And apologies. Slides are dense to serve as future references…

My Goal

Teach you the concepts, the language and provide the references to start googling and implementing your own analysis code

Resources

videolectures.net (Amazing lectures presented at different levels,
from simple to advanced presented…on a poorly designed website).

Simple+ intermediate:

1) Sivia and Skilling, Data Analysis: A Bayesian Introduction, Second edition, 2011.

2) Bishop, Pattern Recognition and Machine Learning, 2006.

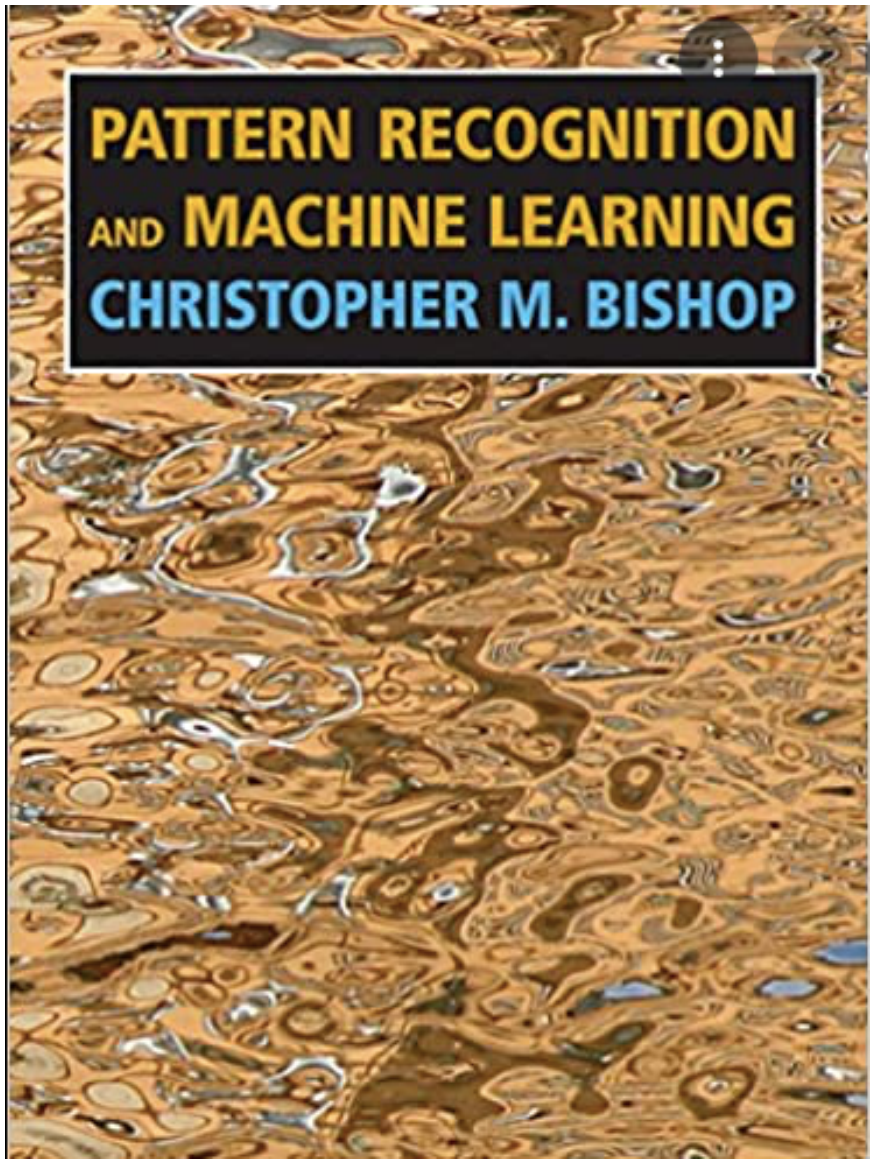
Harder:

3) Gelman et al., Bayesian Data Analysis, Third Edition, 2014.

Tailored for Natural Sciences:

I am writing a book on stochastic processes and data analysis.
One year to go! Until then, see above.

Resources


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

As Physical/Biological Scientists & Engineers here is what we…

Learn

Chemistry

Physics

Biology

Biochemistry

Improvise

Data Analysis

Statistics

⋮

Punchline:

There is only one right way to analyze your data

It is normally not possible to analyze your data in this way

Wrong ways should only be used for computational tractability (provided your conclusions are not qualitatively affected by your short-cuts).

It is only possible to see what approximate way is needed once the correct way is written down.

Why should this sound so controversial?

Why should this sound so controversial?

There is only one right way to write down an electrodynamics problem that satisfies Maxwell's equations (and the built-in Lorentz invariance) and boundary conditions

Wrong ways should only be used for computational tractability (provided your conclusions are not qualitatively affected by your short-cuts).

It is only possible to see what approximate way is needed once the correct way is written down.

Outline

Setting up the problem

System models and observation models

Latent variables and graphical models

Likelihoods and EM algorithm

Bayesian methods, priors

Monte Carlo, Metropolis-Hastings

Setting up the Problem

Imagine a coin flip experiment w HTHHHTH
and we want to determine the probability of heads and tails

The N outcomes of this experiment are random variables.

$$\mathbf{y}_{1:N} = \{y_1, y_2, \cdots y_N\} = \{H, T, H, H, H, \cdots\}$$

In other words y_1 **is heads with probability** p
is tails with probability $1 - p$

Setting up the Problem

To determine the probability of heads/tails, we ask:

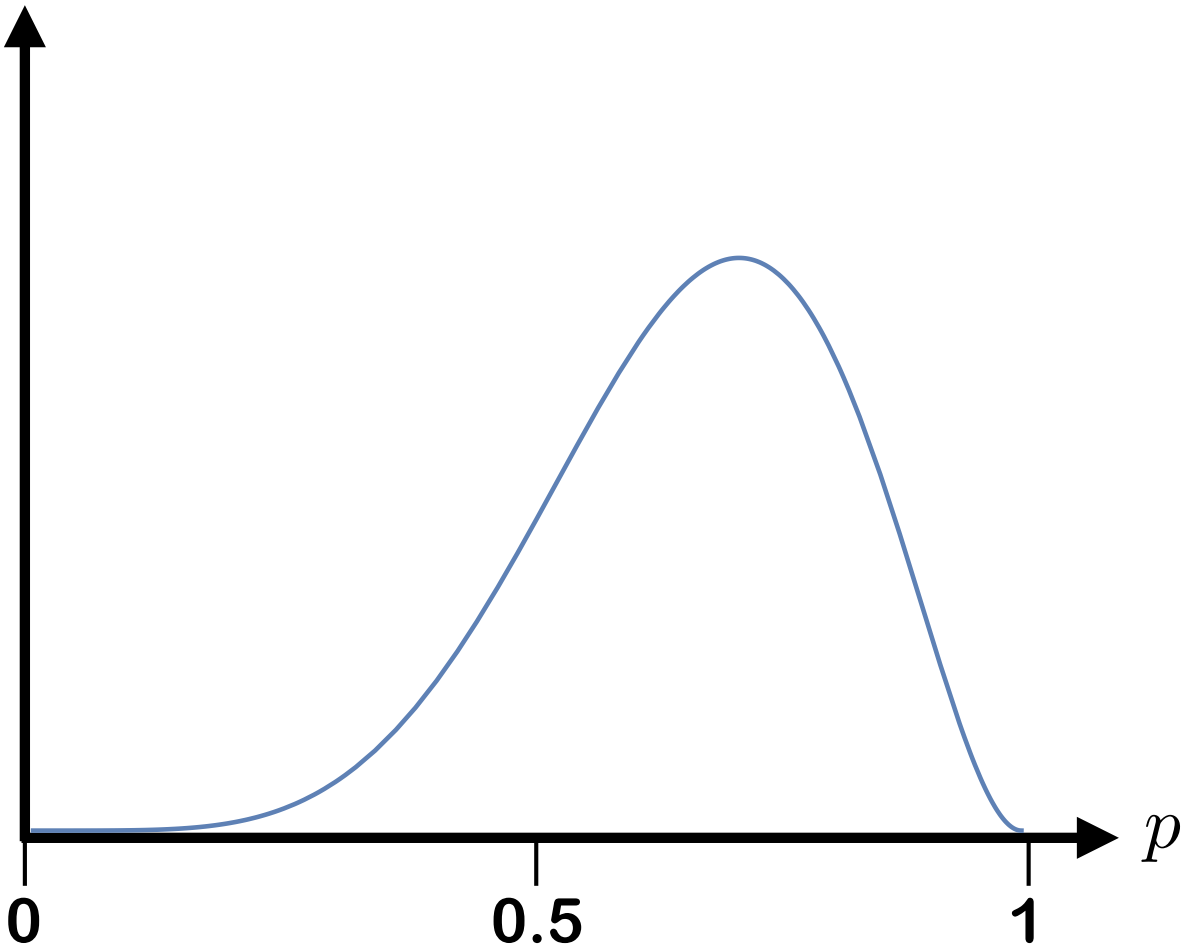
What is the likelihood of having observed the sequence of outcomes HTHHHTH?

likelihood = $p^5(1-p)^2$

Setting up the Problem

likelihood =

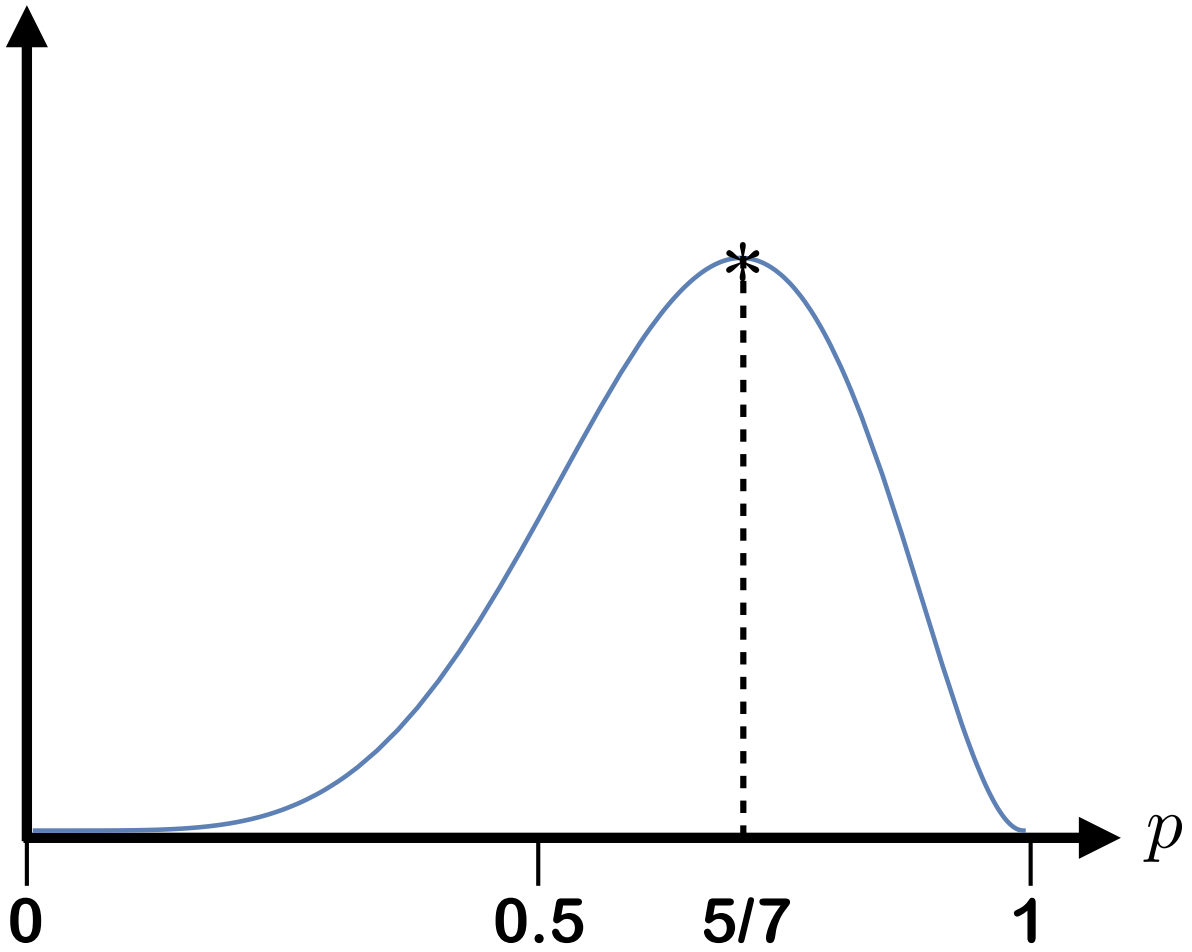
$$p^5(1-p)^2$$



Setting up the Problem

likelihood =

$$p^5(1-p)^2$$



Setting up the Problem

Step 1) Write down the model.

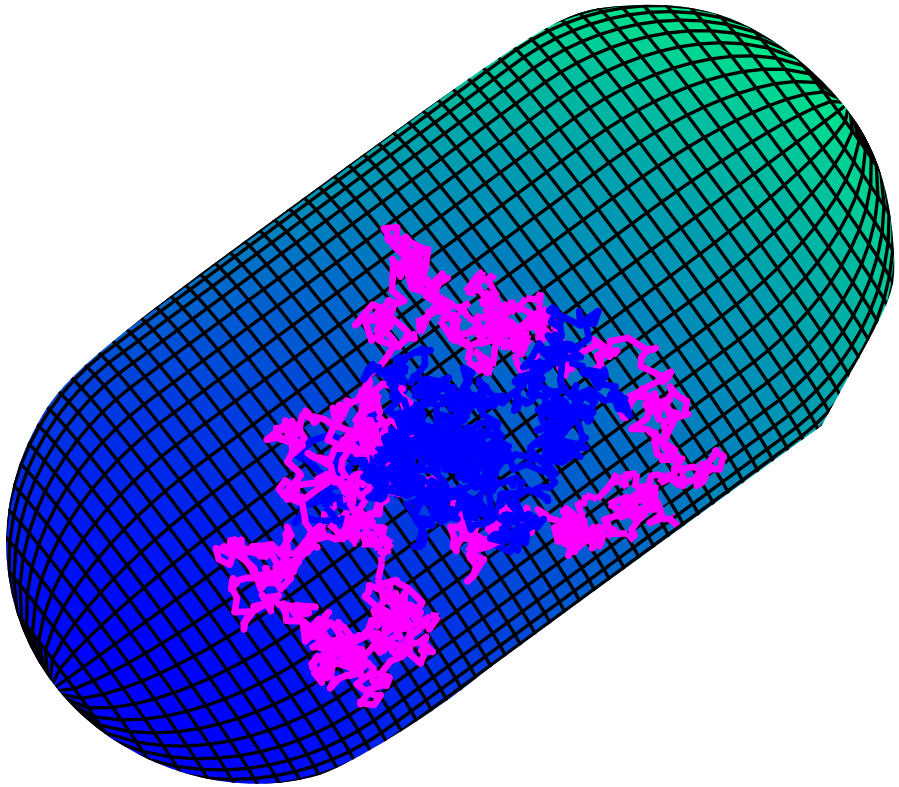
Step 2) Write down the likelihood of your data under the assumption of your model.

