Univariate Bayesian Linear Regression

MLAI Lecture 10

Neil D. Lawrence

Department of Computer Science
Sheffield University

16th October 2012

Outline

Review: Overdetermined Systems

Underdetermined Systems

Bayesian Perspective

Bayesian Regression

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$t_1 = mx_1 + c$$
$$t_2 = mx_2 + c$$

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$t_1 - t_2 = m(x_1 - x_2)$$

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

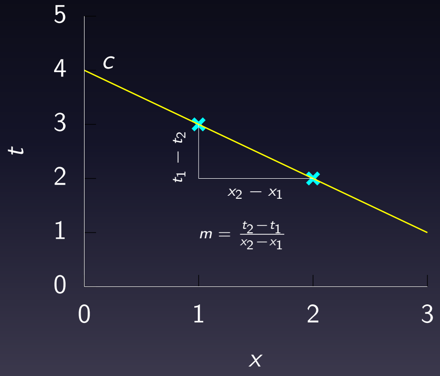
$$\frac{t_1 - t_2}{x_1 - x_2} = m$$

Two Simultaneous Equations

A system of two simultaneous equations with two unknowns.

$$m = \frac{t_2 - t_1}{x_2 - x_1}$$

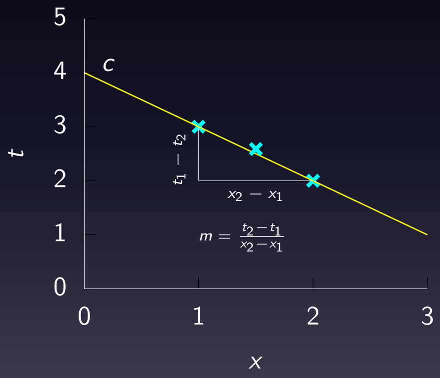
$$c = t_1 - m x_1$$



Two Simultaneous Equations

How do we deal with three simultaneous equations with only two unknowns?

$$t_1 = mx_1 + c$$
$$t_2 = mx_2 + c$$
$$t_3 = mx_3 + c$$



Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
t_1 = & mx_1 + c \\
t_2 = & mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$t_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
t_1 &= mx_1 + c + \epsilon_1 \\
t_2 &= mx_2 + c + \epsilon_2 \\
t_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

Overdetermined System

- With two unknowns and two observations:

$$
\begin{aligned}
t_1 =& mx_1 + c \\
t_2 =& mx_2 + c
\end{aligned}
$$

- Additional observation leads to *overdetermined* system.

$$t_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$
\begin{aligned}
t_1 &= mx_1 + c + \epsilon_1 \\
t_2 &= mx_2 + c + \epsilon_2 \\
t_3 &= mx_3 + c + \epsilon_3
\end{aligned}
$$

Overdetermined System

- With two unknowns and two observations:

$$t_1 = mx_1 + c$$
$$t_2 = mx_2 + c$$

- Additional observation leads to *overdetermined* system.

$$t_3 = mx_3 + c$$

- This problem is solved through a noise model $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$

$$t_1 = mx_1 + c + \epsilon_1$$
$$t_2 = mx_2 + c + \epsilon_2$$
$$t_3 = mx_3 + c + \epsilon_3$$

Noise Models

- We aren't modeling entire system.
- Noise model gives mismatch between model and data.
- Gaussian model justified by appeal to central limit theorem.
- Other models also possible (Student-t for heavy tails).
- Maximum likelihood with Gaussian noise leads to *least squares*.

Outline

Review: Overdetermined Systems

Underdetermined Systems

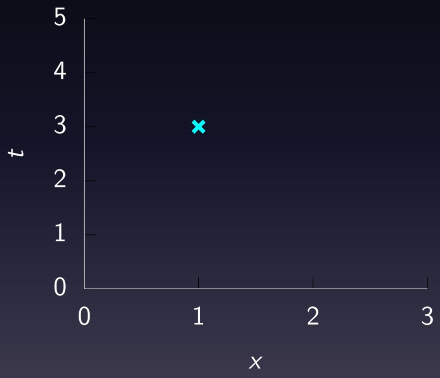
Bayesian Perspective

Bayesian Regression

Underdetermined System

What about two unknowns and *one* observation?

