University of BRISTOL

Machine Learning

Evidence and Graphical Models

Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 29, 2019

http://carlhenrik.com

Introduction

Optimisation

- Classic optimisation

$$\hat{x} = \text{argmin}_x f(x)$$

- Much more common

$$\hat{x} = \text{argmin}_x \text{black-box}(x)$$

Optimisation

- in most cases we have an objective function that we do not know

Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function

Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge

Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge
- the cost of each test is expensive

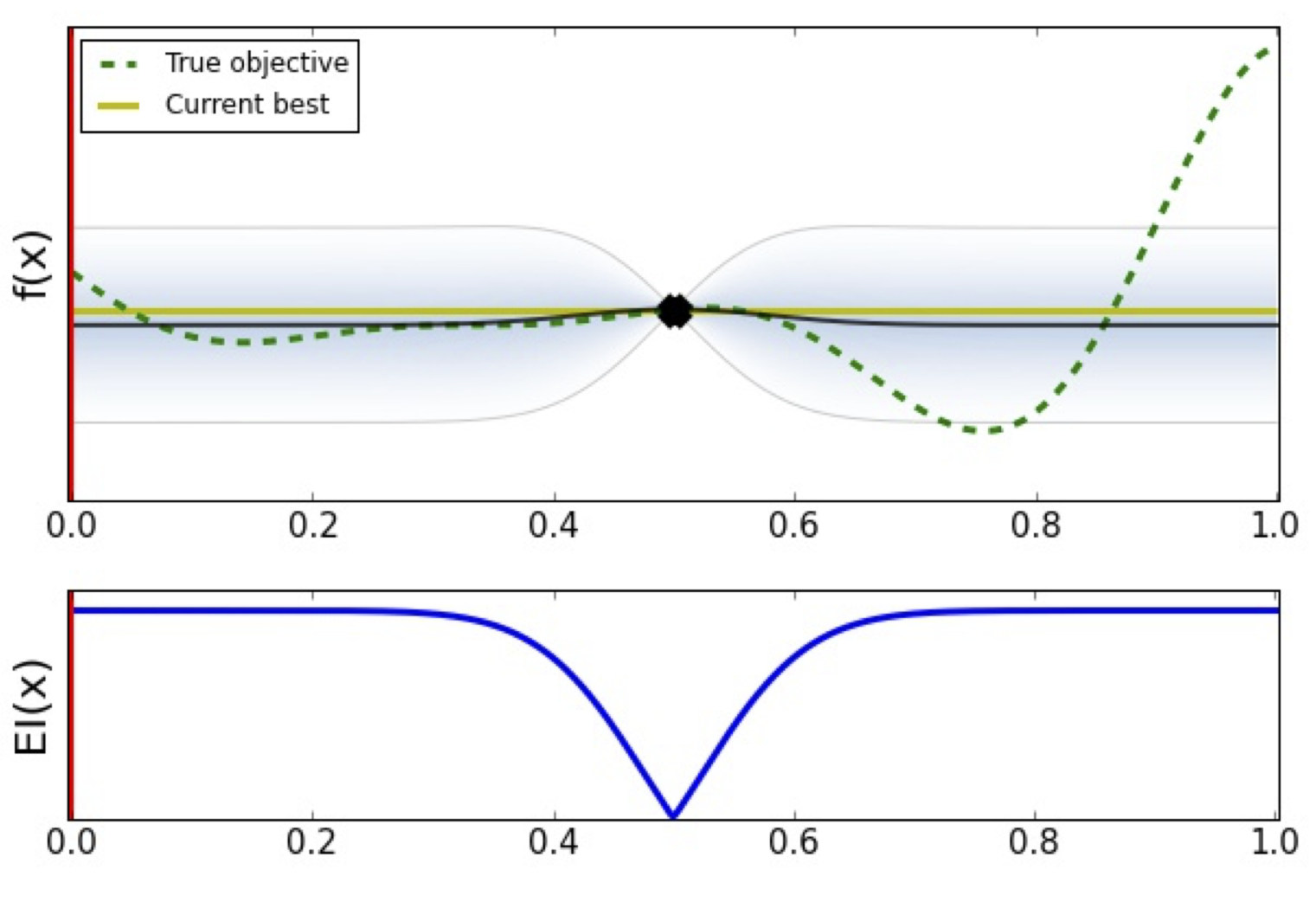
Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge
- the cost of each test is expensive
- the test is noisy

Optimisation

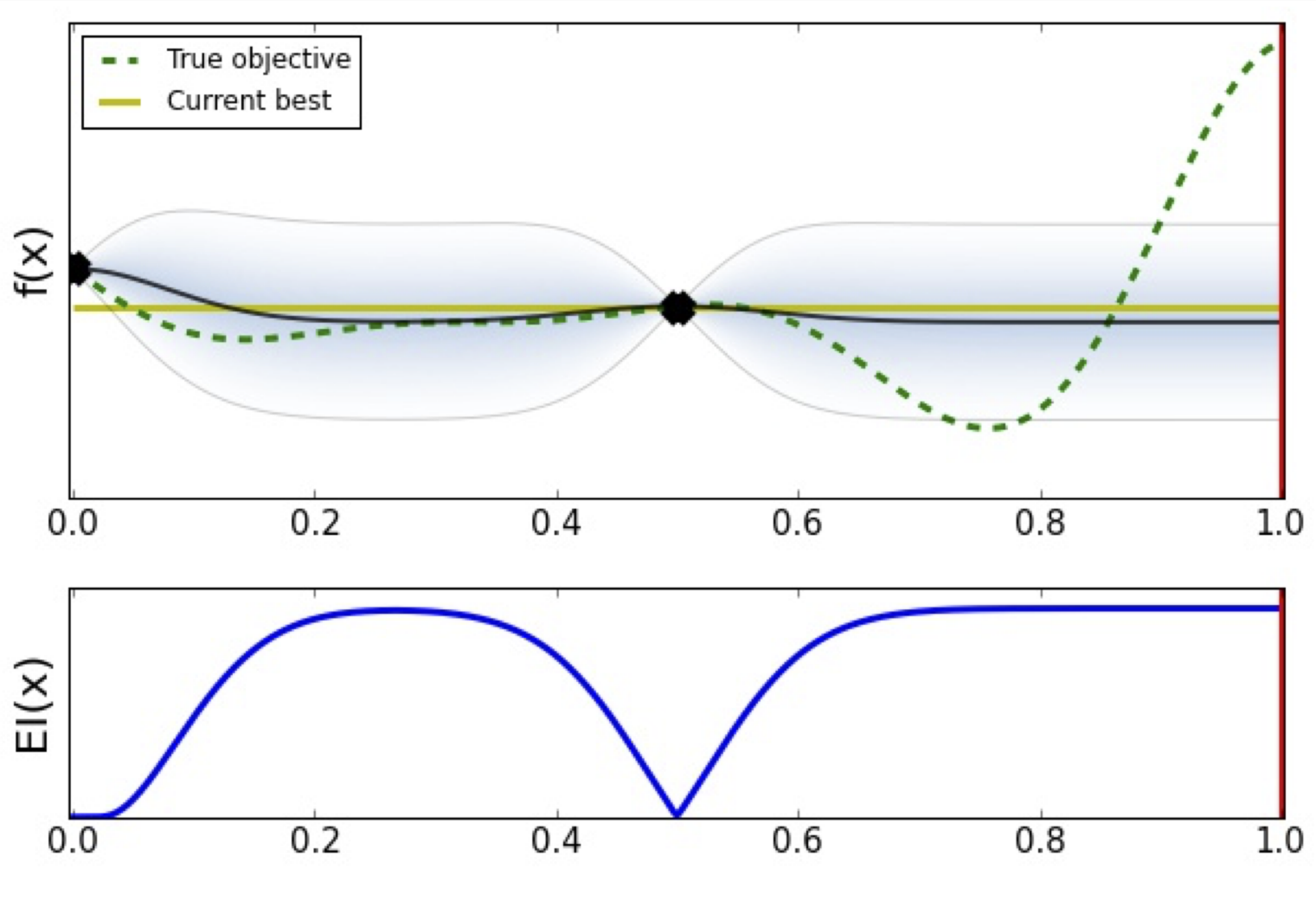
- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge
- the cost of each test is expensive
- the test is noisy
- *can we use machine learning to do this for us?*

Bayesian Optimisation [1]



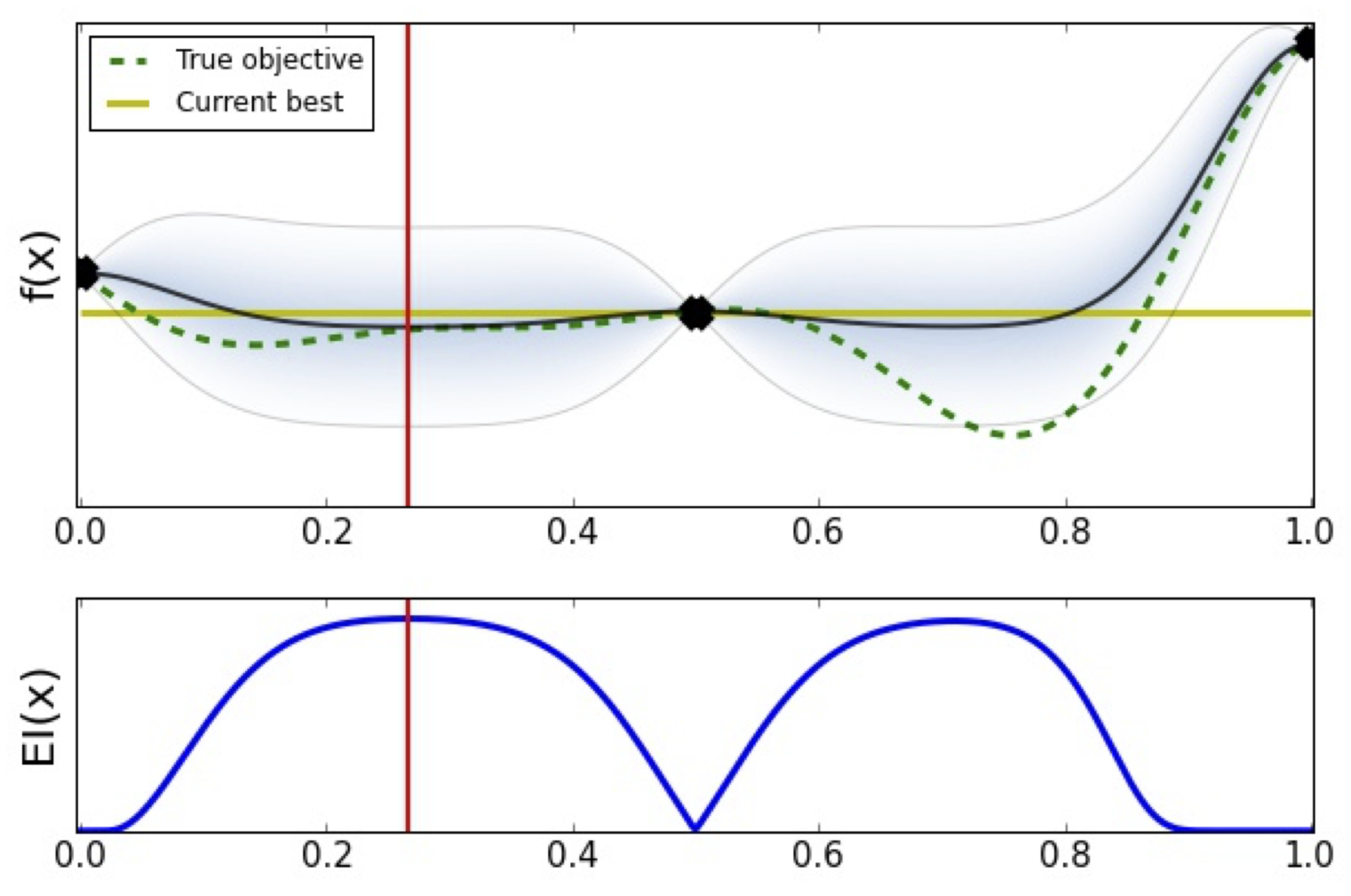
[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



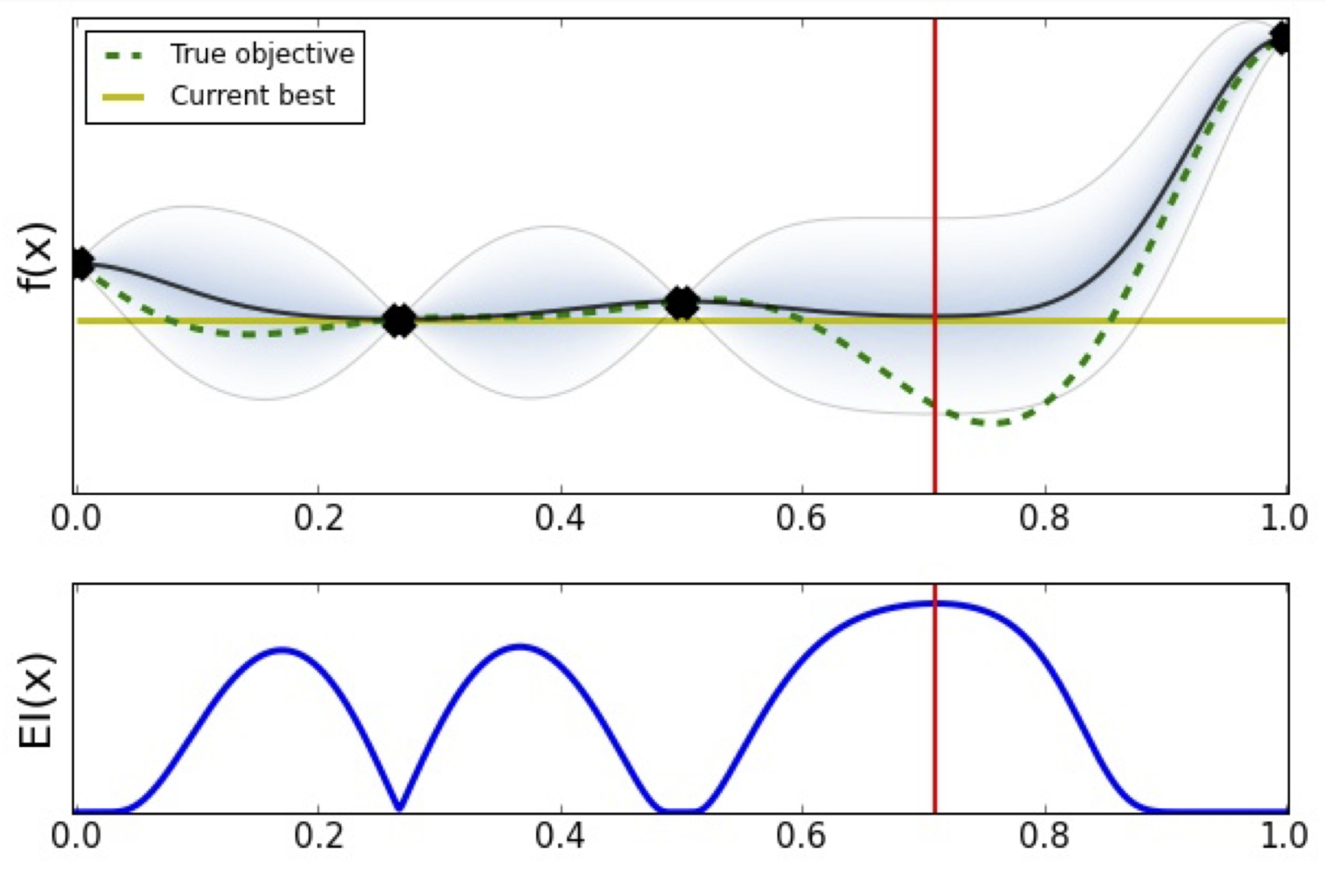
[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



