Machine Learning and the Physical World

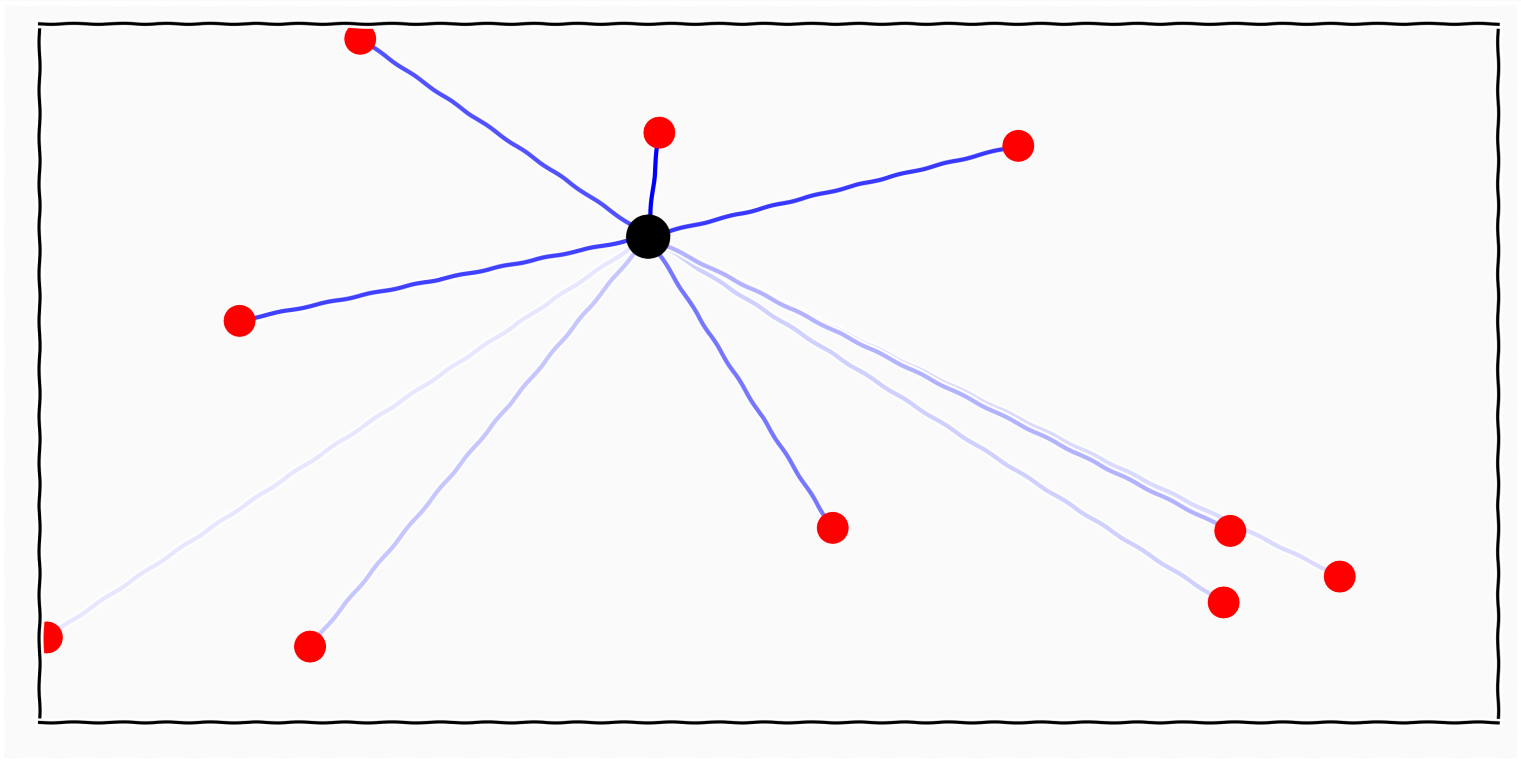
Lecture 4 : Inference and Latent Variables

Carl Henrik Ek - che29@cam.ac.uk

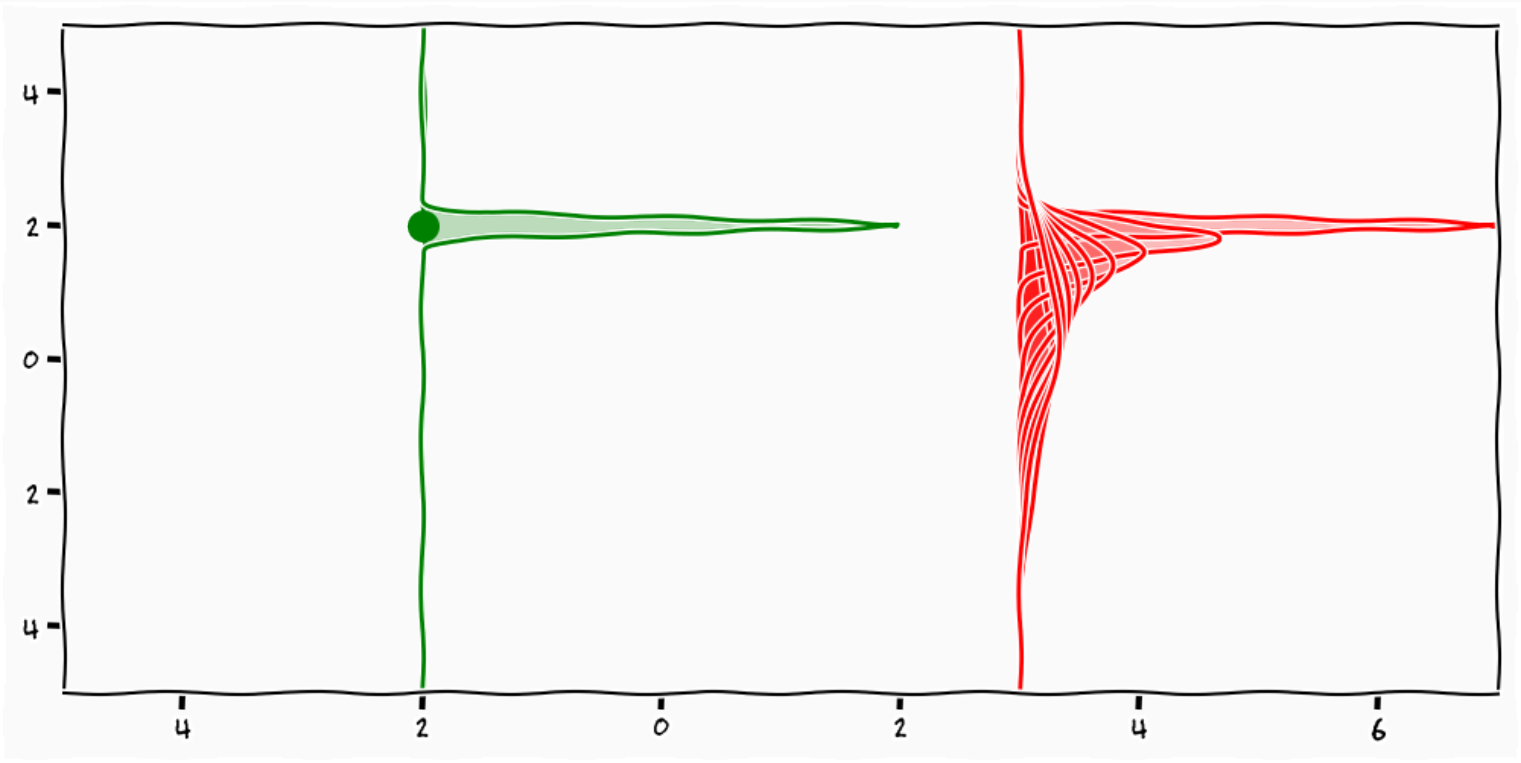
16th of October, 2020

http://carlhenrik.com

Non-parametrics



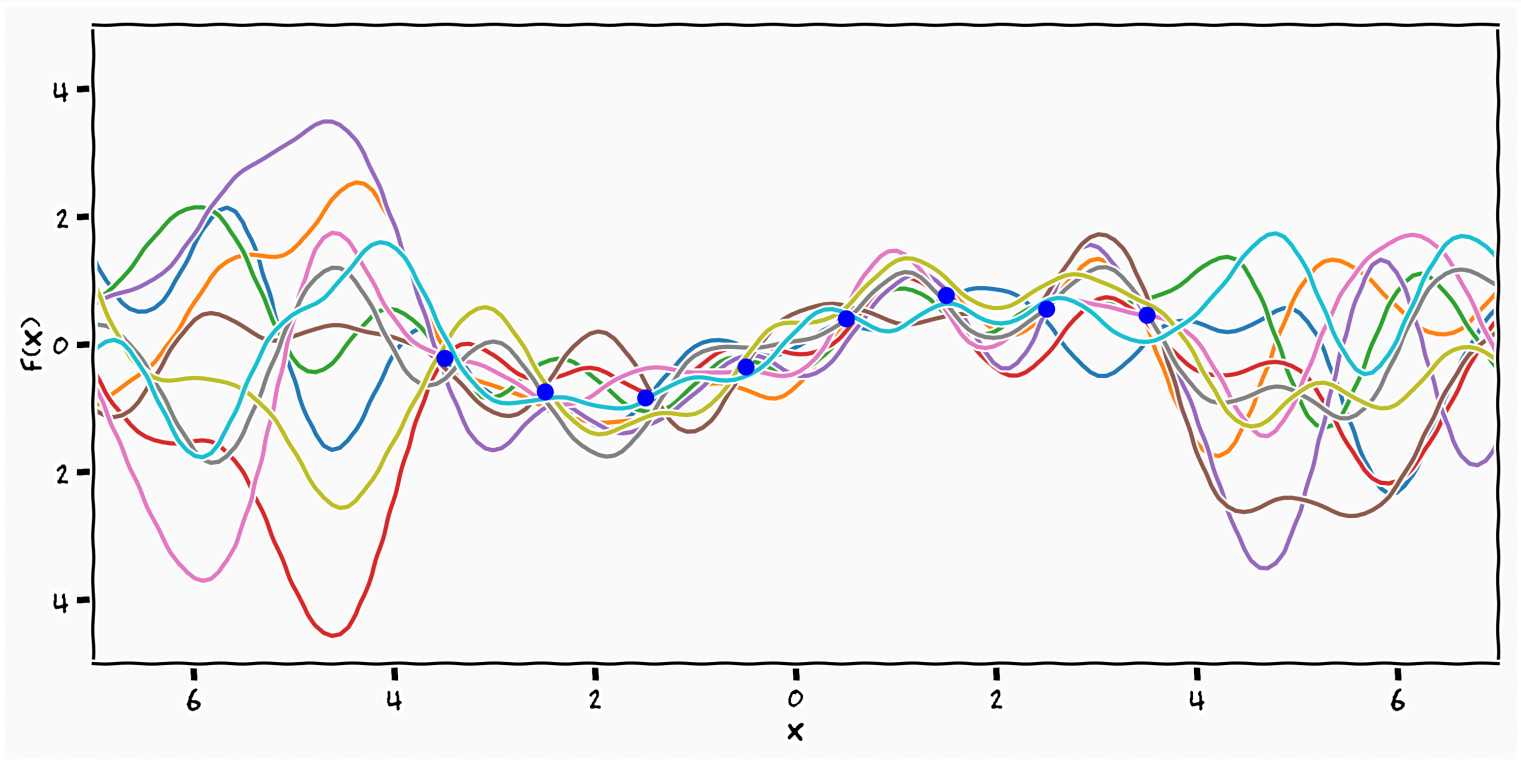
Non-parametric functions



Gaussian Processes: Formalism II

$$p(\mathbf{f}) = \mathcal{N}\left(\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \\ \vdots \end{bmatrix} \middle\| \begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \\ \mu(x_N) \\ \vdots \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \dots & k(x_1,x_N) & \dots \\ k(x_2,x_1) & k(x_2,x_2) & \dots & k(x_2,x_N) & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ k(x_N,x_1) & k(x_N,x_2) & \dots & k(x_N,x_N) & \dots \\ \vdots & \vdots & \dots & \vdots & \ddots \end{bmatrix}\right)$$

Posterior Samples



Gaussian processes

$$\mathcal{GP}(\cdot,\cdot) \qquad M \in \mathbb{R}^{\infty \times N} \qquad \mathcal{N}(\cdot,\cdot)$$

$$\infty \qquad \rightarrow \qquad N$$

The Gaussian distribution is the projection of the infinite Gaussian process

Predictive Posteriors

- Gaussian Process - *Non-parametric formulation*

$$
\begin{aligned}
p(y_* \mid \mathbf{y}, \mathbf{x}_*, \mathbf{X}, \theta) &= \mathcal{N}(y_* \mid \mu_*, K_*) \\
\mu_* &= k(x_*, x)k(x, x)^{-1}\mathbf{y} \\
K_* &= k(x_*, x_*) - k(x_*, x)k(x, x)^{-1}k(x, x_*)
\end{aligned}
$$

Predictive Posteriors

- Gaussian Process - *Non-parametric formulation*

$$
\begin{aligned}
p(y_* \mid \mathbf{y}, \mathbf{x}_*, \mathbf{X}, \theta) &= \mathcal{N}(y_* \mid \mu_*, K_*) \\
\mu_* &= k(x_*, x)k(x, x)^{-1}\mathbf{y} \\
K_* &= k(x_*, x_*) - k(x_*, x)k(x, x)^{-1}k(x, x_*)
\end{aligned}
$$

- Linear Regression - *Parametric formulation*

$$
\begin{aligned}
p(y_*|\mathbf{y}, \mathbf{x}_*, \mathbf{X}, \alpha, \beta) &= \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \alpha, \beta)\mathrm{d}\mathbf{w} \\
&= \mathcal{N}\left((y_* \mid \mu_*(x_*), \Sigma_*(x_*)\right) \\
\mu_* &= \left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}\mathbf{y}\right)^{\mathrm{T}}\mathbf{x}_* \\
&= \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y} \\
\Sigma_* &= \frac{1}{\beta} + \mathbf{x}_*^{\mathrm{T}}(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}
\end{aligned}
$$

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function
  - a basis function per data point

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function
  - a basis function per data point
- if we could parametrise $\Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x}) = k(\mathbf{x}_*, \mathbf{x})$ as a function

Regression Models: Mean

$$k(x_*, x)k(x, x)^{-1}\mathbf{y} \qquad \mathbf{x}_*^{\mathrm{T}}\mathbf{x}\left(\beta(\alpha\mathbf{I} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{x})\right)^{-1}\mathbf{y}$$

- both means are linear with respect to the output data
- how about if we take a basis function approach such that $\mathbf{x}_*^{\mathrm{T}}\mathbf{x} = \Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x})$
- if we think of $\mathbf{x}$ as the center of each basis function
  - a basis function per data point
- if we could parametrise $\Phi(\mathbf{x}_*)^{\mathrm{T}}\Phi(\mathbf{x}) = k(\mathbf{x}_*, \mathbf{x})$ as a function
- this leads to the interpretation of GPs as infinite basis functions

The Marginal Likelihood

Semantics

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int p(y \mid \theta)p(\theta)\mathrm{d}\theta}$$

