Machine Learning
CS342

Lecture 10: Probabilistic (Bayesian) Inference

Dr. Theo Damoulas
T.Damoulas@warwick.ac.uk

Office hours: Mon & Fri 10-11am @ CS 307

The road to full Probabilistic Inference

So far for linear regression we have seen two main approaches:

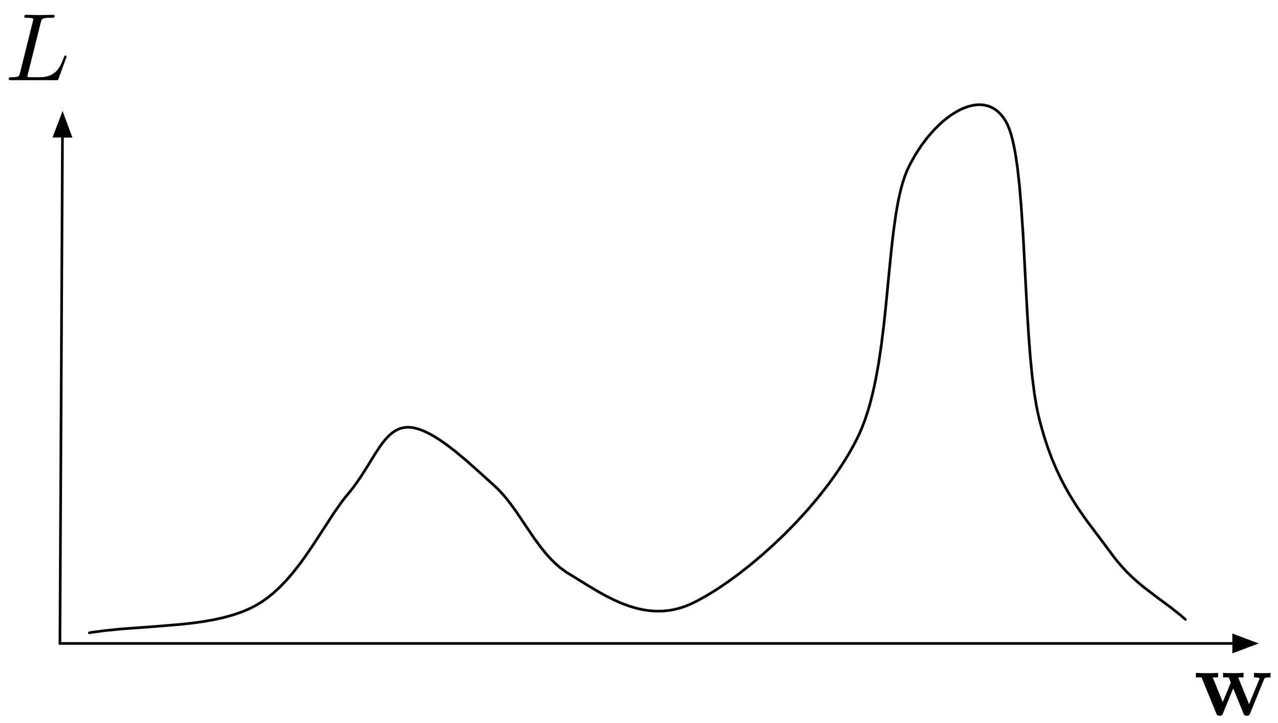
- **Minimise a (Squared Error) Loss function**
  - Ordinary Least Squares [Minimise Loss]
  - Ridge regression [Minimise (Loss + L2)]
  - The Lasso [Minimise (Loss + L1)]
- **Maximise the (Gaussian) Likelihood function**

OLS and Maximum Likelihood (Gaussian noise and likelihood) same 'best' parameter values:

$$\boxed{\widehat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}}$$

The road to full Probabilistic Inference

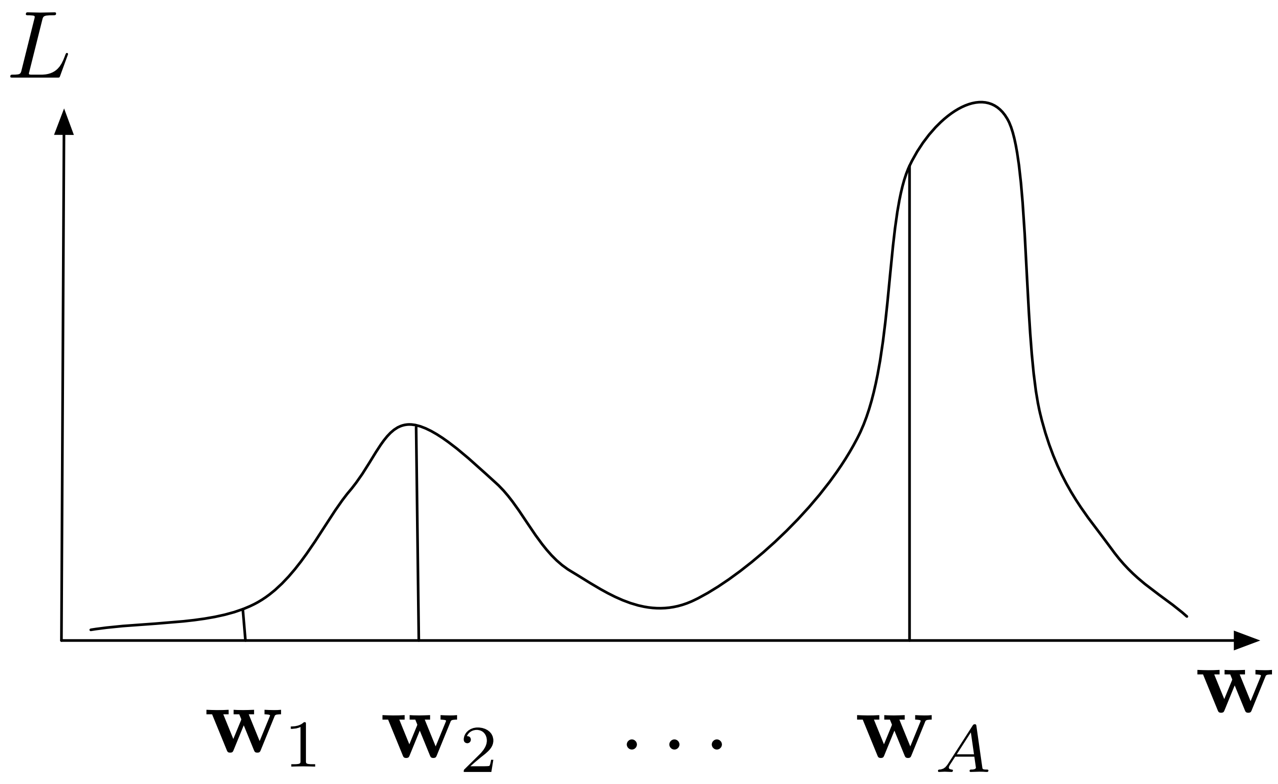
So far we are finding the "best" parameters (Loss/Likelihood)
Is there one best parameter?