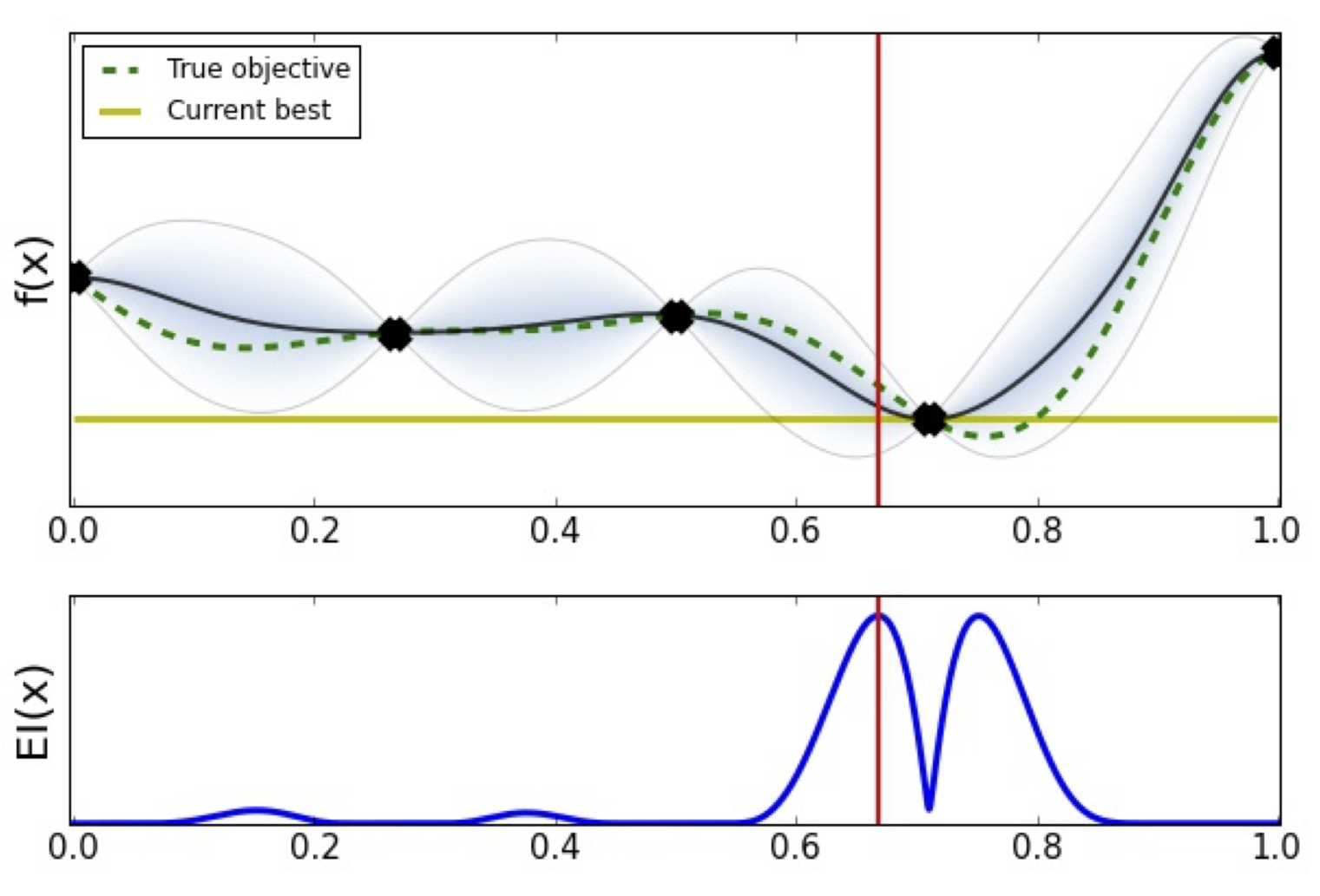
[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



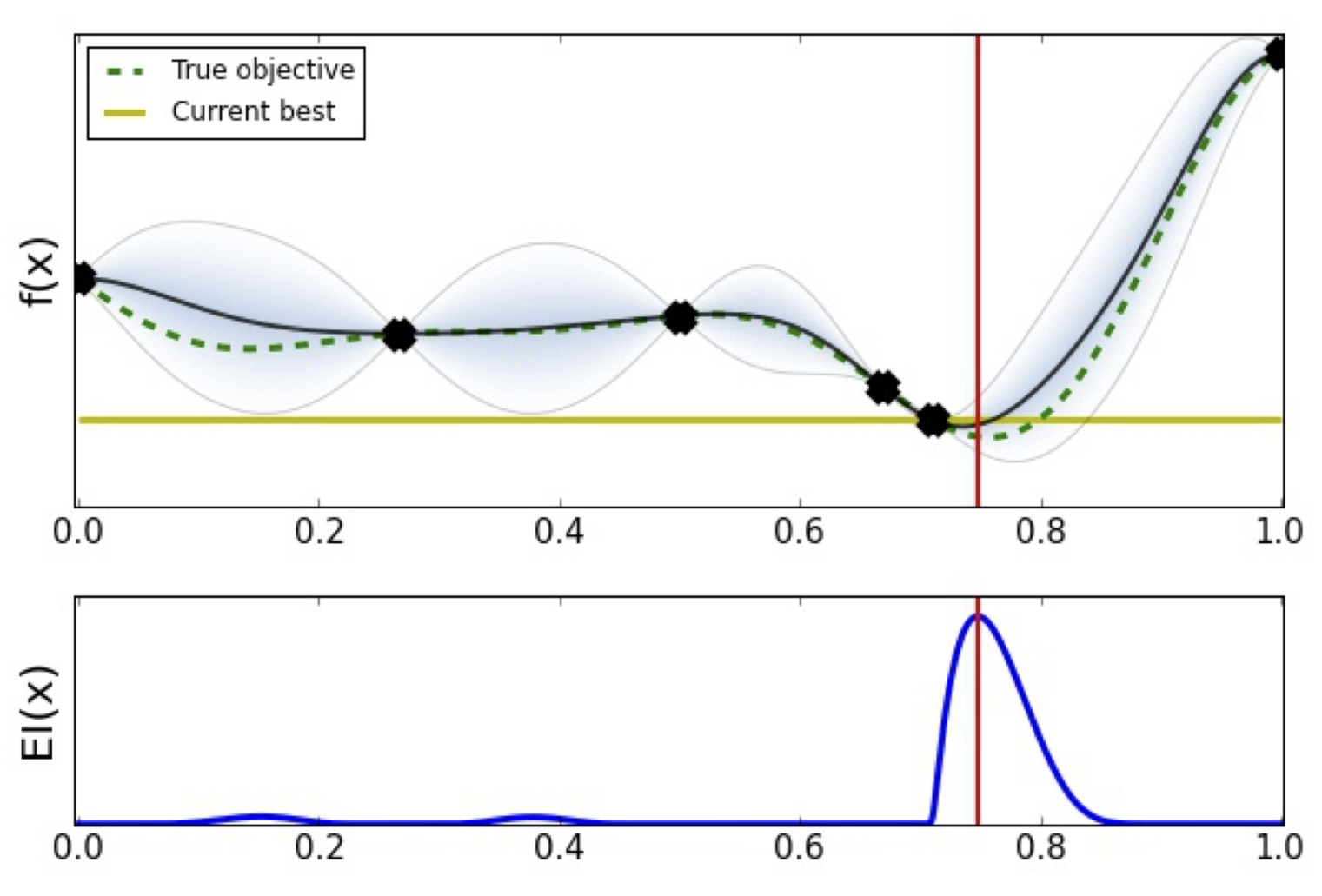
[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



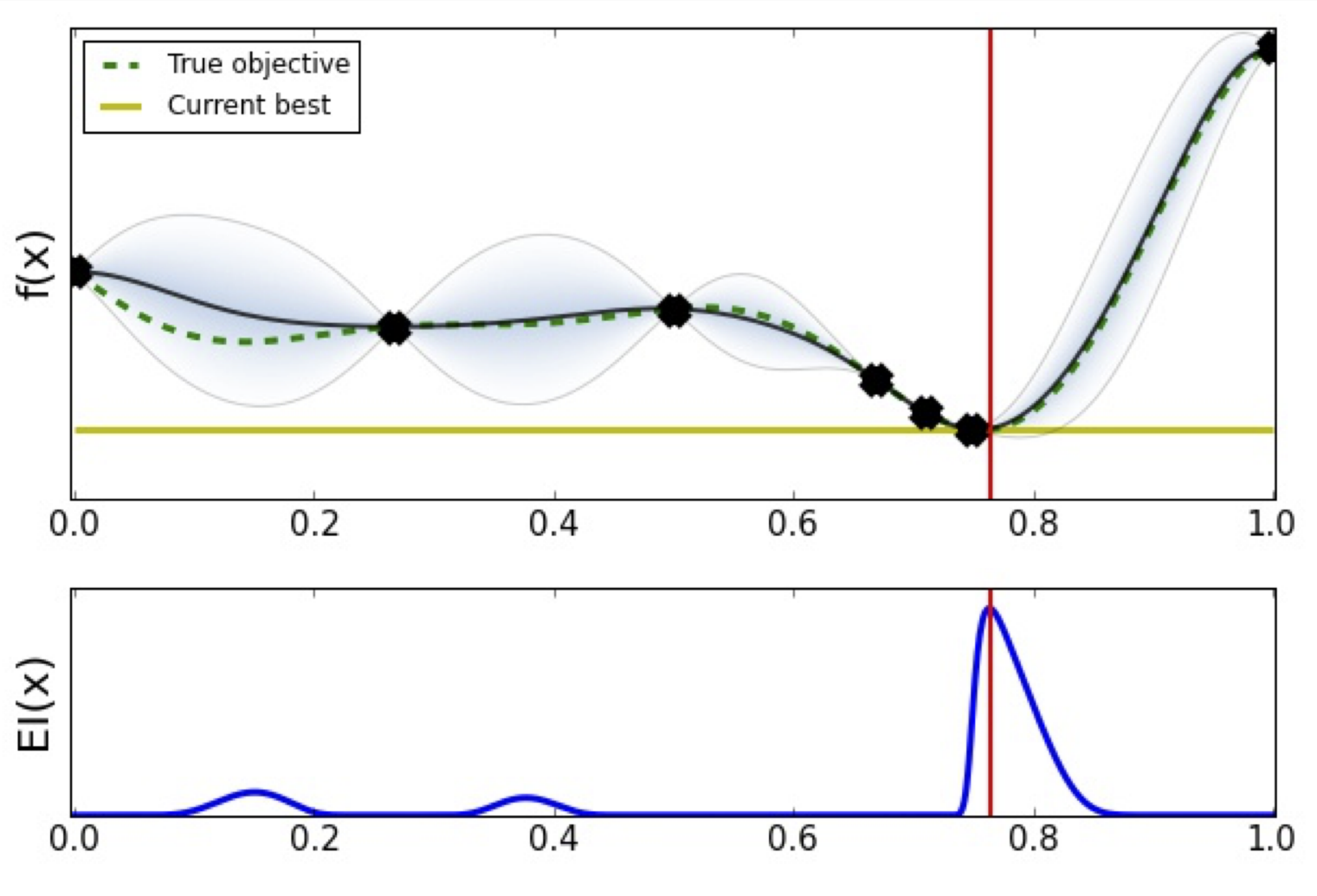
[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



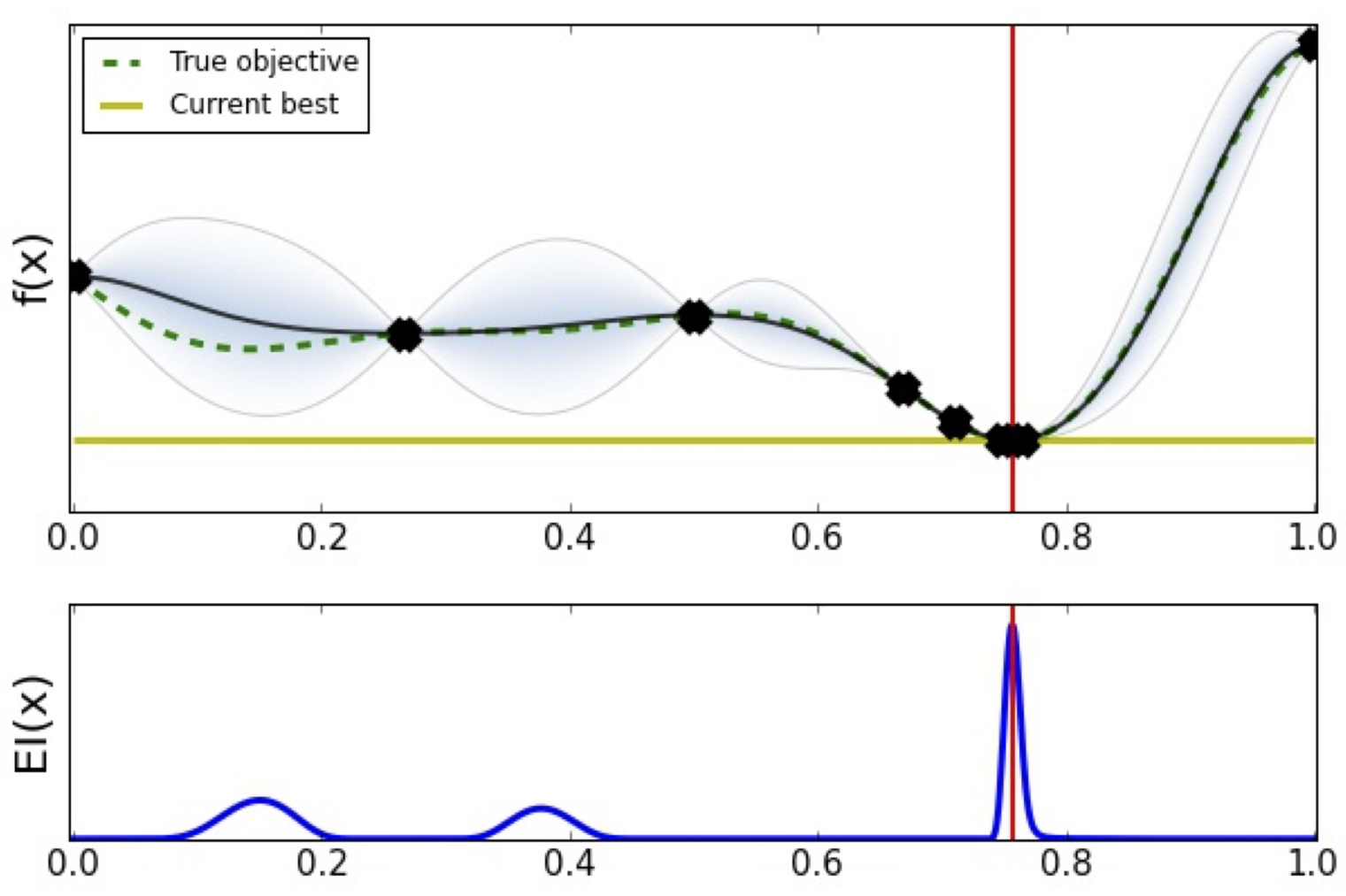
[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optimisation [1]



