

# Introduction to Bayesian inference

Mikkel N. Schmidt  
Technical University of Denmark

DTU Informatics  
Department of Informatics and Mathematical Modeling

---

A collage of mathematical symbols and formulas. The most prominent is the Taylor series expansion:  $f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$ . Other symbols include the integral sign  $\int_a^b$ , Greek letters  $\epsilon$ ,  $\theta$ ,  $\Omega$ ,  $\delta$ ,  $\chi^2$ ,  $\Sigma$ ,  $\infty$ ,  $\sqrt{17}$ ,  $e^{i\pi}$ , and the number  $\{2.7182818284\}$ .

# Sources and references

Parts of these slides are adapted from

- Slides
  - Zoubin Ghahramani and Carl Edward Rasmussen, Slides for machine learning course at Cambridge University.
  - Iain Murray, Tutorial on “Markov chain Monte Carlo.”
  - Jan Larsen, Tutorial on “Basics of Bayesian learning.”
  - Chris Bishop, Tutorial on “Introduction to Bayesian inference.”
- Books
  - Chris Bishop, “Pattern Recognition and Machine Learning.”
  - Andrew Gelman et al., “Bayesian Data Analysis.”
  - Christian Robert, “The Bayesian Choice.”
  - Tom Mitchell, “Machine Learning.”



# Agenda

- Bayesian modeling
- Example: Ordinary linear regression
- Graphical models
- Markov chain Monte Carlo
- Variational inference

**Please ask questions**



# Bayesian modeling

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$$\int_a^b \epsilon \Theta^{\sqrt{17}} + \Omega \int \delta e^{i\pi} =$$

$$= \{2.7182818284\}$$

$\infty$   $\chi^2$   $\Sigma$   $\gg$   $!$

# Bayesian modeling

## Learning objectives

Understand

- Basic probability theory
  - Joint, conditional, and marginal distributions
- The basics of Bayesian modeling
  - Likelihood, priors, and posterior inference
- Bayesian estimation and decision making
- The process of modeling, inference, and evaluation

# Bayesian modeling

## Machine learning

### Definition

- A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at **tasks** in **T**, as measured by **P**, improves with **experience E**.

(Tom Mitchell, "Machine Learning")

### Disciplines

- Artificial intelligence
- Bayesian methods
- Computational complexity theory
- Control theory
- Information theory
- Philosophy
- Psychology and neurobiology
- Statistics

# Bayesian modeling

## Machine learning

- 1st generation: Domain specific ***expert knowledge*** systems
  - Artificial intelligence
  - Hand crafted rules
  - Expert systems
  - Knowledge-based AI
- 2nd generation: Black-box ***statistical learning*** systems
  - Artificial neural networks
  - Support vector machines
- 3rd generation: Integration of ***expert knowledge*** and ***statistical learning***.
  - Bayesian framework
  - Probabilistic graphical models
  - Efficient inference procedures
- 4th generation: ?

# Bayesian modeling

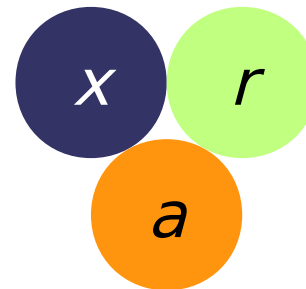
## Advantages of Bayesian modeling

- Principled framework that combines all available uncertain knowledge
  - Data from multiple sources
  - Prior information / domain knowledge
- All assumptions are made explicit
  - Once the model has been specified, inference is “automatic”
- Model selection is integrated in the framework
- Classical learning schemes are special case
- Known to give better performance in many cases
- Gives predictions with error-bars (credible intervals)

# Bayesian modeling

## Three types of learning

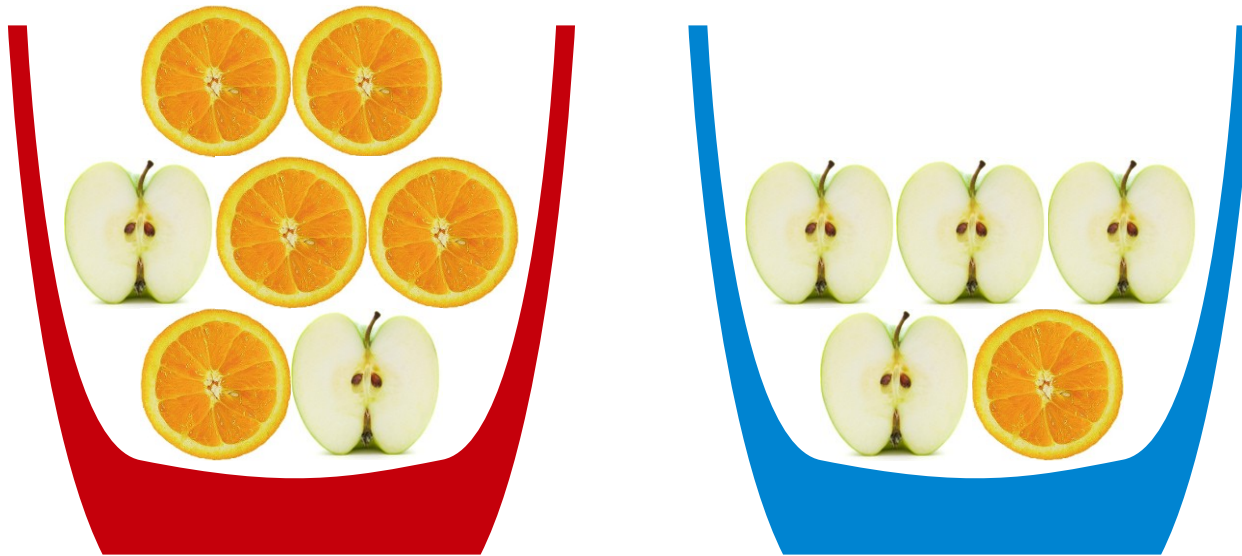
- Unsupervised learning
  - Given inputs,  $x$ , build a model of  $x$
- Supervised learning
  - Given inputs,  $x$ , and outputs,  $y$ , model the relation between  $x$  and  $y$
- Reinforcement learning
  - Given inputs,  $x$ , and rewards,  $r$ , produce actions,  $a$ , to maximize the reward



# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

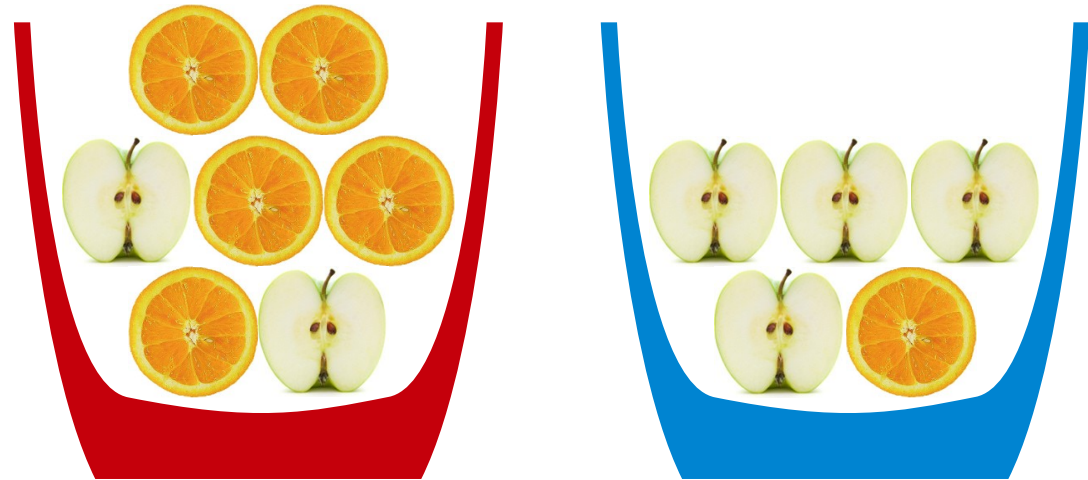


# Bayesian modeling

## Probability theory

Basic rules of probability





- Sum rule 
$$p(x) = \sum_y p(x, y)$$
- Product rule 
$$p(x, y) = p(x|y)p(y)$$
- Bayes' rule 
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

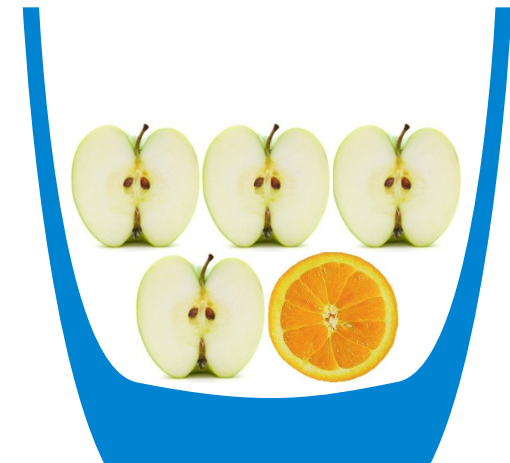
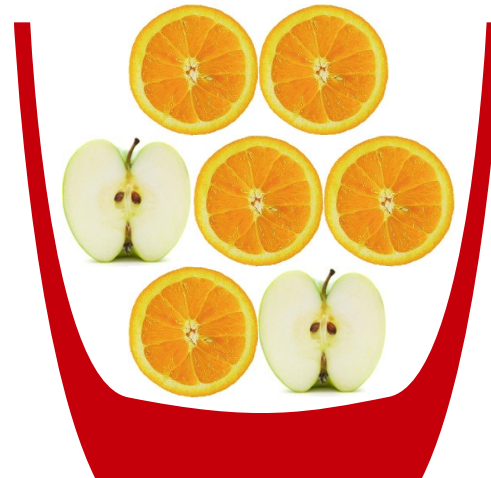


# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?





		
	2	5
	4	1

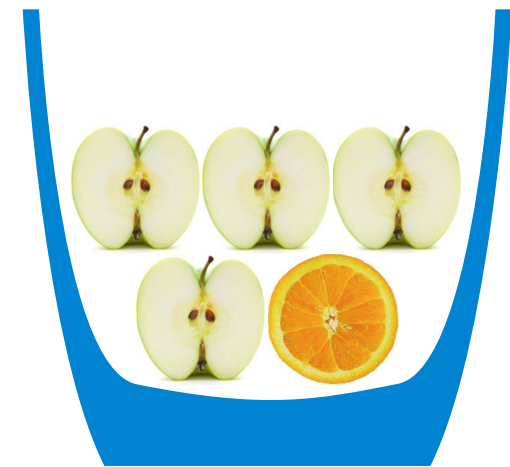
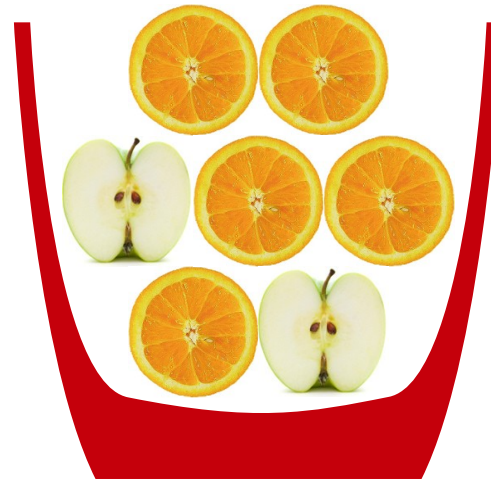


# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?





			
	2	5	7
	4	1	5
	6	6	12



# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?





			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r, o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12





$$p(o|r) = \frac{p(r, o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r, o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r, o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$





$$p(r|o) = \frac{p(r, o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

# Bayesian modeling

## Probability theory

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r, o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r, o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$\begin{aligned}
 &= \frac{p(o|r)p(r)}{p(o)} \\
 &= \frac{5/7 \times 7/12}{6/12} = 5/6
 \end{aligned}$$

# Bayesian modeling

## Basics of Bayesian modeling

- **Model the joint distribution of everything.** Data and parameters are modeled as random with associated probability distributions.

$$p(\mathbf{y}, \boldsymbol{\theta})$$

- **The full posterior** is the conditional distribution assigned after taking the observations into account

$$p(\boldsymbol{\theta}|\mathbf{y})$$

- **Estimands** are unobserved quantities for which statistical inference is made. The posterior for an estimand is computed by averaging over variables that are not of interest

$$p(\gamma|\mathbf{y})$$

# Bayesian modeling

## Basics of Bayesian modeling

- **Model the joint distribution of everything.** Data and parameters are modeled as random with associated probability distributions.

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- **The full posterior** is the conditional distribution assigned after taking the observations into account

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

- **Estimands** are unobserved quantities for which statistical inference is made. The posterior for an estimand is computed by averaging over variables that are not of interest

$$p(\gamma|\mathbf{y}) = \int p(\gamma|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

# Bayesian modeling

## Posterior inference

- **The full posterior** is the conditional distribution assigned after taking the observations into account

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \quad \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

Normalized product of likelihood and prior

- Normalization constant is called the **marginal likelihood**

$$p(\mathbf{y}) = \int p(\mathbf{y}, \boldsymbol{\theta})d\boldsymbol{\theta} = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

# Bayesian modeling

## Likelihood

The probability of the data, given the parameters

$$p(\mathbf{y}|\boldsymbol{\theta})$$

- Think of the likelihood as a generative random process.
  - If we know the parameters, how would random data be generated?

The data affect the posterior inference *only* through the likelihood function

# Bayesian modeling

## Priors

The probability distribution of the parameters, *before observing the data*

$$p(\theta)$$

- Think of the prior as capturing your *knowledge and uncertainty* about the model parameters

# Bayesian modeling

## Priors

Approaches to choosing priors

- Informative / non-informative
- Proper / improper
- Subjective (based on knowledge)
- Empirical (based on past experiments)
- Structural priors
  - Independence
  - Exchangability
  - Invoking constraints
- Convenience priors
  - Functional form that makes computations simple
  - Conjugate priors

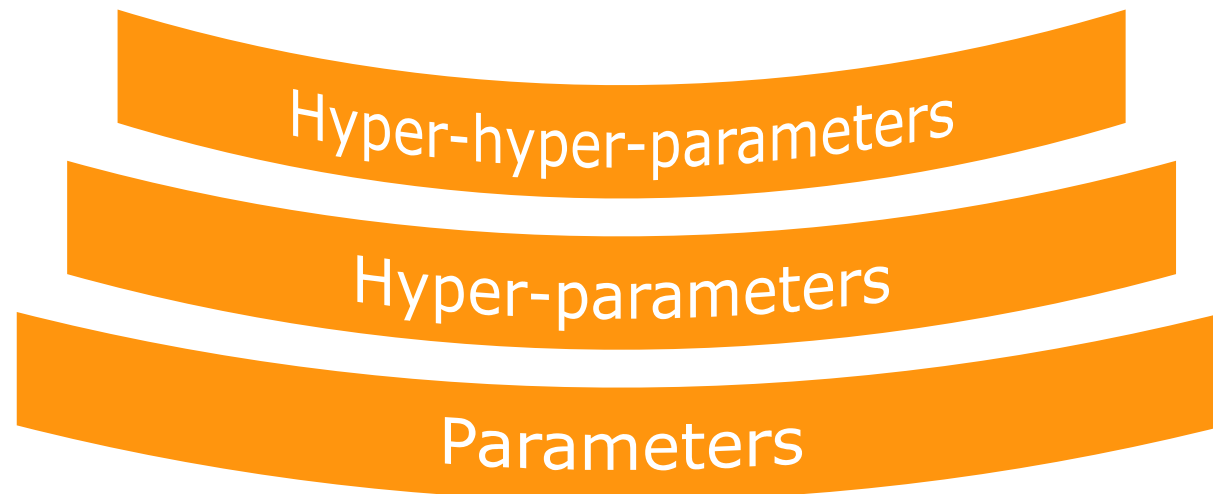
# Bayesian modeling

## Hierarchical models

A more refined hierarchical model can be constructed

- Choose prior distributions that have adjustable (hyper-)parameters
- Treat the hyper-parameters as random variables
- Choose prior distributions for the hyper-parameters

Note: There are no formal distinction between parameters and hyper-parameters (and hyper-hyper-parameters etc.). All are treated equally in Bayesian modeling.




# Bayesian modeling

## Variables

Variables modeled as random

- Observations
  - Data that is measured
- Parameters (and hyper-parameters)
  - Used to define the model
- Hidden variables / latent variables
  - Introduced to facilitate model specification



Principally treated equal  
Given a prior distribution

Variables not modeled as random

- Explanatory variables / covariates / predictors
- Fixed parameters

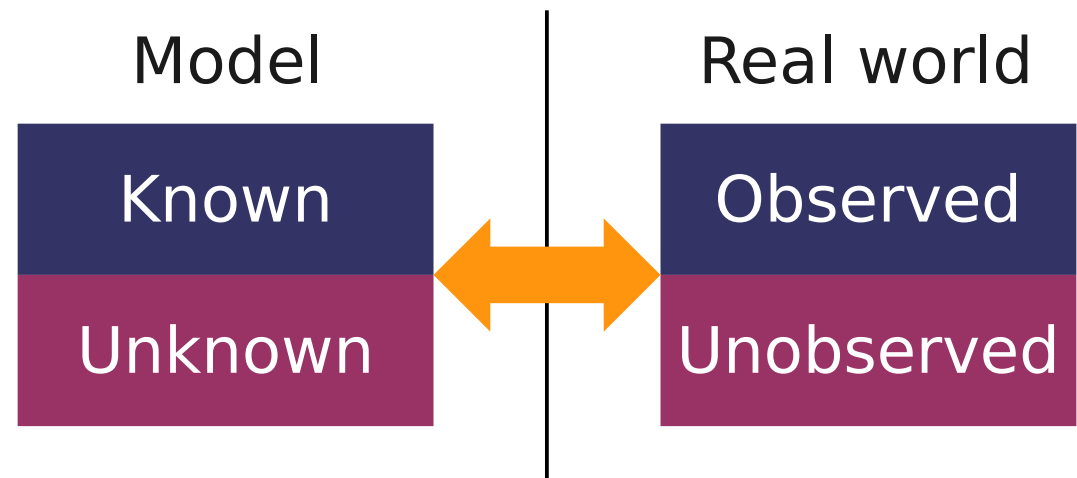
# Bayesian modeling

## Estimands

- **Estimands** are unobserved quantities for which statistical inference is made. The posterior for an estimand is computed by averaging over variables that are not of interest

$$p(\gamma|\mathbf{y}) = \int p(\gamma|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

- Types of estimands
  - Unobservable quantities  
e.g. regression coefficients
  - Potentially observable quantities  
e.g. future observations



# Bayesian modeling

## Predictive distribution

The distribution of a new data point given previously observed data

- The likelihood averaged over the posterior

$$p(y^* | \mathbf{y}) = \int p(y^* | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}$$

# Bayesian modeling

## Estimators

Loss function

$$L(\theta, \hat{\theta})$$

- What would it cost, if the parameters take this value and we estimate that value

Expected loss

$$\rho(\hat{\theta}|\mathbf{x}) = \int L(\theta, \hat{\theta})p(\theta|\mathbf{x})d\theta$$

- Average over the posterior distribution and conditioned on observed data

Bayes estimator

$$\hat{\theta}_{Bayes} = \underset{\hat{\theta}}{\operatorname{argmin}} \rho(\hat{\theta}|\mathbf{x})$$

- Estimator that minimizes the loss function

# Bayesian basics

## Making decisions

Loss function

$$L(\theta, d)$$

- What would it cost, if the parameters take this value and we make that decision

Expected loss

$$\rho(d|\mathbf{x}) = \int L(\theta, d)p(\theta|\mathbf{x})d\theta$$

- Average over the posterior distribution and conditioned on observed data

Bayes estimator

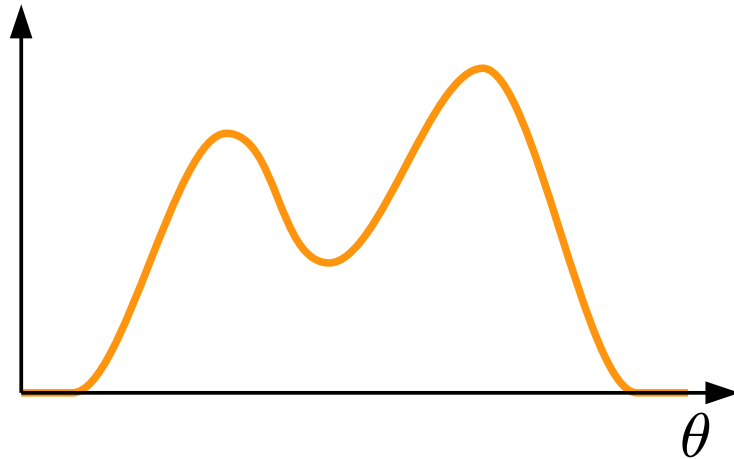
$$d_{Bayes} = \operatorname{argmin}_d \rho(d|\mathbf{x})$$