Might be more then one "best" parameter
Different values might give different predictions
How many values are "best" might be telling us something…
Uncertainty? Evidence?
Parameters as random variables! Place distributions…

The road to full Probabilistic Inference

Rogers & Girolami, Ch. 3

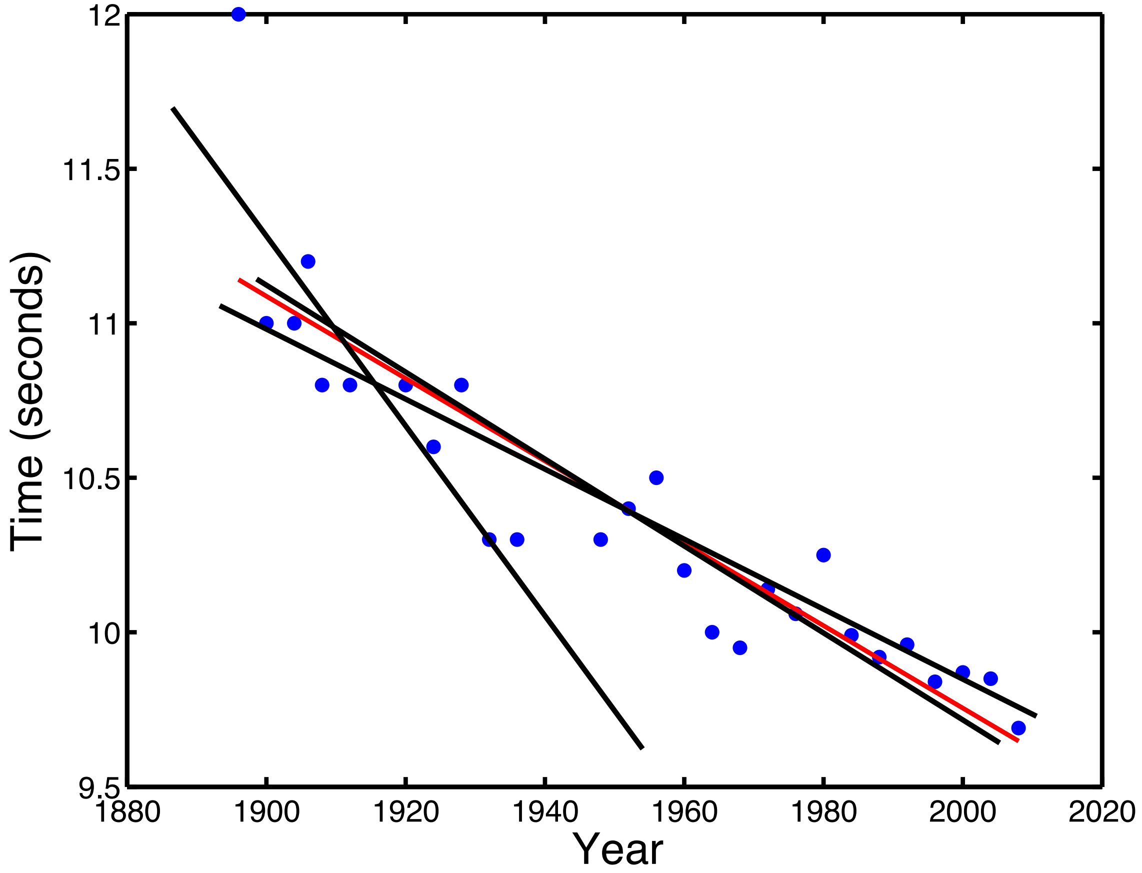


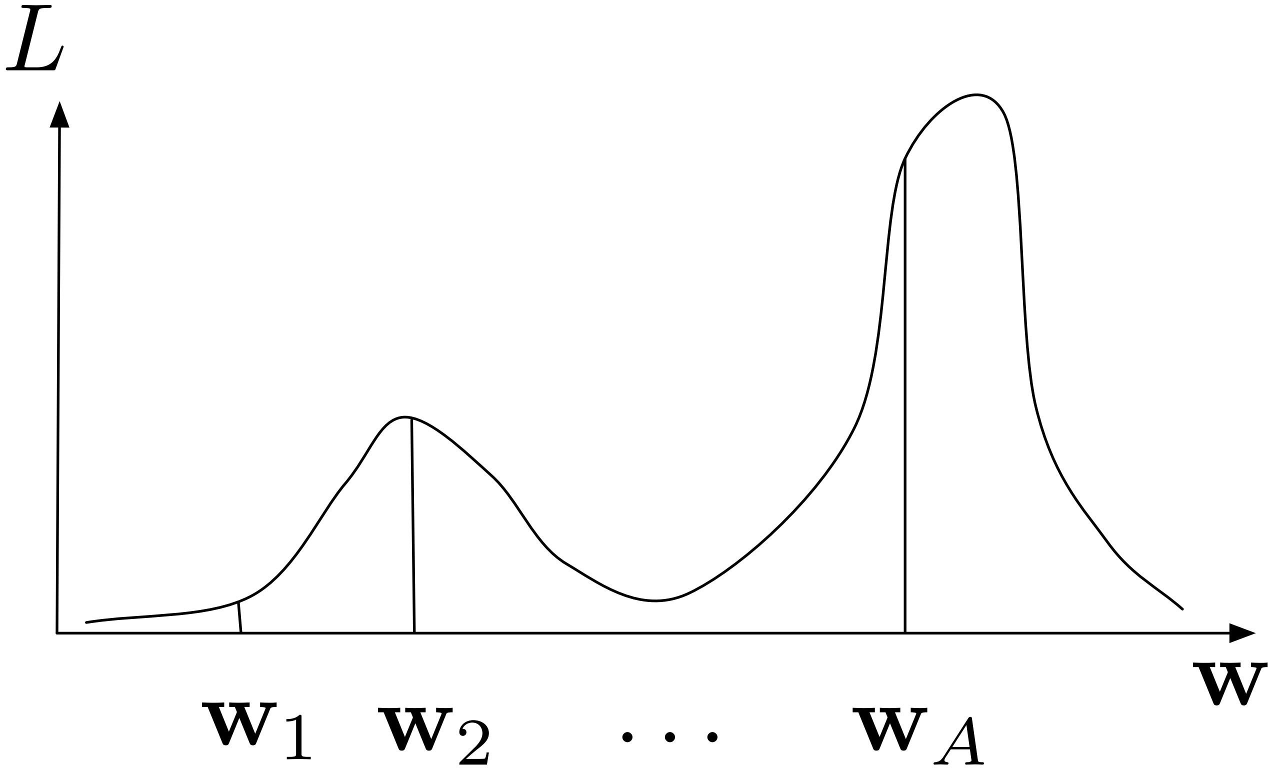
Prediction is some function of w:

$$t^* = \mathbf{x}^* \hat{\mathbf{w}} \quad \text{e.g. univariate case} \quad = \hat{w_0} + \hat{w_1} x_1^*$$

Maybe instead of “single best” parameter we use many?
How?