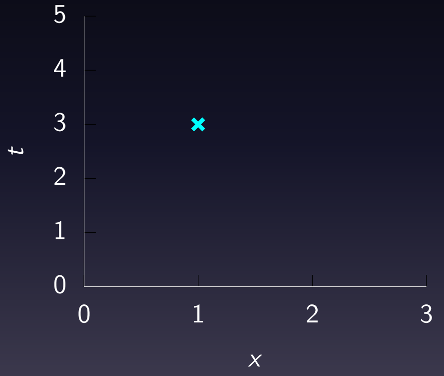
$$t_1 = mx_1 + c$$



Underdetermined System

Can compute m given c.

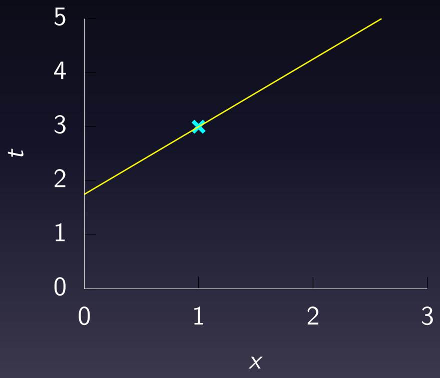
$$m = \frac{t_1 - c}{x}$$



Underdetermined System

Can compute m given c.

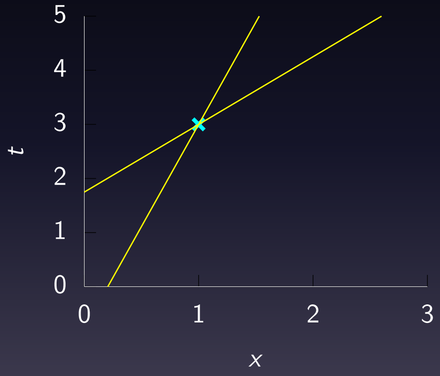
$$c = 1.75 \implies m = 1.25$$



Underdetermined System

Can compute m given c.

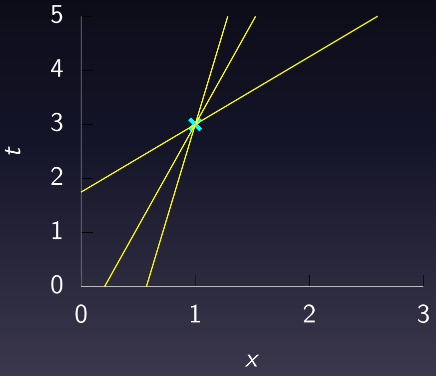
$$c = -0.777 \Longrightarrow m = 3.78$$



Underdetermined System

Can compute m given c.

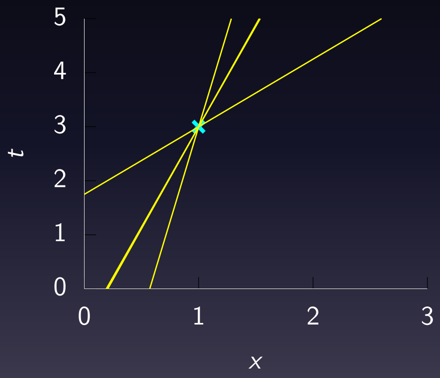
$$c = -4.01 \Longrightarrow m = 7.01$$



Underdetermined System

Can compute m given c.

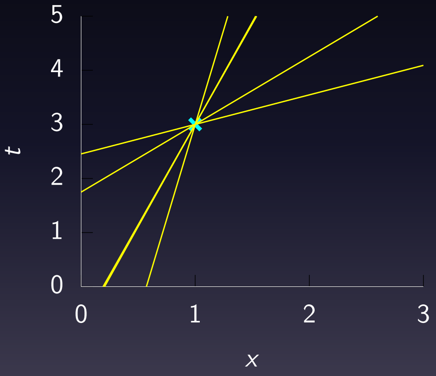
$$c = -0.718 \implies m = 3.72$$



Underdetermined System

Can compute m given c.

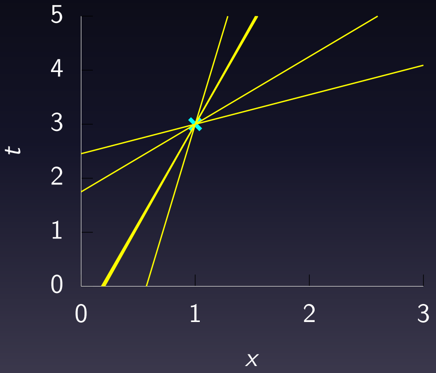
$$c = 2.45 \implies m = 0.545$$



Underdetermined System

Can compute m given c.

