UNIVERSITY OF CAMBRIDGE

Machine Learning and the Physical World

Lecture 2 : Quantification of Beliefs

Carl Henrik Ek - che29@cam.ac.uk

10th of October, 2023

http://carlhenrik.com

Today

- Why understanding our ignorance is not just desirable but necessary for learning

Today

- Why understanding our ignorance is not just desirable but necessary for learning
- Why knowledge is subjective or relative

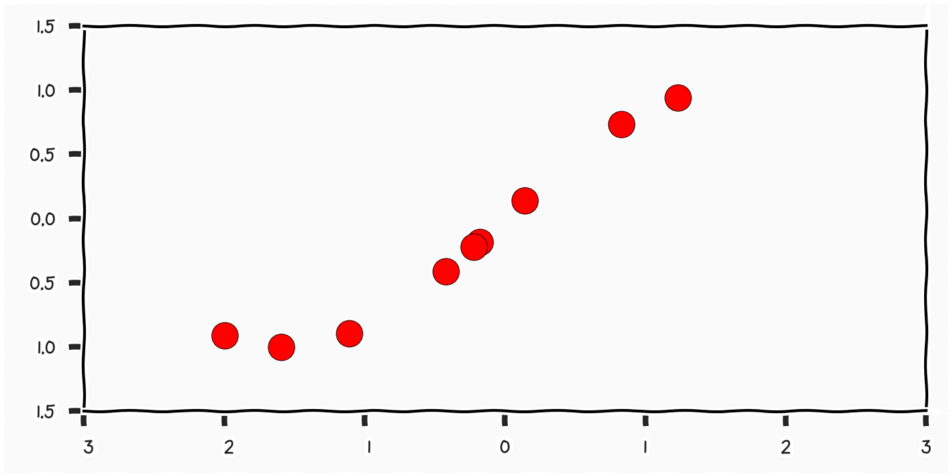
Today

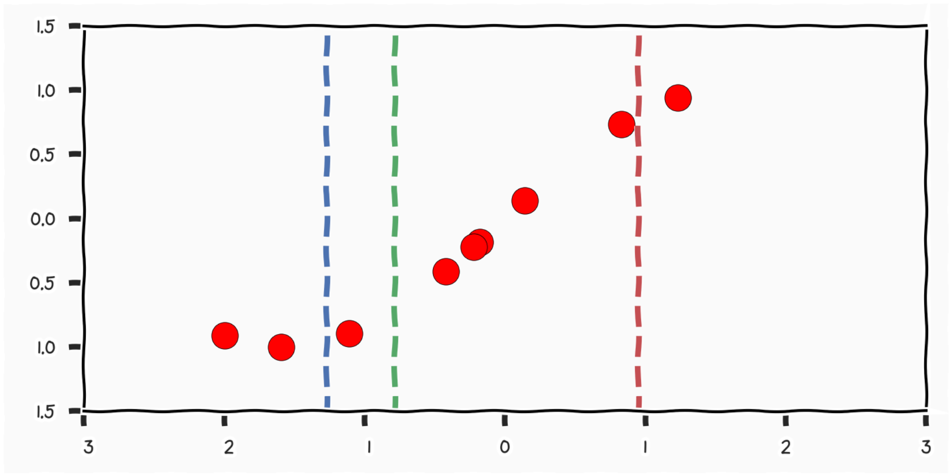
- Why understanding our ignorance is not just desirable but necessary for learning
- Why knowledge is subjective or relative
- Re-cap of linear regression

Inductive Reasoning

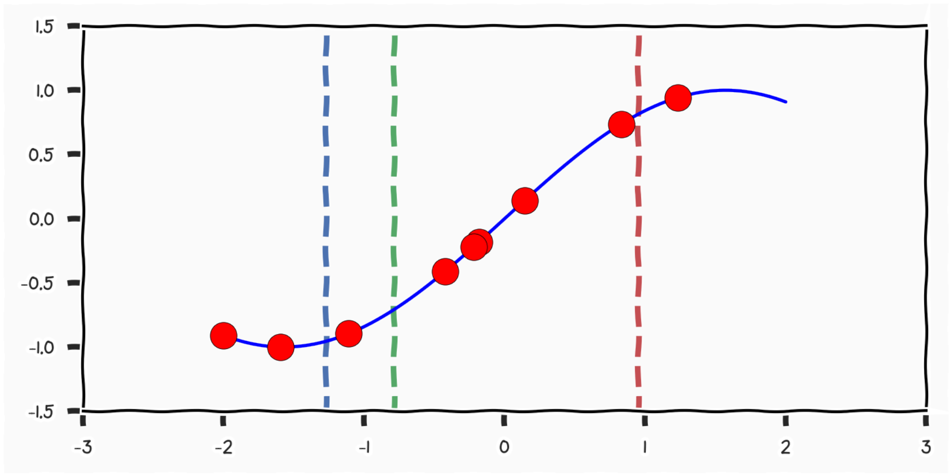
"In inductive inference, we go from the specific to the general. We make many observations, discern a pattern, make a generalization, and infer an explanation or a theory"

– Wassertheil-Smoller


1,5
1,0
0,5
0,0
0,5
1,0
1,5
3
2
1
0
1
2
3

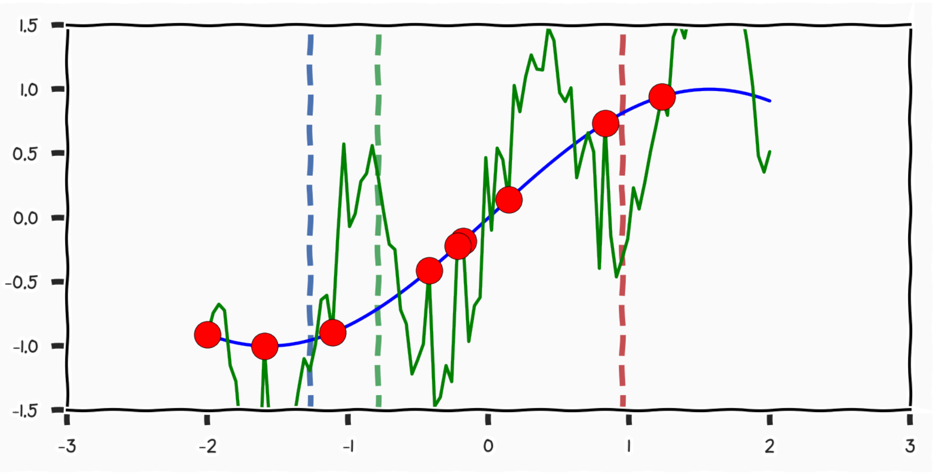


1,5
1,0
0,5
0,0
0,5
1,0
1,5
3
2
1
0
1
2
3


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
1
2
3

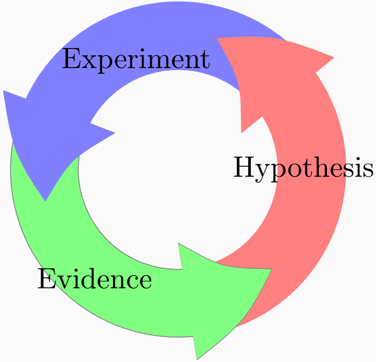
Inductive Reasoning

Unlike deductive arguments, inductive reasoning allows for the possibility that the conclusion is false, even if all of the premises are true.



1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
1
2
3

The Scientific Principle

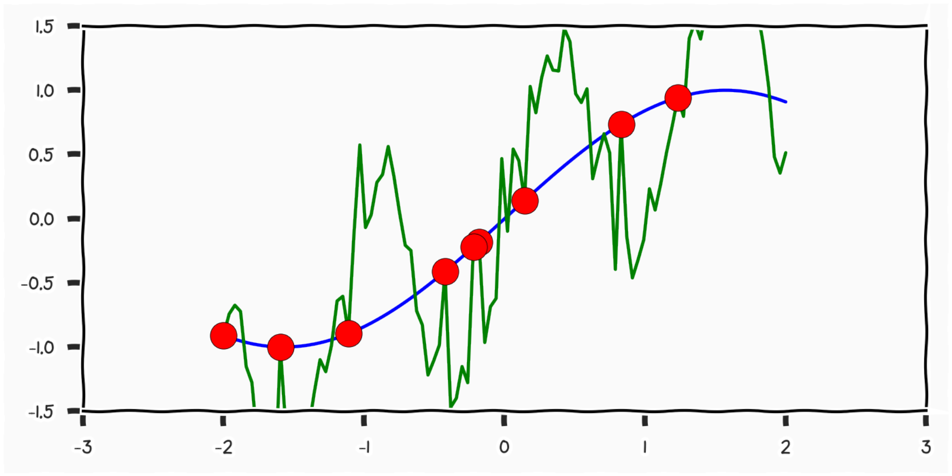


$$\text{Data} + \text{Model} \overbrace{\longrightarrow}^{Compute} \text{Prediction}$$

"The Machine Learning Principle"[1]

"There is a notion of success . . . which I think is novel in the history of science. It interprets success as approximating unanalyzed data."
– Prof. Noam Chomsky

[1] Chomsky et al., 1980



1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
1
2
3

- $\mathcal{H}$ space of Hypothesis

- $\mathcal{H}$ space of Hypothesis
- $\mathcal{A}$ learning algorithm

- $\mathcal{H}$ space of Hypothesis
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

- $\mathcal{H}$ space of Hypothesis
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- $\mathcal{S} \sim P(\mathcal{X} \times \mathcal{Y})$

Learning Theory

- $\mathcal{H}$ space of Hypothesis
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- $\mathcal{S} \sim P(\mathcal{X} \times \mathcal{Y})$
- $\ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x, y)$ loss function

Statistical Learning

$$e(\mathcal{S}, \mathcal{A}, \mathcal{H}) = \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x, y)\right]$$

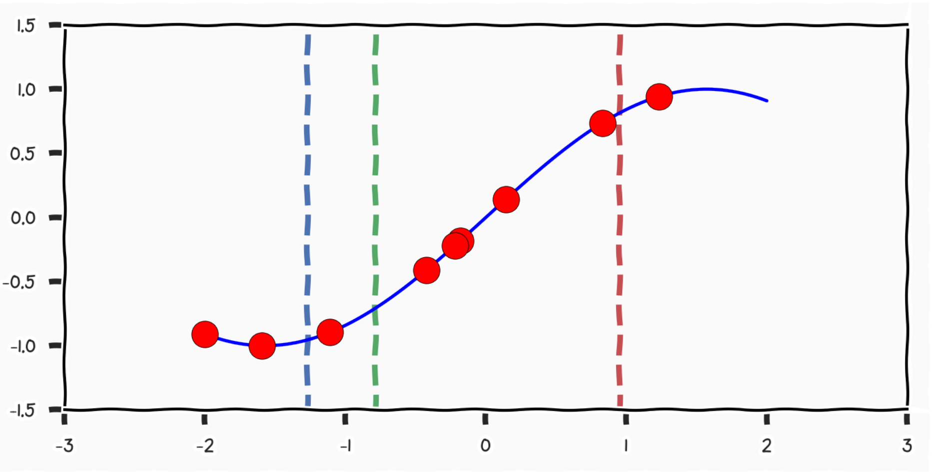
Statistical Learning

$$
\begin{aligned}
e(\mathcal{S}, \mathcal{A}, \mathcal{H}) &= \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x, y)\right] \\
&= \int \ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x, y) p(x, y) \mathrm{d}x \mathrm{d}y
\end{aligned}
$$

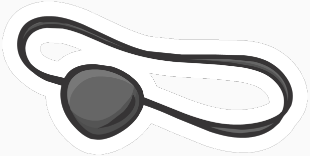
$$\overbrace{e(\mathcal{S}, \mathcal{A}, \mathcal{H})}^{\text{True Risk}} = \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x, y)\right]$$
$$= \int \ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x, y) p(x, y) \mathrm{d}x \mathrm{d}y$$
$$\approx \underbrace{\frac{1}{M} \sum_{n=1}^{M} \ell(\mathcal{A}_{\mathcal{H}}(\mathcal{S}), x_n, y_n)}_{\text{Empirical Risk}}$$

No Free Lunch

We can come up with a combination of $\{\mathcal{S}, \mathcal{A}, \mathcal{H}\}$ that are equvivalent under the empirical risk that makes true risk take an arbitary value


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-3
-2
-1
0
1
2
3

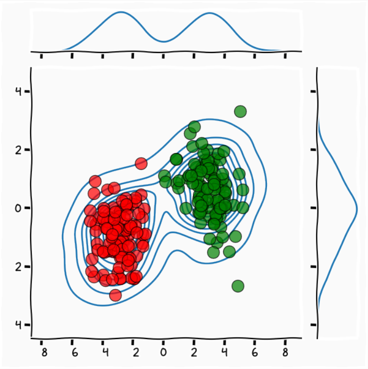
Assumptions: Algorithms



Statistical Learning

$\mathcal{A}_{\mathcal{H}}(\mathcal{S})$

Assumptions: Biased Sample



Statistical Learning

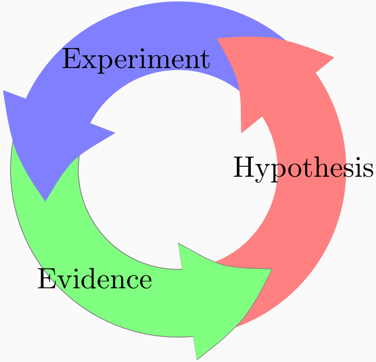
$$\mathcal{A}_{\mathcal{H}}(\mathcal{S})$$

Assumptions: Hypothesis space

Statistical Learning

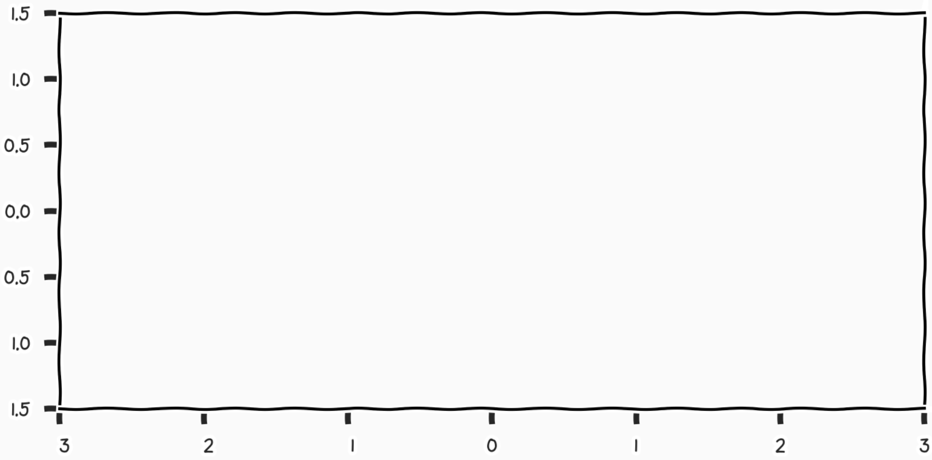
$$\mathcal{A}_{\mathcal{H}}(\mathcal{S})$$

