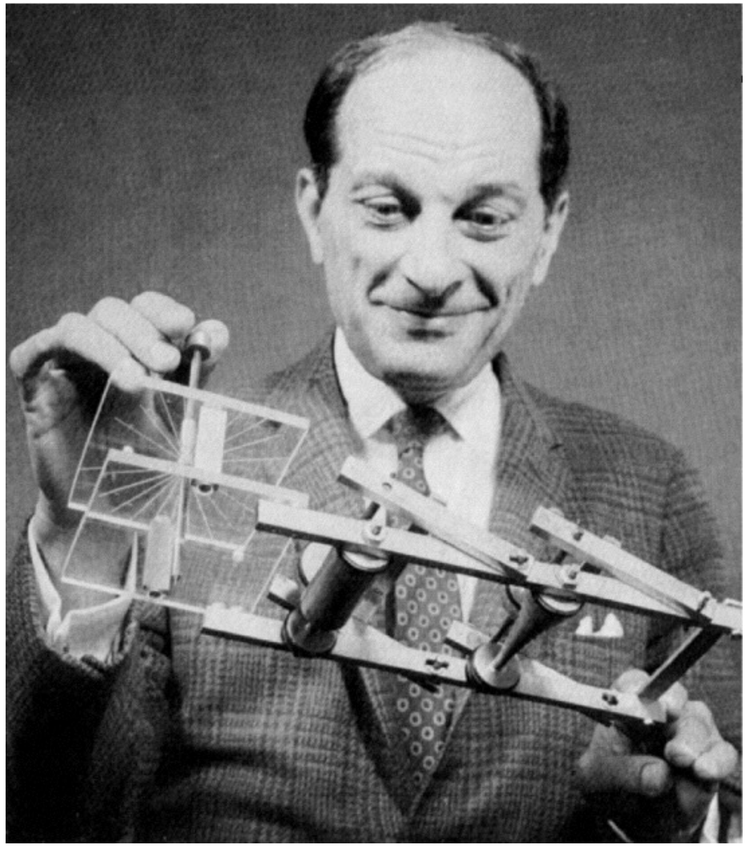
→ reject fair coin at p = 0.008

Statistics for Hackers, Jake Vanderplas PyCon16
https://www.youtube.com/watch?v=Iq9DzN6mvYA

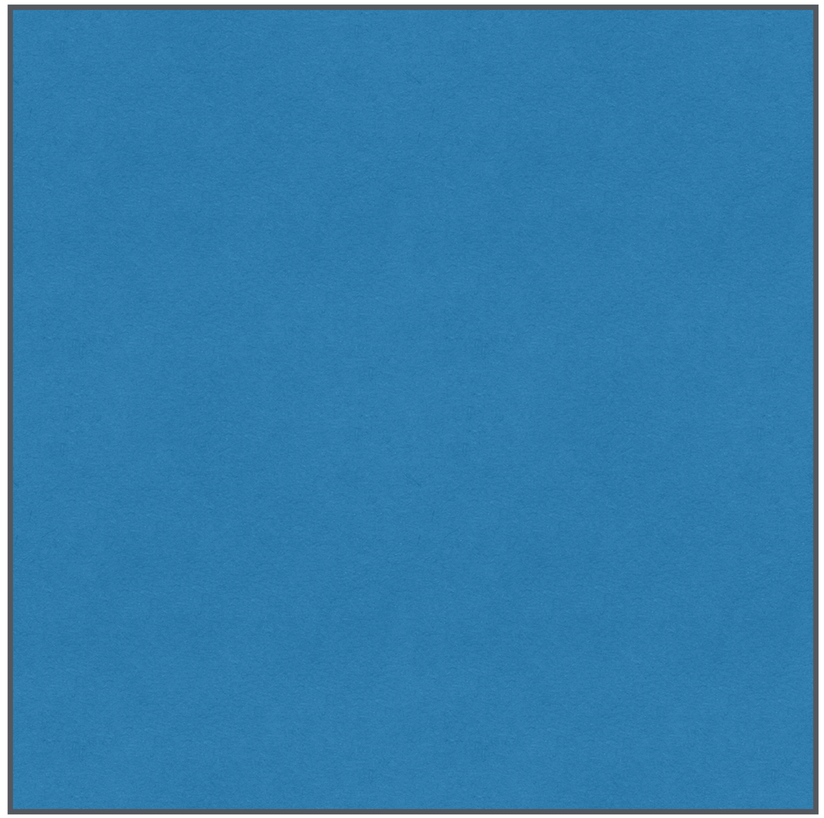
The Fermiac

Enrico Fermi looked really smart with his predictions…



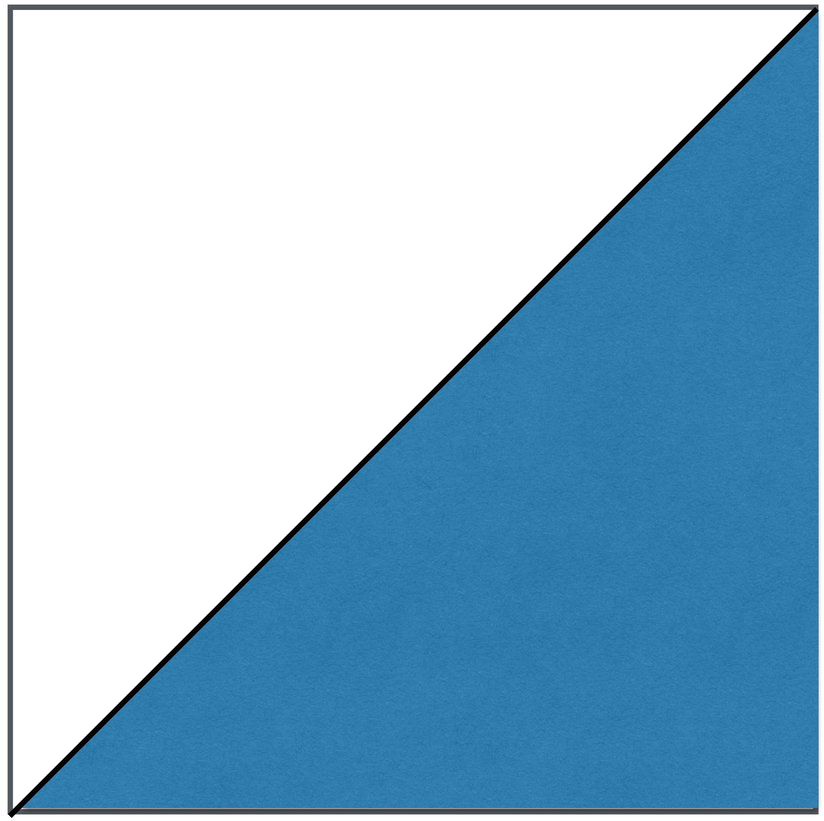


MC - motivation: expectations



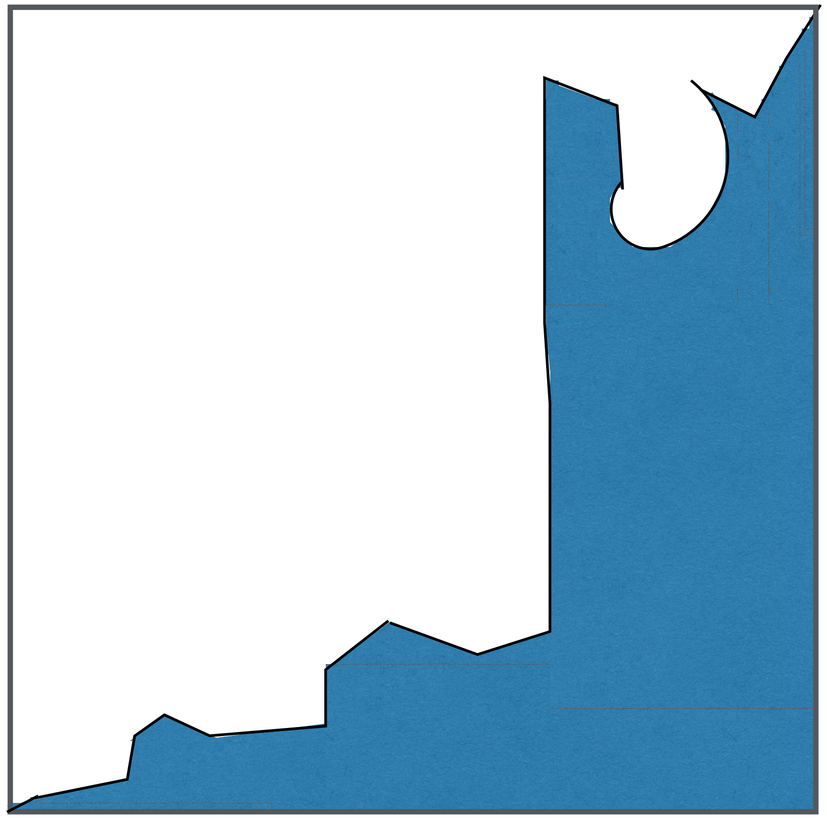
Area: base x height

MC - motivation: expectations



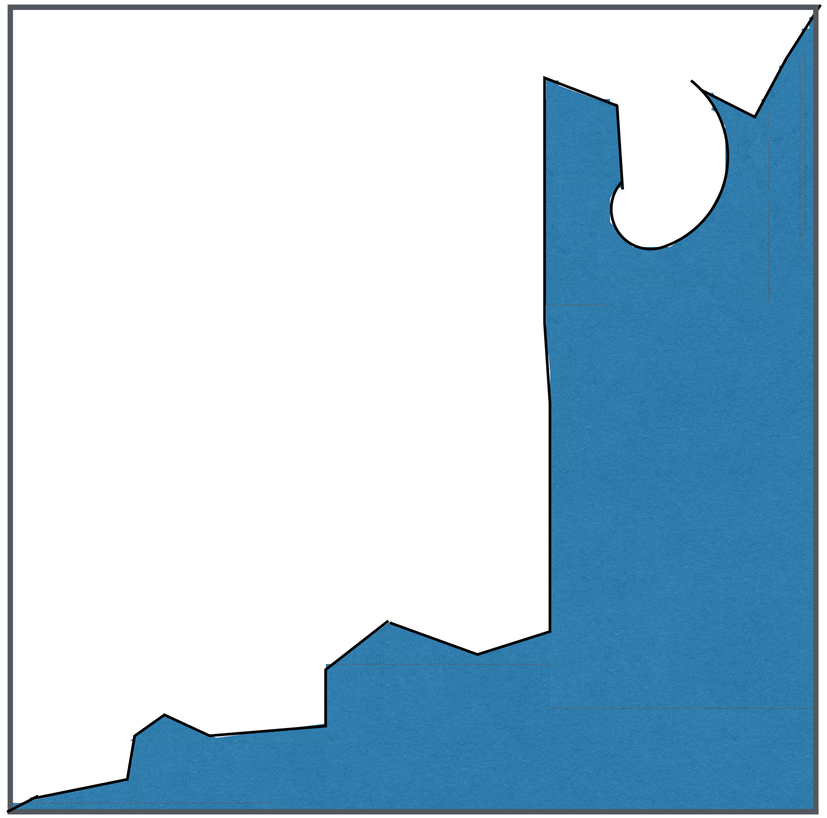
Area: $\frac{\text{base x height}}{2}$

MC - motivation: expectations



Area: ??????

MC - motivation: expectations



Area: ??????