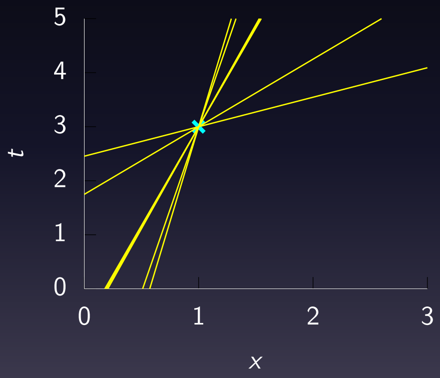
$$c = -0.657 \implies m = 3.66$$



Underdetermined System

Can compute m given c.

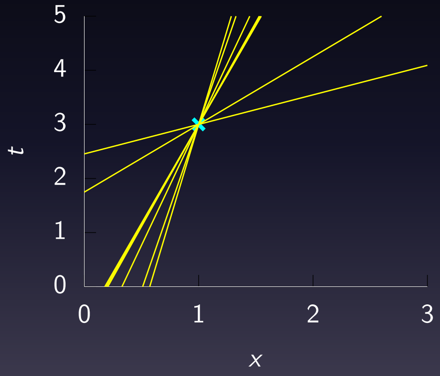
$$c = -3.13 \Longrightarrow m = 6.13$$



Underdetermined System

Can compute m given c.

$$c = -1.47 \implies m = 4.47$$

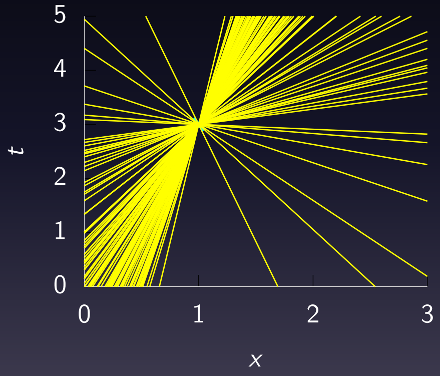


Underdetermined System

Can compute m given c.
Assume

$$c \sim \mathcal{N}(0, 4),$$

we find a distribution of solutions.



Different Types of Uncertainty

- The first type of uncertainty we are assuming is *aleatoric* uncertainty.
- The second type of uncertainty we are assuming is *epistemic* uncertainty.

Aleatoric Uncertainty

- This is uncertainty we couldn't know even if we wanted to. e.g. the result of a football match before it's played.
- Where a sheet of paper might land on the floor.

Outline

Review: Overdetermined Systems

Underdetermined Systems

Bayesian Perspective

Bayesian Regression

Bayesian Approach

- Likelihood for the regression example has the form

$$p(\mathbf{t}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}\left(t_i|\mathbf{w}^\top \phi_i, \sigma^2\right).$$

- Suggestion was to maximize this likelihood with respect to **w**.
- This can be done with gradient based optimization of the log likelihood.
- Alternative approach: integration across **w**.
- Consider expected value of likelihood under a range of potential **w**s.
- This is known as the *Bayesian* approach.

Note on the Term Bayesian

- We will use Bayes' rule to invert probabilities in the Bayesian approach.
    - Bayesian is not named after Bayes' rule (v. common confusion).
    - The term Bayesian refers to the treatment of the parameters as stochastic variables.
    - This approach was proposed by Laplace (1774) and Bayes (1763) independently.
    - For early statisticians this was very controversial (Fisher et al).

Bernoulli Distribution

Bernoulli Distribution

- Jacob Bernoulli described this distribution in terms of an ‘urn’.
- Write as a function

$$P(T = t) = \pi^t (1 - \pi)^{1-t}$$

Bernoulli Distribution

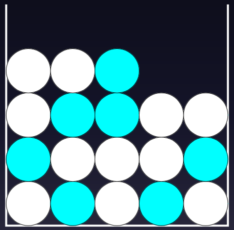
- Jacob Bernoulli described this distribution in terms of an ‘urn’.
- Write as a function

$$P(T = t) = \pi^t (1 - \pi)^{1-t}$$

Bernoulli Distribution

- Jacob Bernoulli described this distribution in terms of an 'urn'.
- Write as a function

$$P(T = t) = \pi^t (1 - \pi)^{1-t}$$

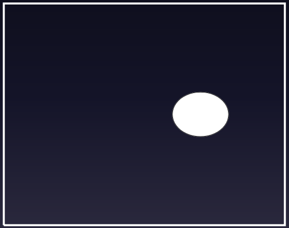


Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.