The road to full Probabilistic Inference: Averaging

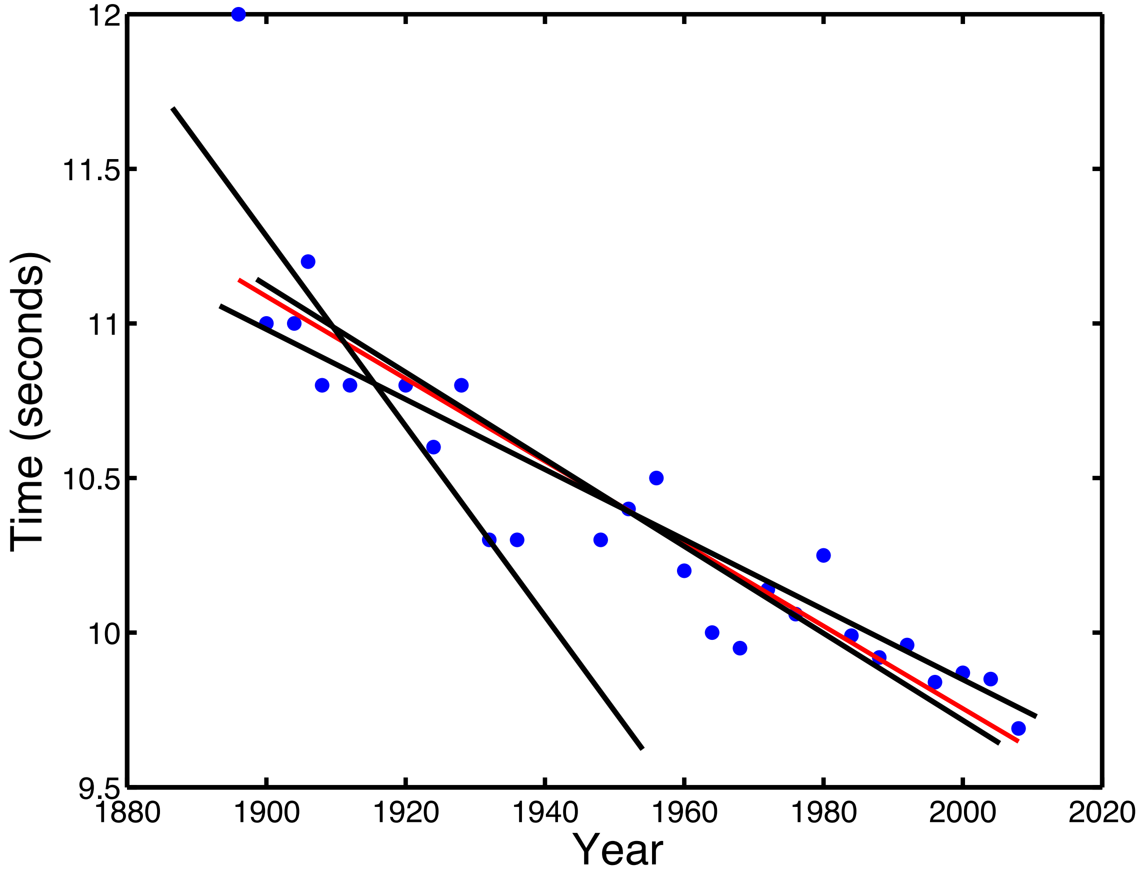


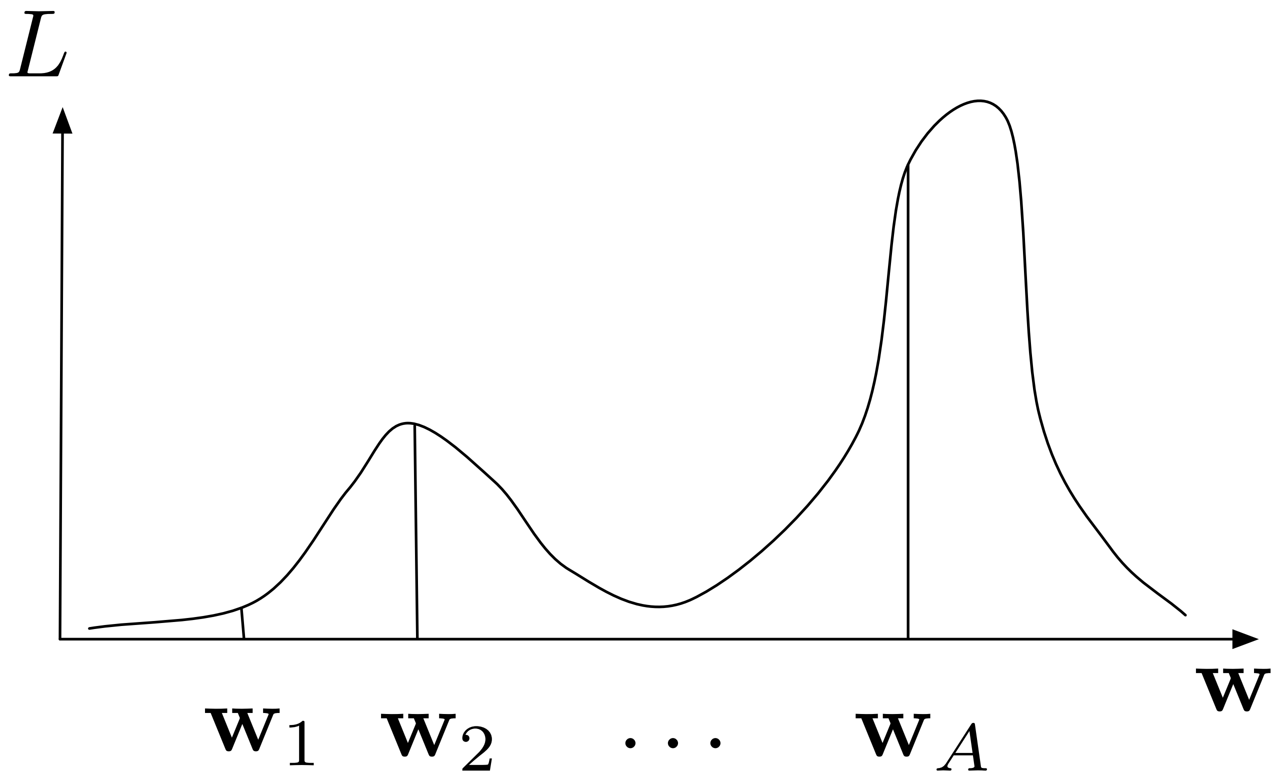


Every parameter setting **w** gives me a hypothesis (line) from my hypothesis space (lines in this case)

Every hypothesis (line) will give a prediction: $f(\mathbf{w})$

The road to full Probabilistic Inference: Averaging





What if **w** is a random variable? $p(\mathbf{w}|\text{stuff})$

Since we have a pdf for **w** we can use every value of **w**!

How do we predict from many “lines” or functions of **w**?

The road to full Probabilistic Inference: Averaging

What do we have so far:

w is a random variable with a pdf $p(\mathbf{w}|\text{stuff})$