Likelihood How much evidence is there in the data for a specific hypothesis

Prior What are my beliefs about different hypothesis

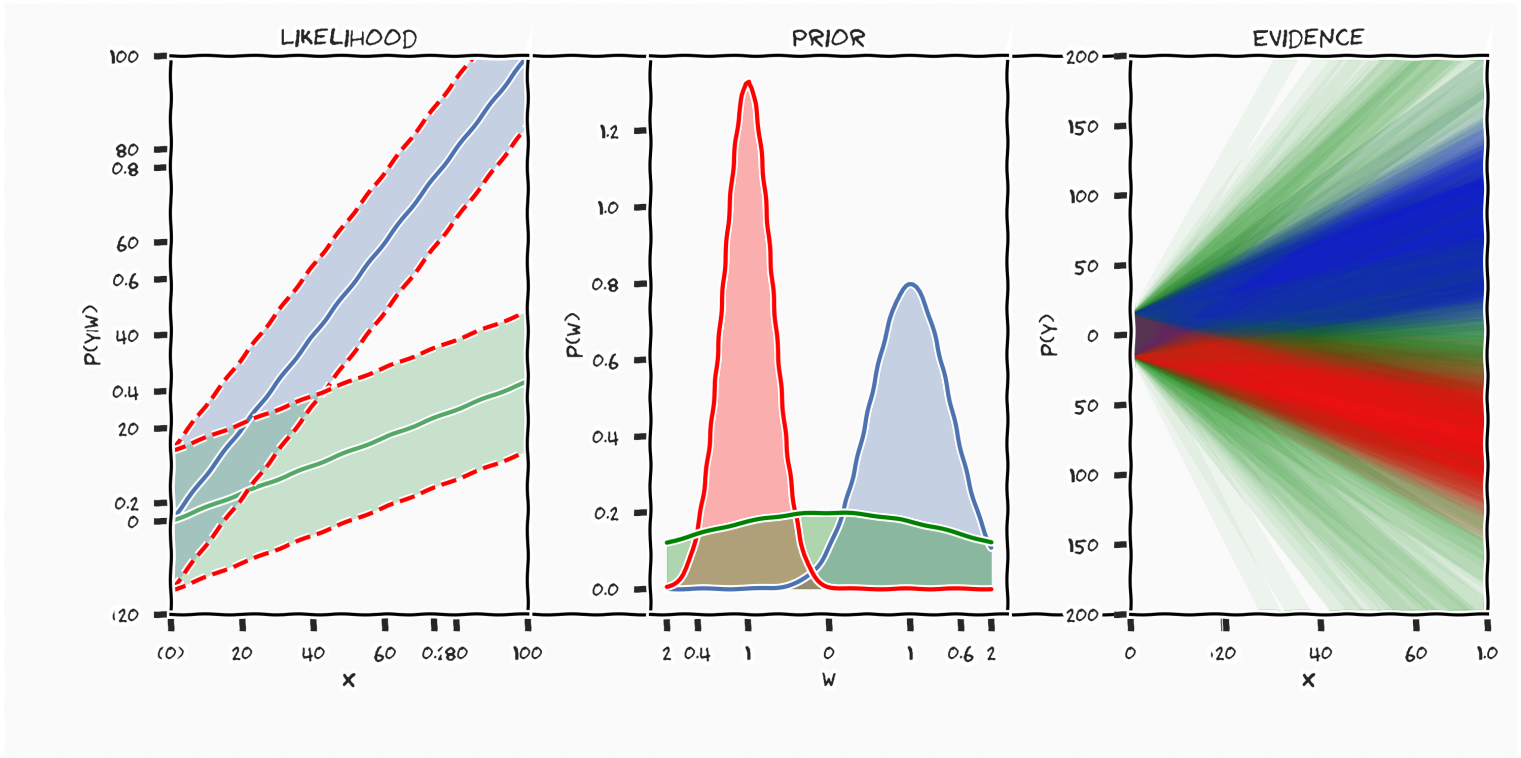
Posterior What is my updated belief after having seen data

Evidence What is my belief about the data

The Compute: Evidence

$$\int p(y \mid \theta) p(\theta) \mathrm{d}\theta$$

Regression Model



Marginalisation

Next time you want to give your friends a compliment, tell them that you have completely marginalised them from your life

Regression Models

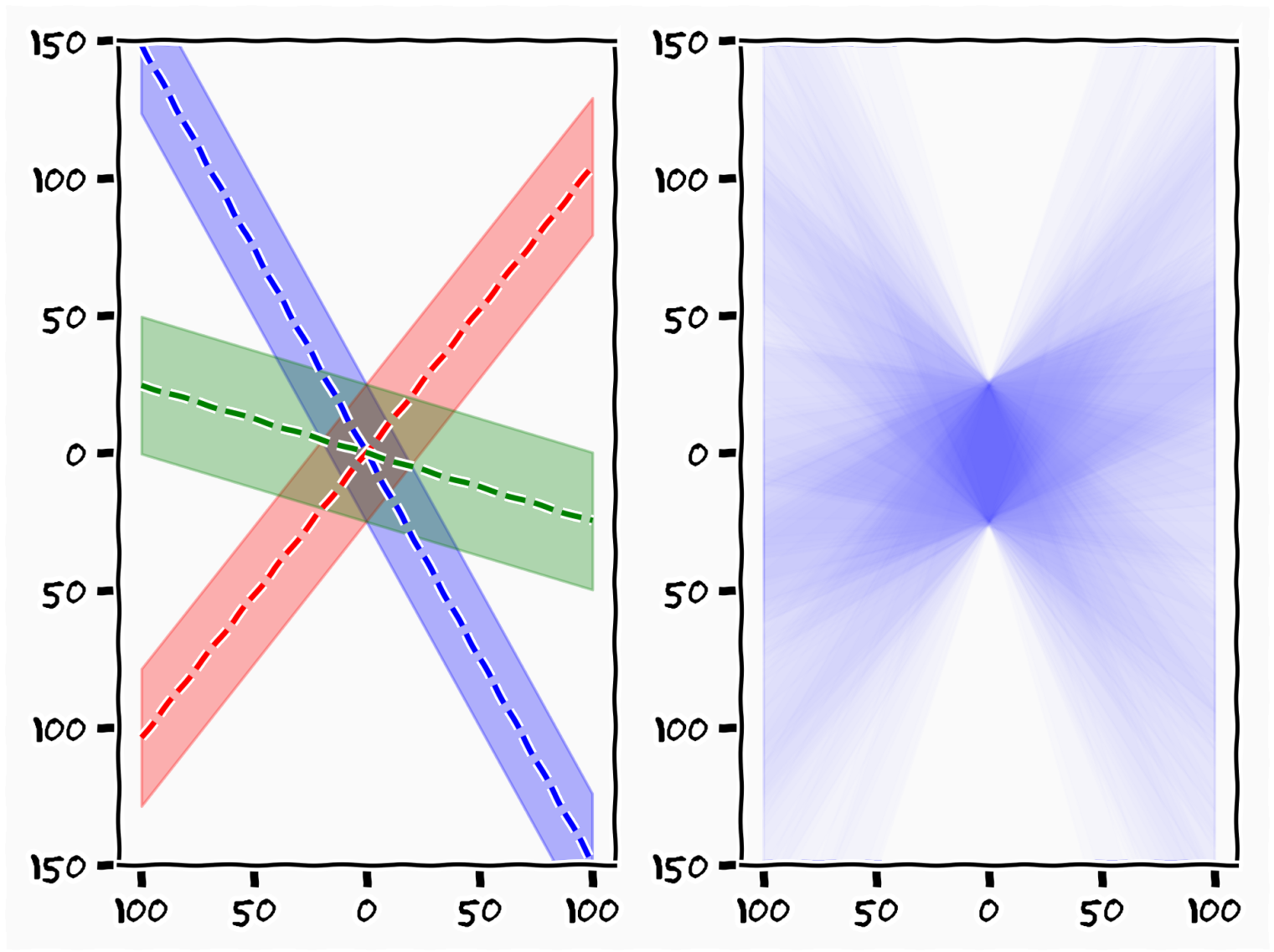
Linear Linear Model

$$p(y_i|x_i, \mathbf{w}) = \mathcal{N}(w_0 + w_1 \cdot x_i, \beta^{-1})$$

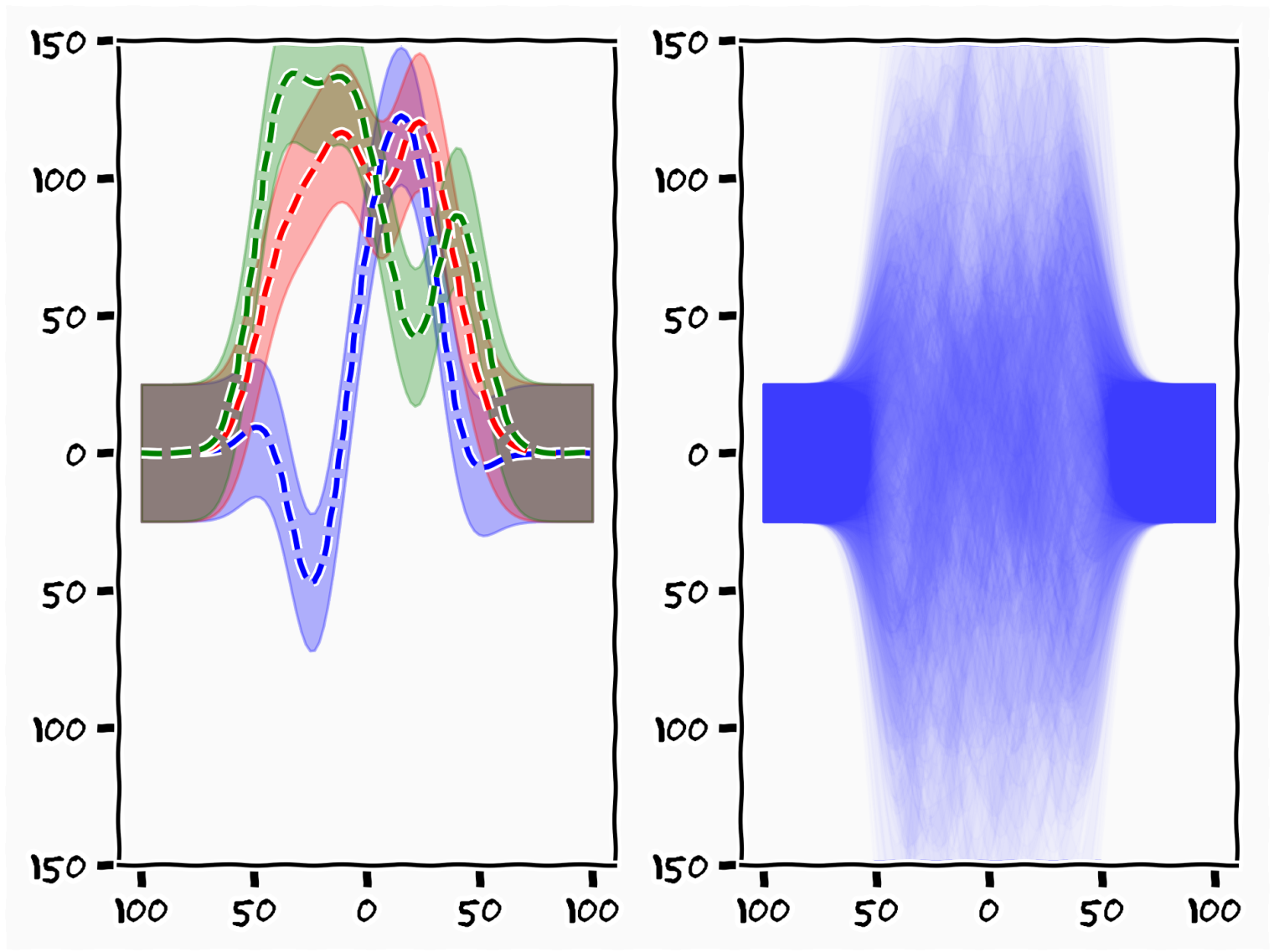
Basis function

$$p(y_i|x_i, \mathbf{w}) = \mathcal{N}(\sum_{i=1}^{6} w_i \phi(x_i), \beta^{-1})$$