Step 3) Maximize your likelihood to determine the parameters of your model

Steps 1-3 are involved in performing "Maximum Likelihood"

Slightly more complicated problem
(Step 2 is harder)

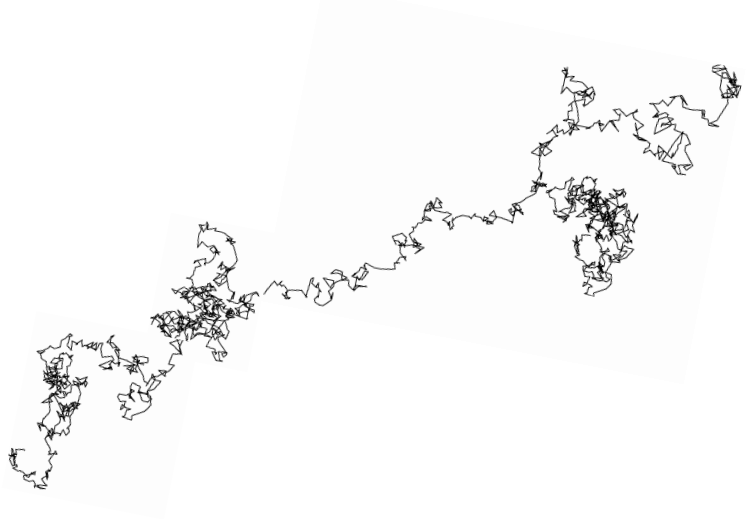
Step 1) Write down the model.



From a single particle track, we want to determine it's diffusion coefficient

Slightly more complicated problem (Step 2 is harder)

Step 1) Write down the model.



$$\frac{\partial p}{\partial t} = D\nabla^2 p$$

Slightly more complicated problem (Step 2 is harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

$$P(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N | D)$$

Now data points are vectors/positions in 3D

$$\mathbf{y_i} = \{x_i, y_i, z_i\}$$

Slightly more complicated problem (Step 2 is harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

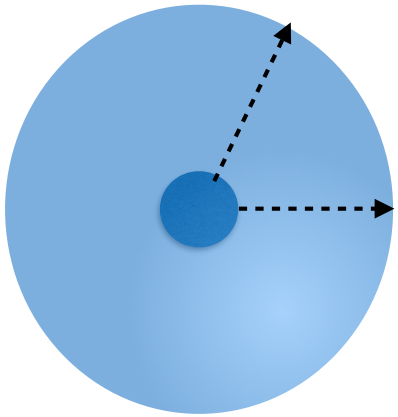
For coin flip, we had…

$$P(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N|p) = P(\mathbf{y}_1|p)P(\mathbf{y}_2|p)\cdots P(\mathbf{y}_N|p)$$

But that doesn't make sense for diffusion. How can we pick positions at random? Where we land at time t depends on where we just were!

Slightly more complicated problem (Step 2 is harder)

Step 2) Write down the likelihood of your data under the assumption of your model.



$$P(\mathbf{y}_1, \mathbf{y}_2|D) = P(\mathbf{y}_2|\mathbf{y}_1, D)P(\mathbf{y}_1)$$

Slightly more complicated problem (Step 2 is harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

$$P(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N | D) = P(\mathbf{y}_N | \mathbf{y}_{N-1}, D) \cdots P(\mathbf{y}_3 | \mathbf{y}_2, D) P(\mathbf{y}_2 | \mathbf{y}_2, D) P(\mathbf{y}_1)$$

What are $P(\mathbf{y}_t | \mathbf{y}_{t-\delta t}, D)$ **?**

Slightly more complicated problem (Step 2 is harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

What are $P(\mathbf{y}_t|\mathbf{y}_{t-\delta t}, D)$ **?**

The model tells us (requires solving the PDE):

$$P(\mathbf{y}_t|\mathbf{y}_{t-\delta t}, D) = \frac{1}{(4\pi D\delta t)^{3/2}} e^{-\frac{(\mathbf{y}_t - \mathbf{y}_{t-\delta t})^2}{4D\delta t}}$$

Slightly more complicated problem (Step 2 is harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

Now we can write down the full-likelihood:

$$p(\mathbf{y}_{1:N}|D) = \frac{1}{(4\pi D\delta t)^{3(N-1)/2}} e^{-\sum_{i=2}^{N} \frac{(\mathbf{y}_i - \mathbf{y}_{i-1})^2}{4D\delta t}} p(\mathbf{y}_1)$$

Slightly more complicated problem (Step 2 is harder)

Step 3) Maximize your likelihood to determine the parameters of your model

$$\log p(\mathbf{y}_{1:N}|D,\delta t) = -\frac{3(N-1)}{2}\log\left(4\pi D\delta t\right) - \sum_{i=2}^{N}\frac{(\mathbf{y}_i-\mathbf{y}_{i-1})^2}{4D\delta t} + \log p(\mathbf{y}_1)$$

Take derivative with respect to D and set to 0

$$-\frac{3(N-1)}{2D} + \sum_{i=2}^{N}\frac{(\mathbf{x}_i-\mathbf{x}_{i-1})^2}{4D^2\delta t} = 0$$

Slightly more complicated problem (Step 2 is harder)

Step 3) Maximize your likelihood to determine the parameters of your model

Take derivative with respect to D and set to 0

$$-\frac{3(N-1)}{2D}+\sum_{i=2}^{N}\frac{(\mathbf{y}_i-\mathbf{y}_{i-1})^2}{4D^2\delta t}=0$$

$$6D=\frac{1}{N-1}\sum_{i=2}^{N}\frac{(\mathbf{y}_i-\mathbf{y}_{i-1})^2}{\delta t}$$

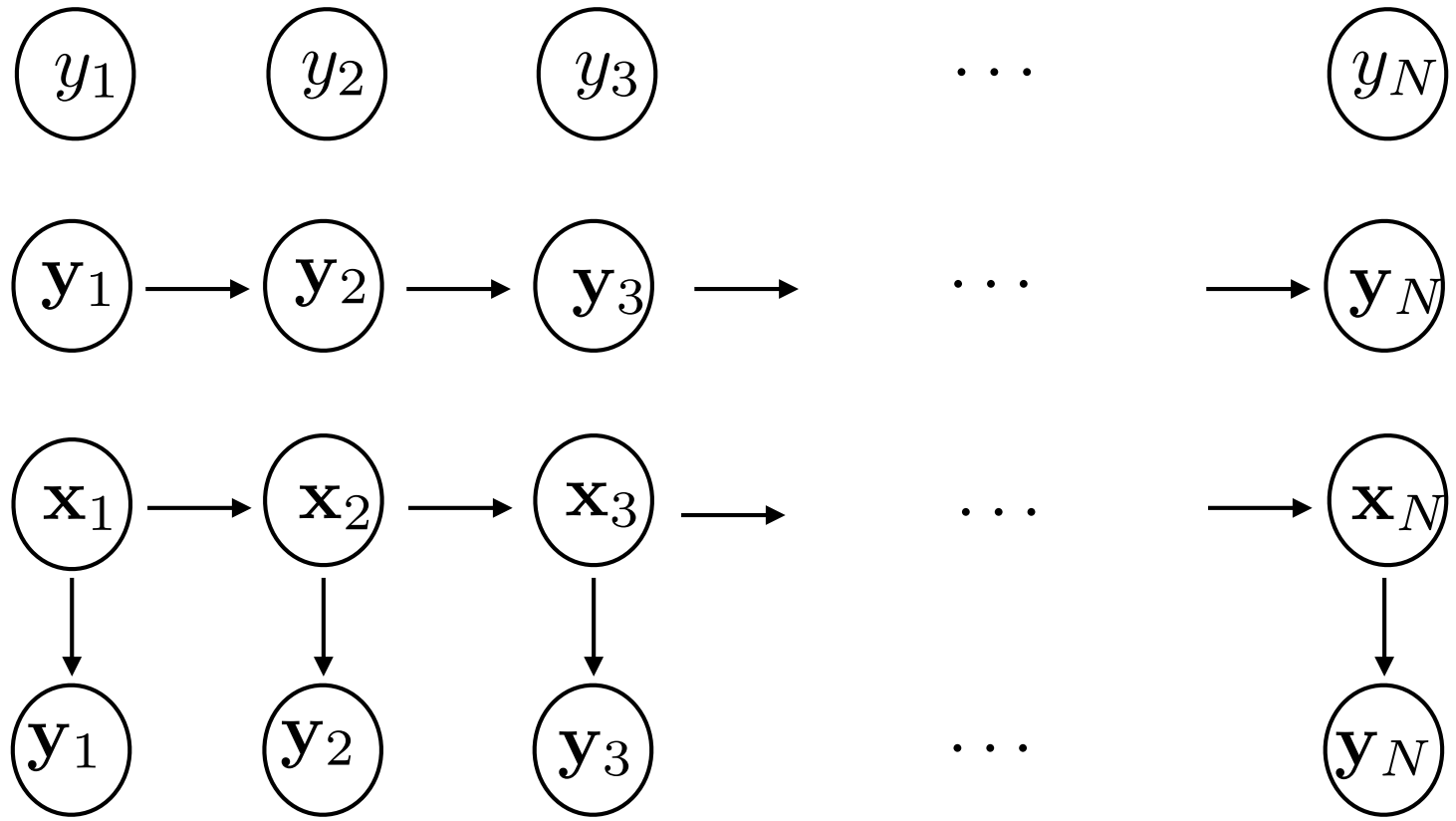
Graphical Models

Coin flips

iid = identical independently distributed

Diffusion

Diffusion w measurement noise



Slightly more complicated problem (Steps 2+3 are harder)

Step 1) Write down the model.

Kinetic model

$$\frac{\partial p}{\partial t} = D\nabla^2 p$$

Slightly more complicated problem (Steps 2+3 are harder)

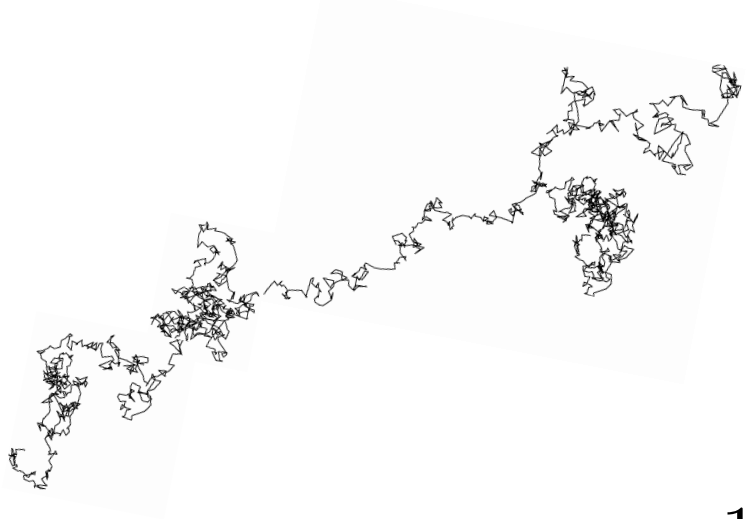
Step 1) Write down the model.

Kinetic model

$$x_t | x_{t-\delta t}, D \sim \frac{1}{\sqrt{4\pi D \delta t}} e^{-\frac{(x_t - x_{t-\delta t})^2}{4D\delta t}}$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 1) Write down the model.



Kinetic model

$$x_t | x_{t-\delta t}, D \sim \frac{1}{\sqrt{4\pi D \delta t}} e^{-\frac{(x_t - x_{t-\delta t})^2}{4D\delta t}}$$

$$y_t | y_{t-\delta t}, D \sim \frac{1}{\sqrt{4\pi D \delta t}} e^{-\frac{(y_t - y_{t-\delta t})^2}{4D\delta t}}$$

$$z_t | z_{t-\delta t}, D \sim \frac{1}{\sqrt{4\pi D \delta t}} e^{-\frac{(z_t - z_{t-\delta t})^2}{4D\delta t}}$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 1) Write down the model.

Kinetic model

$$x_t | x_{t-\delta t}, D \sim \frac{1}{\sqrt{4\pi D \delta t}} e^{-\frac{(x_t - x_{t-\delta t})^2}{4D\delta t}}$$

Observation model

$$\mathbf{y}_t | \mathbf{x}_t, \sigma^2 \sim \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{(\mathbf{y}_t - \mathbf{x}_t)^2}{2\sigma^2}}$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 1) Write down the model.

Kinetic model $$x_t | x_{t-\delta t}, D \sim \frac{1}{\sqrt{4\pi D \delta t}} e^{-\frac{(x_t - x_{t-\delta t})^2}{4D\delta t}}$$

emission distribution

Observation model $$\mathbf{y}_t | \mathbf{x}_t, \sigma^2 \sim \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{(\mathbf{y}_t - \mathbf{x}_t)^2}{\sigma^2}}$$

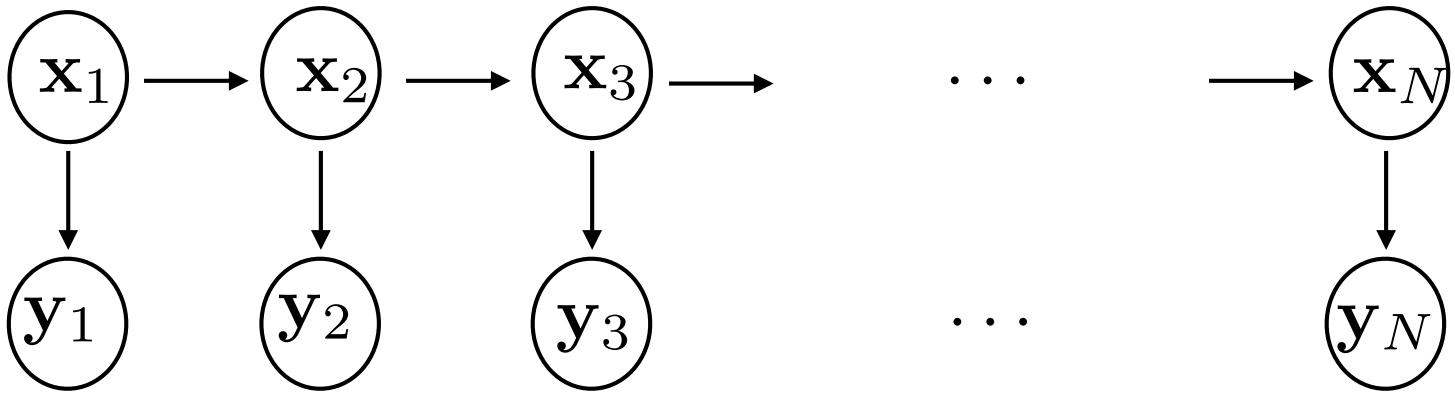
Assumptions: measurement error in all directions is the same. Diffusion is isotropic, only one diffusion coefficient etc…

Slightly more complicated problem (Steps 2+3 are harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

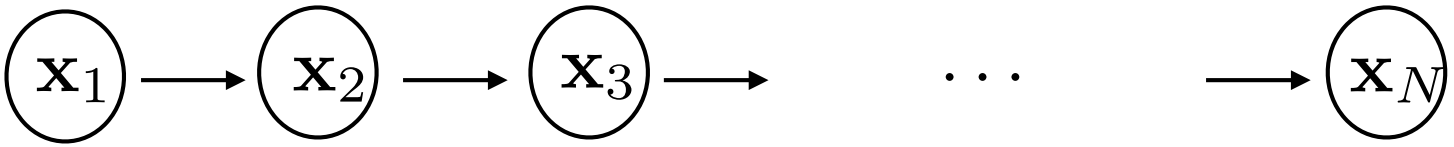
$$P(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N | D, \sigma^2)$$

Diffusion w measurement noise



Slightly more complicated problem (Steps 2+3 are harder)