- Estimator that minimizes the loss function

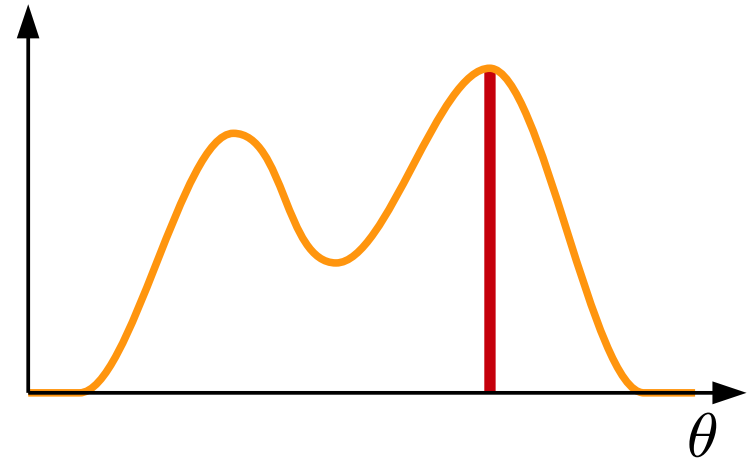
# Bayesian modeling

## (Approximate) inference procedures

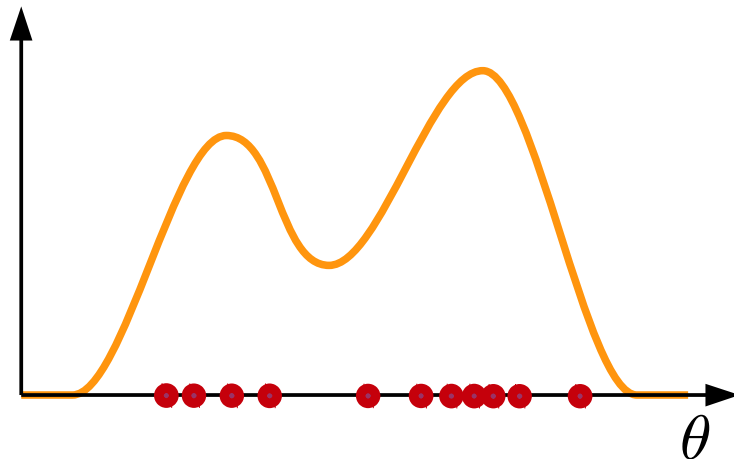
*Exact inference*



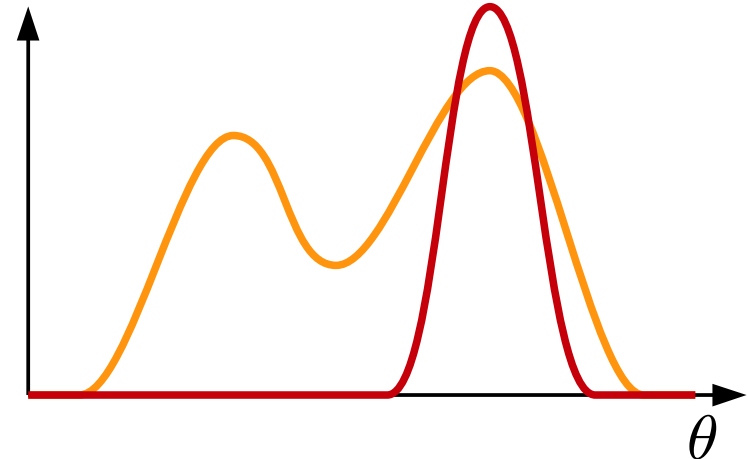
*Maximum a posteriori (MAP)*



*Monte Carlo sampling*

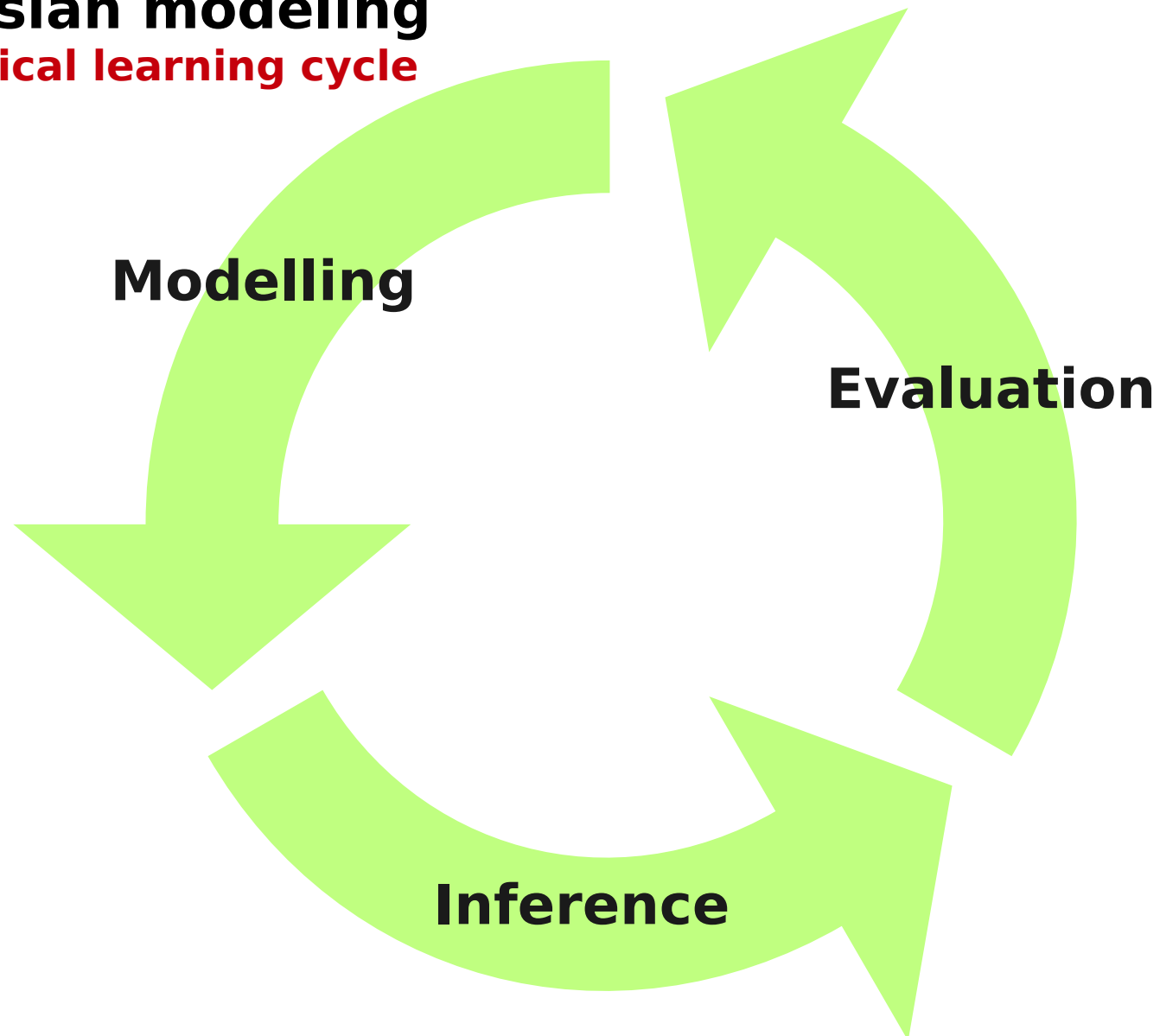


*Variational inference*



# Bayesian modeling

## Statistical learning cycle

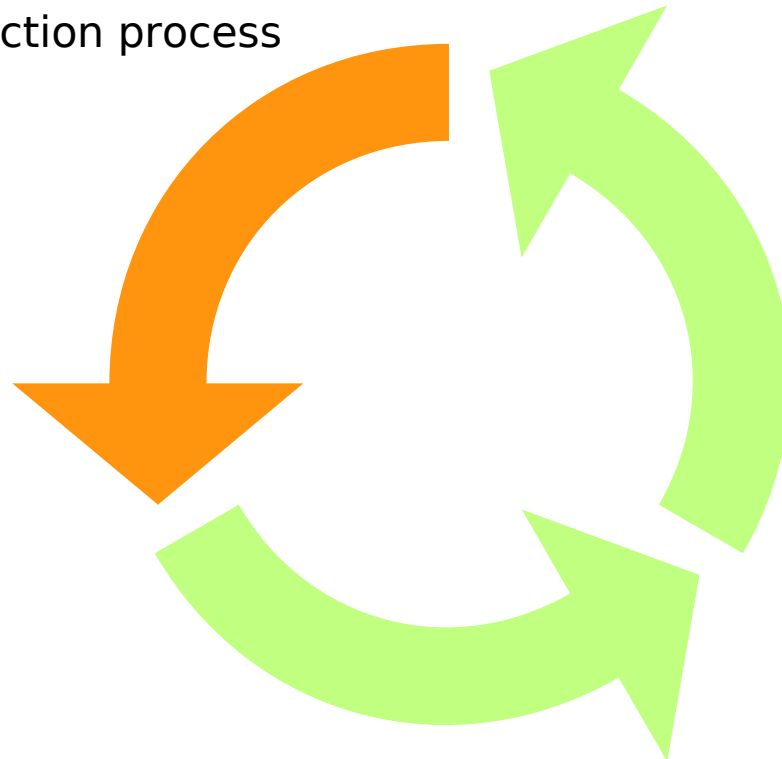


# Bayesian modeling

## Statistical learning cycle

### Modeling

- Joint probability distribution for observable and unobservable quantities
  - data, parameters, hidden variables, etc.
- Consistent with knowledge about
  - the underlying problem and
  - the data collection process

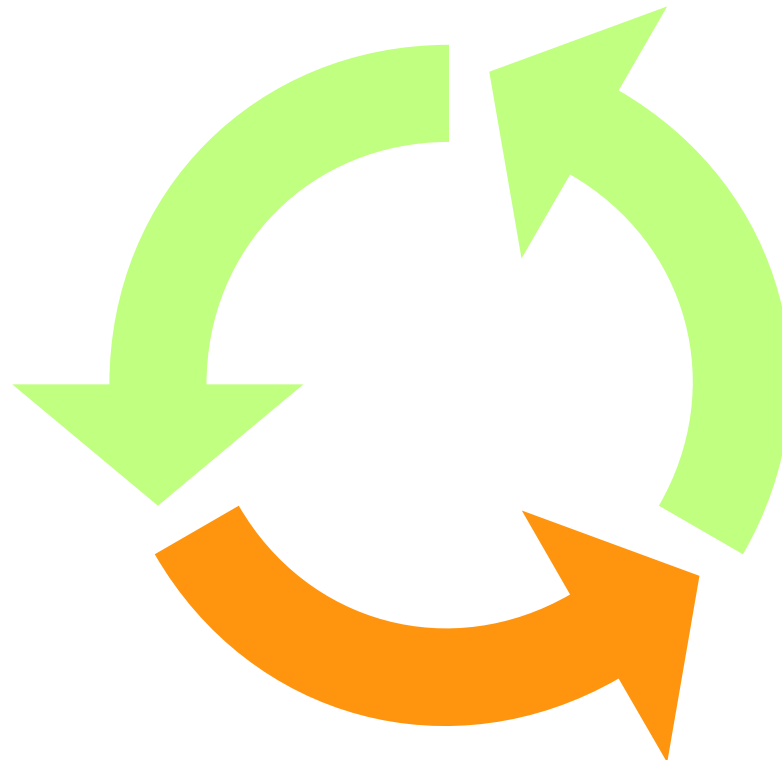


# Bayesian modeling

## Statistical learning cycle

### Inference

- Conditioning on observed data
- Calculating and interpreting the posterior distribution for estimands of interest
  - Involves integrating over the posterior



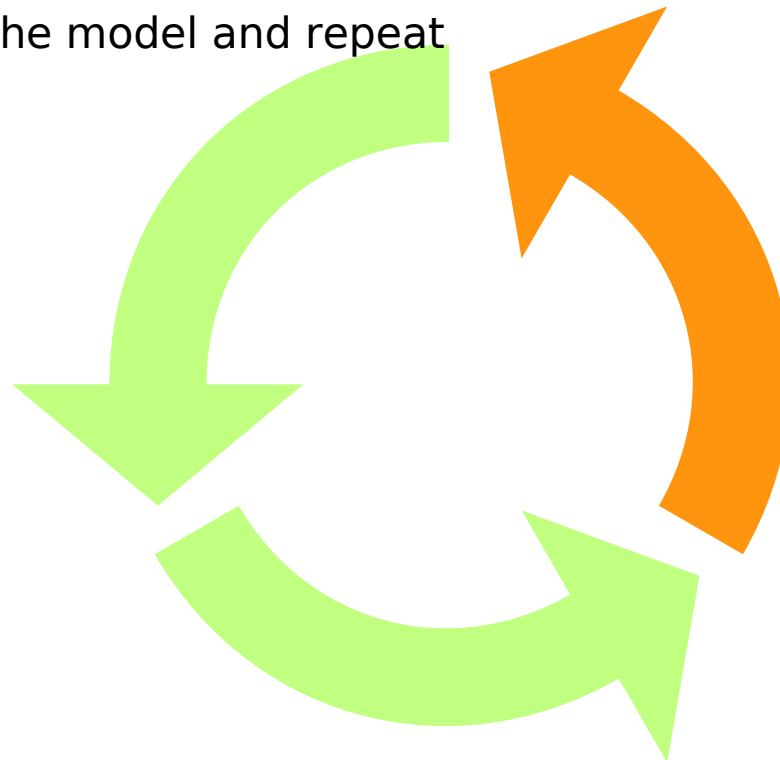
# Bayesian modeling

## Statistical learning cycle

Evaluation

- Does the model fit data?
- Are conclusions reasonable?
- Are results sensitive to model assumptions?

If necessary, alter the model and repeat

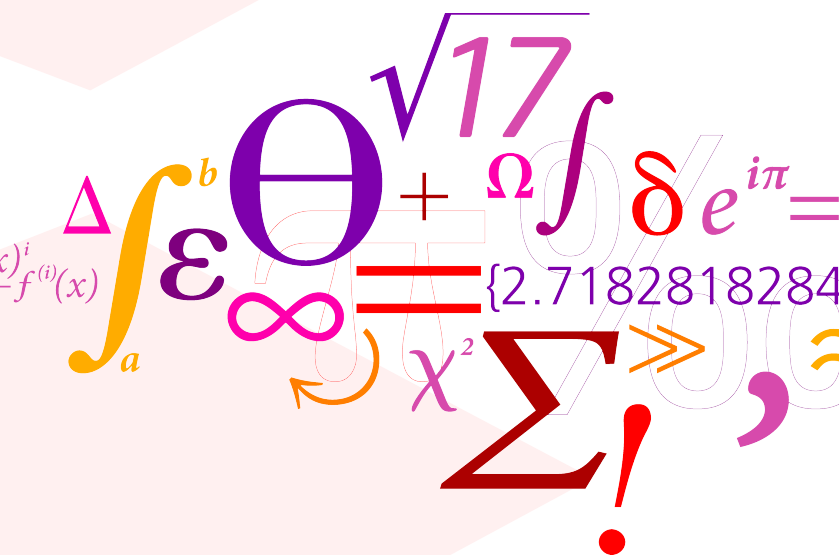


# Bayesian modeling

- Solve Q 1.1 - 1.2



# Ordinary linear regression

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$


# Ordinary linear regression

## Learning objectives

Understand

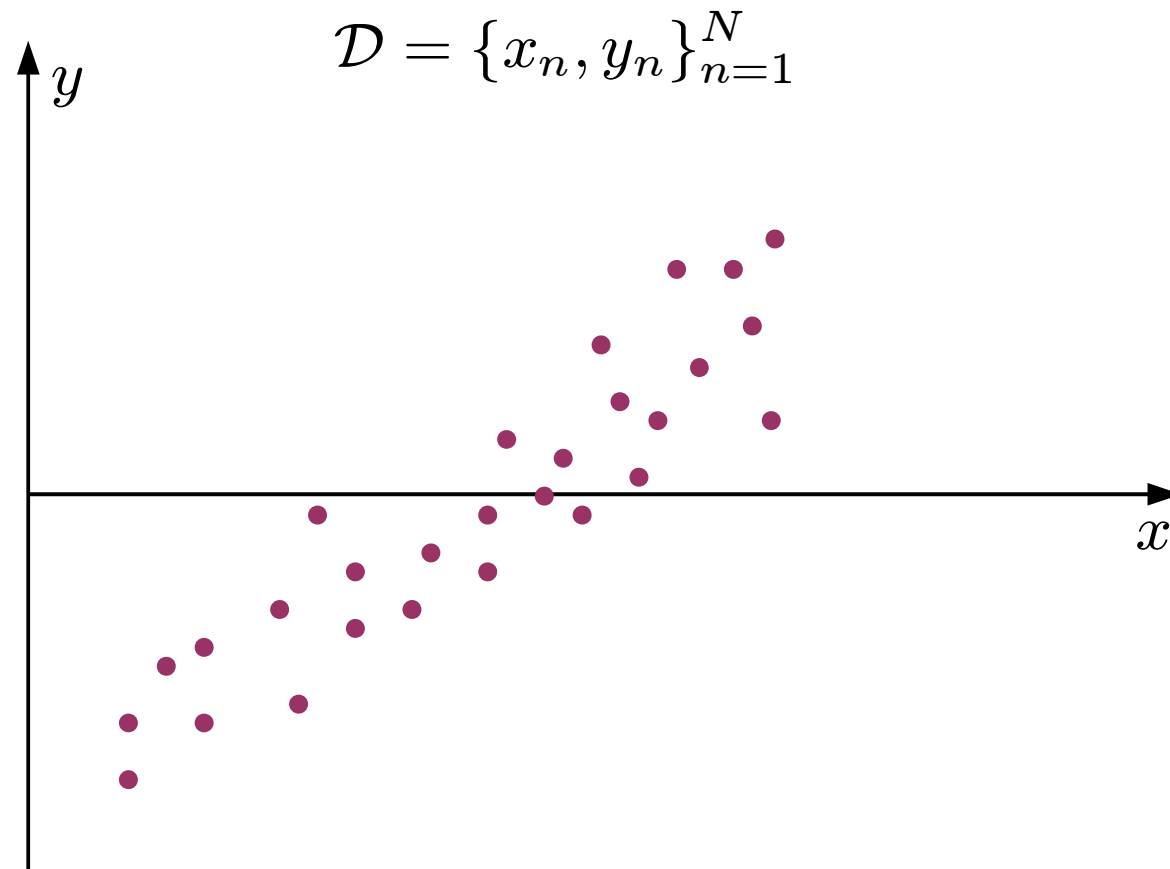
- The ordinary linear regression model
- The relation to central concepts from Bayesian modeling