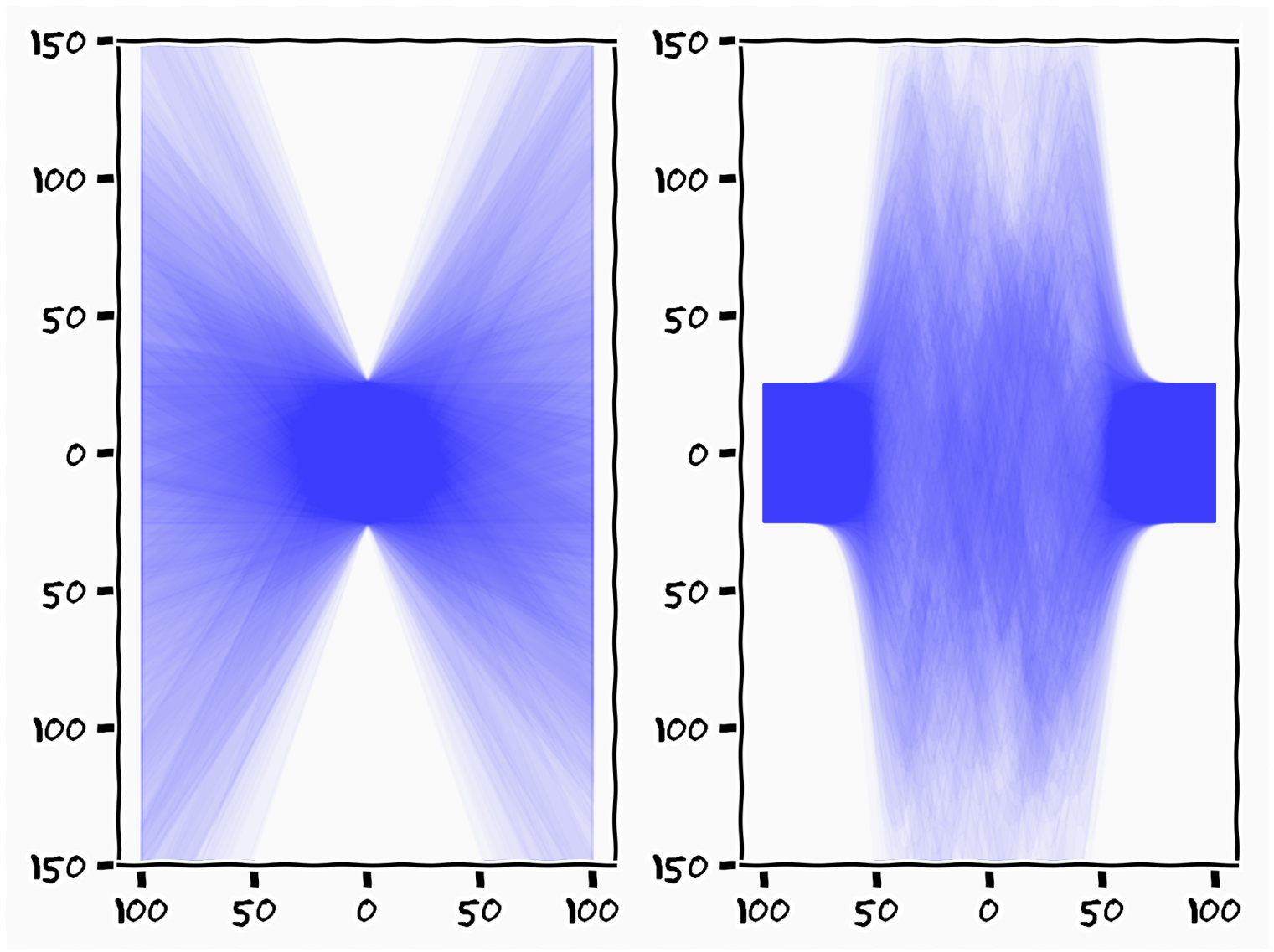
Model 1


150
100
50
0
50
100
150
100
50
0
50
100
150
100
50
0
50
100
150
100
50
0
50
100

Model 2

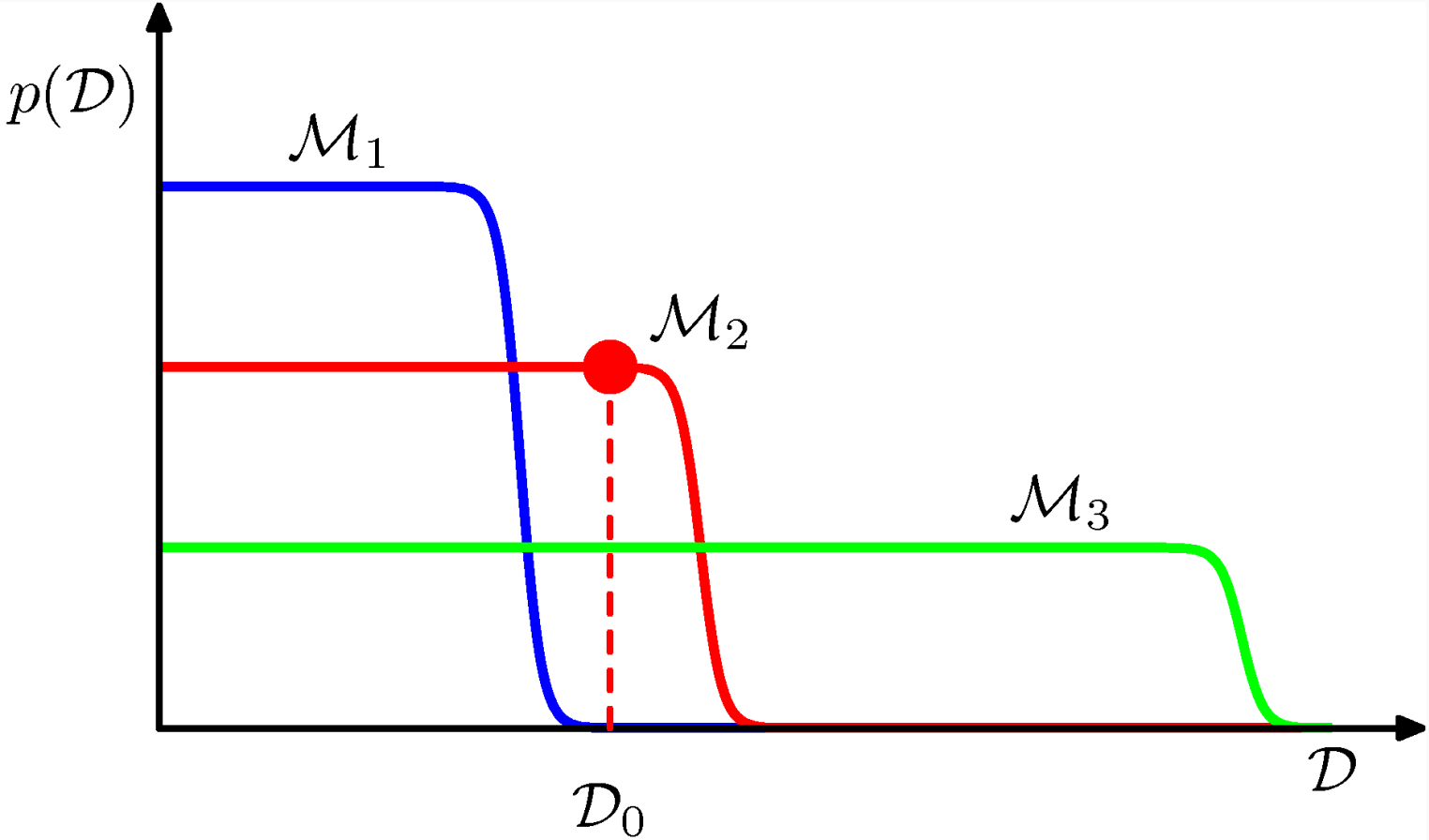

150
100
50
0
50
100
150
100
50
0
50
100
150
100
50
0
50
100
150
100
50
0
50
100

Evidence



Probabilities are a zero-sum game

Model Selection[1]



[1] David MacKay PhD Thesis

Occams Razor

Occams Razor

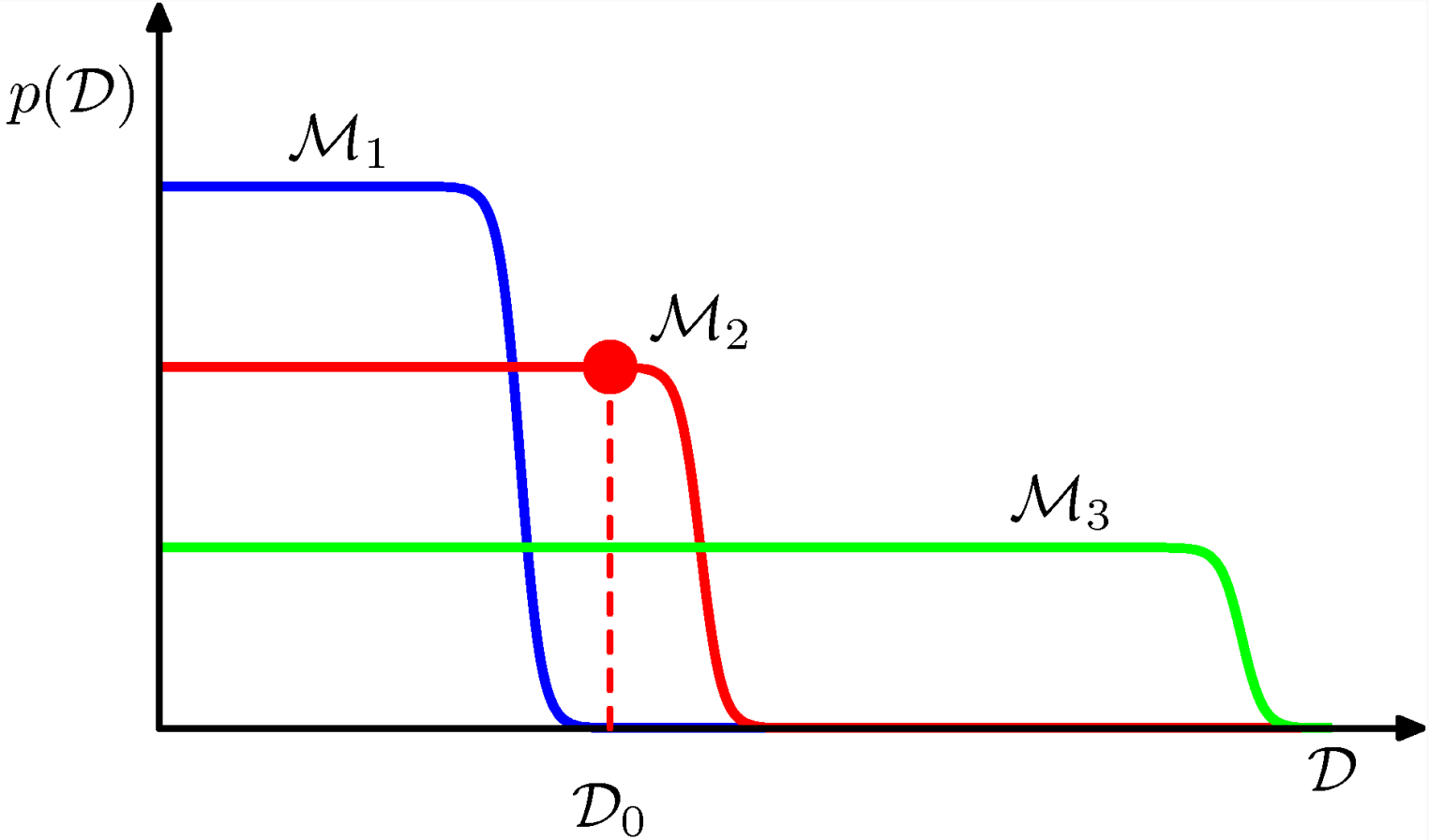
Definition (Occams Razor)
"All things being equal, the simplest solution tends to be the best one"

– William of Ockham

What is Simple?[2]

[2] https://www.imdb.com/title/tt8132700/

Model Selection[1]

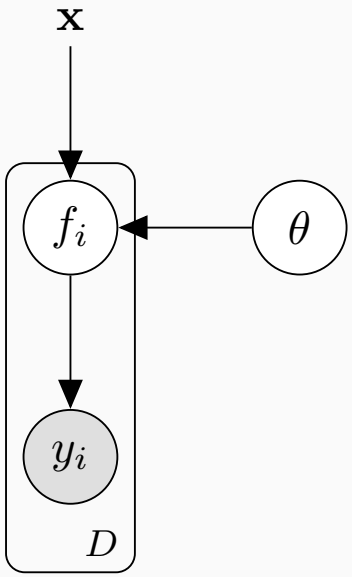


[1] David MacKay PhD Thesis

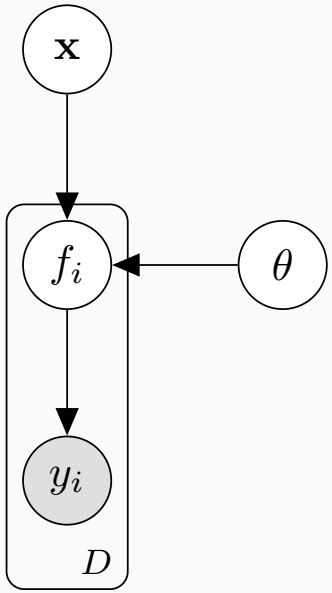
The Compute: Evidence

$$\int p(y \mid \theta) p(\theta) \mathrm{d}\theta$$

Unsupervised Learning



$p(y|x)$



$p(y)$

Gaussian Process Latent Variable Model [3]

- Regression

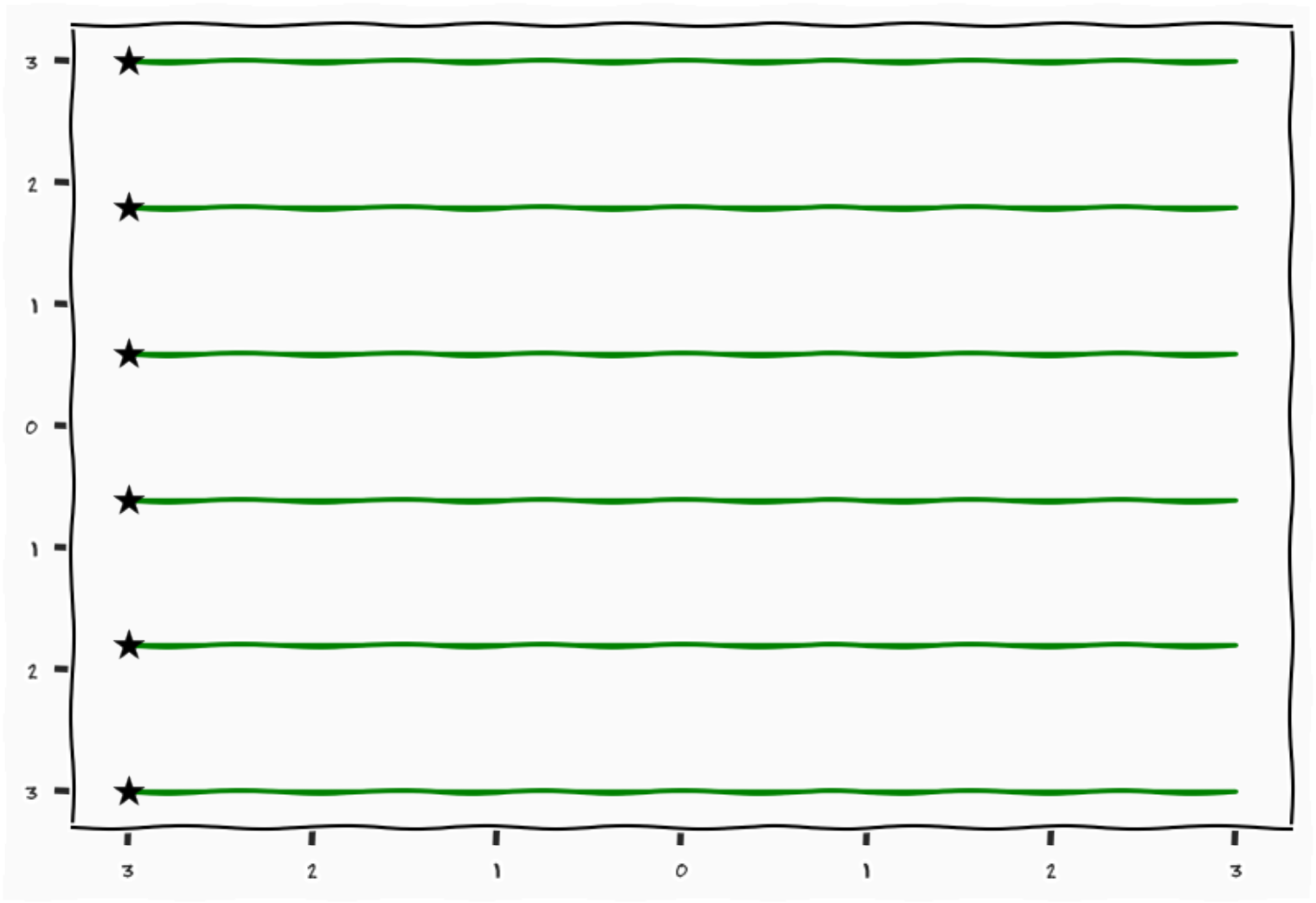
$$p(y \mid x) = \int p(y \mid f)p(f \mid x)\mathrm{d}f$$

- "Unsupervised" Learning

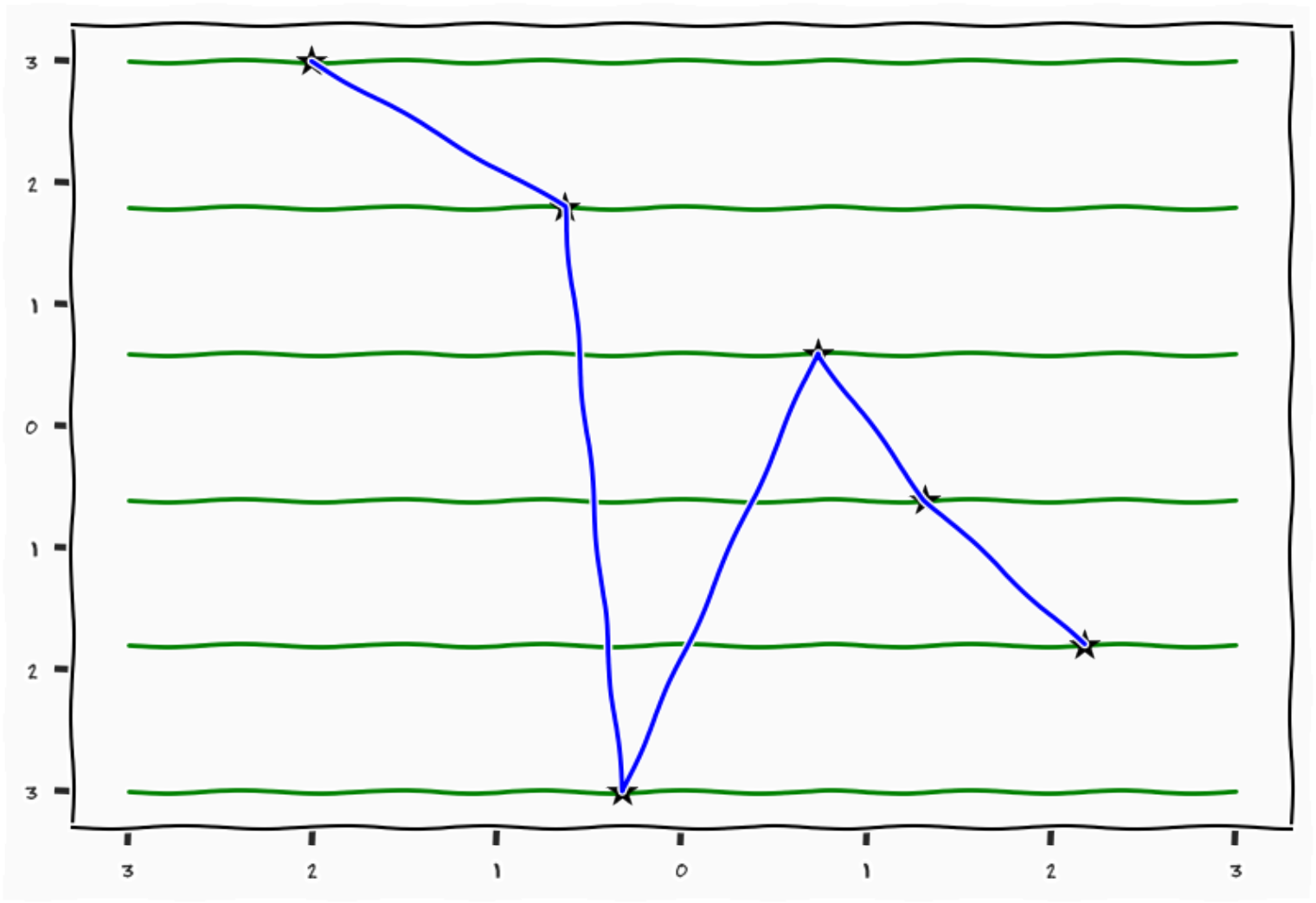
$$p(y) = \int p(y \mid f)p(f \mid x)p(x)\mathrm{d}f\mathrm{d}x$$