[1] https://github.com/jluttine/tikz-bayesnet

Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

- Transform to a series of simpler problems

$$x_{n+1} = \text{argmin}_{x \in \mathcal{X}} \alpha(x; \mathcal{D}_n, \mathcal{M}_n)$$

Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

- Transform to a series of simpler problems

$$x_{n+1} = \text{argmin}_{x \in \mathcal{X}} \alpha(x; \mathcal{D}_n, \mathcal{M}_n)$$

- this will work well if

Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

- Transform to a series of simpler problems

$$x_{n+1} = \text{argmin}_{x \in \mathcal{X}} \alpha(x; \mathcal{D}_n, \mathcal{M}_n)$$

- this will work well if
  - $\alpha(x)$ is cheap to compute

Evidence

Bayes Rule

$$p(\theta|\mathcal{Y}) = p(\mathcal{Y}|\theta)p(\theta)\frac{1}{p(\mathcal{Y})}$$

Evidence

$$p(\mathcal{Y}) = \int p(\mathcal{Y}|\theta)p(\theta)\mathrm{d}\theta$$

Regression Models

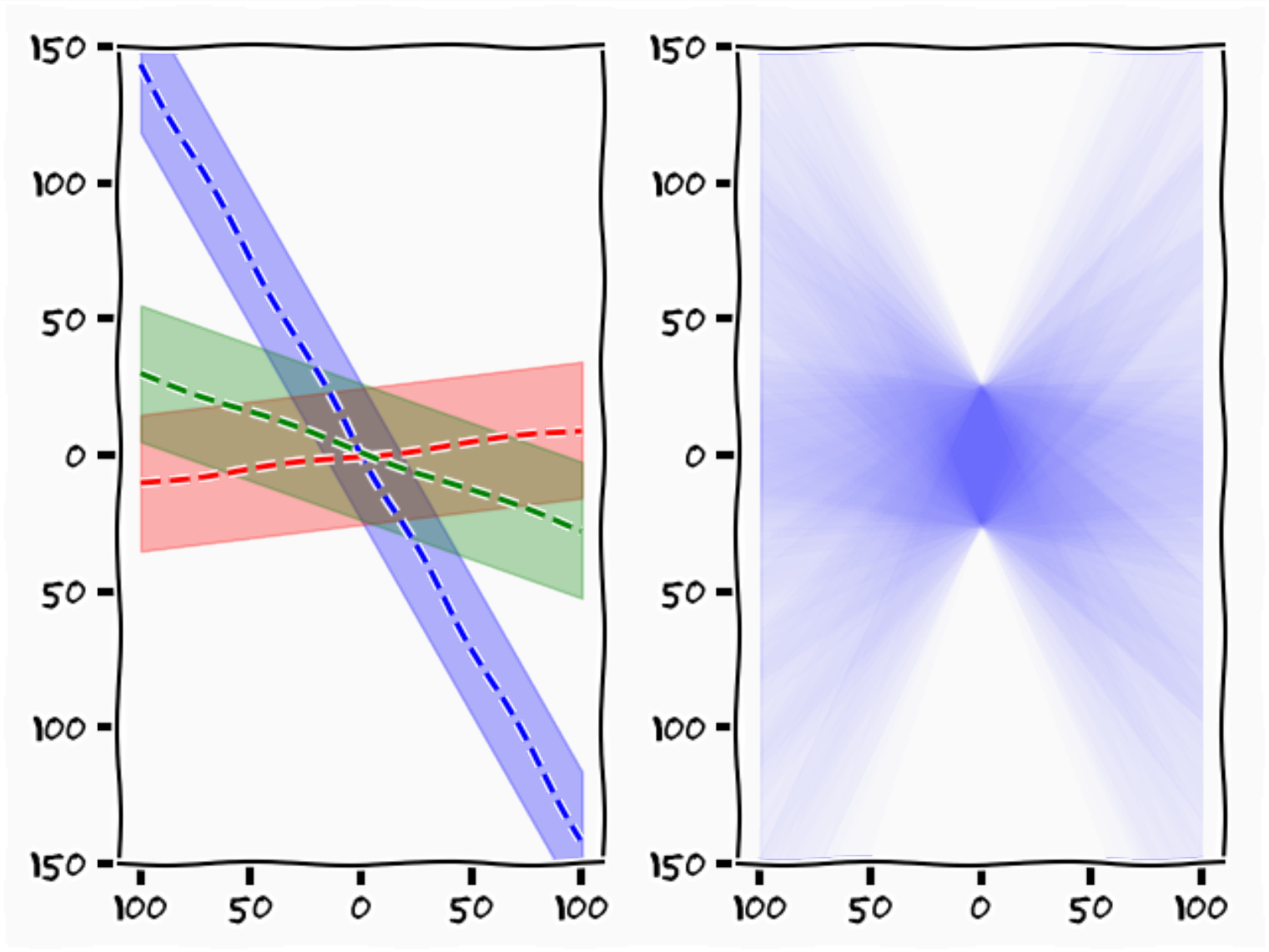
Linear Linear Model

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

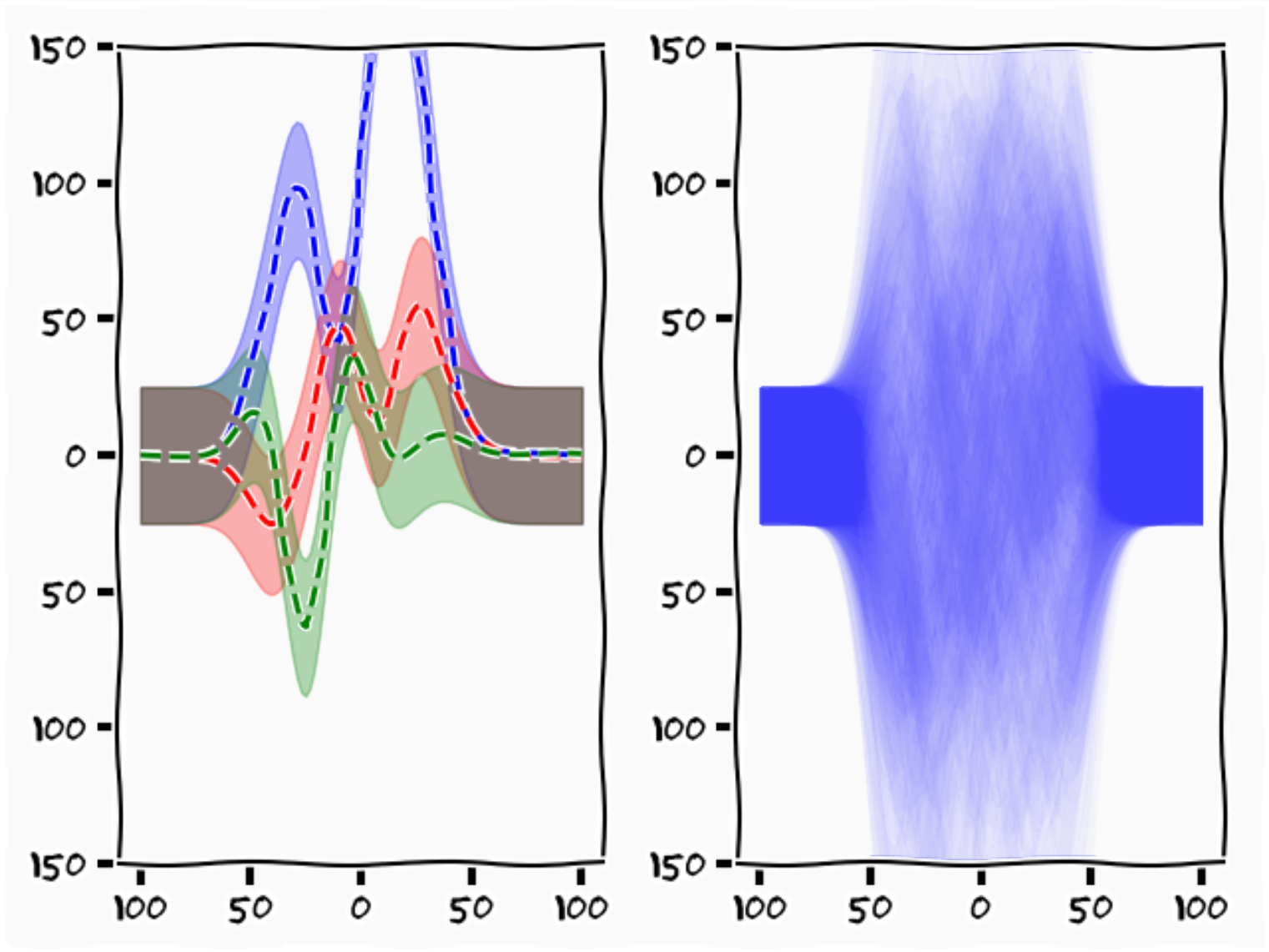
Basis function

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

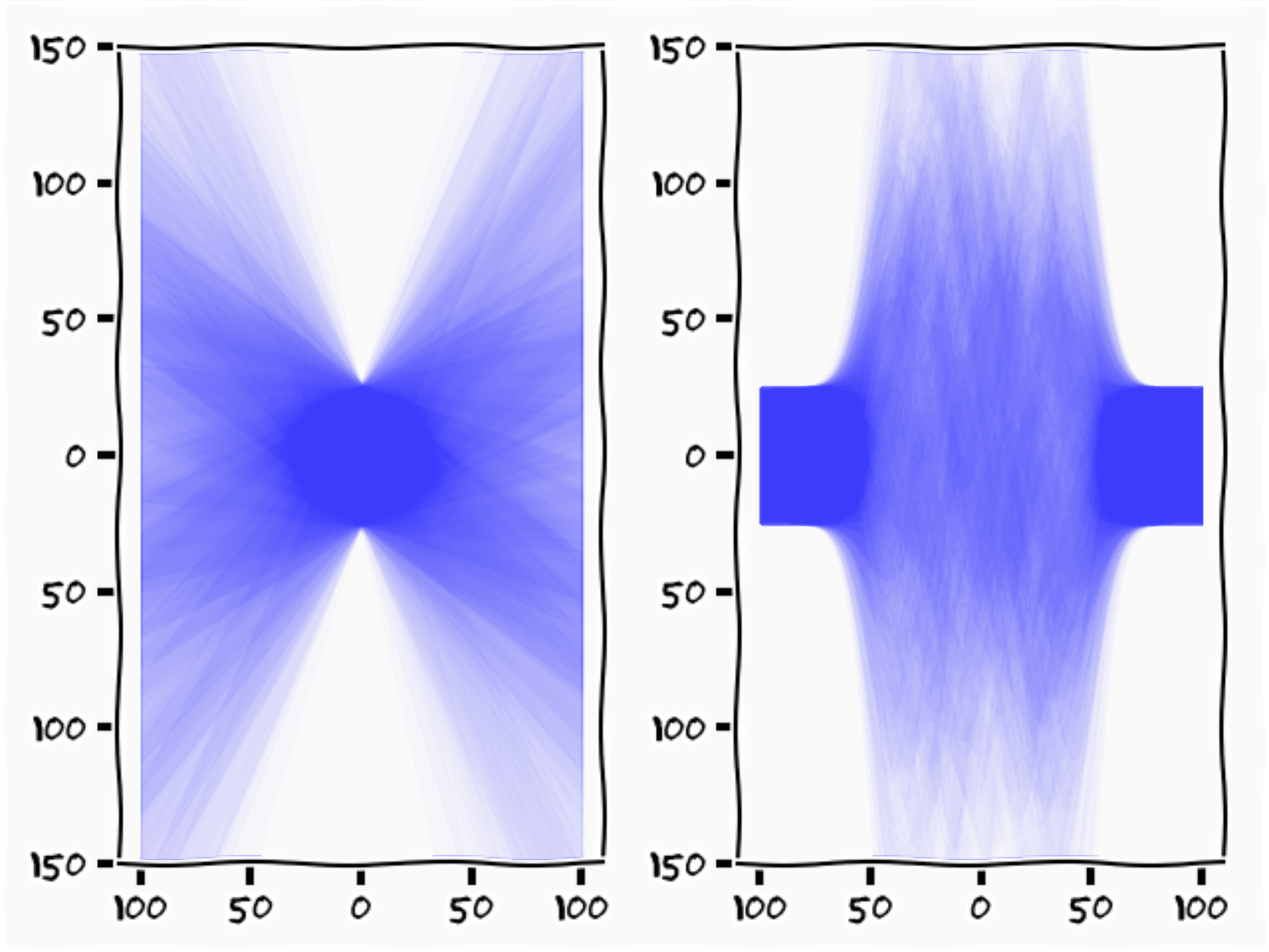
Model 1


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

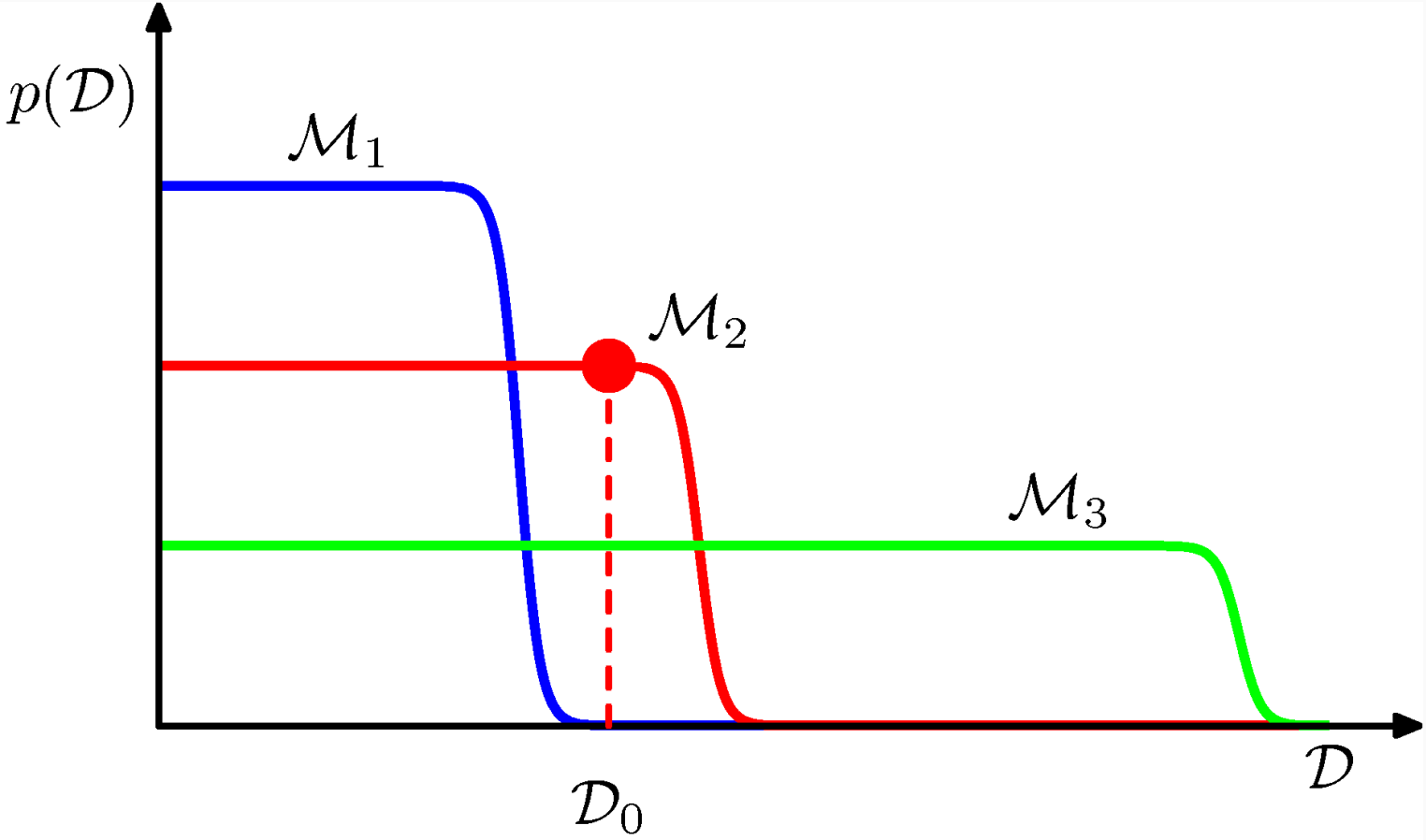
Model 2



Evidence


150
100
