Lecture 11

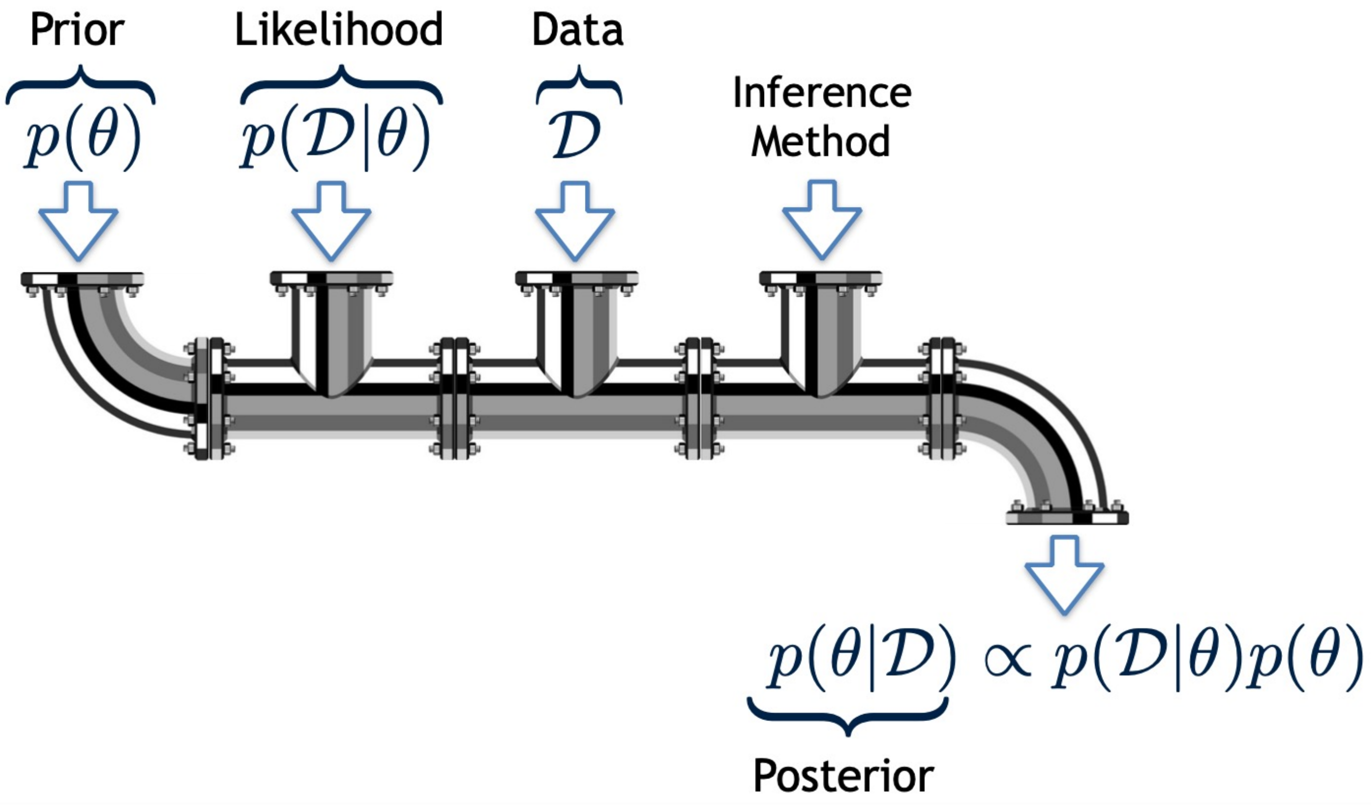
Bayesian Modeling (Part 1)

(Based on slides by **Dr. Tom Rainforth**, HT 2020)

Jiarui Gan

jiarui.gan@cs.ox.ac.uk

Last Lecture: the Bayesian Pipeline



Last Lecture: Bayes’ Rule

Likelihood

Prior

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$$

Posterior

Evidence

*Image Credit: Paul Epps

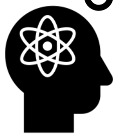
Last Lecture: Coin Flipping Example

$p(H|biased) = 0.2$
$p(T|biased) = 0.8$

biased

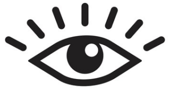
$p(H|fair) = 0.5$
$p(T|fair) = 0.5$

fair



prior belief

$p(biased) = 0.7$
$p(fair) = 0.3$



H

$$p(biased|H) = \frac{0.2\times0.7}{0.2\times0.7+0.5\times0.3} = 0.48$$

$$p(fair|H) = \frac{0.5\times0.3}{0.2\times0.7+0.5\times0.3} = 0.52$$

$$p(H|H) = p(H|biased)\cdot p(biased|H) + p(H|fair)\cdot p(fair|H) = 0.36$$

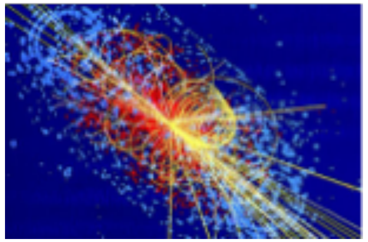
Outline of This Lecture

- What is a Bayesian model?
- Bayesian modeling through the eyes of multiple hypotheses
- Example: Bayesian linear regression

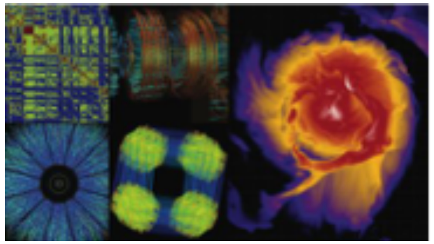
What is a Bayesian Model?

What is a Model?

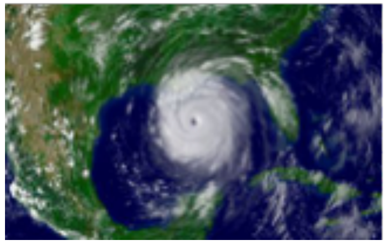
- Models are mechanisms for **reasoning** about the world
- E.g. Newtonian mechanics, simulators, internal models our brain constructs
- Good models balance **fidelity**, **predictive power** and **tractability**
  - E.g. Quantum mechanics is a more accurate model than Newtonian mechanics, but it is actually less useful for everyday tasks