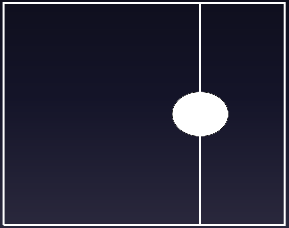
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



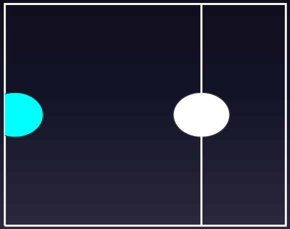
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



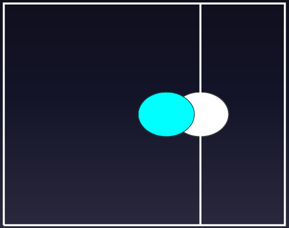
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



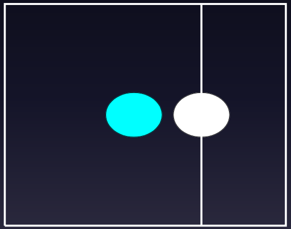
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



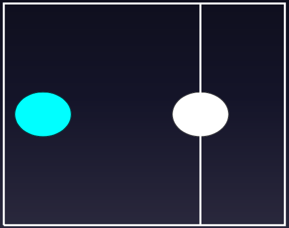
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That ‘parameter’ is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



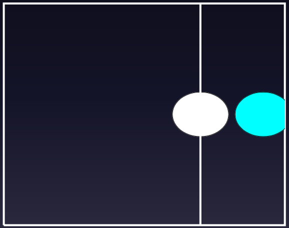
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



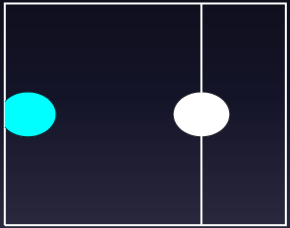
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



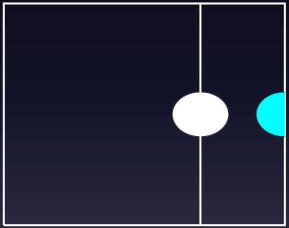
Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



Bernoulli Distribution Revisited

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- The position of the first ball gives the parameter π.
- That 'parameter' is *itself* a random variable.
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



Bayesian Controversy

- Bayesian controversy relates to treating *epistemic* uncertainty as *aleatoric* uncertainty.
- Another analogy:
  - Before a football match the uncertainty about the result is *aleatoric*.
  - If I watch a recorded match *without* knowing the result the uncertainty is *epistemic*.

Simple Bayesian Inference

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

- Four components:
  - Prior distribution: represents belief about parameter values before seeing data.
  - Likelihood: gives relation between parameters and data.
  - Posterior distribution: represents updated belief about parameters after data is observed.
  - Marginal likelihood: represents assessment of the quality of the model. Can be compared with other models (likelihood/prior combinations). Ratios of marginal likelihoods are known as Bayes factors.

Outline

Review: Overdetermined Systems

Underdetermined Systems

Bayesian Perspective

Bayesian Regression

Prior Distribution

- Bayesian inference requires a prior on the parameters.
- The prior represents your belief *before* you see the data of the likely value of the parameters.
- For linear regression, consider a Gaussian prior on the intercept:

$$c \sim \mathcal{N}(0, \alpha_1)$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1} c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (t_i - m x_i - c)^2\right)$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)}{p(\mathbf{t}|\mathbf{x}, m, \sigma^2)}$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1} c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (t_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) = \frac{p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)}{\int p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)\mathrm{d}c}$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) \propto p(\mathbf{t}|\mathbf{x}, c, m, \sigma^2)p(c)$$

$$
\begin{aligned}
\log p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - c - mx_i)^2 - \frac{1}{2\alpha_1}c^2 + \text{const} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - mx_i)^2 - \left(\frac{N}{2\sigma^2} + \frac{1}{2\alpha_1}\right)c^2 \\
&\quad + c\frac{\sum_{i=1}^{N}(t_i - mx_i)}{\sigma^2},
\end{aligned}
$$

complete the square of the quadratic form to obtain

$$
\log p(c|\mathbf{t}, \mathbf{x}, m, \sigma^2) = -\frac{1}{2\varsigma^2}(c - \mu)^2 + \text{const},
$$

where $\varsigma^2 = \left(N\sigma^{-2} + \alpha_1^{-1}\right)^{-1}$ and $\mu = \frac{\varsigma^2}{\sigma^2}\sum_{n=1}^{N}(t_i - mx_i)$.