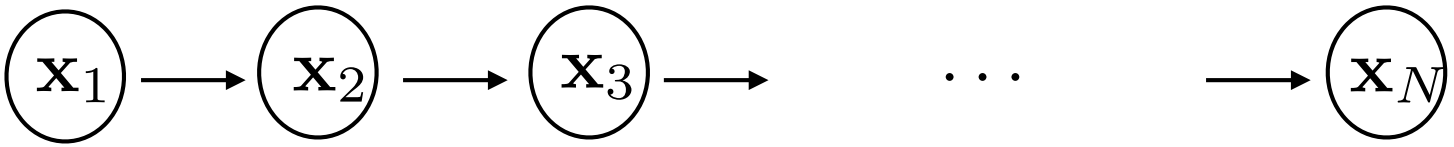
Step 2) Write down the likelihood of your data under the assumption of your model.



$$P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N | D) = P(\mathbf{x}_N | \mathbf{x}_{N-1}, D) \cdots P(\mathbf{x}_2 | \mathbf{x}_1, D) P(\mathbf{x}_1)$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

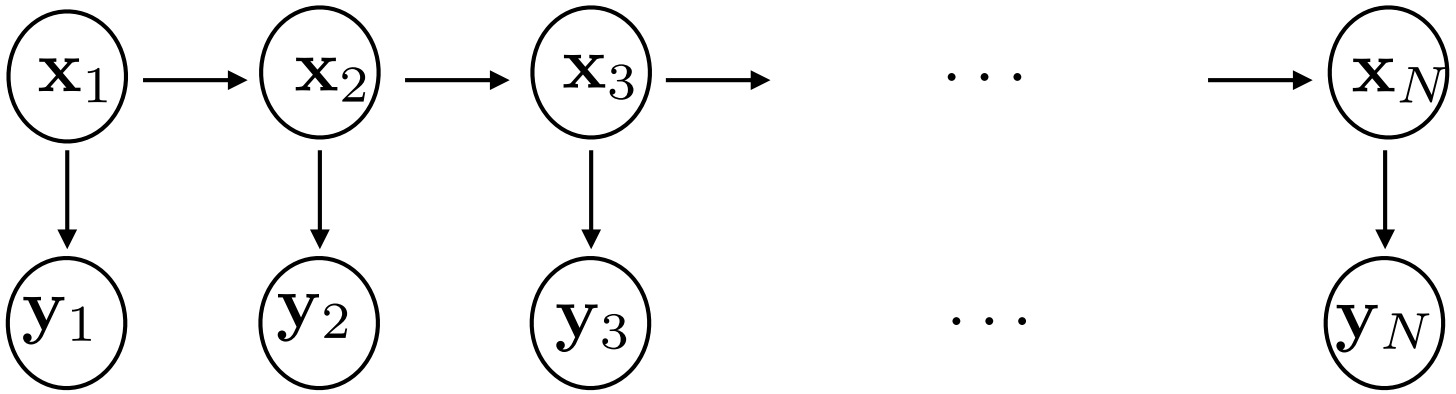


$$P(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N | D) = P(\mathbf{x}_N | \mathbf{x}_{N-1}, D) \cdots P(\mathbf{x}_2 | \mathbf{x}_1, D) P(\mathbf{x}_1)$$

$$P(\mathbf{x}_1) = \delta(\mathbf{x}_1 - 0)$$

Slightly more complicated problem (Steps 2+3 are harder)

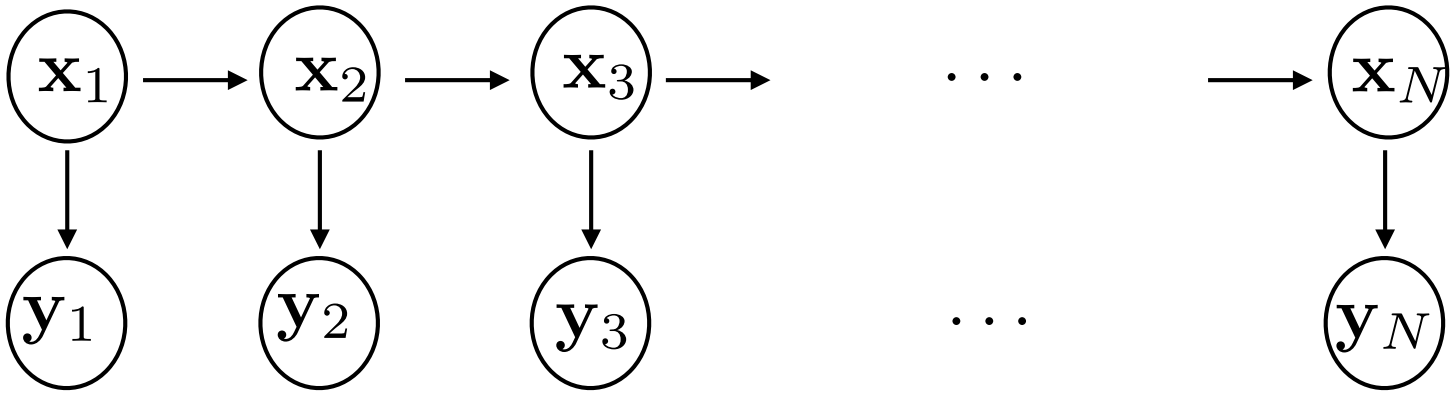
Step 2) Write down the likelihood of your data under the assumption of your model.



$$P(\mathbf{y}_N|\mathbf{x}_N,\sigma^2)P(\mathbf{x}_N|\mathbf{x}_{N-1},D)\cdots P(\mathbf{y}_2|\mathbf{x}_2,\sigma^2)P(\mathbf{x}_2|\mathbf{x}_1,D)P(\mathbf{y}_1|\mathbf{x}_1,\sigma^2)P(\mathbf{x}_1)$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 2) Write down the likelihood of your data under the assumption of your model.



$$P(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{y}_1, \cdots, \mathbf{y}_N | D, \sigma^2) =$$

$$P(\mathbf{y}_N|\mathbf{x}_N, \sigma^2) P(\mathbf{x}_N|\mathbf{x}_{N-1}, D) \cdots P(\mathbf{y}_2|\mathbf{x}_2, \sigma^2) P(\mathbf{x}_2|\mathbf{x}_1, D) P(\mathbf{y}_1|\mathbf{x}_1, \sigma^2) P(\mathbf{x}_1)$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

Complete-data likelihood

$$P(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{y}_1, \cdots, \mathbf{y}_N | D, \sigma^2)$$

Incomplete-data likelihood

$$P(\mathbf{y}_1, \cdots, \mathbf{y}_N | D, \sigma^2)$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 2) Write down the likelihood of your data under the assumption of your model.

$$P(\mathbf{y}_1, \cdots, \mathbf{y}_N | D, \sigma^2)$$

$$= \int d\mathbf{x}_1 \cdots d\mathbf{x}_N \; P(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{y}_1, \cdots, \mathbf{y}_N | D, \sigma^2)$$

Hidden/latent variables

Observation

We marginalize over the complete-data likelihood to get the incomplete-data likelihood

Slightly more complicated problem (Steps 2+3 are harder)

Step 3) Maximize your likelihood to determine the parameters of your model

$$\frac{\partial}{\partial \sigma^2} \log P(\mathbf{y}_1, \cdots, \mathbf{y}_N, | D, \sigma^2) = 0$$

$$\frac{\partial}{\partial D} \log P(\mathbf{y}_1, \cdots, \mathbf{y}_N, | D, \sigma^2) = 0$$

Slightly more complicated problem (Steps 2+3 are harder)

Step 3) Maximize your likelihood to determine the parameters of your model

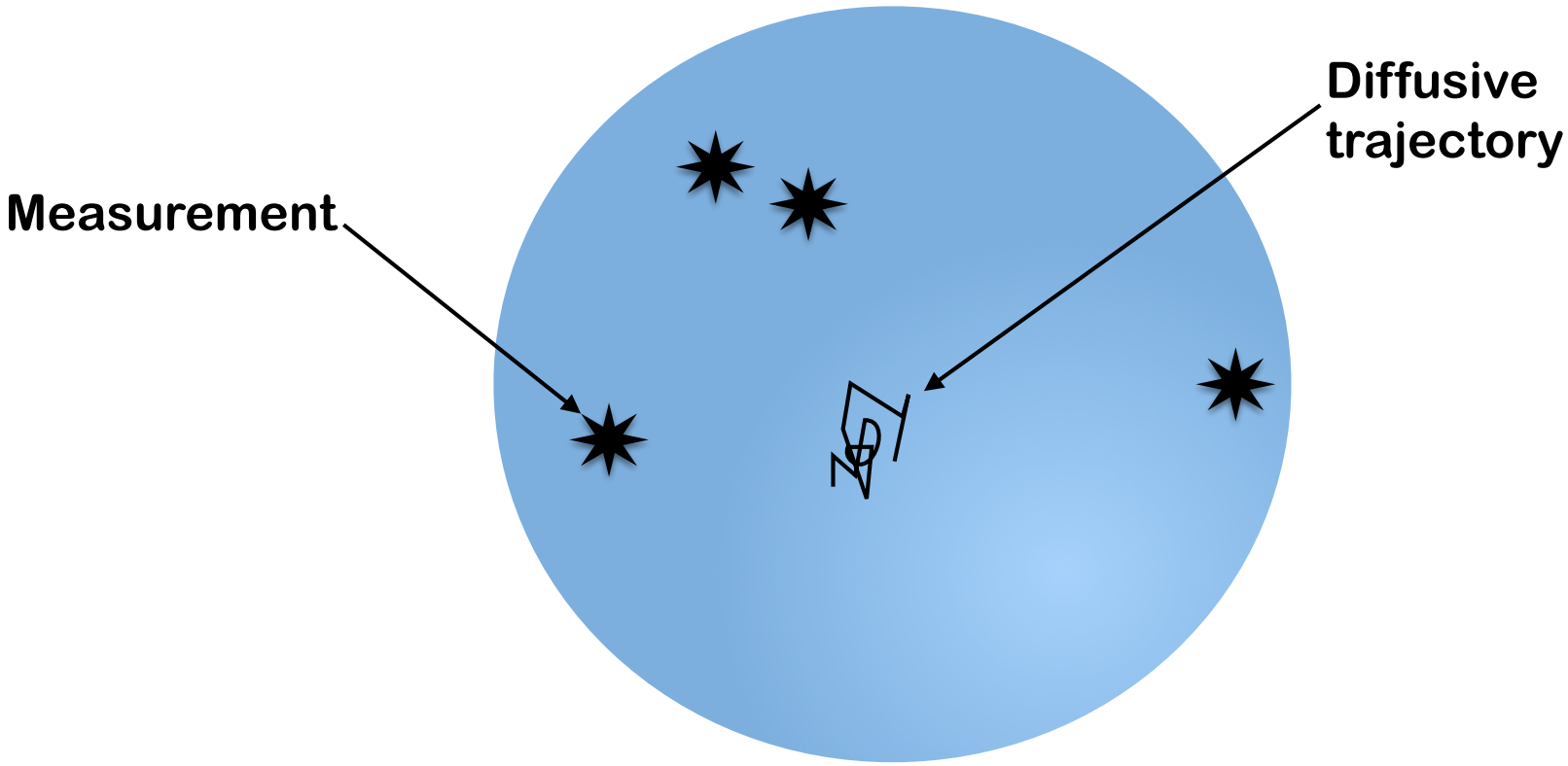
For 1 data point...

$$6D = \frac{(\mathbf{y}_2 - \mathbf{y}_1)^2}{\delta t} \quad - \frac{5\sigma^2}{\delta t}$$

true for small σ^2

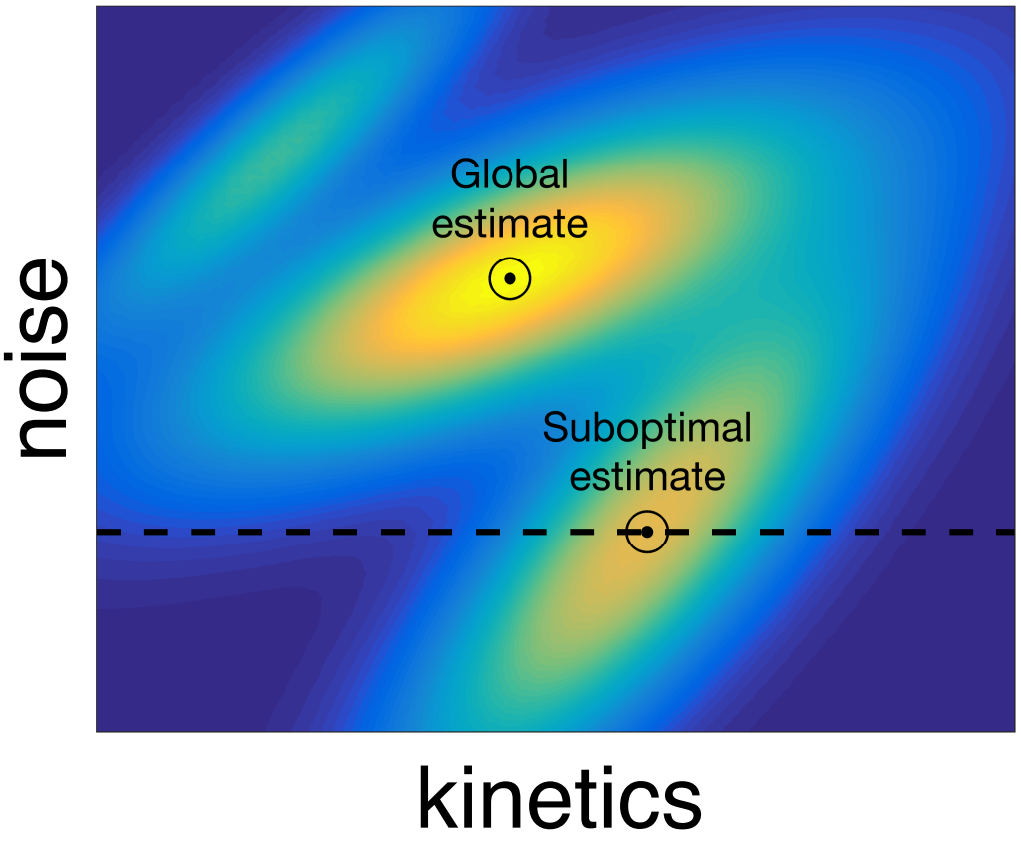
Slightly more complicated problem (Steps 2+3 are harder)

$$6D = \frac{(\mathbf{y}_2 - \mathbf{y}_1)^2}{\delta t} - \frac{5\sigma^2}{\delta t}$$



Intuitively this makes sense. If measurement noise is large, we overestimate diffusion coefficient and have to correct for the fact that the true diffusion coefficient appears artificially large.

Slightly more complicated problem (Steps 2+3 are harder)



In general calculating the incomplete-data likelihood is very difficult

$$\int d\mathbf{x}_{1:N} \;\; P(\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{y}_1, \cdots, \mathbf{y}_N | D, \sigma^2)$$

In general calculating the incomplete-data likelihood is very difficult

$$\int d\mathbf{x}_{1:N} \quad P(\mathbf{x}_{1:N}, \mathbf{y}_{1:N} | \boldsymbol{\theta})$$

Conceptual EM algorithm

E Step:

$$\int d\mathbf{x}_{1:N} p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N}, \boldsymbol{\theta}) \log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}|\boldsymbol{\theta})$$

M Step: Maximize with respect to θ

$$\int d\mathbf{x}_{1:N} p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N}, \boldsymbol{\theta}) \log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}|\boldsymbol{\theta})$$

This maximization is still quite difficult….