50
0
50
100
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 Selection



Occams Razor

Occams Razor

Definition (Occams Razor)

"All things being equal, the simplest solution tends to be the best one"

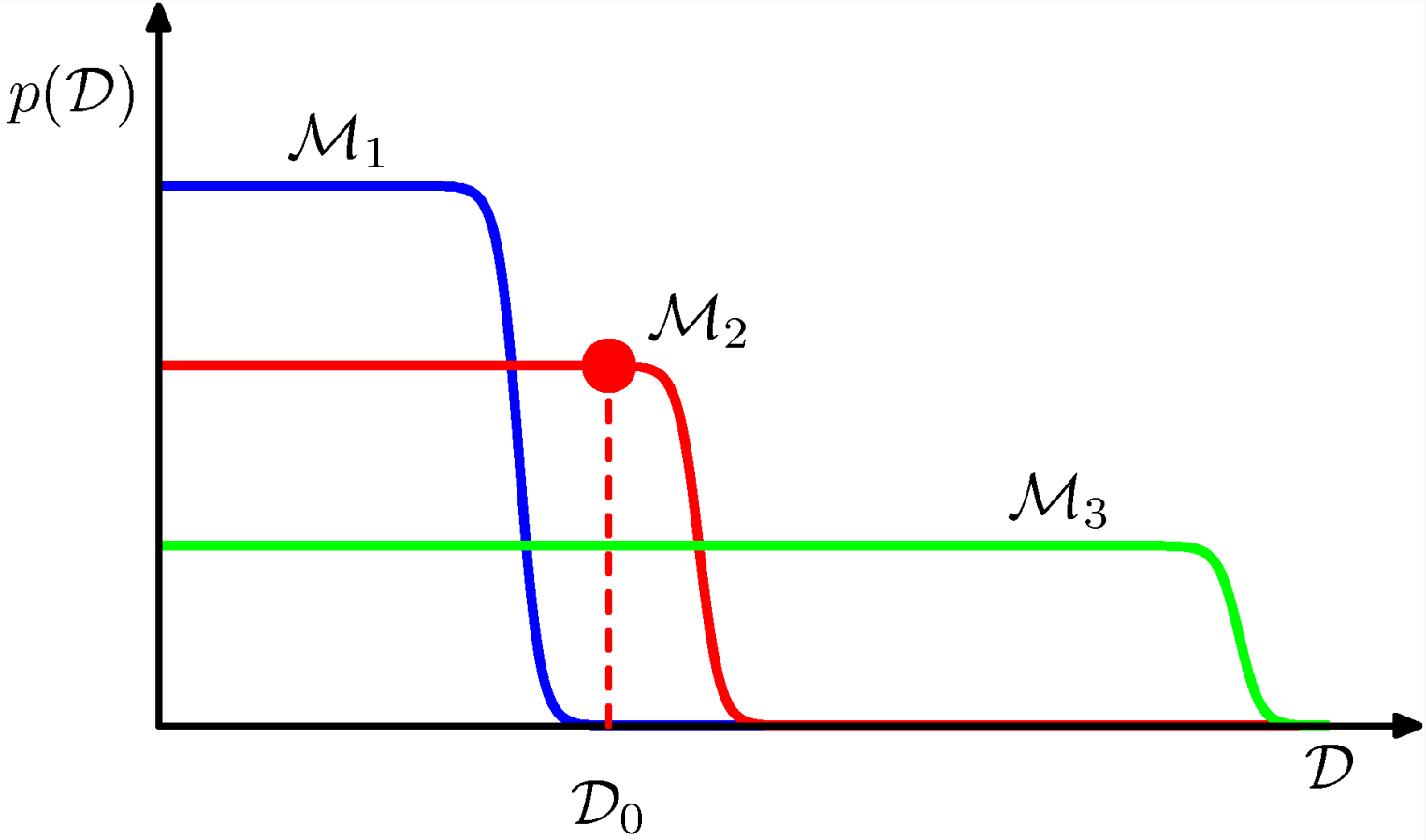
– William of Ockham

"Things must be made as simple as possible, but never simpler"

– Albert Einstein

What is Simple?

Model Selection



Graphical Models

Decisions [2]

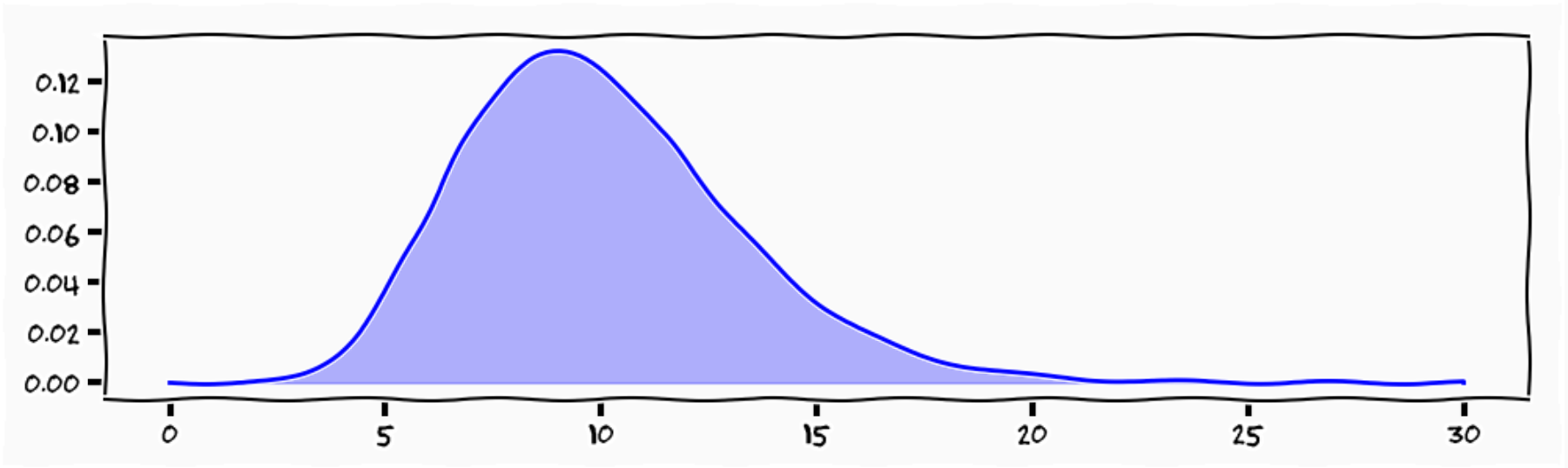
[2] Reservoir Dogs Tipping Scene YouTube

Tip

$$p(y)$$

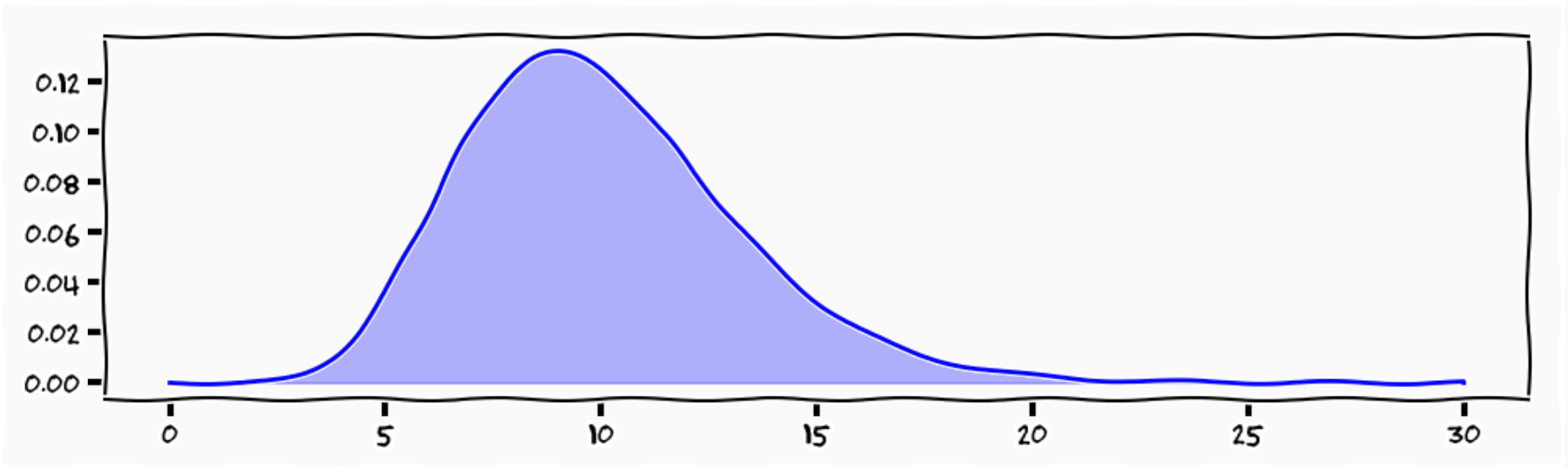
- what do I believe about tip before I see data?
- what is a sensible tip?

Tipping



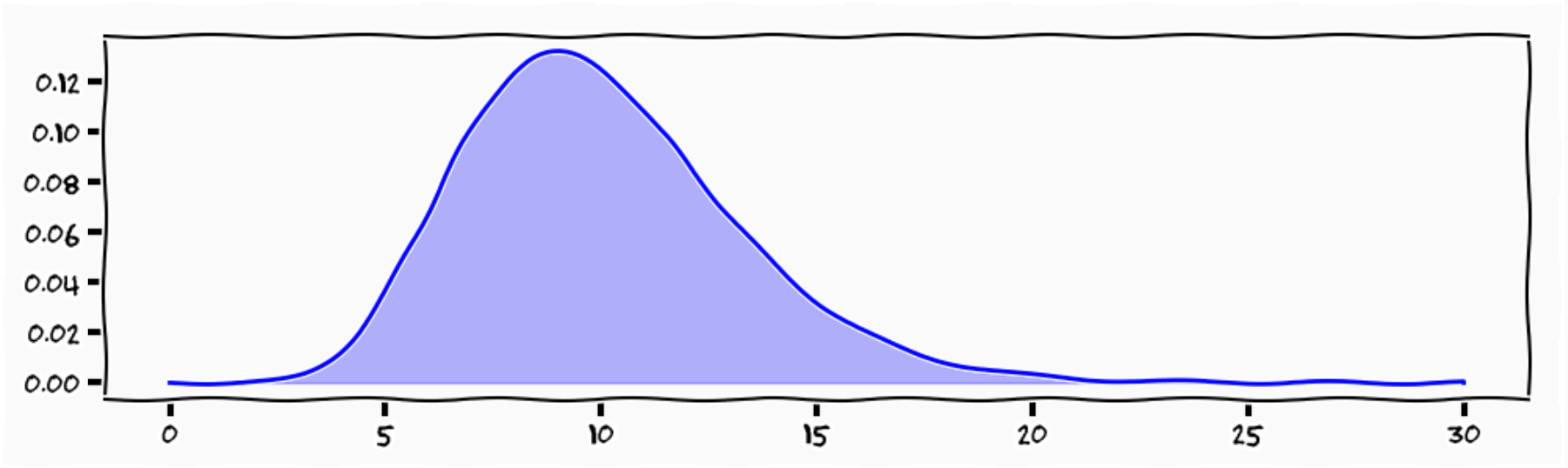
- You cannot tip negative
- There is potentially an upper bound

Tipping



- You cannot tip negative
- There is potentially an upper bound
- This is not a model, its just a belief in a variable

Tipping



- You cannot tip negative
- There is potentially an upper bound
- This is not a model, its just a belief in a variable
- *a model relates new phenomenon to knowledge*

Tipping

- it is quite hard to say something about tip without any other knowledge
- Assumption the value of tip is related to the quality of the food

Likelihoods

$$p(y|x)$$

- how likely do I think the observed data y is to come from this specific x.