MCArea.ipynb

MC - motivation: expectations

Why am I bothering with areas? - Expectation values are related to areas

Mean $$< \vec{x} >= \frac{1}{N} \sum_{i=1}^{N} N(x_i)$$

$\vec{x}$ = [0, 2, 6, 15, 2]

<x> = 25 / 5 = 5

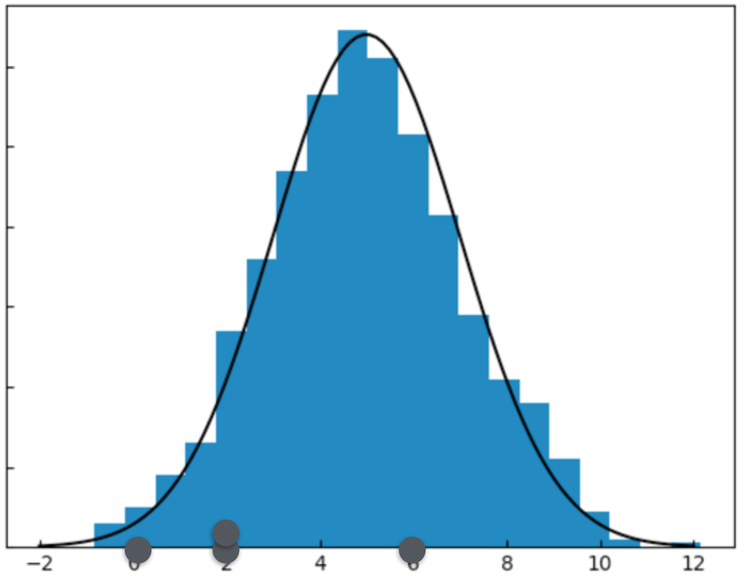
MC - motivation: expectations

Why am I bothering with areas? - Expectation values are related to areas

Mean of a sample

$$< \vec{x} >= \frac{1}{N} \sum_{i=1}^{N} N(x_i)$$

$\vec{x}$ = [0, 2, 6, 15, 2]



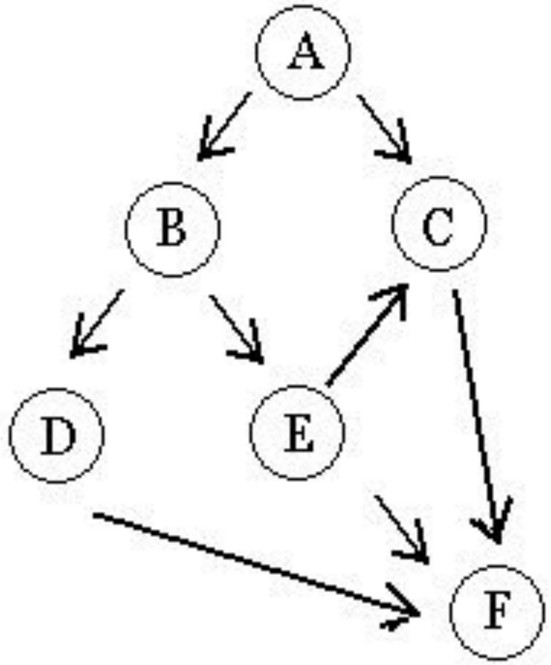
Mean of a continuous distribution

$$mean(X) = E[X] = \int X f(X) dX$$

$$Var(X) = E[X^2] - (E[X])^2.$$

MC - motivation: simulations

Sample



A ~ P(A)

B ~ P(B|A)

C ~ P(C|A,E)

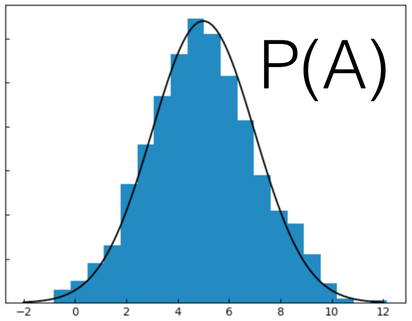
D ~ P(D|B)

E ~ P(E|B)

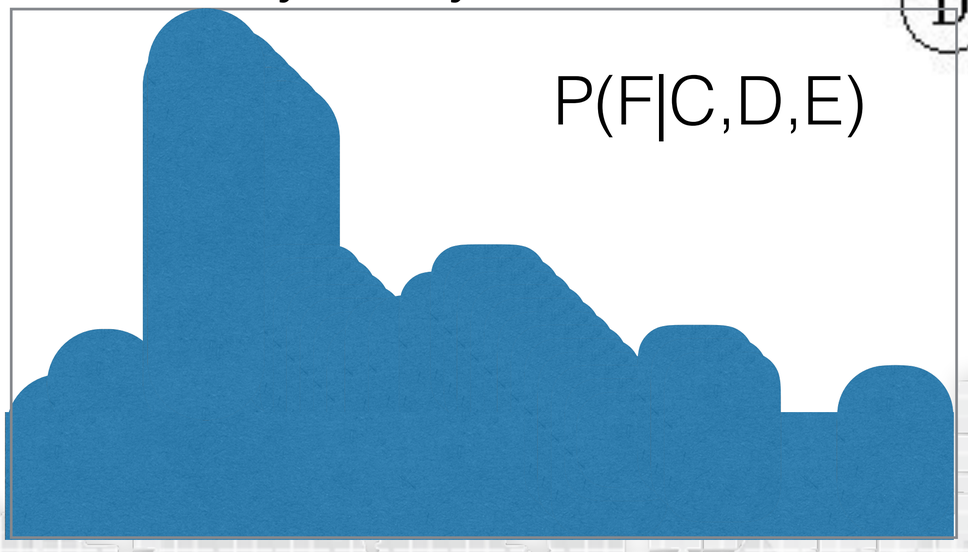
F ~ P(F|C,D,E)

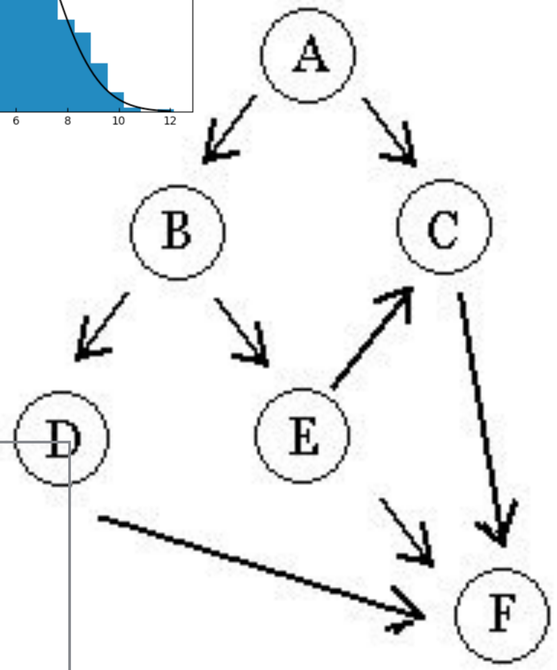
MC - motivation: simulations

The initial probabiliy may be simple



The final probabiliy is likely very complicated (especially if this is a complex system with feedback loops as many urban systems). It may not be tractable analytically but can be simulated





Sample

A ~ P(A)

B ~ P(B|A)

C ~ P(C|A,E)

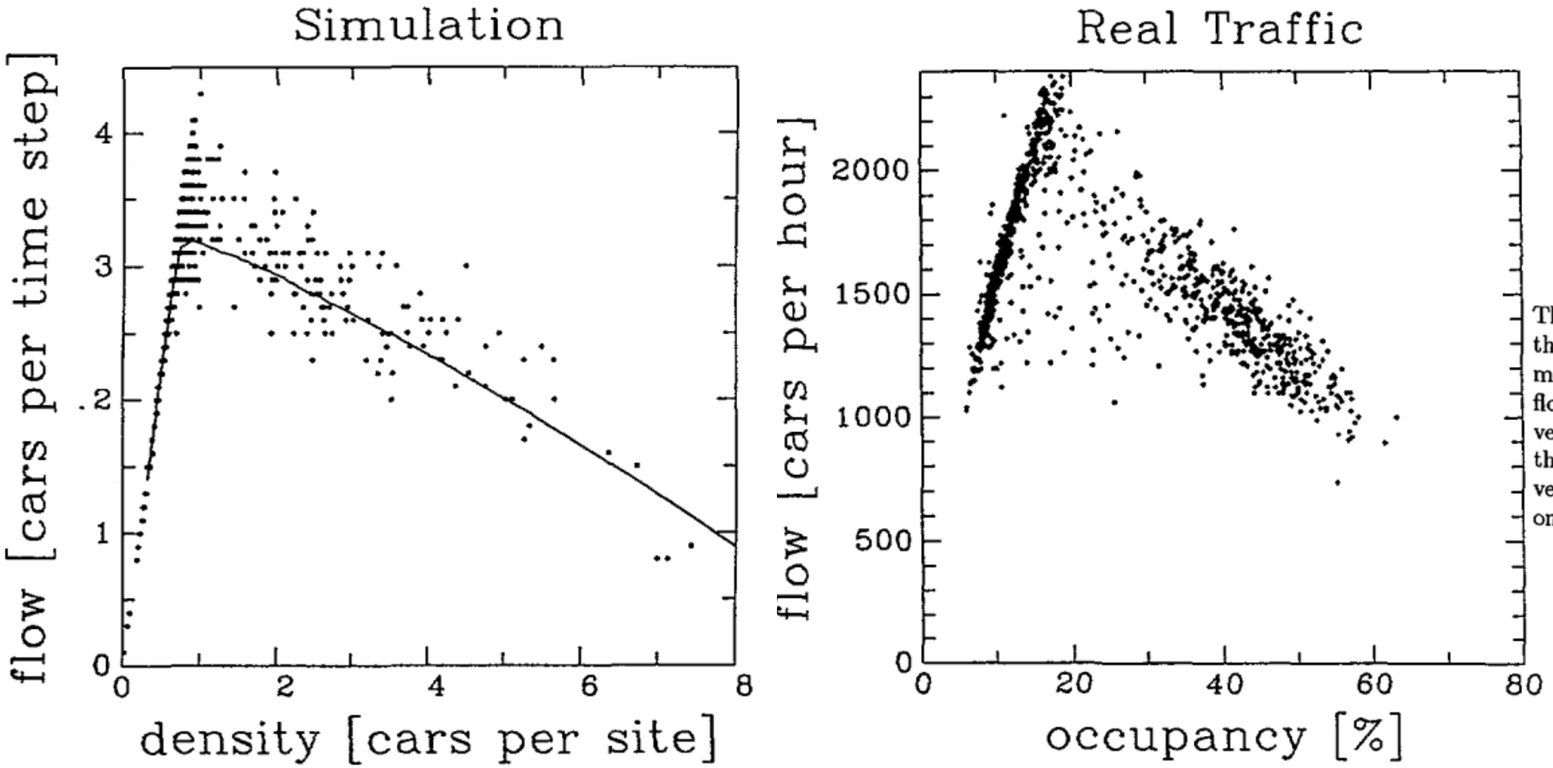
D ~ P(D|B)

E ~ P(E|B)

F ~ P(F|C,D,E)