[3] Lawrence, 2005

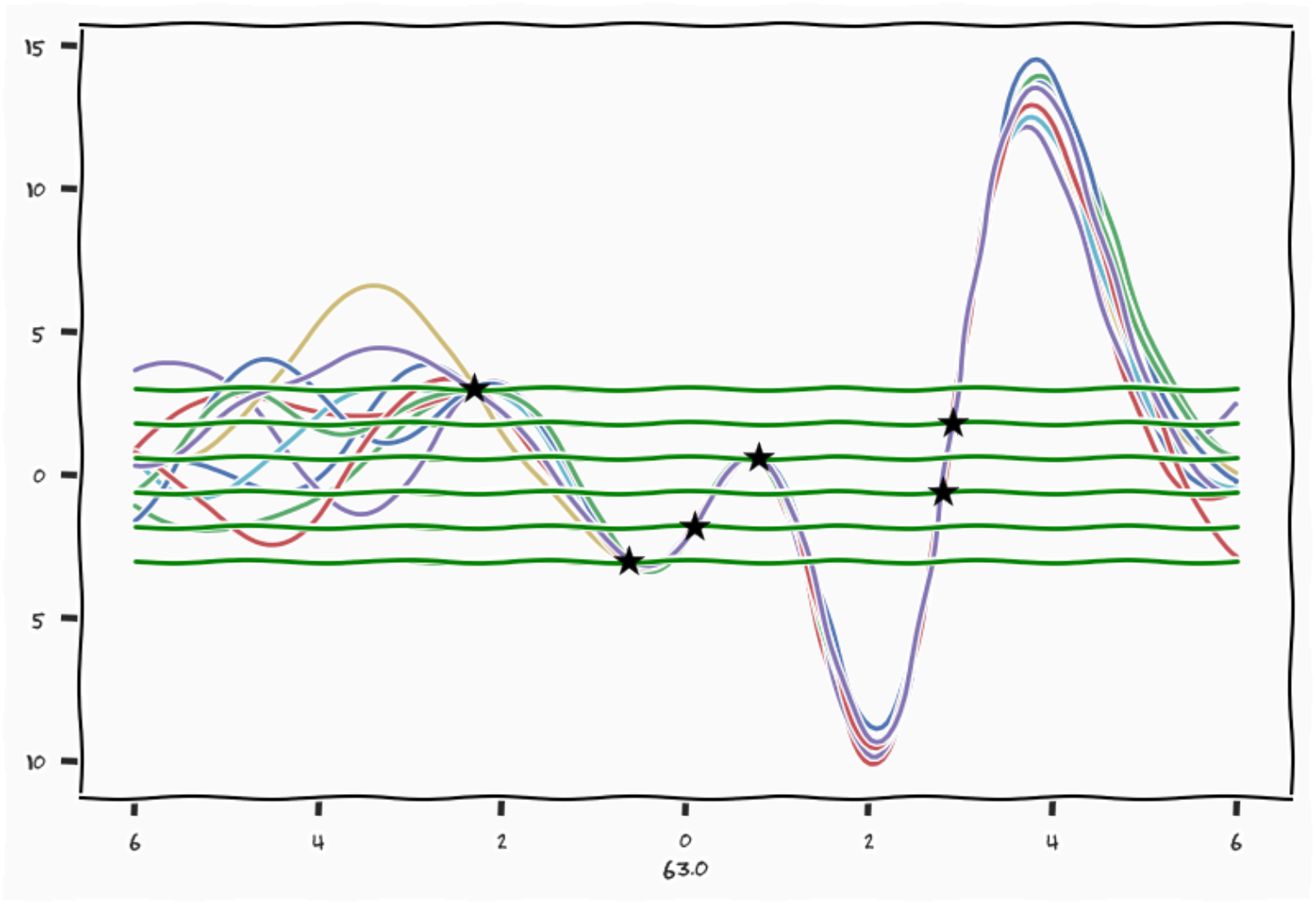
Unsupervised Learning



Unsupervised Learning

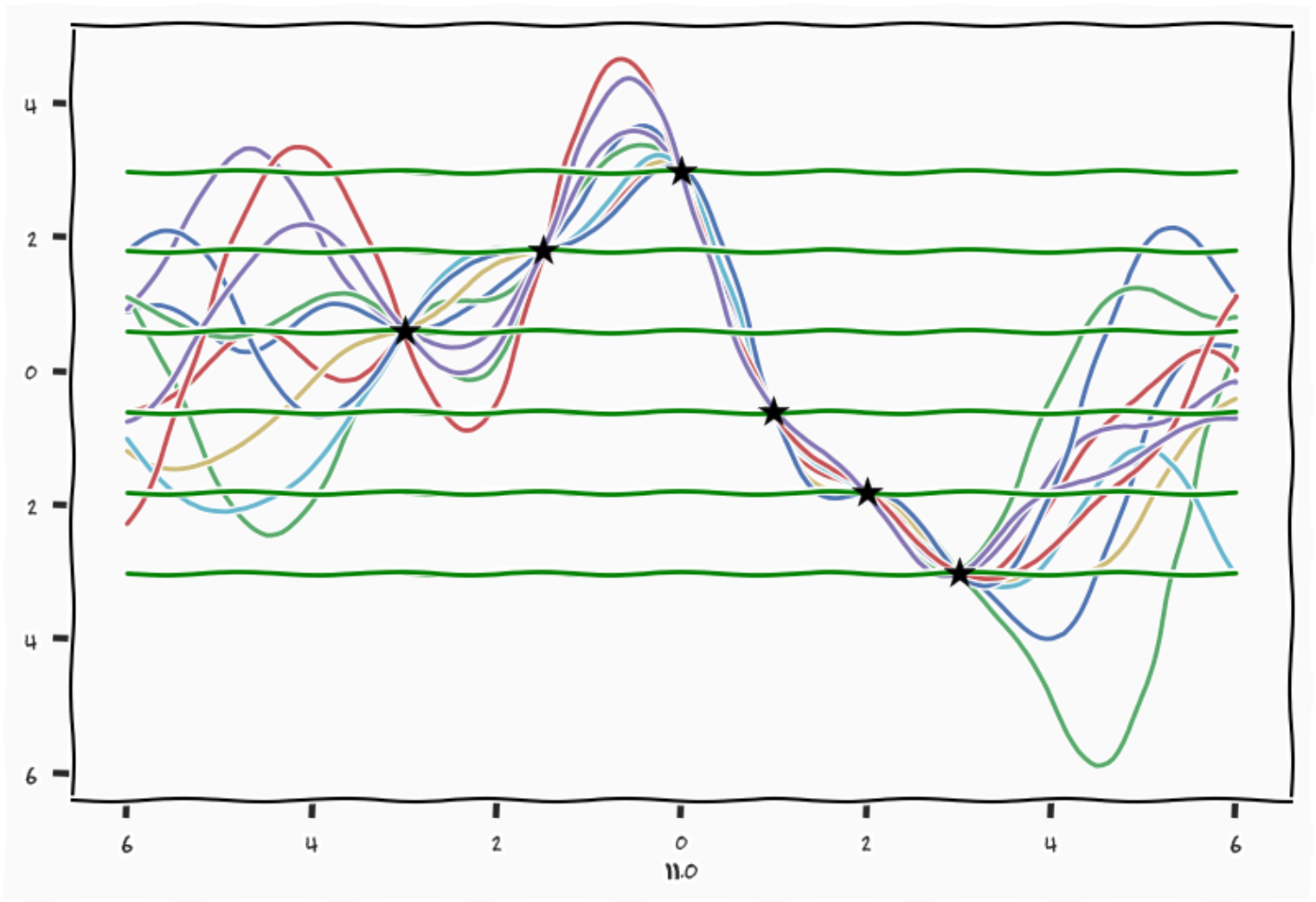


Gaussian Process Latent Variable Model [4]



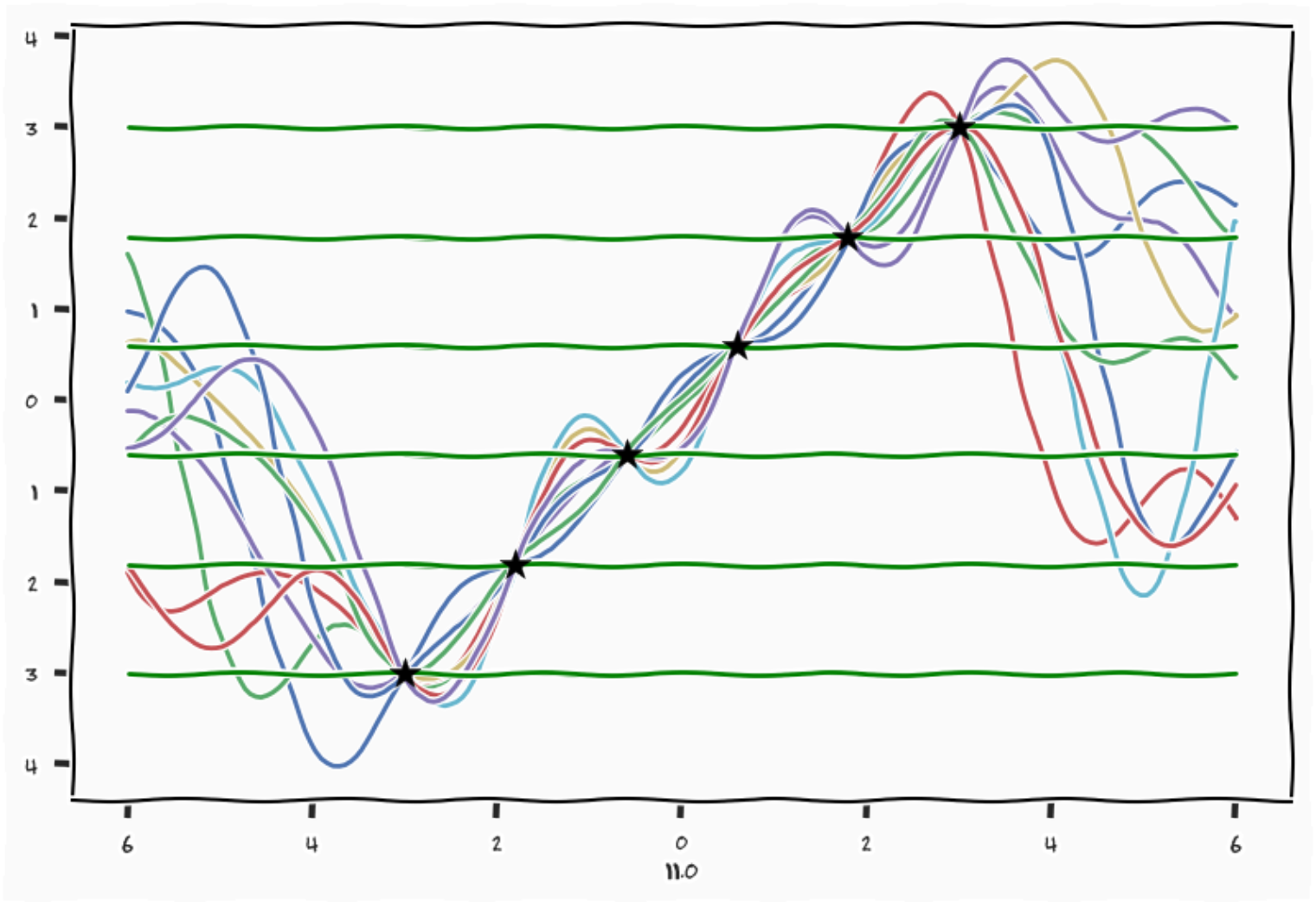
[4] Lawrence, 2005

Gaussian Process Latent Variable Model [4]



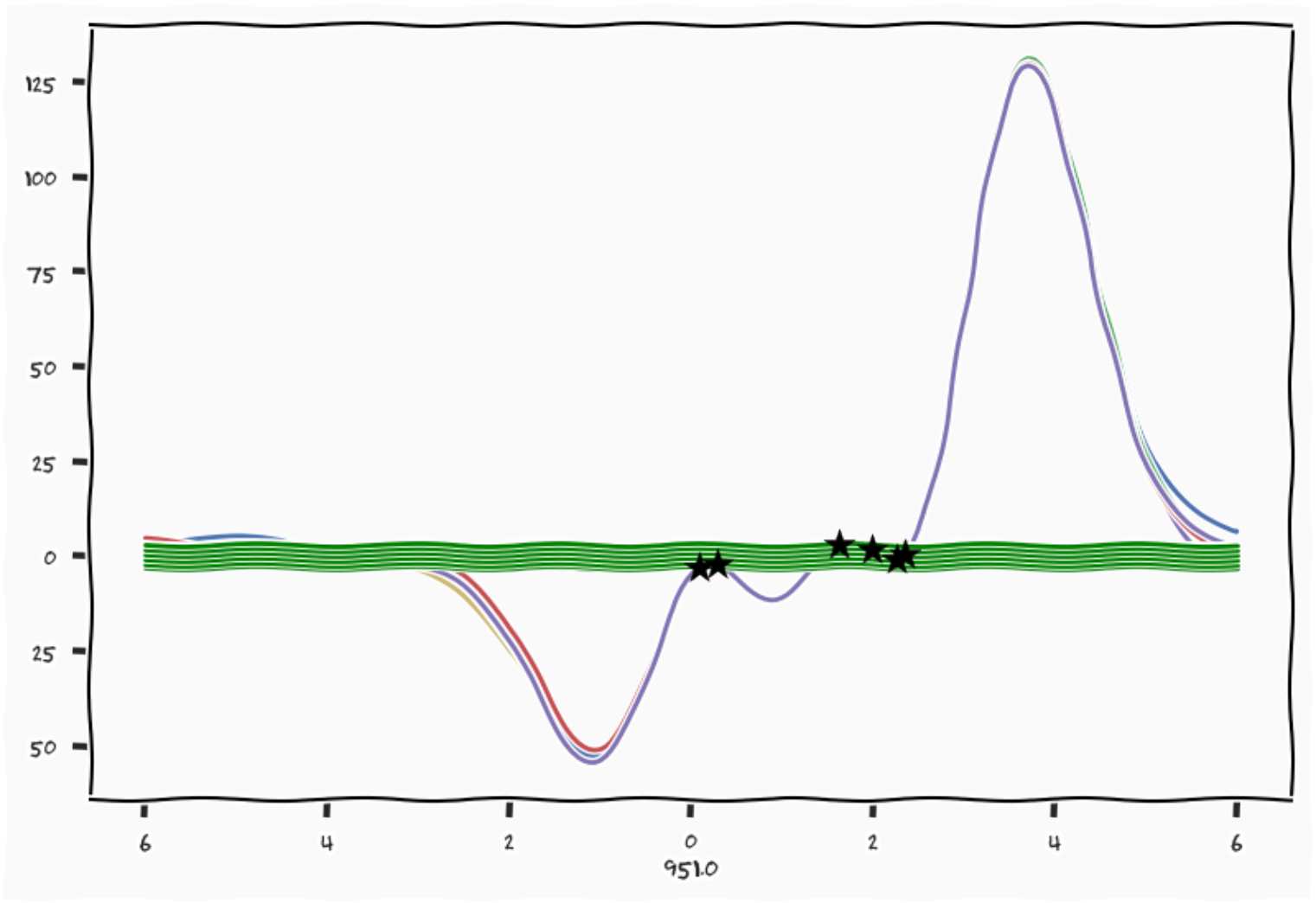
[4] Lawrence, 2005

Gaussian Process Latent Variable Model [4]



[4] Lawrence, 2005

Gaussian Process Latent Variable Model [4]



[4] Lawrence, 2005

Priors

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\mathrm{d}x$$

1. Priors that makes sense
   - **p(f)** describes our belief/assumptions and defines our notion of complexity in the function
   - **p(x)** expresses our belief/assumptions and defines our notion of complexity in the latent space
2. Now lets churn the handle