Gain knowledge about

- How to formulate a Bayesian model in practice

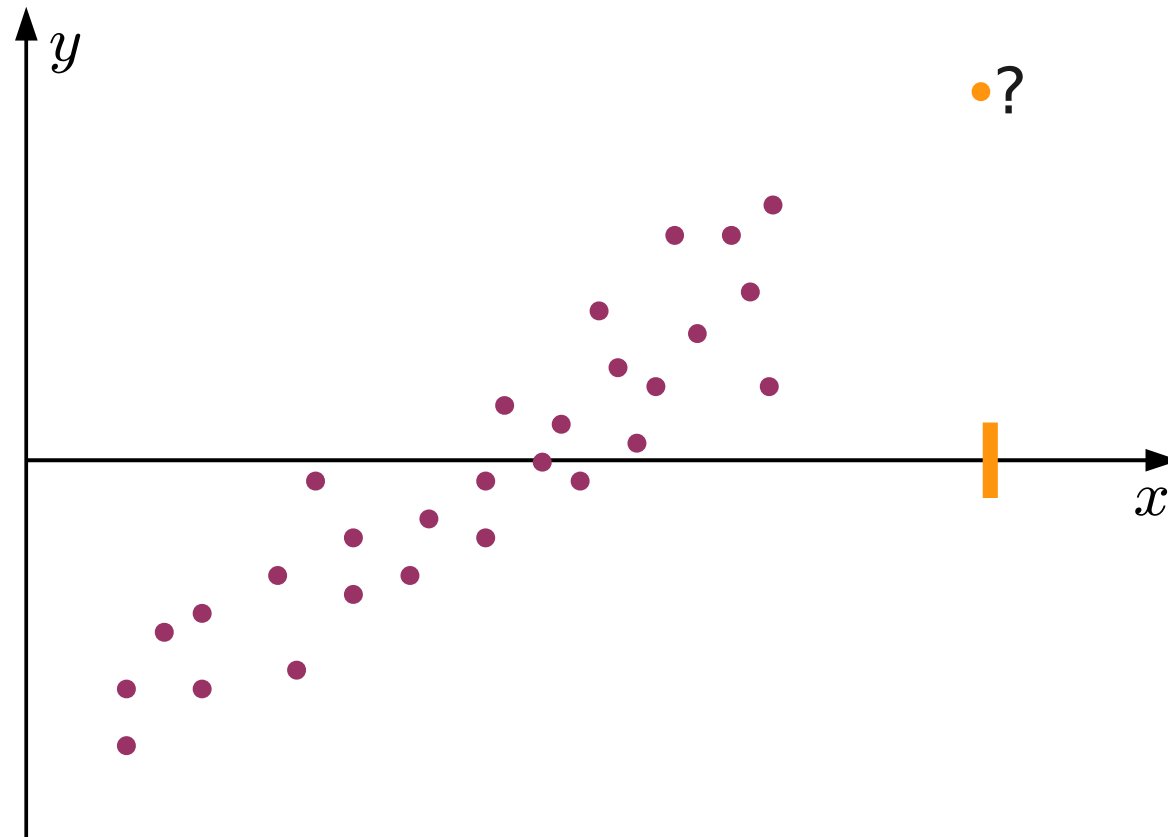
# Ordinary linear regression

## Data



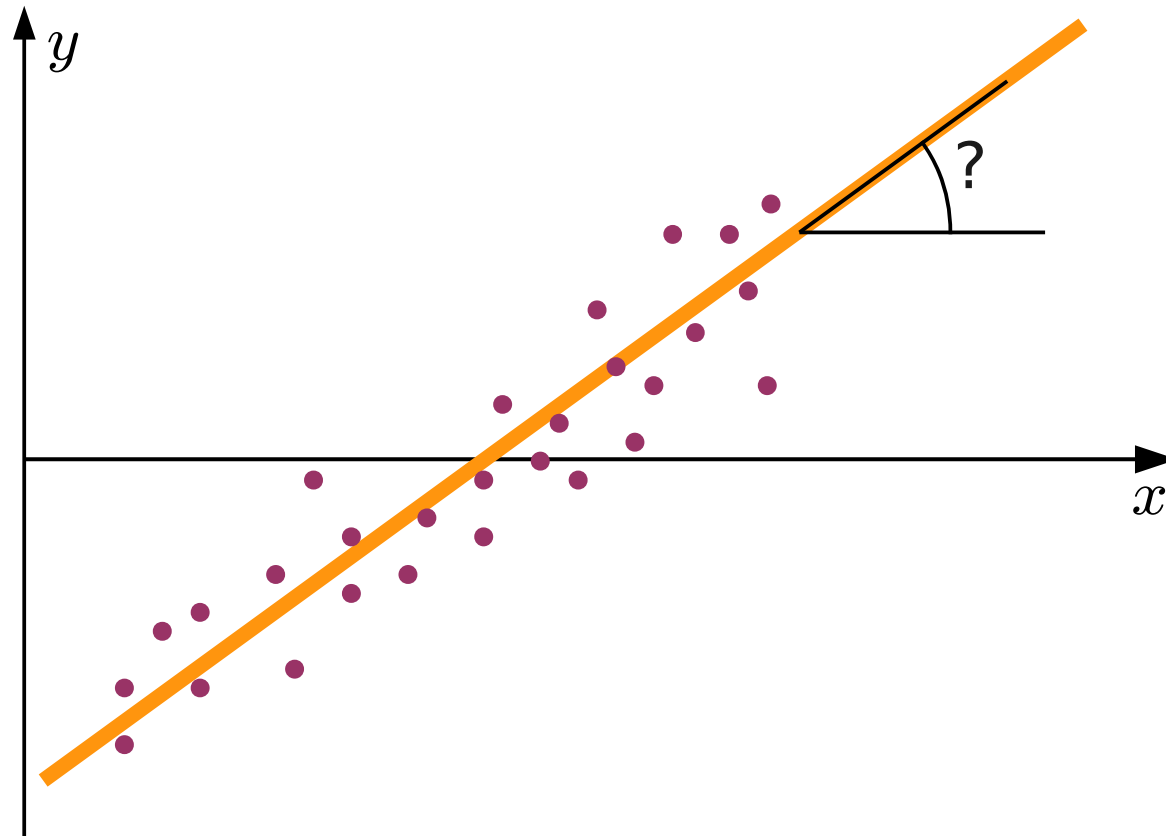
# Ordinary linear regression

## Prediction of future data point



# Ordinary linear regression

## Inference about parameter

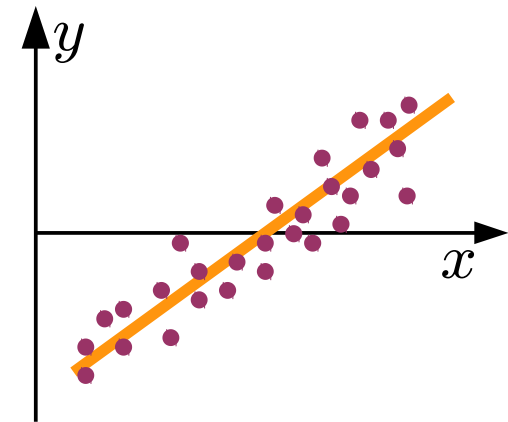


# Ordinary linear regression

## Model

- Linear model

$$\mathbb{E}(y_n) = w_0 + w_1 \cdot x_n$$



$y$  Observations

$x$  Explanatory variables

$w_0, w_1$  Model parameters

# Ordinary linear regression

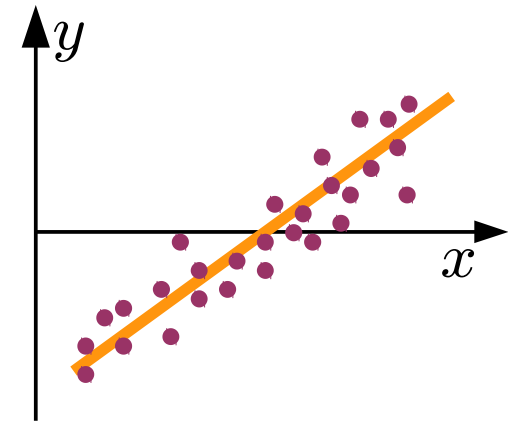
## Model

- Linear model

$$\mathbb{E}(y_n) = w_0 + w_1 \cdot x_n$$

- Model: Joint distribution

$$p(\mathbf{y}, w_0, w_1) = p(\mathbf{y}|w_0, w_1)p(w_0, w_1)$$



$\mathbf{y}$  Observations

$\mathbf{x}$  Explanatory variables

$w_0, w_1$  Model parameters

$p(\mathbf{y}|w_0, w_1)$  Likelihood

$p(w_0, w_1)$  Prior

# Ordinary linear regression

## Model

- Linear model

$$\mathbb{E}(y_n) = w_0 + w_1 \cdot x_n$$

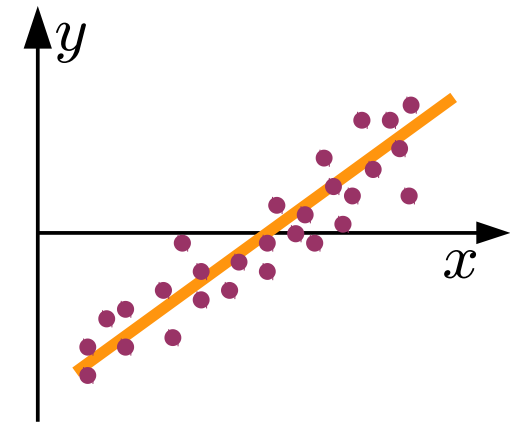
- Model: Joint distribution

$$p(\mathbf{y}, w_0, w_1) = p(\mathbf{y}|w_0, w_1)p(w_0, w_1)$$

- Example of likelihood and prior

$$p(\mathbf{y}|w_0, w_1) = \prod_{n=1}^N \mathcal{N}(y_n | w_0 + w_1 x_n, \sigma^2)$$

$$p(w_0, w_1) \propto 1$$



$\mathbf{y}$  Observations

$\mathbf{x}$  Explanatory variables

$w_0, w_1$  Model parameters

$p(\mathbf{y}|w_0, w_1)$  Likelihood

$p(w_0, w_1)$  Prior

# Ordinary linear regression

## Model with multivariate inputs

- Linear model

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\mathbf{w}$$

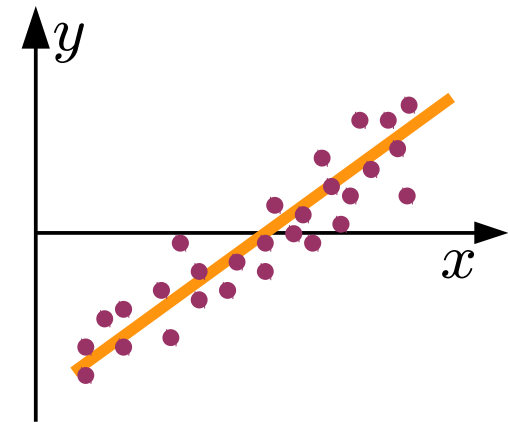
- Model: Joint distribution

$$p(\mathbf{y}, \mathbf{w}) = p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$$

- Example of likelihood and prior

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

$$p(\mathbf{w}) \propto 1$$



$\mathbf{y}$  Observations

$\mathbf{X}$  Explanatory variables

$\mathbf{w}$  Model parameters

$p(\mathbf{y}|\mathbf{w})$  Likelihood

$p(\mathbf{w})$  Prior

# Ordinary linear regression

## Model with multivariate inputs and prior over noise variance

- Linear model

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\mathbf{w}$$

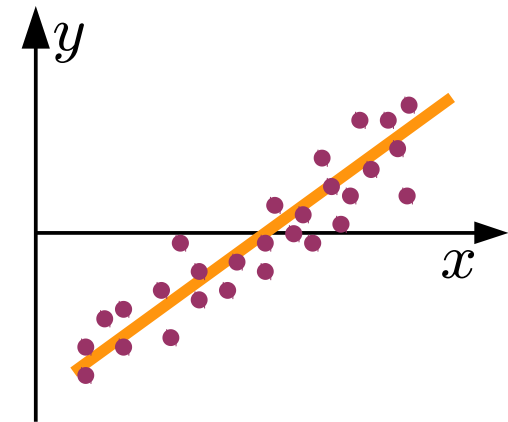
- Model: Joint distribution

$$p(\mathbf{y}, \mathbf{w}, \sigma^2) = p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}, \sigma^2)$$

- Example of likelihood and prior

$$p(\mathbf{y}|\mathbf{w}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

$$p(\mathbf{w}, \sigma^2) \propto \sigma^{-2}$$



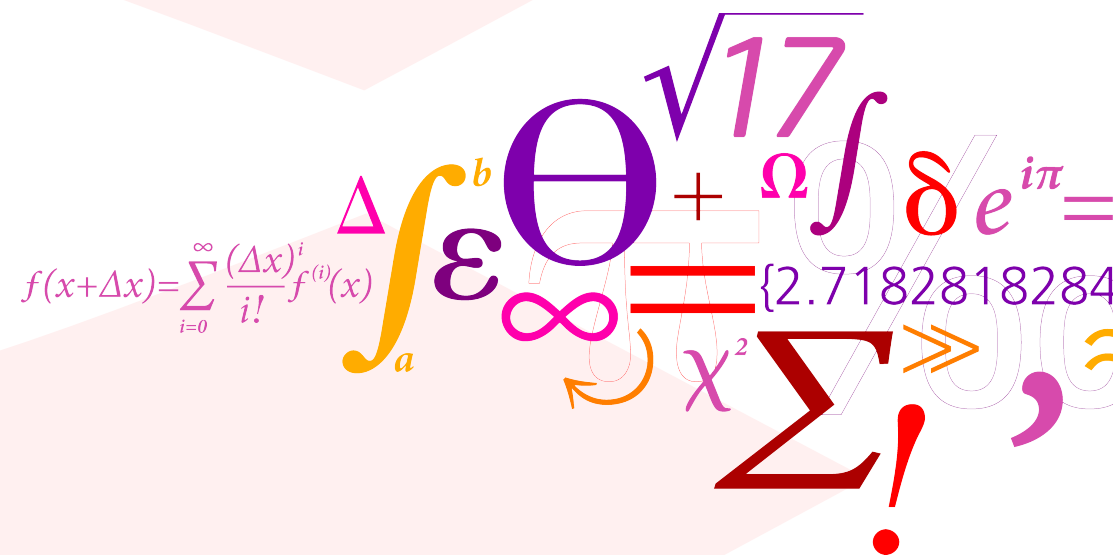
$\mathbf{y}$	Observations
$\mathbf{X}$	Explanatory variables
$\mathbf{w}, \sigma^2$	Model parameters
$p(\mathbf{y} \mathbf{w}, \sigma^2)$	<i>Likelihood</i>
$p(\mathbf{w}, \sigma^2)$	<i>Prior</i>

# Ordinary linear regression

- Solve Q 2.1 - 2.7



# Graphical models



# Graphical models

## Learning objectives

Gain knowledge about

- How the statistical relations between random variables in a model can be represented as a graph
- Factor graphs, undirected graphs, and directed graphs

Understand

- How to generate samples from a model represented as a directed graph

# Graphical models

## Introduction

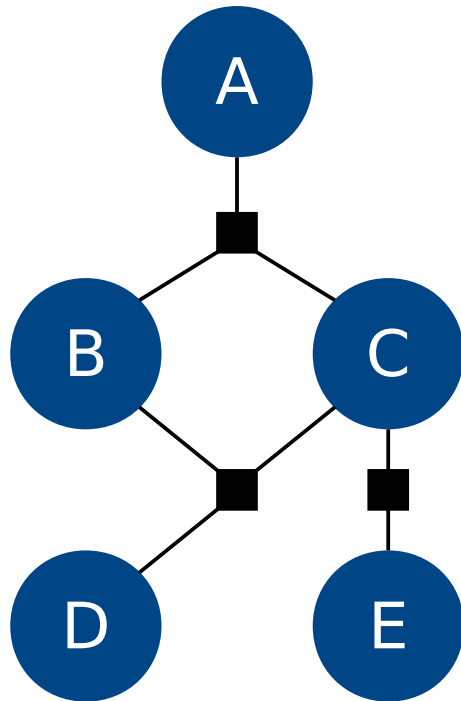
Graphical models are an intuitive way to visualize a Bayesian model

- Represents the statistical relations between variables as edges between nodes
- Makes it easy to see the hierarchical structure in the model
- Allows us to abstract out the conditional independence relations
- Can be used to define message passing algorithms

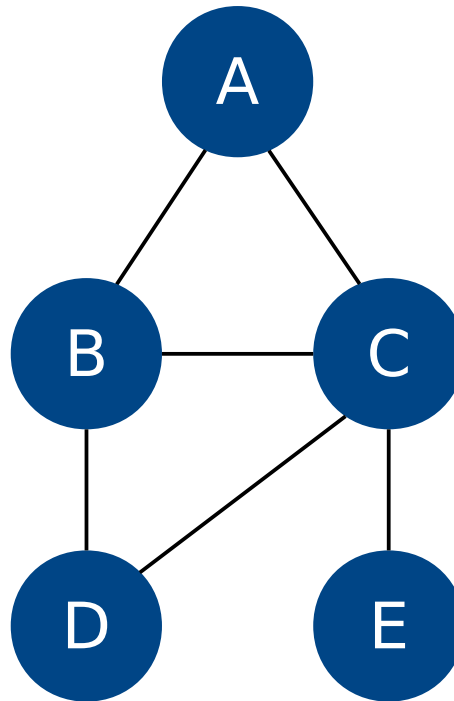
# Graphical models

## Types of graphical models

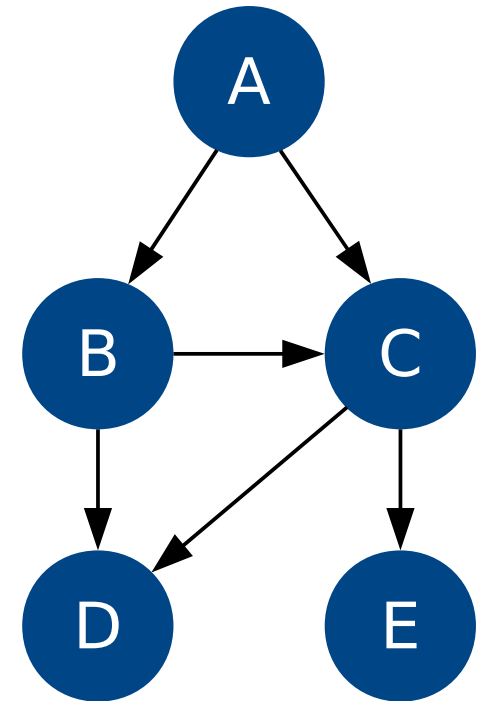
Factor graph



Undirected graph



Directed graph

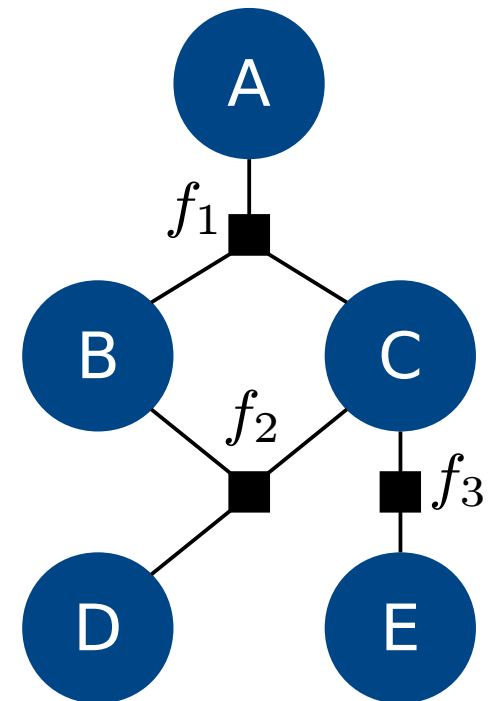


- Nodes are random variables
- Edges represent statistical dependencies

# Graphical models

## Factor graph

- Circles are random variables
- Squares represent factors
  - Factors are non-negative functions (un-normalized probability distributions)
- The joint distribution is proportional to the product of the factors
- Focus on representing the factorization of the joint distribution

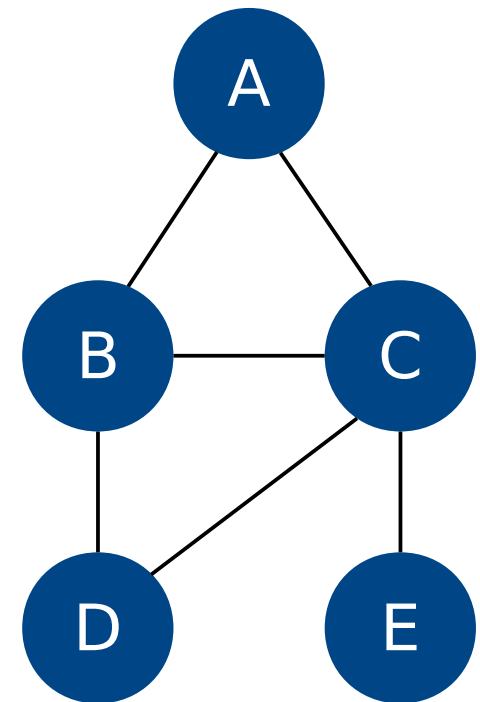


$$p(A, B, C, D, E) \propto f_1(A, B, C) f_2(B, C, D) f_3(C, E)$$

# Graphical models

## Undirected graph

- Circles are random variables
- Nodes  $i$  and  $j$  are connected if there exists a factor to which both  $i$  and  $j$  belongs
- Focus on representing soft constraints between variables

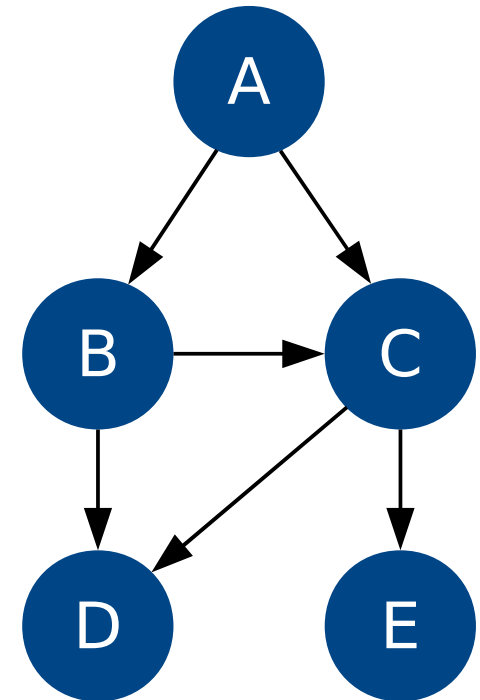


$$p(A, B, C, D, E) \propto f_1(A, B, C) f_2(B, C, D) f_3(C, E)$$

# Graphical models

## Directed graph

- Directed acyclic graph (DAG)
  - also known as a Bayesian network
- Focus on representing the (causal) generative process

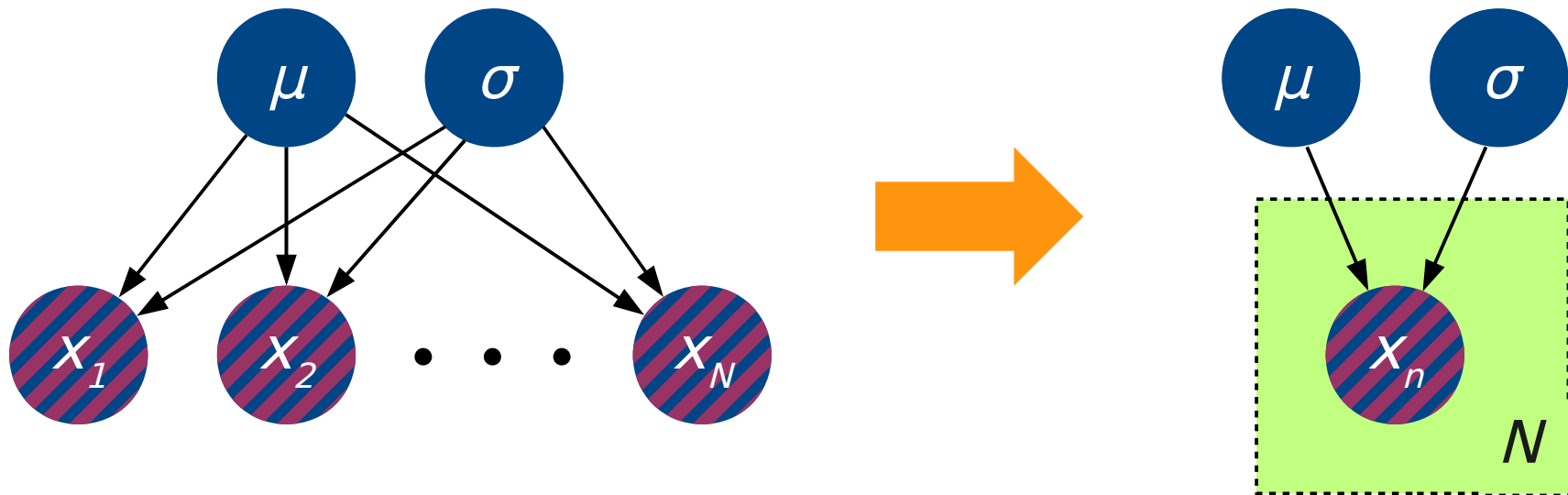


$$p(A, B, C, D, E) = p(A)p(B|A)p(C|A, B)p(D|B, C)P(E|C)$$

# Graphical models

## Special notation

- Plates can be used to denote repeated variables
- Color / shading can be used to denote observed variables



$$p(x_1, \dots, x_N, \mu, \sigma) = p(\mu)p(\sigma) \prod_{n=1}^N p(x_n | \mu, \sigma)$$

# Graphical models

## Sampling from a directed graph

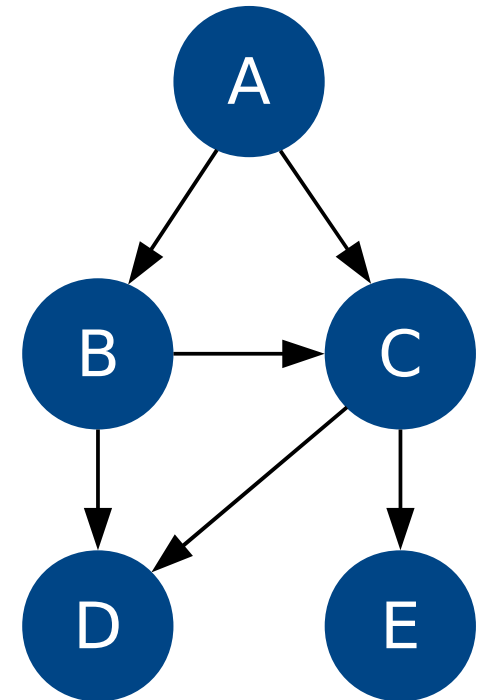
$$A \sim p(A)$$

$$B \sim p(B|A)$$

$$C \sim p(C|A, B)$$

$$D \sim p(D|B, C)$$

$$E \sim p(E|C)$$



$$p(A, B, C, D, E) = p(A)p(B|A)p(C|A, B)p(D|B, C)p(E|C)$$

# Graphical models

- Solve Q 3.1 - 3.2



# Markov chain Monte Carlo

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$\Delta$   $\int_a^b \epsilon$   $\Theta$   $\sqrt{17}$   $+$   $\Omega$   $\int \delta e^{i\pi} =$   
 $\infty$   $\{2.7182818284$   
 $\chi^2$   $\Sigma$   $\gg$   $!$

# Markov chain Monte Carlo

## Learning objectives

Understand

- How expectations can be approximated by Monte Carlo samples
- Basic methods for sampling from random distributions

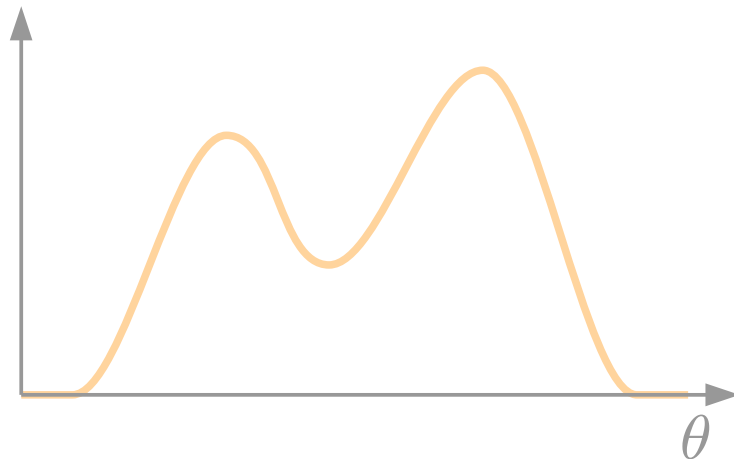
Gain knowledge about

- The basic principles of Markov chain Monte Carlo sampling methods including
  - Metropolis-Hastings
  - Gibbs sampling