A long history of MC simulation in traffic flow analysis

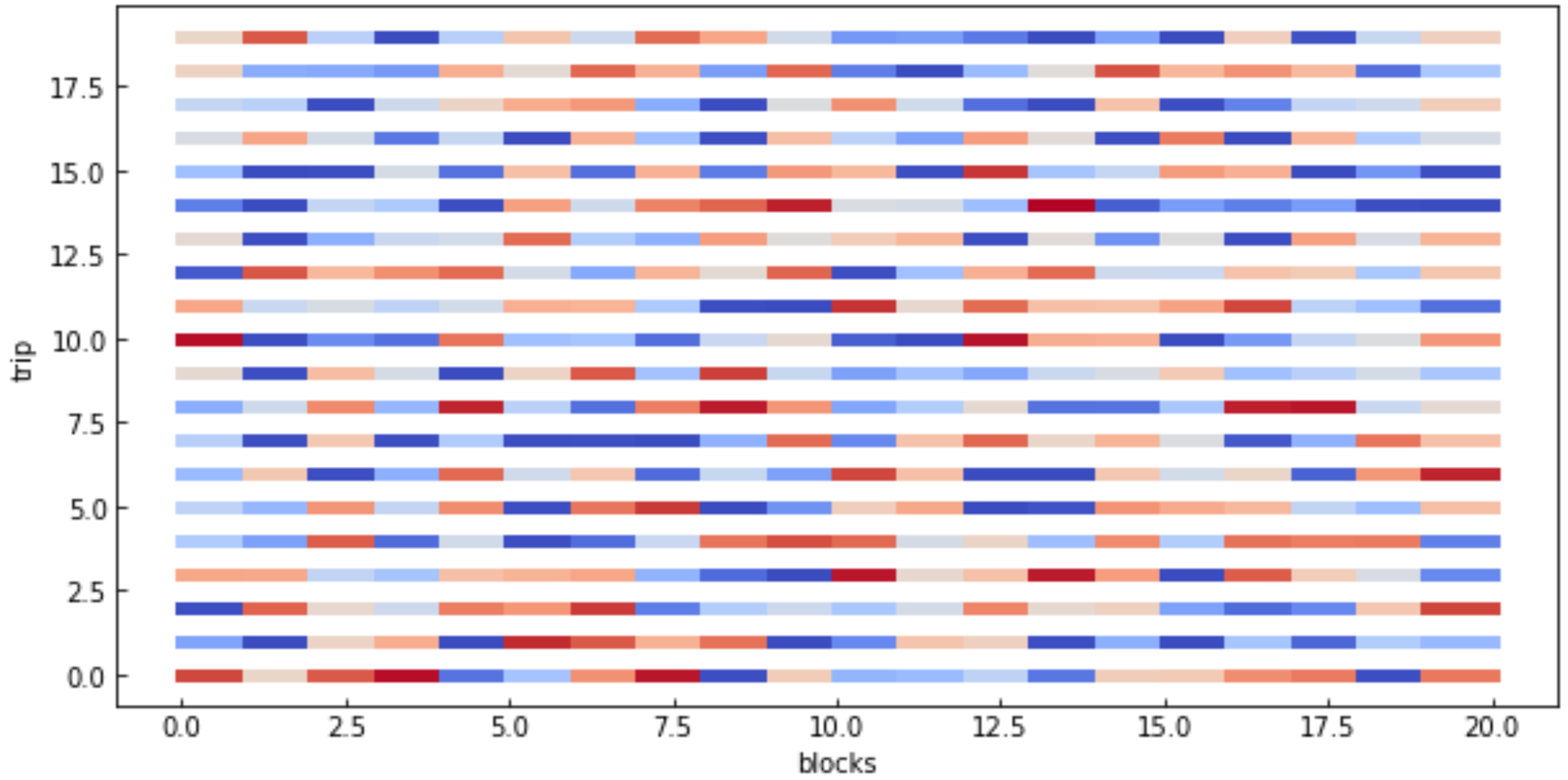
A cellular automaton model for freeway traffic Nagel & Schreckenber 1992



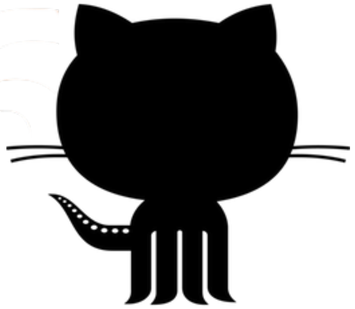
1) **Acceleration:** if the velocity v of a vehicle is lower than v_{max} and if the distance to the next car ahead is larger than $v+1$, the speed is advanced by one $[v \to v+1]$.
2) **Slowing down (due to other cars):** if a vehicle at site i sees the next vehicle at site $i+j$ (with $j \leq v$), it reduces its speed to $j-1$ $[v \to j-1]$.
3) **Randomization:** with probability p, the velocity of each vehicle (if greater than zero) is decreased by one $[v \to v-1]$.
4) **Car motion:** each vehicle is advanced v sites.

Through the steps one to four very general properties of single lane traffic are modelled on the basis of integer valued probabilistic cellular automaton rules [9, 10]. Already this simple model shows nontrivial and realistic behavior. Step 3 is essential in simulating realistic traffic flow since otherwise the dynamics is completely deterministic. It takes into account natural velocity fluctuations due to human behavior or due to varying external conditions. Without this randomness, every initial configuration of vehicles and corresponding velocities reaches very quickly a stationary pattern which is shifted backwards (i.e. opposite the vehicle motion) one site per time step.

MC - Urban Applications



4.71 minutes
9.31 minutes
7.22 minutes
14.64 minutes
9.73 minutes
10.26 minutes
11.16 minutes
7.19 minutes
5.54 minutes
10.01 minutes
12.78 minutes
9.11 minutes
11.65 minutes
8.76 minutes
15.74 minutes
11.09 minutes
13.64 minutes
6.95 minutes
8.79 minutes
5.65 minutes



MCstreetLight.ipynb

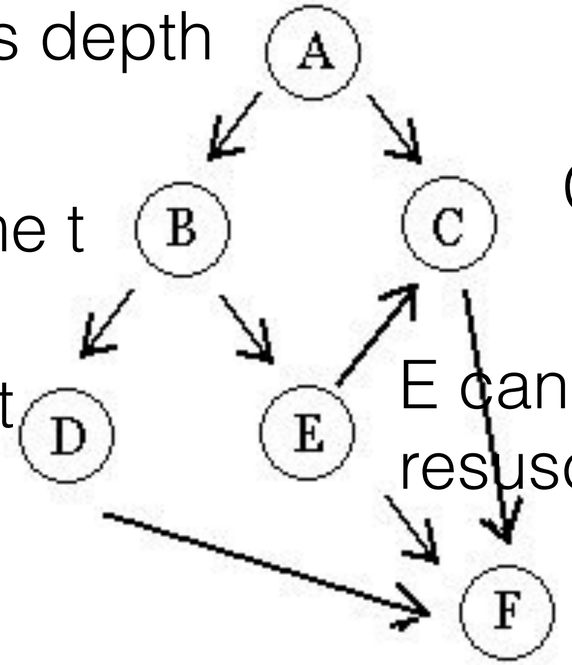
MC - Urban Applications

Sample

A person's depth

B prob to find them at time t

D they are still alive at t



C they survive the avalange

E can be resuscitated at time t

F person survives

A concept for optimizing avalanche rescue strategies using a Monte Carlo simulation approach

Ingrid Reiweger, Manuel Genswein, Peter Paal, Jürg Schweizer

Published: May 3, 2017 • https://doi.org/10.1371/journal.pone.0175877

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175877

MC - motivation: sampling

SetUp 1:

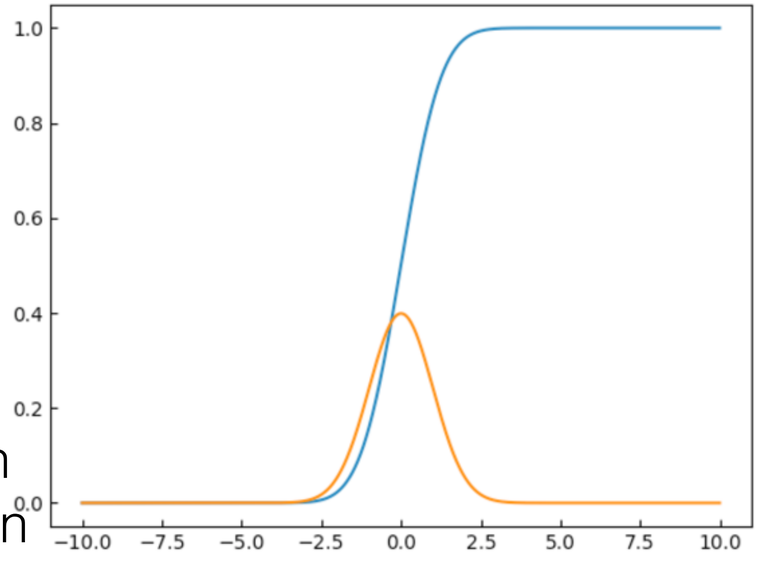
1. I have a distribution described by some formula ***P(x)*** (its PDF)

2. The function can be integrated : *e.g. Gaussian*

$$P(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \int P(x)dx=\frac{1}{2}\left[1+erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

3. If I can *take the integral* of the PDF ***P(x)*** I can calculate the CDF ***F(x)***

4. If I know *and can invert* ***F(x)*** *(i.e. calculate* $\boldsymbol{F^{-1}(u)}$*)* I know at which percentile a value is and I can directly sample from the distribution



MC - motivation: sampling

SetUp 1:

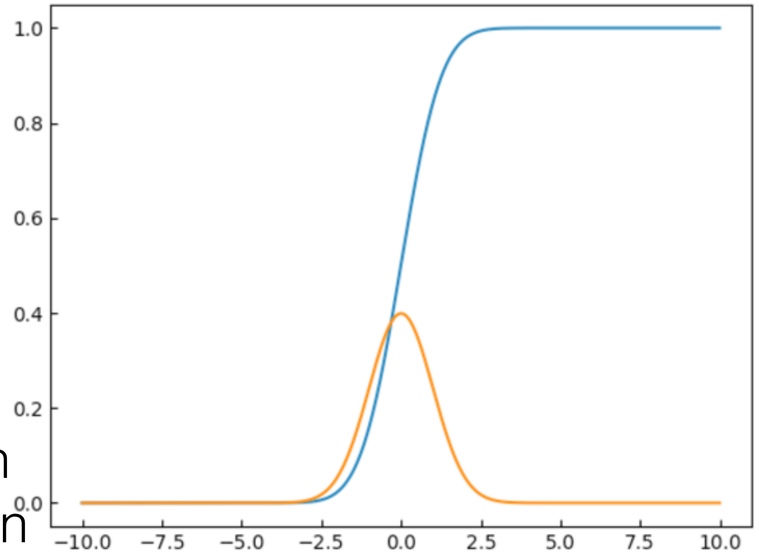
1. I have a distribution described by some formula ***P(x)***

2. The function can be integrated : *e.g. Gaussian*

$$P(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \int P(x)dx=\frac{1}{2}\left[1+erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

3. If I can *take the integral* of the PDF ***P(x)*** I can calculate the CDF ***F(x)***

4. If I know *and can invert* ***F(x)*** *(i.e. calculate* ***$F^{-1}(u)$****)* I know at which percentile a value is and I can directly sample from the distribution