Relationship between x and data

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\mathrm{d}x$$

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

Relationship between x and data

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\mathrm{d}x$$

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

- Likelihood

$$p(y|f) \sim N(y|f, \beta) \propto e^{-\frac{1}{2\beta}\mathrm{tr}(y-f)^{\mathrm{T}}(y-f)}$$

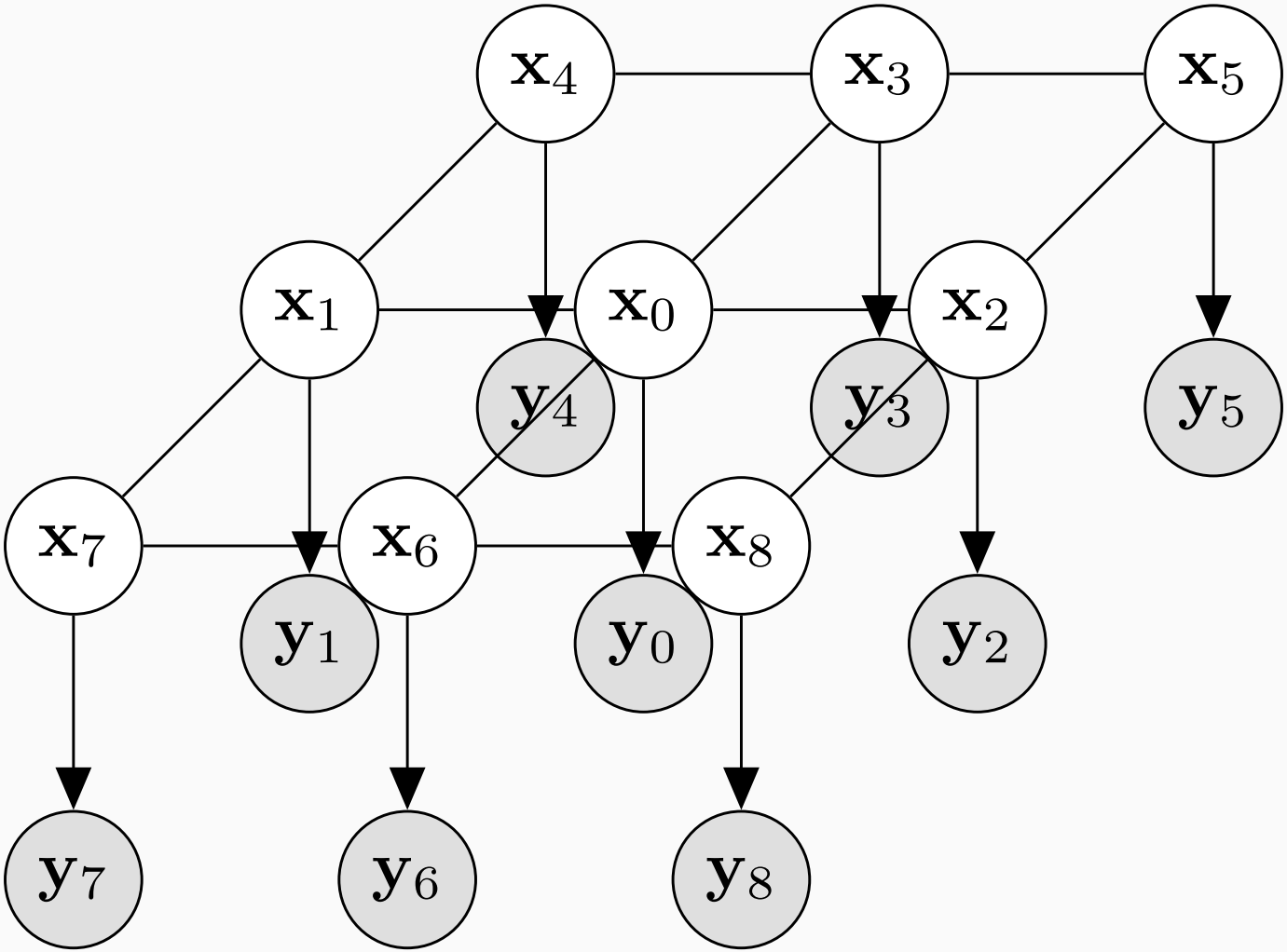
Laplace Integration

"Nature laughs at the difficulties of integrations"

– Simon Laplace

Image Segmentation

Markov Random Field



Markov Random Field

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \sum_{i}^{N} p(\mathbf{y}|\mathbf{x}_i)p(\mathbf{x}_i)$$

- $\mathbf{x}_i$ is a specific binary images

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

$$10^{80} \approx (2^{\frac{10}{3}})^{80} \approx 2^{267}$$

Numbers

- Possible black and white 3 Megapixel images

$$2^{3145728}$$

- Number of atoms in the universe

$$10^{80} \approx (2^{\frac{10}{3}})^{80} \approx 2^{267}$$

- Age of the universe in seconds

$$4.35 \cdot 10^{17} \approx 2^{59}$$

Intractability

- For most interesting problems computing the evidence is intractable
- Computational intractability: there are too many states to sum over
- Analytic: no closed form exists for the distribution

Today

- We have been living in a fairy land so far, the world is not conjugate

Today

- We have been living in a fairy land so far, the world is not conjugate
- We cannot give one lecture on approximate inference

Today

- We have been living in a fairy land so far, the world is not conjugate
- We cannot give one lecture on approximate inference
- We hope to give you a flavour of approaches

Today

- We have been living in a fairy land so far, the world is not conjugate
- We cannot give one lecture on approximate inference
- We hope to give you a flavour of approaches
    - Stochastic

Stochastic Inference

Basic Sampling

$$\mathbb{E}_{p(x)}[f] = \int f(x)p(x)\mathrm{d}x \approx \frac{1}{L}\sum_{l=1}^{L} f(x^{(l)}) = \hat{f}$$

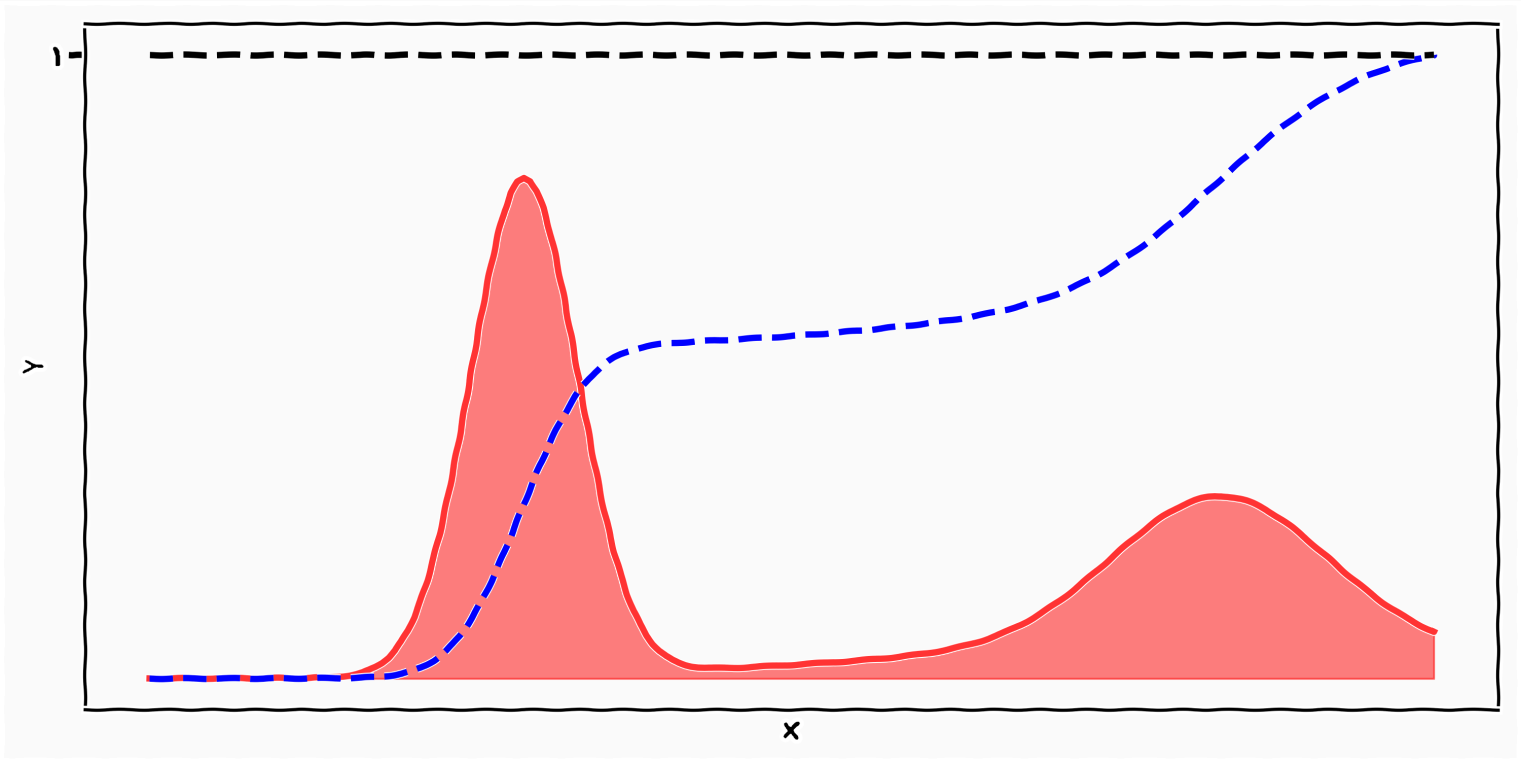
$$x^{(l)} \sim p(x)$$

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

$$\mathsf{var}[\hat{f}] = \frac{1}{L}\mathbb{E}\left[(f(x) - \mathbb{E}[f])^2\right]$$

- Approximation not dependent on dimensionality of x
- Variance of estimator shrinks with number of samples

Basic Sampling

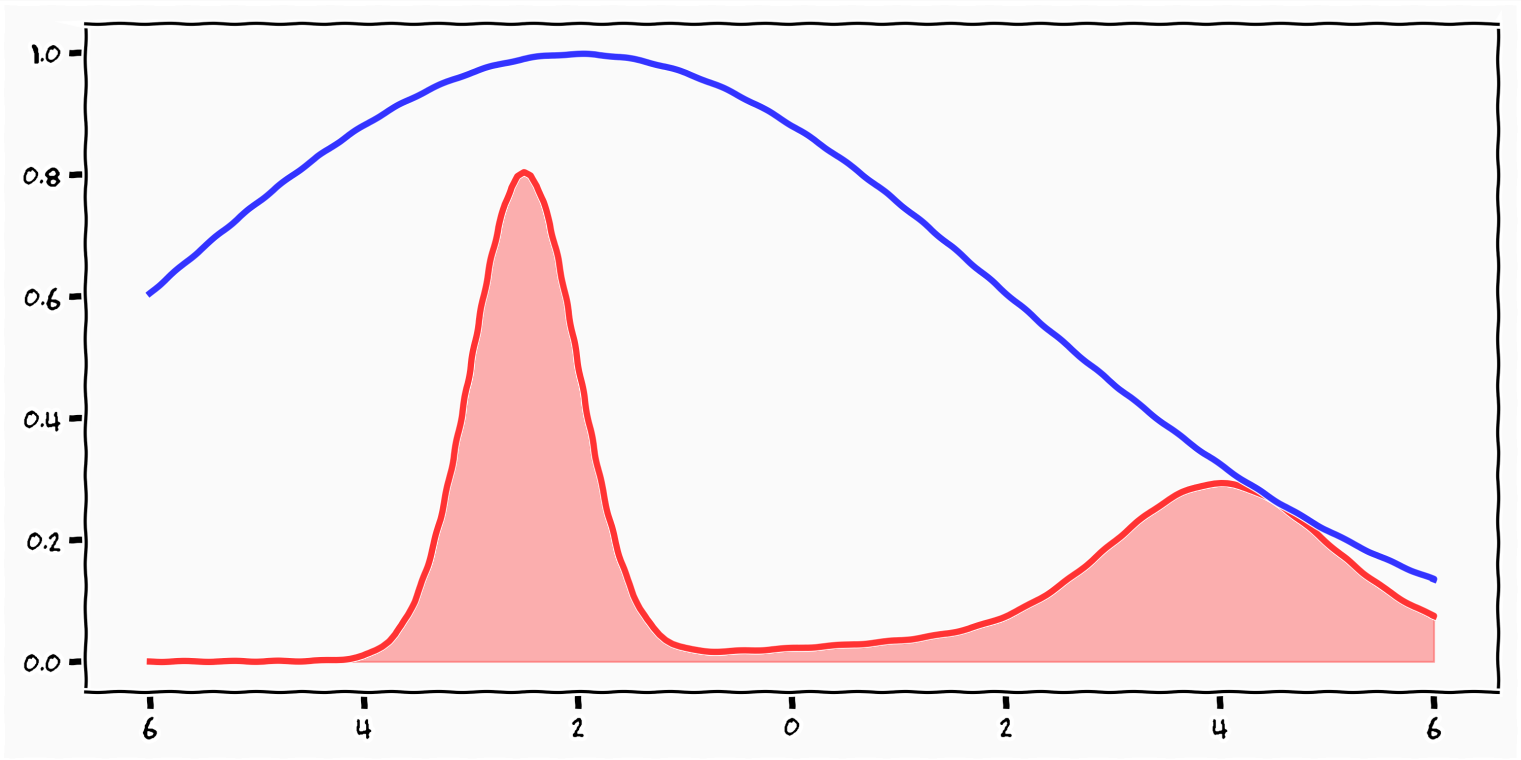


Sampling

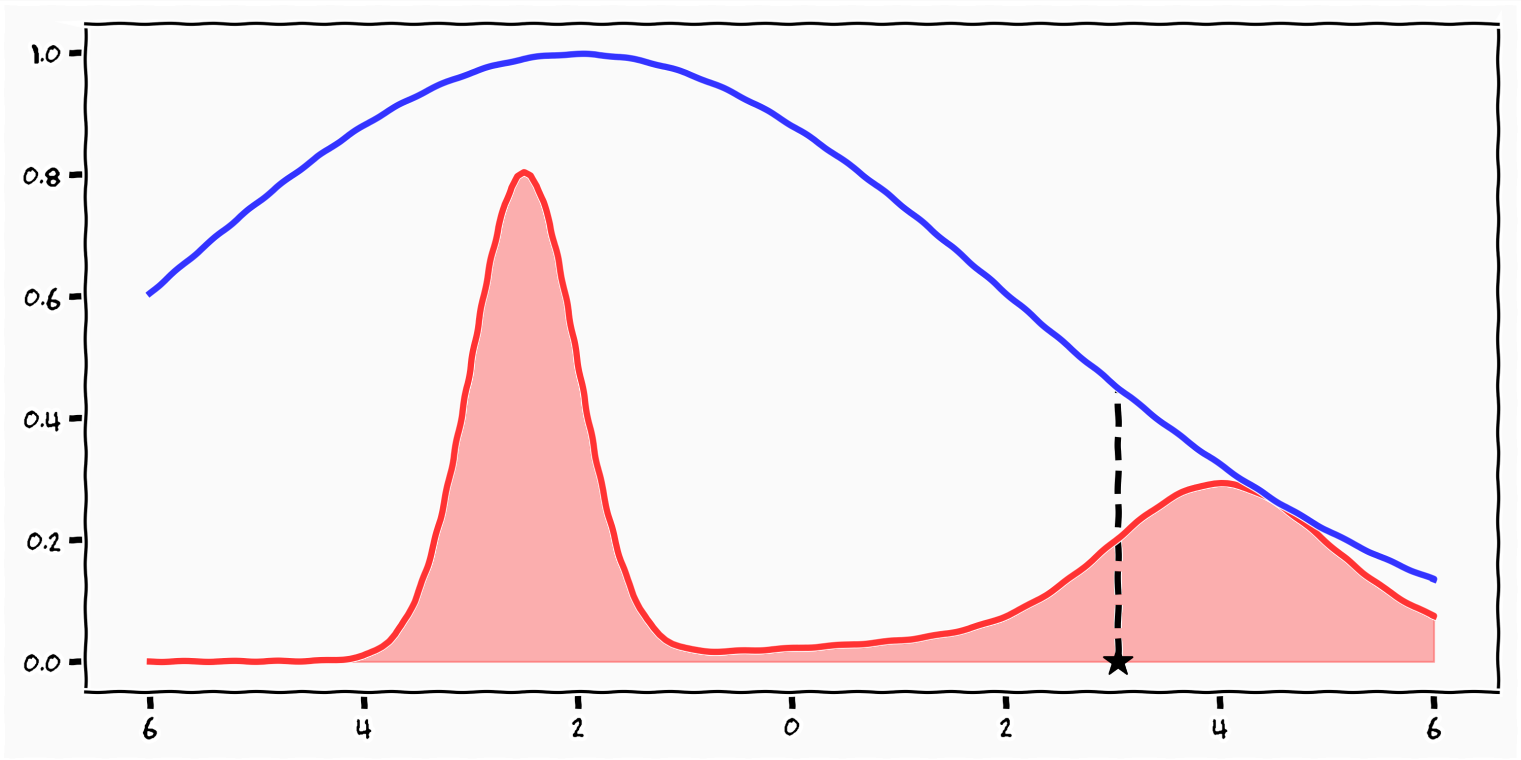
- We know how to transform samples from uniform to any distribution we can formulate the cummulative distribution
- Can we sample from distrubitons we do not know the form of?