The Scientific Principle

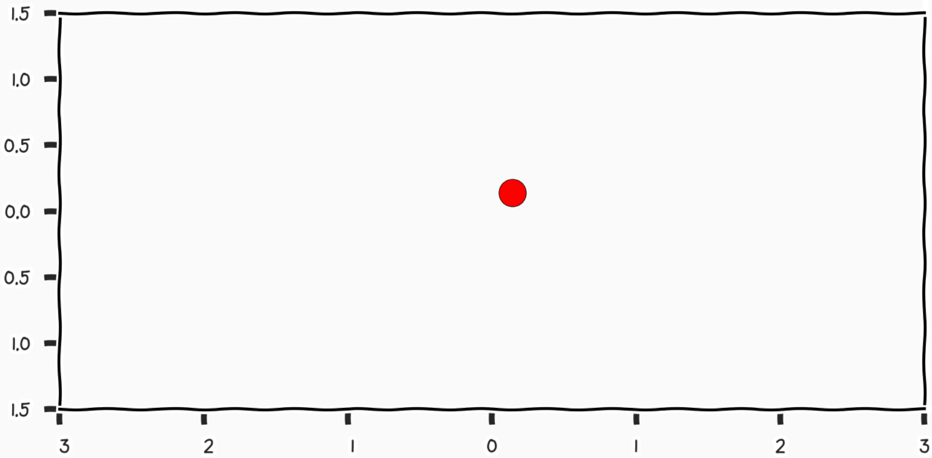


$$\text{Data + Model} \overbrace{\longrightarrow}^{Compute} \text{Prediction}$$

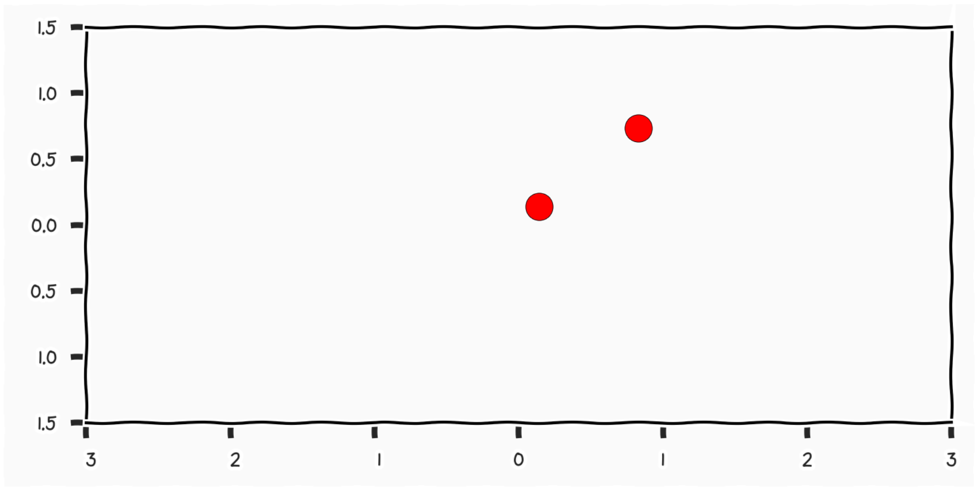
Example

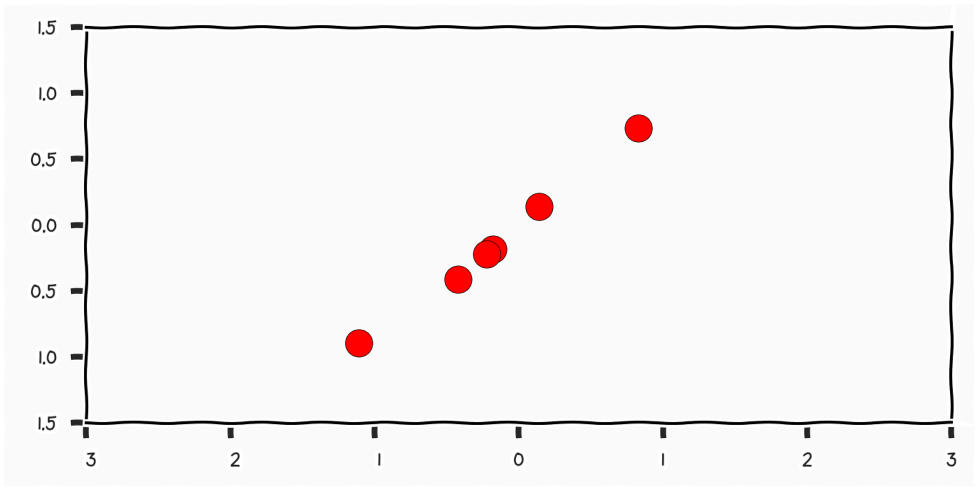


Example

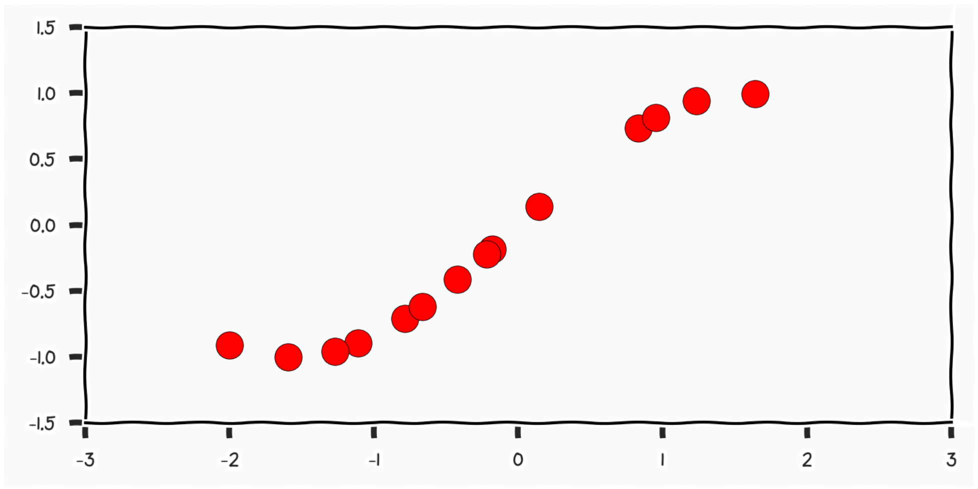


Example

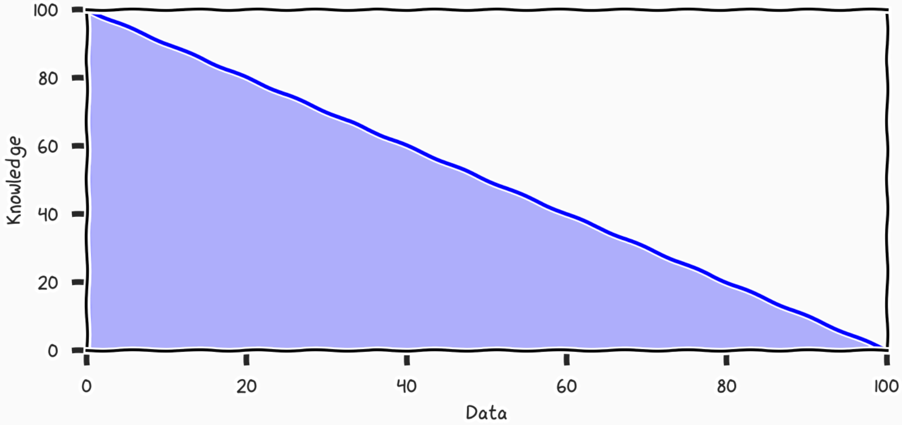



1.5
1.0
0.5
0.0
0.5
1.0
1.5
3
2
1
0
1
2
3

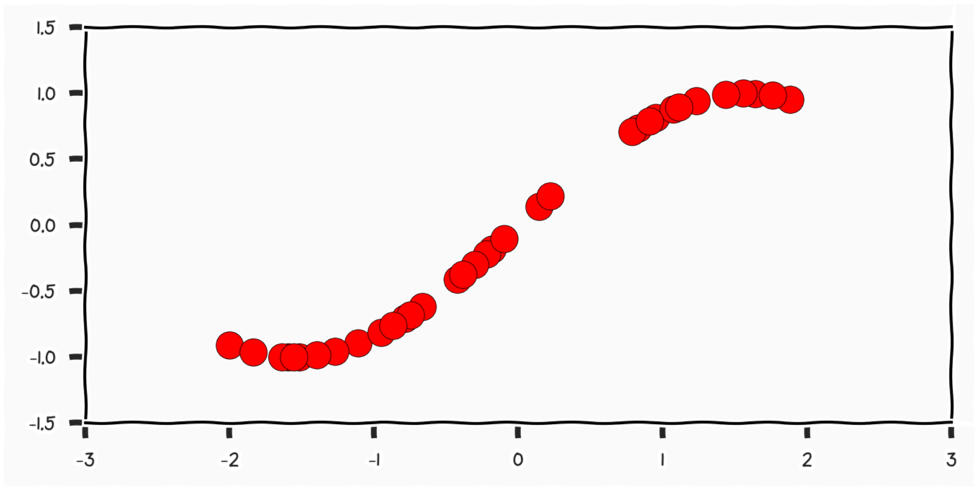
Example



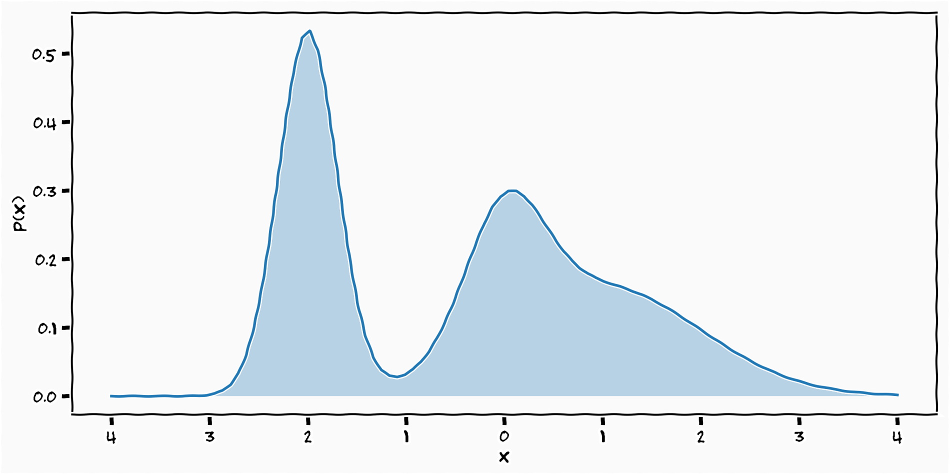
Data and Beliefs



Example



Encoding Beliefs



Manipulation of Beliefs

Sum Rule

$$p(y) = \begin{cases} \sum_{\forall \theta \in \Theta} p(y, \theta) \\ \\ \int p(y, \theta) \mathrm{d}\theta \end{cases}$$

Product Rule

$$p(y, \theta) = p(y \mid \theta) p(\theta)$$

Bayes' "Rule"

$$p(y, \theta) = p(y|\theta)p(\theta)$$

Bayes' "Rule"

$$p(y, \theta) = p(y|\theta)p(\theta)$$
$$p(y, \theta) = p(\theta|y)p(y)$$

Bayes' "Rule"

$$p(y, \theta) = p(y|\theta)p(\theta)$$
$$p(y, \theta) = p(\theta|y)p(y)$$
$$p(\theta|y)p(y) = p(y|\theta)p(\theta)$$

Bayes' "Rule"

$$
\begin{aligned}
p(y, \theta) &= p(y|\theta)p(\theta) \\
p(y, \theta) &= p(\theta|y)p(y) \\
p(\theta|y)p(y) &= p(y|\theta)p(\theta) \\
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)}
\end{aligned}
$$

Bayes' "Rule"

$$
\begin{aligned}
p(y, \theta) &= p(y|\theta)p(\theta) \\
p(y, \theta) &= p(\theta|y)p(y) \\
p(\theta|y)p(y) &= p(y|\theta)p(\theta) \\
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\
&= \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\mathrm{d}\theta}
\end{aligned}
$$

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

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

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

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

Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

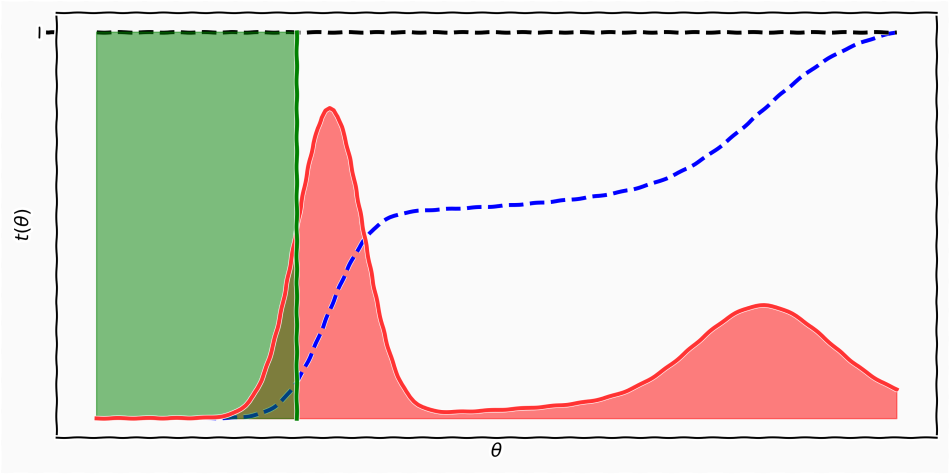
Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \mathrm{d}\theta$$

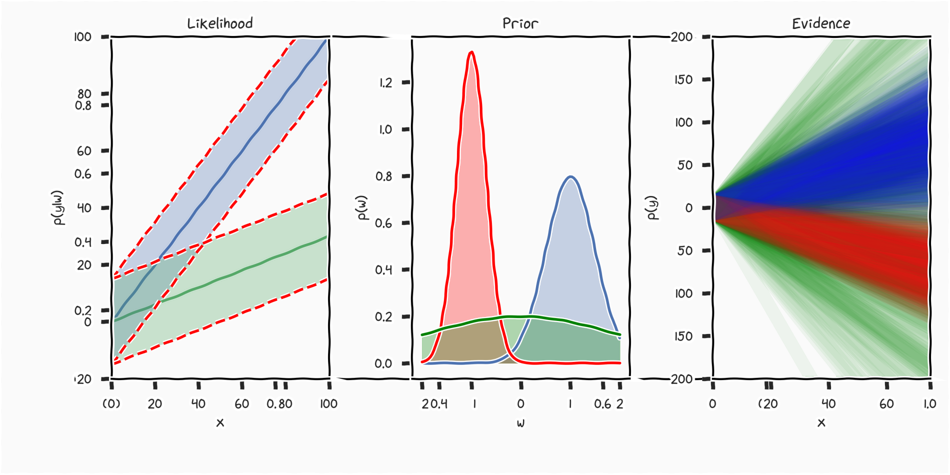
Marginalisation

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) \underbrace{p(\theta)\mathrm{d}\theta}_{\mathrm{d}t(\theta)}$$

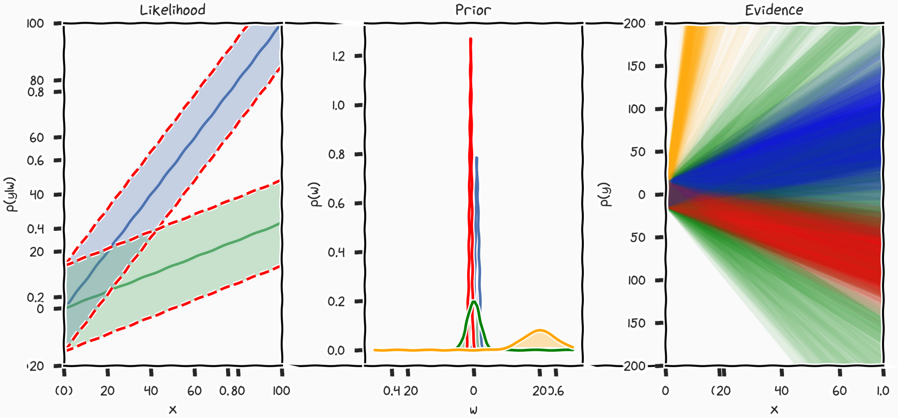
Marginalisation



Marginalisation



Marginalisation



Laplace, 1814

"One sees, from this Essay, that the theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it."