WHILE convergence: // ***P(x)*** is filled in

draw a uniform random number ***u ~ Uniform[0,1]***

calculate ***x*** = ***$F^{-1}(u)$*** *//x is a sample from* ***P***

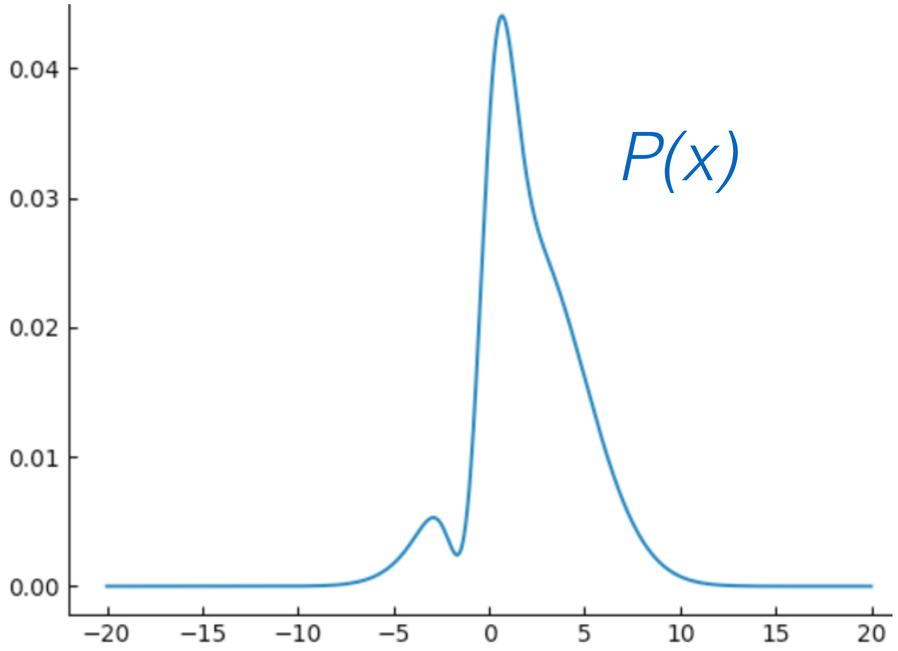
Slides on sampling from distributions
Paul E. Johnson 2015

MC - Rejection Sampling

SetUp 2:

1. I have a distribution described by some formula ***P(x)***

2. The function *cannot* be (easily) integrated :
   I dont know how to draw samples
   but I can calculate its value at every x



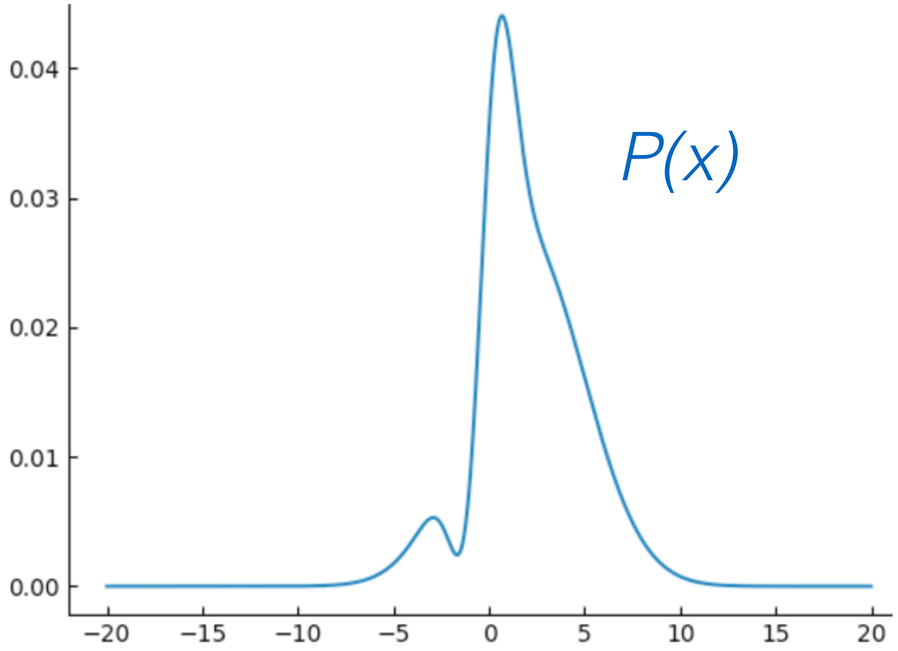
MC - Rejection Sampling

SetUp 2:

1. I have a distribution described by some formula ***P(x)***

2. The function *cannot* be (easily) integrated :
   I dont know how to draw samples
   but I can calculate its value at every x

3. There exist distributions - Q(x) - that are higher than the P(x) at every x



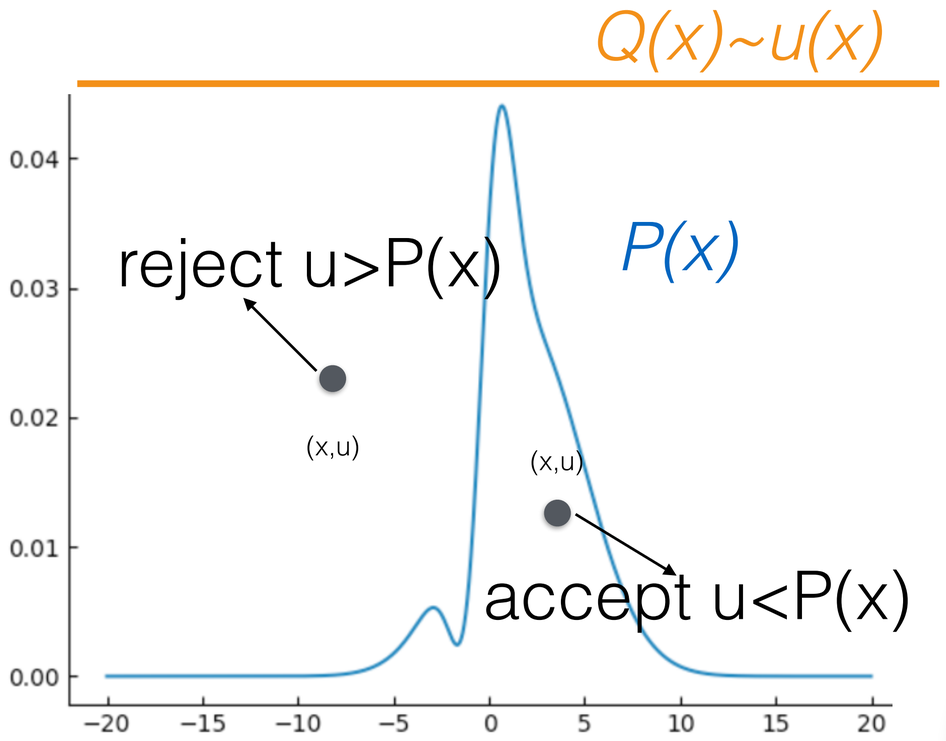
MC - Rejection Sampling

SetUp 2:

1. I have a distribution described by some formula ***P(x)***

2. The function *cannot* be (easily) integrated :
   I dont know how to draw samples
   but I can calculate its value at every x

3. There exist distributions - Q(x) - that are higher than the P(x) at every x : e.g. *Uniform distribution!*



MC - Rejection Sampling

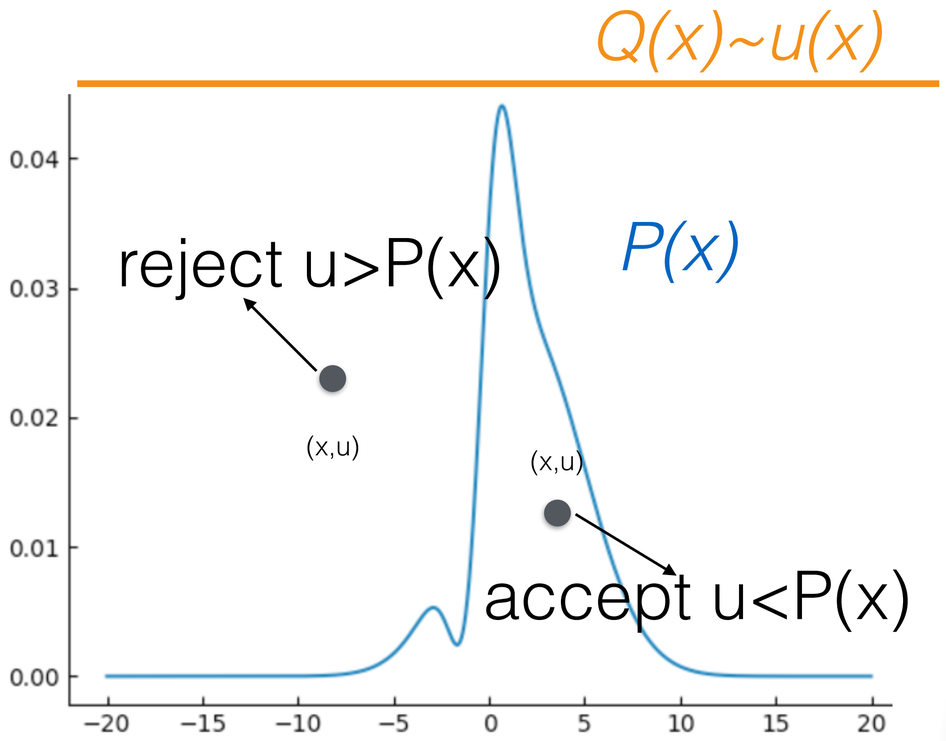
SetUp 2:

1. I have a distribution described by some formula ***P(x)***

2. The function *cannot* be (easily) integrated :
   I dont know how to draw samples
   but I can calculate its value at every x

3. There exist distributions - Q(x) - that are higher than the P(x) at every x : e.g. *Uniform distribution!*

WHILE convergence: // ***P(x)*** is filled in
- draw a point ***x*** from ***Q(x)***
- calculate ***P(x)***
- draw a height ***u ~ Uniform[0,Q(x)]***
- IF : ***u <= P(x)***
  - accept // point is sample of ***P(x)***
- ELSE :
  - reject



MC - Rejection Sampling

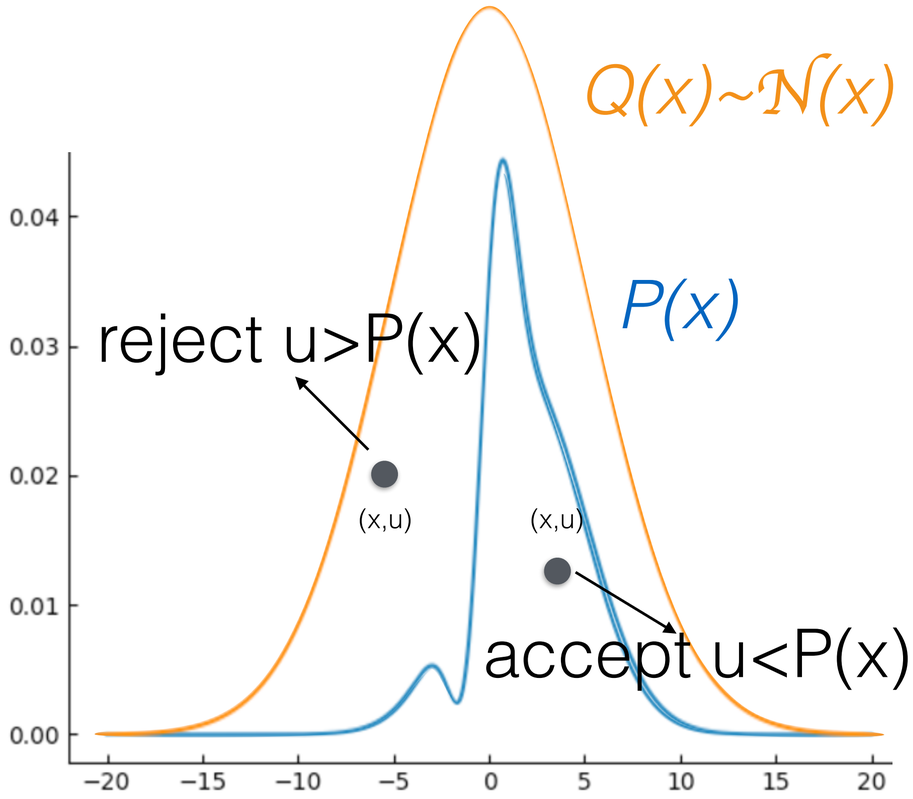
SetUp 2:

1. I have a distribution described by some formula ***P(x)***

2. The function *cannot* be (easily) integrated :
   I dont know how to draw samples
   but I can calculate its value at every x

3. There exist distributions - Q(x) - that are higher than the P(x) at every x : e.g. *Gaussian distribution!*

WHILE convergence: // ***P(x)*** is filled in
- draw a point ***x*** from ***Q(x)***
- calculate ***P(x)***
- draw a height ***u ~ Uniform[0,Q(x)]***
- IF : ***u <= P(x)***
  - accept // point is sample of ***P(x)***
- ELSE :
  - reject



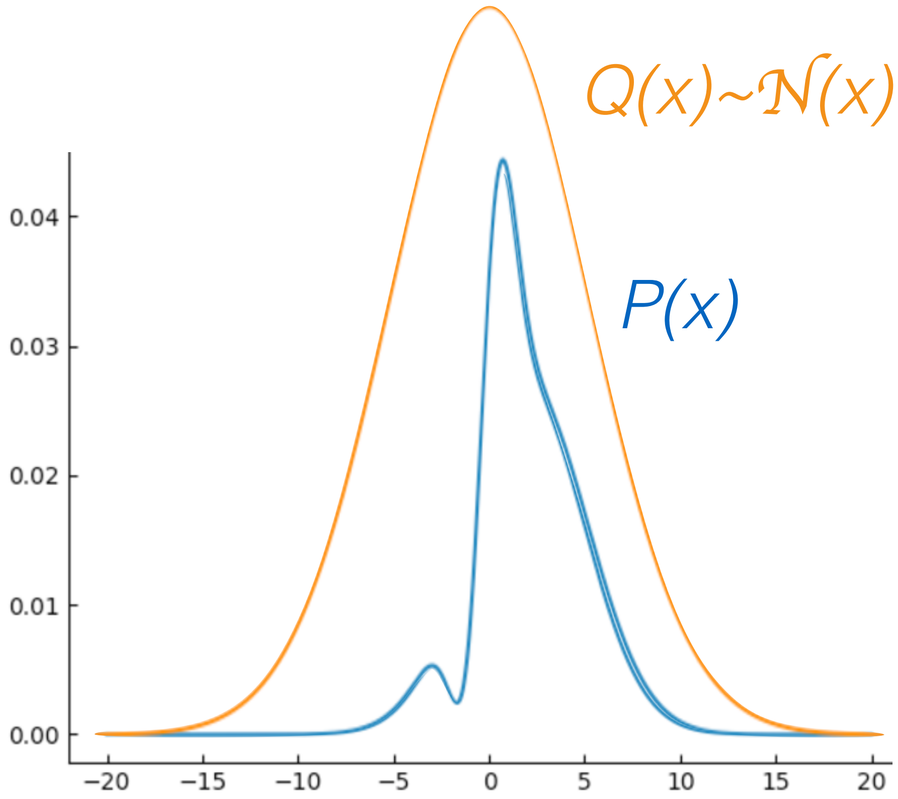
MC - Importance Sampling

SetUp 2:

1. I have a distribution described by some formula ***P(x)***

2. The function *cannot* be (easily) integrated :
   I dont know how to draw samples
   but I can calculate its value at every x

3. There exist distributions - Q(x)



$$\int f(x)P(x)dx = \int f(x)\frac{P(x)}{Q(x)}Q(x)dx, \quad (Q(x)>0 \ \ if P(x)>0)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)\frac{P(x_s)}{Q(x_s)}, \quad x(s)\sim Q(x)$$

Q(x) <=> P(x) guarantees that the integral does not diverge
chose Q(x) s.t. Q(x) is large where f(x) P(x) is large

Markov Chain Monte Carlo

Markov Chain

Markov Chain

memory-less stochastic process:
make predictions for the future of the process based solely on its present state independently from the previous history;
i.e. the next state of the process is based on a chosen distribution (e.g. gaussian) with parameters that depend only on the current state (e.g. with mean at the current state)

Markov processes

Markov Chain

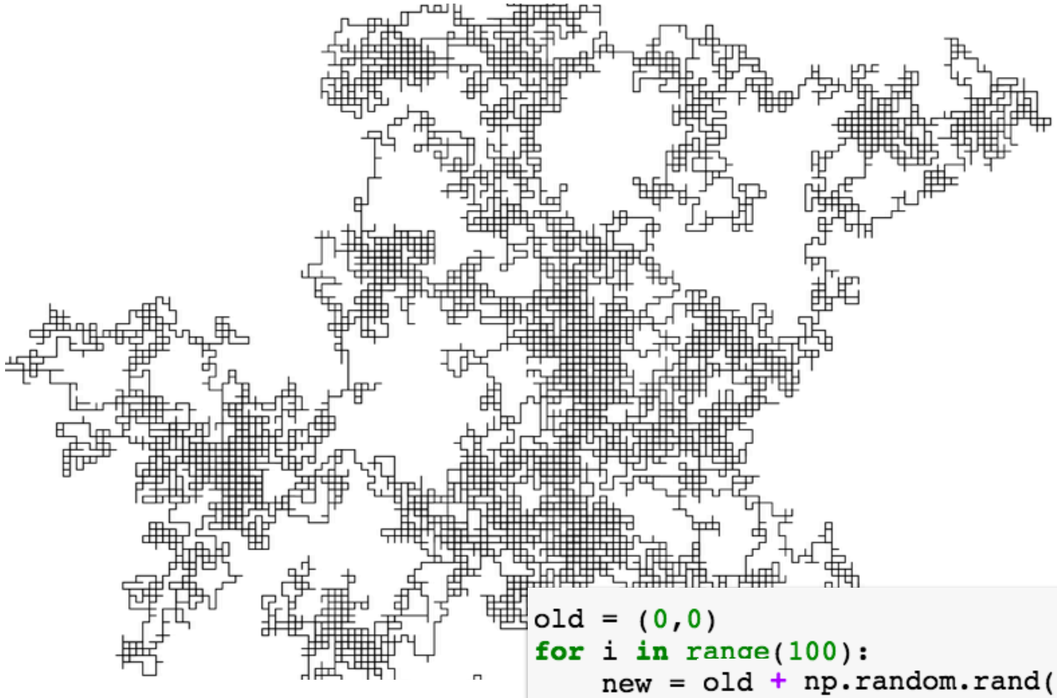
memory-less stochastic process:

e.g.:

Random Walk -> *next position is a stochastic perturbation over current position*

Gamblers' ruin

Markov processes



Markov Chain

memory-less stochastic process:

e.g.:

Random Walk -> choose next position as a gaussian perturbation over the current

Gamblers' ruin

```
old = (0,0)
for i in range(100):
    new = old + np.random.rand(2)
    old = new
```