Given a specific setting **w** we predict with $f(\mathbf{w})$

Overall prediction then should be this expectation:

$$\mathbb{E}_{p(\mathbf{w}|\text{stuff})} f(\mathbf{w}) = \int f(\mathbf{w}) p(\mathbf{w}|\text{stuff}) d\mathbf{w}$$

This is an average of predictions from each possible w weighted by how likely that w value is!

Expectations refresher

When I have discrete RVs, e.g. a dice roll, the expected value is just the sum of all possible events weighted by their probability of happening

$$\mathbb{E}_{P(X)} X = \sum_x P(X = x) x$$

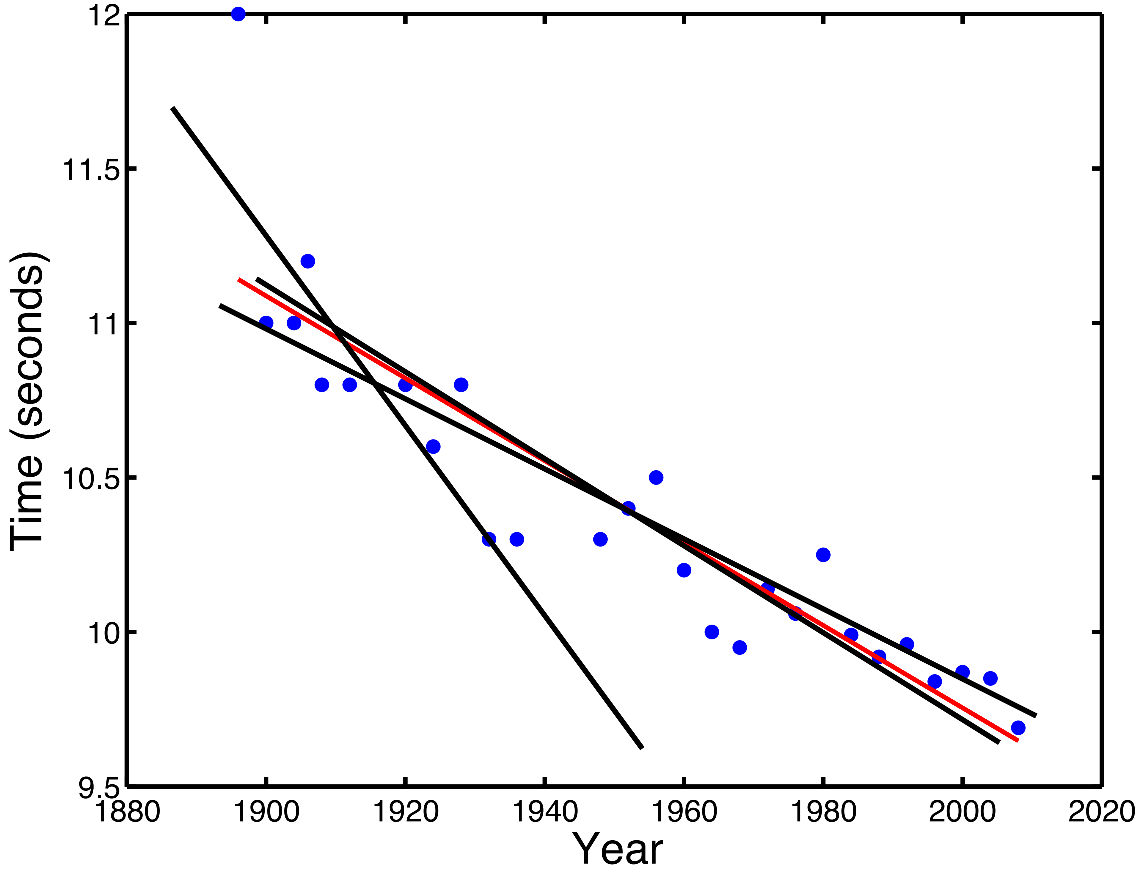
For continuous RVs, I have infinite many events and a density function