– Simon Laplace

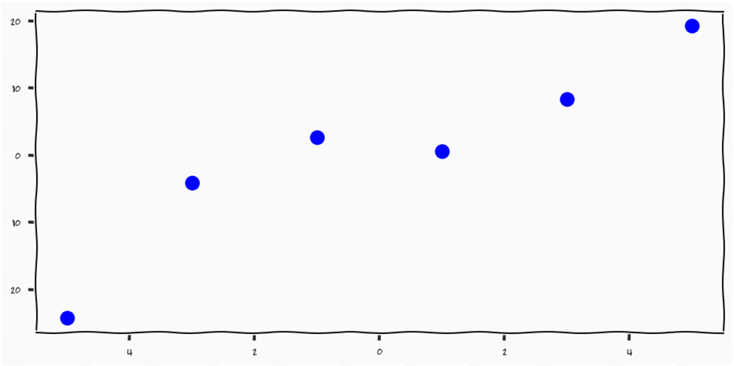
Parametrising our Ignorance

Data Today

Model Today and Thursday

Computation Week 4

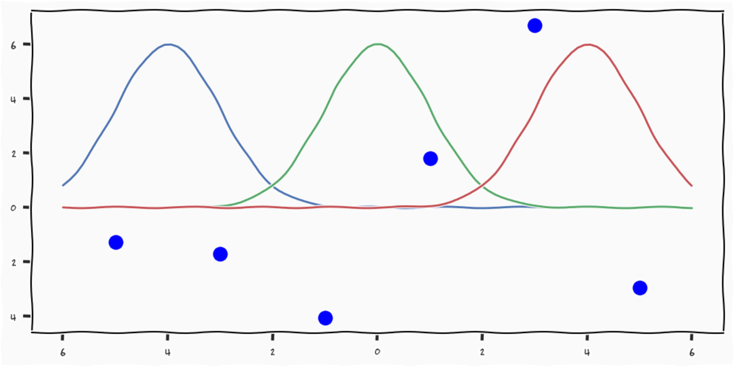
Linear Regression



- Linear function in both parameters and data

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots w_D x_D = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \{D = 1\} w_0 + w_1 \cdot x$$

Linear Regression



- Linear function only in parameters

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Linear Regression

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

- Given observations of data pairs $\mathcal{D} = \{y_i, \mathbf{x}_i\}_{i=1}^{N}$ can we infer what $\mathbf{w}$ should be

Task 1 define a likelihood (model)

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 2 define an assumption/belief over all hypothesis (model)

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 2 define an assumption/belief over all hypothesis (model)

what types of models do I think are more probable than others

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 2 define an assumption/belief over all hypothesis (model)

what types of models do I think are more probable than others

Task 3 update my belief with new observations (data)

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 2 define an assumption/belief over all hypothesis (model)

what types of models do I think are more probable than others

Task 3 update my belief with new observations (data)

formulate posterior (compute)

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 2 define an assumption/belief over all hypothesis (model)

what types of models do I think are more probable than others

Task 3 update my belief with new observations (data)

formulate posterior (compute)

Task 4 predict using my new belief (predict)

Task 1 define a likelihood (model)

what output do I consider likely under a given hypothesis?

Task 2 define an assumption/belief over all hypothesis (model)

what types of models do I think are more probable than others

Task 3 update my belief with new observations (data)

formulate posterior (compute)

Task 4 predict using my new belief (predict)

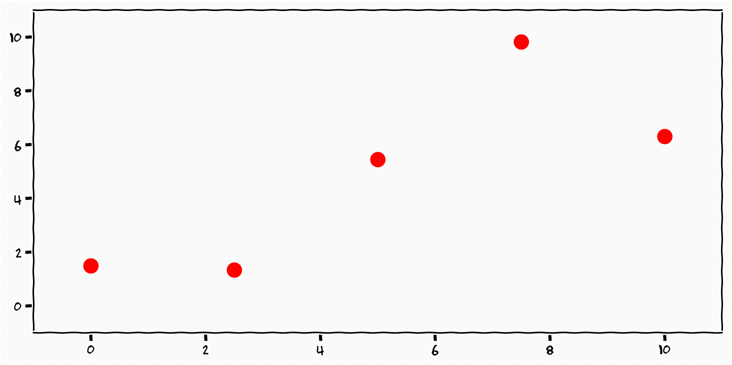
formulate predictive distribution

Linear Regression

$$y = f(\mathbf{x}, \mathbf{w}) + \epsilon = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \beta^{-1} I)$$

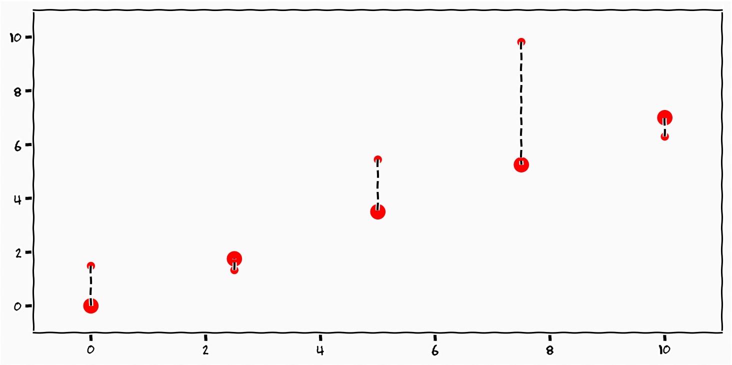
- We assume that we have been given data pairs $\{y_i, \mathbf{x}_i\}_{i=1}^{N}$ corrupted by addative noise
- We assume that the distribution of the noise follows a Gaussian

Explaining Away



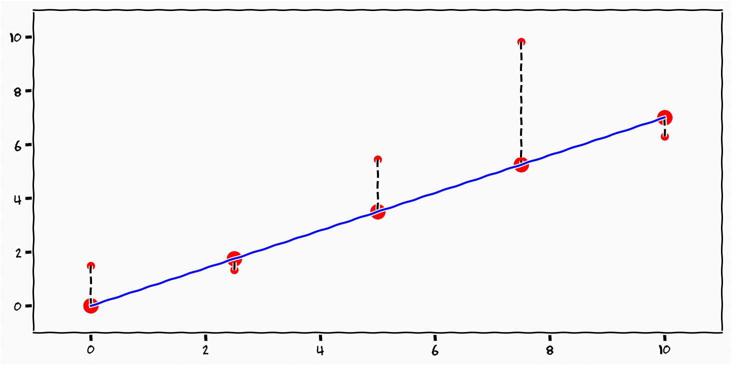
$$y = \mathbf{w}^{\mathrm{T}} x + \epsilon$$

Explaining Away



$$y - \epsilon = \mathbf{w}^{\mathrm{T}} x$$

Explaining Away



$$\tilde{y} = \mathbf{w}^{\mathrm{T}} x$$

Likelihood

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

Likelihood

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

Likelihood

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} = \epsilon$$

$$y - \mathbf{w}^{\mathrm{T}}\mathbf{x} \sim \mathcal{N}(\epsilon | 0, \beta^{-1} I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon - 0)\beta(\epsilon - 0)}$$

Likelihood

$$
\begin{aligned}
y &= \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon \\
y - \mathbf{w}^{\mathrm{T}}\mathbf{x} &= \epsilon \\
y - \mathbf{w}^{\mathrm{T}}\mathbf{x} &\sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)} \\
\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) &= \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})\beta(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})}
\end{aligned}
$$

Likelihood

$$
\begin{aligned}
y &= \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon \\
y - \mathbf{w}^{\mathrm{T}}\mathbf{x} &= \epsilon \\
y - \mathbf{w}^{\mathrm{T}}\mathbf{x} &\sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)} \\
\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) &= \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})\beta(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})} \\
\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) &= \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)
\end{aligned}
$$

Likelihood

$$
\begin{aligned}
y &= \mathbf{w}^{\mathrm{T}}\mathbf{x} + \epsilon \\
y - \mathbf{w}^{\mathrm{T}}\mathbf{x} &= \epsilon \\
y - \mathbf{w}^{\mathrm{T}}\mathbf{x} &\sim \mathcal{N}(\epsilon|0, \beta^{-1}I) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\epsilon-0)\beta(\epsilon-0)} \\
\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) &= \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})\beta(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})} \\
\Rightarrow \mathcal{N}(y - \mathbf{w}^{\mathrm{T}}\mathbf{x}|0, \beta^{-1}I) &= \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I) \\
\Rightarrow p(y|\mathbf{w}, \mathbf{x}) &= \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)
\end{aligned}
$$

Likelihood

- Likelihood

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}\left(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}\right)$$

- Independence

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(y_n|\mathbf{w}^{\mathrm{T}}\mathbf{x}_n, \beta^{-1}\right)$$

Assume each output to be independent given the input and the parameters

- Likelihood is Gaussian in $\mathbf{w}$

$$p(y|\mathbf{w},\mathbf{x}) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

- Likelihood is Gaussian in $\mathbf{w}$

$$p(y|\mathbf{w},\mathbf{x}) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

- Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Likelihood is Gaussian in $\mathbf{w}$

$$p(y|\mathbf{w},\mathbf{x}) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

- Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Posterior

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\right)$$

Linear Regression

- Likelihood is Gaussian in $\mathbf{w}$

$$p(y|\mathbf{w}, \mathbf{x}) = \mathcal{N}(y|\mathbf{w}^{\mathrm{T}}\mathbf{x}, \beta^{-1}I)$$

- Conjugate Prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$$

- Posterior

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N\right)$$

- $\mathbf{m}_N, \mathbf{S}_N$ is the mean and the co-variance of the posterior after having seen N data-points

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

- Identification

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{\underbrace{\int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}}_{p(\mathbf{y}|\mathbf{X})}} \propto p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

Posterior

- Posterior is Gaussian

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)$$

- Identification

$$p(\mathbf{w}|\mathbf{y},\mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{\underbrace{\int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}}_{p(\mathbf{y}|\mathbf{X})}} \propto p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

- Posterior

$$\mathbf{m}_N = \left(\mathbf{S}_0^{-1} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left(S_0^{-1}\mathbf{m}_0 + \beta\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)$$
$$\mathbf{S}_N = \left(\mathbf{S}_0^{-1} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

Posterior

- Assumption Zero mean isotropic Gaussian

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

Posterior

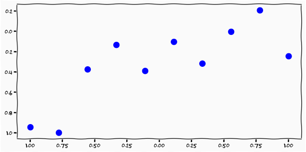
- Assumption Zero mean isotropic Gaussian

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I})$$

- Posterior

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}(\mathbf{w}|\beta \left(\alpha\mathbf{I} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}, \left(\alpha\mathbf{I} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1})$$

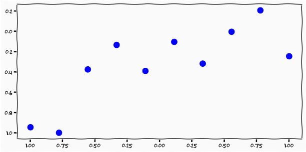
Linear Regression Example



- Model

$$f(x, \mathbf{w}) = w_0 + w_1 x$$

Linear Regression Example



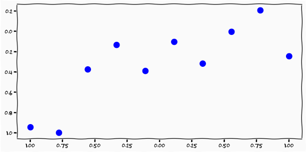
- Model

$$f(x, \mathbf{w}) = w_0 + w_1 x$$

- Data

$$f(x, \mathbf{a}) = a_0 + a_1 x, \ \{a_0, a_1\} = \{-0.3, 0.5\}$$
$$y = f(x, \mathbf{a}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, 0.2^2)$$

Linear Regression Example



- Model

$$f(x, \mathbf{w}) = w_0 + w_1 x$$

- Data

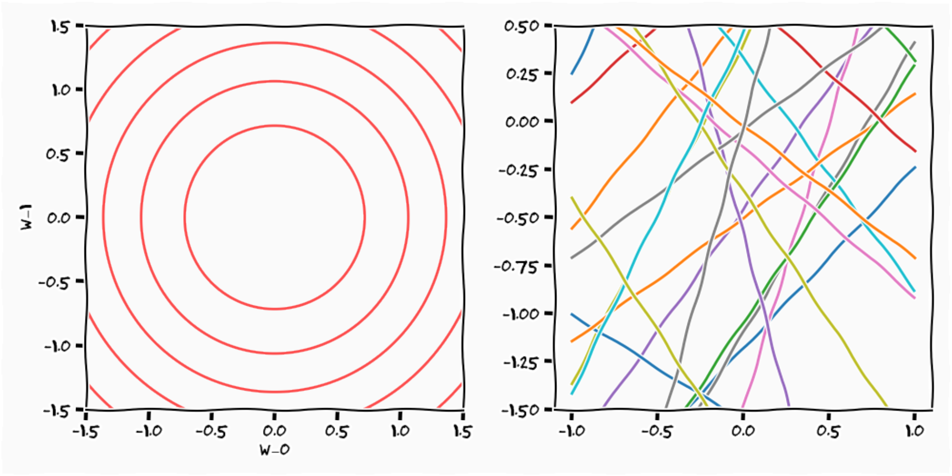
$$f(x, \mathbf{a}) = a_0 + a_1 x, \ \{a_0, a_1\} = \{-0.3, 0.5\}$$

$$y = f(x, \mathbf{a}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, 0.2^2)$$