    1. Rejection Sampling
    2. Importance Sampling

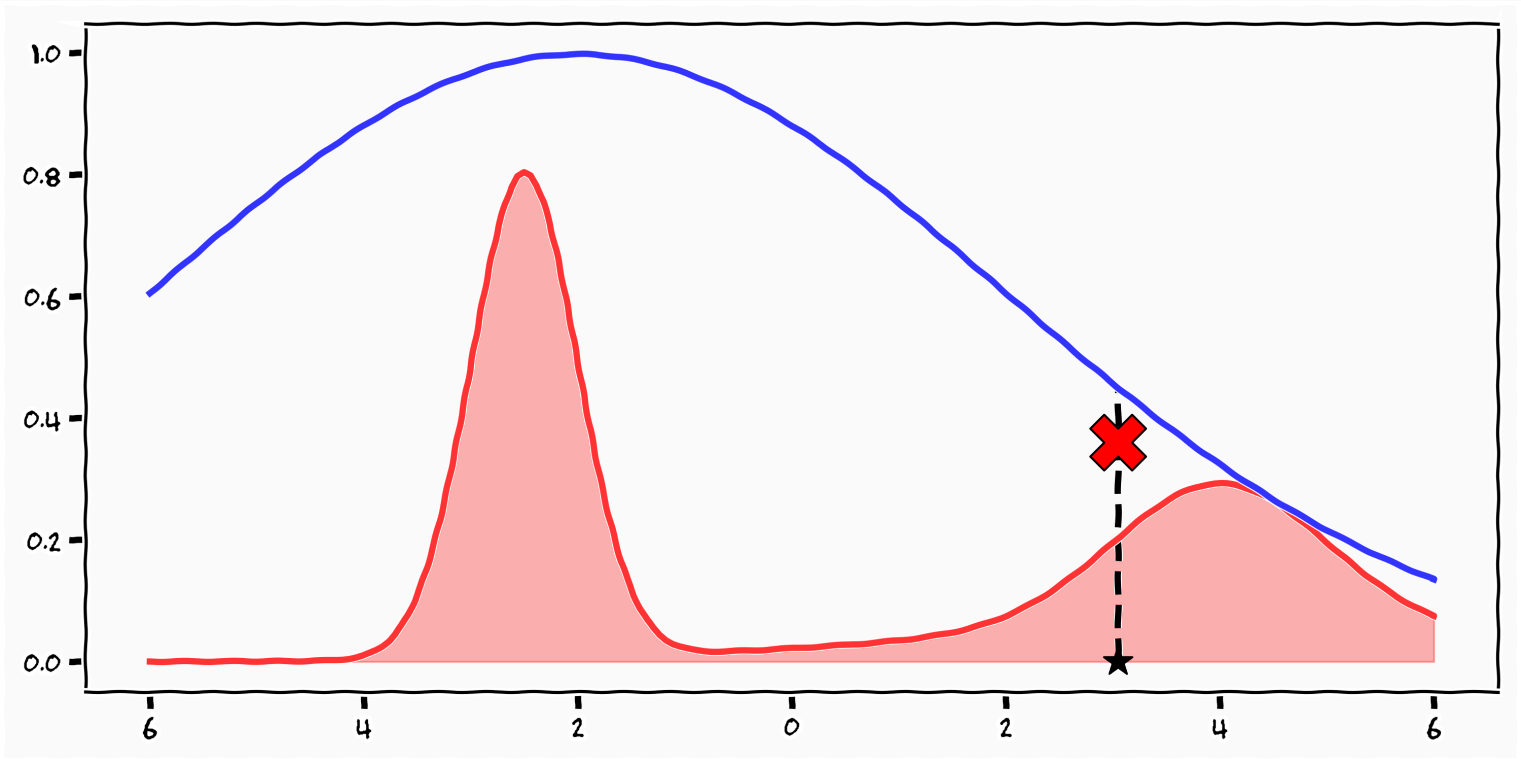
Rejection Sampling



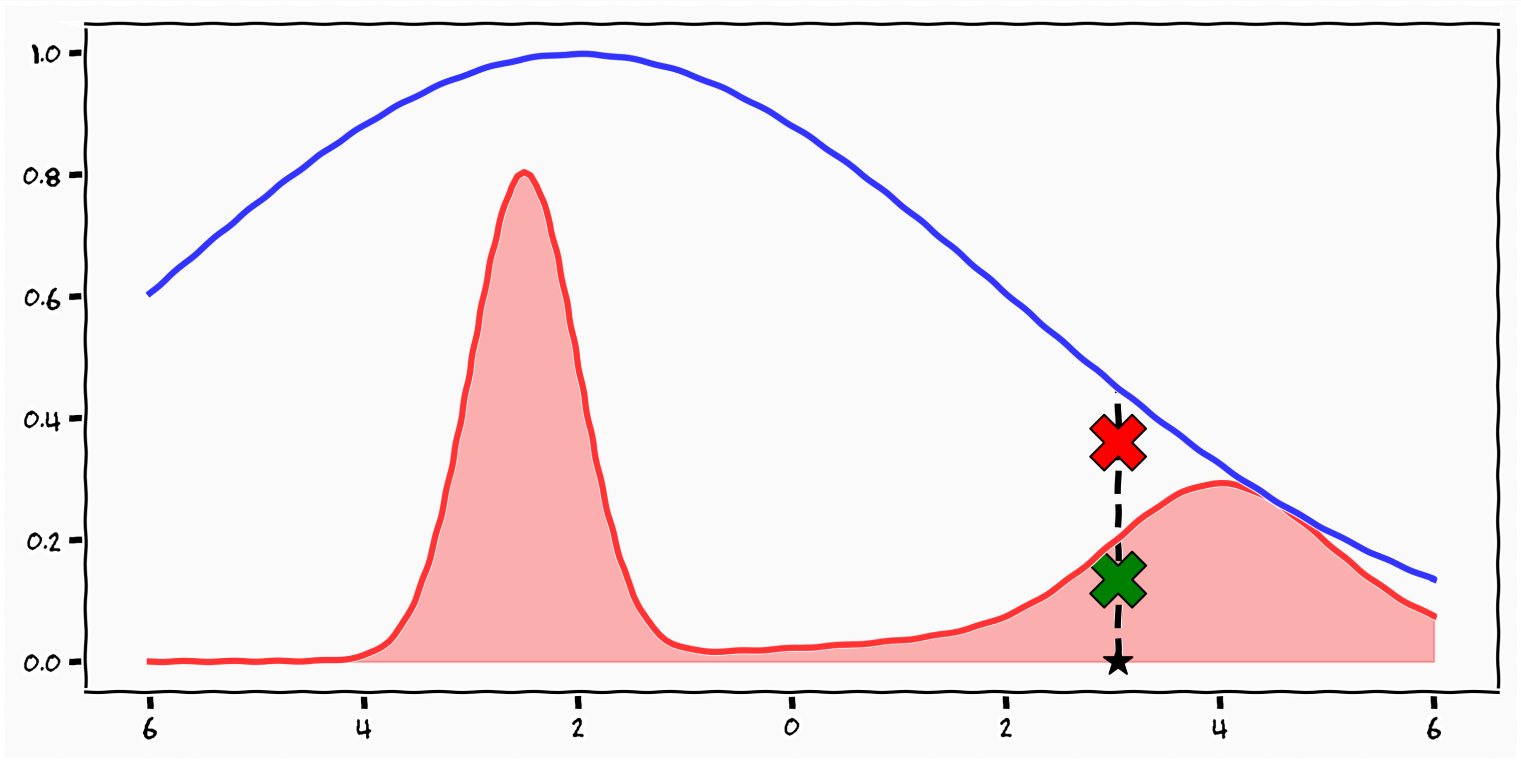
Rejection Sampling



Rejection Sampling



Rejection Sampling



Importance Sampling

$$\mathbb{E}_{p(x)}[f] = \int f(x)p(x)\mathrm{d}x$$

Importance Sampling

$$\mathbb{E}_{p(x)}[f] = \int f(x)p(x)\mathrm{d}x = \int f(x)p(x)\frac{q(x)}{q(x)}\mathrm{d}x$$

Importance Sampling

$$\begin{aligned}\mathbb{E}_{p(x)}[f] &= \int f(x)p(x)\mathrm{d}x = \int f(x)p(x)\frac{q(x)}{q(x)}\mathrm{d}x \\ &= \int f(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x\end{aligned}$$

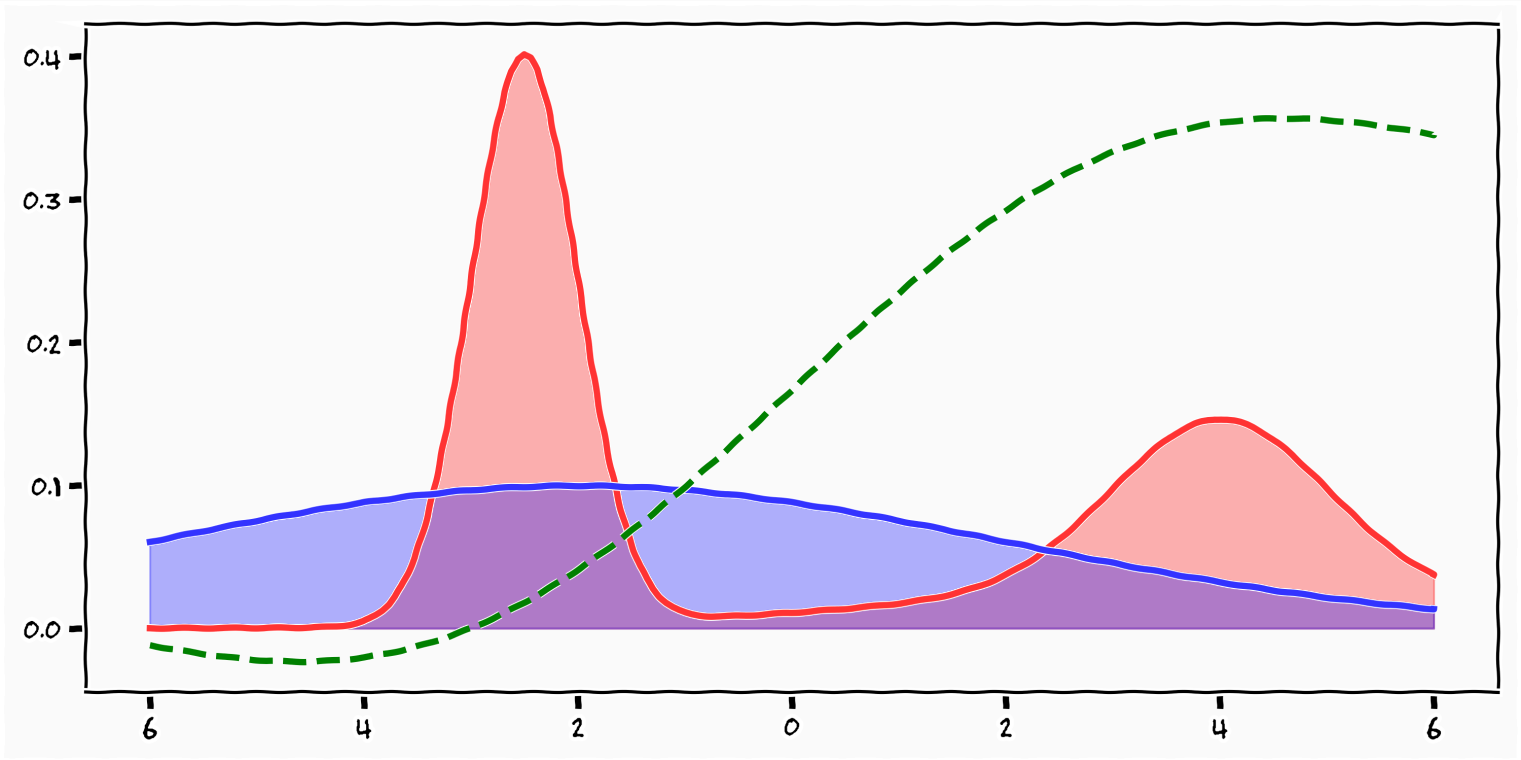
Importance Sampling

$$\begin{aligned}\mathbb{E}_{p(x)}[f] &= \int f(x)p(x)\mathrm{d}x = \int f(x)p(x)\frac{q(x)}{q(x)}\mathrm{d}x \\ &= \int f(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x = \mathbb{E}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right]\end{aligned}$$

Importance Sampling

$$
\begin{aligned}
\mathbb{E}_{p(x)}[f] &= \int f(x)p(x)\mathrm{d}x = \int f(x)p(x)\frac{q(x)}{q(x)}\mathrm{d}x \\
&= \int f(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x = \mathbb{E}_{q(x)}\left[f(x)\frac{p(x)}{q(x)}\right] \\
&\approx \frac{1}{L}\sum_{l=1}^{L} f(x^{(l)})\frac{p(x^{(l)})}{q(x^{(l)})}
\end{aligned}
$$

Importance Sampling



Markov Chain Monte Carlo

- Sample from a proposal distribution
- Remembers the state and samples from a conditional
- Can lead to much better exploration of the space

Metropolis Sampling

1. start with state $x^{(0)}$

Metropolis Sampling

1. start with state $x^{(0)}$
2. sample from conditional proposal distribution $q(x^* \mid x^{(0)})$

Metropolis Sampling

1. start with state $x^{(0)}$
2. sample from conditional proposal distribution $q(x^* \mid x^{(0)})$
3. compute acceptance probability

$$A(x^*, x^{(0)}) = \min\left(1, \frac{\tilde{p}(x^*)}{\tilde{p}(x^{(0)})}\right)$$

Metropolis Sampling

1. start with state $x^{(0)}$
2. sample from conditional proposal distribution $q(x^* \mid x^{(0)})$
3. compute acceptance probability

$$A(x^*, x^{(0)}) = \min\left(1, \frac{\tilde{p}(x^*)}{\tilde{p}(x^{(0)})}\right)$$

4. Draw uniform random number $u \sim \text{Uniform}(0, 1)$

Metropolis Sampling

1. start with state $x^{(0)}$
2. sample from conditional proposal distribution $q(x^* \mid x^{(0)})$
3. compute acceptance probability

$$A(x^*, x^{(0)}) = \min\left(1, \frac{\tilde{p}(x^*)}{\tilde{p}(x^{(0)})}\right)$$

4. Draw uniform random number $u \sim \text{Uniform}(0, 1)$
   - if $A(x^*, x^{(0)}) > u \rightarrow x^{(1)} = x^*$

Metropolis Sampling

1. start with state $x^{(0)}$
2. sample from conditional proposal distribution $q(x^* \mid x^{(0)})$
3. compute acceptance probability

$$A(x^*, x^{(0)}) = \min\left(1, \frac{\tilde{p}(x^*)}{\tilde{p}(x^{(0)})}\right)$$

4. Draw uniform random number $u \sim \text{Uniform}(0, 1)$
   - if $A(x^*, x^{(0)}) > u \rightarrow x^{(1)} = x^*$
   - otherwise reject x^* and start over

Gibbs Sampling

- Often 1D samples are easy to get
- Gibbs sampling exploits this to create a very simple Markov Chain
- Sample each variable in turn conditioned on the others and cycle through

Gibbs Sampling

1. Initialise $x^{(0)}$

Gibbs Sampling

1. Initialise $x^{(0)}$
2. Pick single variable $x_i \in x$

Gibbs Sampling

1. Initialise $x^{(0)}$
2. Pick single variable $x_i \in x$
3. Formulate posterior $p(x_i|x_{\neg i})$

Gibbs Sampling

1. Initialise $x^{(0)}$
2. Pick single variable $x_i \in x$
3. Formulate posterior $p(x_i|x_{\neg i})$
4. Sample from posterior

$$x_i^{(1)} \sim p(x_i|\mathbf{x}_{\neg i})$$

Gibbs Sampling

1. Initialise $x^{(0)}$
2. Pick single variable $x_i \in x$
3. Formulate posterior $p(x_i|x_{\neg i})$
4. Sample from posterior

$$x_i^{(1)} \sim p(x_i|\mathbf{x}_{\neg i})$$

5. cycle through variables

Why is this easier?