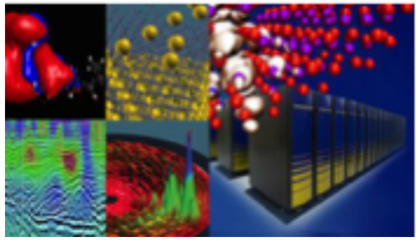
Particle physics



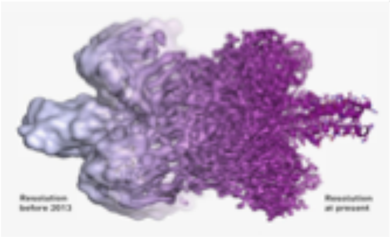
Nuclear physics



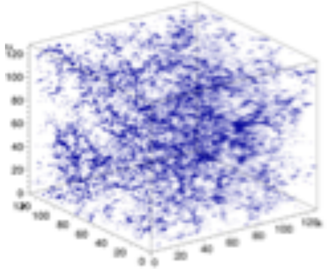
Weather



Material design



Drug discovery



Cosmology

Example Model: Poler Players' Reasoning about Each Other

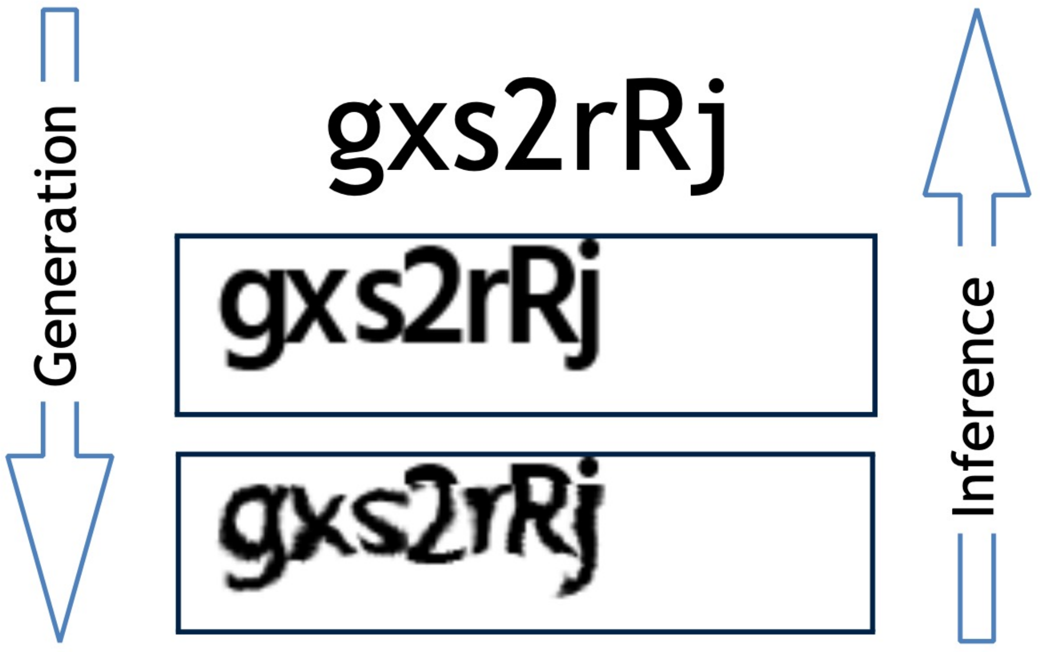
What is a Bayesian Model?

A **probabilistic generative model** $p(\theta, \mathcal{D})$ over **latents** θ and **data**$\mathcal{D}$

- It forms a probabilistic “simulator” for generating data that we might have seen
- Almost any stochastic simulator can be used as a Bayesian model (we will return to this idea in more detail when we cover probabilistic programming)

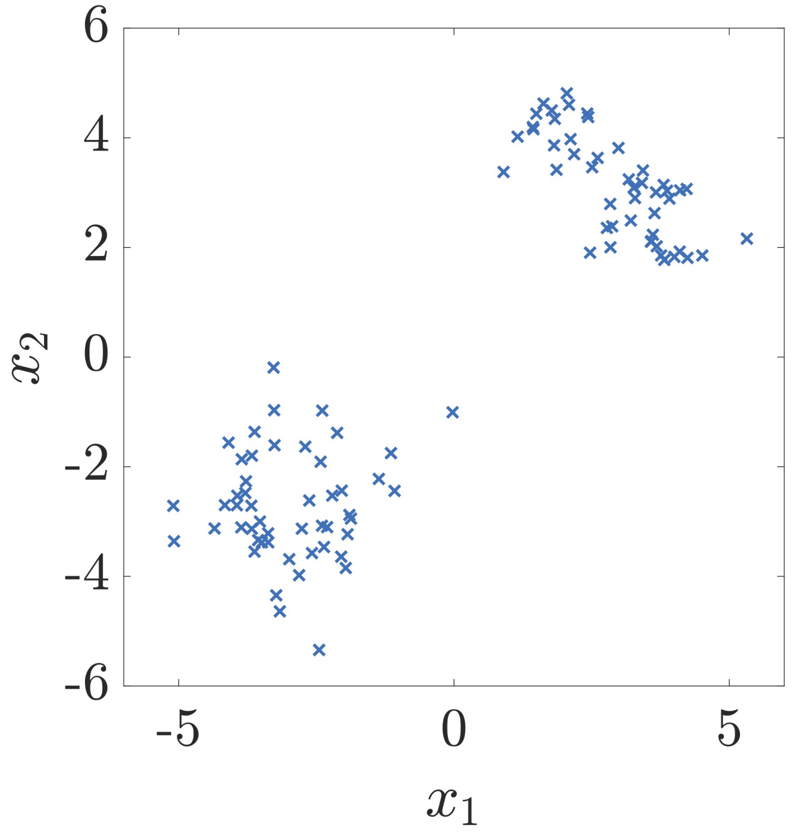
Example Bayesian Model: Captcha Simulator

$p(\text{letters}|\text{image})$



$p(\text{image}|\text{letters})$

Example Bayesian Model: Gaussian Mixture Model



Example Bayesian Model: Gaussian Mixture Model

Gaussian 1:

$$\mu_1 = [-3, -3], \Sigma_1 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}$$

Gaussian 2:

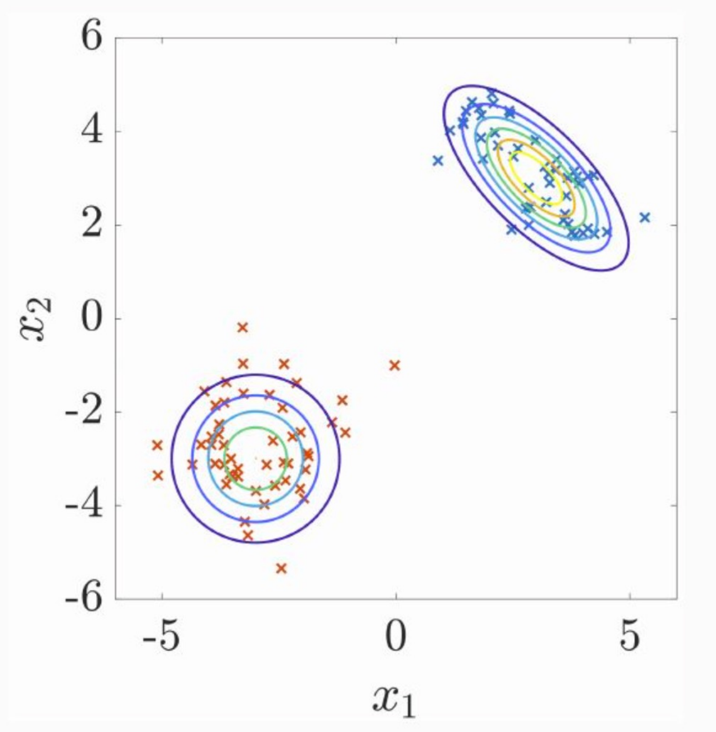
$$\mu_2 = [3,3], \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Generative mode:

$$\theta \sim \text{Categorical}([0.5, 0.5])$$

$$x \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$$

$$p(\mathcal{D}|\theta) = \prod_{n-1}^{N} p(x_n|\theta)$$



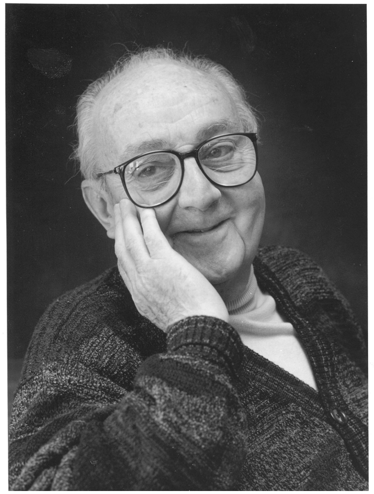
A Fundamental Assumption

- An assumption made by virtually all Bayesian models is that data points are **conditionally independent** given the parameter values.
- In other words, if our data is given by $\mathcal{D} = \{x_n\}_{n=1}^{N}$, we assume that the likelihood factorizes as:

$$p(\mathcal{D}|\theta) = \prod_{n-1}^{N} p(x_n|\theta)$$

- Effectively equates to assuming that our model captures all information relevant to prediction
- For more details, see the lecture notes

"All models are wrong,
but some are useful"



George Box
(1919—2013)

"All models are wrong, but some are useful"

- The purpose of a model is to help provide insights into a target problem or data and sometimes to further use these insights to make predictions
- Its purpose is **not** to try and fully encapsulate the "true" generative process or perfectly describe the data
- There are infinite different ways to generate any given dataset. Trying to uncover the "true" generative process is not even a well-defined problem
- In any real–world scenario, no Bayesian model can be "correct". The posterior is inherently subjective
- It is still important to criticize—models can be very wrong! E.g. we can use frequentist methods to falsify the likelihood

Bayesian Modeling Through the Eyes of Multiple Hypotheses

Bayesian Modeling as Multiple Hypotheses

Bayesian models are rooted in hypotheses:

- Each instance of our parameters θ is a hypothesis. Given a θ, we can **simulate** data using the likelihood model $p(D|\theta)$
- Bayesian inference allows us to reason about these hypothesis, giving the probability that each is true given the actual data we observe
- The **posterior predictive** is a weighted sum of the predictions from all possible hypotheses, where these weights are how likely that hypothesis is to be true

Recap: Coin Flipping

Hypotheses

$p(H|biased) = 0.2$
$p(T|biased) = 0.8$

biased

$p(H|fair) = 0.5$
$p(T|fair) = 0.5$

fair

prior belief

$p(biased) = 0.7$
$p(fair) = 0.3$

H

Posterior

$$p(biased|H) = \frac{0.2\times0.7}{0.2\times0.7+0.5\times0.3} = 0.48$$

$$p(fair|H) = \frac{0.5\times0.3}{0.2\times0.7+0.5\times0.3} = 0.52$$

Posterior predictive

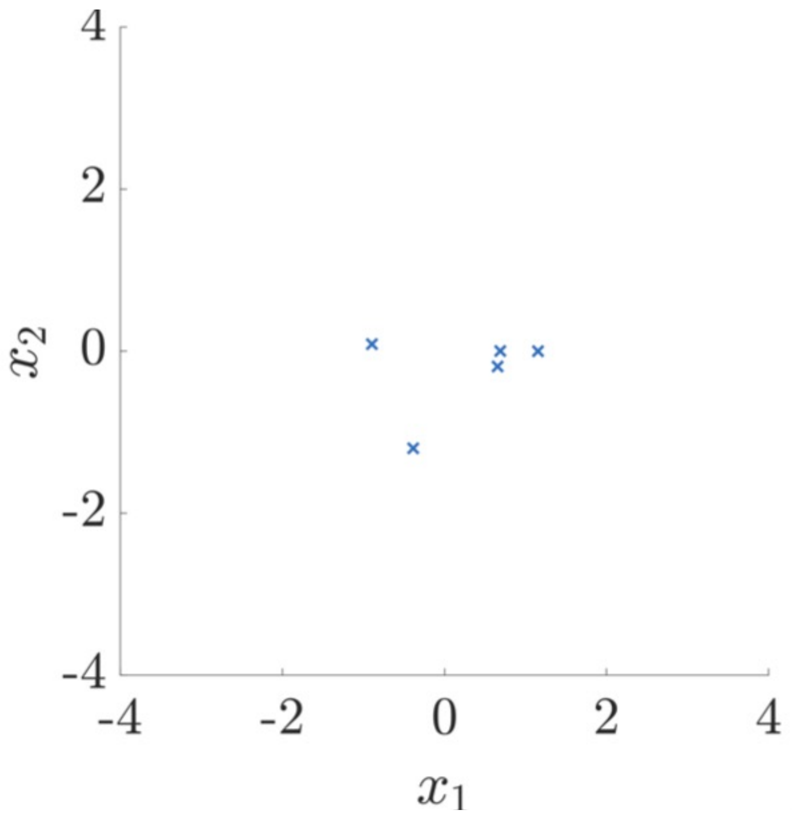
$$p(H|H) = p(H|biased)\cdot p(biased|H) + p(H|fair)\cdot p(fair|H) = 0.36$$

Example: Density Estimation

- Suppose that we decide to use an isotropic Gaussian likelihood with unknown mean θ to model the data on the right:

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \mathcal{N}(x_n; \theta, I)$$

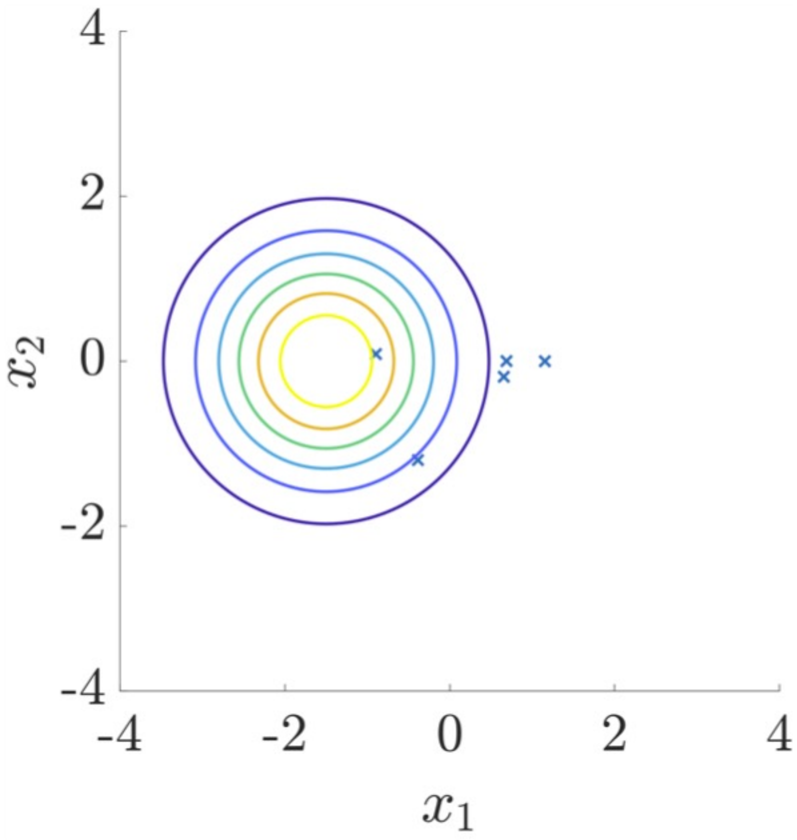
where I is a two-dimensional identity matrix



Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$p(\mathcal{D}|\theta = [-2,0]) = 0.00059 \times 10^{-5}$



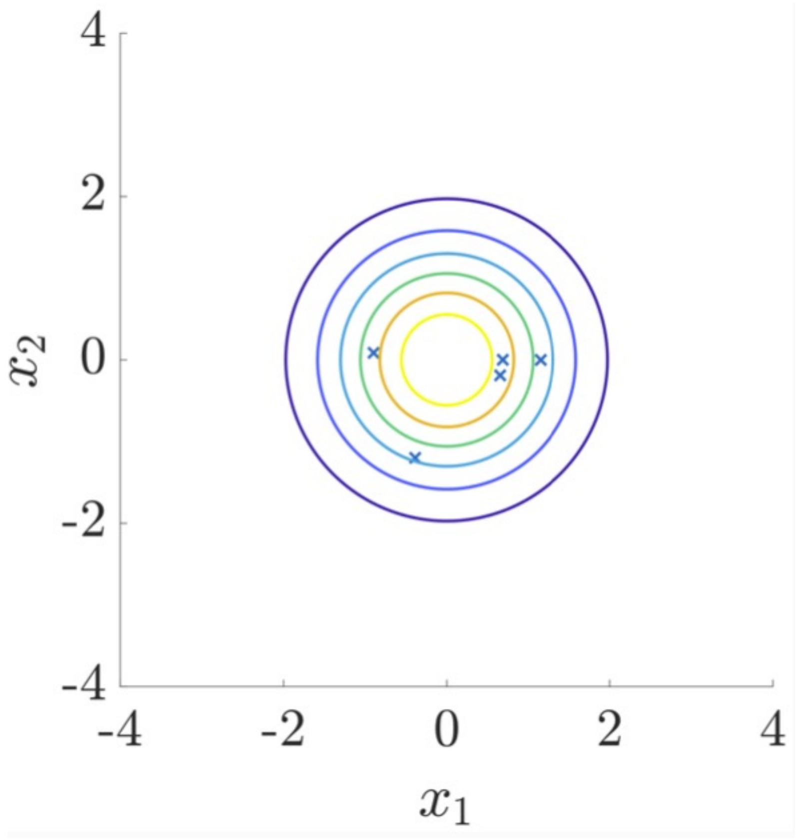
Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$p(\mathcal{D}|\theta = [-2,0]) = 0.00059{\times}10^{-5}$

Hypothesis 2: $\theta = [0, 0]$

$p(\mathcal{D}|\theta = [0,0]) = 0.99{\times}10^{-5}$



Example: Density Estimation

Highest likelihood

Hypothesis 1: $\theta = [-2, 0]$

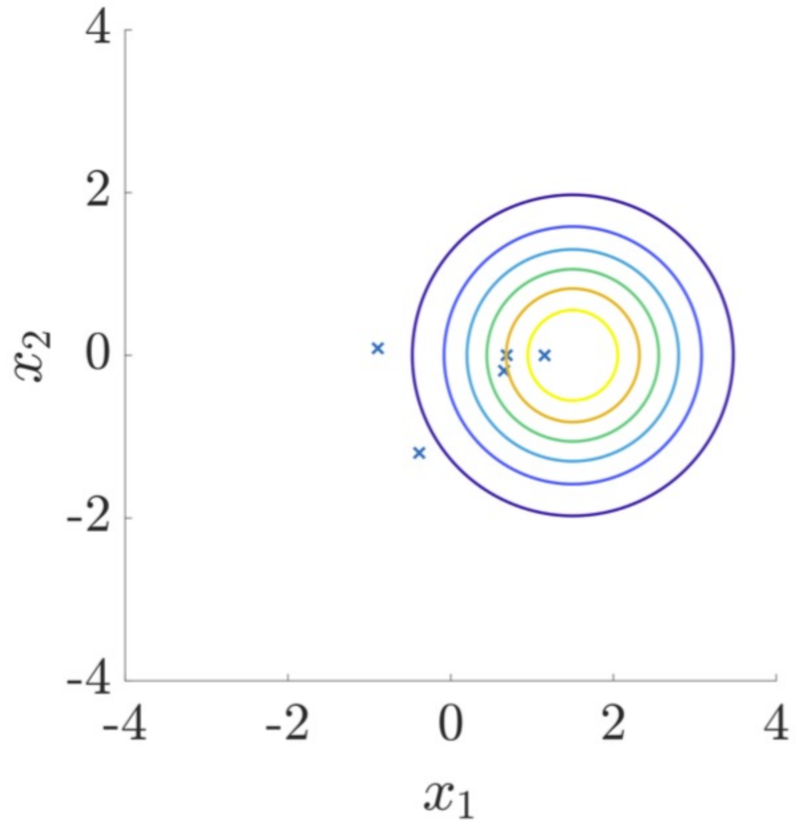
$p(\mathcal{D}|\theta = [-2,0]) = 0.00059 \times 10^{-5}$

Hypothesis 2: $\theta = [0, 0]$

$p(\mathcal{D}|\theta = [0,0]) = 0.99 \times 10^{-5}$

Hypothesis 3: $\theta = [2, 0]$

$p(\mathcal{D}|\theta = [2,0]) = 0.021 \times 10^{-5}$



The Posterior Predictive Averages over Hypotheses (1)

- The posterior predictive distribution allows us to average over each of our hypotheses, weighting each by their posterior probability.
- For example, in our density estimation example, lets introduce (the rather unusual but demonstrative) prior:

$$p(\theta) = \begin{cases} 0.05 & \text{if} \quad \theta = [-2, 0] \\ 0.05 & \text{if} \quad \theta = [0, 0] \\ 0.9 & \text{if} \quad \theta = [2, 0] \\ 0 & \text{otherwise} \end{cases}$$

The Posterior Predictive Averages over Hypotheses (2)

- Then we have:

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\theta)p(\theta|\mathcal{D})d\theta \\
&= \frac{1}{p(\mathcal{D})}\int p(x|\theta)p(\theta,\mathcal{D})d\theta \\
&= \frac{1}{p(\mathcal{D})}\Big(\mathcal{N}(x;[-2,0],I)\times 0.05\times p(\mathcal{D}|\theta=[-2,0]) \\
&\qquad + \mathcal{N}(x;[0,0],I)\times 0.05\times p(\mathcal{D}|\theta=[0,0]) \\
&\qquad + \mathcal{N}(x;[2,0],I)\times 0.9\times p(\mathcal{D}|\theta=[2,0])\Big)
\end{aligned}
$$