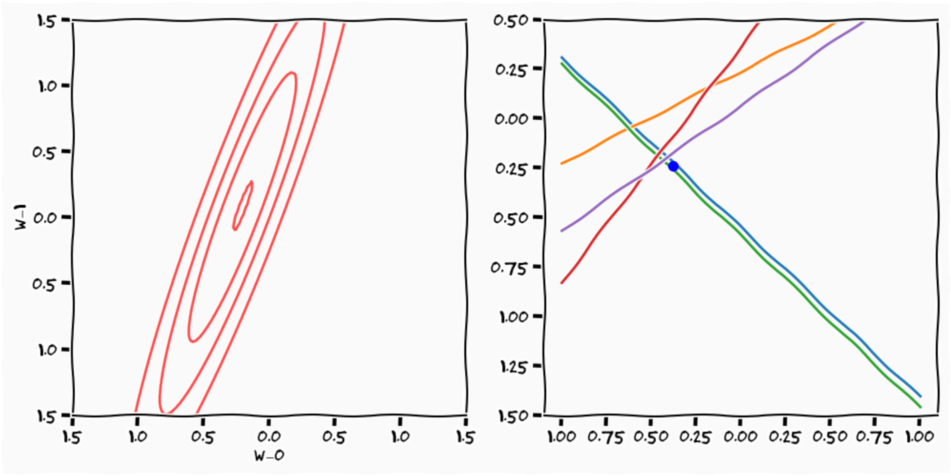
- Prior

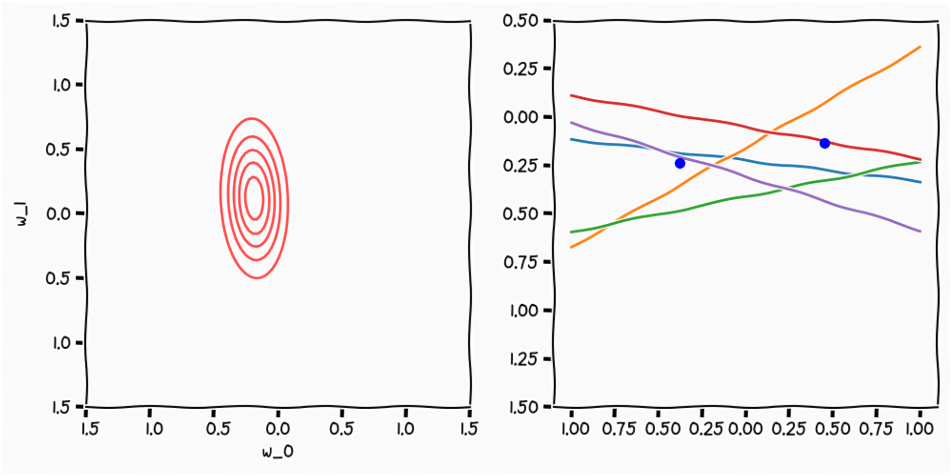
$$p(\mathbf{w}) = \mathcal{N}(\boldsymbol{w} | \mathbf{0}, 2.0 \cdot \mathbf{I})$$

Linear Regression Example

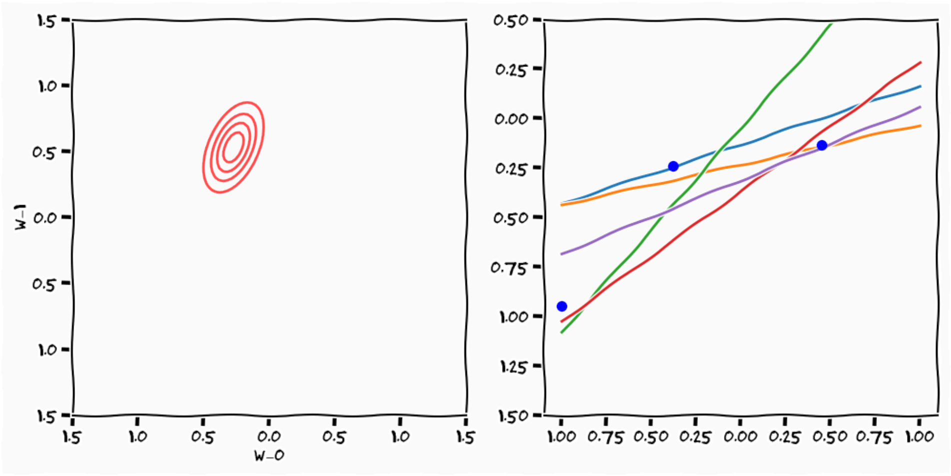


Linear Regression Example

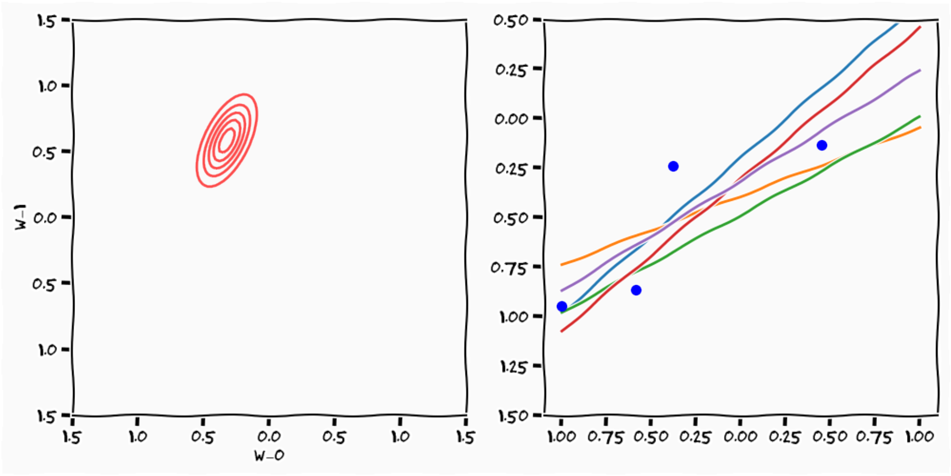



w_1
w_0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
0.50
0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.00 0.75 0.50 0.25 0.00 0.25 0.50 0.75 1.00