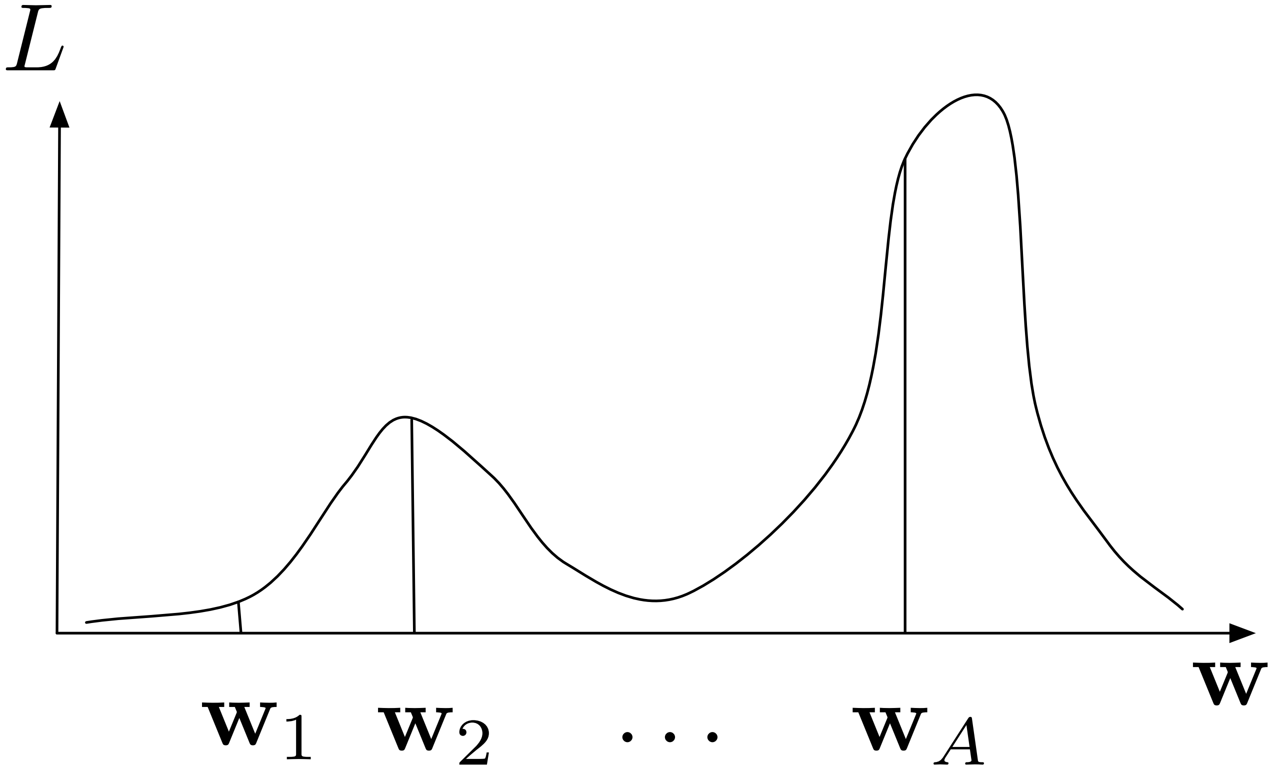
$$\mathbb{E}_{p(x)} x = \int x p(x) dx$$

So we are just applying the same principle

$$\mathbb{E}_{p(\mathbf{w}|\text{stuff})} f(\mathbf{w}) = \int f(\mathbf{w}) p(\mathbf{w}|\text{stuff}) d\mathbf{w}$$

The road to full Probabilistic Inference





$$\mathbb{E}_{p(\mathbf{w}|\text{stuff})} f(\mathbf{w}) = \int f(\mathbf{w}) p(\mathbf{w}|\text{stuff}) d\mathbf{w}$$

Great!! But what is this “stuff”? What should our parameters depend on? How do we compute p(**w** l stuff)?

Bayes rule

“Stuff” should include data: **X**,**t**
i.e. what we know about **w** *after we have seen the evidence* (data)

$$p(\mathbf{w}|\mathbf{X},\mathbf{t})$$

We have also seen our Likelihood function:

$$p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)$$

Lets ignore the noise variance term for now and see if we can use our likelihood to find the pdf of **w**

Bayes rule: $$\boxed{p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}}$$

Bayes rule

$$p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

Some naming

- **Posterior** distribution = the pdf of **w** *after we have seen* the data

$p(\mathbf{w}|\mathbf{X},\mathbf{t})$ ***We are after this***

- **Prior** distribution = the pdf of **w** *before seeing* the data

$p(\mathbf{w})$ ***What is this?***

- **Likelihood** function

$p(\mathbf{t}|\mathbf{w},\mathbf{X})$ ***We have seen this***

- **Marginal Likelihood** = Normalising constant = Partition function

$p(\mathbf{t}|\mathbf{X})$ ***Ensures*** p(w|X,t) **integrates to 1**

Computing the Posterior

$$\boxed{p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}}$$

- Unfortunately, computing the posterior is usually very hard
- Because the marginal likelihood p(**t**l**X**) is hard to compute:

$$p(\mathbf{t}|\mathbf{X}) = \int p(\mathbf{t}|\mathbf{w},\mathbf{X})p(\mathbf{w})d\mathbf{w}$$

- In some cases (models), we can compute it exactly (This lecture)
- In most cases we can't and we resort to **approximations**:
  - Approximate the posterior distribution with something else
  - Sample from it!
    (Can sample from it even if we cannot compute it!)

Intuitively: Probability as “belief”

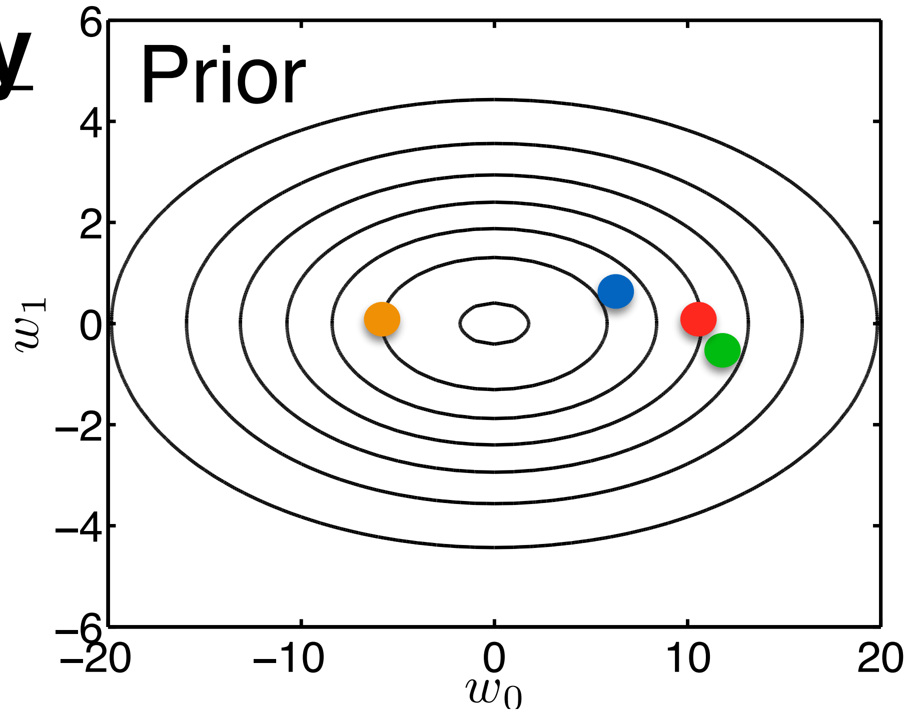
- I have some initial beliefs about something (prior)
- I “see” some evidence (data)
- I update my initial beliefs to “informed” new beliefs (posterior)