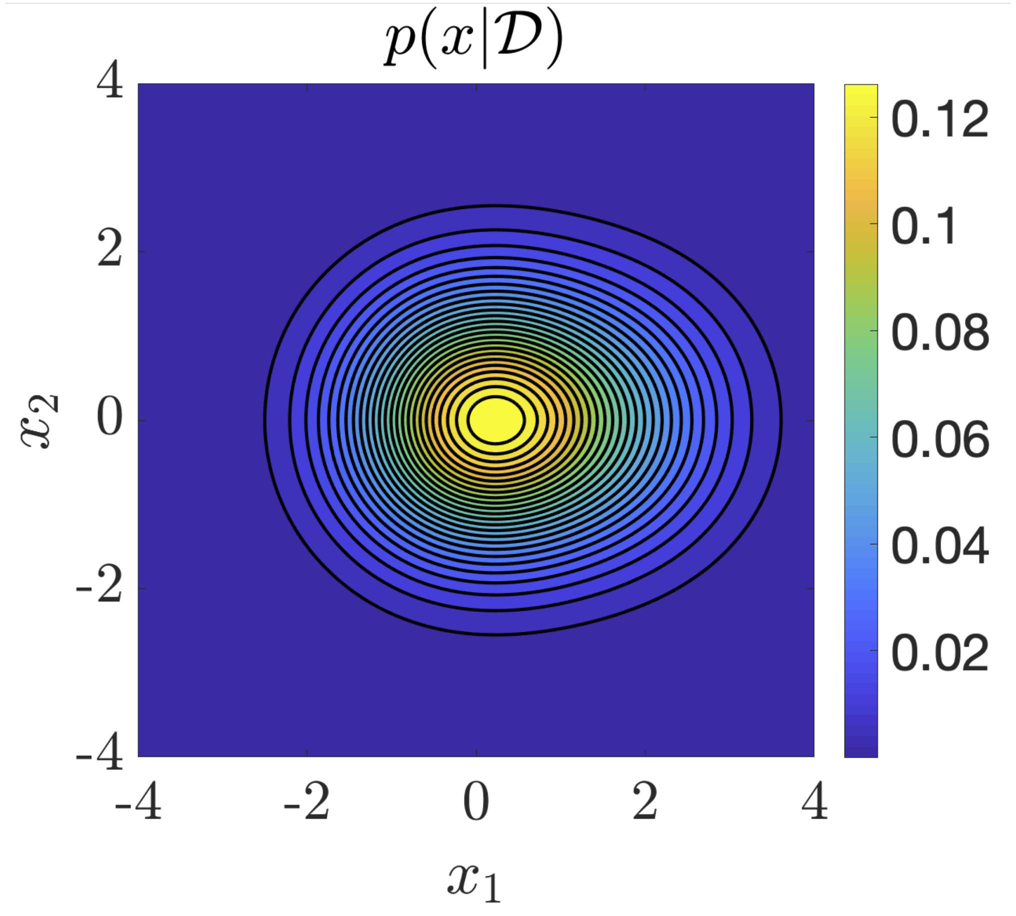
The Posterior Predictive Averages over Hypotheses (3)

- Inserting our likelihoods from earlier and trawling through the algebra gives

$$
\begin{aligned}
p(x|\mathcal{D}) =& 0.0004 \times \mathcal{N}(x; [-2, 0], I) \\
&+0.716 \times \mathcal{N}(x; [0, 0], I) \\
&+0.283 \times \mathcal{N}(x; [2, 0], I)
\end{aligned}
$$

We thus have that the posterior predictive is a weighted sum of the three possible predictive distributions

The Posterior Predictive Averages over Hypotheses



Some Subtleties

- Even though we average over θ, a Bayesian model is still implicitly assuming that there is still a single true θ
  - The averaging over hypotheses is from **our own uncertainty** as which one is correct
  - This can be problematic with lots of data given our model is an approximation
- In the limit of large data, the posterior is guaranteed to collapse to a point estimate:

$$p(\theta|x_{1:N}) \rightarrow \delta\left(\theta = \hat{\theta}\right) \text{ as } N \rightarrow \infty$$

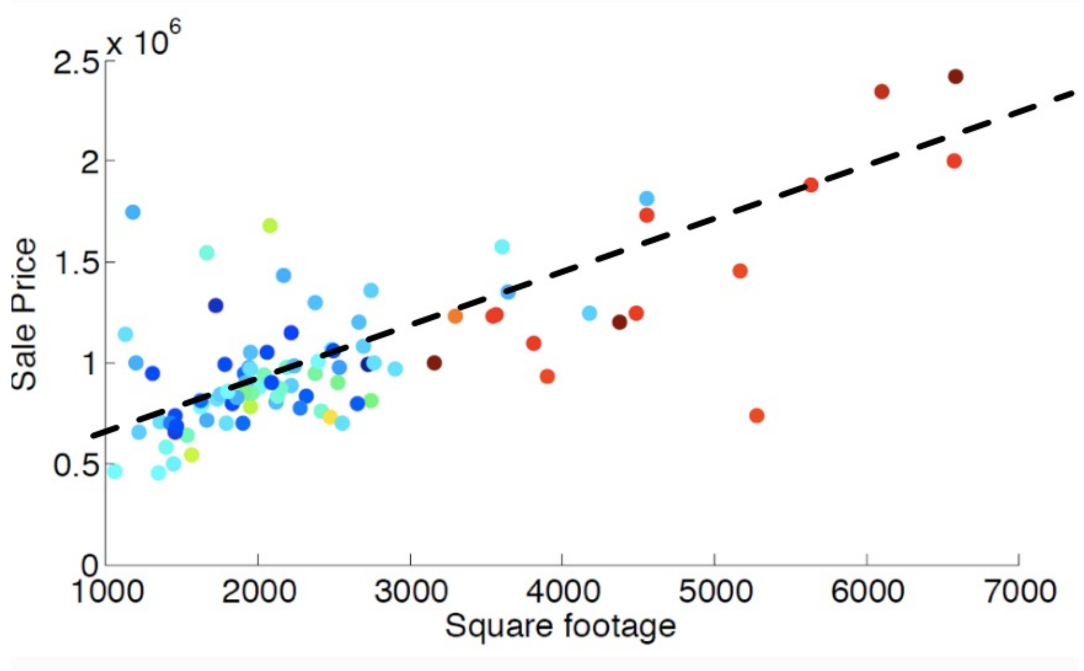
- The value of $\hat{\theta}$ and the exact nature of this convergence is dictated by the Bernstein–von Mises Theorem (see the lecture notes)
- Note that, subject to mild assumptions, $\hat{\theta}$ is independent of the prior: with enough data, the likelihood always dominates the prior

Example:
Bayesian Linear Regression

Linear Regression

House size is a good linear predictor for price (ignore the colors)