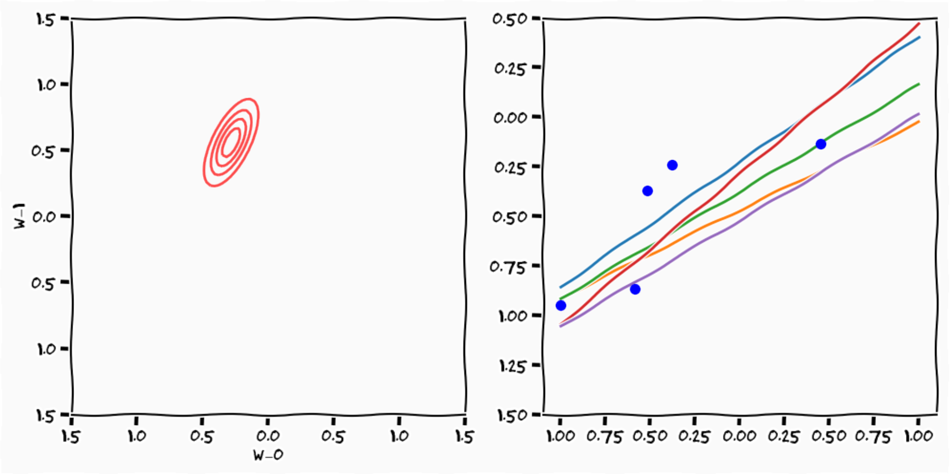
Linear Regression Example



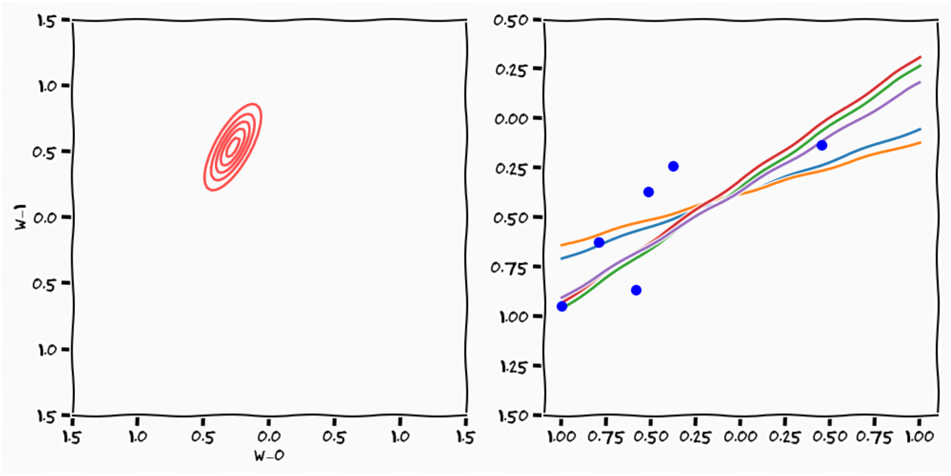
Linear Regression Example



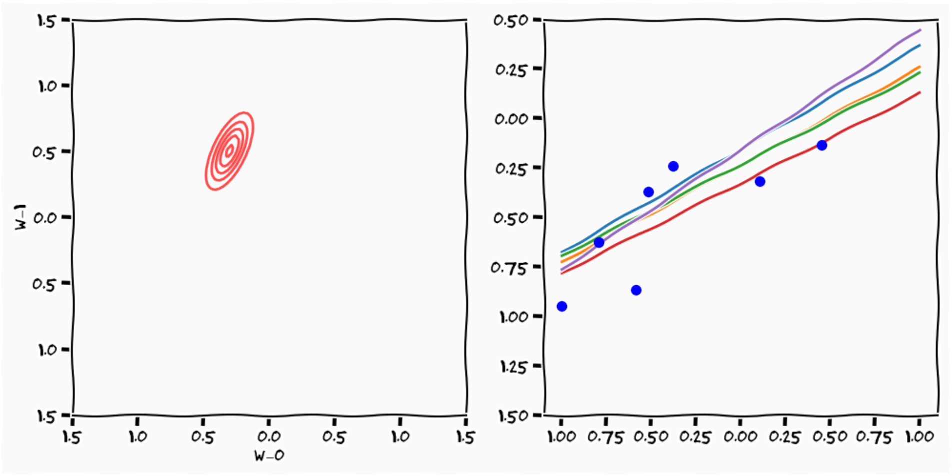
Linear Regression Example



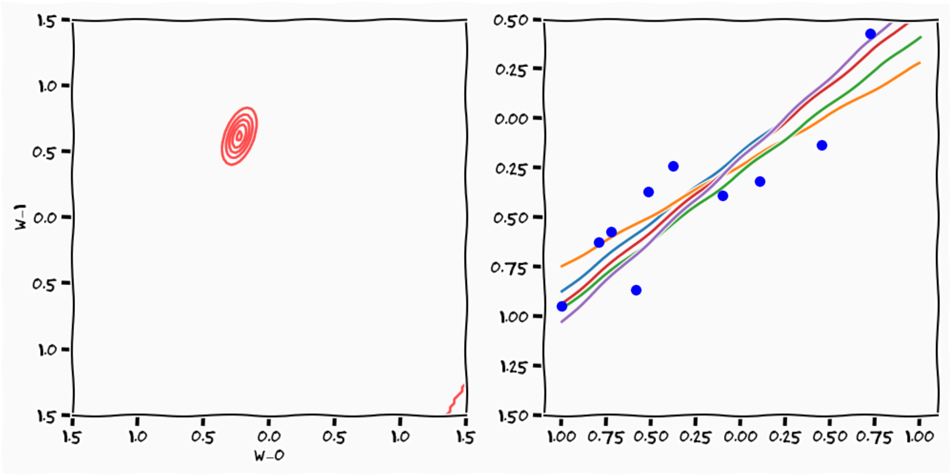
Linear Regression Example



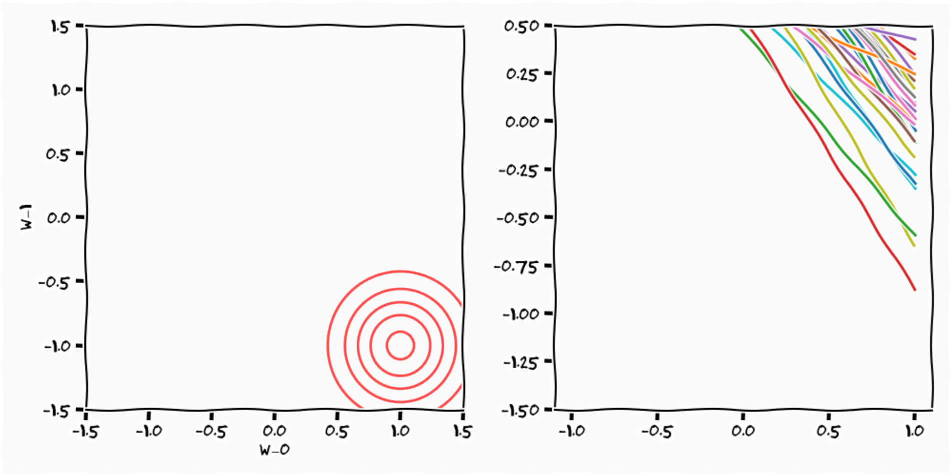
Linear Regression Example



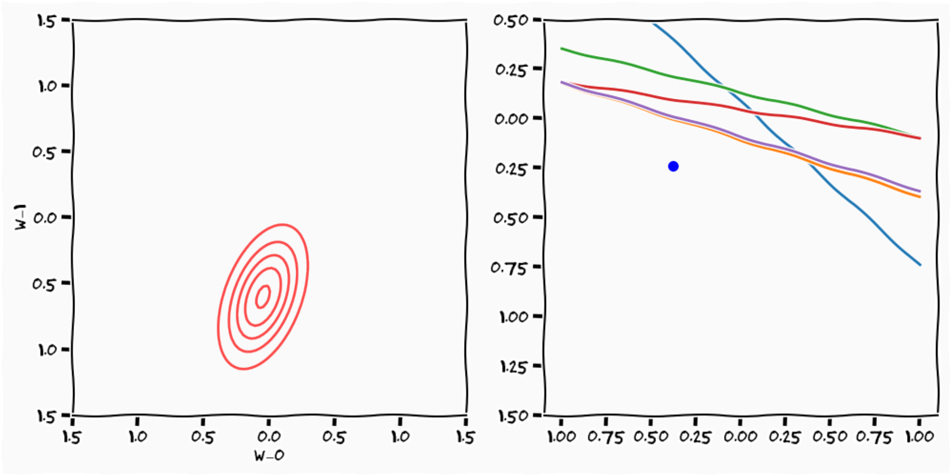
Linear Regression Example



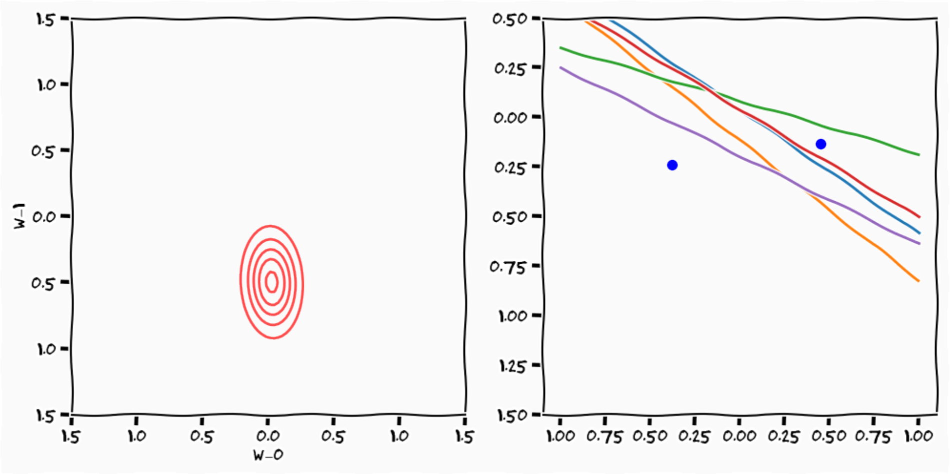
Linear Regression Example



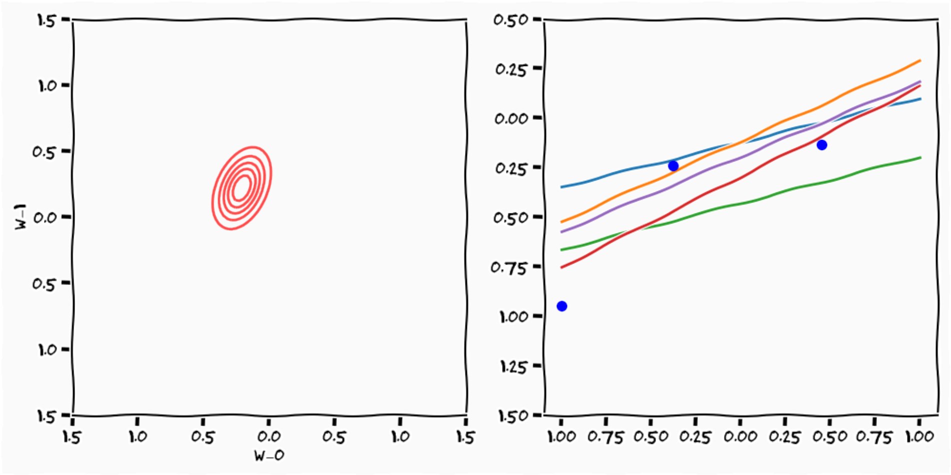
Linear Regression Example



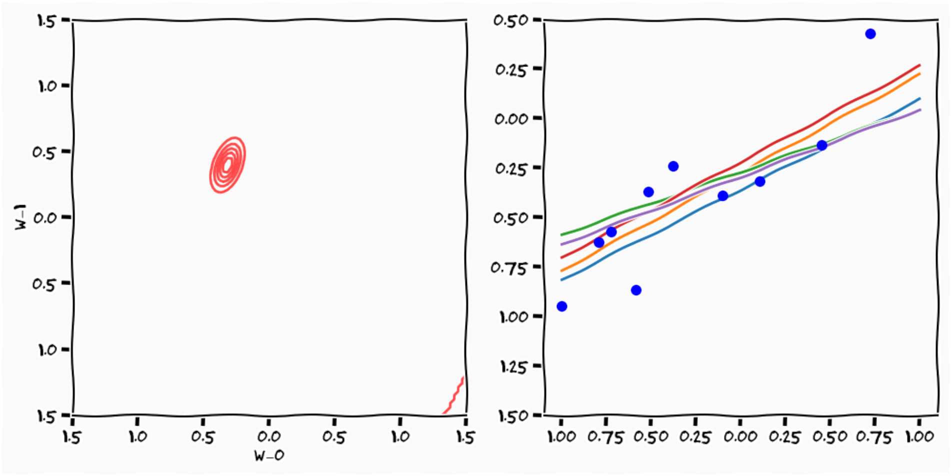
Linear Regression Example



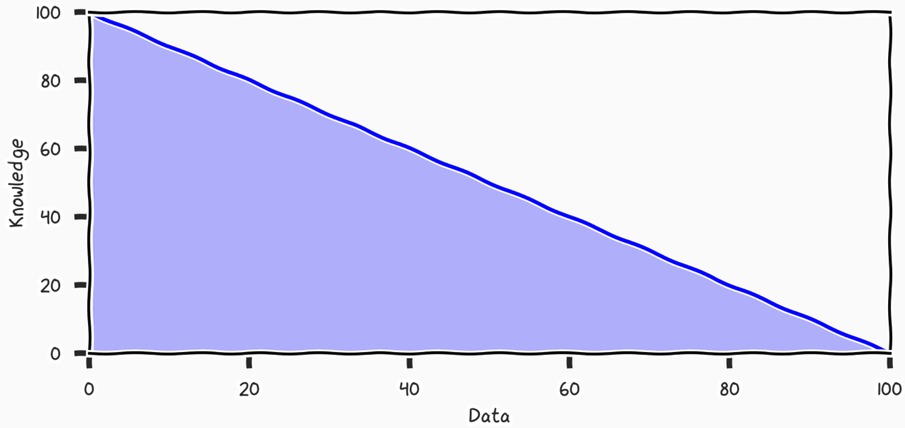
Linear Regression Example



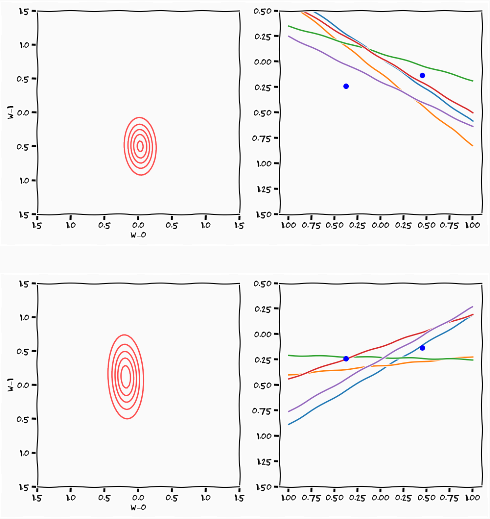
Linear Regression Example



Data and Beliefs



Knowledge is Relative



Statistics or Machine Learning

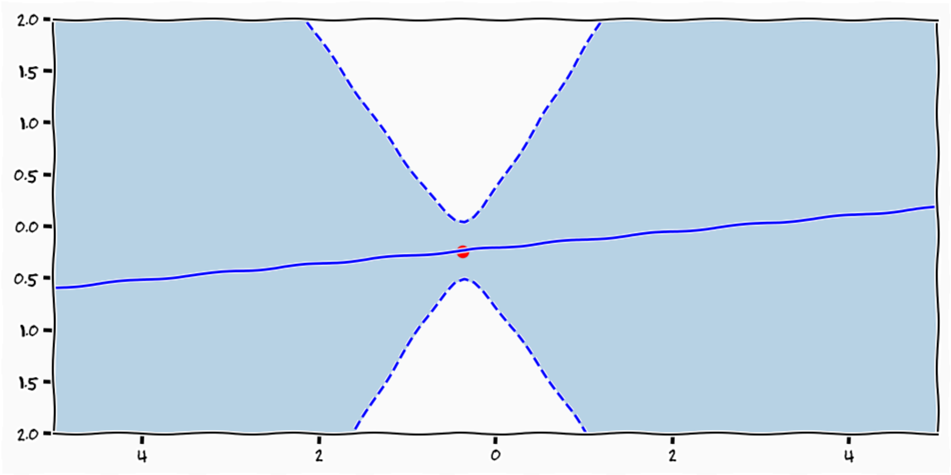
"The difference between statistics and machine learning is that the former cares about parameters while the latter cares about prediction"
– Prof. Neil D. Lawrence

Prediction

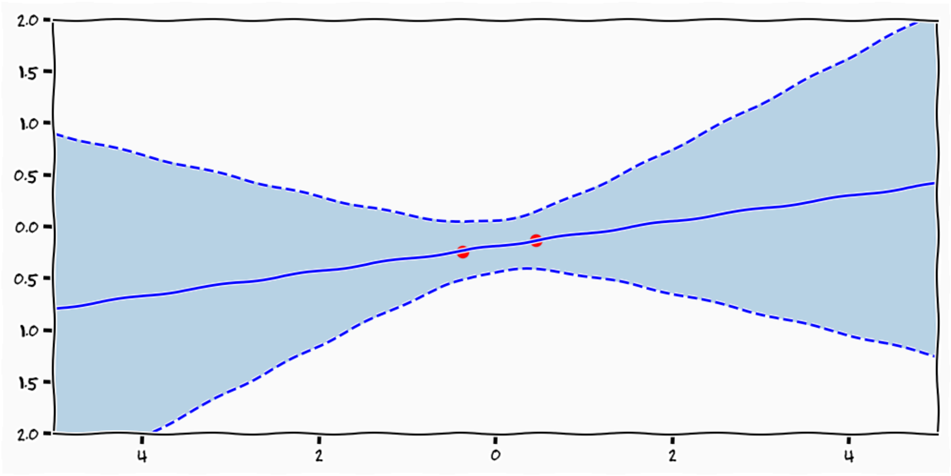
$$p(y_*|\mathbf{y}, \mathbf{x}_*, \mathbf{X}, \alpha, \beta) = \int p(y_*|\mathbf{x}_*, \mathbf{w}, \beta)p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \alpha, \beta)\mathrm{d}\mathbf{w}$$

- we do not really care about the value of $\mathbf{w}$ we care about new prediction y_* at location $\mathbf{x}_*$
- look at the marginal distribution, i.e. when we average out the weight