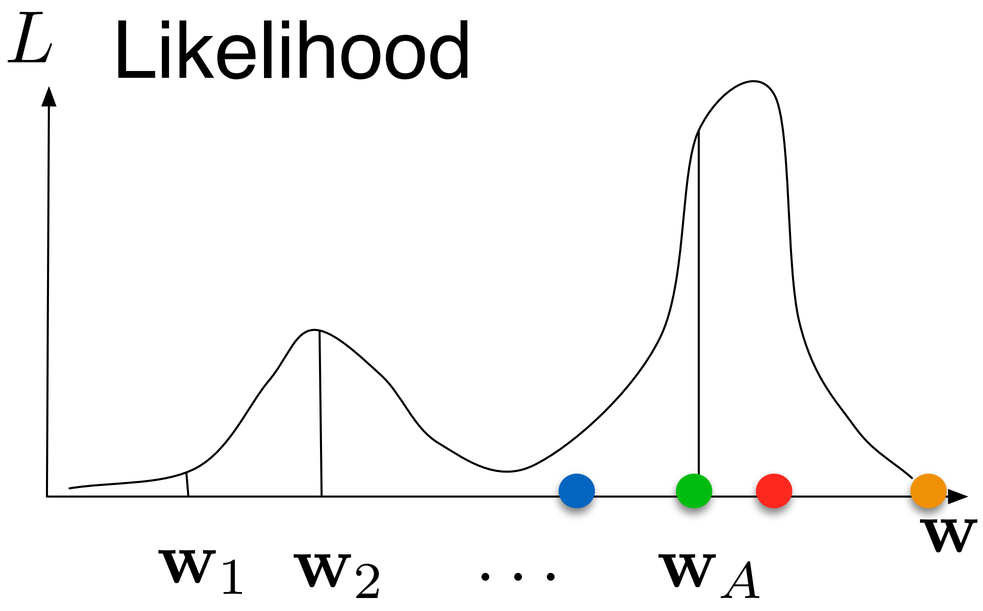
Reasoning under uncertainty

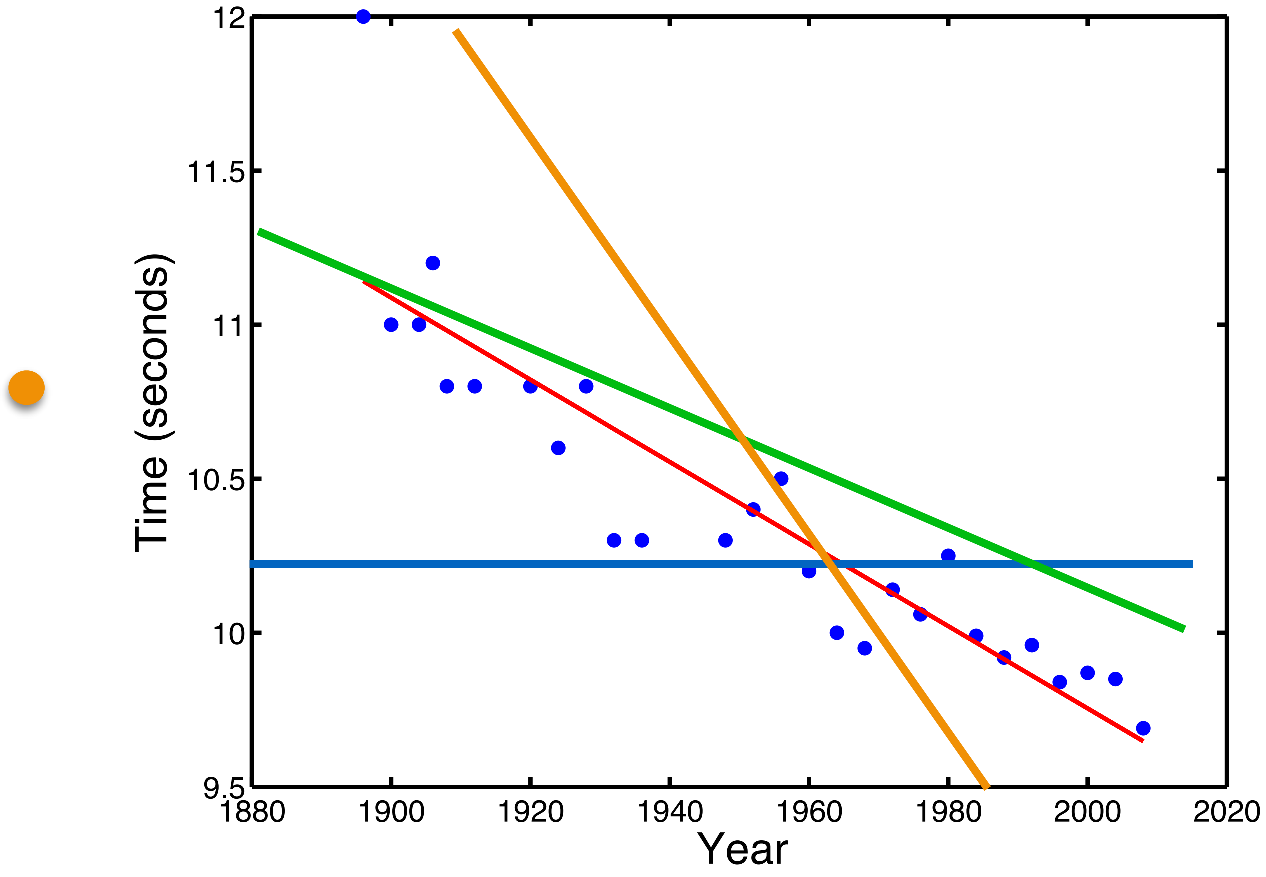
Lets take linear regression as a running example:

- I have some initial belief about the lines/hypothesis/parameters
  - Could be pretty “uninformative” = vague
    - e.g. Not expecting too complex a hyper-plane
    - It implies that I place some initial prior distribution over **w**
- I observe some training data
  - Evidence in favour of some parameter ranges
  - This informs me via the Likelihood function
- I update my beliefs about the lines/hypothesis/parameters
  - Implies that I get an updated distribution over **w**
  - **Remember I am not looking for a single “best” setting**