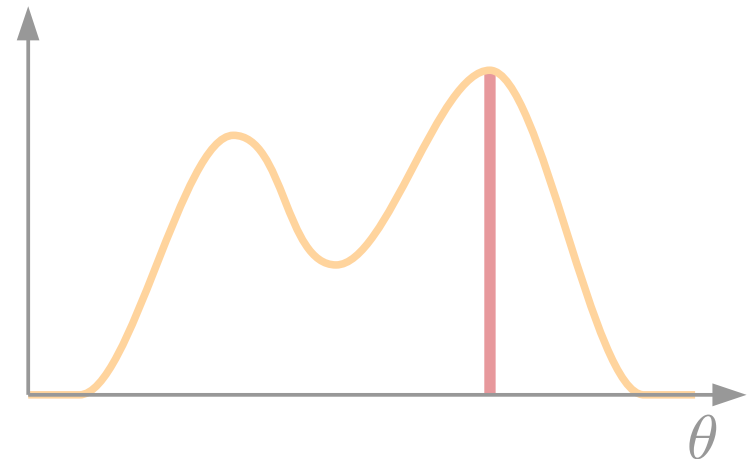
# Markov chain Monte Carlo

(Approximate) inference procedures

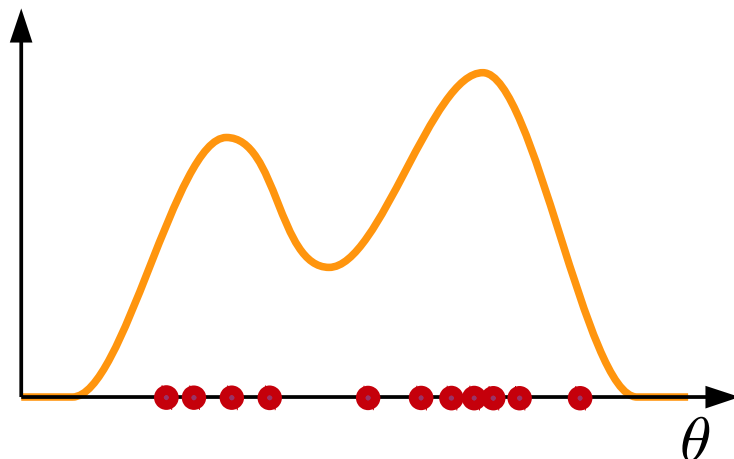
*Exact inference*



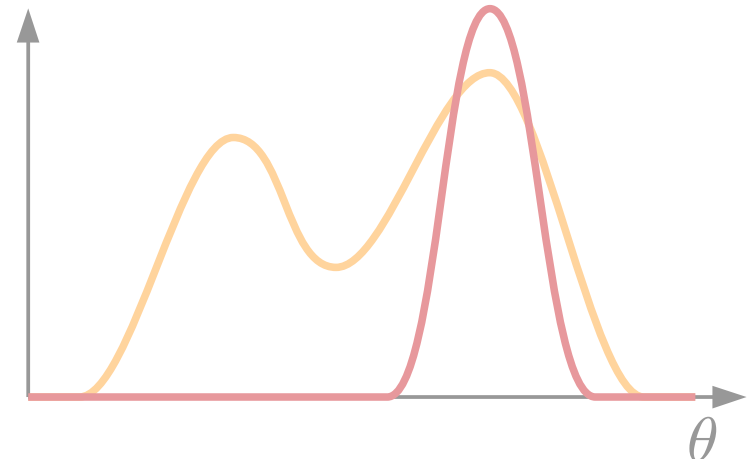
*Maximum a posteriori (MAP)*



*Monte Carlo sampling*



*Variational inference*



# Markov chain Monte Carlo

## Monte Carlo sampling

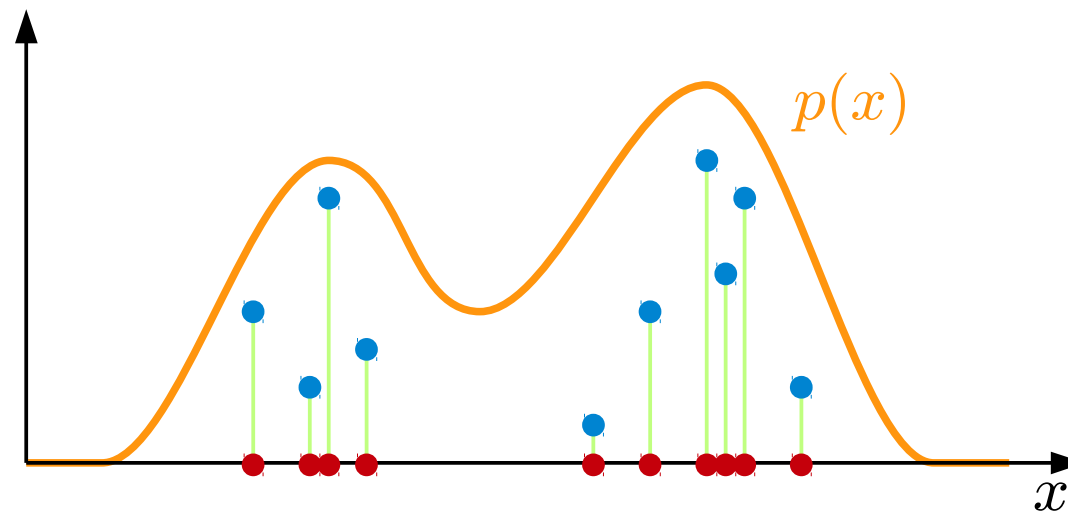
Any expectation can be approximated by sampling

$$\mathbb{E}[f(x)] = \int f(x)p(x)dx \approx \frac{1}{J} \sum_{j=1}^J f(x^{(j)}), \quad x^{(j)} \sim p(x)$$

# Markov chain Monte Carlo

## Sampling from distributions

Draw points uniformly under the probability density curve



# Markov chain Monte Carlo

## Inverse transform sampling

If the cumulative distribution is known and invertible



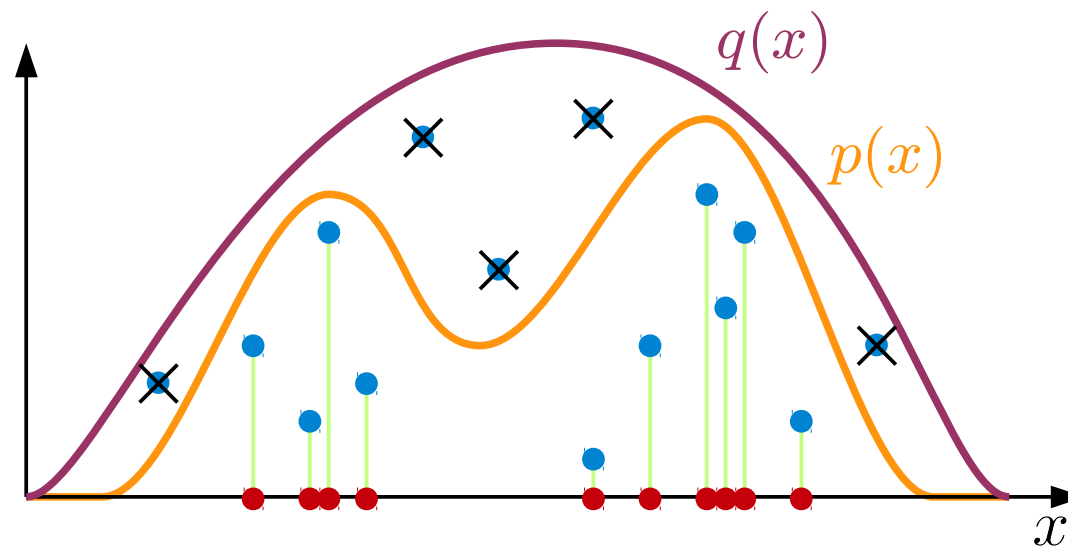
Sample  $u \sim \mathcal{U}(u|0, 1)$

Compute  $x = h^{-1}(u)$

# Markov chain Monte Carlo

## Rejection sampling

Sample from a simple distribution



Sample from a simple distribution  $x \sim q(x) \geq p(x)$

Sample  $u \sim \mathcal{U}(u|0, q(x))$

Reject if  $u \geq p(x)$

# Markov chain Monte Carlo

## Sampling using a Markov chain

- Start at some random initial value

$$\mathbf{x}^{(0)}$$

- A Markov chain generates samples conditioned on the previous sample

$$\mathbf{x}^{(m)} \sim p(\mathbf{x}^{(m)} | \mathbf{x}^{(m-1)})$$

- The chain is constructed such that its stationary distribution is the desired distribution
  - In the limit, samples will be drawn from the desired distribution
- Output is a set of samples

$$\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$$

- In practice: Check for convergence and throw away initial samples (burn in)

# Markov chain Monte Carlo

## Metropolis-Hastings

Task: Generate samples from the multivariate distribution

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_N)$$

- Choose a random initial point

$$\mathbf{x} = \mathbf{x}^{(0)}$$

- Generate a random sample from a proposal distribution

$$\mathbf{x}^* \sim q(\mathbf{x}^* | \mathbf{x})$$

and accept it with probability

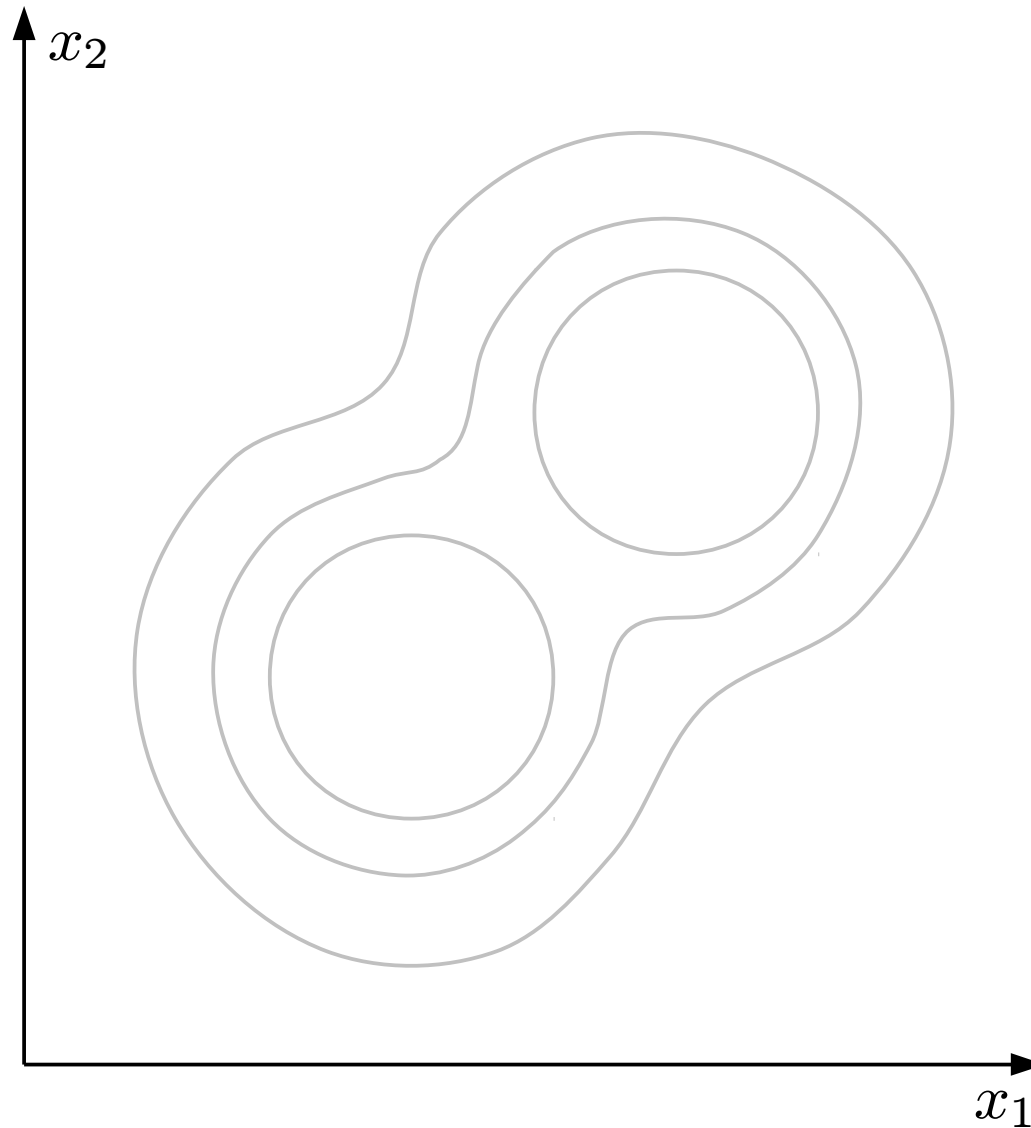
$$\alpha = \frac{p(\mathbf{x}^*)q(\mathbf{x} | \mathbf{x}^*)}{p(\mathbf{x})q(\mathbf{x}^* | \mathbf{x})}$$