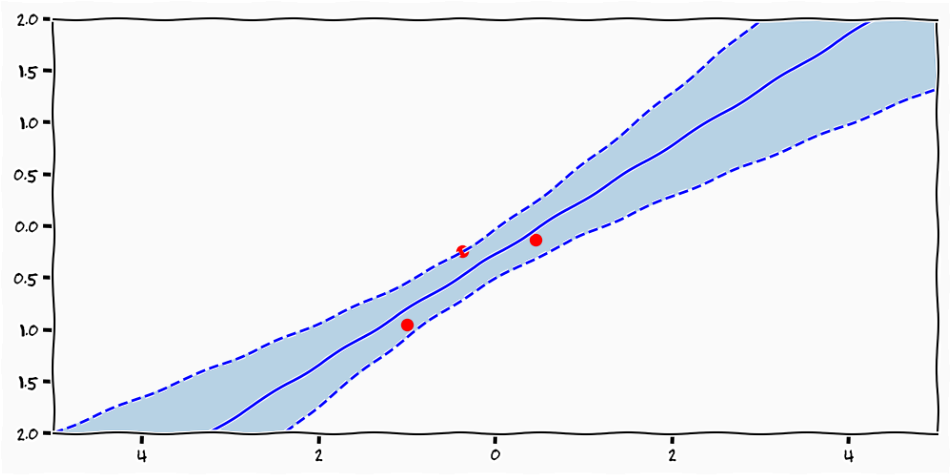
Predictive Posterior

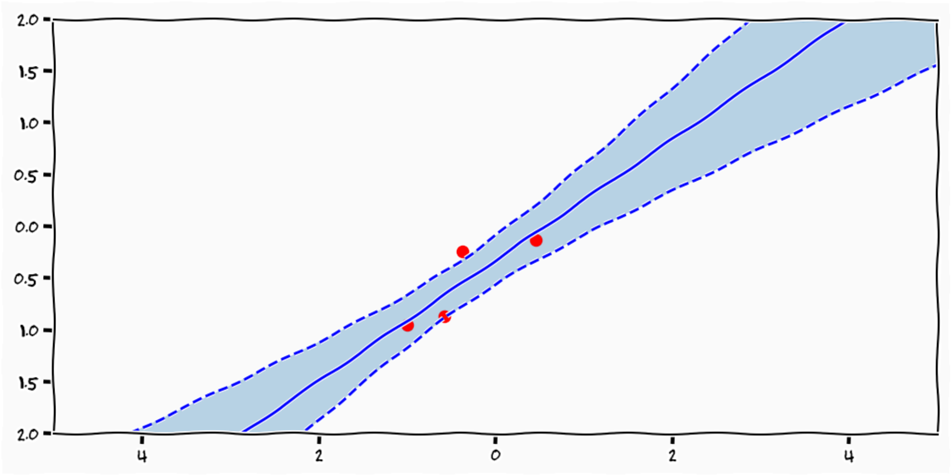


Predictive Posterior

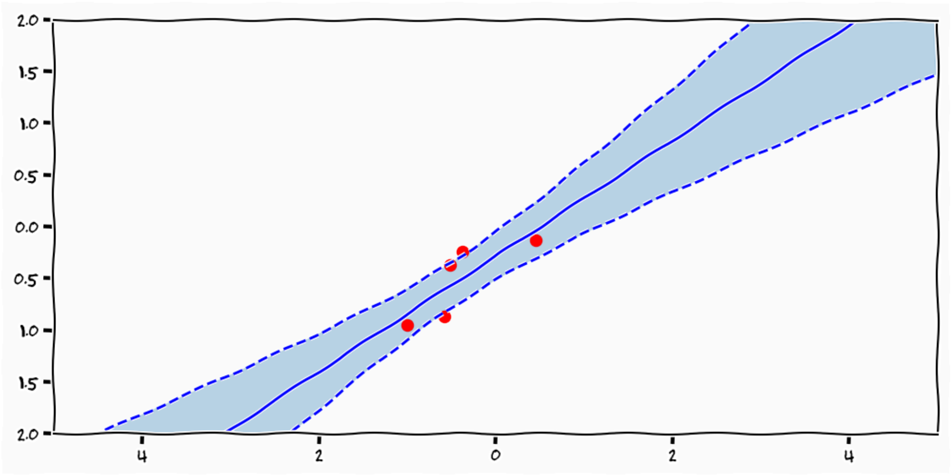


Predictive Posterior

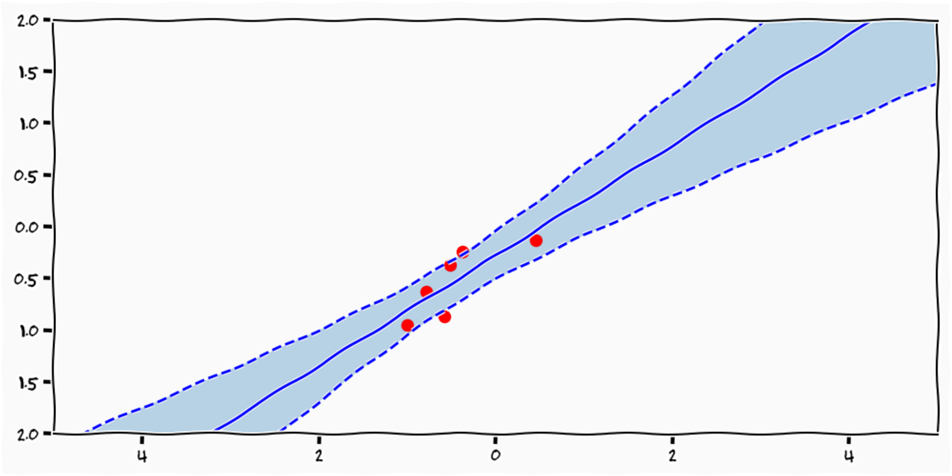



2.0
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
4
2
0
2
4

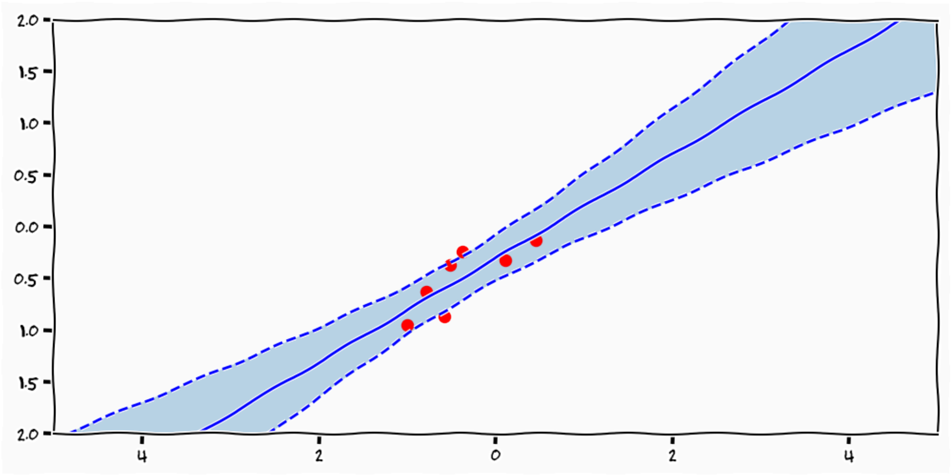
Predictive Posterior



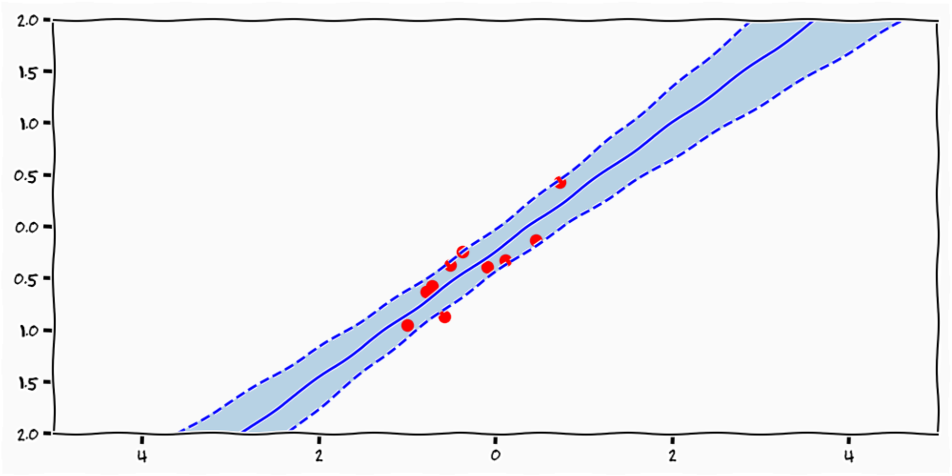
Predictive Posterior



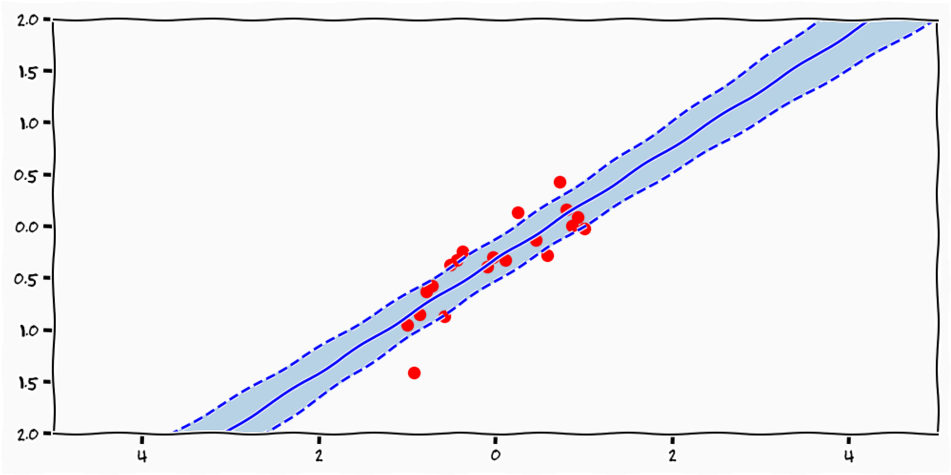
Predictive Posterior



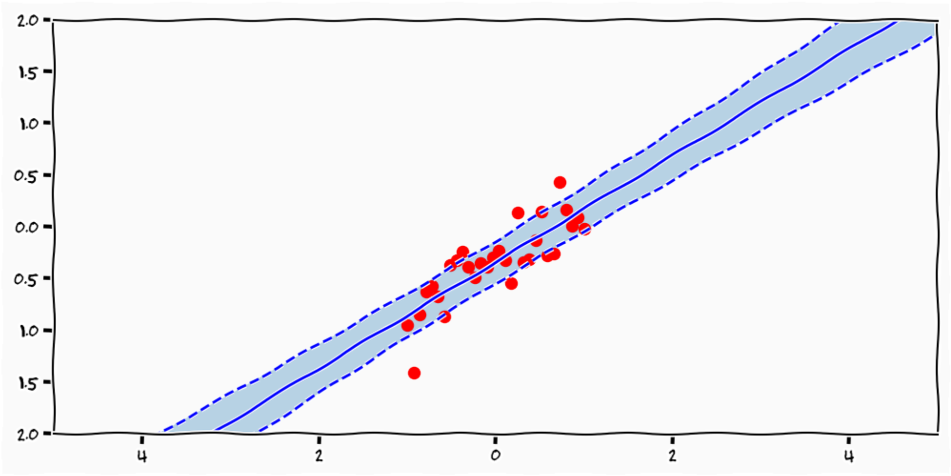
Predictive Posterior



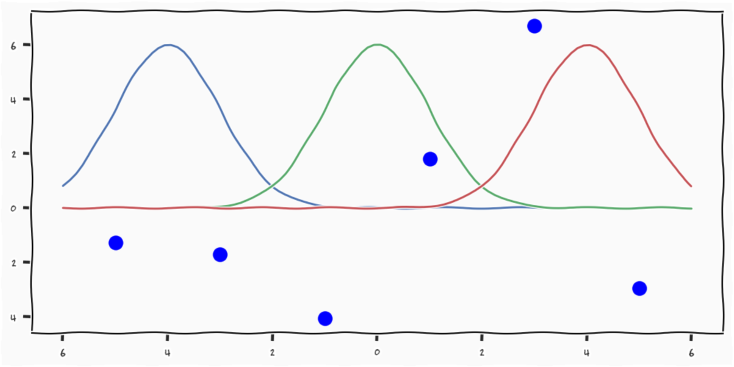
Predictive Posterior



Predictive Posterior



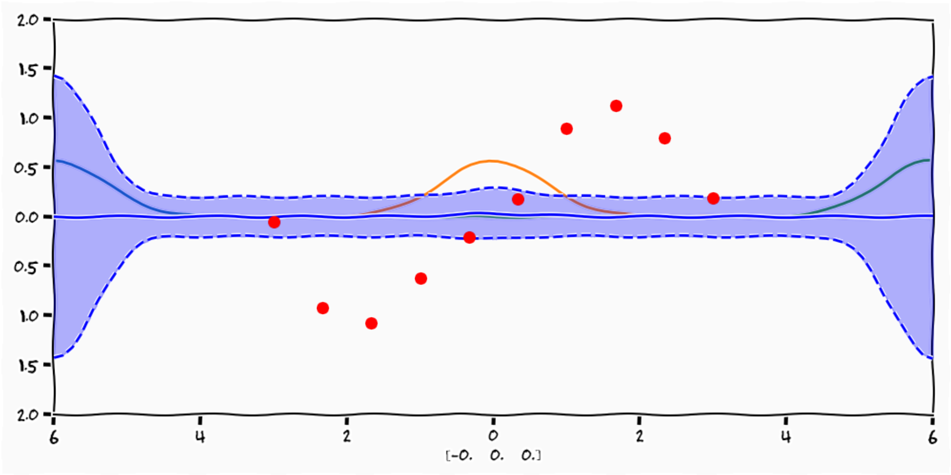
Linear Regression



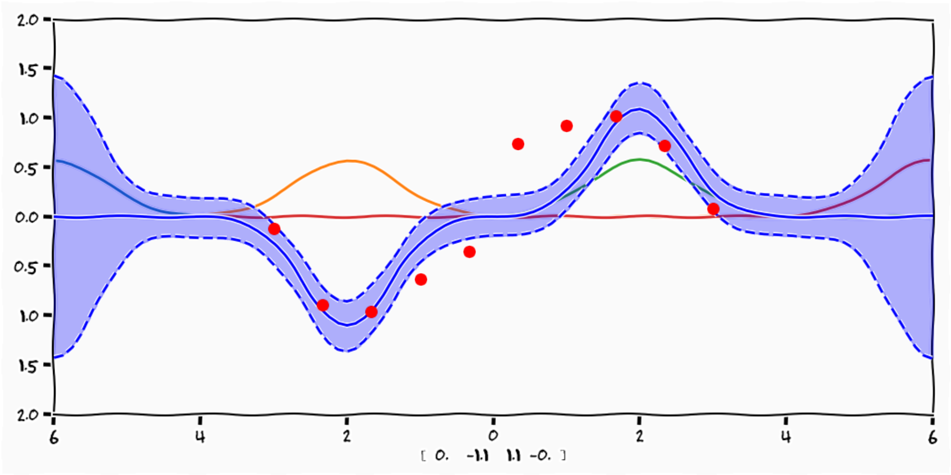
- Linear function only in parameters

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

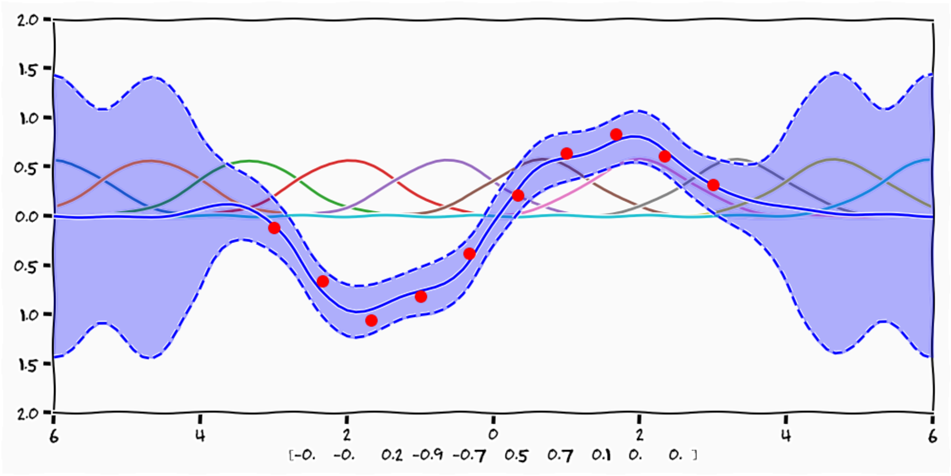
Non-Linear Basis Functions



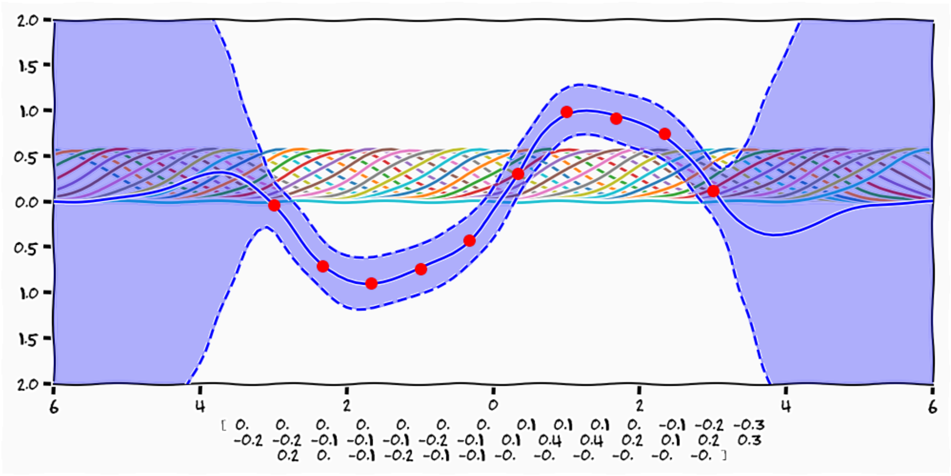
Non-Linear Basis Functions



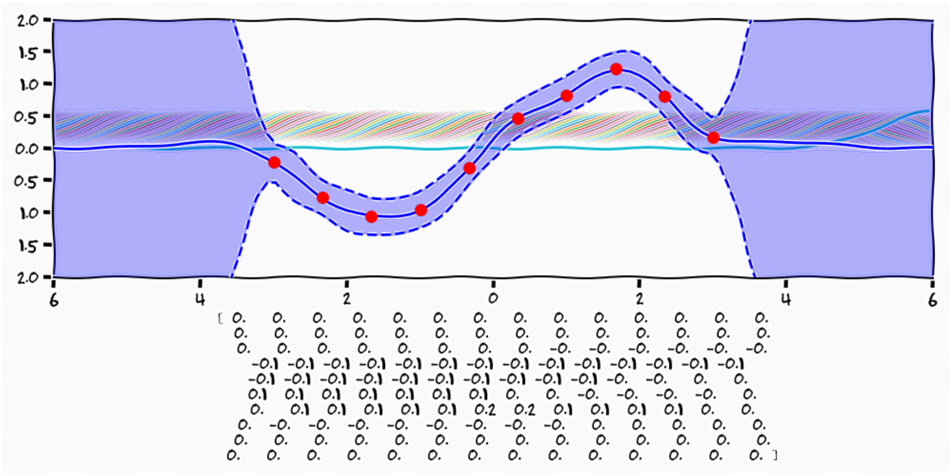
Non-Linear Basis Functions



Non-Linear Basis Functions



Non-Linear Basis Functions



Summary

Summary

- *That was a lot of philosphical nonsense to do something I did in school when I was 12*

[2] we really hope so :-)