Tipping if I know the quality of the food what do I believe the tip should be

Tipping

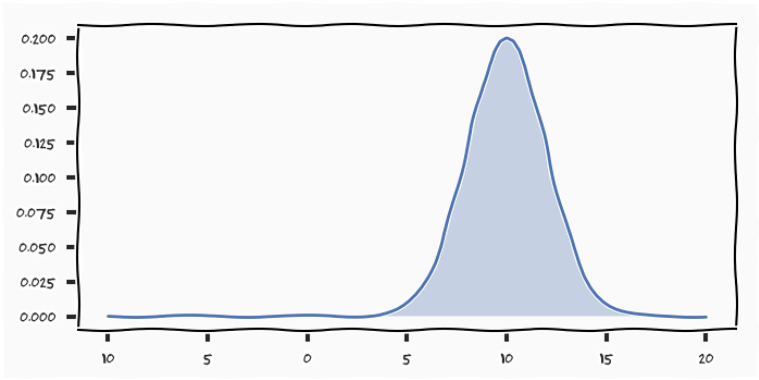
$$p(x|c)$$

Hierarchical distribution

- Its quite hard to think of a prior over quality of food
- Can we parametrise the quality?

Tipping

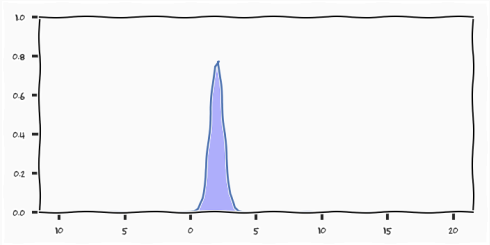
$$p(x|c) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_c)(x-\mu_c)}{2\sigma^2}}$$

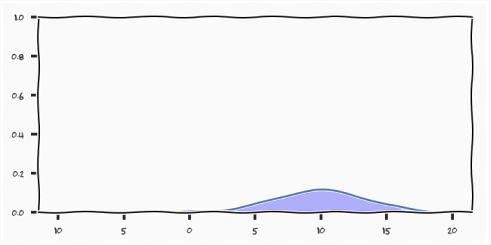


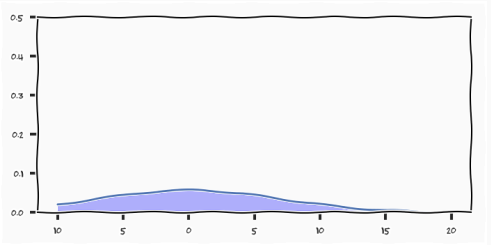
Hierarchical distribution

- Its quite hard to think of a prior over quality of food
- Can we parametrise the quality?
- Lets assume that if we know the cusine we have an idea

Tipping







Swedish Italian Uzbeki

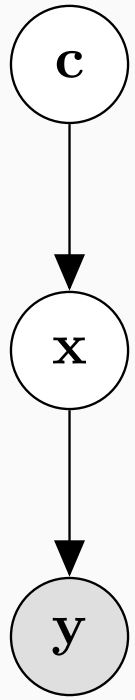
Hierarchical distribution

- Its quite hard to think of a prior over quality of food
- Can we parametrise the quality?
- Lets assume that if we know the cusine we have an idea
- *Relating to knowledge!*

Tipping model

$$p(y, x, c) = p(y|x)p(x|c)p(c)$$

- Graphical Model shows dependency structure
- Shows "minimal" factorisation of joint distribution (model)



What is the tip that I should expect to get?

$$\mathbb{E}_{p(x,c)}[p(y|x)p(x|c)] = \int p(y|x)p(x|c)p(c)\mathrm{d}x\mathrm{d}c = p(y)$$

- What should I expect to get in tip
- I have an idea of the general distribution of quality of food
- *Understanding is when we can relate knowledge to new phenomenon*

Graphical Models

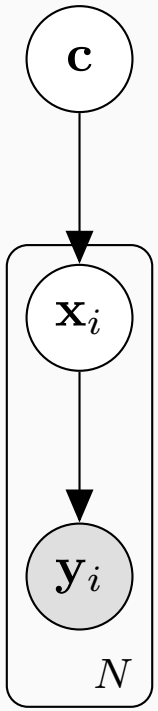
node/vertice random variable

edge stochastic relationship

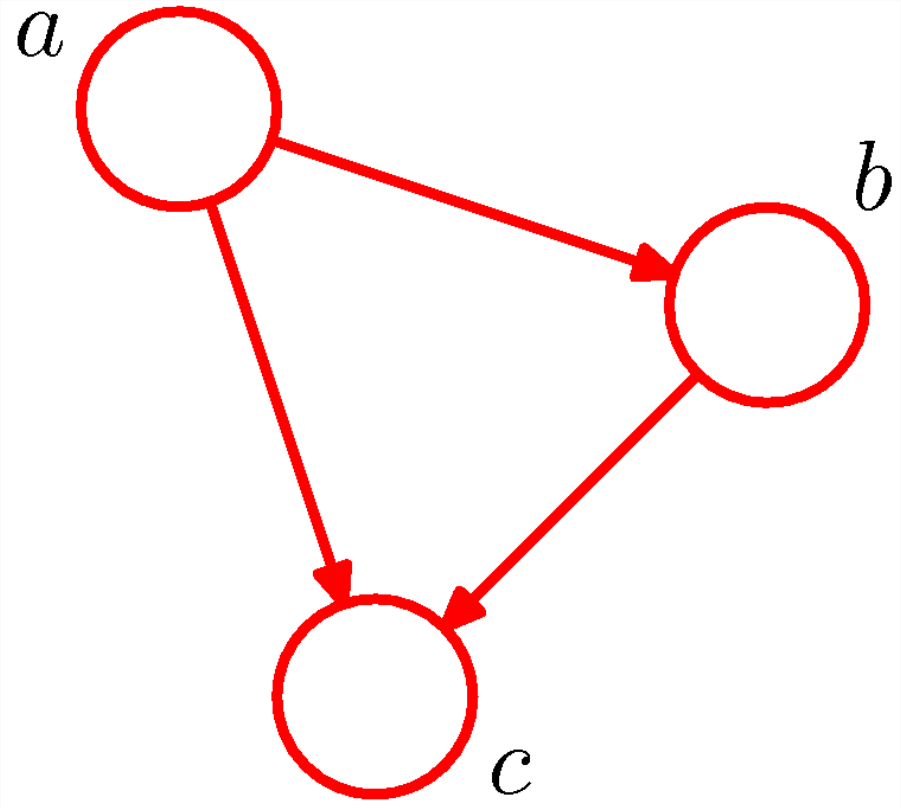
plate product

directed graph often known as Bayesian network

undirected graph often known as Markov Random Field

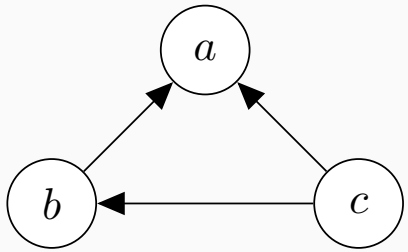


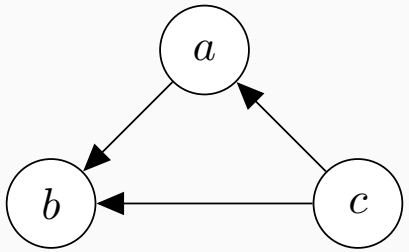
Directed Graphs

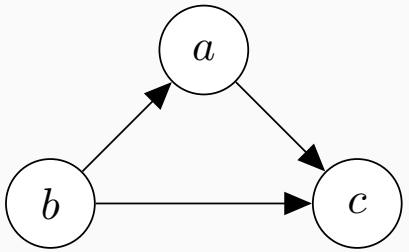


$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

Equivivalence

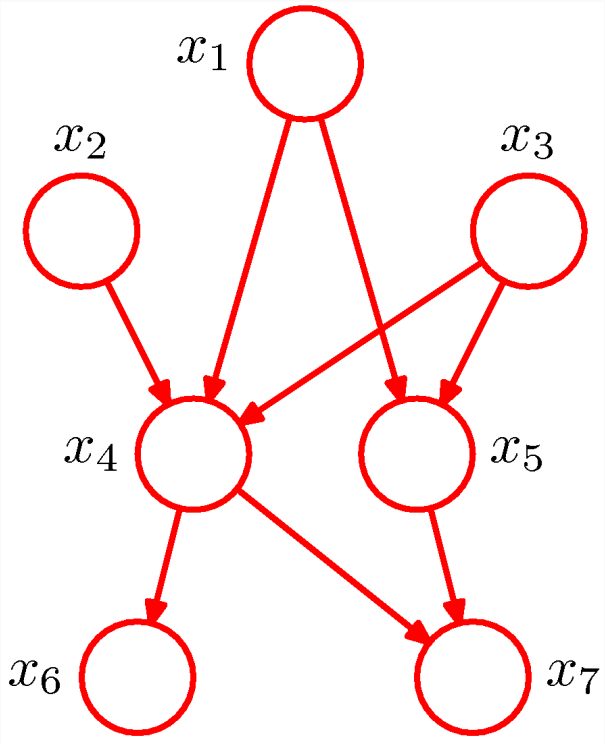






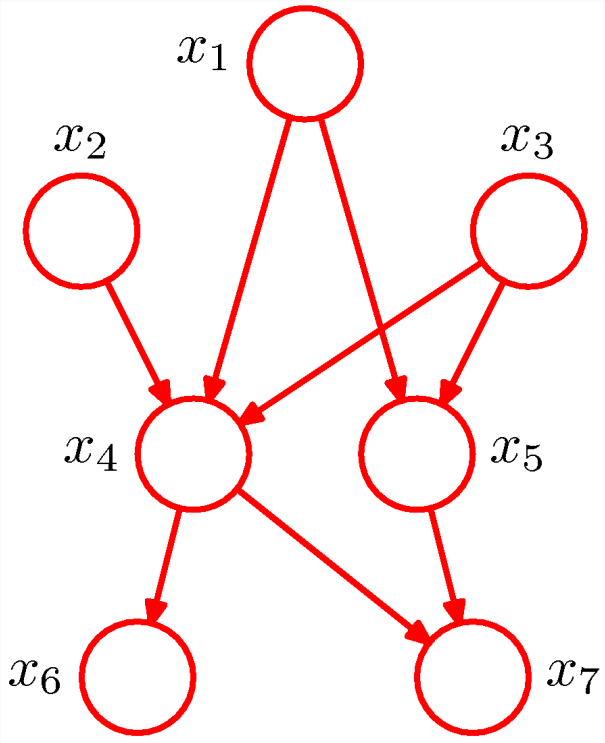
$p(a|b,c)p(b|c)p(c)$ $p(b|a,c)p(a|c)p(c)$ $p(c|a,b)p(a|b)p(b)$

Directed Graphs



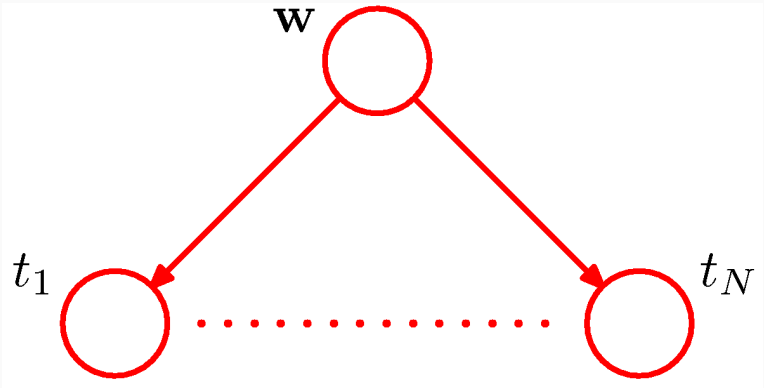
$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_5, x_4)$$

Directed Graphs



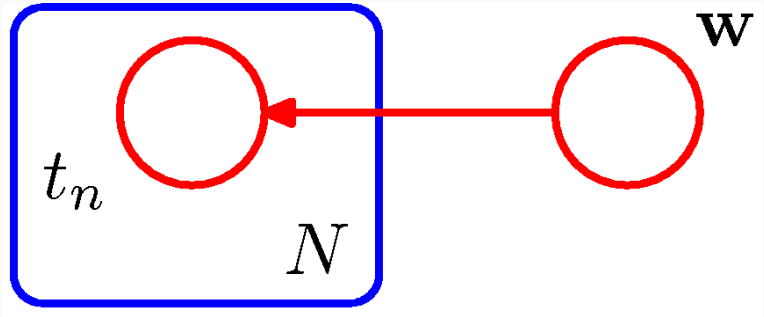
$$p(\mathbf{x}) = \prod_i p(x_i|\mathrm{pa}_i)$$

Linear Regression



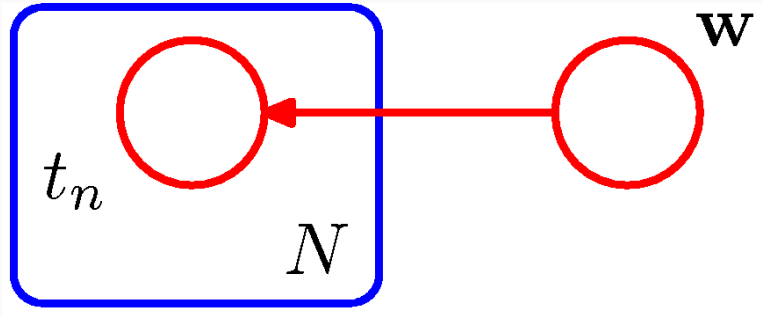
$$p(\mathbf{t}, \mathbf{W}|\mathbf{X}, \alpha, \beta) = \prod_{i}^{N} p(t_i|\mathbf{W}, x_i, \beta) p(\mathbf{W}|\alpha)$$

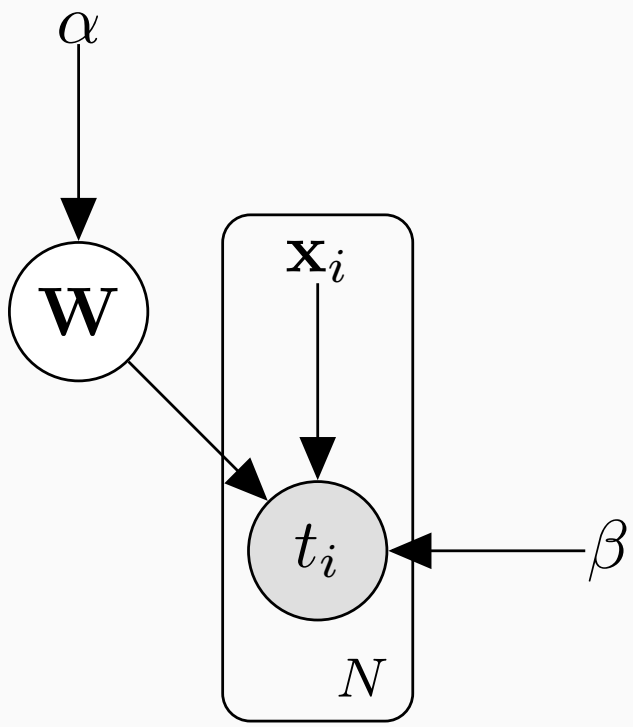
Linear Regression



$$p(\mathbf{t}, \mathbf{W}|\mathbf{X}, \alpha, \beta) = \prod_{i}^{N} p(t_i|\mathbf{W}, x_i, \beta) p(\mathbf{W}|\alpha)$$