Multivariate case

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

$$p(\mathbf{y}) = \sum_i p(\mathbf{y}|\mathbf{x}^{(i)}) p(\mathbf{x}^{(i)})$$

Why is this easier?

Multivariate case

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x})}{p(\mathbf{y})}$$
$$p(\mathbf{y}) = \sum_i p(\mathbf{y}|\mathbf{x}^{(i)}) p(\mathbf{x}^{(i)})$$

1D case

$$p(x_i|\mathbf{x}_{\neg i}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}_{\neg i}, \mathbf{y})}$$

Why is this easier?

Multivariate case

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{x})}{p(\mathbf{y})}$$
$$p(\mathbf{y}) = \sum_i p(\mathbf{y}|\mathbf{x}^{(i)}) p(\mathbf{x}^{(i)})$$

1D case

$$p(x_i|\mathbf{x}_{\neg i}, \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}_{\neg i}, \mathbf{y})}$$
$$p(\mathbf{x}_{\neg i}, \mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) \mathrm{d}x_i = \sum_{x_i \in [1,-1]} p(x_i, \mathbf{x}_{\neg i}, \mathbf{y})$$
$$= p(x_i = 1, \mathbf{x}_{\neg i}, \mathbf{y}) + p(x_i = -1, \mathbf{x}_{\neg i}, \mathbf{y})$$

Summary

- There are lots and lots of sampling methods

Summary

- There are lots and lots of sampling methods
- Most provide quarantees in the limit

Summary

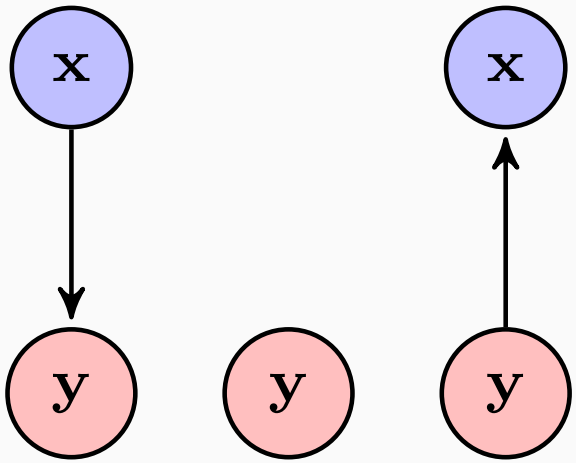
- There are lots and lots of sampling methods
- Most provide quarantees in the limit
- Often requires a lot of problem knowledge to make efficient

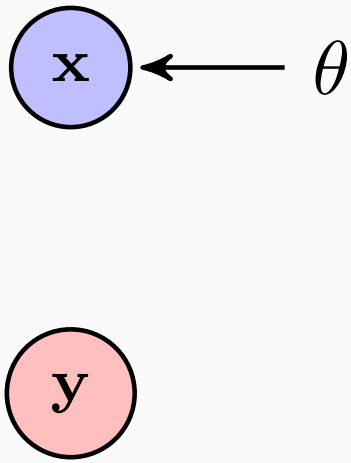
Summary

- There are lots and lots of sampling methods
- Most provide quarantees in the limit
- Often requires a lot of problem knowledge to make efficient
- Hard to get a measure of how well you are doing

Deterministic Inference

Lower Bound





$$p(y) = \int_x p(y|x)p(x) = \frac{p(y|x)p(x)}{p(x|y)}$$

$$q_\theta(x) \approx p(x|y)$$

Variational Bayes

$$p(y)$$

Variational Bayes

$$\log p(y)$$

Variational Bayes

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$
\begin{aligned}
\log p(y) &= \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

Variational Bayes

$$
\begin{aligned}
\log p(y) &= \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{1}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

Variational Bayes

$$\log p(y) = \log p(y) + \int \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

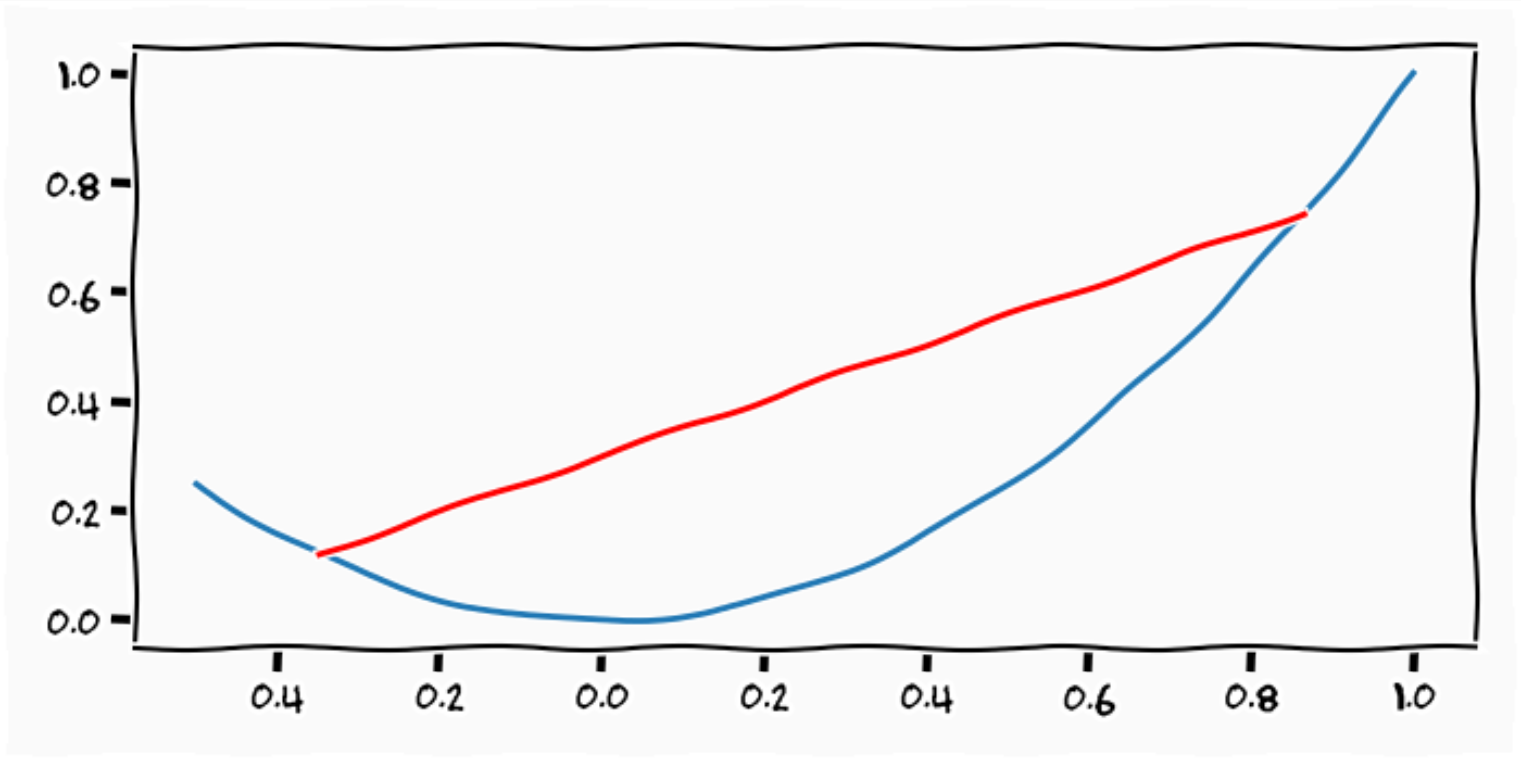
$$= \int q(x) \log p(y) \mathrm{d}x + \int q(x) \log \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{1}{p(x|y)} \mathrm{d}x$$

$$= \int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x,y) \mathrm{d}x + \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

Jensen Inequality



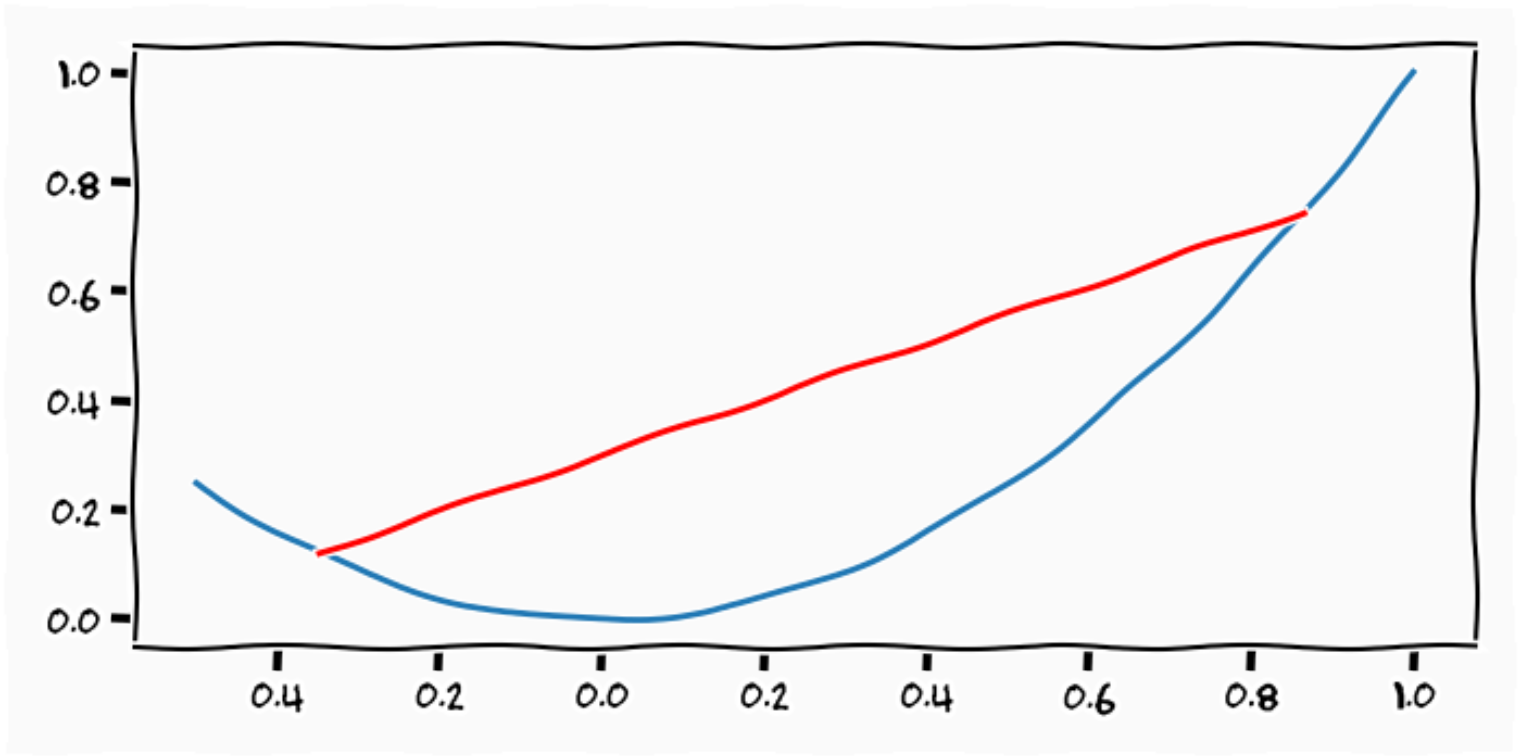
Convex Function

$$\lambda f(x_0) + (1-\lambda) f(x_1) \geq f(\lambda x_0 + (1-\lambda) x_1)$$

$$x \in [x_{min}, x_{max}]$$

$$\lambda \in [0, 1]]$$

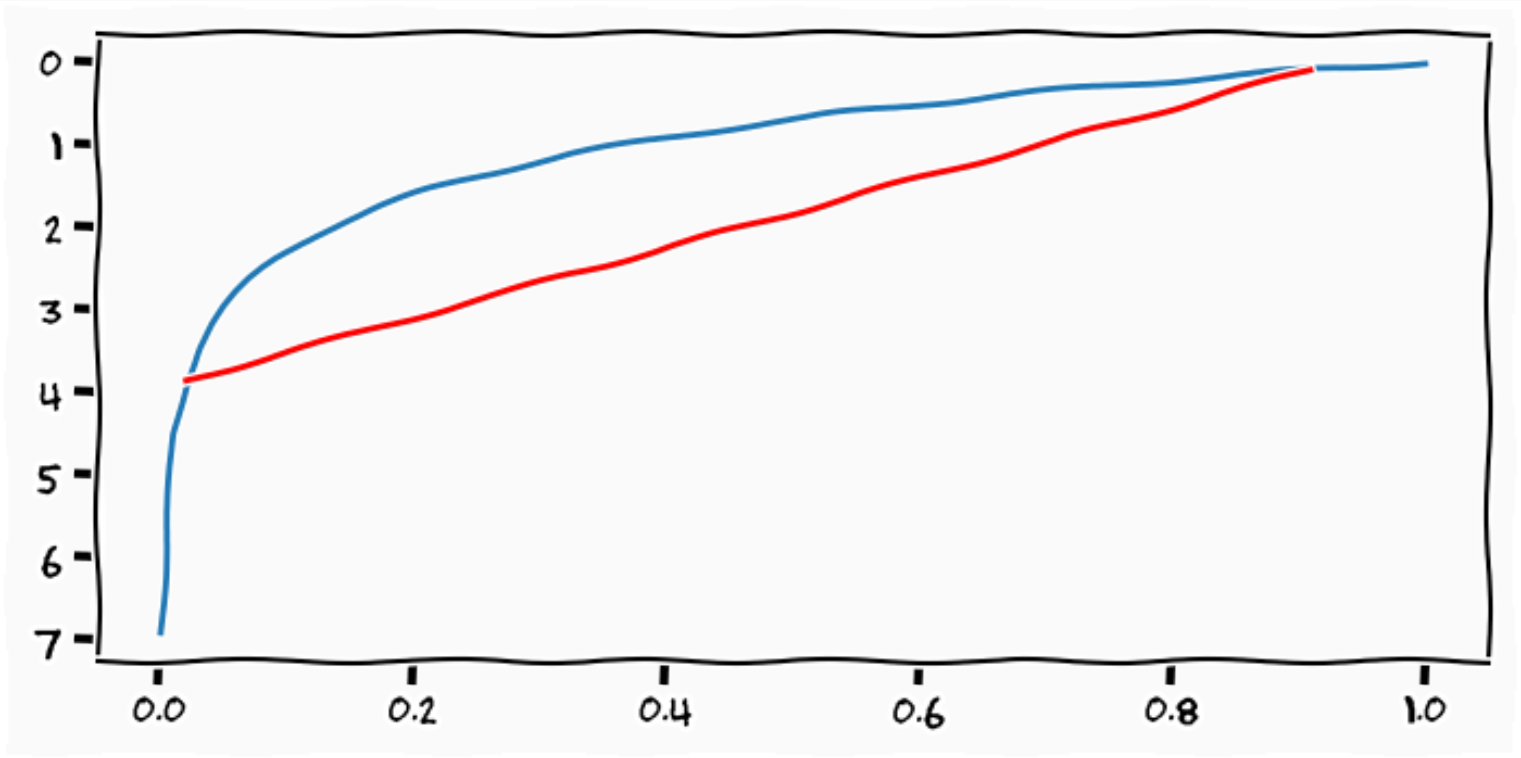
Jensen Inequality



$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

$$\int f(x)p(x)\mathrm{d}x \geq f\left(\int xp(x)\mathrm{d}x\right)$$

Jensen Inequality in Variational Bayes



$$\int \log(x)p(x)\mathrm{d}x \leq \log\left(\int xp(x)\mathrm{d}x\right)$$

moving the log inside the the integral is a lower-bound on the integral

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$= -\int q(x) \log \frac{p(x|y)}{q(x)} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$= -\int q(x) \log \frac{p(x|y)}{q(x)} \mathrm{d}x$$

$$\geq -\log \int p(x|y) \mathrm{d}x = -\log 1 = 0$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y) \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x$$