- Learn a **function** that maps size to price



Linear Regression

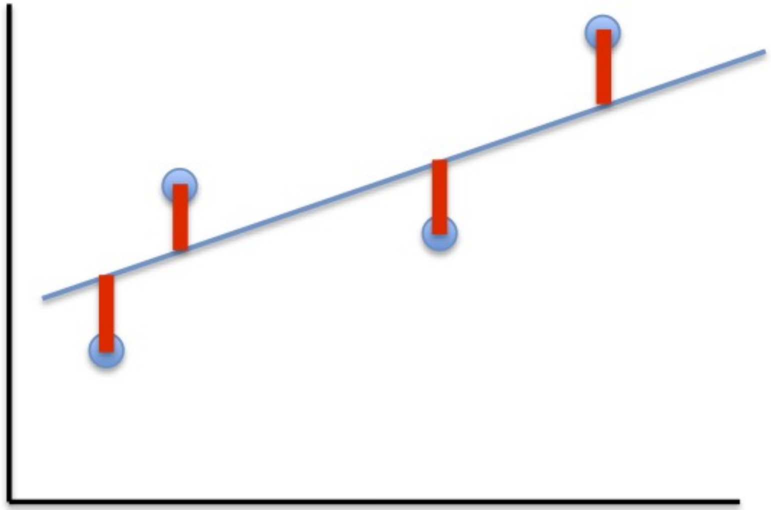
- **Inputs**: $x \in \mathbb{R}^D$ (where $D = 1$ for this example)
- **Outputs**: $y \in \mathbb{R}$
- **Data**: $D = \{x_n, y_n\}_{n=1}^N$
- **Regression model**: $y \approx x^{\mathrm{T}} w + b$ where $w \in \mathbb{R}^D$ and $b \in \mathbb{R}$

  We can simplify this notation by redefining $x \leftarrow [1, x^{\mathrm{T}}]^{\mathrm{T}}$ and $w \leftarrow [b, w^{\mathrm{T}}]^{\mathrm{T}}$, so that the model becomes $y \approx x^{\mathrm{T}} w$

Classical **least squares** linear regression is a discriminative method aiming to minimize the empirical mean squared error

$$L(w) = \frac{1}{N} \sum_{n=1}^{N} \left(y_n - x_n^{\mathrm{T}} w\right)^2$$

Linear Regression



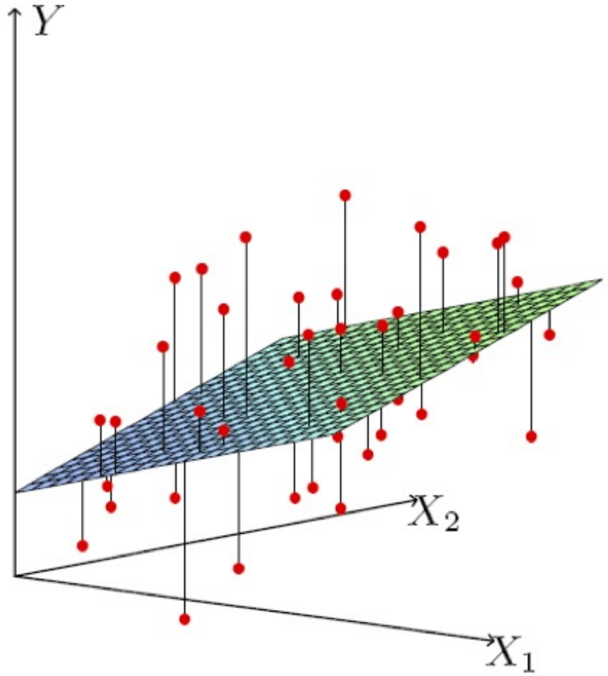


*Image credit: Pier Palamara

Bayesian Linear Regression

- Least square provides a point estimate without **uncertainty**

$$w^* = \underset{w}{\operatorname{argmin}}\, L(w)$$

- Bayesian method introduces uncertainty by building a **probabilistic** generic model based around linear regression and then being **Bayesian about the weights**

Bayesian Linear Regression: **Prior and Likelihood**

- For example, prior of w is a zero-mean Gaussian with a fixed covariance C

$$p(w) = \mathcal{N}(w; 0, C)$$

- And given input x, the output is $y = x^{\mathrm{T}}w$ plus a Gaussian noise, and datapoints are independent of each other:

$$p(y|x,w) = \prod_{n=1}^{N} p(y_n|x_n, w)$$

$$= \prod_{n=1}^{N} \mathcal{N}(y_n; x_n^{\mathrm{T}}w, \sigma^2)$$

where σ is a fixed standard deviation

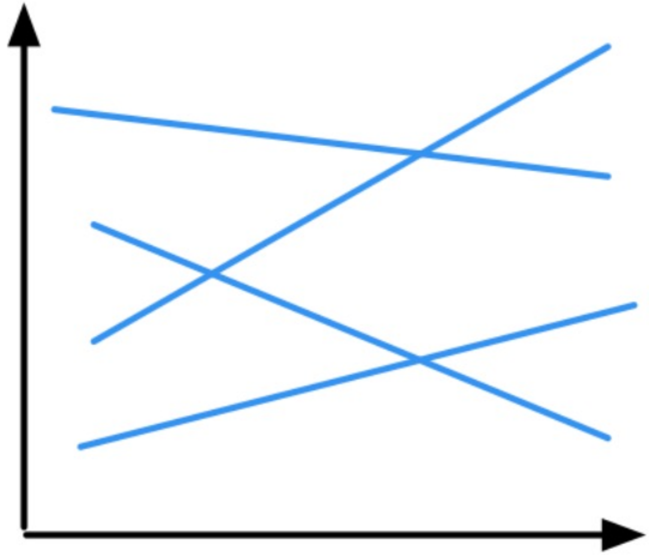


*Image credit: Roger Grosse

Bayesian Linear Regression: **Posterior**

- Using Bayes' rule (and some math) to derive the posterior. See Bishop, *Pattern recognition and machine learning,* 2006, Chapter 3

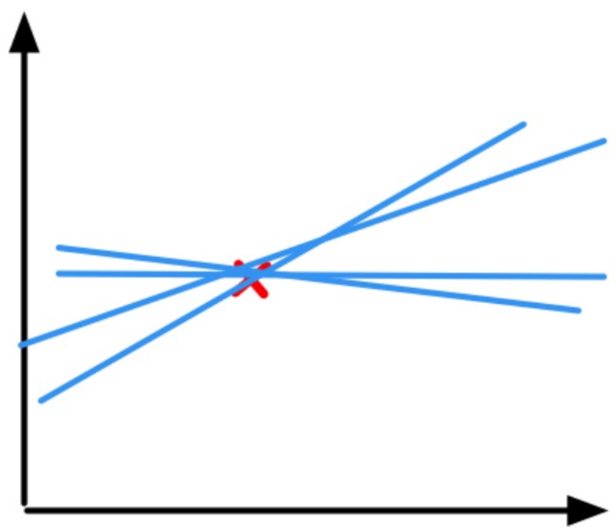
$$
\begin{aligned}
p(w|\mathbf{x},\mathbf{y}) &\propto p(w)p(\mathbf{y}|\mathbf{x},w) \\
&= \mathcal{N}(w;0,C)\prod_{n=1}^{N}\mathcal{N}(y_n;x_n^T w,\sigma^2)
\end{aligned}
$$