Gaussian Noise

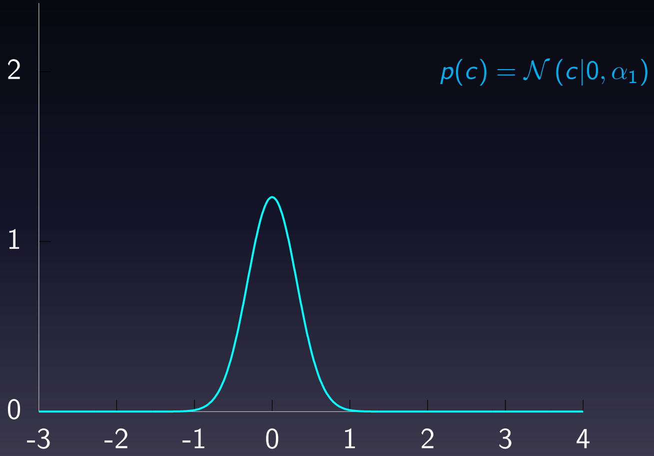


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Gaussian Noise

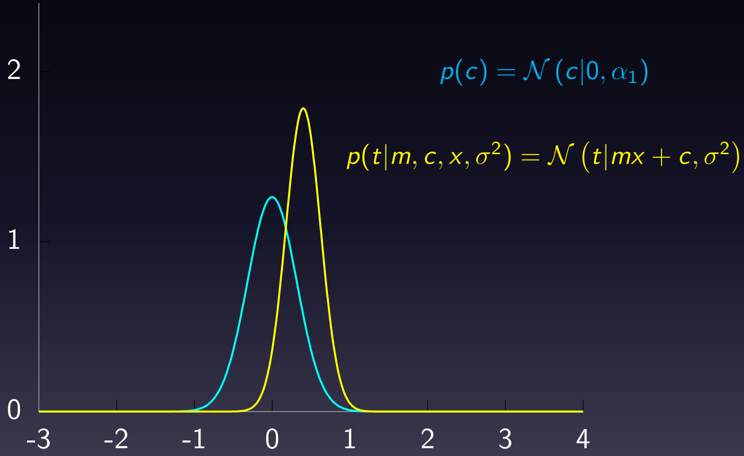


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Gaussian Noise

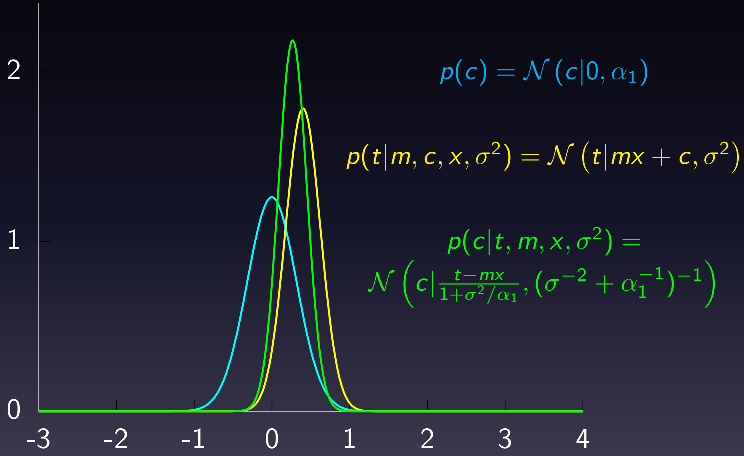


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

The Joint Density

- Really want to know the *joint* posterior density over the parameters c *and* m.
- Could now integrate out over m, but it's easier to consider the multivariate case.

Aleatoric Uncertainty

- This is uncertainty we couldn't know even if we wanted to. e.g. the result of a football match before it's played.
- Where a sheet of paper might land on the floor.

Epistemic Uncertainty

- This is uncertainty we could in principal know the answer too. We just haven't observed enough yet, e.g. the result of a football match *after* it's played.
- What colour socks your lecturer is wearing.

Reading

- Bishop Section 1.2.3 (pg 21–24).
- Bishop Section 1.2.6 (start from just past eq 1.64 pg 30-32).
- Rogers and Girolami use an example of a coin toss for introducing Bayesian inference Chapter 3, Sections 3.1-3.4 (pg 95-117). Although you also need the beta density which we haven't yet discussed. This is also the example that Laplace used.

References I

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763. [DOI].

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

P. S. Laplace. Mémoire sur la probabilité des causes par les évènemens. In *Mémoires de mathèmatique et de physique, presentés à lAcadémie Royale des Sciences, par divers savans, & lù dans ses assemblées 6*, pages 621–656, 1774. Translated in Stigler (1986).

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .

S. M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378, 1986.