Markov Chain Monte Carlo

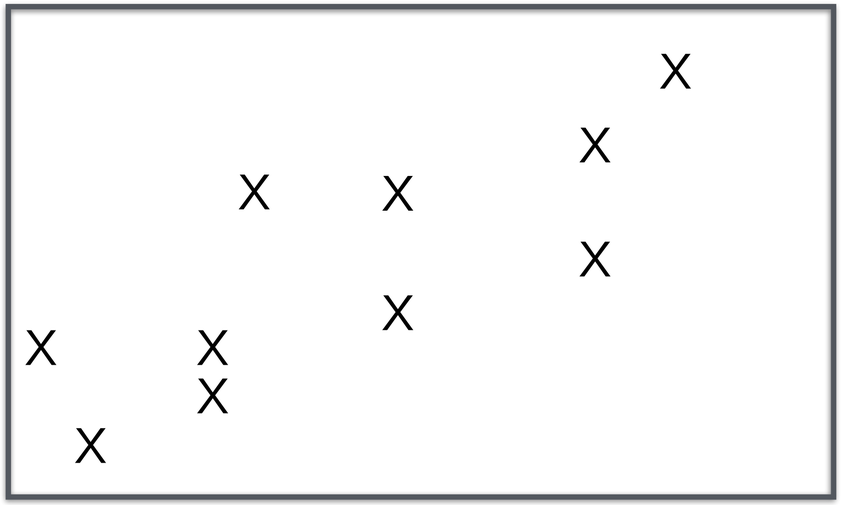
MCMC - motivation

I have a model and I want to find the best parameters to describe my data

Data ***D***

Model - some function $f(\boldsymbol{\theta})$

Paremeters $\boldsymbol{\theta}$



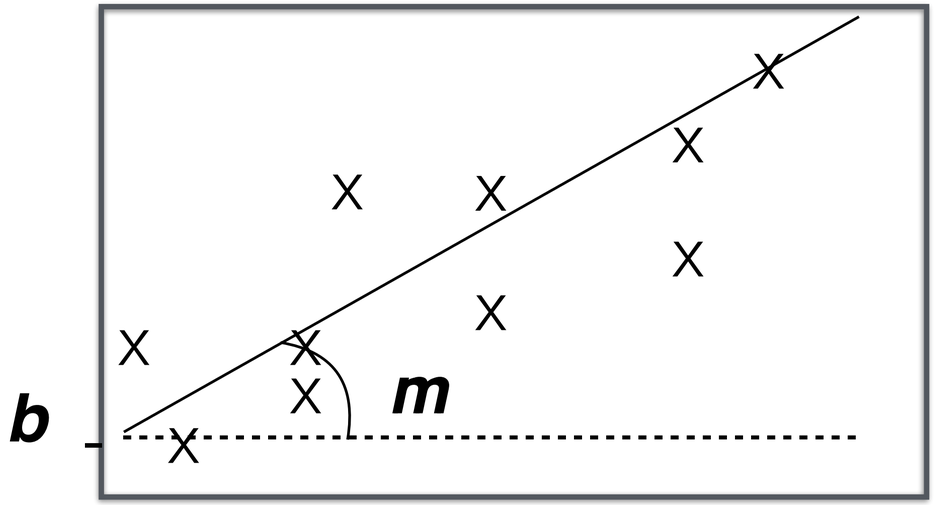
MCMC - motivation

I have a model and I want to find the best parameters to describe my data

Data ***D***

Model $f(\boldsymbol{m}, \boldsymbol{b}) = mx + b$

Paremeters $\boldsymbol{\theta = (m,b)}$



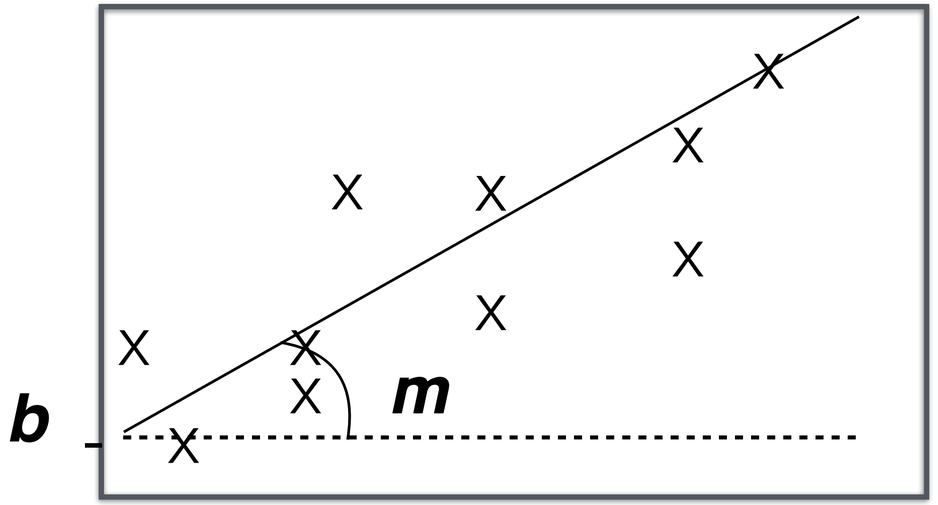
MCMC - motivation

I have a model and I want to find the best parameters to describe my data

Data ***D***

Model $f(\boldsymbol{m}, \boldsymbol{b}) = mx + b$

Paremeters $\boldsymbol{\theta = (m,b)}$



To find the best model parameters:
maximize likelihood: $\boldsymbol{\theta}$ ***such that*** $\boldsymbol{P(D|\theta)}$ ***is max***

https://github.com/fedhere/PUI2016_fb55/blob/master/HW6_fb55/building_nrg_solution.ipynb

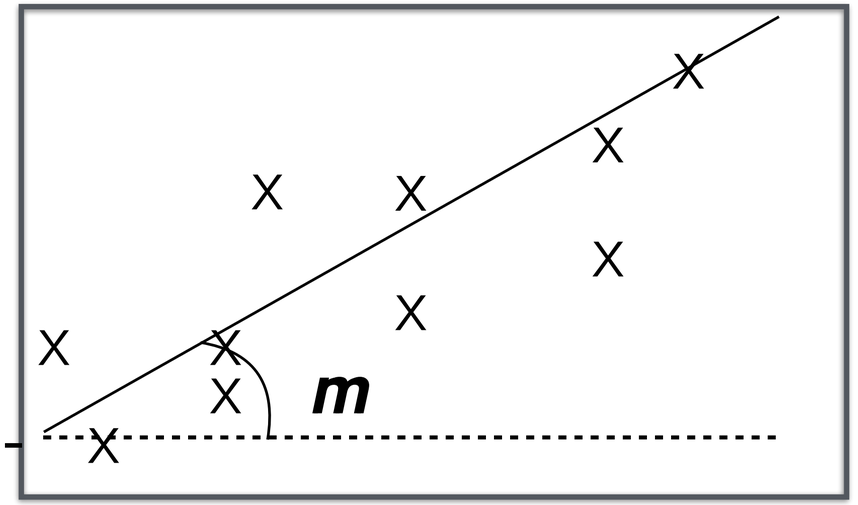
MCMC - motivation

I have a model and I want to find the best parameters to describe my data

Data ***D***

Model $f(\boldsymbol{m}, \boldsymbol{b}) = mx + b$

Paremeters $\boldsymbol{\theta = (m,b)}$



Refresher about *likelihood:*

- *The likelihood of a distribution has the same form as its PDF*
- *the likelohood of a Gaussian distribution is:* $L \equiv P(D|N,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \Pi_{i=1}^{N} exp\left(-\frac{(\mu-x)^2}{2\sigma^2}\right)$
- *generally we like to work in log space with likelihoods because they can be very large numbers and finding the maximum is equivalent to finding a 0 in log space*

$$log(L) = -\frac{1}{2}\sum_{i=1}^{N} log(2\pi) - \sum_{i=1}^{N} log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N} \frac{(\mu-x)^2}{\sigma^2} = C - \frac{1}{2}\chi^2$$

PUI2016 HW 6, lecture 6

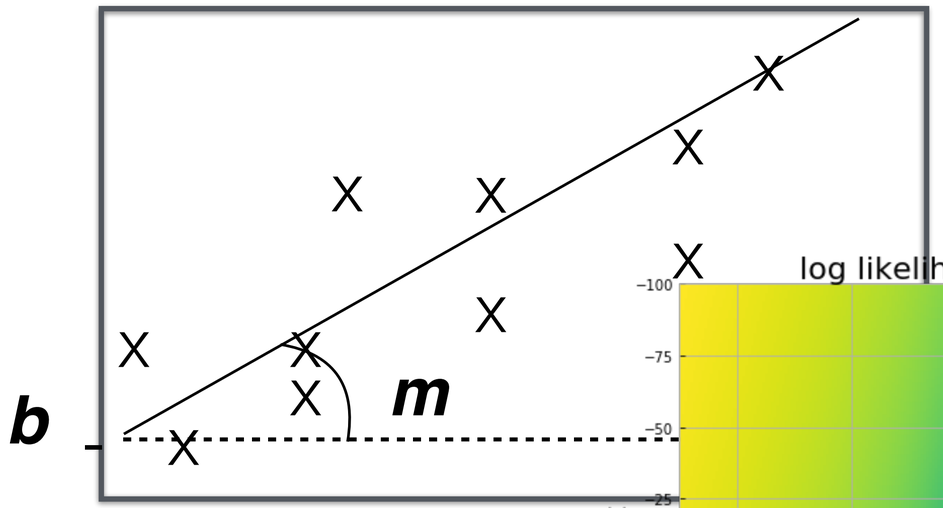
MCMC - motivation

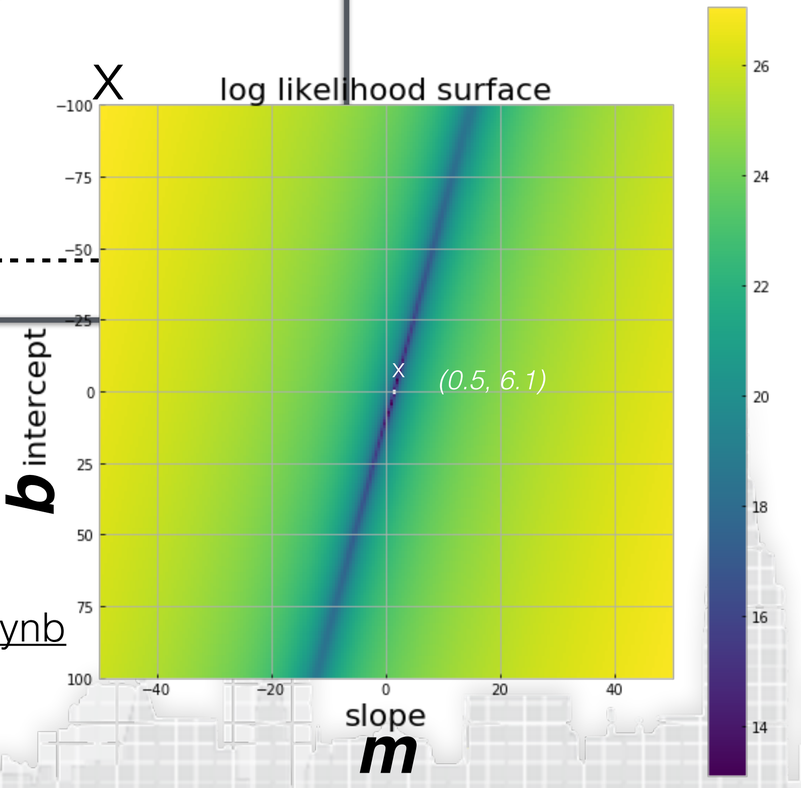
I have a model and I want to find the best parameters to describe my data

Data ***D***

Model $f(\boldsymbol{m}, \boldsymbol{b}) = mx + b$

Paremeters $\boldsymbol{\theta = (m,b)}$





To find the best model parameters:
maximize likelihood: $\boldsymbol{\theta}$ ***such that*** $\boldsymbol{P(D|\theta)}$ ***is max***

https://github.com/fedhere/PUI2016_fb55/blob/master/HW6_fb55/building_nrg_solution.ipynb

Bayes theorem:

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{P(D)}$$

Bayes theorem:

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

Definitions:

posterior: joint probability distributin of a parameter set (*m*, *b*) condition upon some data *D* and a model hypothesys *f*

Bayes theorem:

$$P(\theta|D,f) = \frac{P(D|\theta,f)\,P(\theta,f)}{P(D|f)}$$

Definitions: **posterior**

posterior: joint probability distributin of a parameter set (*m*, *b*) condition upon some data *D* and a model hypothesys *f*

Bayes theorem:

Definitions:

posterior: joint probability distributin of a parameter set (*m*, *b*) condition upon some data *D* and a model hypothesys *f*

Bayes theorem:

$$P(\theta|D,f) = \frac{P(D|\theta,f)\,P(\theta,f)}{P(D|f)}$$

$P(\theta|D,f)$: **posterior**

$P(\theta,f)$: **prior**

Definitions:

posterior: joint probability distributin of a parameter set (*m, b*) condition upon some data *D* and a model hypothesys *f*

prior: “intellectual” knowledge about the model parameters

MCMC - motivation

Bayes theorem:

$$P(\theta|D,f) = \frac{P(D|\theta,f)\,P(\theta,f)}{P(D|f)}$$

$P(\theta|D,f)$: **posterior**

$P(\theta,f)$: **prior**

Definitions:

posterior: joint probability distributin of a parameter set (*m, b*) condition upon some data *D* and a model hypothesys *f*

prior: "intellectual" knowledge about the model parameters

e.g.: energy consumption increaded w number of units: m>0

Bayes theorem:

$$P(\theta|D,f)=\frac{P(D|\theta,f)P(\theta,f)}{P(D|f)}$$

$P(\theta|D,f)$: **posterior**

$P(\theta,f)$: **prior**

$P(D|f)$: **evidence**

Definitions:

posterior: joint probability distributin of a parameter set (*m*, *b*) condition upon some data *D* and a model hypothesys *f*