Thus we want to iteratively determine the parameters where we call j the iteration index.

EM algorithm

Initiate $\boldsymbol{\theta}_{j=0} = \boldsymbol{\theta}_0$

E Step:

$$Q(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_j) = \int d\mathbf{x}_{1:N} p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N}, \boldsymbol{\theta}_{j-1}) \log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}|\boldsymbol{\theta}_j)$$

M Step: Maximize $Q(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_j)$ **with respect to** $\boldsymbol{\theta}_j$

EM algorithm

Initiate $\boldsymbol{\theta}_{j=0} = \boldsymbol{\theta}_0$

E Step:

$$Q(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_j) = \int d\mathbf{x}_{1:N} p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N}, \boldsymbol{\theta}_{j-1}) \log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}|\boldsymbol{\theta}_j)$$

M Step: Maximize $Q(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_j)$ **with respect to** $\boldsymbol{\theta}_j$

$$\boldsymbol{\theta}_j = f(\boldsymbol{\theta}_{j-1})$$

Iterate away until

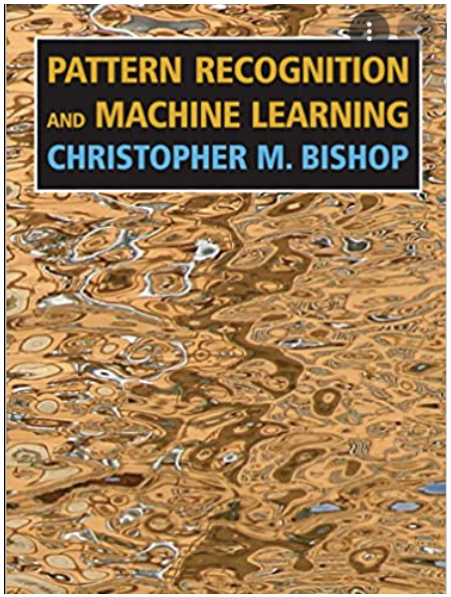
$$|\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}| < \epsilon$$

EM algorithm

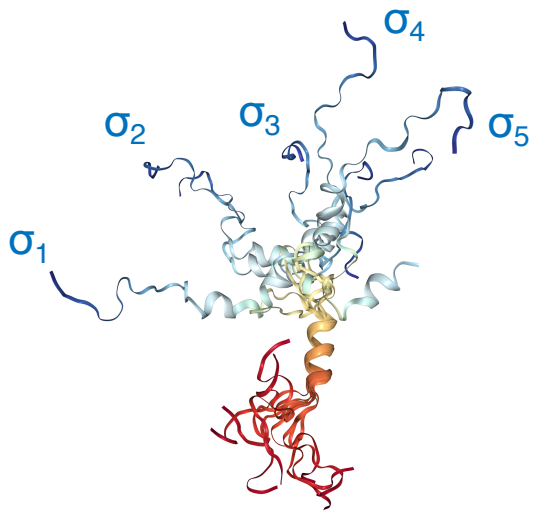
$$Q(\boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_j) = \int d\mathbf{x}_{1:N} p(\mathbf{x}_{1:N}|\mathbf{y}_{1:N}, \boldsymbol{\theta}_{j-1}) \log p(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}|\boldsymbol{\theta}_j)$$

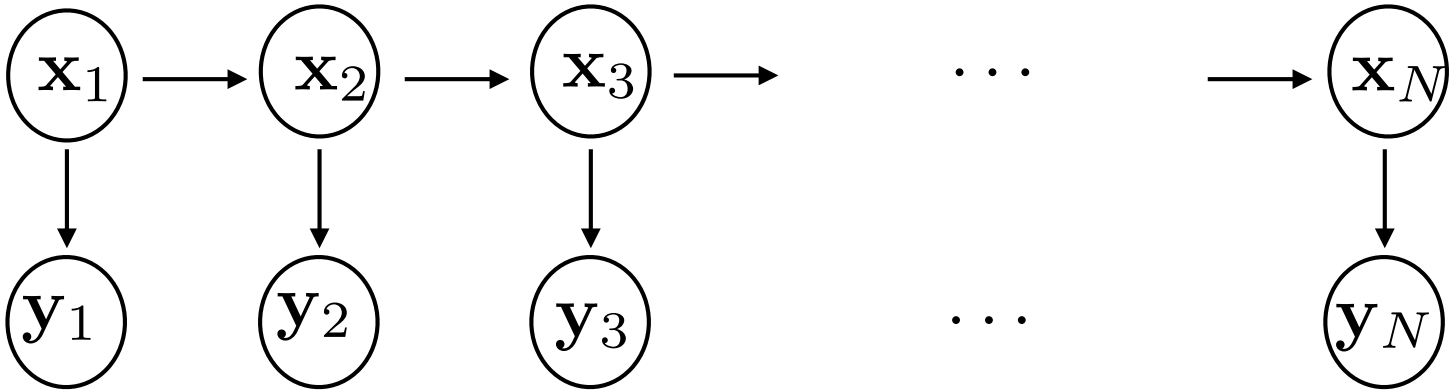
In general this is hard to calculate…

For dynamical systems, it will require "filters"

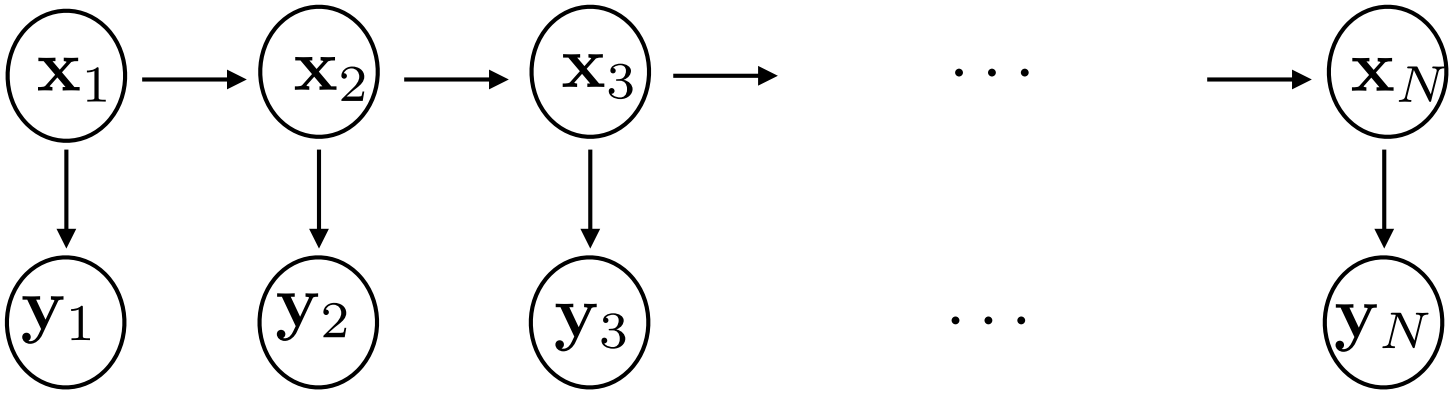


Slightly more complicated problem (Step 3 is hard)



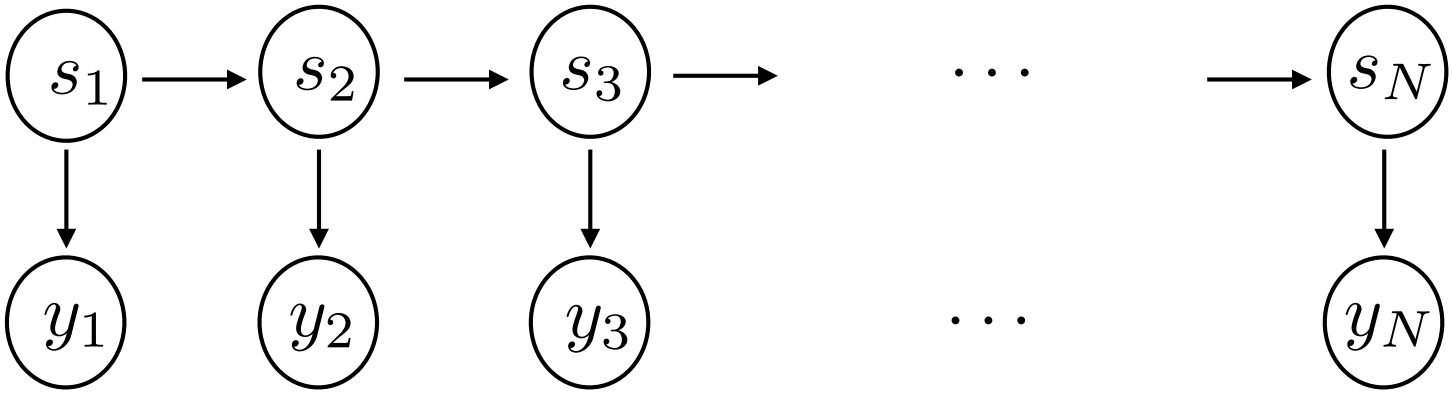


Slightly more complicated problem (Step 3 is hard)



Different from diffusion only in so far that space is discrete and in 1D

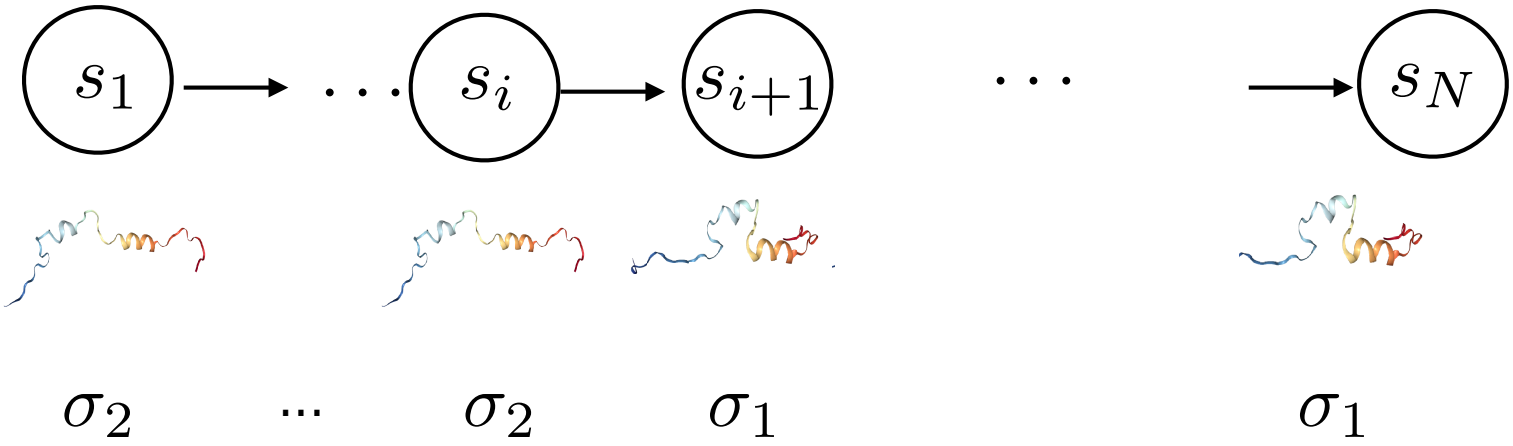
Slightly more complicated problem (Step 3 is hard)



Different from diffusion only in so far that space is discrete and in 1D

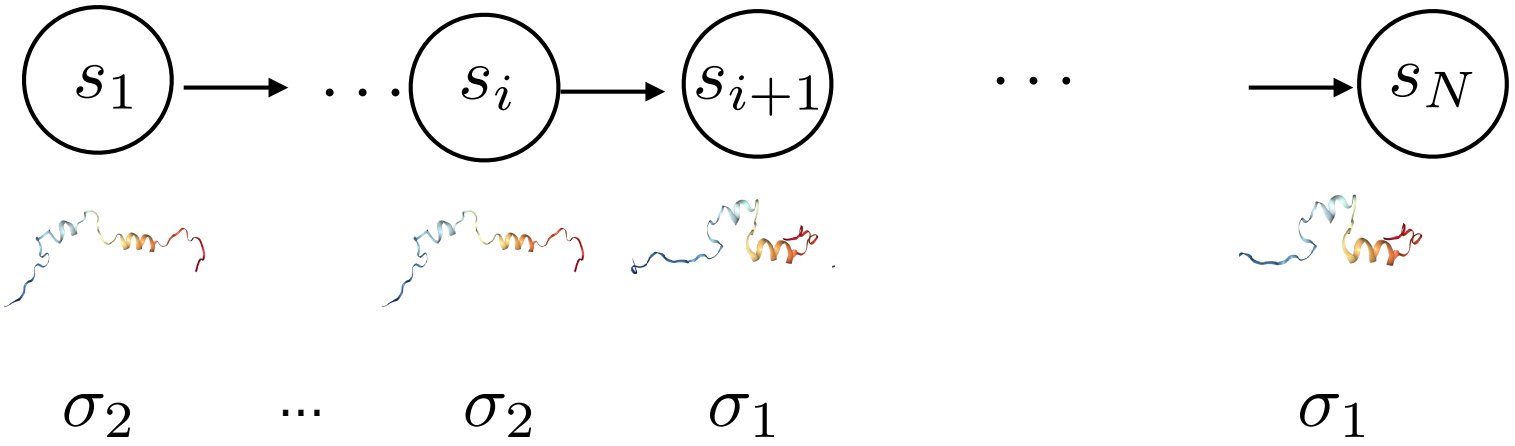
Slightly more complicated problem (Step 3 is hard)

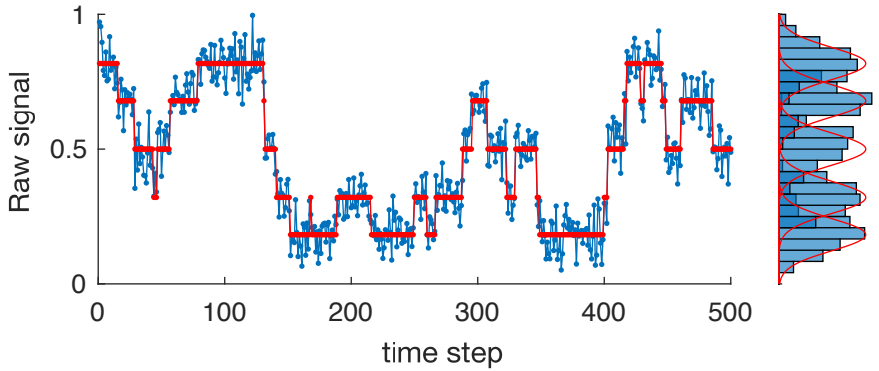
The Hidden Markov Model



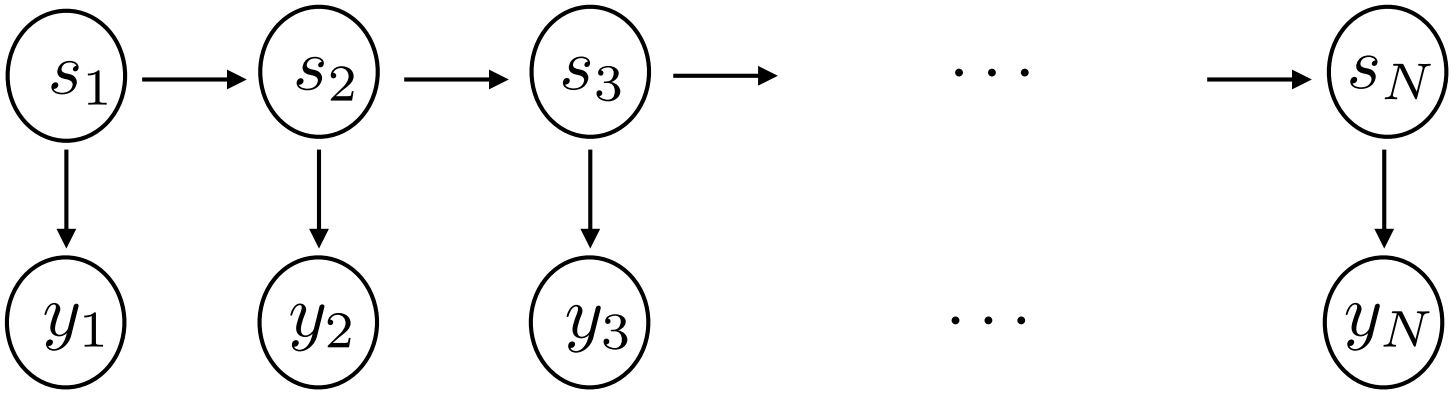
Slightly more complicated problem (Step 3 is hard)