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

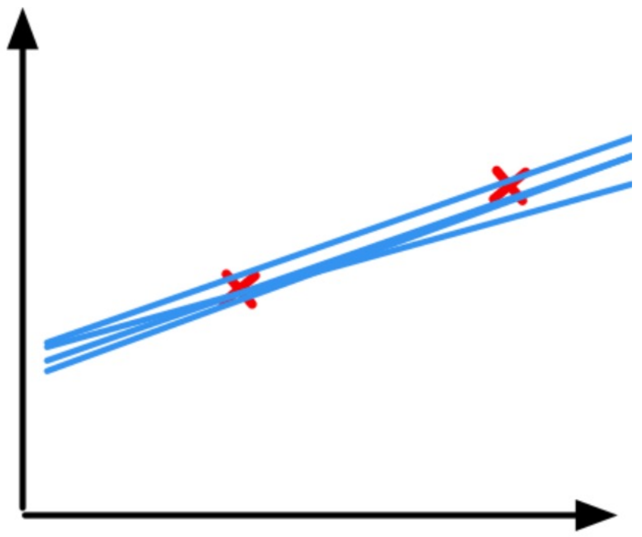
$$
\text{where} \quad m = S^{-1}\mathbf{x}^T\mathbf{y}/\sigma^2 \quad \text{and} \quad S = \left(C^{-1} + \frac{\mathbf{x}^T\mathbf{x}}{\sigma^2}\right)^{-1}
$$

Bayesian Linear Regression: **Posterior**

- Note here that the fact the prior and posterior share the same form is **highly special** case. This is known as a **conjugate distribution** and it is why we were able to find an analytic solution for the posterior.