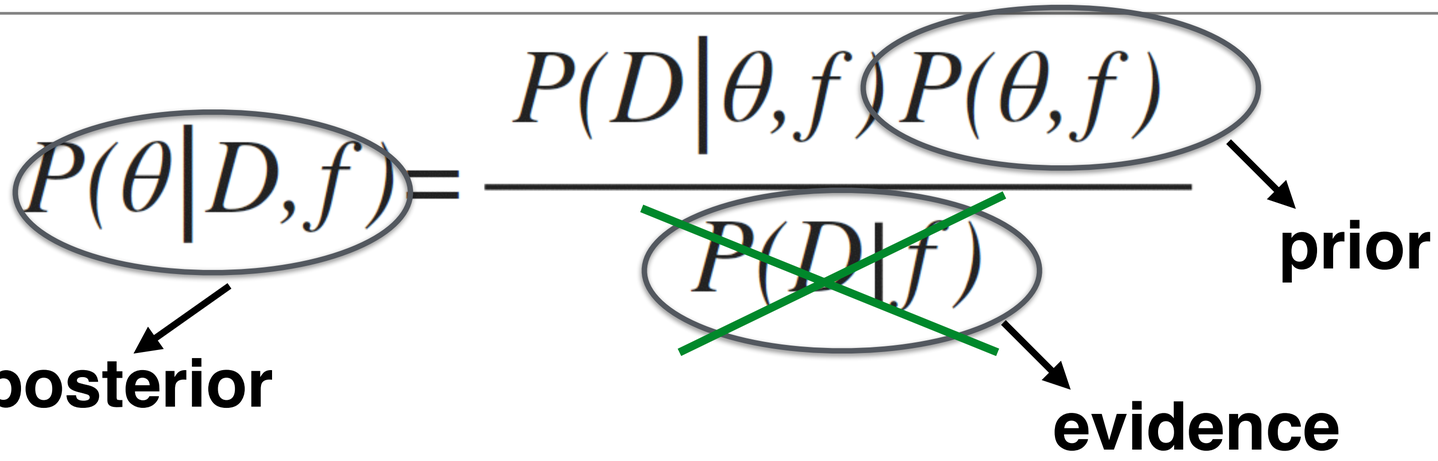
prior: "intellectual" knowledge about the model parameters

evidence: marginal likelihood of data under the model

$$P(D|f)=\int P(D|\theta,f)P(\theta|f)d\theta$$

MCMC - motivation

Bayes theorem:



Definitions:

posterior: joint probability distributin of a parameter set (*m, b*) condition upon some data *D* and a model hypothesys *f*

prior: "intellectual" knowledge about the model parameters

evidence: marginal likelihood of data under the model

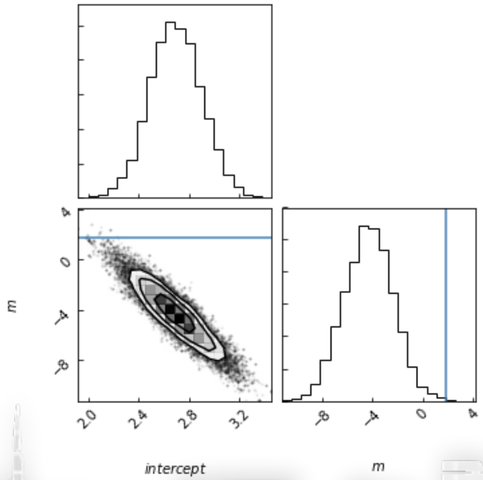
its constant in θ so we can ignore it $P(D|f)=\int P(D|\theta,f)P(\theta|f)d\theta$

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Definitions: **posterior**

posterior: joint probability distributin of a parameter set (*m, b*) condition upon some data *D* and a model hypothesys *f*

triangle plot



MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

A nice tutorial on MCMC by Thomas Wiecki (Quantopian)

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

*/*proposal*/*
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

Questions:

1. how do I choose the next point?

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

Questions:

1. how do I choose the next point?

   Any *Markovian* process

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

Questions:

1. how do I choose the next point?

   Any *Markovian* process

   Any *ergodic* process
   (with enough time all locations will be explored)

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

*/*proposal*/*
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

Questions:

1. how do I choose the next point?

Any *Markovian* process

Any *ergodic* process
CN: detailed balance

detailed balance $\pi(x_1)p(x_2|x_1)=\pi(x_2)p(x_1|x_2)$

Metropolis Rosenbluth Rosenbluth Teller 1953 - Hastings 1970

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

choose a starting point **current** = $\theta_0 = (m,b)$

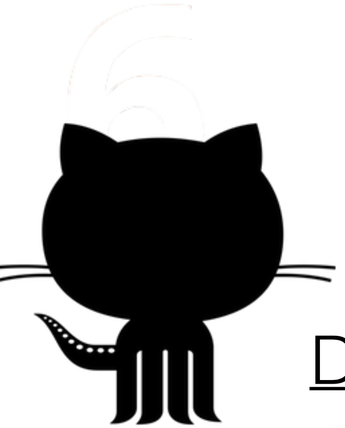
WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing



DYI_MCMC.ipynb

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r < *post_new / post_orig:***
current = new
ELSE:
pass //do nothing

Questions:

1. how do I choose the next point?

Any *Markovian* process

Any *ergodic* process
CN: detailed balance

detailed balance $\pi(x_1)p(x_2|x_1)=\pi(x_2)p(x_1|x_2)$

Metropolis Rosenbluth Rosenbluth Teller 1953 - Hastings 1970

MCMC - Gibbs sampler

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/

chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***

current = new

ELSE***:***

/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */

r = random uniform number [0,1]

IF **r < *post_new / post_orig:***

current = new

ELSE:

pass //do nothing

Questions:

1. how do I choose the next point?

   Gibbs sampling:
   Metropolis-Hastings proposal distribution with change *along a single direction at a time => always accept*
   must know the integral P(D|f) along that direction

detailed balance $\pi(x_1)p(x_2|x_1)=\pi(x_2)p(x_1|x_2)$

Metropolis Rosenbluth Rosenbluth Teller 1953 - Hastings 1970

MCMC - Gibbs sampler

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

Questions:

1. how do I choose the next point?

Other options:
simulated annealing (good for multimodal)
parallel tempering (good for multimodal)
differential evoution (good for covariant spaces)

detailed balance $\pi(x_1)p(x_2|x_1)=\pi(x_2)p(x_1|x_2)$

Metropolis Rosenbluth Rosenbluth Teller 1953 - Hastings 1970

MCMC - Gibbs sampler

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

Questions:

1. how do I choose the next point?

Other options:
affine invariant : EMCEE package

detailed balance $\pi(x_1)p(x_2|x_1)=\pi(x_2)p(x_1|x_2)$

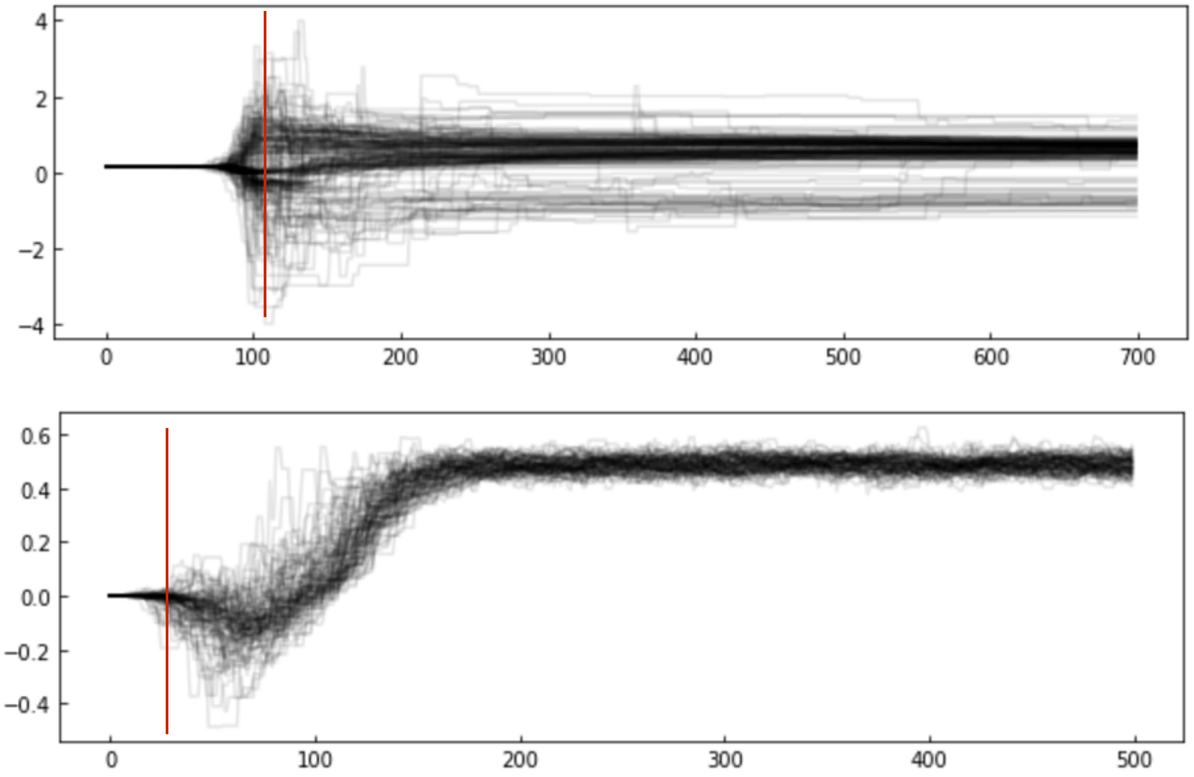
Metropolis Rosenbluth Rosenbluth Teller 1953 - Hastings 1970

MCMC - EMCEE

MCMC - convergence

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution



Questions:

1. how do I choose the next point?
2. when have I sampled the posterior adequatly?
   has your chain *burned-in* ?

MCMC - convergence

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

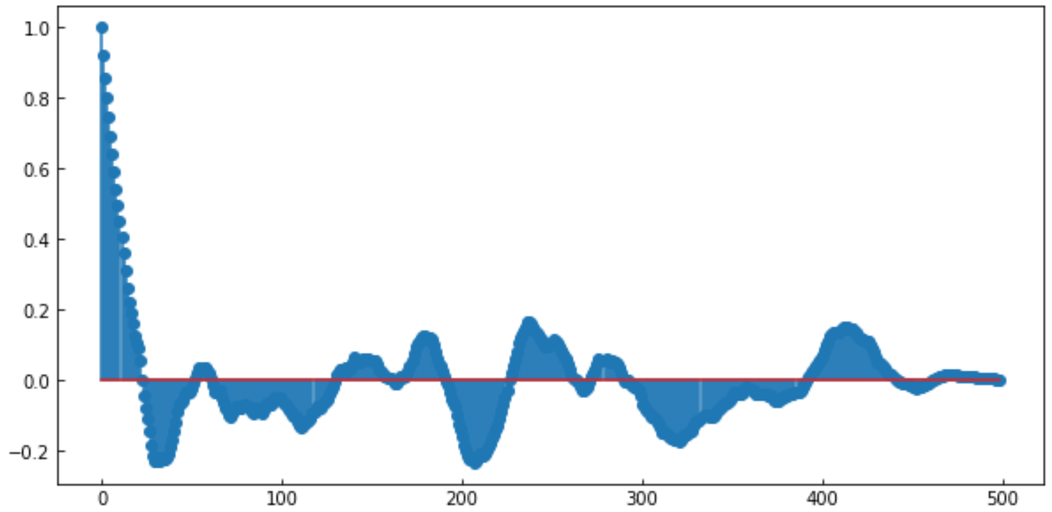
Questions:

1. how do I choose the next point?
2. when have I sampled the posterior adequatly?
   has your MCMC *converged* ?

MCMC - convergence

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution



Questions:

1. how do I choose the next point?