The Hidden Markov Model





Slightly more complicated problem (Step 3 is hard)



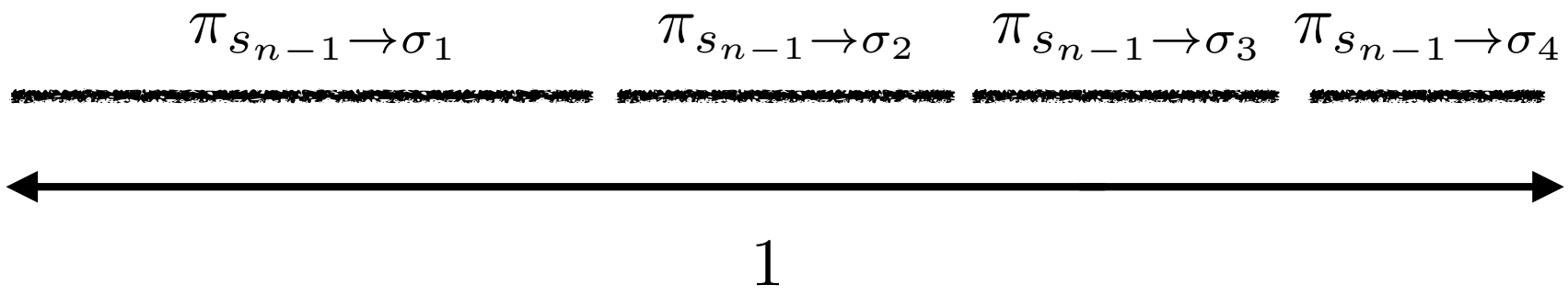
Kinetic model

$$s_n | s_{n-1} \sim Categorical(\pi_{s_{n-1} \to \sigma_1}, \cdots, \pi_{s_{n-1} \to \sigma_K})$$

Observation model $y_n | s_n \sim F_{s_n}(\tilde{\phi}_{s_n})$

Slightly more complicated problem (Step 3 is hard)

$$s_n | s_{n-1} \sim Categorical(\pi_{s_{n-1} \to \sigma_1}, \cdots, \pi_{s_{n-1} \to \sigma_K})$$



Slightly more complicated problem (Step 3 is hard)

$$\tilde{\tilde{\pi}} = \begin{pmatrix} \pi_{\sigma_1 \to \sigma_1} & \pi_{\sigma_1 \to \sigma_2} & \pi_{\sigma_1 \to \sigma_3} & \pi_{\sigma_1 \to \sigma_4} \\ \pi_{\sigma_2 \to \sigma_1} & \pi_{\sigma_2 \to \sigma_2} & \pi_{\sigma_2 \to \sigma_3} & \pi_{\sigma_2 \to \sigma_4} \\ \pi_{\sigma_3 \to \sigma_1} & \pi_{\sigma_3 \to \sigma_2} & \pi_{\sigma_3 \to \sigma_3} & \pi_{\sigma_3 \to \sigma_4} \\ \pi_{\sigma_4 \to \sigma_1} & \pi_{\sigma_4 \to \sigma_2} & \pi_{\sigma_4 \to \sigma_3} & \pi_{\sigma_4 \to \sigma_4} \end{pmatrix}$$

Slightly more complicated problem (Step 3 is hard)

The idea is always the same…

Write down complete-data likelihood

$$P(y_{1:N}, s_{1:N} | \tilde{\tilde{\pi}}, (\sigma^2)_{1:K}, \mu_{1:K}, \ldots)$$

But we are interested in maximizing the incomplete-data likelihood

$$\sum_{s_{1:N}} P(y_{1:N}, s_{1:N} | \tilde{\tilde{\pi}}, (\sigma^2)_{1:K}, \mu_{1:K}, \ldots)$$

If we are just interested in the most probable state sequence we use the "Viterbi algorithm"

Slightly more complicated problem (Step 3 is hard)

But we are interested in maximizing the incomplete data likelihood

$$\sum_{s_{1:N}} P(y_{1:N}, s_{1:N} | \tilde{\tilde{\pi}}, (\sigma^2)_{1:K}, \mu_{1:K}, \ldots)$$

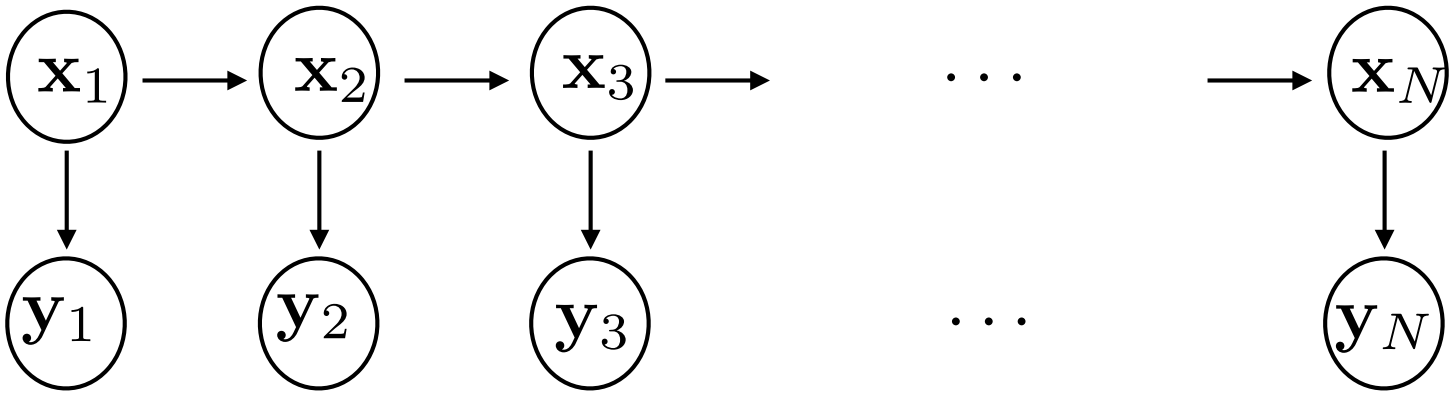
Since we cannot do this exactly, we will use EM

N.B. for HMM

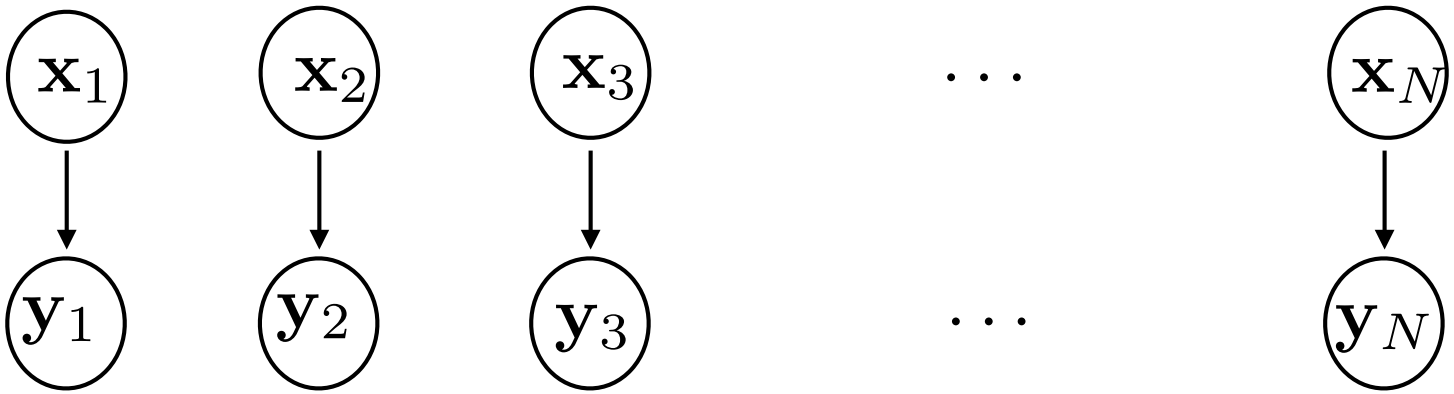
-People normally use EM to approximately evaluate the maximum likelihood and (within EM) use filtering.

-You need to put in by hand: the number of states, specify the emission distribution

Clustering is another example of a latent variable model

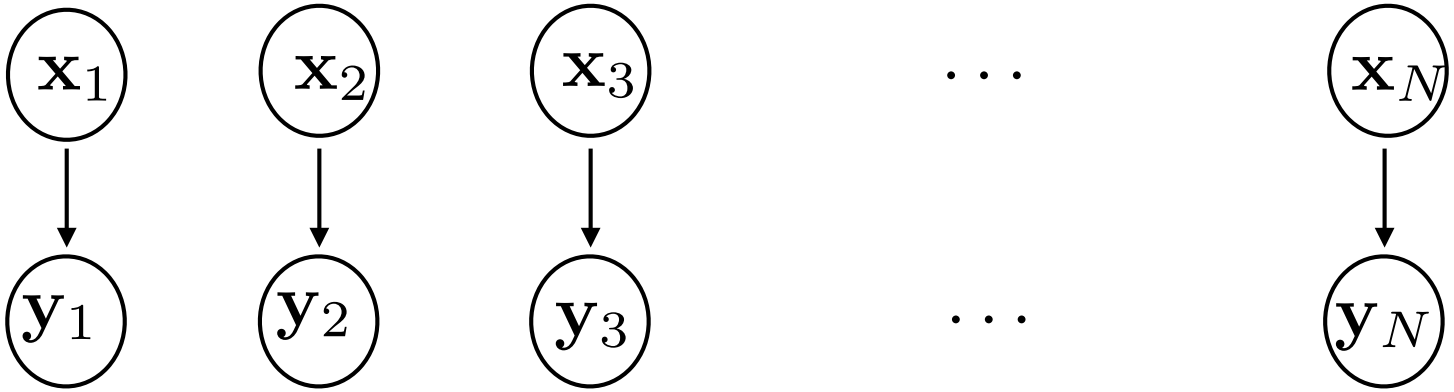


Clustering is another example of a latent variable model



no dynamics…

Clustering is another example of a latent variable model

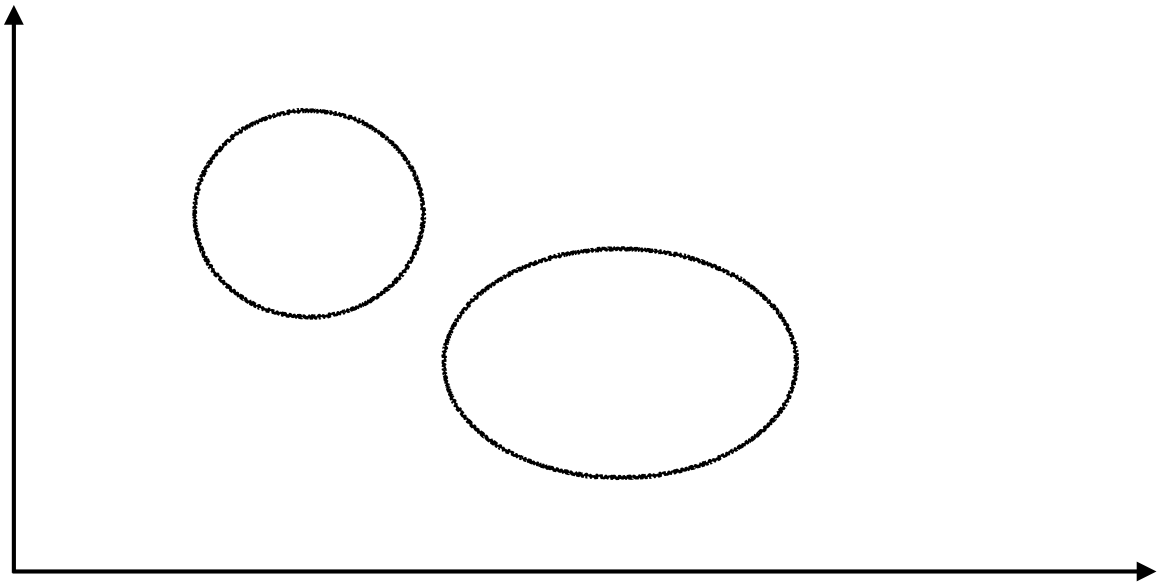


System model $x_n|\tilde{\pi} \sim Categorical(\pi_1, \pi_2, \cdots, \pi_K)$

Observation model $\mathbf{y}_n|\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n^2 \sim \frac{1}{(2\pi\sigma_n)^{3/2}} e^{-\frac{(\mathbf{y}_n - \boldsymbol{\mu}_n)^2}{2\sigma_n^2}}$

Clustering is another example of a latent variable model

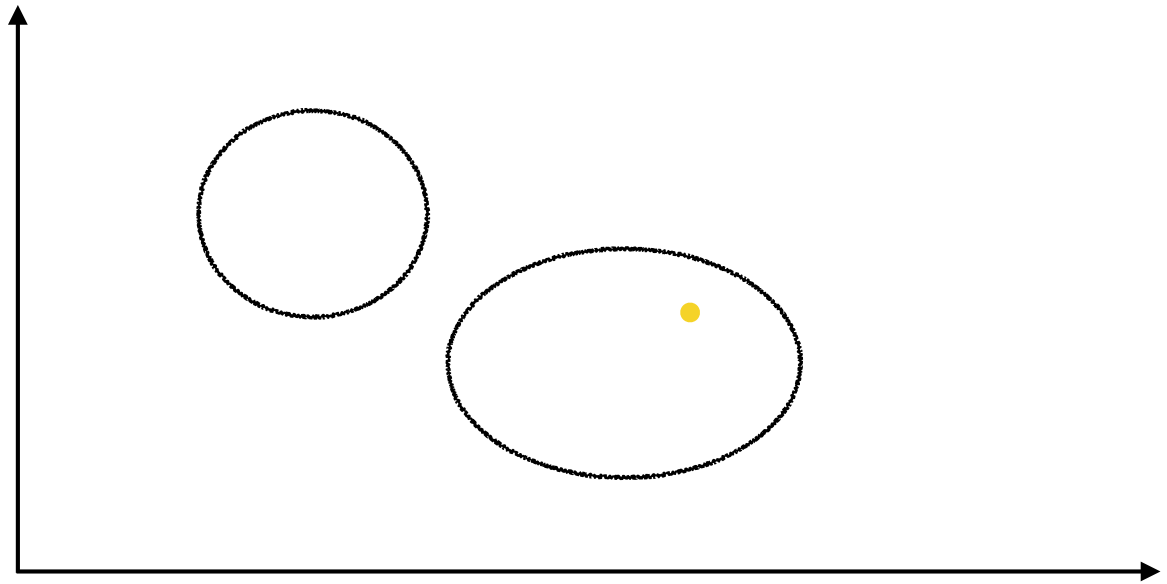
System model $x_n|\tilde{\pi} \sim Categorical(\pi_1, \pi_2, \cdots, \pi_K)$