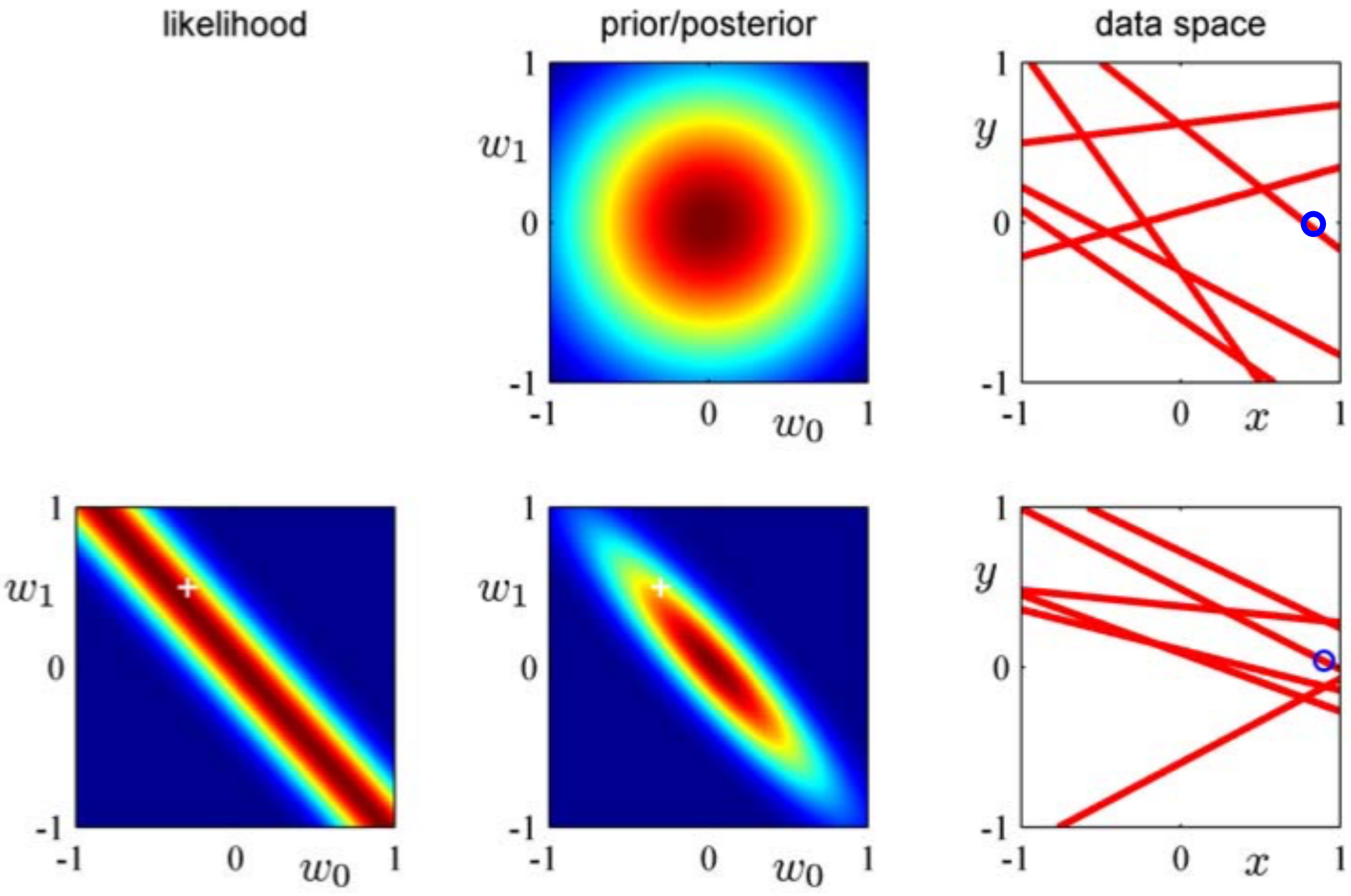


Posterior after 1 observation



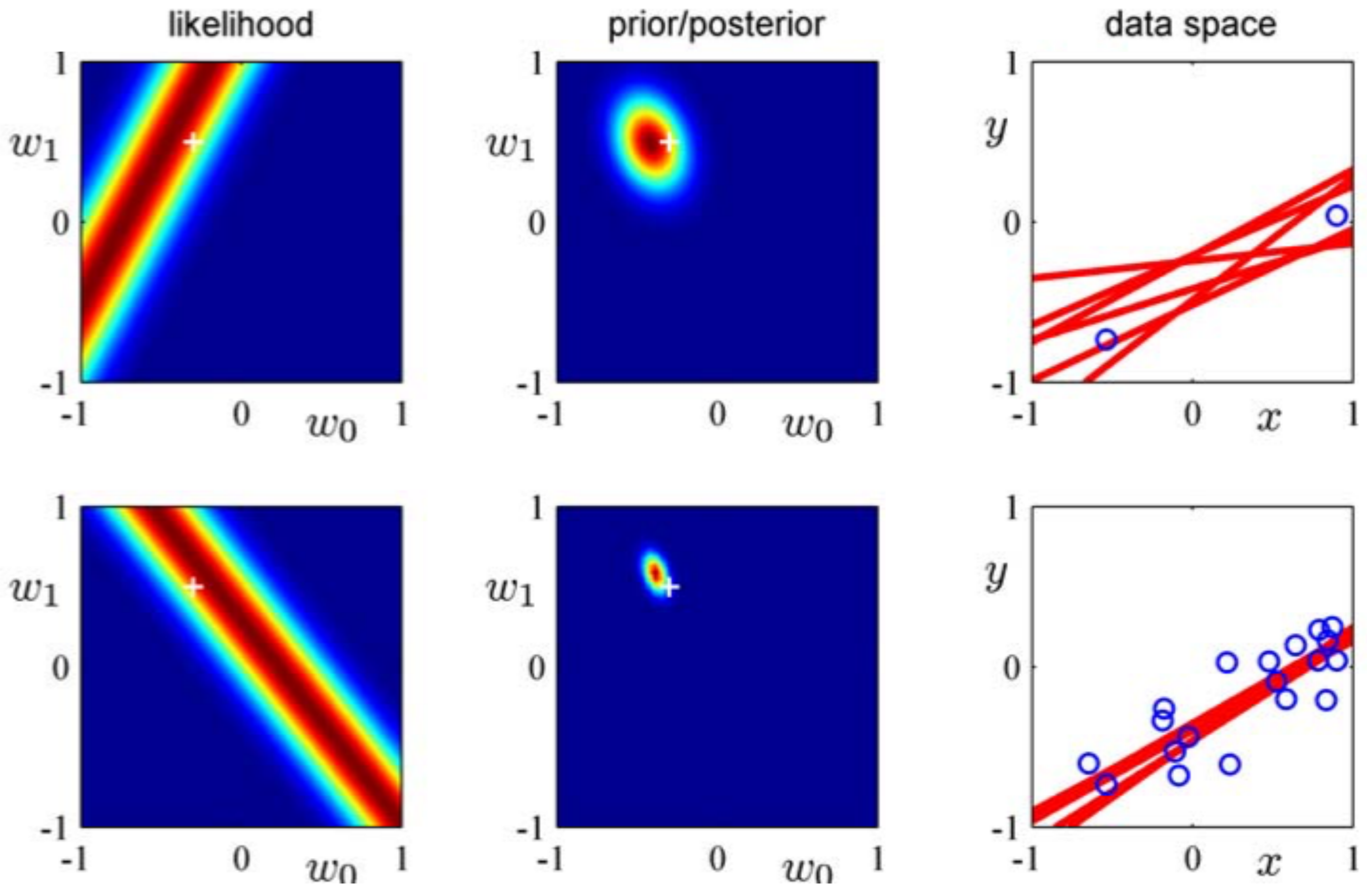
Posterior after 2 observations

Bayesian Linear Regression: Posterior



*Bishop, *Pattern recognition and machine learning*.

Bayesian Linear Regression: Posterior



*Bishop, *Pattern recognition and machine learning*.

Bayesian Linear Regression: **Posterior Predictive**

- Some more math to derive the posterior predictive… where the result is again a consequence of Gaussian identities, and m and S are as before

$$
\begin{aligned}
p(\tilde{y}|\tilde{x}, \mathbf{x}, \mathbf{y}) &= \int p(\tilde{y}|\tilde{x}, w)p(w|\mathbf{x}, \mathbf{y})dw \\
&= \int \mathcal{N}(\tilde{y}; \tilde{x}^T w, \sigma^2)\, \mathcal{N}(w; m, S)\, dw \\
&= \mathcal{N}\left(\tilde{y}\ ;\ \tilde{x}^T m,\ \left(\tilde{x}^T S^{-1} \tilde{x} + \frac{1}{\sigma^2}\right)^{-1}\right)
\end{aligned}
$$

Bayesian Linear Regression: **Posterior Predictive**

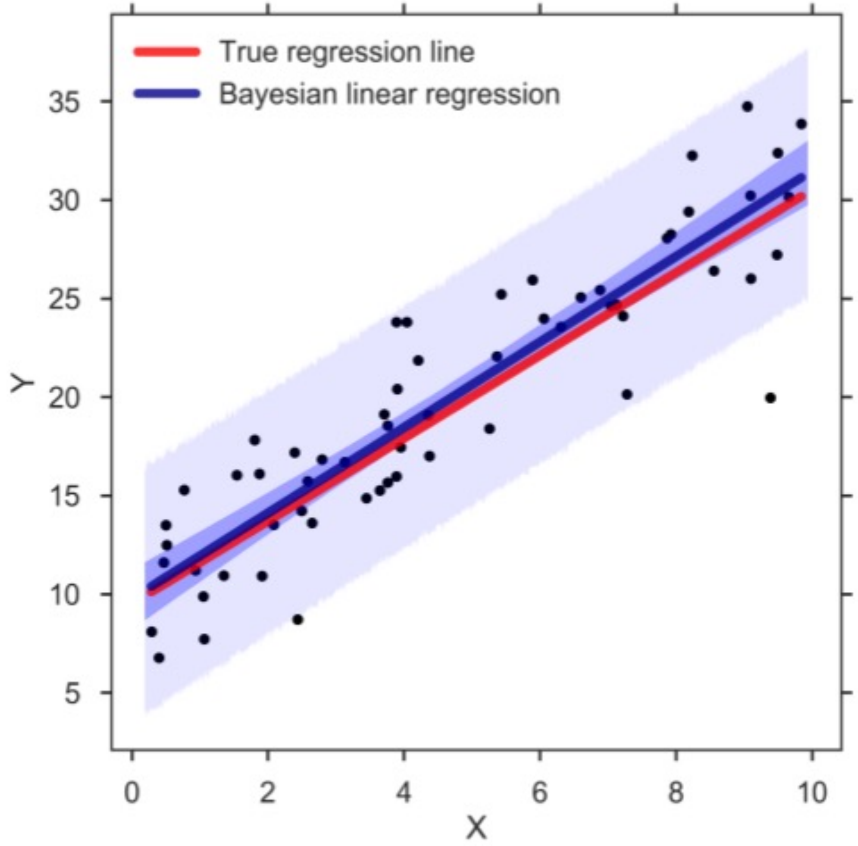


*Image credit: https://www.dataminingapps.com/2017/09/simple- linear- regression- do- it- the- bayesian- way/

Further Reading

- Information on non-parametric models and Gaussian processes in course notes
- Bishop, Pattern recognition and machine learning, Chapters 1-3
- K P Murphy. Machine learning: a probabilistic perspective. 2012, Chapter 5
- D Barber. Bayesian reasoning and machine learning. 2012, Chapter 12
- T P Minka. “Bayesian model averaging is not model combination”. In: (2000)
- Zoubin Ghahramani on Bayesian machine learning (there are various alternative variations of this talk): https://www.youtube.com/watch?v=y0FgHOQhG4w
- Iain Murray on Probabilistic Modeling https://www.youtube.com/watch?v=pOtvyVYAuW4