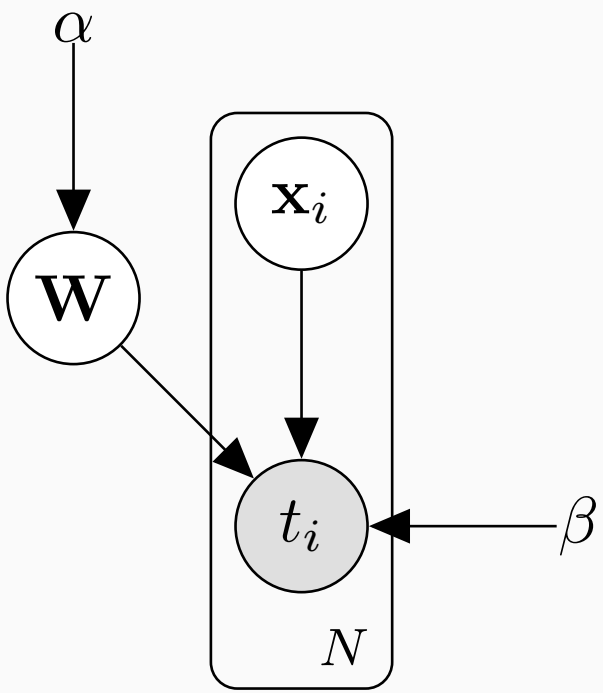
Linear Regression





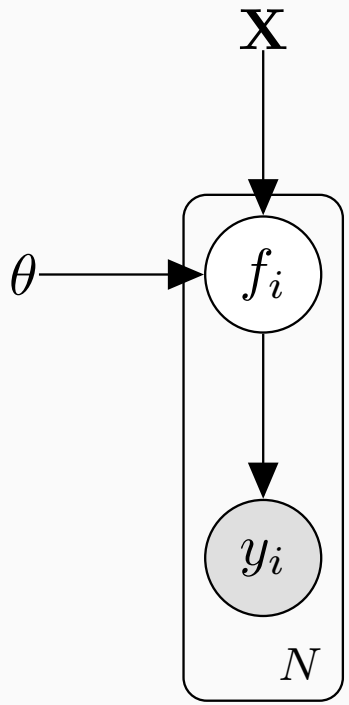
$$p(\mathbf{t}, \mathbf{W}|\mathbf{X}, \alpha, \beta) = \prod_{i}^{N} p(t_i|\mathbf{W}, x_i, \beta) p(\mathbf{W}|\alpha)$$

Unsupervised Linear Regression



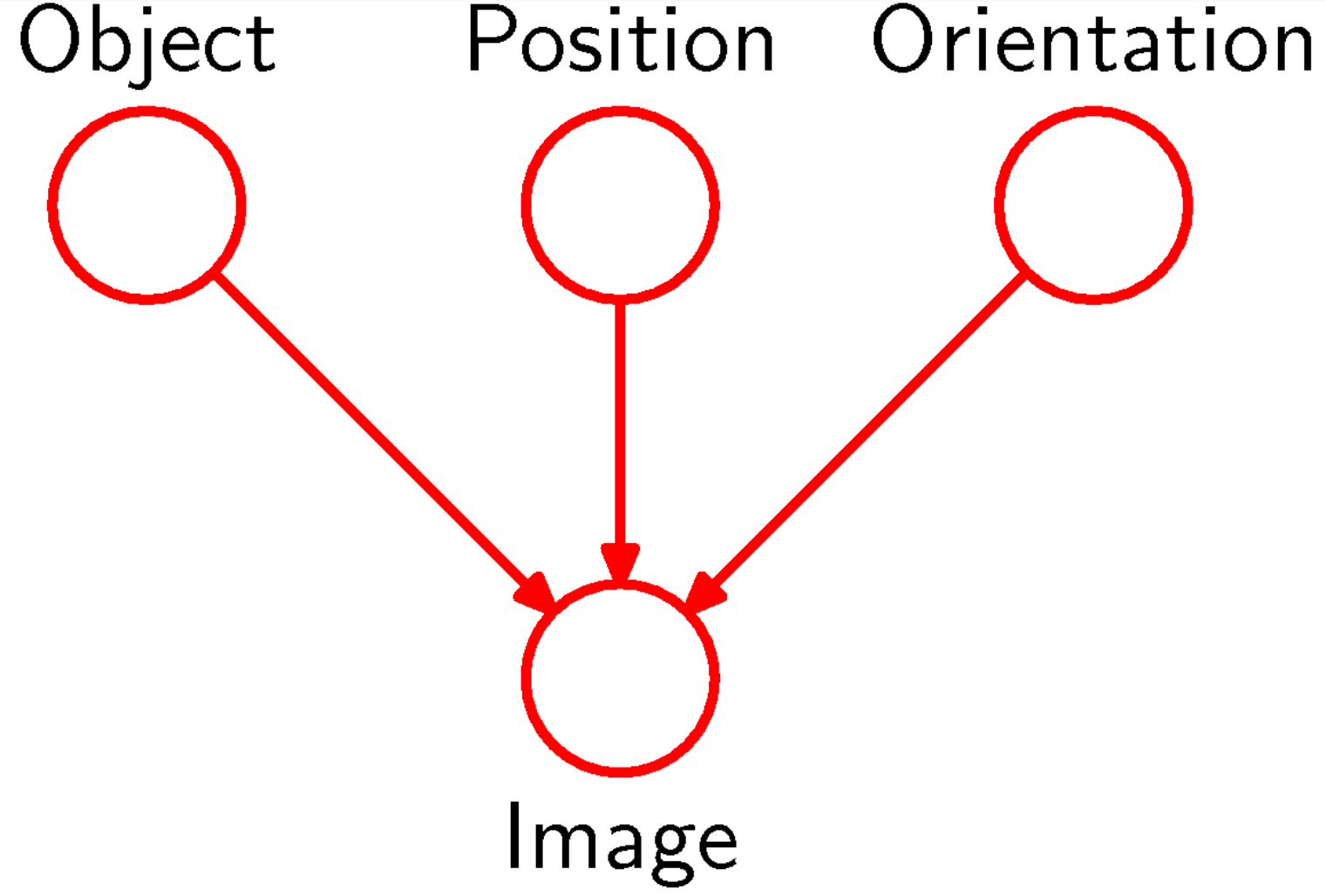
$$p(\mathbf{t}, \mathbf{x}, \mathbf{W}|\alpha, \beta) = \prod_{i}^{N} p(t_i|\mathbf{W}, x_i, \beta)p(\mathbf{W}|\alpha)p(\mathbf{x})$$

Gaussian process regression

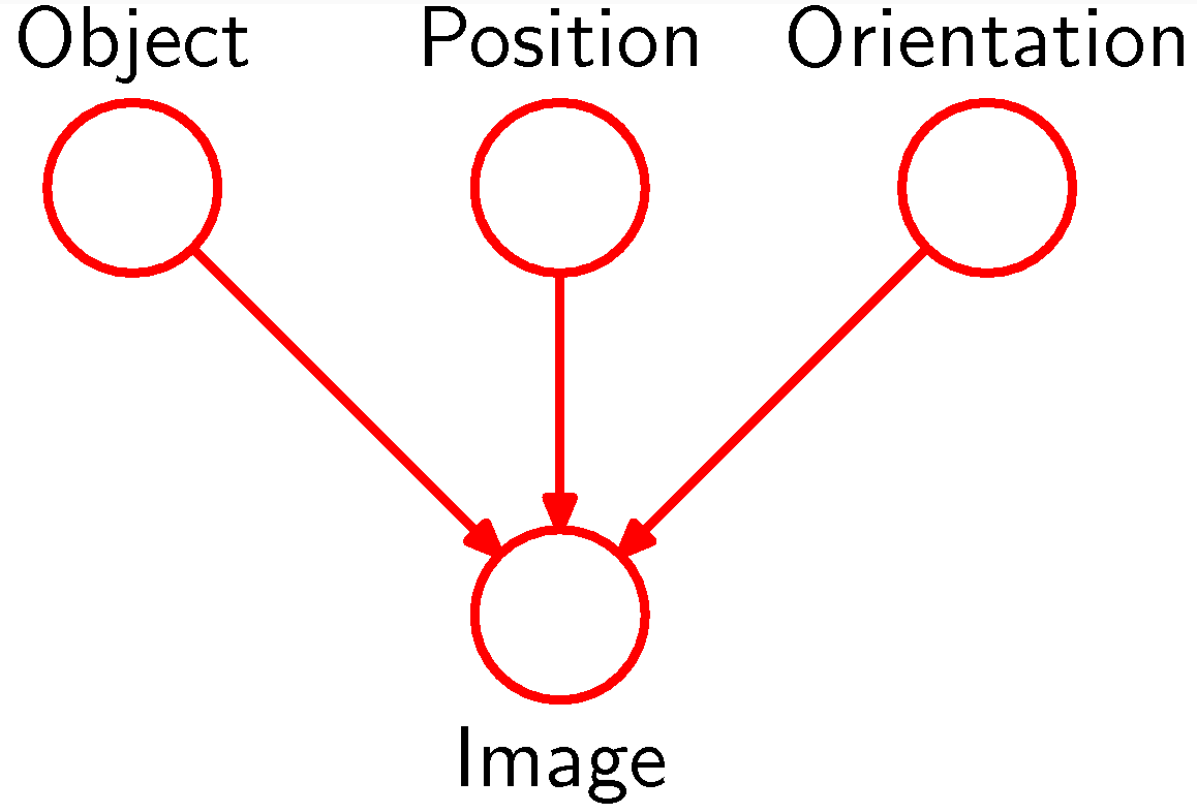


$$p(\mathbf{y}, \mathbf{f}|\mathbf{X}, \theta) = \prod_{i=1}^{N} p(y_i|f_i)p(f_i|\mathbf{X}, \theta)$$

Generative Models

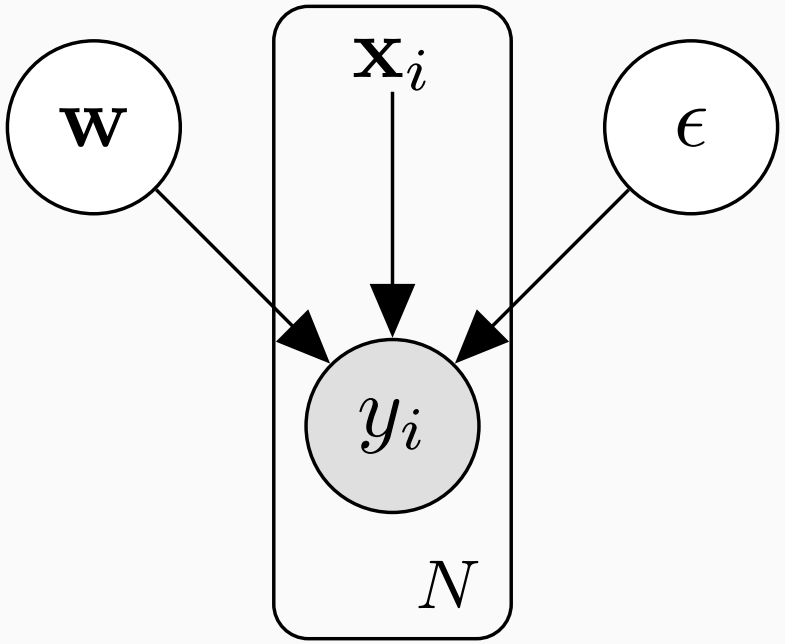


Explaining Away



- The Object variable *explains away* variance associated with objects from the image
  - $\rightarrow$ position won't contain object variations
  - $\rightarrow$ orientation won't contain object variations

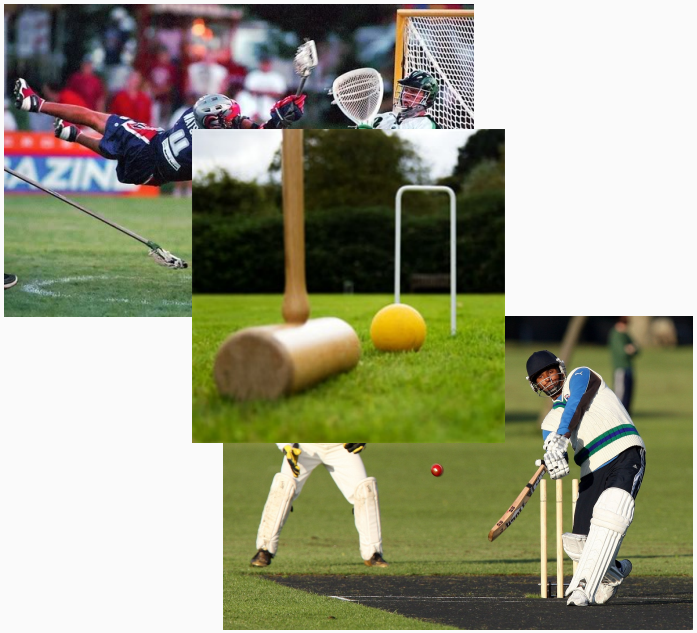
Explaining Away



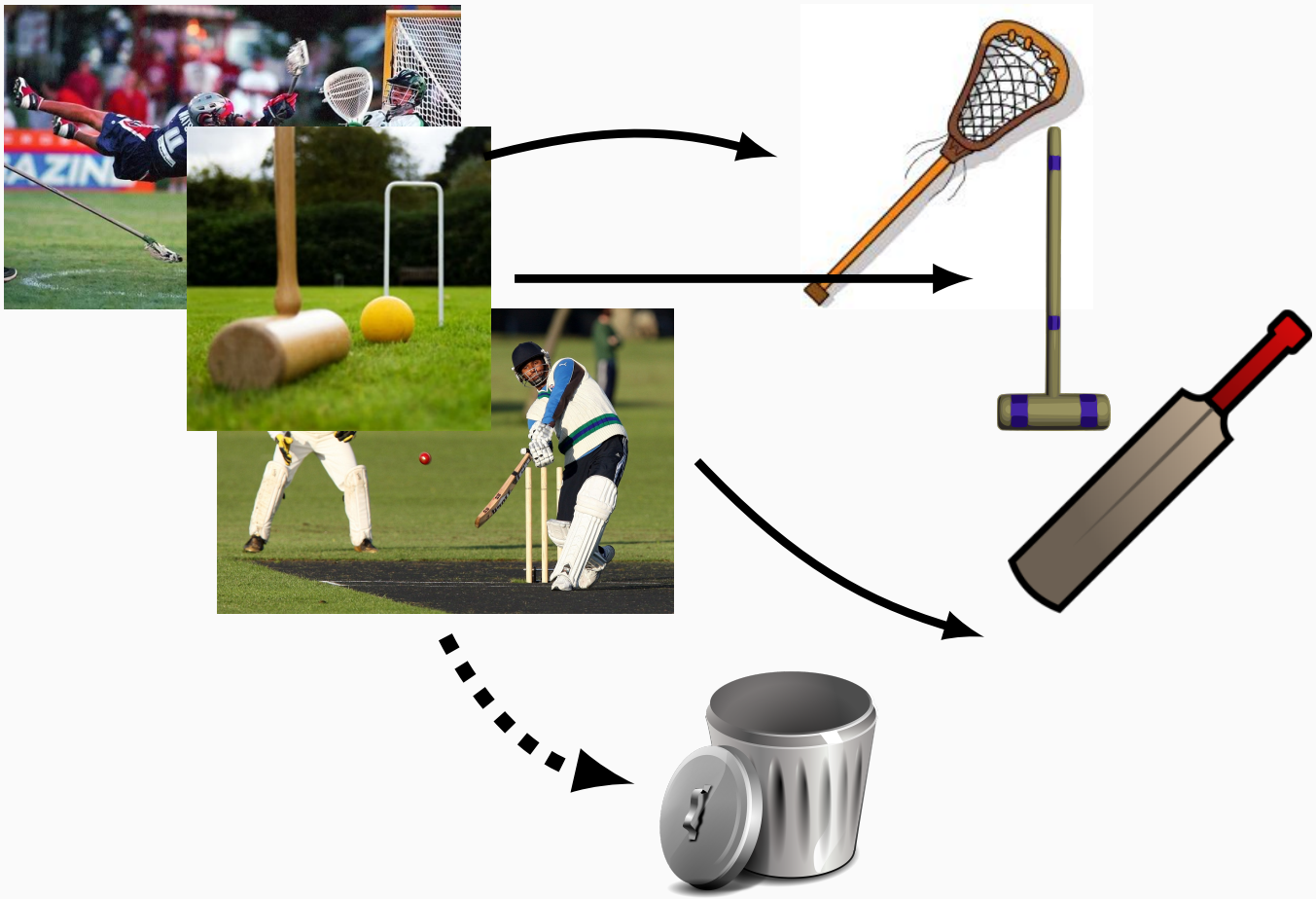
$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \epsilon \quad y_i - \epsilon = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i$$