Clustering is another example of a latent variable model

System model $x_n|\tilde{\pi} \sim Categorical(\pi_1, \pi_2, \cdots, \pi_K)$

Observation model $\mathbf{y}_n|\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n^2 \sim \frac{1}{(2\pi\sigma_n)^{3/2}} e^{-\frac{(\mathbf{y}_n - \boldsymbol{\mu}_n)^2}{2\sigma_n^2}}$



In clustering, just as with HMMs, we build complete-data likelihoods, then derive incomplete-data likelihoods to be maximized (this can all be done exactly or approximately, e.g. through variational methods such as EM)

$$\mathbf{y}_{1:N} \rightarrow \pi_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\sigma}_{1:K}$$

Big assumptions: number of clusters is inputed by hand, the emission distribution is specific by hand, etc…

Likelihoods need not only be maximized

e.g. likelihood ratio test can be used to compare parameter values

$$\frac{P(\mathbf{y}_{1:N}|D = 10\mu m^2/s)}{P(\mathbf{y}_{1:N}|D = 5\mu m^2/s)}$$

e.g. likelihoods's curvature near maximum value tells you something about estimate uncertainty (how sharp the likelihood is around the maximum)

$$\frac{\partial^2}{\partial D^2} P(\mathbf{y}_{1:N}|D)|_{D=D*}$$

Physics dictates likelihoods.
A good understanding of the data collection process and underlying physics can be used to approximate likelihoods

A proper understanding and use of likelihoods avoids having to de-noise the time trace, average down the data etc…

e.g. of more sophisticated models we cover in my class

$$\zeta d\mathbf{x} = \mathbf{v}dt + BdW_t$$

$$W_t - W_{t-dt} \sim Normal(0, dt)$$

$$\mathbf{y}_t = \frac{1}{t_E}\int_{t-t_E}^{t} d\mathbf{x}(t) + B'(W_t - W_{t-t_E})$$

-the above is relevant if you have finite exposure time

e.g. of more sophisticated models we cover in my class

$$\zeta d\mathbf{x} = \mathbf{v}dt + BdW_t$$
$$W_t - W_{t-dt} \sim Normal(0, dt)$$
$$\mathbf{y}_t = \frac{1}{t_E}\int_{t-t_E}^{t} d\mathbf{x}(t) + B'(W_t - W_{t-t_E})$$

$$\sigma_n|\sigma_{n-1}, \tilde{\pi}_{n-1} \sim Cat(\tilde{\pi}_{n-1})$$
$$W_t - W_{t-dt} \sim Normal(0, dt)$$
$$\zeta d\mathbf{x} = -\nabla U_{\sigma_n}(\mathbf{x})dt + B_{\sigma_n}dW_t.$$
$$\mathbf{y}_t|\mathbf{x}_t \sim p(\mathbf{y}_t|\mathbf{x}_t, \boldsymbol{\theta}_o)$$

-the above is relevant if you switch between behaviors

Everything we have done so far is frequentist.

-we have assumed that there exists true parameter values (as opposed to assuming that parameters themselves are random variables distributed according to some probability distribution)

-frequentist (at least as our discussion here goes) means maximum likelihood

However…

-we may want to calculate a full distribution over parameters … like $P(D|\mathbf{y}_{1:N})$ instead of just D

-we may want to bias our estimates for the parameter by inputting prior knowledge (e.g. we may have a range to within a order of magnitude what the diffusion coefficient should be).

-we may want to grow the dimensionality of our model based on the the data…

The Bayesian paradigm…from Laplace!

$$P(\boldsymbol{\theta}) \;\rightarrow\; P(\boldsymbol{\theta}|\mathbf{y}_{1:N})$$

prior **posterior**

The Bayesian paradigm…from Laplace!

Bayes' theorem

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N})P(\mathbf{y}_{1:N}) = P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

posterior likelihood prior

The Bayesian paradigm…from Laplace!

Bayes' theorem

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N})P(\mathbf{y}_{1:N}) = P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

posterior — **likelihood** — **prior**

normalization

what we want

dictated by physics of the experiment at hand

specified by hand - or dictated by physics outside the experiment generating the data

The Bayesian paradigm…from Laplace!

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \quad \propto \quad P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

Imagine repeated experiments…

$$P(\boldsymbol{\theta}|\mathbf{y}_1) \propto P(\mathbf{y}_1|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

The Bayesian paradigm…from Laplace!

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \quad \propto \quad P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

Imagine repeated experiments…

$$P(\boldsymbol{\theta}|\mathbf{y}_1) \propto P(\mathbf{y}_1|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

$$P(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) \propto P(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{y}_1)P(\boldsymbol{\theta}|\mathbf{y}_1)$$

The Bayesian paradigm…from Laplace!

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \quad \propto \quad P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

Imagine repeated experiments…

$$P(\boldsymbol{\theta}|\mathbf{y}_1) \propto P(\mathbf{y}_1|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

$$P(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) \propto P(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{y}_1)P(\boldsymbol{\theta}|\mathbf{y}_1)$$

$$P(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2) \propto P(\mathbf{y}_2|\boldsymbol{\theta}, \mathbf{y}_1)P(\mathbf{y}_1|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

The Bayesian paradigm…from Laplace!

This motivates the idea that all "priors" (which become posteriors for the next iteration) should have the same form…

$$P(\boldsymbol{\theta})$$

$$P(\boldsymbol{\theta}|\mathbf{y}_1)$$

$$P(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2)$$

$$P(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)$$

The Bayesian paradigm…from Laplace!

"26. La probabilité de la plupart des événements simples est inconnue: en la considérant *a priori*, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais, si l'on a observé un résultat composé de plusieurs de ces événements, la manière dont ils y entrent rend quelques-unes de ces valeurs plus probables que les autres. Ainsi, à mesure que le résultat observé se compose par le développement des événements simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans des limites qui, se resserrant sans cesse, finiraient par coïncider, si le nombre des événements simples devenait infini.

$P(\boldsymbol{\theta})$

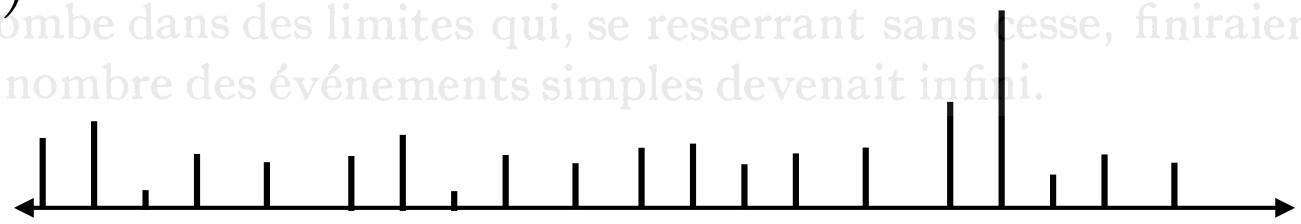
Model parameters, θ

Laplace Book II, Chapter VI, Page 370

The Bayesian paradigm…from Laplace!

"26. La probabilité de la plupart des événements simples est inconnue: en la considérant *a priori*, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais, si l'on a observé un résultat composé de plusieurs de ces événements, la manière dont ils y entrent rend quelques-unes de ces valeurs plus probables que les autres. Ainsi, à mesure que le résultat observé se compose par le développement des événements simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans des limites qui, se resserrant sans cesse, finiraient par coïncider, si le nombre des événements simples devenait infini.

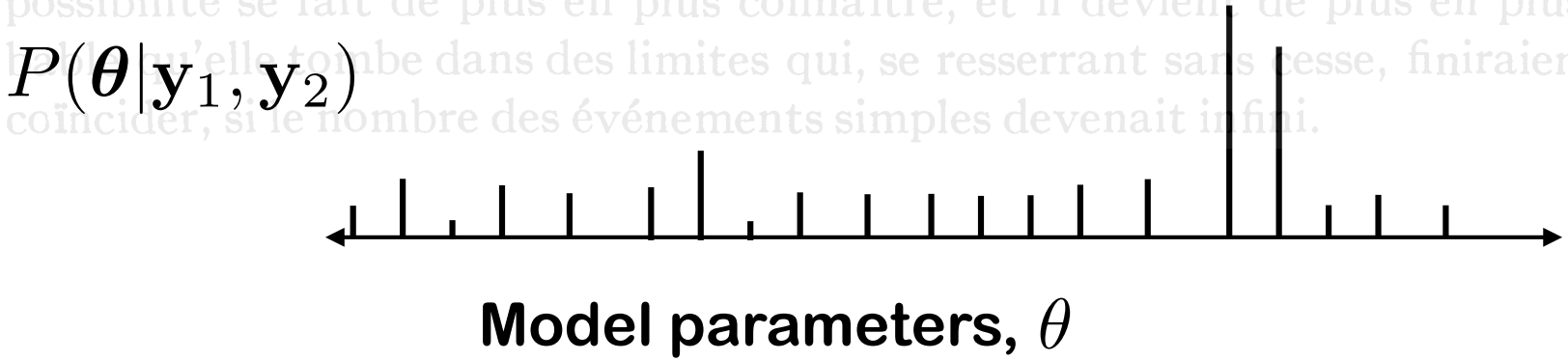
$P(\boldsymbol{\theta}|\mathbf{y}_1)$



Model parameters, θ

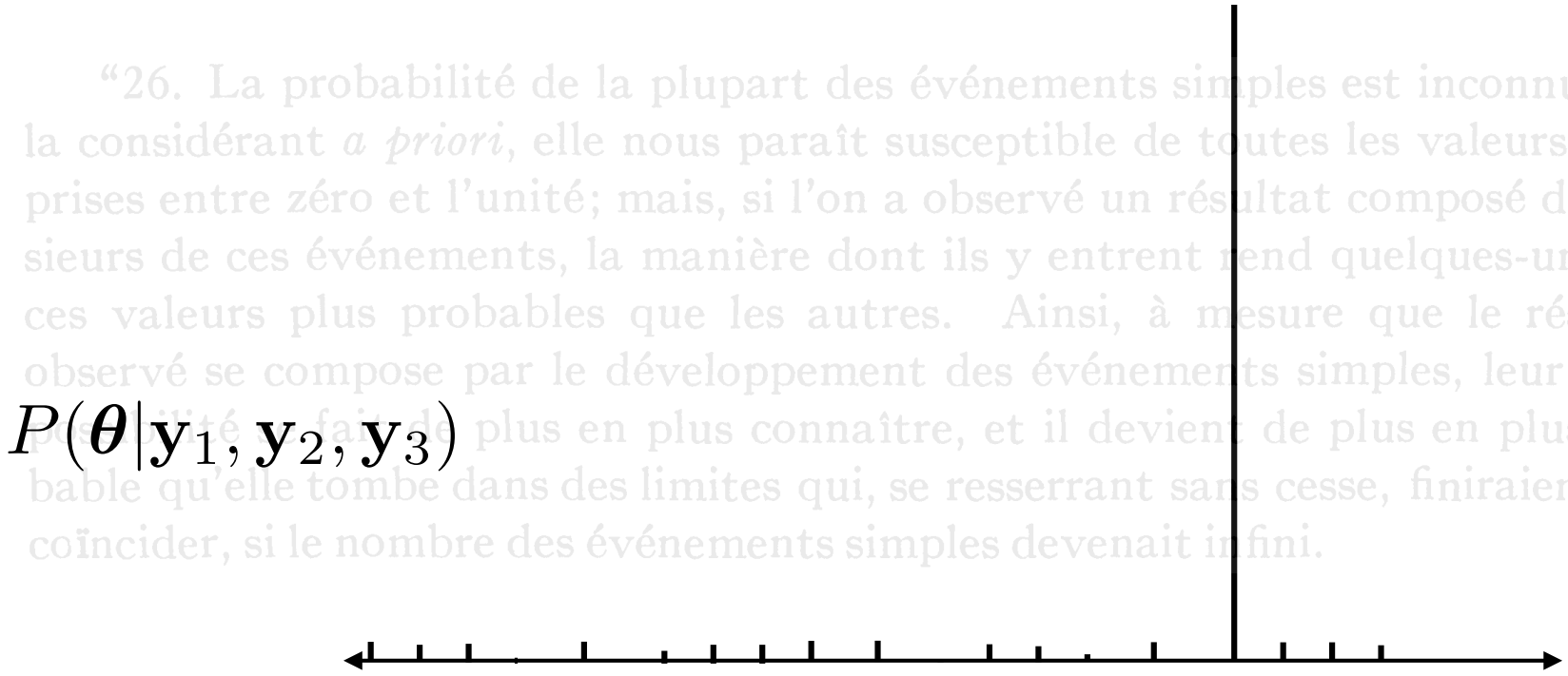
Laplace Book II, Chapter VI, Page 370

The Bayesian paradigm…from Laplace!



Laplace Book II, Chapter VI, Page 370

The Bayesian paradigm…from Laplace!



Laplace Book II, Chapter VI, Page 370

The Bayesian paradigm

Thus — once the physics that dictates the likelihood is set — it is convenient to select a prior that is conjugate to the likelihood. Meaning, we seek a prior that, once multiplied by the likelihood, yields a posterior of the same mathematical form as the prior. This simplifies computation considerably.

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \quad \propto \quad P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

The Bayesian paradigm

Thus — once the physics that dictates the likelihood is set — it is convenient to select a prior that is conjugate to the likelihood. Meaning, we seek a prior that, once multiplied by the likelihood, yields a posterior of the same mathematical form as the prior. This simplifies computation considerably.

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \propto P(\mathbf{y}_{1:N}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

e.g.

Likelihood	Model parameters	Conjugate prior distribution
Bernoulli	p (probability)	Beta
Binomial	p (probability)	Beta
Negative binomial with known failure number, r	p (probability)	Beta
Poisson	λ (rate)	Gamma
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet

Likelihood	Model parameters	Conjugate prior distribution
Normal with known variance σ^2	μ (mean)	Normal
Normal with known precision τ	μ (mean)	Normal
Normal with known mean μ	σ^2 (variance)	Inverse gamma
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared
Normal with known mean μ	τ (precision)	Gamma

https://en.wikipedia.org/wiki/Conjugate_prior

N.B. Normalizations

$$\int d\boldsymbol{\theta} \quad P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \quad = 1$$

$$\int d\boldsymbol{\theta} \quad P(\boldsymbol{\theta}) \quad = 1$$

$$\int d\mathbf{y}_{1:N} \quad P(\mathbf{y}_{1:N}|\boldsymbol{\theta}) \quad = 1$$

Setting up the problem within the Bayesian paradigm

Imagine coin flip experiment w HTHHHTH and we want to determine the probability of heads and tails

The N outcomes of this experiment are random variables.

$$\mathbf{y}_{1:N} = \{y_1, y_2, \cdots y_N\} = \{H, T, H, H, H, \cdots\}$$

In other words y_1 **is heads with probability** p
is tails with probability $1 - p$

The likelihood is a Bernoulli distribution. So our conjugate prior will be Beta distribution.