otherwise keep an extra copy of the old sample

Converges to a sample from the desired distribution

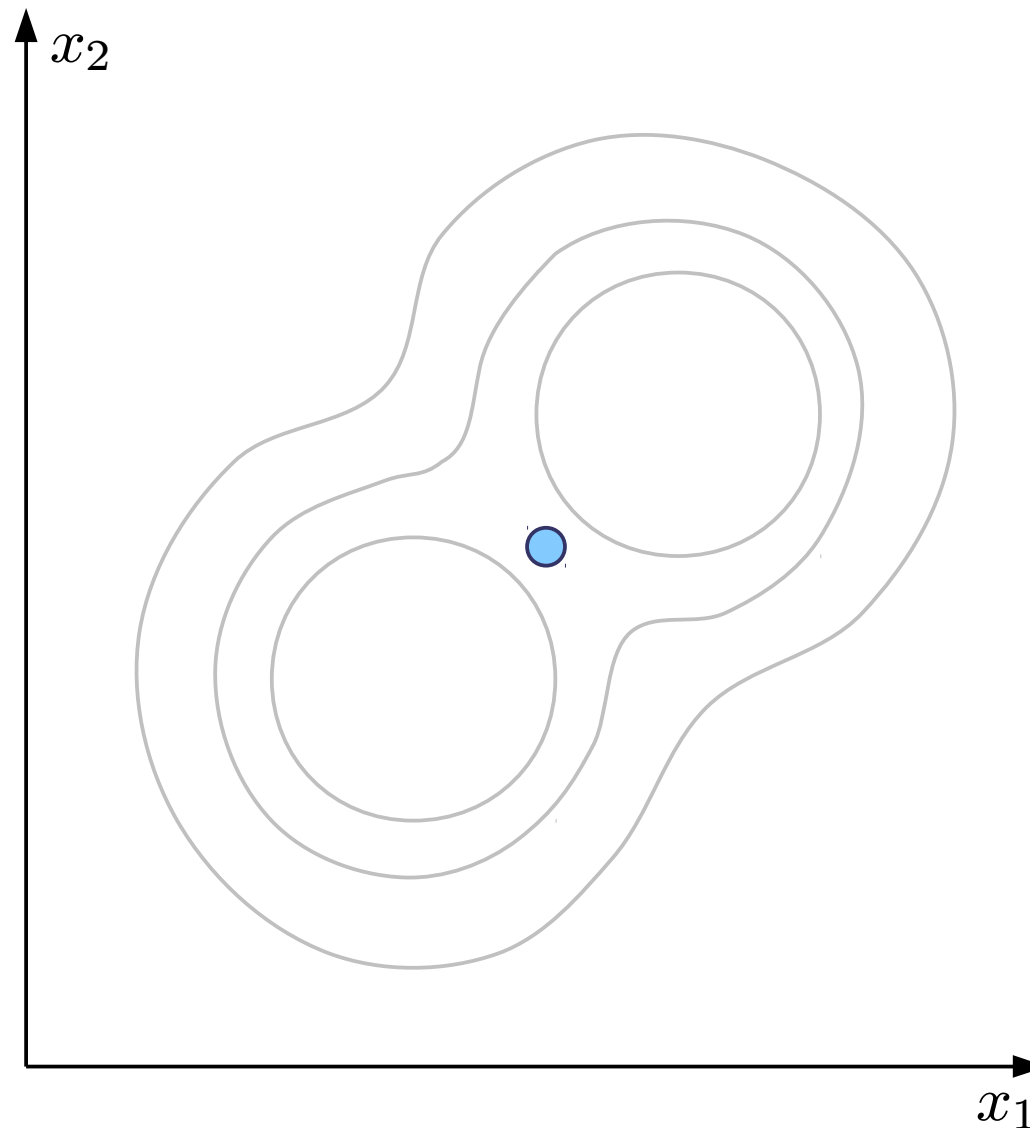
# Markov chain Monte Carlo

## Metropolis-Hastings



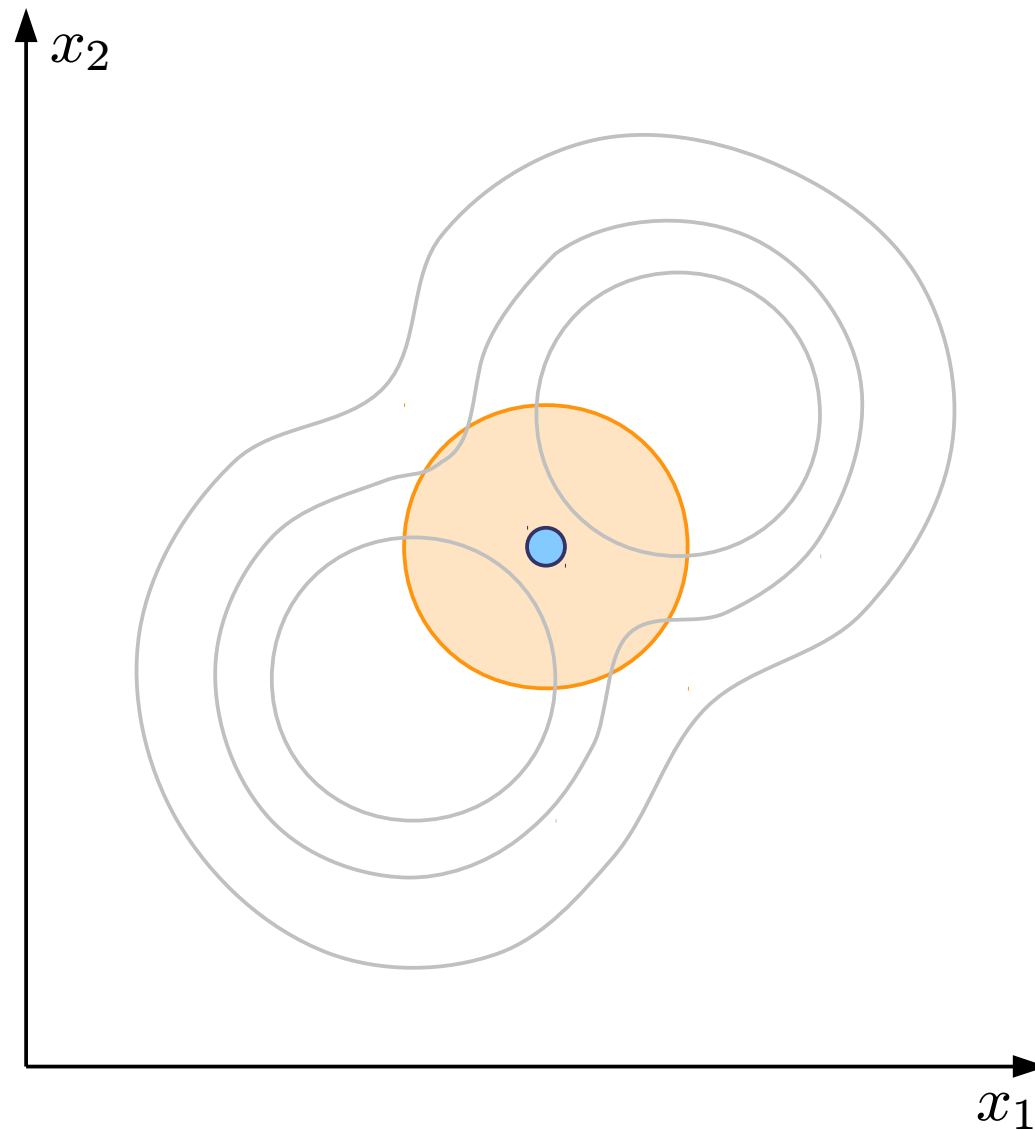
# Markov chain Monte Carlo

## Metropolis-Hastings



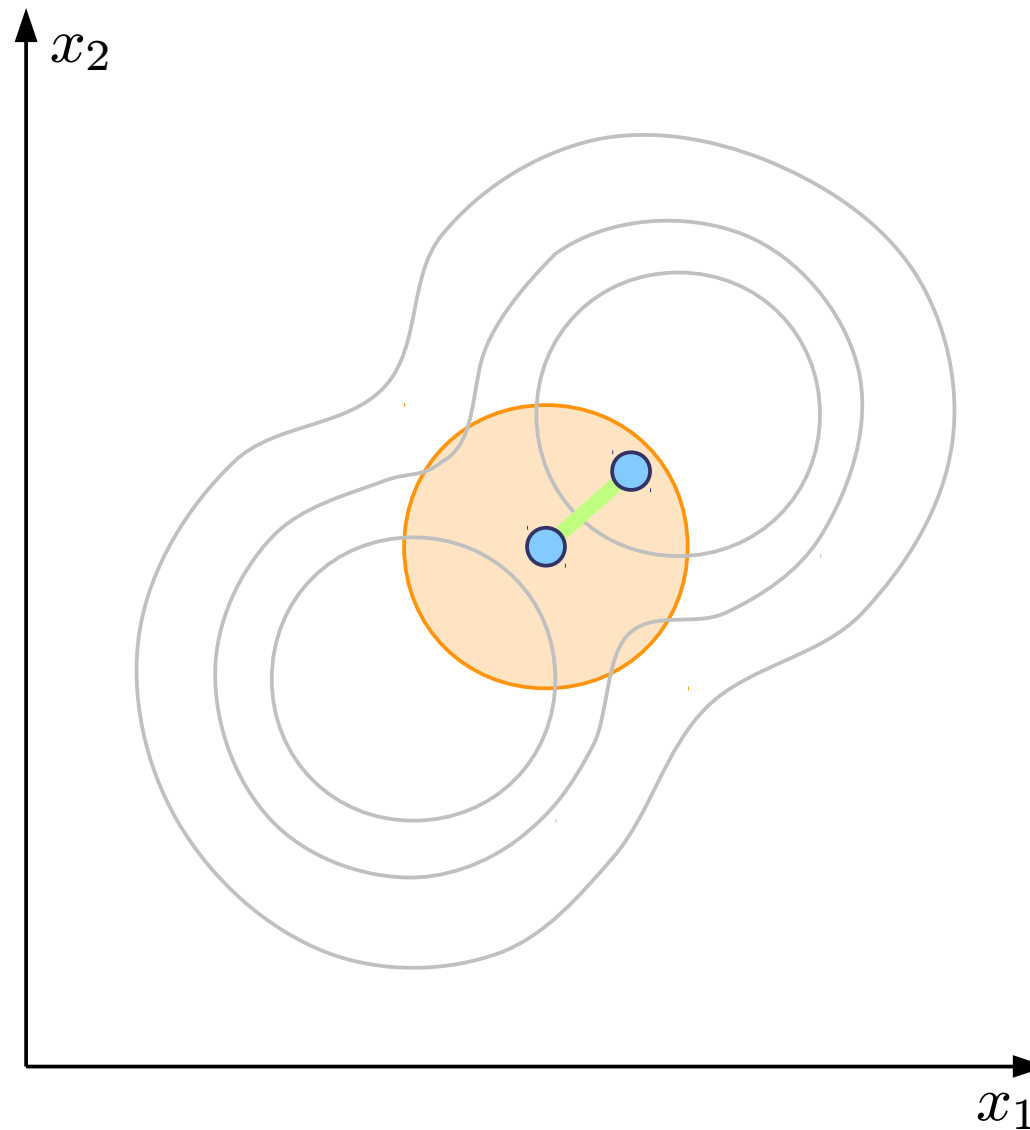
# Markov chain Monte Carlo

## Metropolis-Hastings



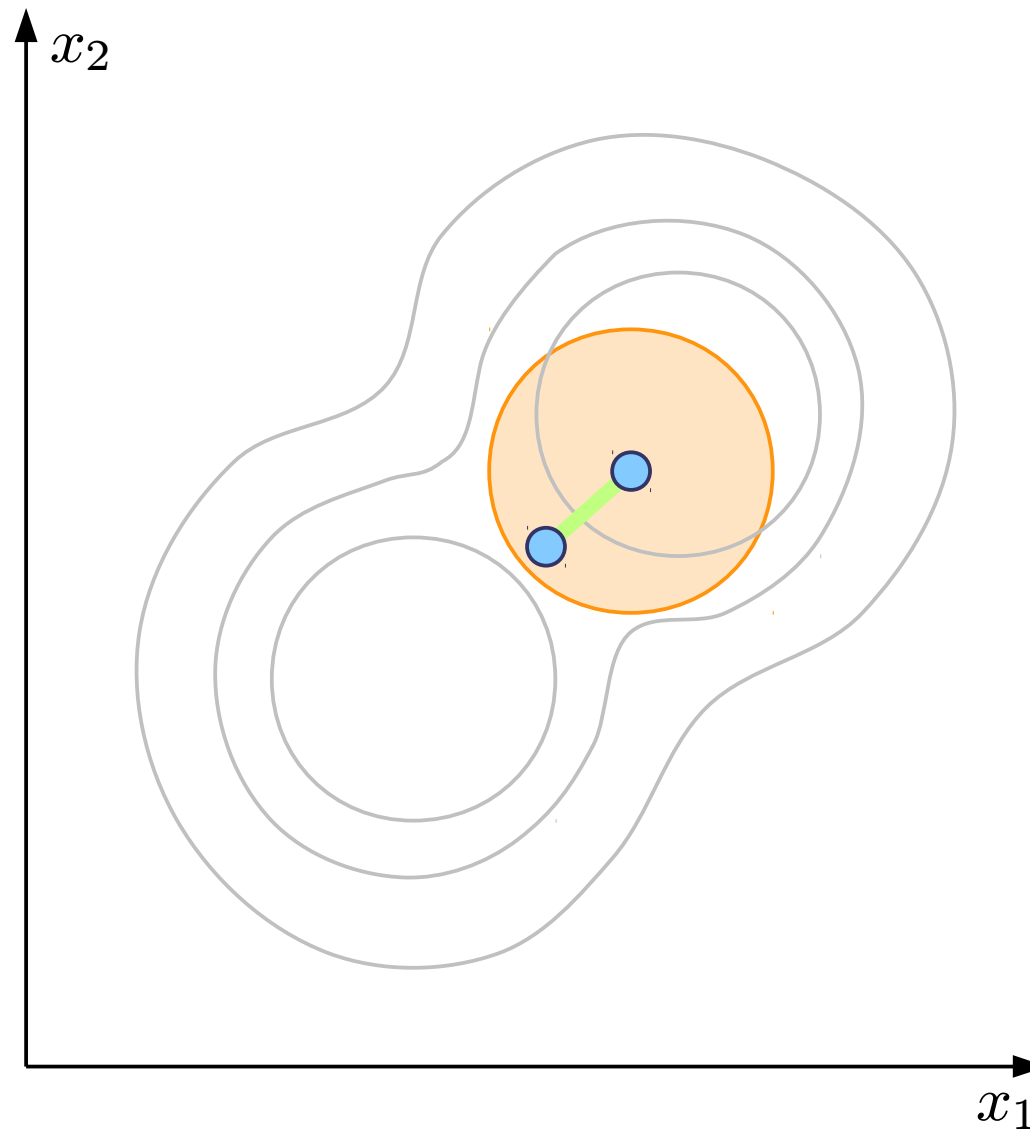
# Markov chain Monte Carlo

## Metropolis-Hastings



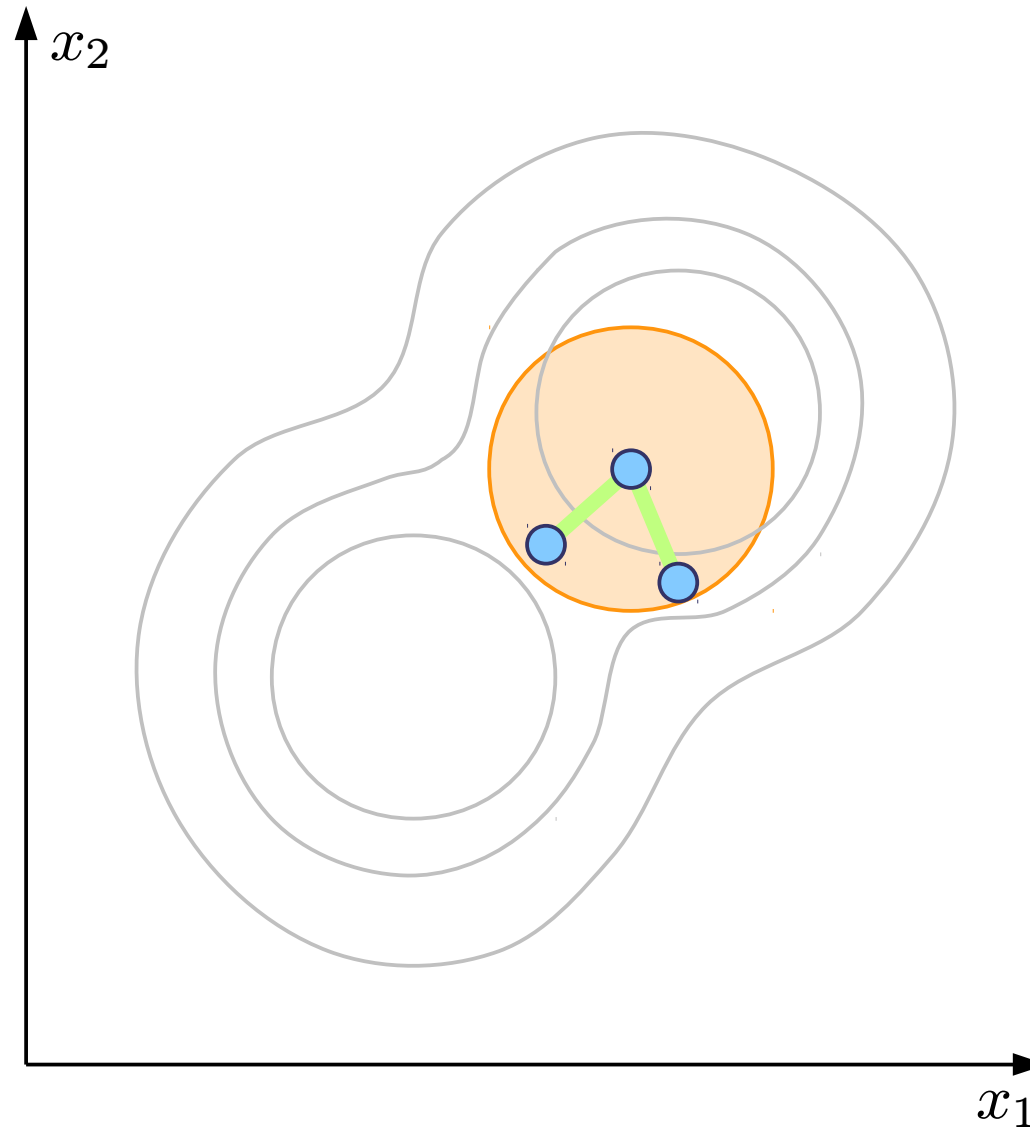
# Markov chain Monte Carlo

## Metropolis-Hastings



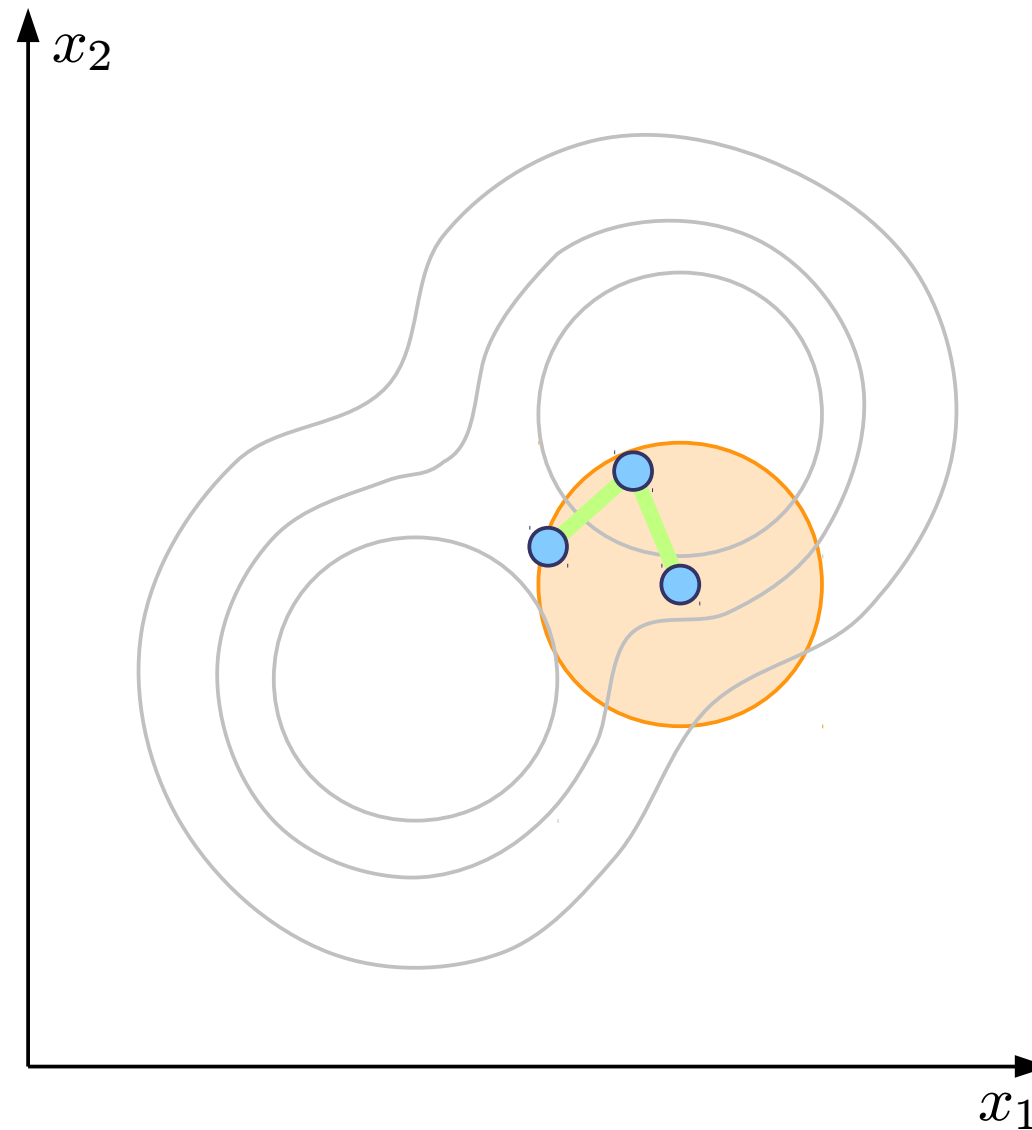
# Markov chain Monte Carlo

## Metropolis-Hastings



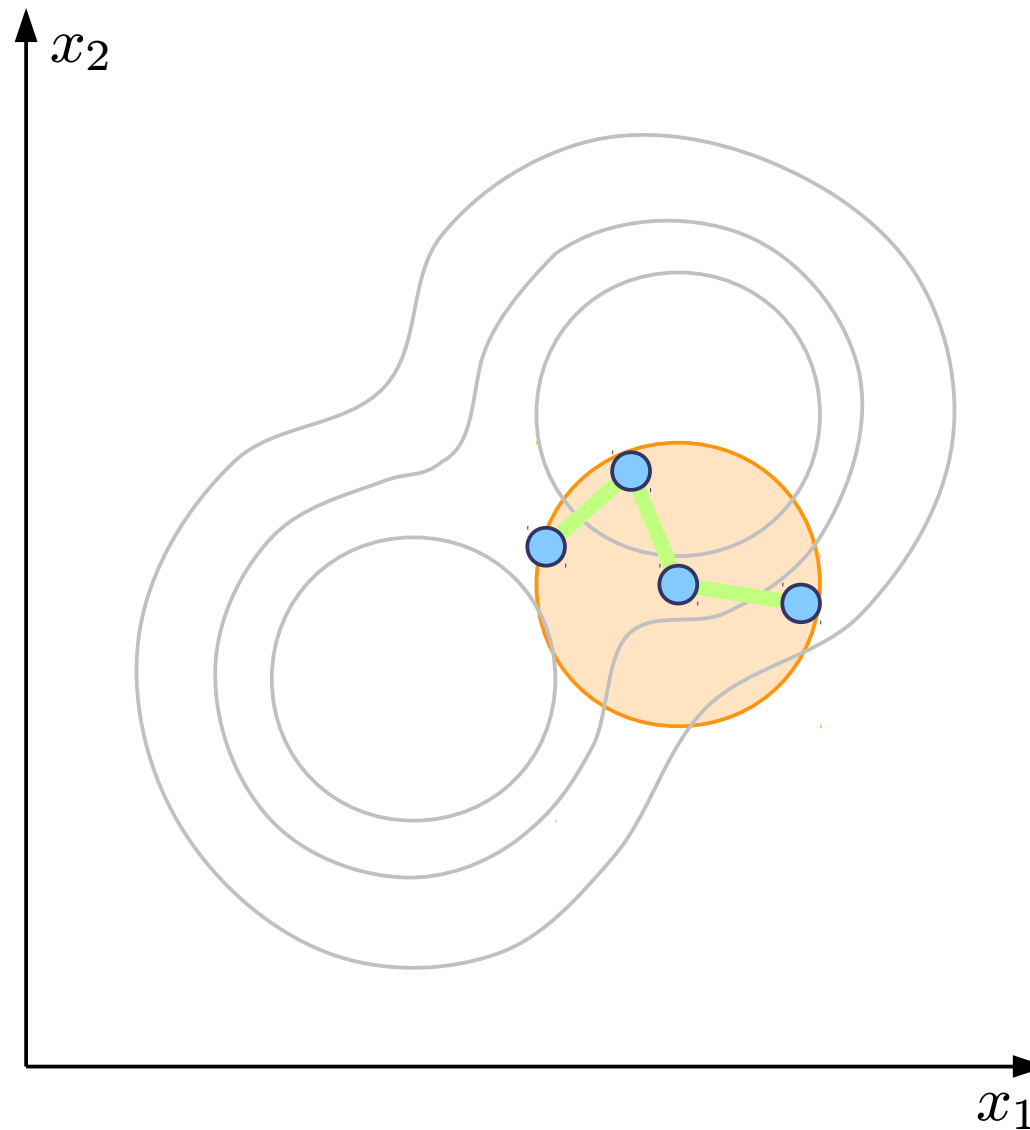
# Markov chain Monte Carlo

## Metropolis-Hastings



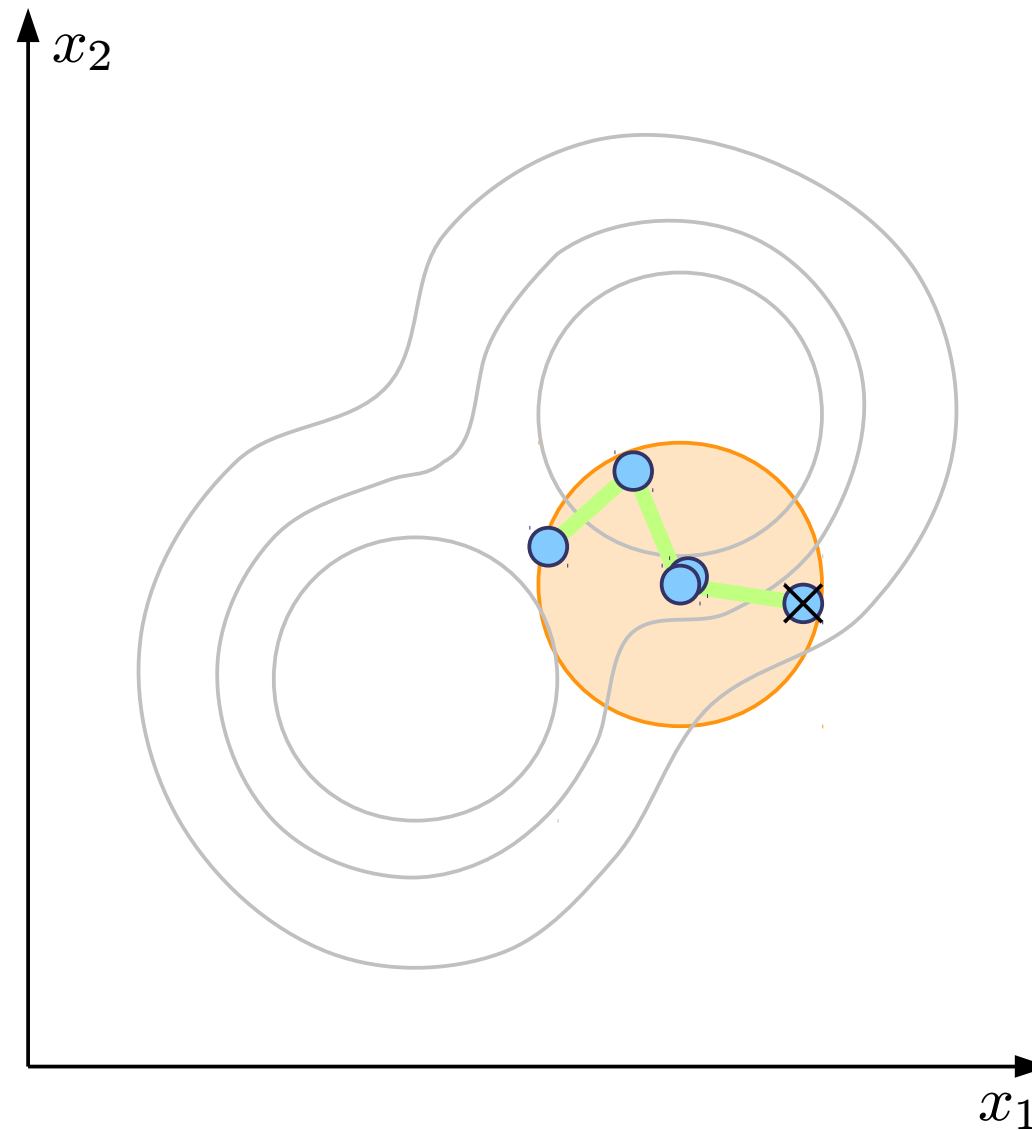
# Markov chain Monte Carlo

## Metropolis-Hastings



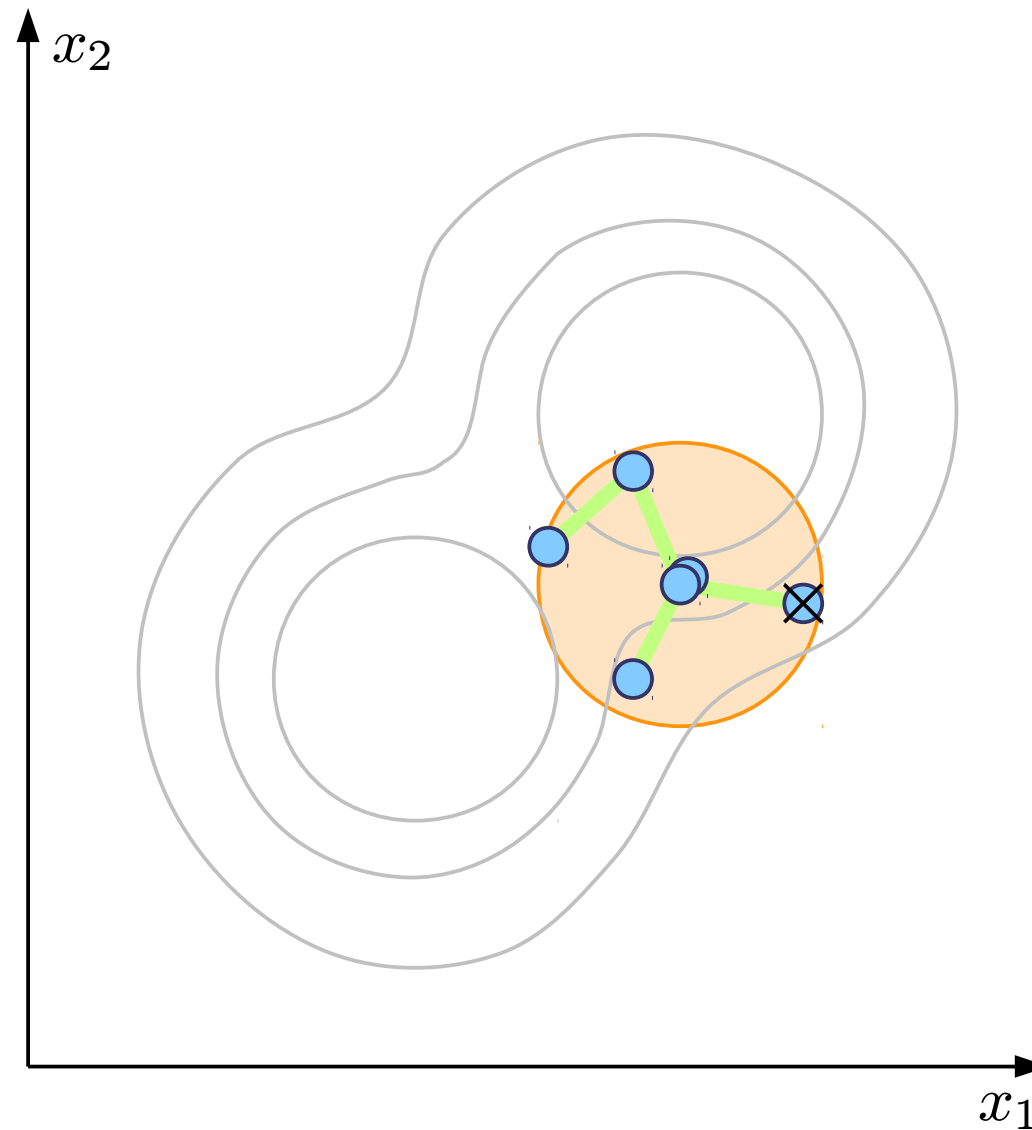
# Markov chain Monte Carlo

## Metropolis-Hastings



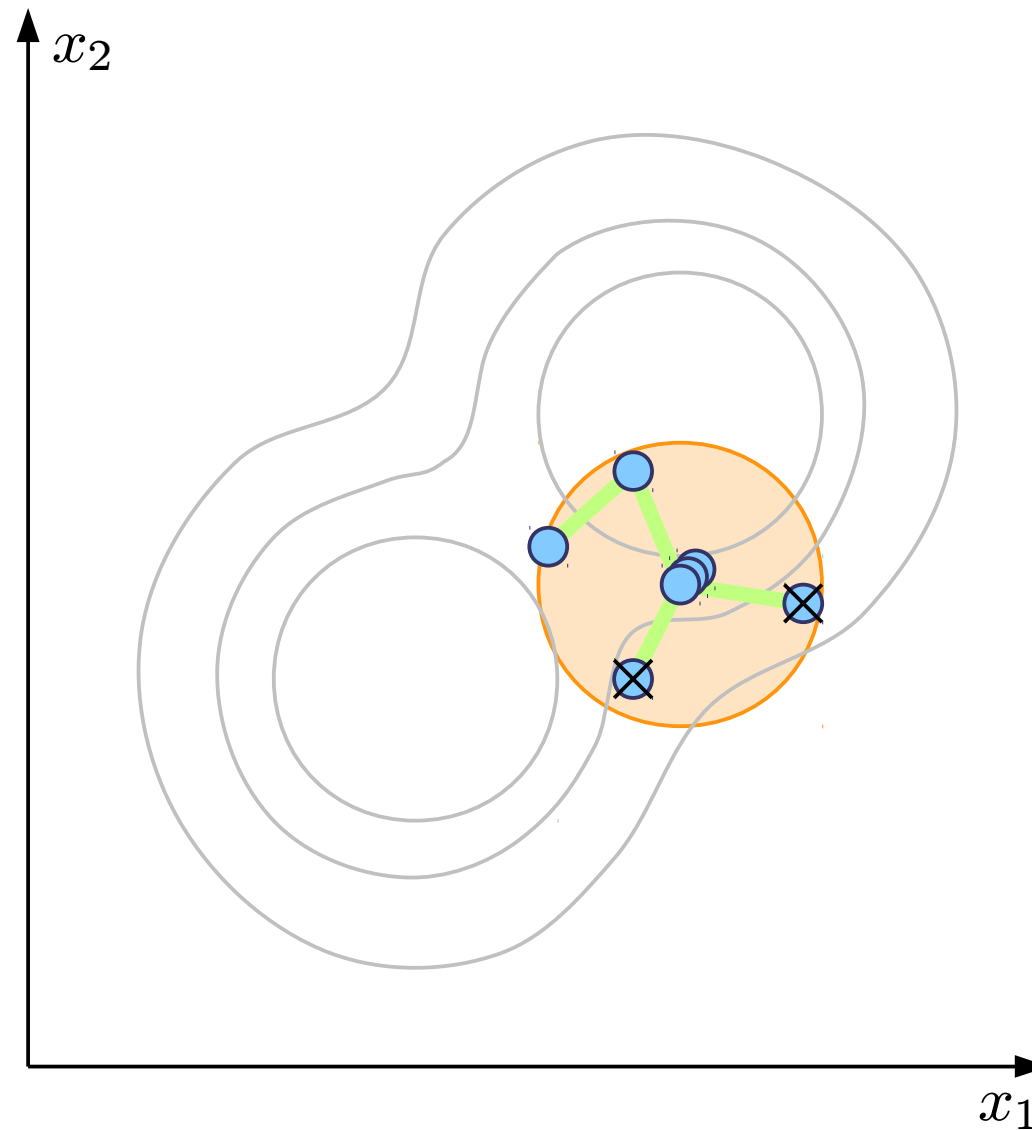