- ϵ *explains away* the noise from the data so that $\mathbf{w}$ can represent the signal

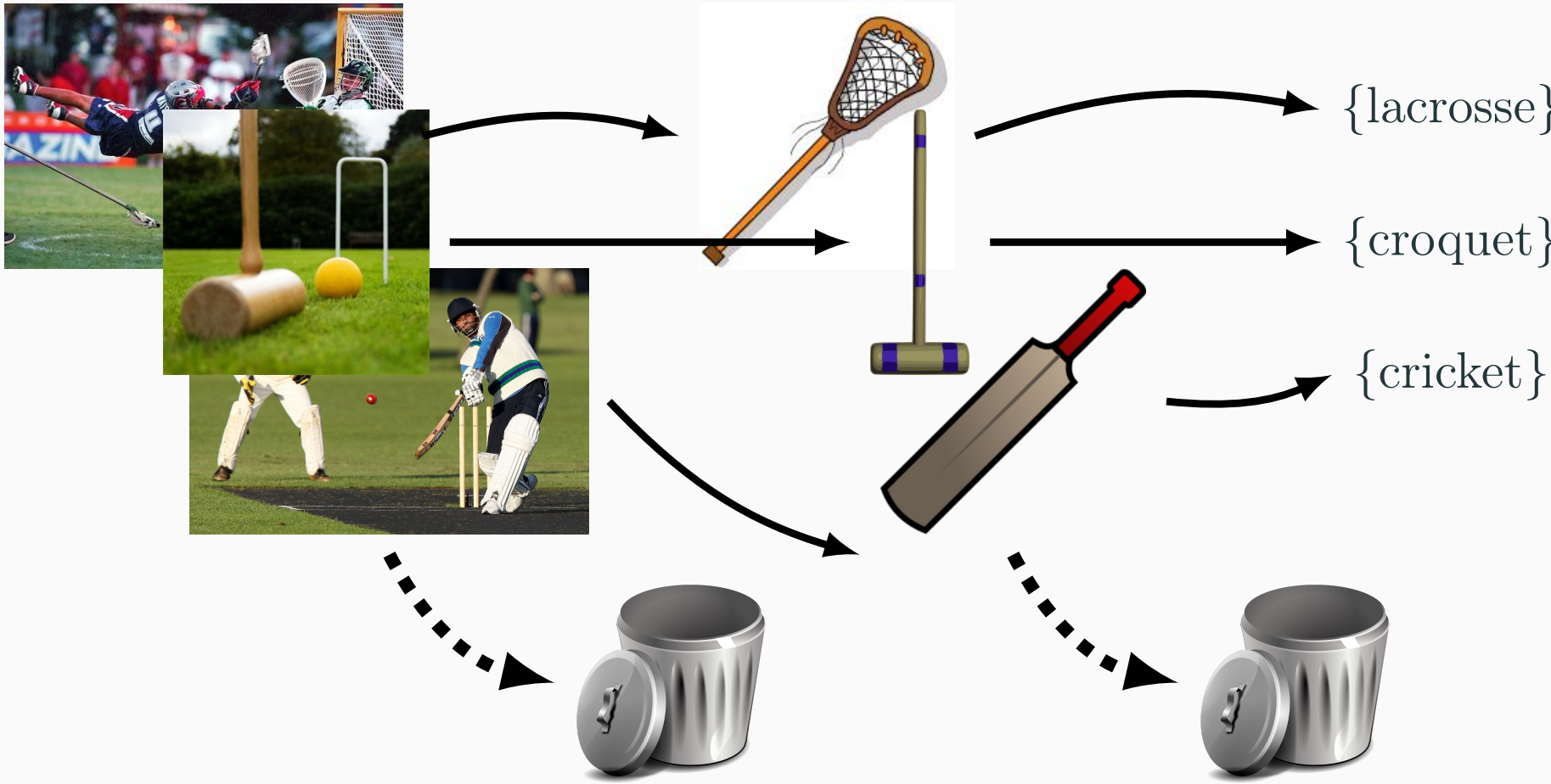
Explaining Away



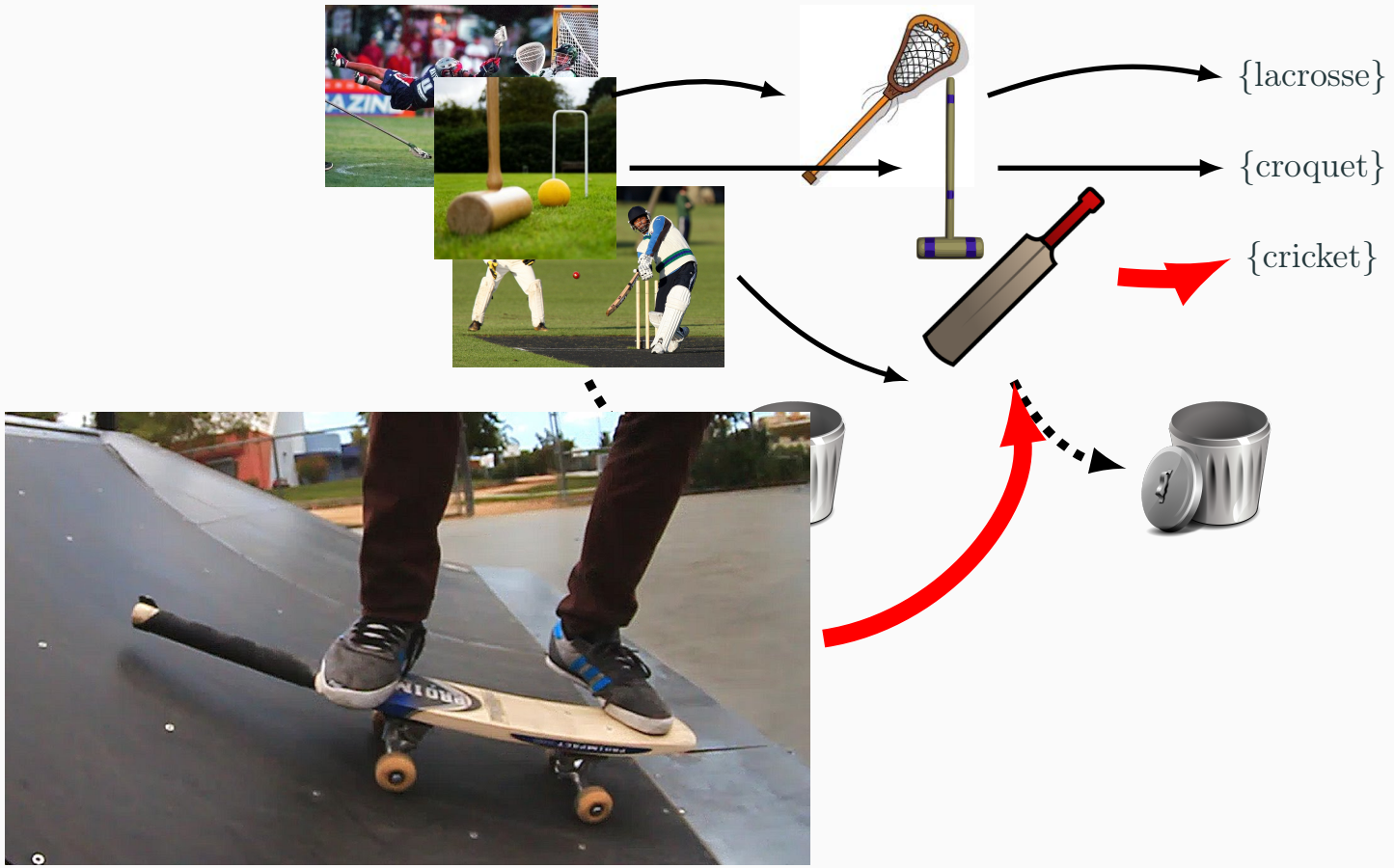
Explaining Away



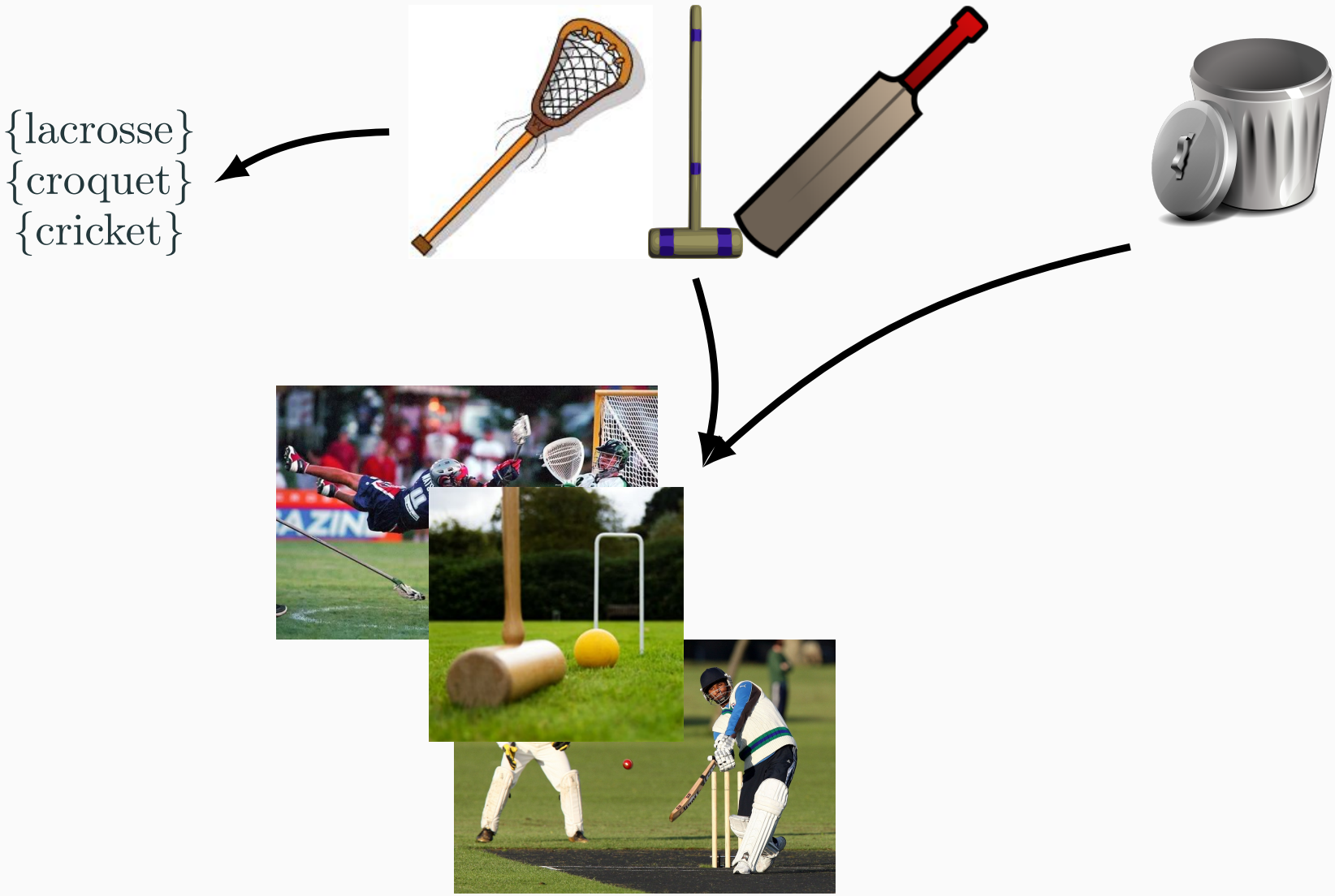
Explaining Away



Explaining Away

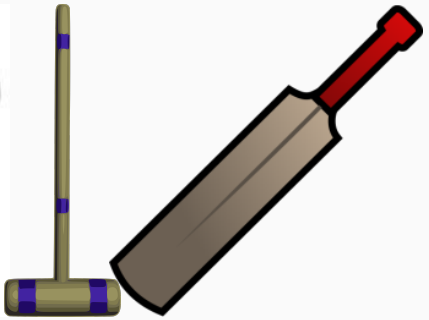


Explaining Away



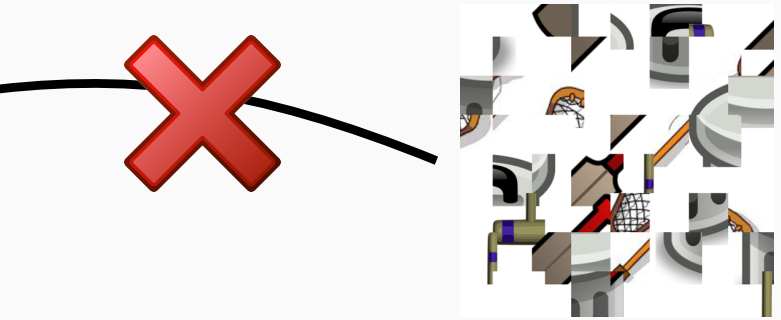
Explaining Away

{lacrosse}
{croquet}
{cricket}

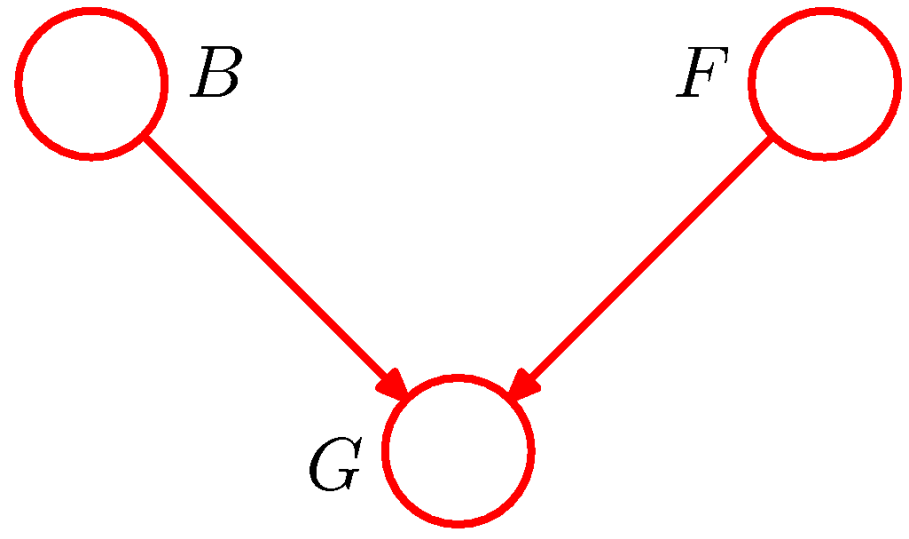


Explaining Away

{lacrosse}
{croquet}
{cricket}



Example p. 377 [1]