Setting up the problem within the Bayesian paradigm

To determine the posterior probability of heads/tails, we ask

What is the likelihood of having observed the sequence of outcomes HTHHHTH?

likelihood = $p^5(1-p)^2$

We set our prior

prior = $\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}$

Setting up the problem within the Bayesian paradigm

To determine the posterior probability of heads/tails, we ask

What is the likelihood of having observed the sequence of outcomes HTHHHTH?

likelihood = $p^5(1-p)^2$

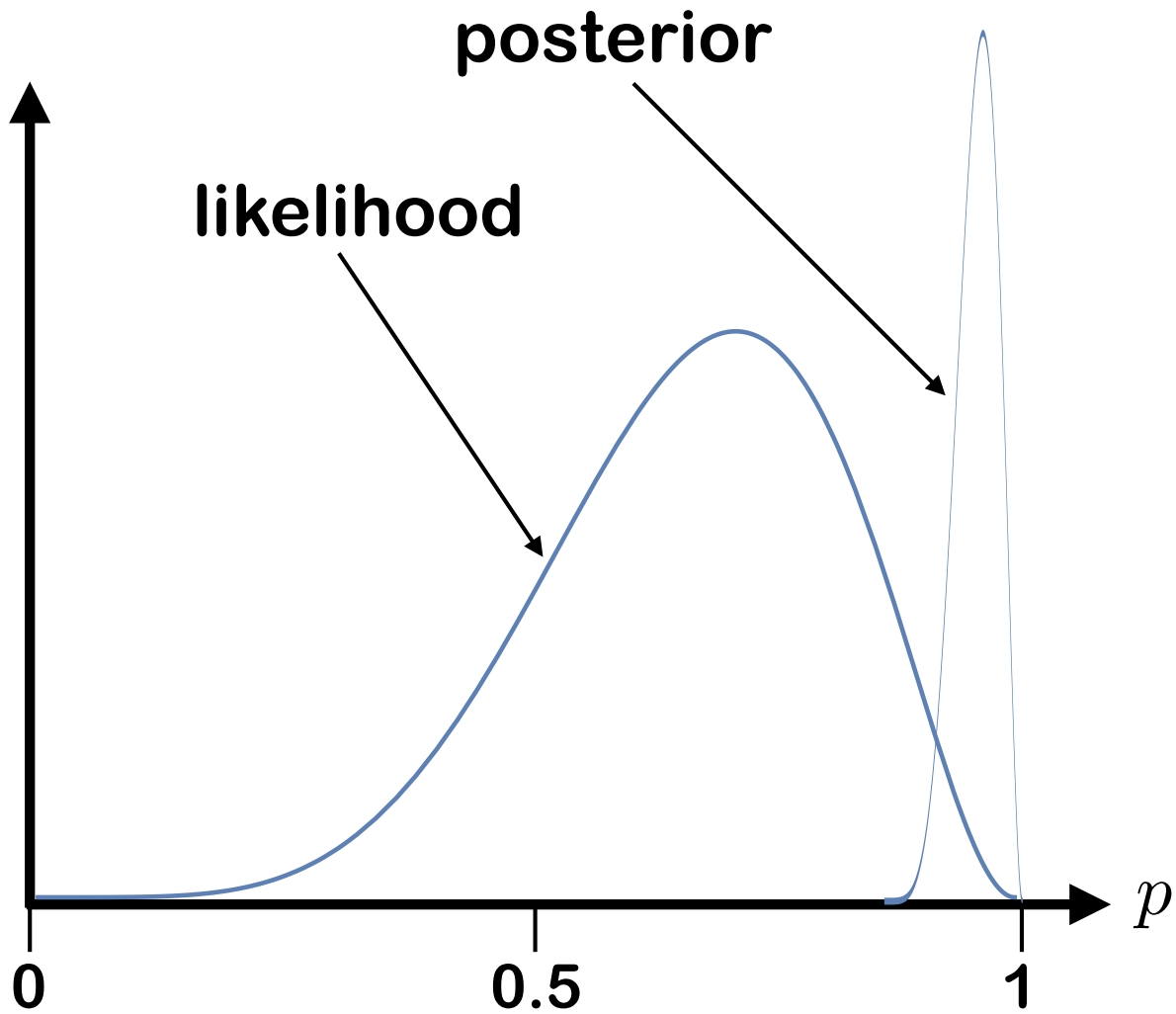
hyperparameters

We set our prior

prior = $\frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

Setting up the problem within the Bayesian paradigm

$$\text{posterior} \propto p^5(1-p)^2 \times p^{\alpha-1}(1-p)^{\beta-1}$$



Setting up the problem within the Bayesian paradigm

posterior $\propto \quad p^5(1-p)^2 \times p^{\alpha-1}(1-p)^{\beta-1}$

$\frac{\partial}{\partial p}$ **likelihood = 0** $\rightarrow \qquad p = \frac{5}{5+2} = \frac{5}{7}$

$\frac{\partial}{\partial p}$ **posterior = 0** $\rightarrow \qquad p = \frac{5+\alpha-1}{5+2+\alpha-1+\beta-1} = \frac{5+(\alpha-1)}{5+\alpha+\beta}$

Setting up the problem within the Bayesian paradigm

posterior $\propto \quad p^5(1-p)^2 \times p^{\alpha-1}(1-p)^{\beta-1}$

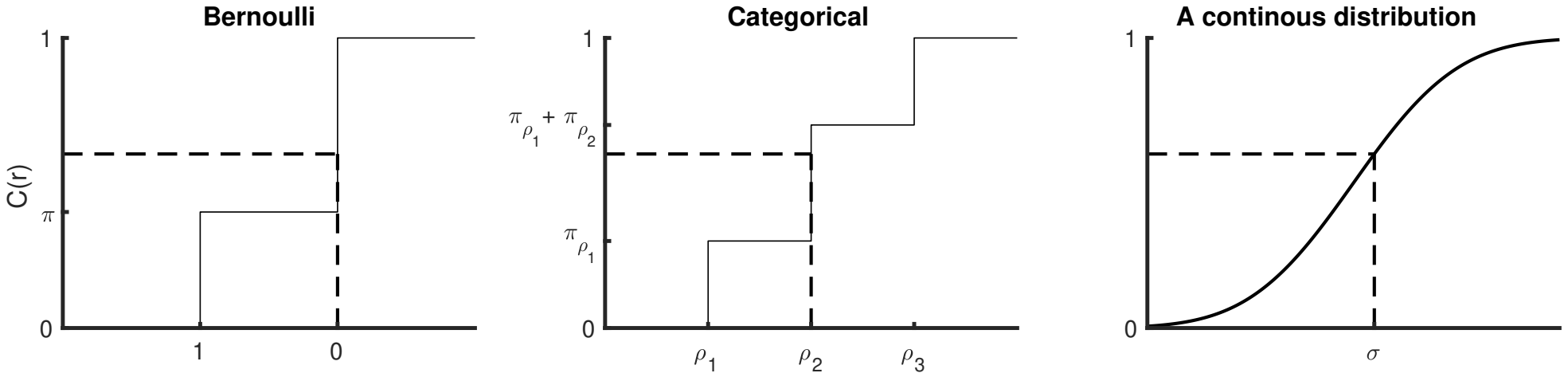
$\frac{\partial}{\partial p}$ **likelihood = 0** $\rightarrow \qquad p = \frac{5}{5+2} = \frac{5}{7}$

$\frac{\partial}{\partial p}$ **posterior = 0** $\rightarrow \qquad p = \frac{5+\alpha-1}{5+2+\alpha-1+\beta-1} = \frac{5+(\alpha-1)}{5+\alpha+\beta}$

You can do much more than maximize a posterior.
You can obtain a full distribution over all unknowns where the error is rigorously propagated from your emission distribution that contains all features of the measurement model.

Exact Sampling

Most simple functions can be sampled from directly using the inverse cdf method



Exact Sampling

Most simple functions can be sampled from directly using the inverse cdf method

$$\theta \sim P(\theta)$$

e.g. $P(\theta) = b^{-1} e^{-\theta/b}$

$$cdf(A) = \int_0^A d\theta P(\theta)$$

Exact Sampling

Most simple functions can be sampled from directly using the inverse cdf method

$$\theta \sim P(\theta)$$

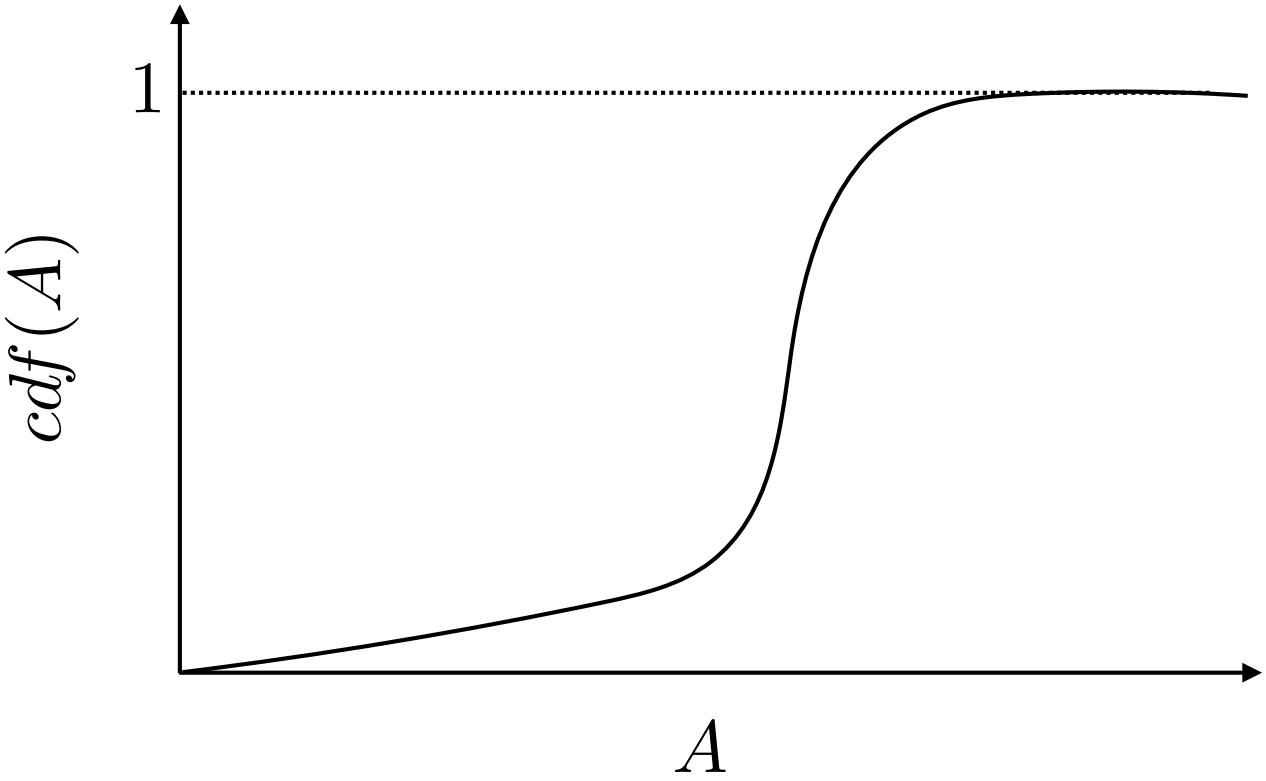
e.g. $P(\theta) = b^{-1} e^{-\theta/b}$

$$cdf(A) = \int_0^A d\theta P(\theta)$$

For the exponential the y starts from 0. However for a Gaussian spanning all real numbers the integral would start from $-\infty$.

Exact Sampling

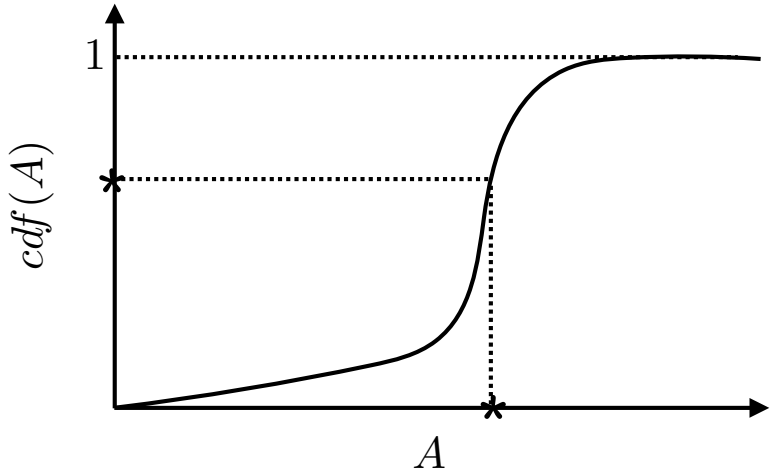
Most simple functions can be sampled from directly using the inverse cdf method



Exact Sampling

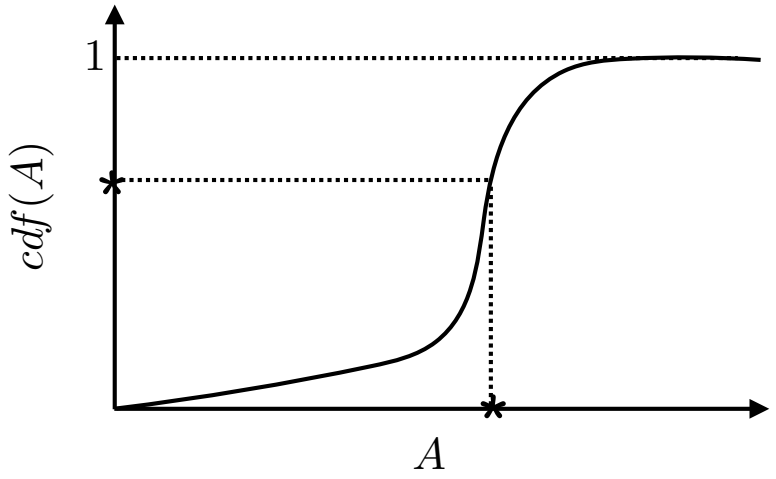
Most simple functions can be sampled from directly using the inverse cdf method

Step 1) Sample random #, x, uniformly from 0 to 1



Step 2) Find the A to which this corresponds

Exact Sampling



Step 2) Find the A to which this corresponds

$$P(\theta) = b^{-1} e^{-\theta/b}$$

$$cdf(A) = 1 - e^{-A/b}$$

Exact Sampling

Step 2) Find the A to which this corresponds

$$P(\theta) = b^{-1} e^{-\theta/b}$$

$$cdf(A) = 1 - e^{-A/b}$$

$$x = 1 - e^{-A/b}$$

$$\boxed{A = b \ln \frac{1}{1-x}}$$

Exact Sampling

Step 2) Find the A to which this corresponds

$$A = b \ln \frac{1}{1-x}$$

You have now converted a uniform random number (x), which we know how to sample on a computer, into an exponential random variable, A.