The "posterior" term

$$\int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y) \log \underbrace{\frac{p(x|y)}{p(x|y)}}_{=1} \mathrm{d}x$$

$$= 0$$

Kullback-Leibler Divergence

$$KL(q(x)||p(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

- Measure of divergence between distributions
- Not a metric (not symmetric)
- $KL(q(x)||p(x|y)) = 0 \Leftrightarrow q(x) = p(x|y)$
- $KL(q(x)||p(x|y)) \geq 0$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$
$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log\frac{p(x,y)}{p(x|y)}\mathrm{d}x$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log\frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log\frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log\frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log\frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

$$= \int p(x|y)\log\underbrace{\frac{p(x|y)}{p(x|y)}}_{=1}\mathrm{d}x + \int p(x|y)\log p(y)\mathrm{d}x$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log\frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log\frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

$$= \int p(x|y)\log\underbrace{\frac{p(x|y)}{p(x|y)}}_{=1}\mathrm{d}x + \int p(x|y)\log p(y)\mathrm{d}x$$

$$= \underbrace{\int p(x|y)\mathrm{d}x}_{=1}\log p(y)$$

The "other terms"

$$\int q(x)\log\frac{1}{q(x)}\mathrm{d}x + \int q(x)\log p(x,y)\mathrm{d}x =$$

$$= \int q(x)\log\frac{p(x,y)}{q(x)}\mathrm{d}x$$

$$= \{\text{Lets assume that } q(x) = p(x|y)\}$$

$$= \int p(x|y)\log\frac{p(x,y)}{p(x|y)}\mathrm{d}x = \int p(x|y)\log\frac{p(x|y)p(y)}{p(x|y)}\mathrm{d}x$$

$$= \int p(x|y)\log\underbrace{\frac{p(x|y)}{p(x|y)}}_{=1}\mathrm{d}x + \int p(x|y)\log p(y)\mathrm{d}x$$

$$= \underbrace{\int p(x|y)\mathrm{d}x}_{=1}\log p(y) = \log p(y)$$

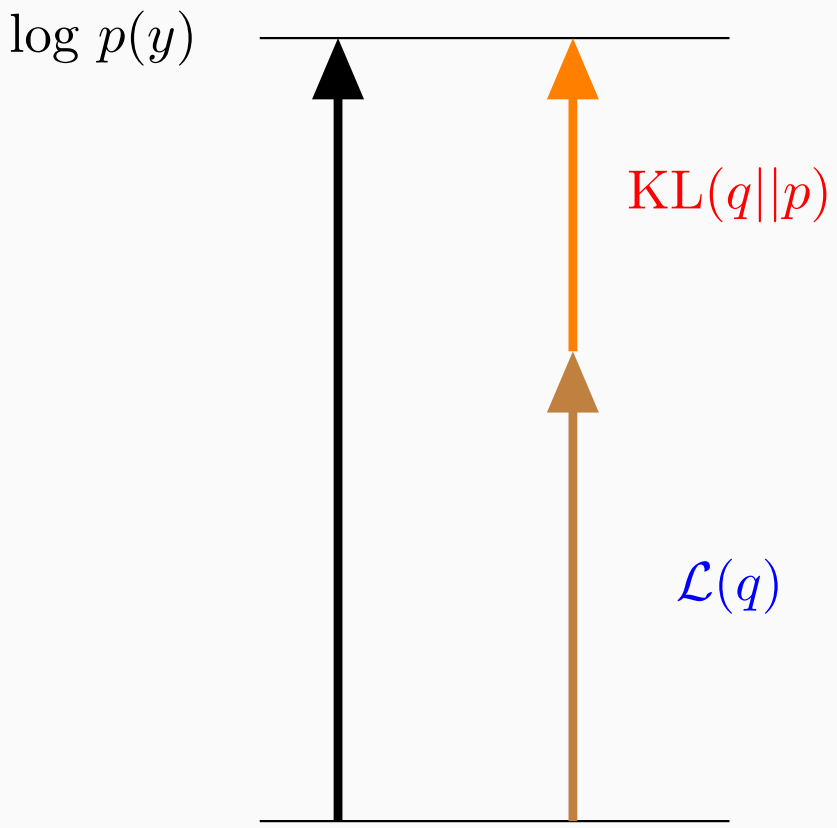
Variational Bayes

$$\log p(y) = \int q(x) \log \frac{1}{q(x)} \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x + \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

$$\geq - \int q(x) \log q(x) \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x$$

- The Evidence Lower BOnd
- Tight if $q(x) = p(x|y)$

Deterministic Approximation

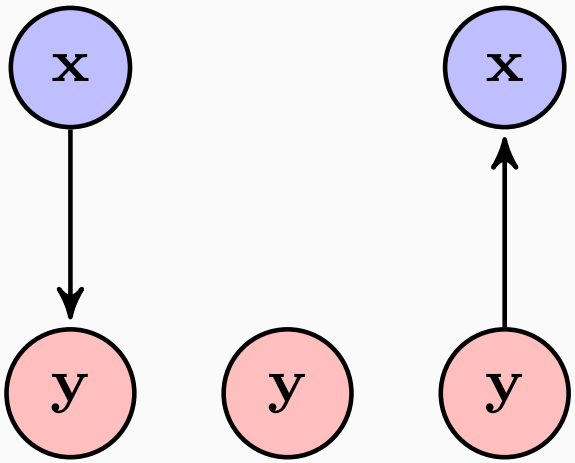


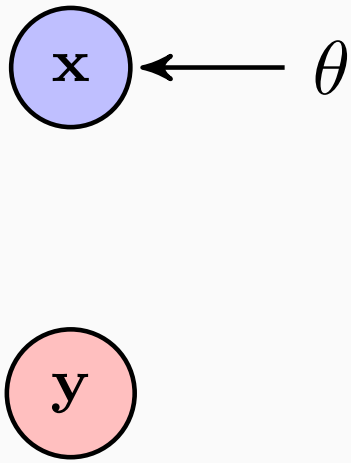
ELBO

$$\log p(y) \geq -\int q(x) \log q(x) \mathrm{d}x + \int q(x) \log p(x, y) \mathrm{d}x$$
$$= \mathbb{E}_{q(x)} \left[\log p(x, y) \right] - H(q(x)) = \mathcal{L}(q(x))$$

- if we maximise the ELBO we,
  - find an approximate posterior
  - lower bound the marginal likelihood
- *maximising* $p(y)$ is learning
- finding $q(x) \approx p(x|y)$ is prediction

Lower Bound





$$p(y) = \int_x p(y|x)p(x) = \frac{p(y|x)p(x)}{p(x|y)}$$

$$q_\theta(x) \approx p(x|y)$$

Why is this useful?

Why is this a sensible thing to do?

– Ryan Adams[5]

[5] Talking Machines Podcast

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do

– Ryan Adams[5]

[5] Talking Machines Podcast

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation

– Ryan Adams[5]

[5] Talking Machines Podcast

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation
- We are allowed to choose the distribution to take the expectation over

– Ryan Adams[5]

[5] Talking Machines Podcast

How to choose Q?

$$\mathcal{L}(q(x)) = \mathbb{E}_{q(x)}\left[\log\, p(x, y)\right] - H(q(x))$$

- We have to be able to compute an expectation over the joint distribution
- The second term should be trivial

Mean Field Approximation

$$q(\mathbf{X}) = \prod_i q_i(\mathbf{x}_i)$$

$$\mathcal{L}(q_j) = \mathcal{L}_j(q_j) + \mathcal{L}_{\neg j}(q_{\neg j}),$$

- Model originating if Physics
- We model marginals rather than the full distribution
- We can update each distribution in turn and cycle

Summary

- Variational Bayes is never exact
- It can be made very fast
- There are exciting works using *implicit* variational distributions

Summary

Summary

- The evidence allows you to do model selection
  - "simple" is a relative concept that needs to be defined
- Most interesting things are intractable
- This leads to the field of approximative inference
  - you have gotten a "hint" of what is out there and some scaffolding to build upon

But what...

- *Now how about if we would treat the whole thing as a machine learning/inference problem instead?*

But what. . .

- *Now how about if we would treat the whole thing as a machine learning/inference problem instead?*
- you can sample the integrand $\rightarrow$ data

But what...

- *Now how about if we would treat the whole thing as a machine learning/inference problem instead?*
- you can sample the integrand $\rightarrow$ data
- you kinda have an idea of the integral $\rightarrow$ prior

But what. . .

- *Now how about if we would treat the whole thing as a machine learning/inference problem instead?*
- you can sample the integrand $\rightarrow$ data
- you kinda have an idea of the integral $\rightarrow$ prior
- you would be able to match the integral to data $\rightarrow$ likelihood

But what...

- *Now how about if we would treat the whole thing as a machine learning/inference problem instead?*
- you can sample the integrand $\rightarrow$ data
- you kinda have an idea of the integral $\rightarrow$ prior
- you would be able to match the integral to data $\rightarrow$ likelihood
- and this we might talk about later in the course :-)

eof

References

Lawrence, Neil D (2005). "Probabilistic non-linear principal component analysis with Gaussian process latent variable models". In: *Journal of Machine Learning Research* 6, pp. 1783–1816.

Mean Field Approximation

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X}$$

Mean Field Approximation

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} = \int \prod_i q_i(\mathbf{x}_i) \log \frac{p(\mathbf{Y}, \mathbf{X})}{\prod_k q_k(\mathbf{x}_k)} \mathrm{d}\mathbf{X}$$

Mean Field Approximation

$$\mathcal{L}(q) = \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}, \mathbf{X})}{q(\mathbf{X})} \mathrm{d}\mathbf{X} = \int \prod_i q_i(\mathbf{x}_i) \log \frac{p(\mathbf{Y}, \mathbf{X})}{\prod_k q_k(\mathbf{x}_k)} \mathrm{d}\mathbf{X}$$

$$= \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right)$$

Mean Field Approximation

$$\mathcal{L}(q_j) = \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}$$

Mean Field Approximation

$$\mathcal{L}(q_j) = \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}$$

$$= \int_j \int_{\neg j} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left(\log\, p(\mathbf{X}, \mathbf{Y}) - \sum_k \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j$$

Mean Field Approximation

$$\mathcal{L}(q_j) = \int \prod_i q_i(\mathbf{x}_i) \left(\log p(\mathbf{Y}, \mathbf{X}) - \sum_k \log q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}$$

$$= \int_j \int_{\neg j} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left(\log\, p(\mathbf{X}, \mathbf{Y}) - \sum_k \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j$$

$$= \int_j q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \log\, p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}}_{\log\, f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \left(\log\, q_j(\mathbf{x}_j) + \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \log\, p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}}_{\log\, f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \left(\log\, q_j(\mathbf{x}_j) + \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \log\, p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}}_{\log\, f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \int_{\neg j} \prod_{i \neq j} q(\mathbf{x}_i) \left(\log\, q_j(\mathbf{x}_j) + \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \right) \mathrm{d}\mathbf{x}_{\neg j} \mathrm{d}\mathbf{x}_j$$

$$= \int_j q_j(\mathbf{x}_j) \log\, f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \left(\log\, q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \sum_{k \neq j} \log\, q_k(\mathbf{x}_k) \mathrm{d}\mathbf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right)$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \left(\log q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \sum_{k \neq j} \log q_k(\mathbf{x}_k) \mathrm{d}\mathbf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right) \mathrm{d}x_j$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j$$

$$- \int_j q_j(\mathbf{x}_j) \left(\log q_j(\mathbf{x}_j) \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \mathrm{d}\mathbf{x}_{\neg j}}_{=1} + \underbrace{\int_{\neg j} \prod_{i \neq j} q_i(\mathbf{x}_i) \sum_{k \neq j} \log q_k(\mathbf{x}_k) \mathrm{d}\mathbf{x}_{\neg j}}_{\text{constant w.r.t. } q_j} \right) \mathrm{d}x_j$$

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1}$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1}$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1}$$

$$= \int_j q_j(\mathbf{x}_j) \log \frac{f_j(\mathbf{x}_j)}{q_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1}$$

$$= \int_j q_j(\mathbf{x}_j) \log \frac{f_j(\mathbf{x}_j)}{q_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

$$= -\int_j q_j(\mathbf{x}_j) \log \frac{q_j(\mathbf{x}_j)}{f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1}$$

$$= \int_j q_j(\mathbf{x}_j) \log \frac{f_j(\mathbf{x}_j)}{q_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

$$= -\int_j q_j(\mathbf{x}_j) \log \frac{q_j(\mathbf{x}_j)}{f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

$$= -\mathrm{KL}(q_j(\mathbf{x}_j) || f_j(\mathbf{x}_j)) + \text{const.}$$

Mean Field Approximation

$$= \int_j q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j - \int_j q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j + \text{const.} \cdot \underbrace{\int_j q_j(\mathbf{x}_j) \mathrm{d}\mathbf{x}_j}_{=1}$$

$$= \int_j q_j(\mathbf{x}_j) \log \frac{f_j(\mathbf{x}_j)}{q_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

$$= -\int_j q_j(\mathbf{x}_j) \log \frac{q_j(\mathbf{x}_j)}{f_j(\mathbf{x}_j)} \mathrm{d}\mathbf{x}_j + \text{const.}$$

$$= -\mathrm{KL}(q_j(\mathbf{x}_j) || f_j(\mathbf{x}_j)) + \text{const.} = \mathcal{L}(q_j)$$

Mean Field Approximation

$$\mathcal{L}(q_j) = -\mathsf{KL}(q_j(\mathbf{x}_j)||f_j(\mathbf{x}_j)) + \text{const.}$$

- Want to maximise lower bound
- Negative KL $\rightarrow$ minimise KL term
- *we are free to choose the form of the distribution*

Mean Field Approximation

$$\log q_j(\mathbf{x}_j) = \log f_j(\mathbf{x}_j) = \int_{\neg j} \underbrace{\prod_{i \neq j} q_i(\mathbf{x}_i)}_{q_{\neg j}(\mathbf{x}_{\neg j})} \log p(\mathbf{Y}, \mathbf{X}) \mathrm{d}\mathbf{x}_{\neg j}$$

$$= \mathbb{E}_{q_{\neg j}(\mathbf{x}_{\neg j})} \left[\log p(\mathbf{Y}, \mathbf{X})\right]$$

- Choose the marginal distribution that makes the bound tight
- Will not make the bound tight in general though

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters
2. Formulate fully factorised approximative posterior over latent variables

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters
2. Formulate fully factorised approximative posterior over latent variables
3. Fit marginal approximation by making bound tight

Mean Field Variational Bayes

1. Formulate joint distribution over data and latent parameters
2. Formulate fully factorised approximative posterior over latent variables
3. Fit marginal approximation by making bound tight
4. Iterate through variables

Taking stock

- We have derived a general update for the mean-field approximation

Taking stock

- We have derived a general update for the mean-field approximation
- Looks an awful lot like Expectation-Maximisation
    - EM is VB (E-local, M-global)

Taking stock

- We have derived a general update for the mean-field approximation
- Looks an awful lot like Expectation-Maximisation
  - EM is VB (E-local, M-global)
- Looks an awful lot like Gibbs sampling
  - Gibbs: sample from complete conditional
  - VB: variational expectations of natural parameters of the complete conditional