Summary

- *That was a lot of philosphical nonsense to do something I did in school when I was 12*
- The important thing was not "least squares" but how we reasoned to get to the result

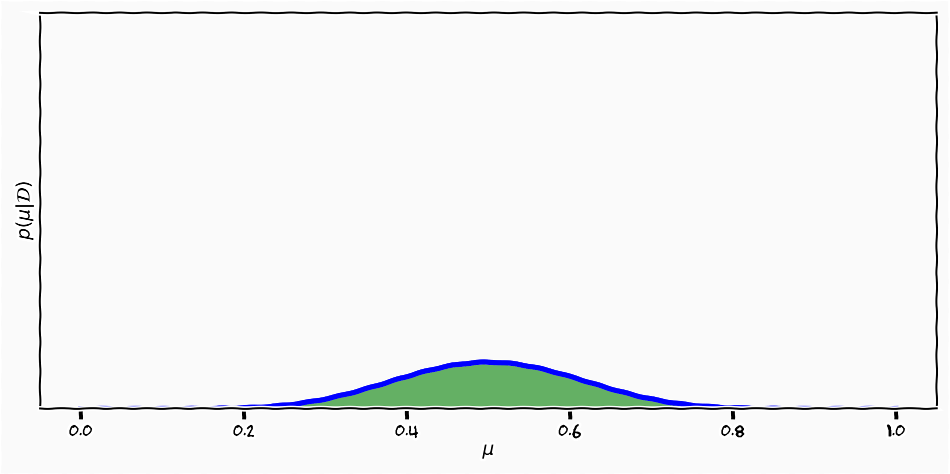
[2] we really hope so :-)

Summary

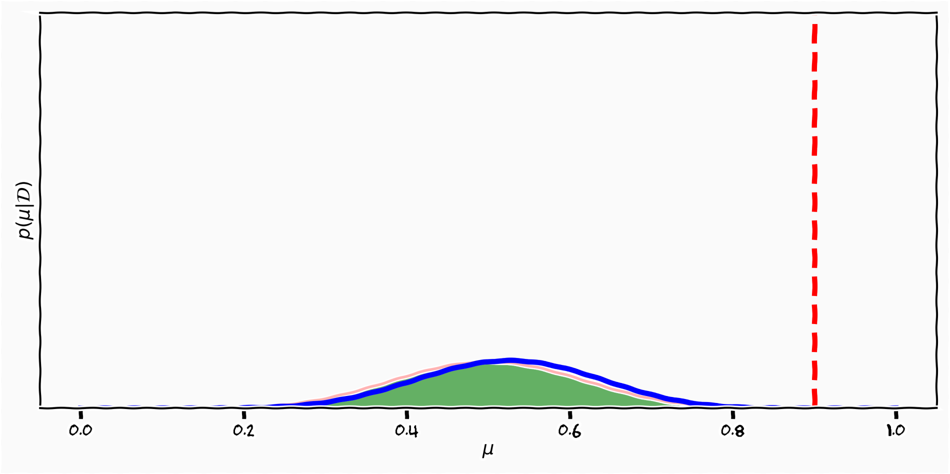
- *That was a lot of philosphical nonsense to do something I did in school when I was 12*
- The important thing was not "least squares" but how we reasoned to get to the result
- This reasoning will stay consistent through the course[2]

[2] we really hope so :-)