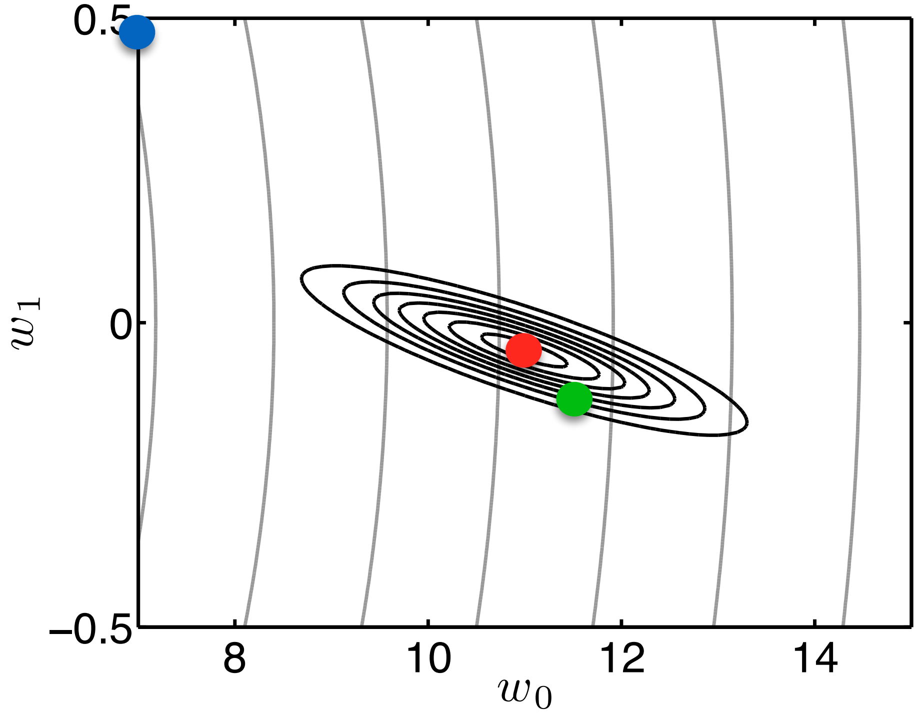
Schematically







Posterior



Probability Theory: The logic of science

E.T. Jaynes, “Probability theory the logic of science”
The only coherent and systematic way to reason under uncertainty

Arguably, you follow similar cognitive principles when reasoning

- Before taking the ML module had some prior beliefs about ML/instructor
- Now you have seen some evidence/data after few lectures
- You have now updated your beliefs about ML/instructor but still have some uncertainty
- You have an informed opinion even under uncertainty

Cool? This ties in to decision theory and risk very nicely

Bayesian Linear regression

Full details and derivation: Rogers & Girolami, Ch. 3

$$p(\mathbf{w}|\mathbf{X},\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

We have a Gaussian Likelihood (previous lectures) due to noise term

Likelihood $$p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n\mathbf{w},\sigma^2)$$

Prior? We don't know much about **w** a priori (+ve and -ve)

Do we have a preference for the parameters?
Think of Lasso, Ridge and L1/L2regularisation…any ideas?

Bayesian Linear regression

Choosing an "easy" prior distribution to work with

Conjugacy (definition)
A prior p(w) is said to be conjugate to a likelihood if it results (Prior x Likelihood) in a posterior of the same type of density as the prior

Examples:

Prior Gaussian & Likelihood Gaussian then Posterior is Gaussian
Prior Beta & Likelihood Binomial then Posterior is Beta
Prior Dirichlet & Likelihood Multinomial then Posterior is Dirichlet
many more… see:
https://en.wikipedia.org/wiki/Conjugate_prior

Bayesian Linear regression

Effects and considerations of Prior distributions

- Prior effect will diminish as more data arrive
- When we don't have much data prior is very important
- Prior knowledge:
  - Data type: real, integer, string, etc..
  - Expert knowledge
  - Occam’s razor (simplicity)
  - Computational considerations (not as important nowadays)
  - If we know nothing - very broad prior, e.g. uniform density

Bayesian Linear regression

Rogers & Girolami, Ch. 3

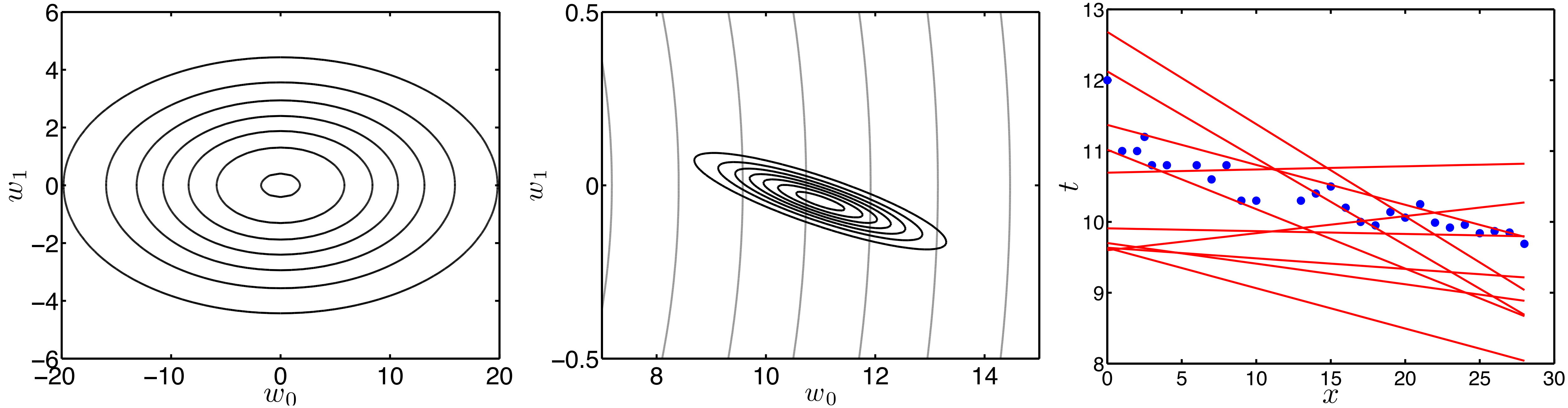
Likelihood (Gaussian)

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2)$$

Prior (Gaussian)

$$p(\mathbf{w}) = \prod_{d=1}^{D} \mathcal{N}(0, \sigma_d^2) = \mathcal{N}(\mathbf{0}, \mathbf{S})$$

Conjugacy therefore posterior $p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2) = \mathcal{N}(\mu, \mathbf{\Sigma})$ is Gaussian



Bayesian Linear regression

We are done! We have the full posterior distribution of the parameters

Back to the start: Overall prediction is this expectation:

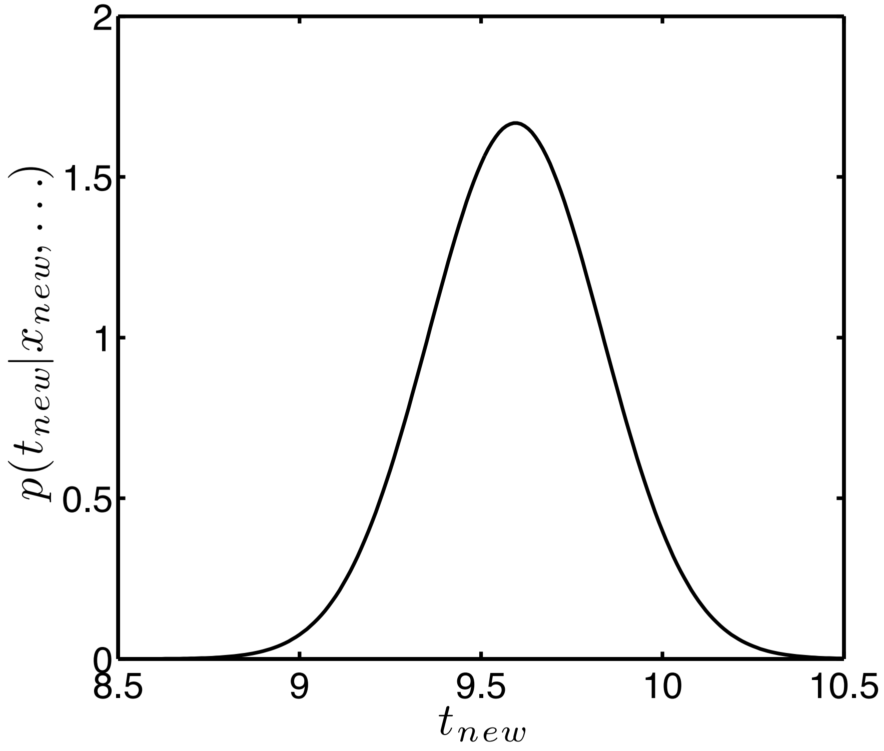
$$\mathbb{E}_{p(\mathbf{w}|\text{stuff})} f(\mathbf{w}) = \int f(\mathbf{w}) p(\mathbf{w}|\text{stuff}) d\mathbf{w}$$

We can now compute such expectations and predict by averaging!

What is f(**w**) if we want a **predictive density** like that below?

$$p(t^*|\mathbf{x}^*, \mathbf{X}, \mathbf{t}, \sigma^2)$$

e.g. for predicting for 2020



Bayesian Linear regression

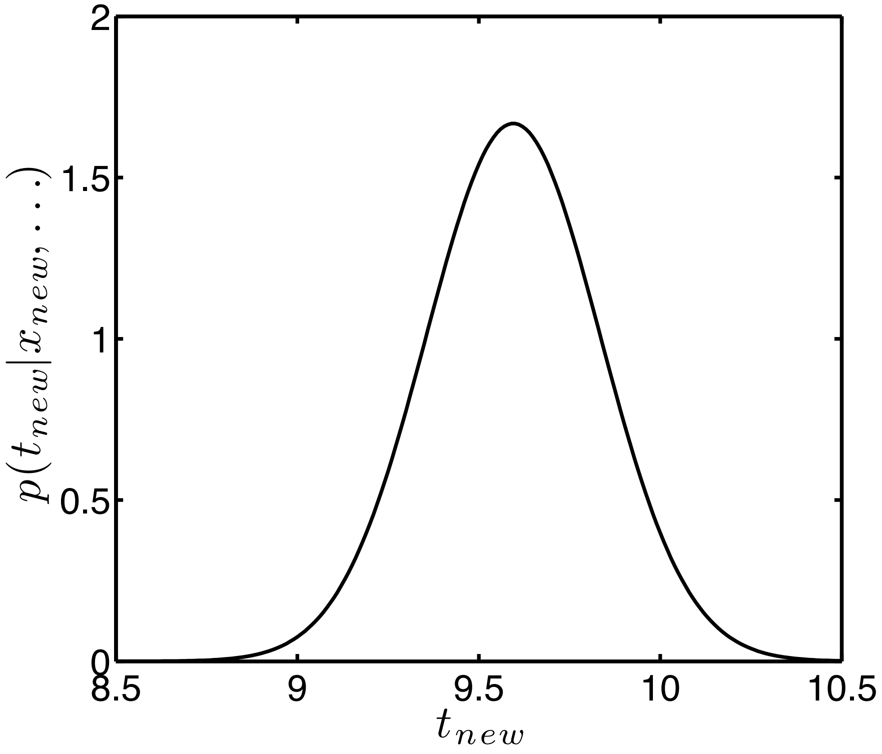
To get a predictive density, f(**w**) is a density (same form as likelihood) itself:

$$f(\mathbf{w}) = \mathcal{N}_{t^*}(\mathbf{x}^*\mathbf{w}, \sigma^2)$$

Which makes our expectation that we need to compute:

$$p(t^*|\mathbf{x}^*, \mathbf{X}, \mathbf{t}, \sigma^2) = \int p(t^*|\mathbf{x}^*, \mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{X}, \mathbf{t})d\mathbf{w}$$

$$p(t^*|\mathbf{x}^*, \mathbf{X}, \mathbf{t}, \sigma^2)$$



Marginal Likelihood: Model selection!

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

So what does this “marginal likelihood” tells us?

$$p(\mathbf{t}|\mathbf{X}) = \int p(\mathbf{t}|\mathbf{w}, \mathbf{X})p(\mathbf{w})d\mathbf{w}$$

It tells us how good our model is **across all possible parameter settings**

So we can use this quantity for **model selection** and comparison!

Advantages and Drawbacks

Advantages:

- Coherent framework to reason under uncertainty
- Outputs are densities and probabilities, e.g. p(t*|x*,X,t) or P(t=+1|…)
- Not sensitive to small parameter perturbations
- Hard to overfit! (prior acts as regulariser)
- Uncertainty quantification via posterior (co)-variance
- Natural decision making mechanisms (Bayesian Decision Theory)
- Natural way to encode prior knowledge

Drawbacks:

- Many solutions - **High computational complexity**
- To the limit of infinite evidence - no uncertainty? big data?
- Hard to compute posterior - resort to approximations
- Choice of prior distribution/beliefs can strongly influence posterior
- Can be easier misused…
- Not as developed/automated tools as optimisation - point estimators