# Markov chain Monte Carlo

## Metropolis-Hastings



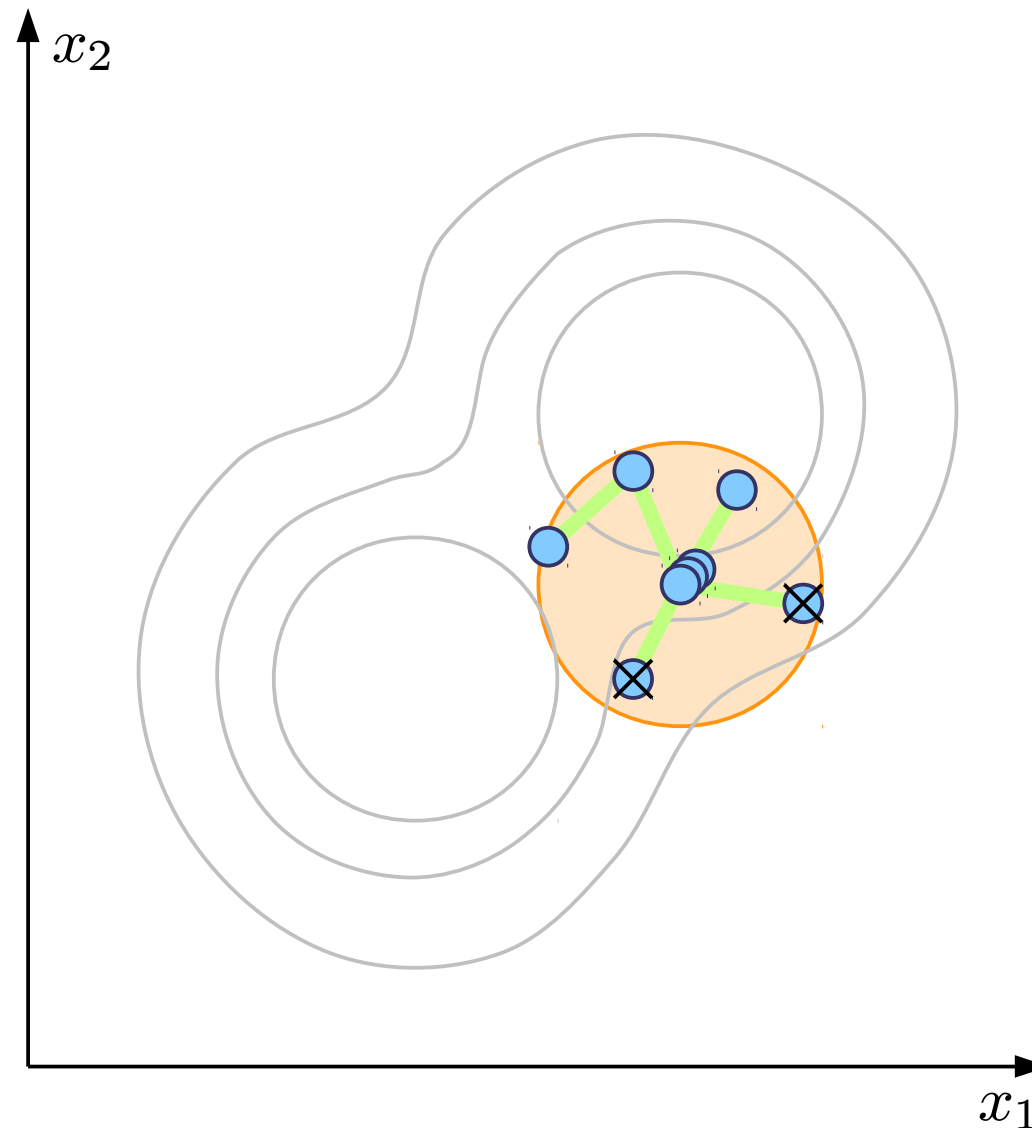
# Markov chain Monte Carlo

## Metropolis-Hastings



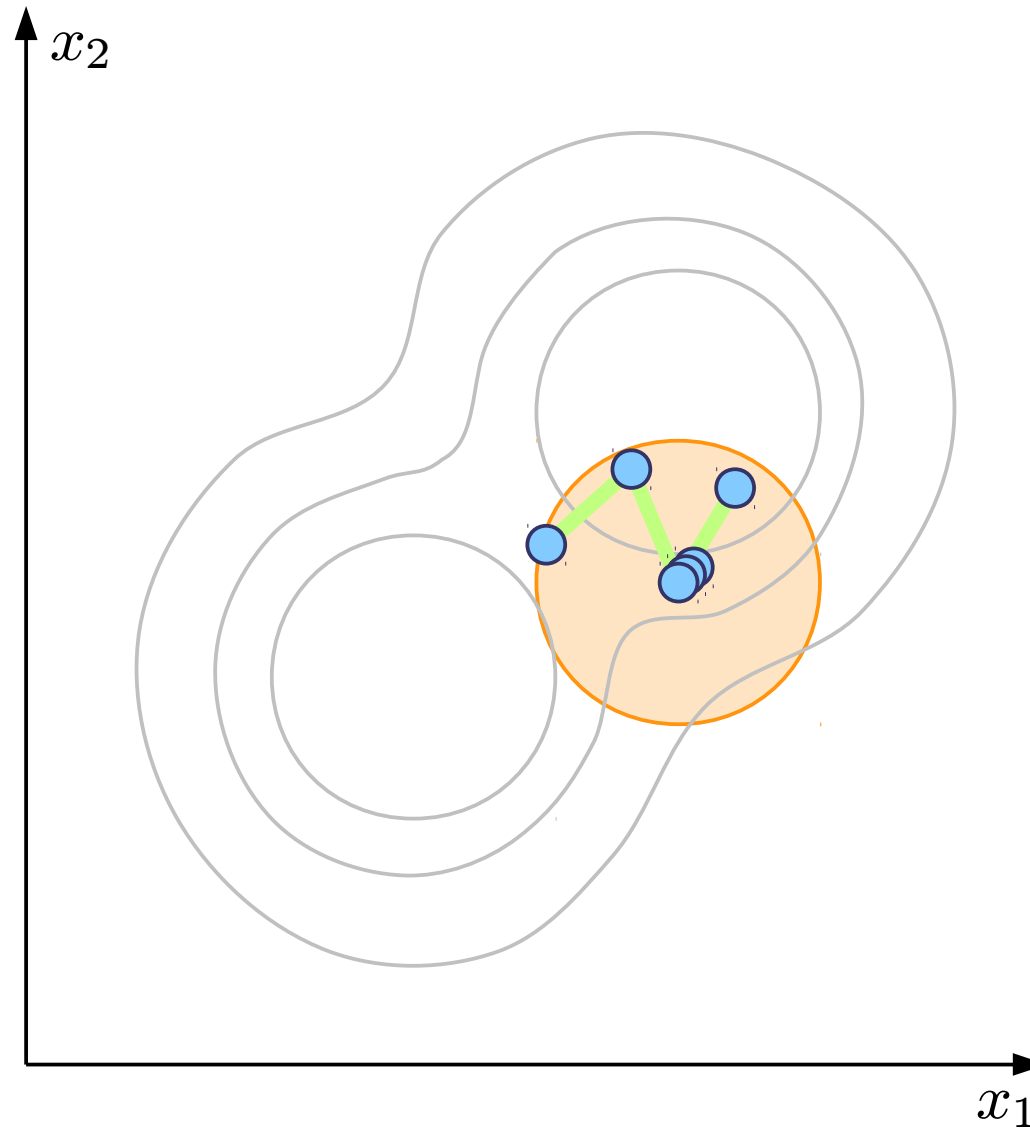
# Markov chain Monte Carlo

## Metropolis-Hastings



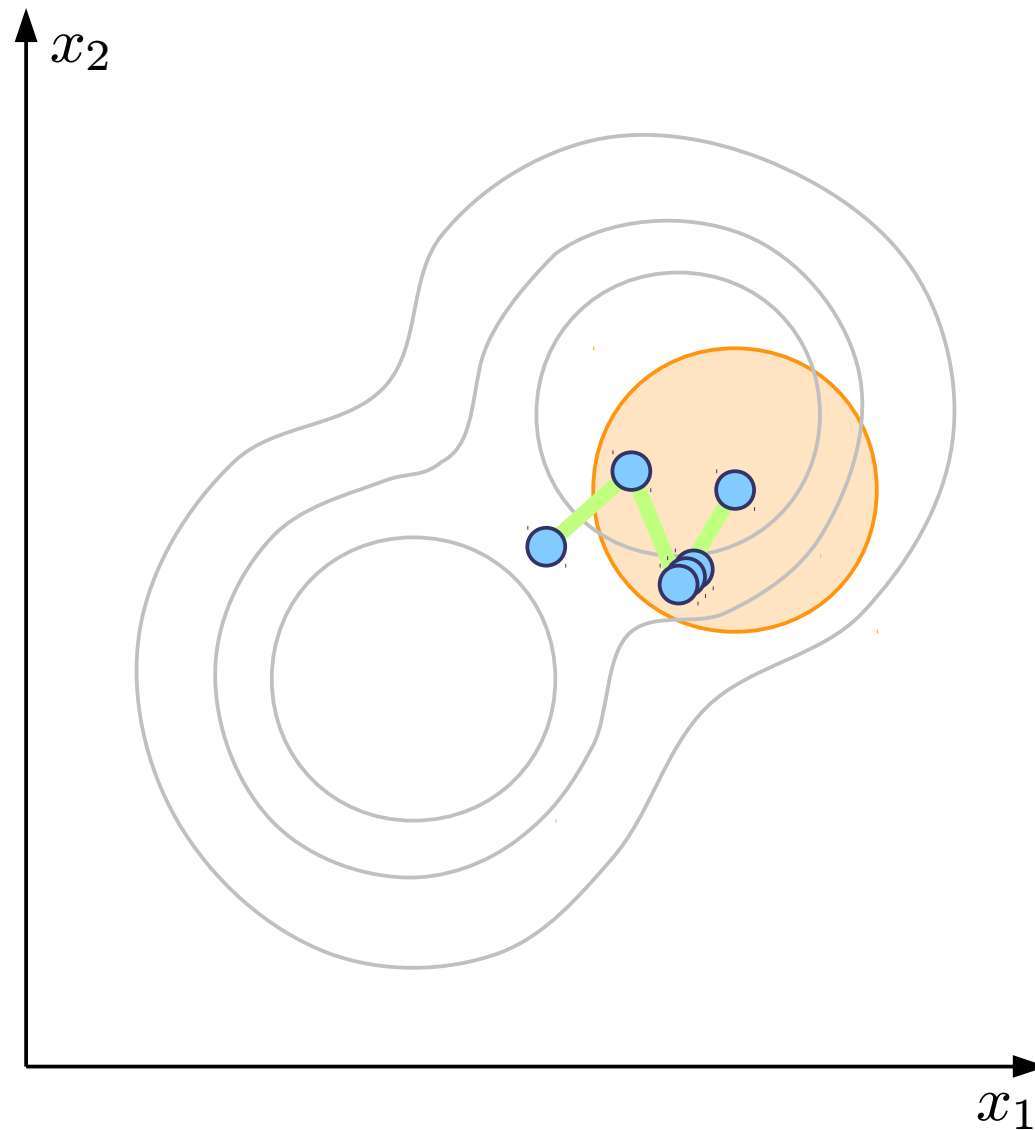
# Markov chain Monte Carlo

## Metropolis-Hastings



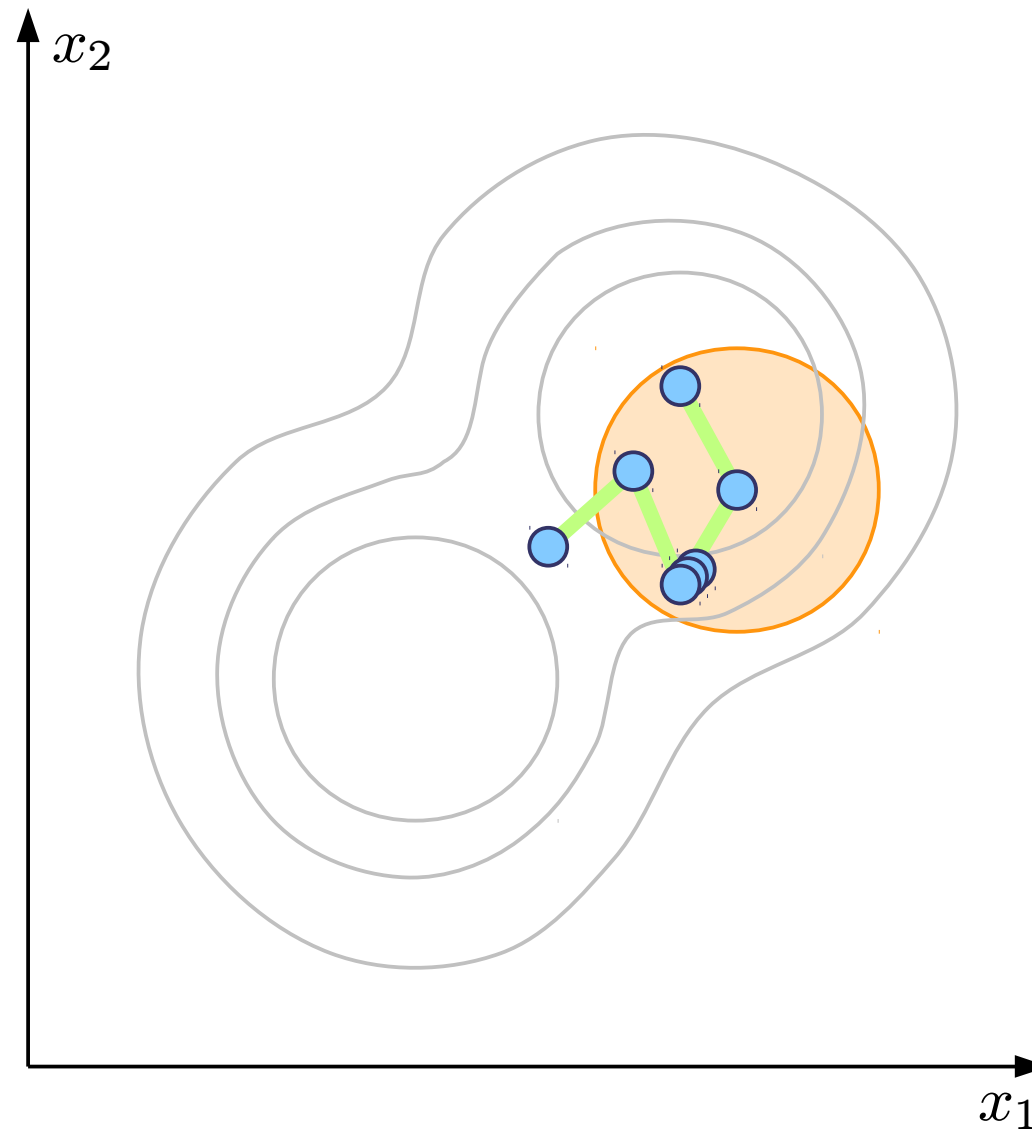
# Markov chain Monte Carlo

## Metropolis-Hastings



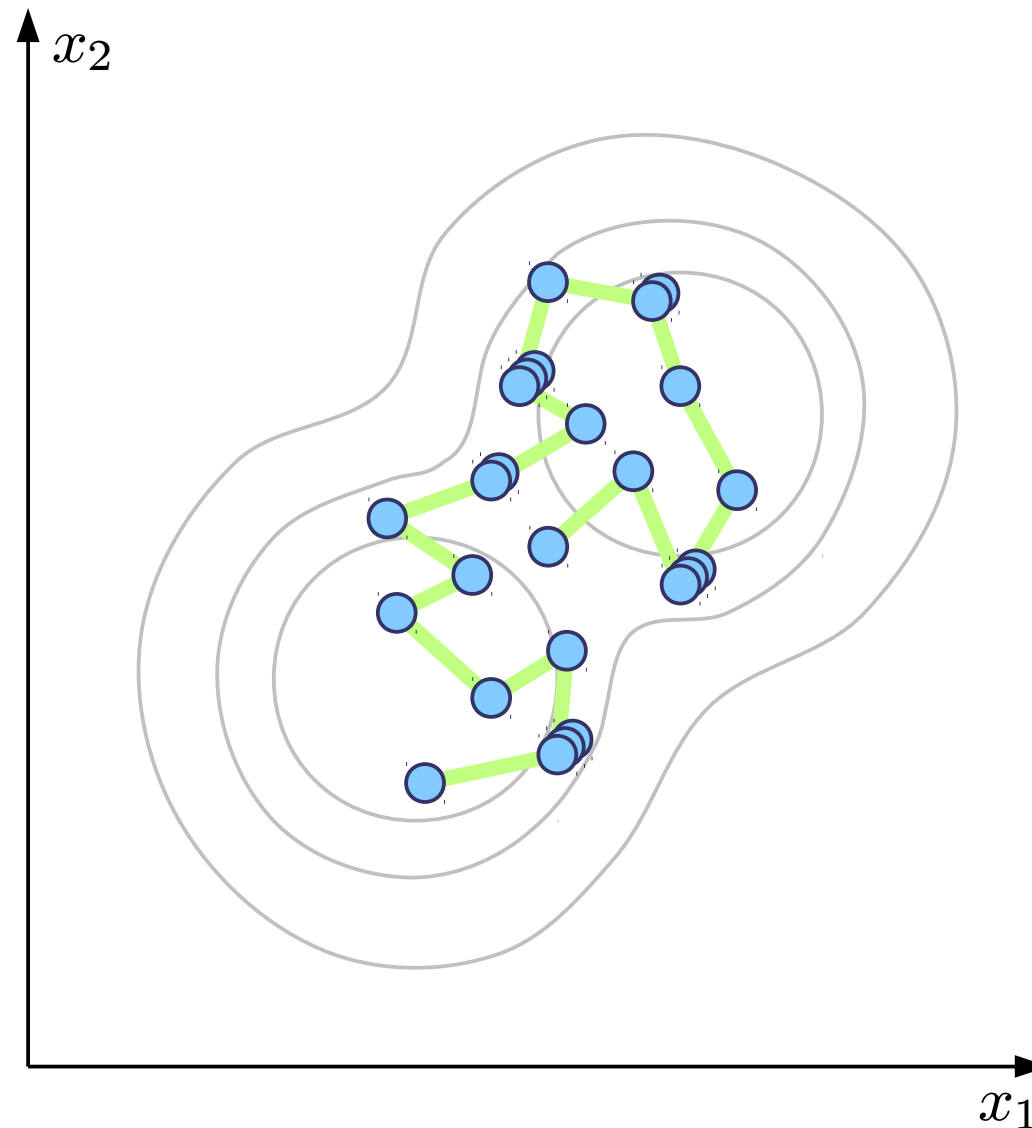
# Markov chain Monte Carlo

## Metropolis-Hastings



# Markov chain Monte Carlo

## Metropolis-Hastings



# Markov chain Monte Carlo

## Metropolis-Hastings

Task: Generate samples from the multivariate distribution

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_N)$$

- Choose a random initial point

$$\mathbf{x} = \mathbf{x}^{(0)}$$

- Generate a random sample from a proposal distribution

$$\mathbf{x}^* \sim q(\mathbf{x}^* | \mathbf{x})$$

and accept it with probability

$$\alpha = \frac{p(\mathbf{x}^*)q(\mathbf{x} | \mathbf{x}^*)}{p(\mathbf{x})q(\mathbf{x}^* | \mathbf{x})}$$