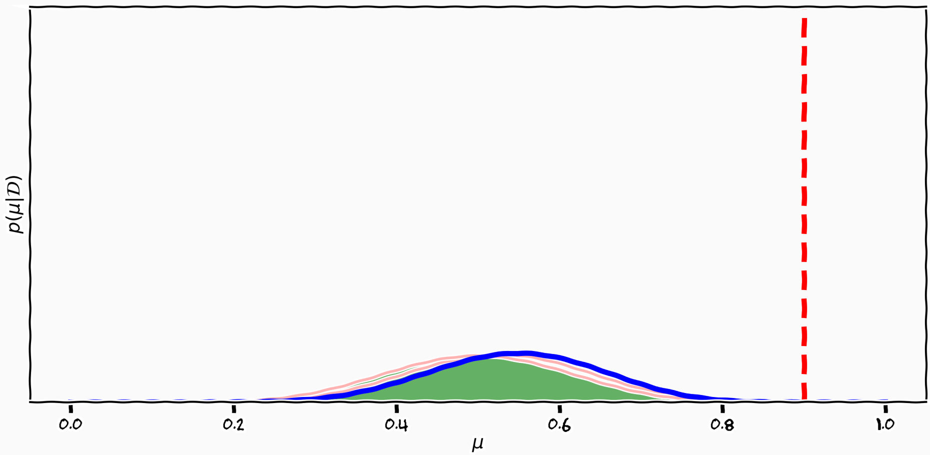
Bernoulli Trial



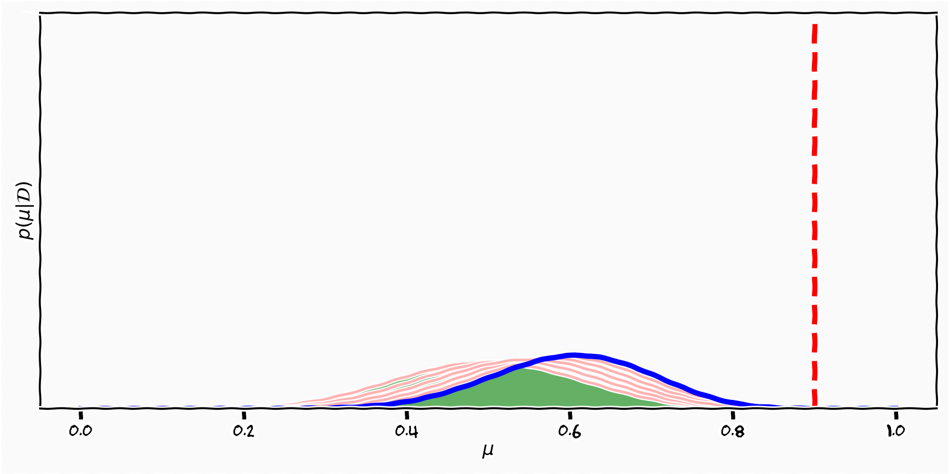
Bernoulli Trial



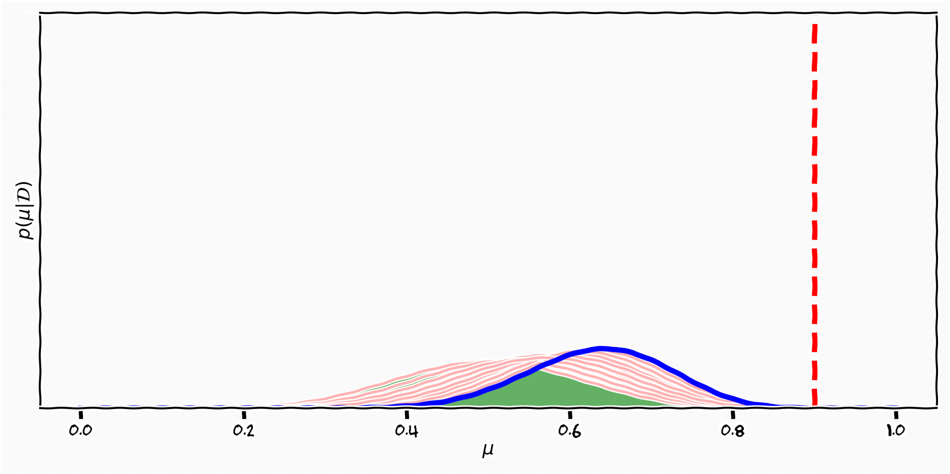
Bernoulli Trial



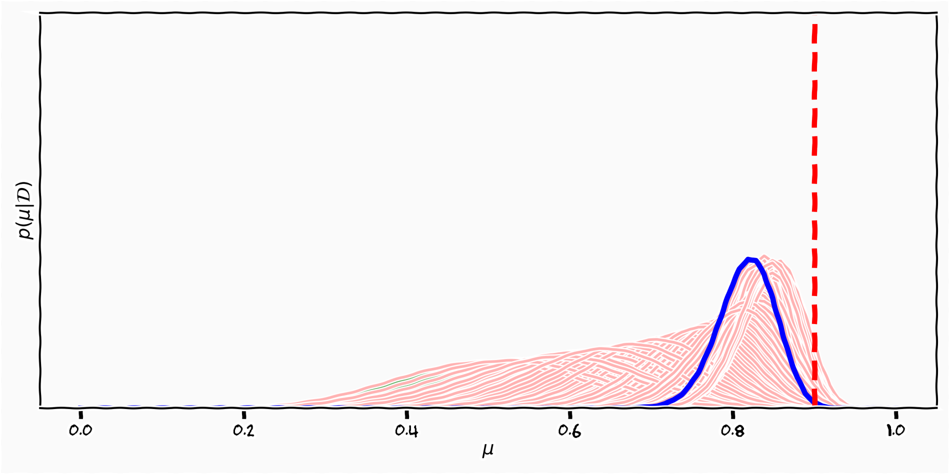
Bernoulli Trial



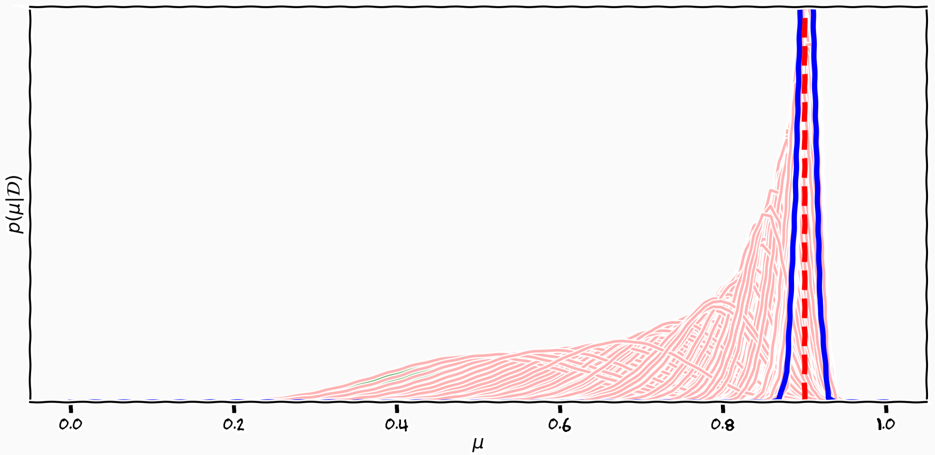
Bernoulli Trial



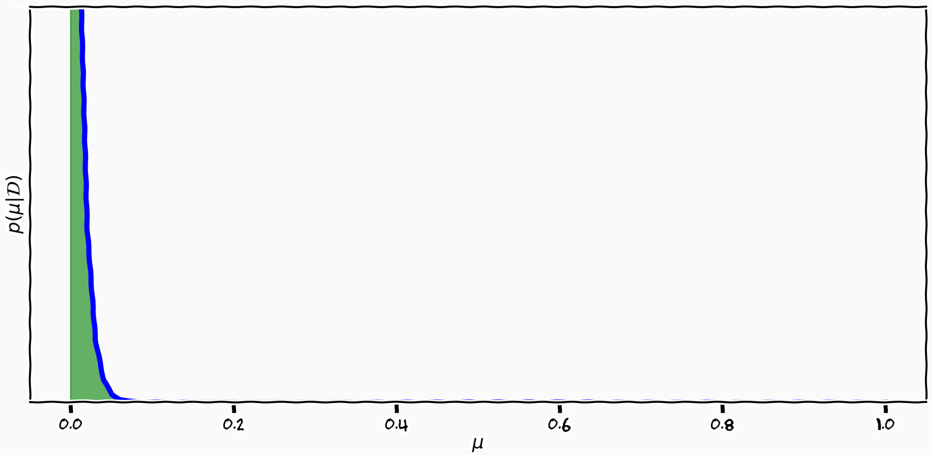
Bernoulli Trial



Bernoulli Trial



Bernoulli Trial

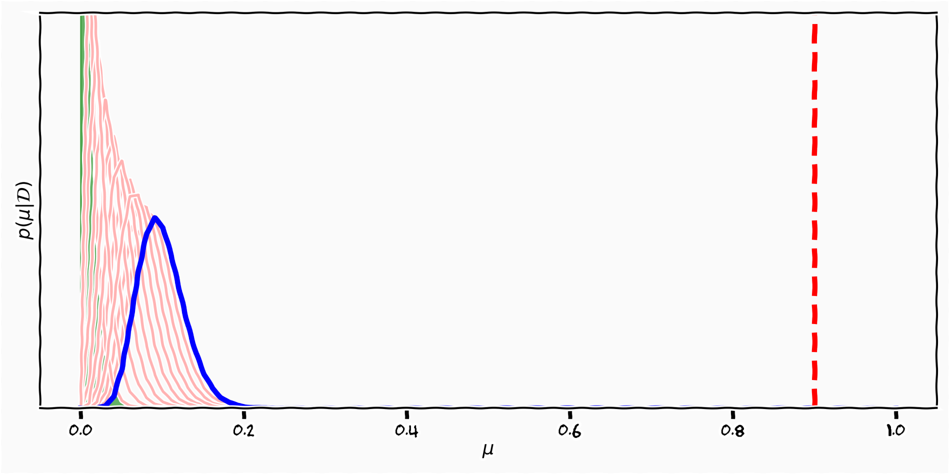


Bernoulli Trial

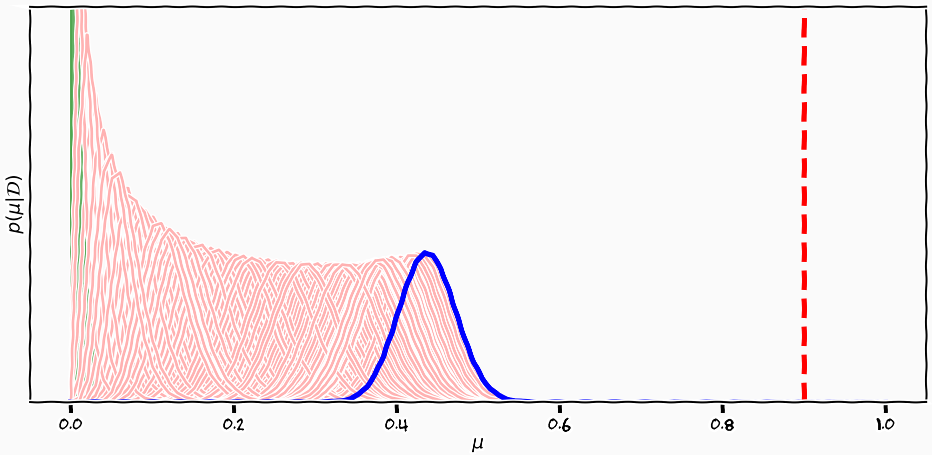
Bernoulli Trial

Bernoulli Trial

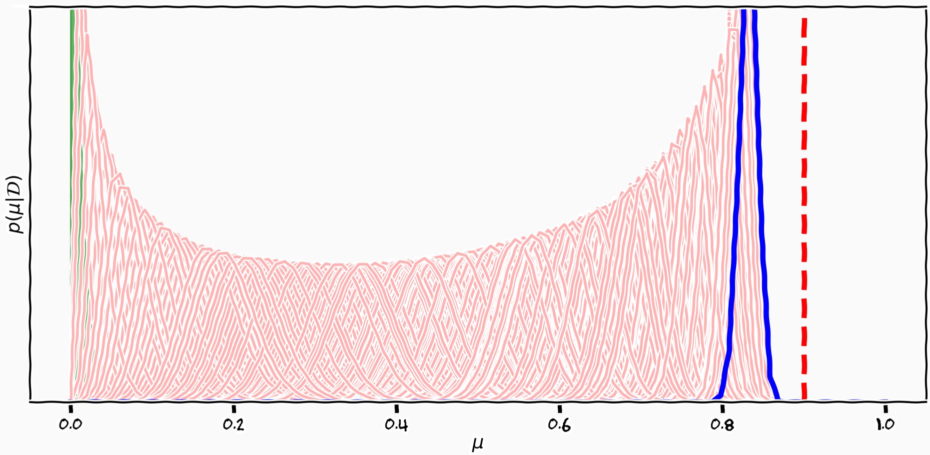
Bernoulli Trial



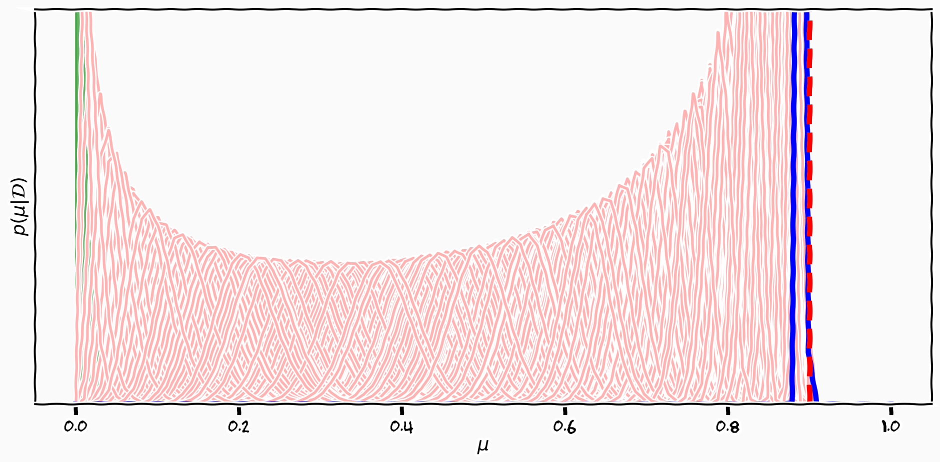
Bernoulli Trial

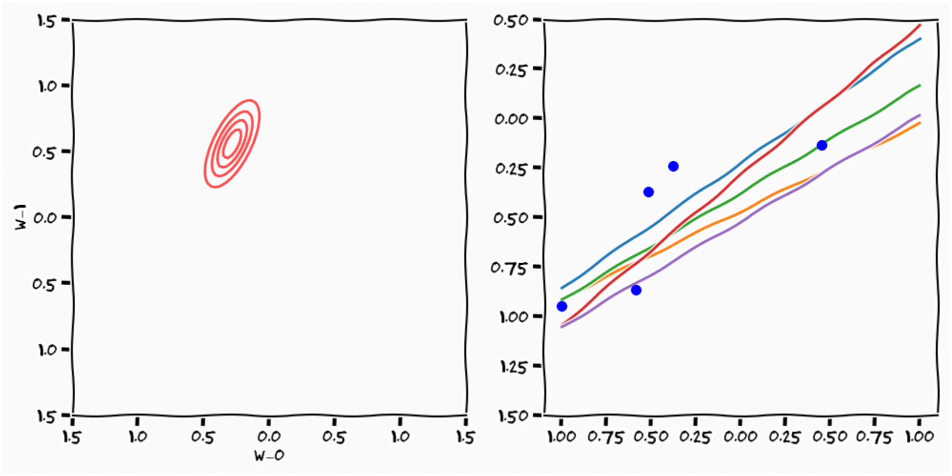


Bernoulli Trial



Bernoulli Trial




1.5
1.0
0.5
0.0
0.5
1.0
1.5
w_1
1.5
1.0
0.5
0.0
0.5
1.0
1.5
w_0
0.50
0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.00
0.75
0.50
0.25
0.00
0.25
0.50
0.75
1.00

Gaussian Identities

$$p(x_1, x_2) \quad p(x1) \quad p(x1 \mid x_2)$$

eof

References

References

- Chomsky, Noam A and Jerry A Fodor (1980). "The inductivist fallacy." In: *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*.
- Laplace, Pierre Simon (1814). *A philosophical essay on probabilities*.

Does this make sense?

Posterior Variance

$$\mathbf{S}_N = \left(\mathbf{I}\alpha + \beta \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

Posterior Mean

$$\mathbf{m}_N = \left(\frac{1}{\alpha}\mathbf{I} + \beta \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \beta \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Posterior Variance

$$
\begin{aligned}
\mathbf{S}_N &= \left(\mathbf{I}\alpha + \beta \mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{I}\alpha + \beta \begin{bmatrix} \sum_i^N 1 & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}\right)^{-1} = \begin{bmatrix} \beta N + \alpha & \beta \sum_i x_i \\ \beta \sum_i x_i & \alpha + \beta \sum_i x_i^2 \end{bmatrix}^{-1} \\
&= \frac{1}{(\beta N + \alpha)(\alpha + \beta \sum_i x_i^2) - (\beta \sum_i x_i)^2} \begin{bmatrix} \alpha + \beta \sum_i x_i^2 & -\beta \sum_i x_i \\ -\beta \sum_i x_i & \beta N + \alpha \end{bmatrix}
\end{aligned}
$$

Posterior Variance

$$\mathbf{S}_N = \frac{1}{(\beta N + \alpha)(\alpha + \beta \sum_i x_i^2) - (\beta \sum_i x_i)^2} \begin{bmatrix} \alpha + \beta \sum_i x_i^2 & -\beta \sum_i x_i \\ -\beta \sum_i x_i & \beta N + \alpha \end{bmatrix}$$

Posterior Variance

$$\mathbf{S}_N = \frac{1}{(\beta N + \alpha)(\alpha + \beta \sum_i x_i^2) - (\beta \sum_i x_i)^2} \begin{bmatrix} \alpha + \beta \sum_i x_i^2 & -\beta \sum_i x_i \\ -\beta \sum_i x_i & \beta N + \alpha \end{bmatrix}$$

- Lets assume input is centered $\Rightarrow \sum_i x_i = 0$

$$\mathbf{S}_N = \frac{1}{(\beta N + \alpha)(\alpha + \beta \sum_i x_i^2)} \begin{bmatrix} \alpha + \beta \sum_i x_i^2 & 0 \\ 0 & \beta N + \alpha \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{\beta N + \alpha} & 0 \\ 0 & \frac{1}{\alpha + \beta \sum_i x_i^2} \end{bmatrix}$$

Posterior Mean

$$
\begin{aligned}
\mathbf{m}_N &= \left(\alpha\mathbf{I} + \beta\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\beta\mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&= \beta\mathbf{S}_N \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \\
&= \beta\mathbf{S}_N \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{bmatrix}
\end{aligned}
$$

Posterior Mean

$$\mathbf{m}_N = \beta \mathbf{S}_N \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{bmatrix}$$

- Lets assume input is centered $\Rightarrow \sum_i x_i = 0$

$$\begin{aligned} \mathbf{m}_N &= \beta \begin{bmatrix} \frac{1}{\beta N + \alpha} & 0 \\ 0 & \frac{1}{\alpha + \beta \sum_i x_i^2} \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{bmatrix} \\ &= \begin{bmatrix} \frac{\beta \sum_i y_i}{\beta N + \alpha} \\ \frac{\beta \sum_i y_i x_i}{\alpha + \beta \sum_i x_i^2} \end{bmatrix} \end{aligned}$$

Posterior Mean Slope

$$\tilde{w}_0 = \frac{\beta \sum_i y_i}{\beta N + \alpha}$$

$$p(w_0) = \mathcal{N}(w_0 | 0, \frac{1}{\alpha})$$

$$p(\epsilon) = \mathcal{N}(\epsilon | 0, \frac{1}{\beta})$$

Which Parametrisation

- Should I use a line, polynomial, quadratic basis function?
- How many basis functions should I use?
- Likelihood won't help me
- How do we proceed?

Regression Models

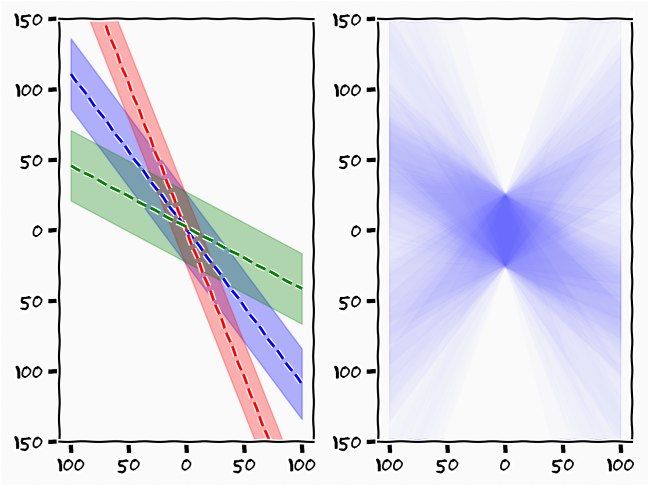
Linear Linear Model

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

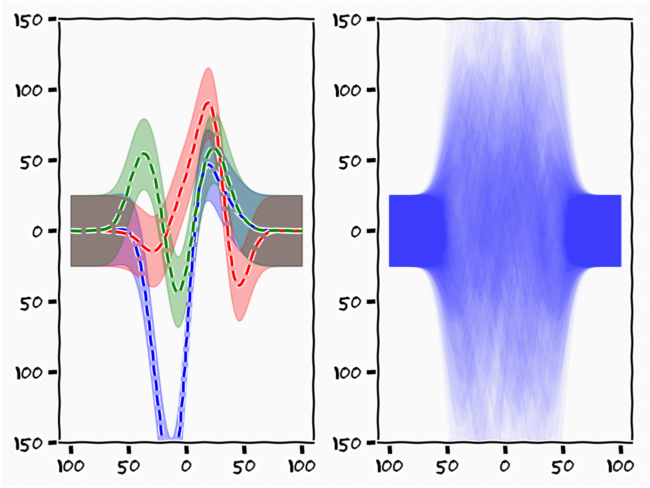
Basis function

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

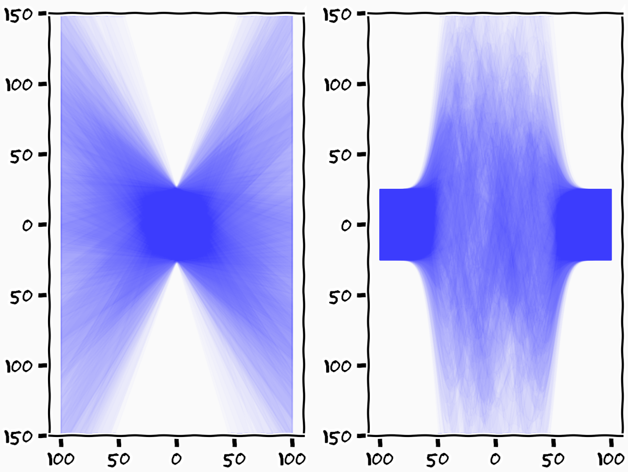
Model 1



Model 2

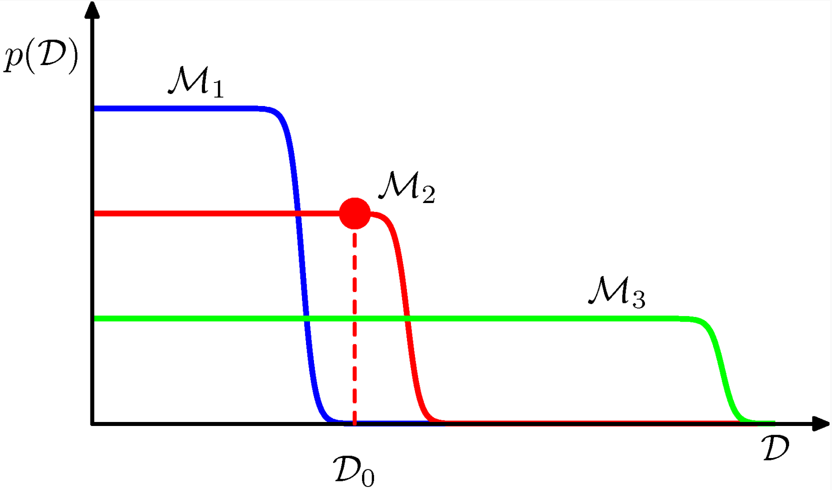

150
100
50
0
50
100
150
100
50
0
50
100
150
100
50
0
50
100
150
100
50
0
50
100

Evidence



$\int$

Model Selection[3]



[3] David MacKay PhD Thesis

Occams Razor

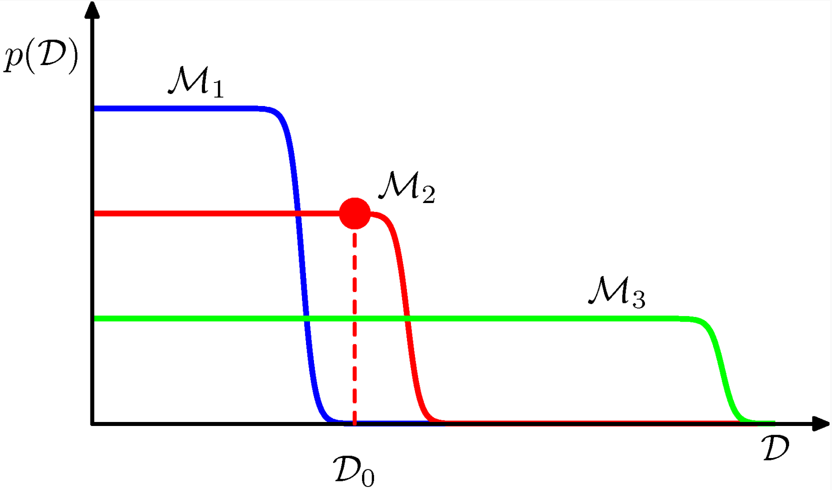
Definition (Occams Razor)
"All things being equal, the simplest solution tends to be the best one"

– William of Ockham

What is Simple?[4]

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

Model Selection[3]



[3] David MacKay PhD Thesis