2. when have I sampled the posterior adequatly?
   has your MCMC *converged* ?
   a. **check autocorrelation within a chain (*Raftery*)**
   b. check that all chains coverged to same region (a stationary distribution *GelmanRubin*)
   c. mean at beginning = mean at end (*Geweke*)
   d. check that entire chain reached stationary distribution (or a final fraction of the chain, *Heidelberg-Welch* using Cramer-von-Mises statistic)

MCMC - convergence

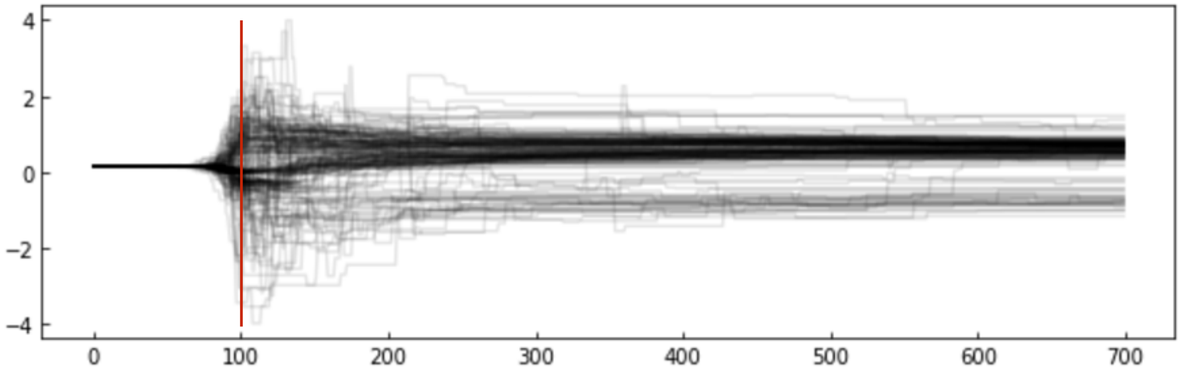
Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

Questions:

1. how do I choose the next point?

2. when have I sampled the posterior adequatly?
   has your MCMC *converged* ?

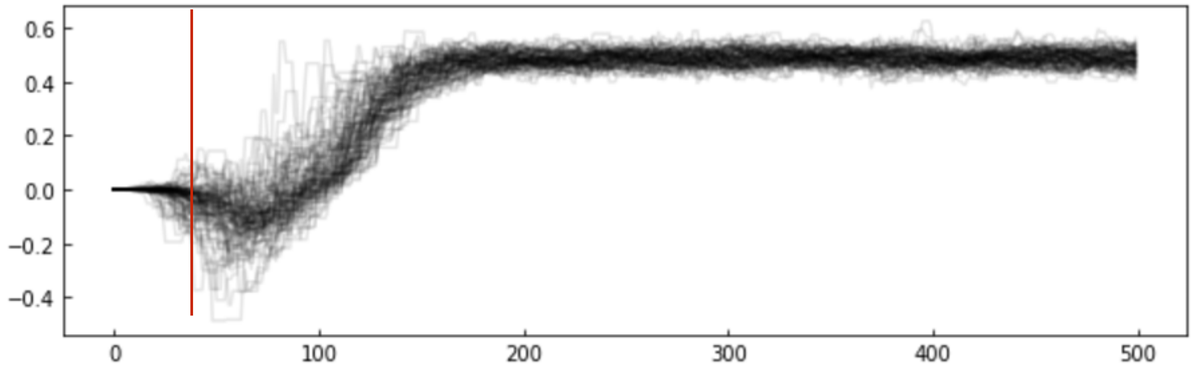


   a. check autocorrelation within a chain (*Raftery*)
   b. check that all chains coverged to same region (a stationary distribution *GelmanRubin*)
   c. mean at beginning = mean at end (*Geweke*)
   d. check that entire chain reached stationary distribution (or a final fraction of the chain, *Heidelberg-Welch* using Cramer-von-Mises statistic)

MCMC - convergence

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution



Questions:

1. how do I choose the next point?

2. when have I sampled the posterior adequatly?
   has your MCMC *converged* ?
   a. check autocorrelation within a chain (*Raftery*)
   b. check that all chains coverged to same region (a stationary distribution *GelmanRubin*)
   c. mean at beginning = mean at end (*Geweke*)
   d. check that entire chain reached stationary distribution (or a final fraction of the chain, *Heidelberg-Welch* using Cramer-von-Mises statistic)

MCMC - Metropolis Hastings algorithm

Bayes theorem: $P(\theta|D,f) \propto P(D|\theta,f)P(\theta,f)$

Goal: sample the posterior distribution

Questions:

1. how do I choose the next point?
2. when have I sampled the posterior adequatly?
3. how can it be-the samples are *not independent!*

good point!…

choose a starting point **current** = $\theta_0 = (m,b)$

WHILE convergence criterion is met:

calculate current posterior **post_curr** = $P(D|\theta,f)$

/*proposal*/
chose a new set of parameters **new** = $\theta_{new} = (m,b)$

calculate the new posterior **post_new**=$P(D|\theta_{new},f)$

IF ***post_new > post_curr:***
current = new
ELSE***:***
/*accept with probability $P(D|\theta_{new},f) / P(D|\theta,f)$ */
r = random uniform number [0,1]
IF **r** ***< post_new / post_orig:***
current = new
ELSE:
pass //do nothing

Resources Markov Chain Monte Carlo

Information Theory, Inference, and Learning Algorithms
David J.C. MacKay, 2003

Numerical Recipes
Bill Press+ 1992 (+)

Ensemble samplers with affine invariance
Jonathan Goodman and Jonathan Weare 2010

Resources Markov Chain Monte Carlo

Slides on sampling from distributions
Paul E. Johnson 2015

EMCEE readme
provides high level discussion, references, suggestion on parameter choices
D. Foreman-Mackey, D. Hogg, D. Lang, J. Goodman+ 2012

Bill Press (Numerical Recipes) Video
proving how Metropolis-Hastings sutisfied Detail Balance

Quick Glossary

- **Stochastic**: random, following any distribution
- **PDF**: probability distribution function *P(x)* describes the *relative* likelihood of sample *x* compared
- **CDF**: cumulative distribution function - the probability that a value drawn from a distribution will be smaller than *x*

$$F(x)=\int_{-}^{x} P(x)$$

- **Marginalize**: integrate along a dimension
- **Gaussian distribution**: a distribution with PDF

$$N(\mu,\sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

- **Chi Squared χ^2:** a model fitting method based on the provable fact that (under proper assumption) the function follows a χ^2 distribution

$$\sum_{i=1}^{N}\frac{(M-D)^2}{\sigma^2} \sim \chi^2_{DOF}$$

- **Likelihood**: in Bayes theorem its the term indicating the probability of the data under the model for a choice of parameters. More generally it can be thought of the probability of the parameters given the data
- **Posterior**: the probability of data given model calculated by Beyes theorem as likelihood * prior / evidence
- **Evidence**: the probability of the data given a model marginalized over all parameters
- **Prior**: prior, or otherwise obtained, knowledge about the problem which indicates how likeli the model parameter are for any value
- **Markovian process**: a process whose next stage depends stochastically on the current state only
- **Ergodic**: a process that given enough time would visit all location of the space
- **Markov Chain**: an N dimensional sequence of values of each parameter of the N-dim parameter space that is explored by an MCMC

Summary: Stochastic methods

Stochastic Processes in Science Inference: starting with the advent of computers simulations became a valuable alternative to analytical derivation to solve complex scientific problems and the only way to solve untractable problems. Events that occurr with a known probability can be simulated, the possible outcome would be simulated with a frequency corresponding to the probability.
Applications: Instances of the evolution of a complex systems can be simulated and from this synthetic (simulated) sample solutions can be generalized as they would from a sample extracted from a population:
e.g.. *simulate traffic flow to determine the average trip duration instead of measuring many trips to estimate the trip duration*,
or a better scheme would be: *simulate traffic flow and validate your simulation by comparing the average trip duration for a synthetic sample and from a sample from the real system, then simulate proposed changes to traffic to validate planning options before implementing them.*
Simulations require drawing samples from distributions.
Drawing samples from a distribution: can be done directly if the PDF *P(X)* can be integrated *analytically* to find a CDF *F(x)* and if this CDF is invertible ($F^{-1}(u)$ *can be calculated analytically)* . The algorithm is:

1. draw a *uniformly distributed* number **u** between 0-1
2. invert the CDF of your distribution $\boldsymbol{x=F^{-1}(u)}$ *is a sample from the desired PDF (i.e. x are drawn at a frequency P(x))*

If *F(x)* of $F^{-1}(u)$ cannot be calculated analytically Rejection Sampling allows to sample from the desired *P(x)* . The algorithm is:

1. find a function *Q(x)* that is larger than *P(x)* for every x and that has an analytical, integrable, invertible form
2. draw samples x from *Q(x)* (see above)
3. draw a *uniformly distributed* number **u** between 0-Q(x)
4. only accept x where u < *P(x)*

If your proposal distribution is poorly chosen (much higher than P(x) in some regions) this can be an extremely wasteful process. The higher the problem dimensionality the more this issue is concerning. Alternatives include Importance

Summary: MCMC, background concepts

Markovian processes: A process is Markovian if the next state of the system is determined stochastically as a perturbation of the current state of the system, and only the current state of the system, i.e. the system has no memory of earlier states (a *memory-less* process).
A state being a stochastic perturbation of the previous state means that given the conditions of the state at time *t* (e.g. *A(t)* = (position +velocity)) the *next* set of conditions *A(t+1)* (updated position+velocity) will be drawn from a distribution related to the earlier state. For example the *next* velocity can be a sample from a Gaussian distribution with mean equal to the *current* velocity. $A(t+1) \sim \mathcal{N}(A(t), s)$

Bayes theorem: relates observed data to proposed models by allowing to calculate the *posterior distribution of model parameters* for a given prior and observed dataset (see glossary for term definition).

Posterior(data, model-parameters) = Likelihood(data, model-parameters) * Prior(model-parameters) / Evidence(data)

$$P(\theta|D,f) = \frac{P(D|\theta,f)\,P(\theta,f)}{P(D|f)}$$

Summary: MCMC

Markov Chain Monte Carlo: Is a method to sample a parameter space that is based on Bayes theorem. The MCMC samples the ***joint posterior*** of the parameters in the model (up to a constant, the *evidence,* probability of observing your data under any model parameter choice, which is generally not calculable). Thus we can get posterior median, confidence intervals, covariance, etc… The algorithm is:

1. starting at some location in the parameter space propose a new location as a Markovian perturbation of the current location
2. if the proposal posterior is better than the posterior at the current location update your position (and save the new position in the chain)
3. if the proposal posterior is worse than the posterior at the current location update your position with some probability ***a***

The choice of the proposal distribution and rule ***a*** for accepting the new step in the chain have to satisfy the ***ergodic*** condition, that is: given enough time the entire parameter space would be sampled. (***Detailed Balance*** is a sufficient condition for ergodicity)

If the chain is Markovian and the proposal distribution is *ergodic* *the entire parameter space is sample, given enough time, with sampling frequency proportional to the posterior distribution*

Different MCMC algorithms: while all MCMC algorithm share the structure above the choice of proposal and the acceptance probability are different for different MCMC algorithms.

Metropolis Hastings MCMC first, most common MCMC with acceptance proportional to the ratio of posteriors ***a~posteriorNew/posteriorCurrent.*** This becomes problematic when the posterior has multiple peaks (may not explore them all) or parameter are highly covariant (may take a very long time to converge)

Convergence: It is crucial to confirm that your chains have converged and your parameter space is properly sampled, but it is also very difficult to do it. Methods include checking for stationarity of the chain means and low auto correlation in the chains. The beginning of the chain is typically removed as the chains require a minimum number of steps to move away from the initial position effectively.