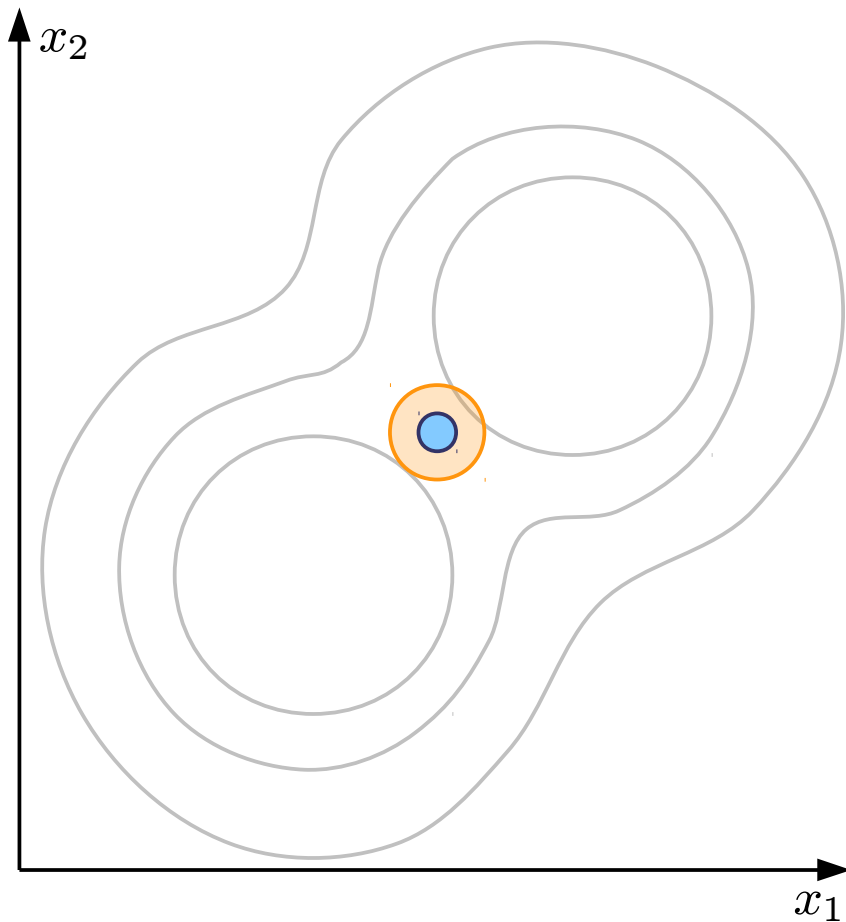
otherwise keep an extra copy of the old sample