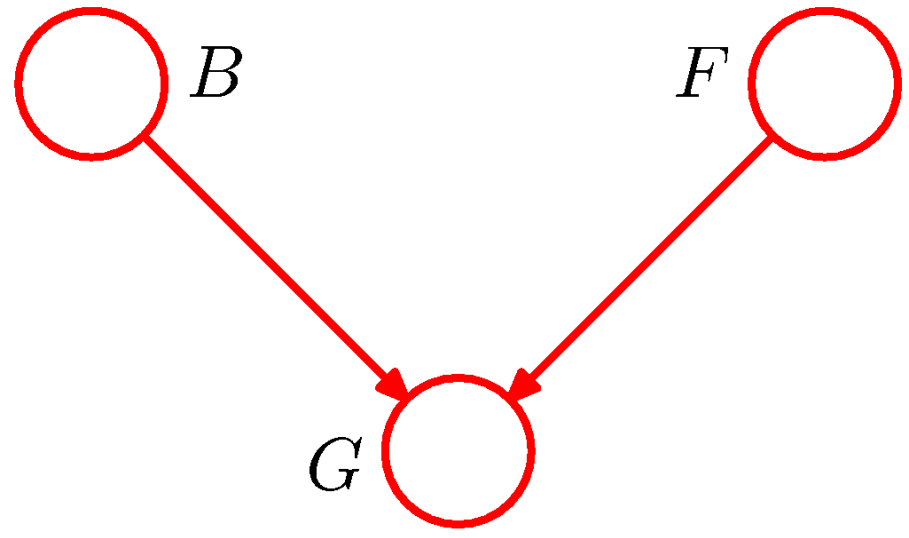


B Battery: $1 \rightarrow$ Full, $0 \rightarrow$ Empty

F Fuel Tank: $1 \rightarrow$ Full, $0 \rightarrow$ Empty

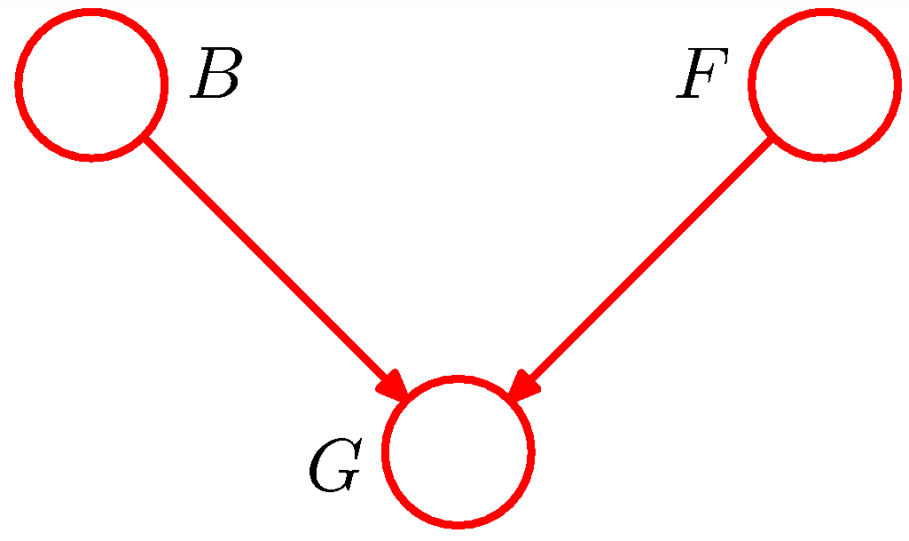
G Fuel Gauge: $1 \rightarrow$ Indicates Full, $0 \rightarrow$ Indicates Empty

Example p. 377 [1]



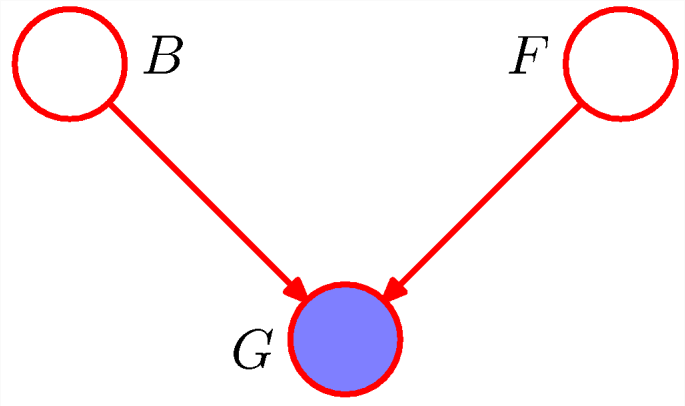
$$p(B = 1) = 0.9$$
$$p(F = 1) = 0.9$$

Example p. 377 [1]



$$p(G = 1|B = 1, F = 1) = 0.8$$
$$p(G = 1|B = 1, F = 0) = 0.2$$
$$p(G = 1|B = 0, F = 1) = 0.2$$
$$p(G = 1|B = 0, F = 0) = 0.1$$

Example p. 377 [1]



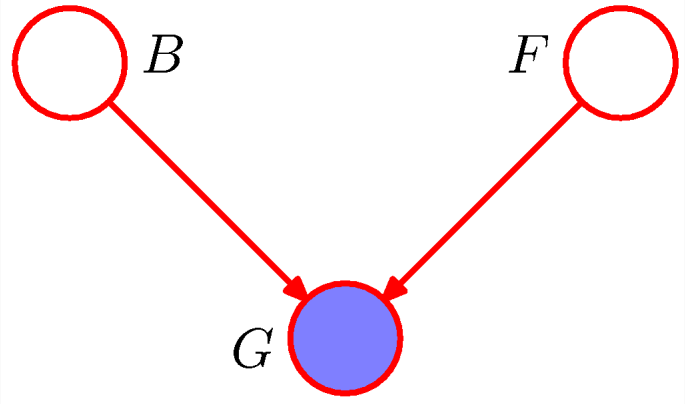
We observe an empty fuel tank $G = 0$

$$p(G = 0) = \sum_{B \in \{0,1\}} \sum_{F \in \{0,1\}} p(G = 0|B, F)p(B)p(F) = 0.315$$

$$p(G = 0|F = 0) = \sum_{B \in \{0,1\}} p(G = 0|B, F = 0)p(B) = 0.81$$

$$p(F = 0|G = 0) = \frac{p(G = 0|F = 0)p(F = 0)}{p(G = 0)} \approx 0.257$$

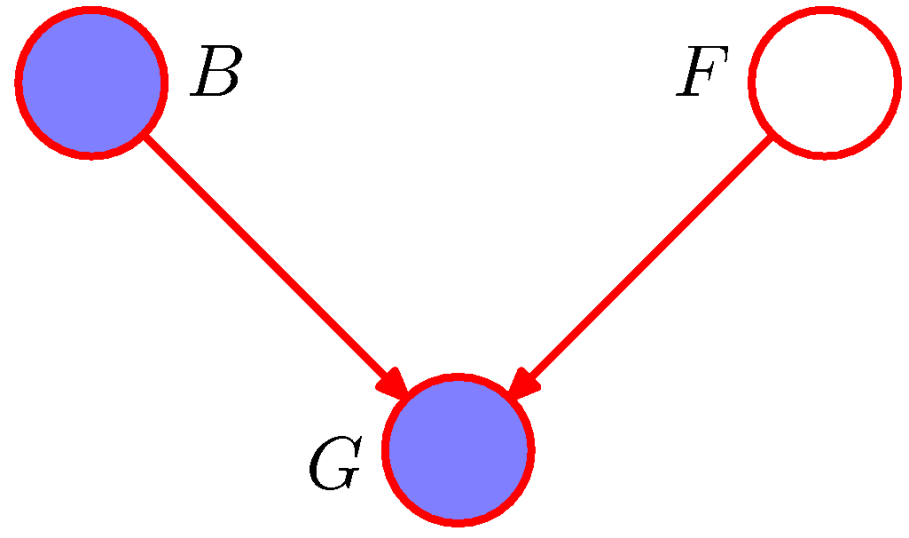
Example p. 377 [1]



$$p(F = 0|G = 0) > p(F = 0)$$

The gauge does provide information about the tank

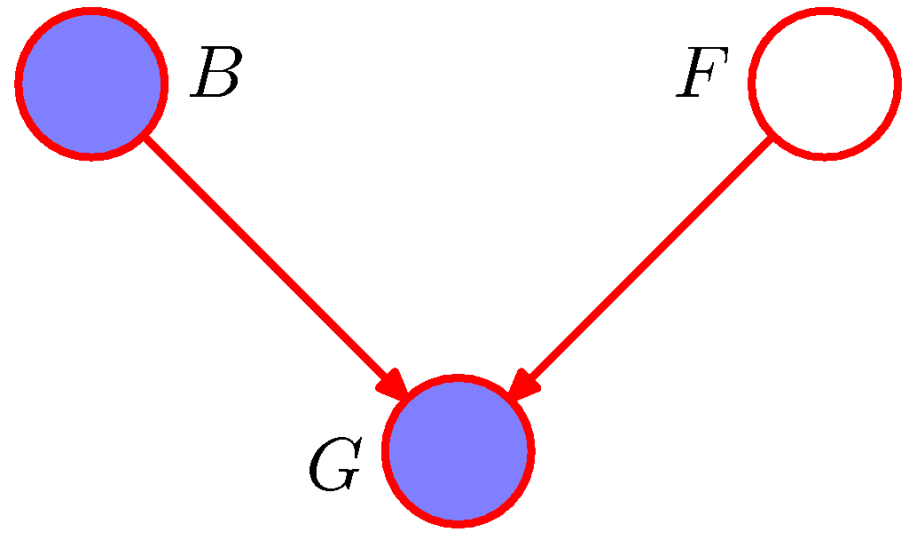
Example p. 377 [1]



We observe an empty fuel tank $G = 0$ and Battery empty $B = 0$

$$p(F = 0|G = 0, B = 0) = \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F \in \{0,1\}} p(G = 0|B = 0, F)p(F)} \approx 0.111$$

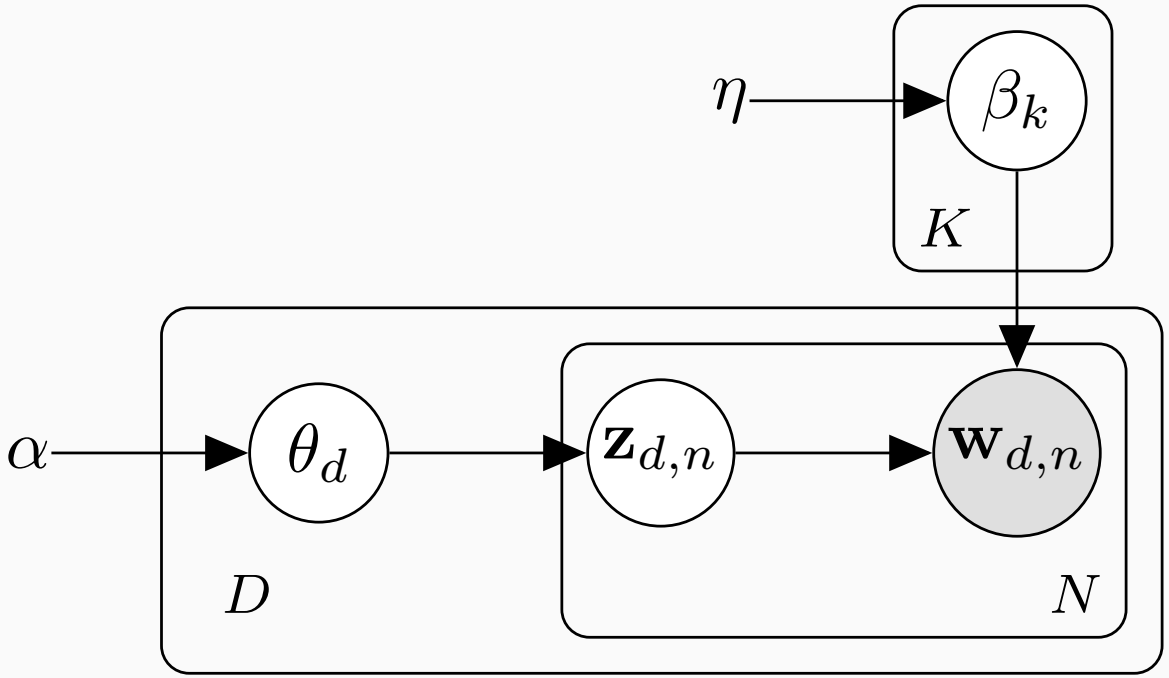
Example p. 377 [1]



$$p(F = 0|G = 0) > p(F = 0|G = 0, B = 0) > P(F = 0)$$

Knowing that the battery is empty explains away the information about the Gauge indicating empty

Graphical Models



$$p(w, z, \theta, \beta|\eta, \alpha) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{n=1}^{N} p(w_{d,n}|\beta, z_{d,n}) p(z_{d,n}|\theta_d)$$

Summary

Summary

- The marginal likelihood implements Occams Razor, but what is simple?
- Graphical models is just a language of what we have been doing
- Much easier to talk about when thinking of new models
- Directed graphical models implies building up conditional probabilities

eof

References


Christopher M. Bishop.
Pattern Recognition and Machine Learning (Information Science and Statistics).
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.