Approximate Sampling

Goal: sample from a target distribution, $\pi\left(r\right)$, whose cdf cannot be computed

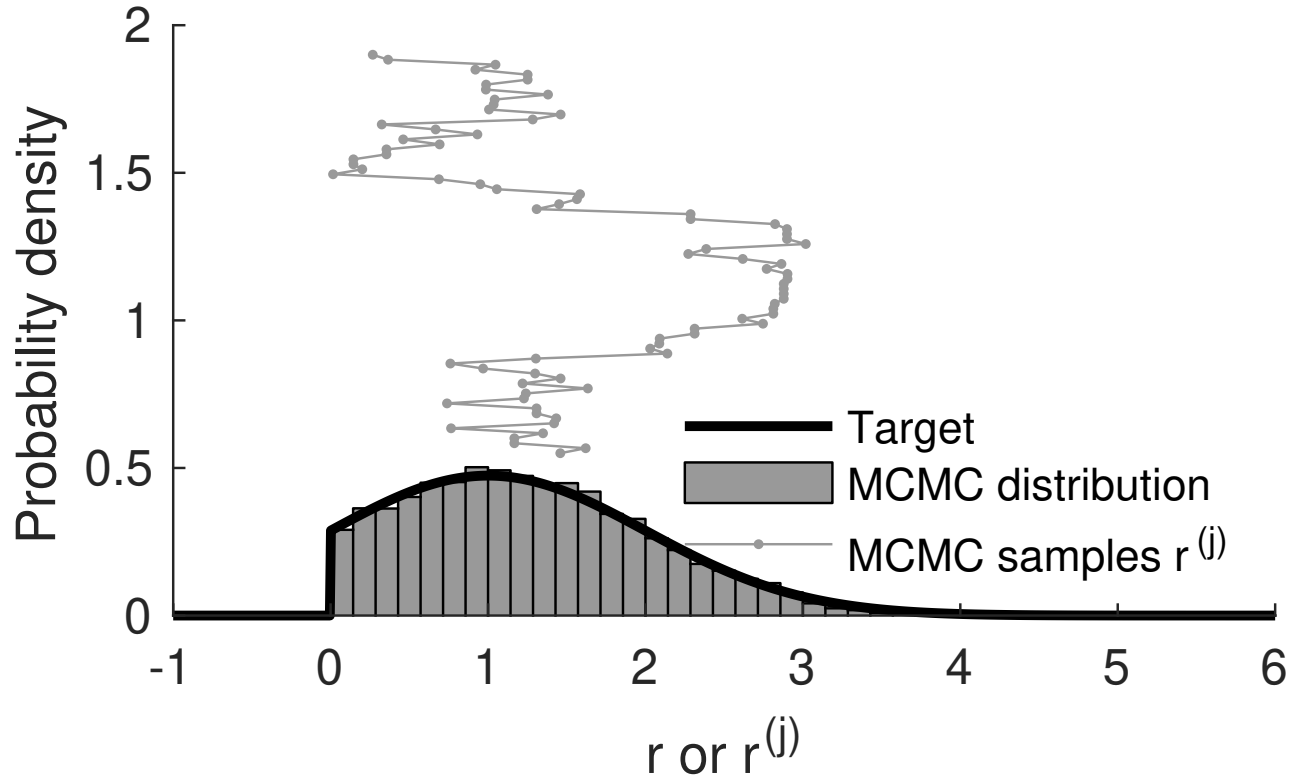
As a result, we generate a Markov chain of samples using Markov Chain Monte Carlo (MCMC)

$$r^{(0)} \rightarrow r^{(1)} \rightarrow r^{(2)} \rightarrow \ldots \rightarrow r^{(j)} \rightarrow r^{(j+1)} \rightarrow \ldots \rightarrow r^{(J)}$$

Approximate Sampling

Markov chain

$$r^{(0)} \rightarrow r^{(1)} \rightarrow r^{(2)} \rightarrow \ldots \rightarrow r^{(j)} \rightarrow r^{(j+1)} \rightarrow \ldots \rightarrow r^{(J)}$$



How we do Metropolis-Hastings (a type of MCMC)

Consider a proposal distribution, Q

$$Q_{r^{\text{old}}}\left(r^{\text{prop}}\right) = p\left(r^{\text{prop}} \middle| r^{\text{old}}\right)$$

Conditions on Q:

- the simulation of random variables $r^{\text{prop}}|r^{\text{old}} \sim \mathbb{Q}_{r^{\text{old}}}$ is possible,
- the simulation of random variables $r^{\text{prop}}|r^{\text{old}} \sim \mathbb{Q}_{r^{\text{old}}}$ allows the generation of any feasible value.

Write down the acceptance ratio:

$$A_{r^{\text{old}}}\left(r^{\text{prop}}\right) = \underbrace{\frac{\bar{\pi}\left(r^{\text{prop}}\right)}{\bar{\pi}\left(r^{\text{old}}\right)}}_{\text{target}} \underbrace{\frac{Q_{r^{\text{prop}}}\left(r^{\text{old}}\right)}{Q_{r^{\text{old}}}\left(r^{\text{prop}}\right)}}_{\text{proposal}}$$

How we do Metropolis-Hastings (a type of MCMC)

Write down the acceptance ratio:

$$A_{r^{\text{old}}}\left(r^{\text{prop}}\right) = \underbrace{\frac{\bar{\pi}\left(r^{\text{prop}}\right)}{\bar{\pi}\left(r^{\text{old}}\right)}}_{\text{target}} \underbrace{\frac{Q_{r^{\text{prop}}}\left(r^{\text{old}}\right)}{Q_{r^{\text{old}}}\left(r^{\text{prop}}\right)}}_{\text{proposal}}$$

Algorithm 5.1: Metropolis-Hastings sampler for arbitrary targets

Given a target $\bar{\pi}(r)$, a proposal $Q_{r^{\text{old}}}\left(r^{\text{prop}}\right)$, and a feasible initial sample $r^{(0)}$, the Metropolis-Hastings sampler proceeds as follows:
For each j from 1 to J:

- Generate a proposal $r^{\text{prop}} \sim \mathbb{Q}_{r^{(j-1)}}$.
- Compute the acceptance ratio $A_{r^{(j-1)}}\left(r^{\text{prop}}\right)$.
- Generate $u \sim \text{Uniform}_{[0,1]}$.
- If $u < A_{r^{(j-1)}}\left(r^{\text{prop}}\right)$; set $r^{(j)} = r^{\text{prop}}$, else set $r^{(j)} = r^{(j-1)}$.

Example 5.2: Two Metropolis-Hastings schemes for the truncated Normal distribution

Consider a random variable R distributed according to a Normal distribution with mean μ and variance σ^2 *truncated* below 0. That is, R has a probability density given by

$$\pi(r) \propto \bar{\pi}(r) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right), & r \geq 0 \\ 0, & r < 0 \end{cases}.$$

Example 5.2: Two Metropolis-Hastings schemes for the truncated Normal distribution

Consider a random variable R distributed according to a Normal distribution with mean μ and variance σ^2 *truncated* below 0. That is, R has a probability density given by

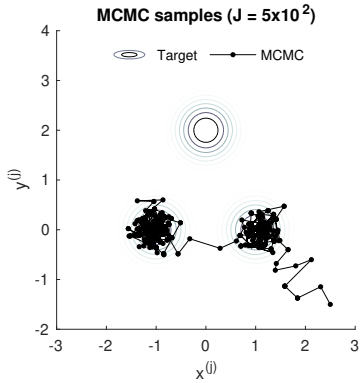
$$\pi(r) \propto \bar{\pi}(r) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right), & r \geq 0 \\ 0, & r < 0 \end{cases}.$$

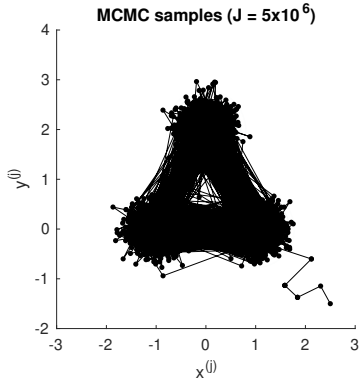
$$\mathbb{Q}_{r^{\text{old}}} = \text{Normal}(r^{\text{old}}, \lambda^2)$$

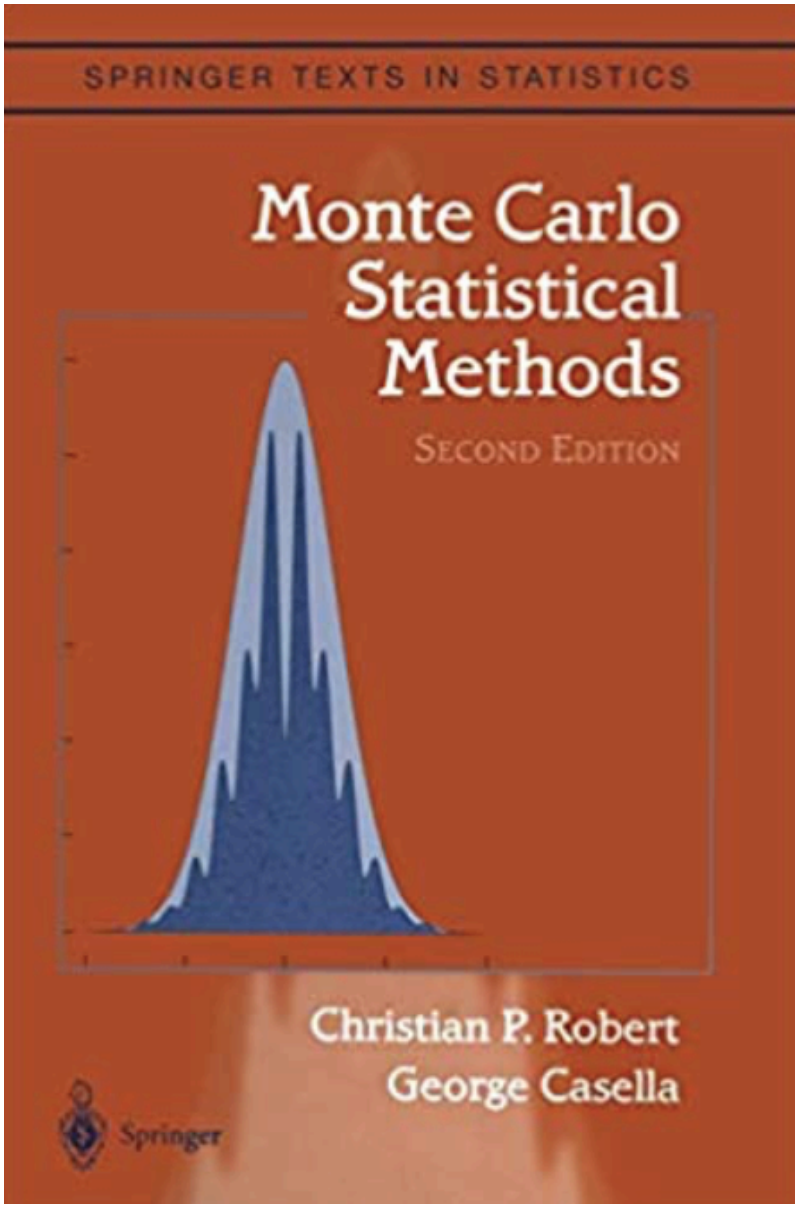
$$A_{r^{\text{old}}}(r^{\text{prop}}) = \begin{cases} \exp\left(\frac{(r^{\text{old}}-\mu)^2-(r^{\text{prop}}-\mu)^2}{2\sigma^2}\right), & r^{\text{prop}} \geq 0 \\ 0, & r^{\text{prop}} < 0 \end{cases}.$$

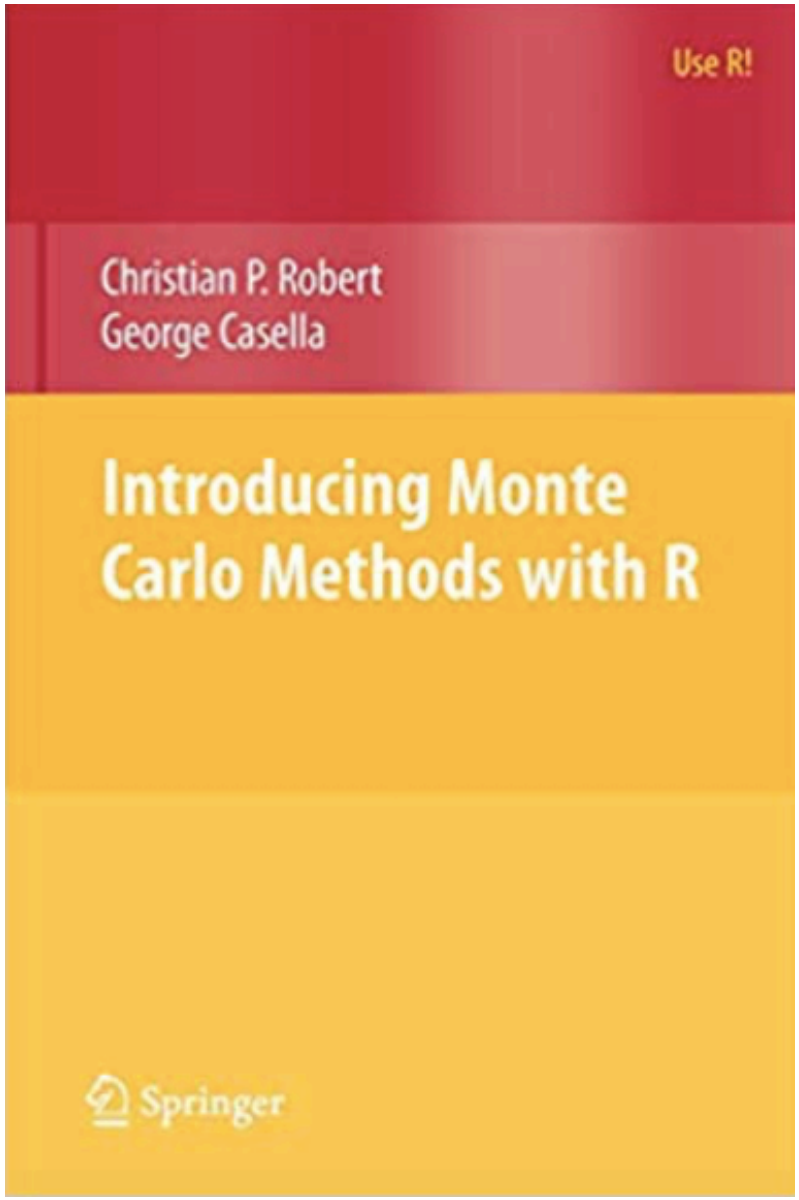
Example 5.4: Choice of proposals in MCMC

$$\pi = 0.3\,\text{Normal}_2\left(\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}0.25^2 & 0\\0 & 0.25^2\end{bmatrix}\right)$$
$$+ 0.3\,\text{Normal}_2\left(\begin{bmatrix}+1\\0\end{bmatrix}, \begin{bmatrix}0.25^2 & 0\\0 & 0.25^2\end{bmatrix}\right)$$
$$+ 0.4\,\text{Normal}_2\left(\begin{bmatrix}0\\+1\end{bmatrix}, \begin{bmatrix}0.25^2 & 0\\0 & 0.25^2\end{bmatrix}\right)$$






SPRINGER TEXTS IN STATISTICS
Monte Carlo Statistical Methods
SECOND EDITION
Christian P. Robert
George Casella
Springer


Use R!
Christian P. Robert
George Casella
Introducing Monte Carlo Methods with R
Springer

Setting up the problem within the Bayesian paradigm

Step 1) Write down the model.

Step 2) Write down the likelihood of your data under the assumption of your model. Pick a prior which is either conjugate or otherwise informed by some physics.

Step 3) Compute your posterior to find the probability of your model parameters

Steps 1-3 are collectively called Bayesian model learning/training

Setting up the problem within the Bayesian paradigm

Step 3) Compute your posterior to find the probability of your model parameters

$$P(\boldsymbol{\theta}|\mathbf{y}_{1:N}) = P(\theta_1, \cdots, \theta_K|\mathbf{y}_{1:N})$$

Thanks!

Data Modeling in the Sciences

Applications, Basics, Computations

Steve Pressé and Ioannis Sgouralis