Converges to a sample from the desired distribution

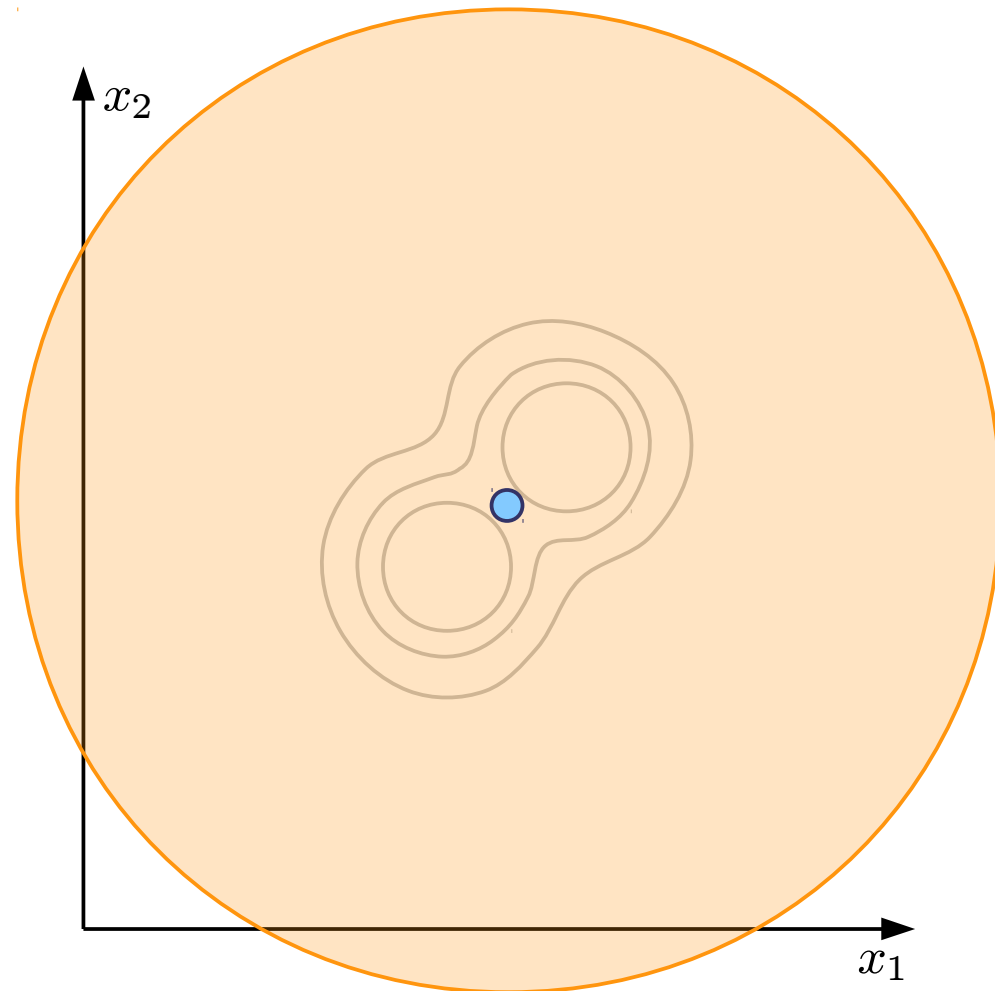
# Markov chain Monte Carlo

## Metropolis-Hastings

Narrow proposal: Small steps



Wide proposal: Many rejections



# Markov chain Monte Carlo

## Gibbs sampling

Task: Generate samples from the multivariate distribution

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_N)$$

- Choose a random initial point

$$\mathbf{x} = \mathbf{x}^{(0)}$$

- Sequentially sample each variable conditional on all other variables

$$x_1 \sim p(x_1 | x_2, \dots, x_N)$$

$$x_2 \sim p(x_2 | x_1, x_3, \dots, x_N)$$

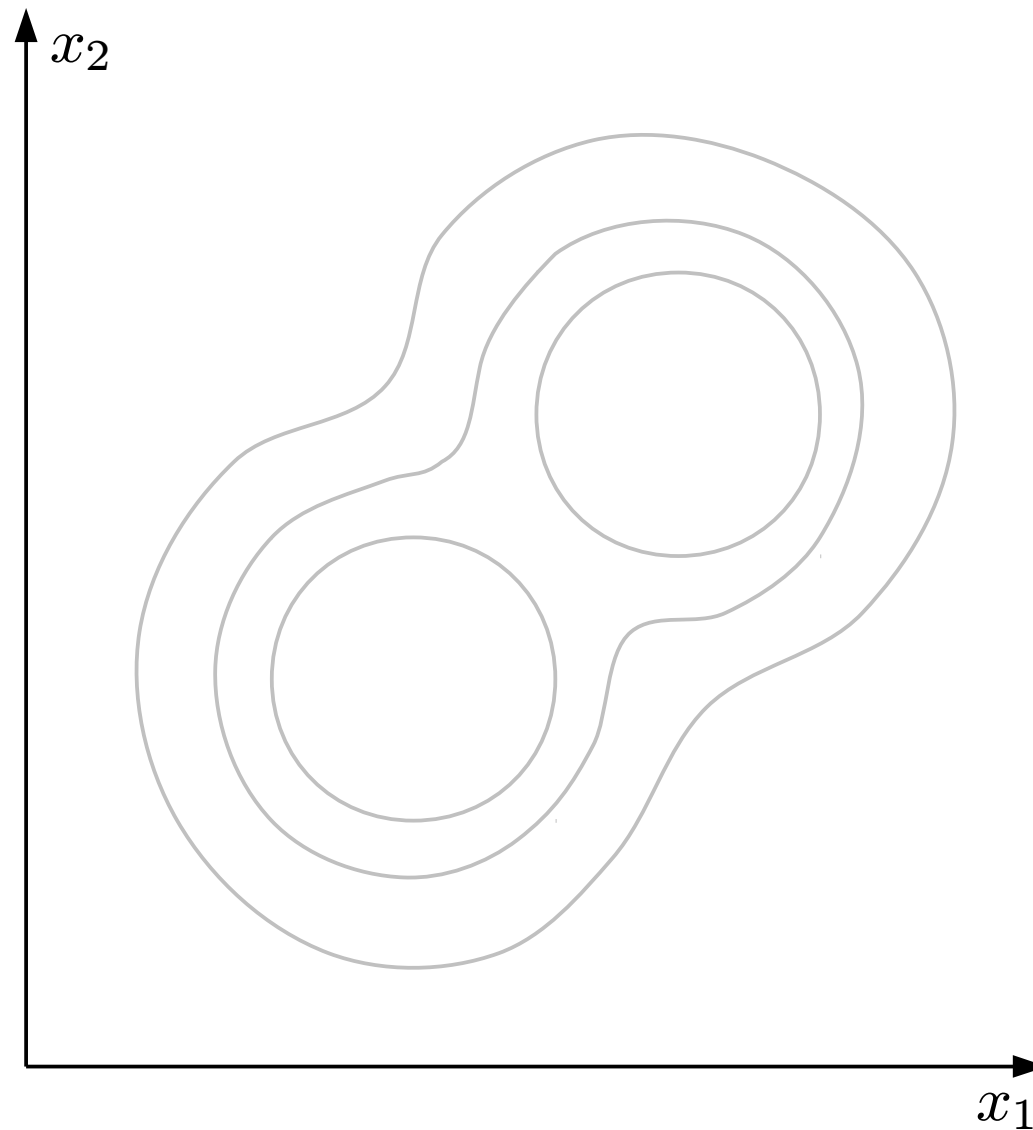
$$\vdots$$

$$x_N \sim p(x_N | x_1, x_2, \dots, x_{N-1})$$

Converges to a sample from the desired distribution

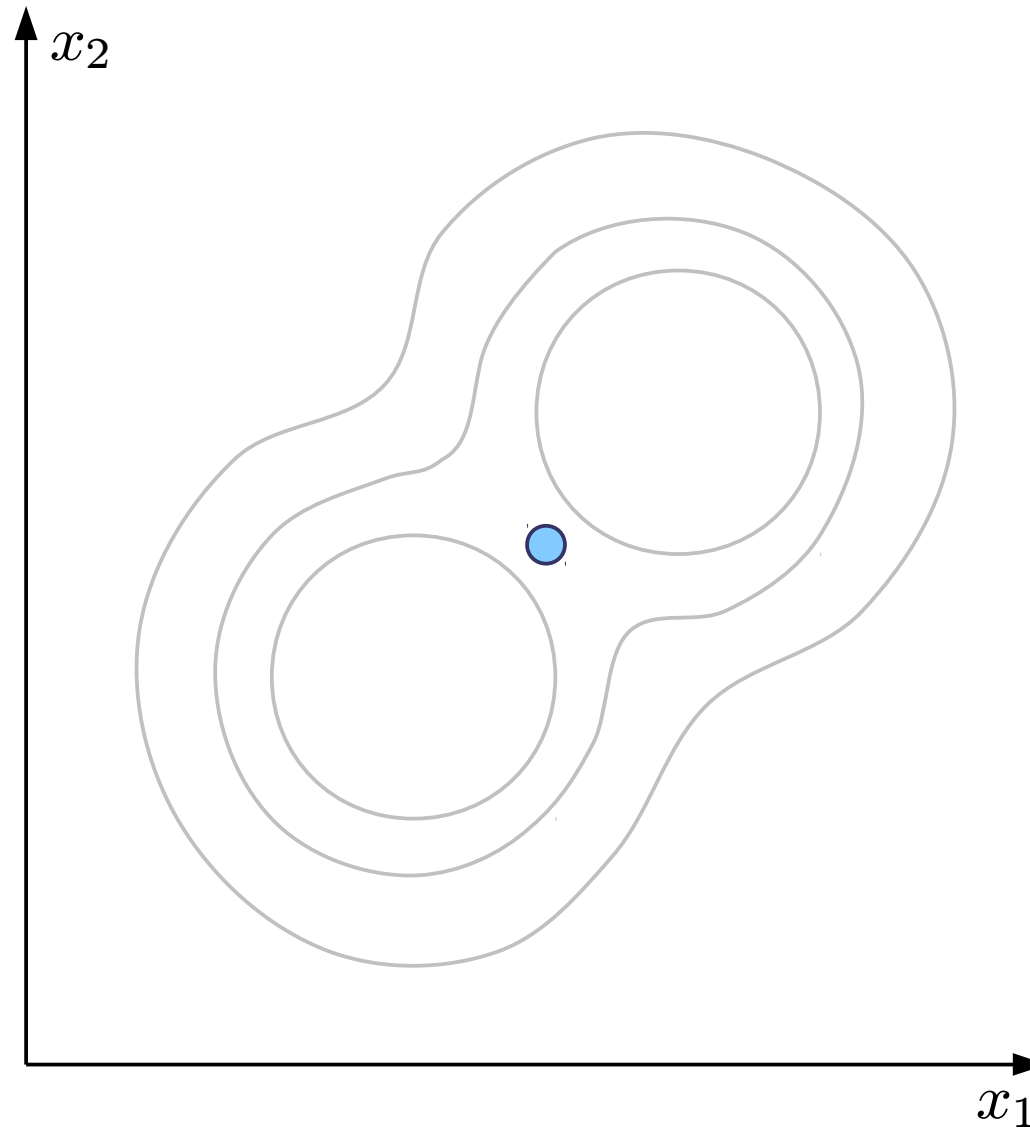
# Markov chain Monte Carlo

## Gibbs sampling



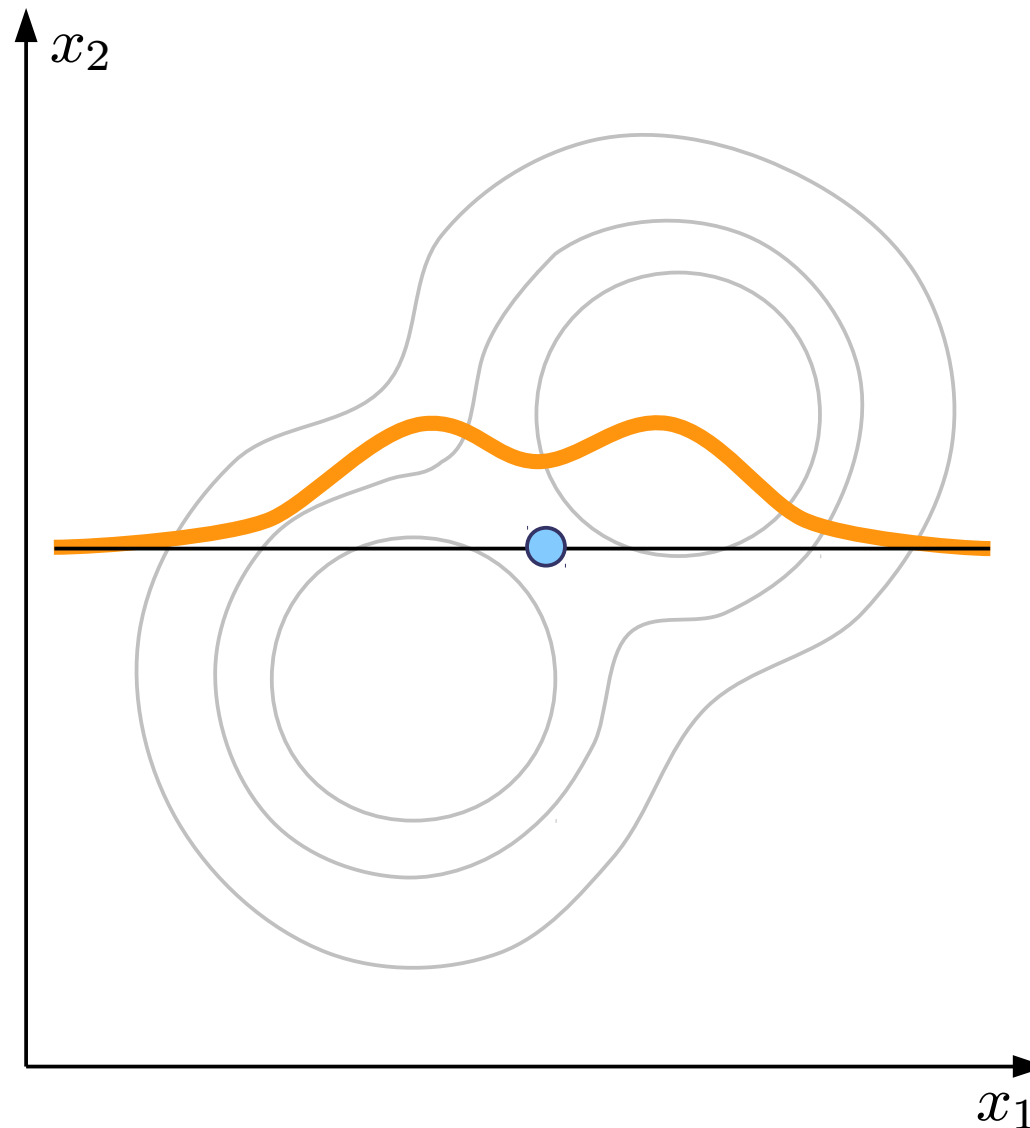
# Markov chain Monte Carlo

## Gibbs sampling



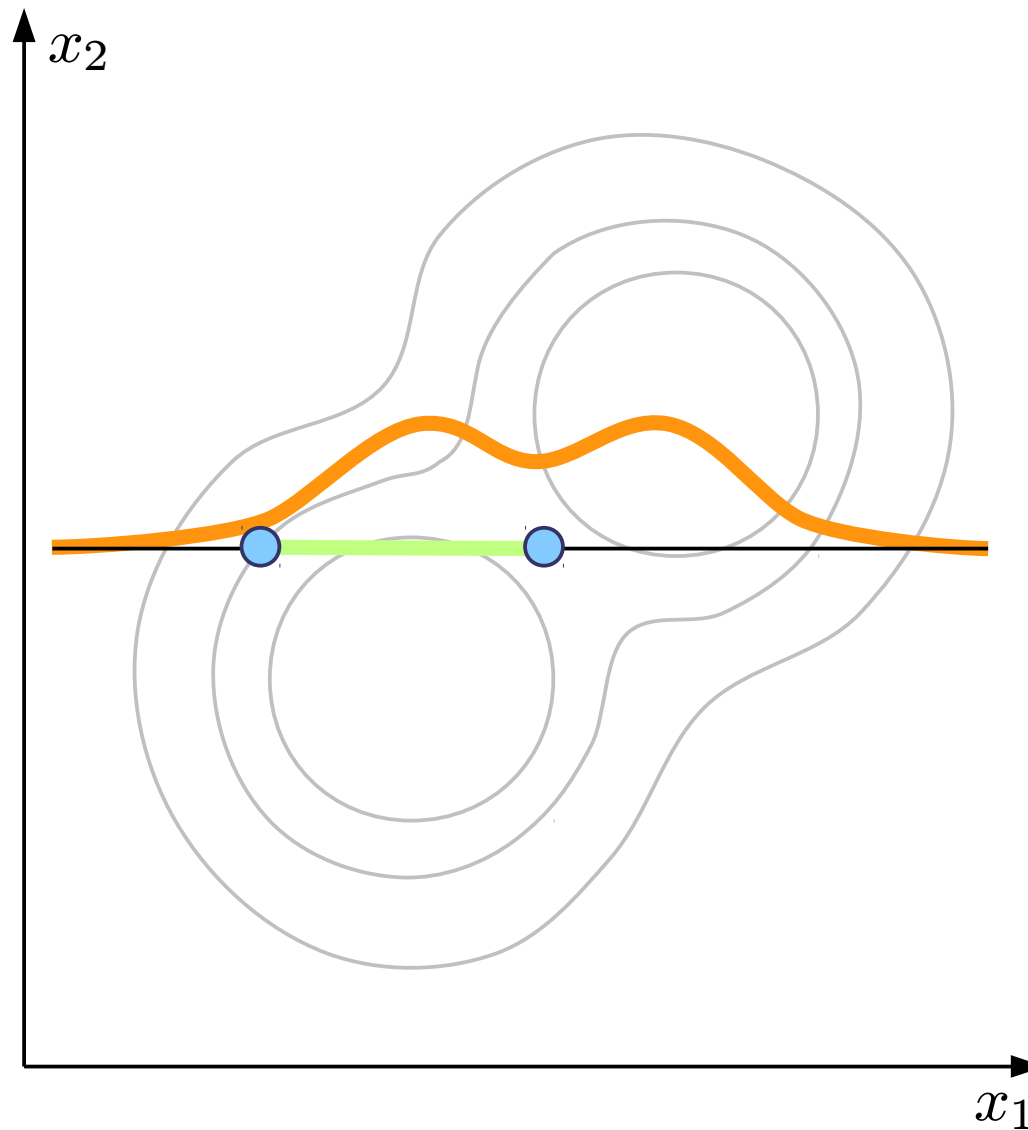
# Markov chain Monte Carlo

## Gibbs sampling



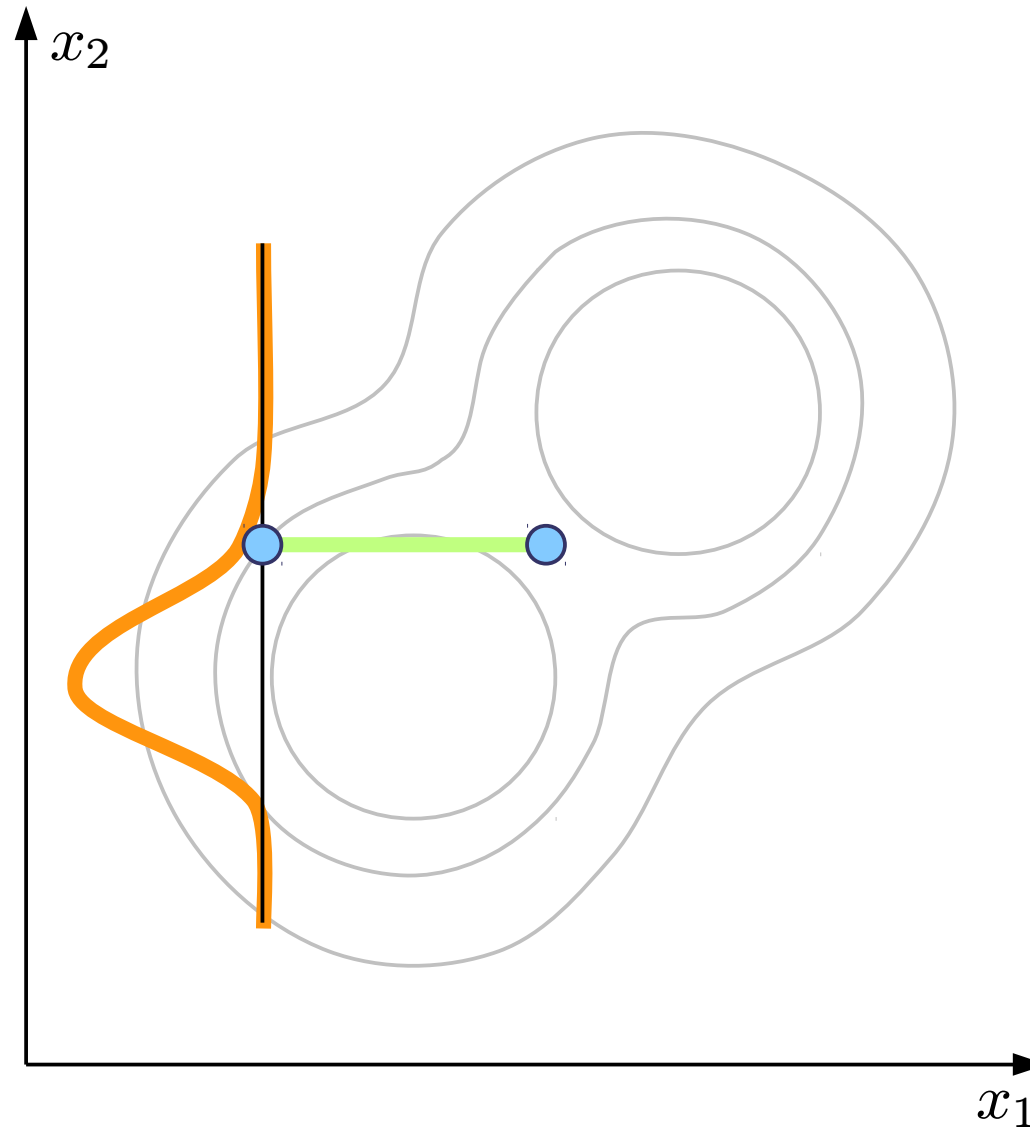
# Markov chain Monte Carlo

## Gibbs sampling



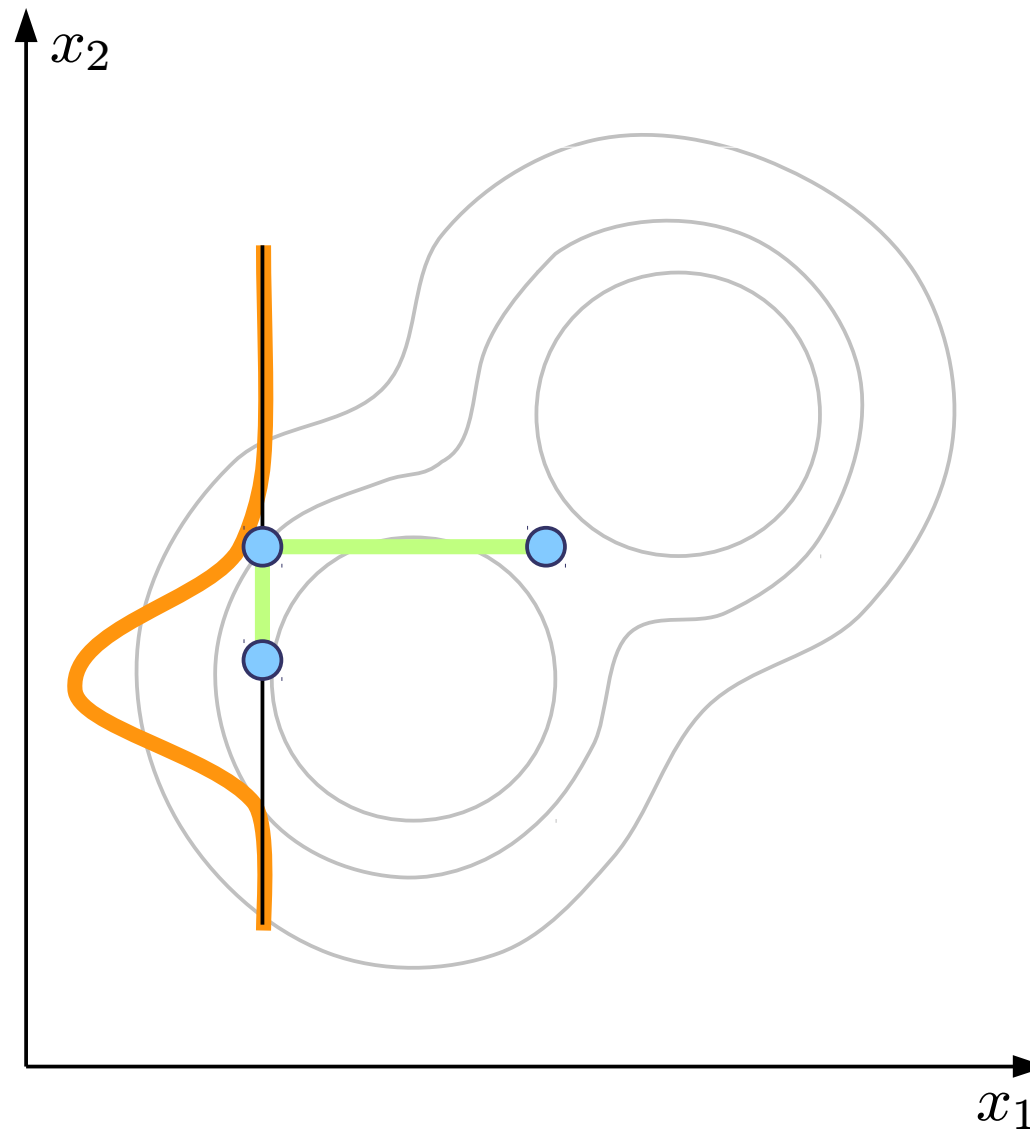
# Markov chain Monte Carlo

## Gibbs sampling



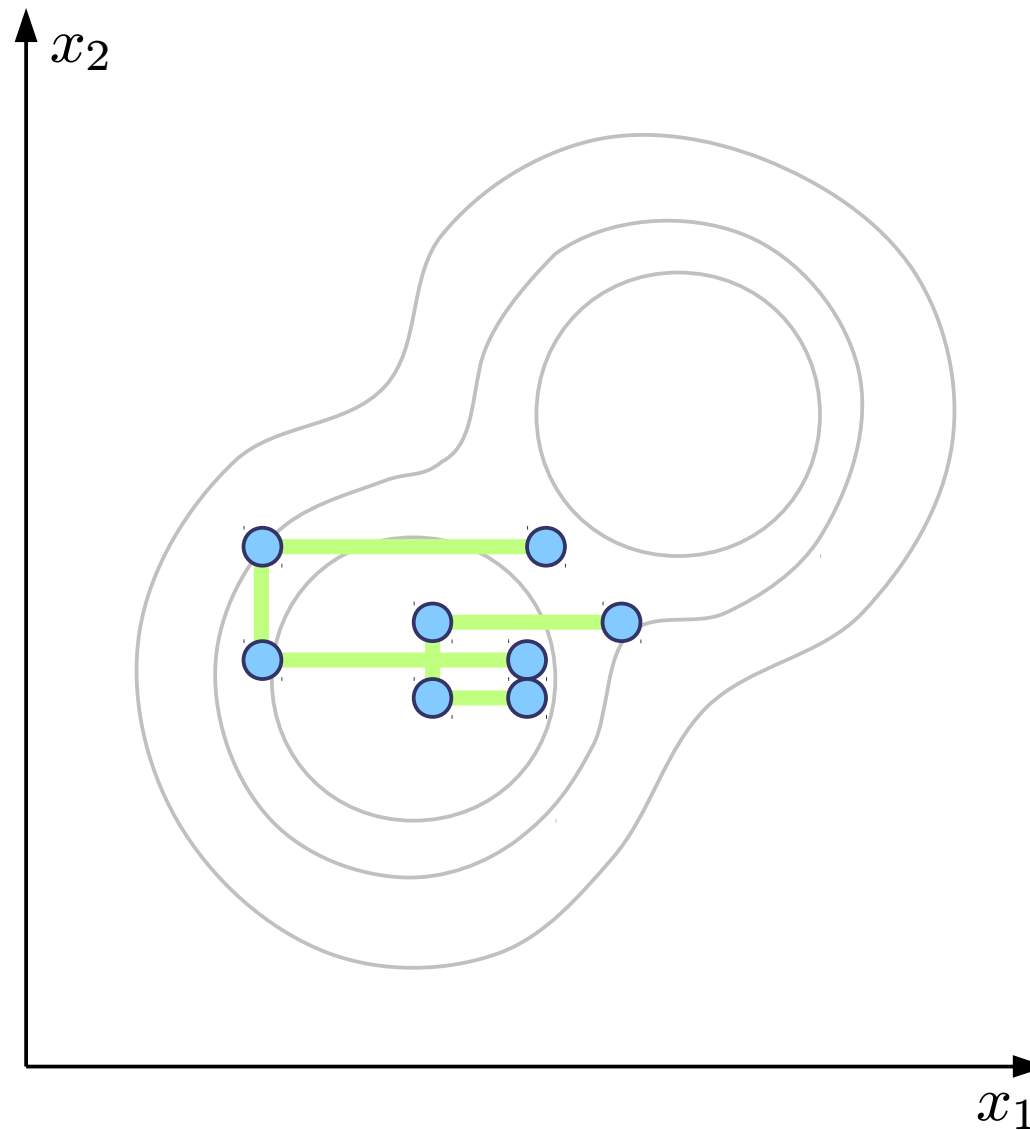
# Markov chain Monte Carlo

## Gibbs sampling



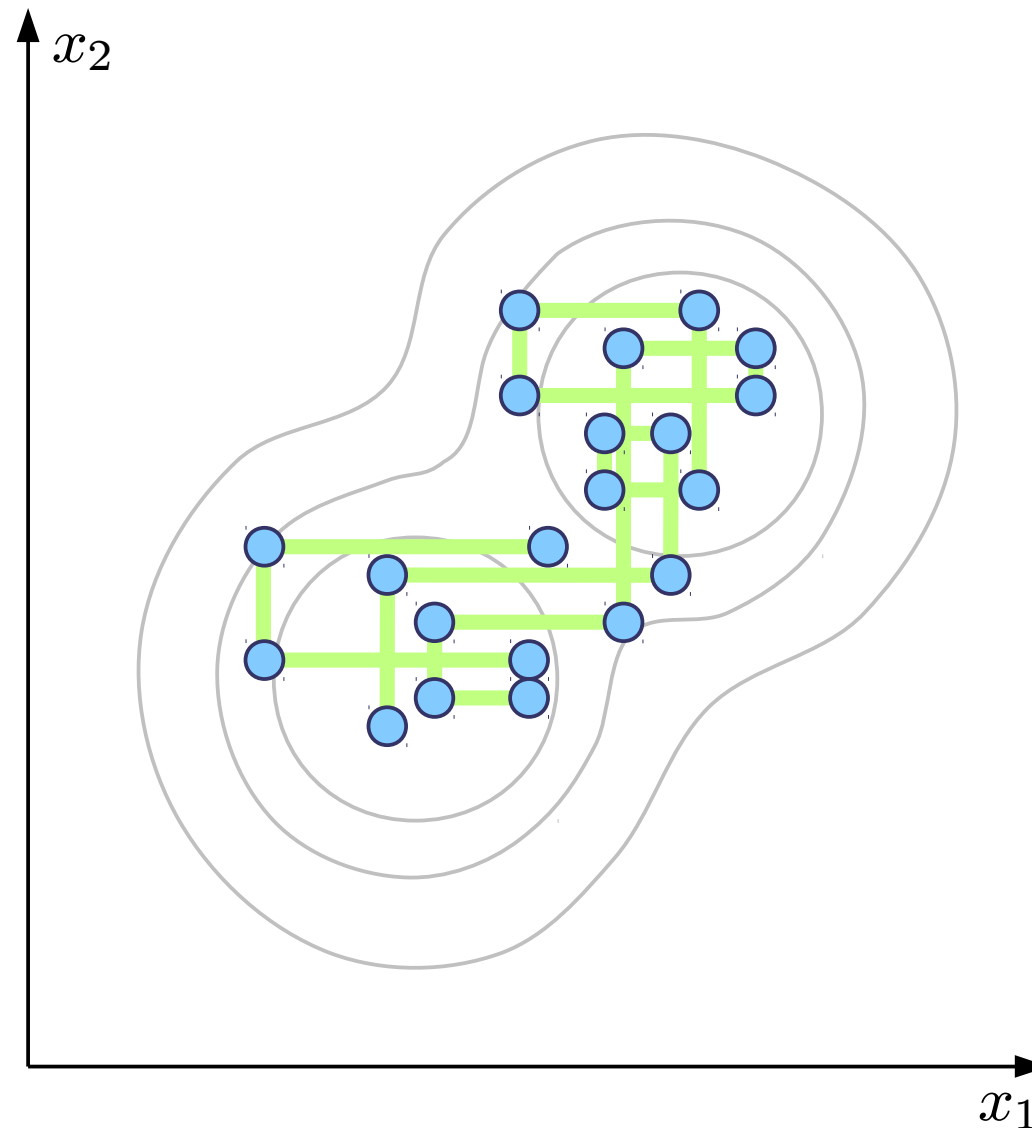
# Markov chain Monte Carlo

## Gibbs sampling



# Markov chain Monte Carlo

## Gibbs sampling



# Markov chain Monte Carlo

## Gibbs sampling

Task: Generate samples from the multivariate distribution

$$p(\mathbf{x}) = p(x_1, x_2, \dots, x_N)$$

- Choose a random initial point

$$\mathbf{x} = \mathbf{x}^{(0)}$$

- Sequentially sample each variable conditional on all other variables

$$x_1 \sim p(x_1 | x_2, \dots, x_N)$$

$$x_2 \sim p(x_2 | x_1, x_3, \dots, x_N)$$

$$\vdots$$

$$x_N \sim p(x_N | x_1, x_2, \dots, x_{N-1})$$

Converges to a sample from the desired distribution

# Markov chain Monte Carlo

## Convergence diagnostics

- Use common sense
  - Check if model predictions make sense
- Multiple chains
  - Run multiple simulations from different starting points and check if they mix and converge to the same distribution
  - Make statistical tests to verify that distributions are identical
- Autocorrelation plots
  - Plot the autocorrelation of some derived quantity to assess the speed of mixing

# Markov chain Monte Carlo

- Solve Q 4.1 - 4.3



# Variational inference

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$$\int_a^b \epsilon \Theta^{\sqrt{17}} + \Omega \int \delta e^{i\pi} = \{2.7182818284\}$$

$\infty$   $\chi^2$   $\Sigma$   $\gg$   $!$

# Variational inference

## Learning objectives

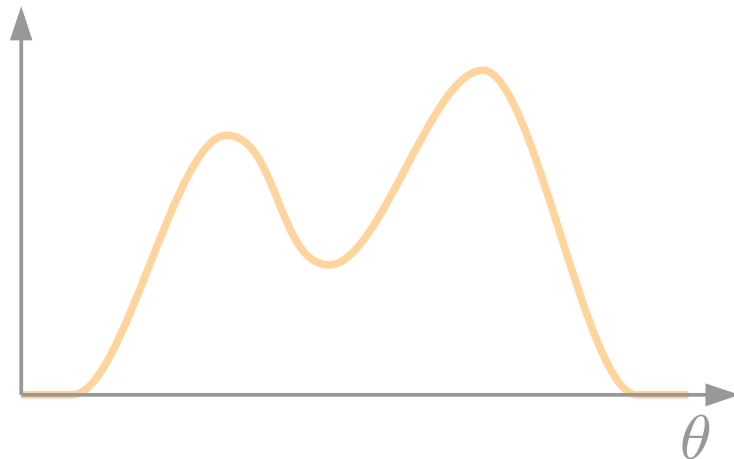
Gain knowledge about

- The basic idea behind variational inference

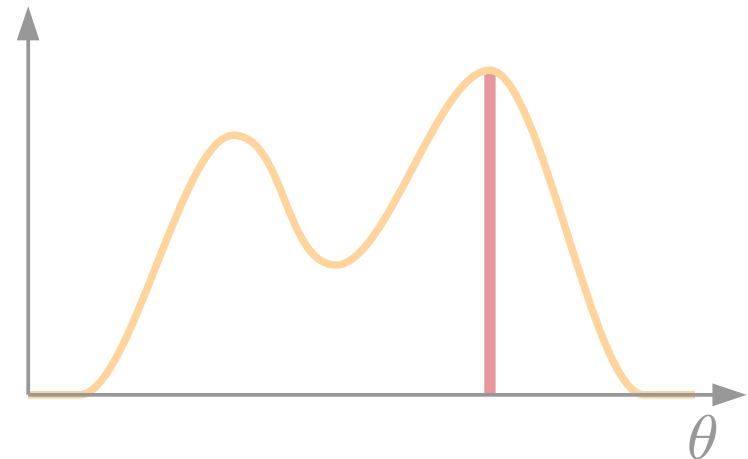
# Variational inference

(Approximate) inference procedures

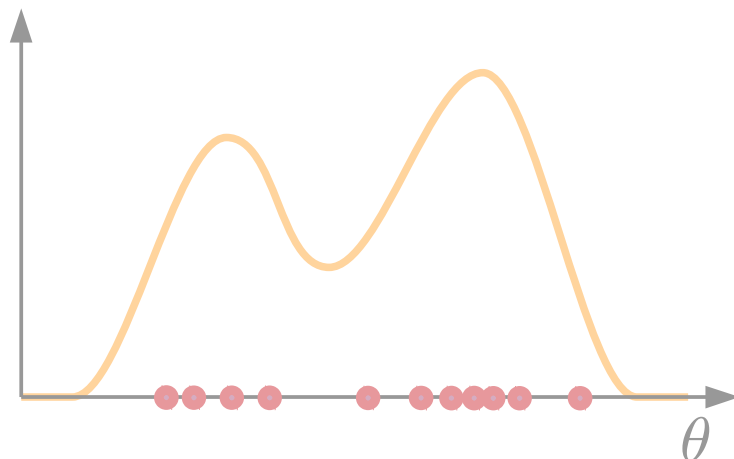
*Exact inference*



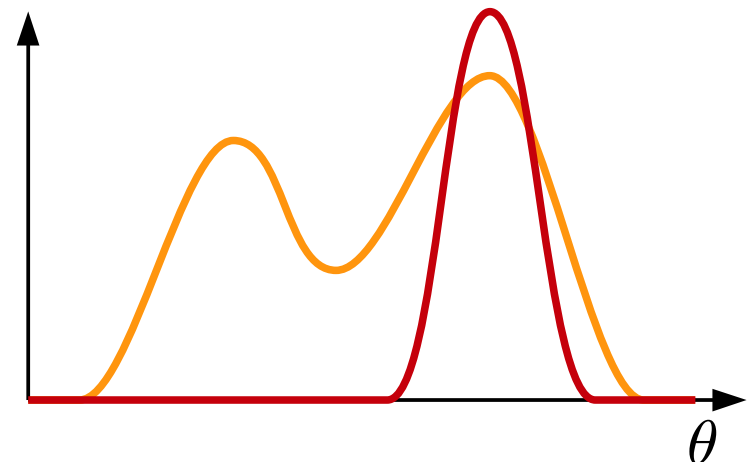
*Maximum a posteriori (MAP)*



*Monte Carlo sampling*



*Variational inference*



# Variational inference

## Minimizing Kullback-Leibler divergence

- Given a probabilistic model

$$p(\mathbf{x}, \boldsymbol{\theta})$$

- We wish to find a tractable distribution that approximates the posterior as well as possible

$$q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta}|\mathbf{x})$$

- To measure the difference between the two distributions we use the Kullback-Leibler divergence (relative entropy)

$$\text{KL}[q||p] = - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{x})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

# Variational inference

## Minimizing Kullback-Leibler divergence

- To measure the difference between the two distributions we use the Kullback-Leibler divergence (relative entropy)

$$\text{KL}[q||p] = - \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{x})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

- We could e.g. choose a distribution governed by a set of parameters

$$q(\boldsymbol{\theta}|\boldsymbol{\psi})$$

- We can use standard numerical optimization techniques to estimate the parameters

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\text{argmin}} \text{KL}[q(\boldsymbol{\theta}|\boldsymbol{\psi}) || p(\boldsymbol{\theta}|\mathbf{x})]$$

# Variational inference

## Factorized distribution

- Partition the parameters in disjoint groups

$$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \}$$

- Choose a factorized distribution (no further assumptions)

$$q(\boldsymbol{\theta}) = \prod_{k=1}^K q_k(\boldsymbol{\theta}_k)$$

- Minimize the KL divergence

$$q(\boldsymbol{\theta}) = \underset{\psi}{\operatorname{argmin}} \operatorname{KL} [q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{x})]$$

- Conditioned all other groups, the optimum for one group is given by

$$\log q_k^*(\boldsymbol{\theta}_k) = \underset{\prod_{i \neq k} q(\boldsymbol{\theta}_i)}{\mathbb{E}} [\log p(\mathbf{x}, \boldsymbol{\theta})] + c$$

# Questions

