COMP90051 **Statistical Machine Learning**

Semester 2, 2017

Lecturer: Trevor Cohn

17. Bayesian inference; Bayesian regression

THE UNIVERSITY OF MELBOURNE

Training == optimisation (?)

Stages of learning & inference:

- Formulate model

Regression

$$p(y|\mathbf{x}) = \mathrm{sigmoid}(\mathbf{x}'\mathbf{w}) \qquad p(y|\mathbf{x}) = \mathrm{Normal}(\mathbf{x}'\mathbf{w}; \sigma^2)$$

- Fit parameters to data

$$\hat{\mathbf{w}} = \mathrm{argmax}_{\mathbf{w}}\, p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

ditto

- Make prediction

$$p(y_*|\mathbf{x}_*) = \mathrm{sigmoid}(\mathbf{x}_*'\hat{\mathbf{w}}) \qquad E[y_*] = \mathbf{x}_*'\hat{\mathbf{w}}$$

$\hat{w}$ referred to as a '*point estimate*'

Bayesian Alternative

Nothing special about $\widehat{\boldsymbol{w}}$… use more than one value?

- Formulate model

Regression

$$p(y|\mathbf{x}) = \mathrm{sigmoid}(\mathbf{x}'\mathbf{w}) \qquad p(y|\mathbf{x}) = \mathrm{Normal}(\mathbf{x}'\mathbf{w}; \sigma^2)$$

- Consider the space of likely parameters – those that fit the training data well

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

- Make ‘expected’ prediction

$$p(y_*|\mathbf{x}_*) = E_{p(\mathbf{w}|\mathbf{X}_i, \mathbf{y})}\left[\mathrm{sigmoid}(\mathbf{x}'\mathbf{w})\right]$$

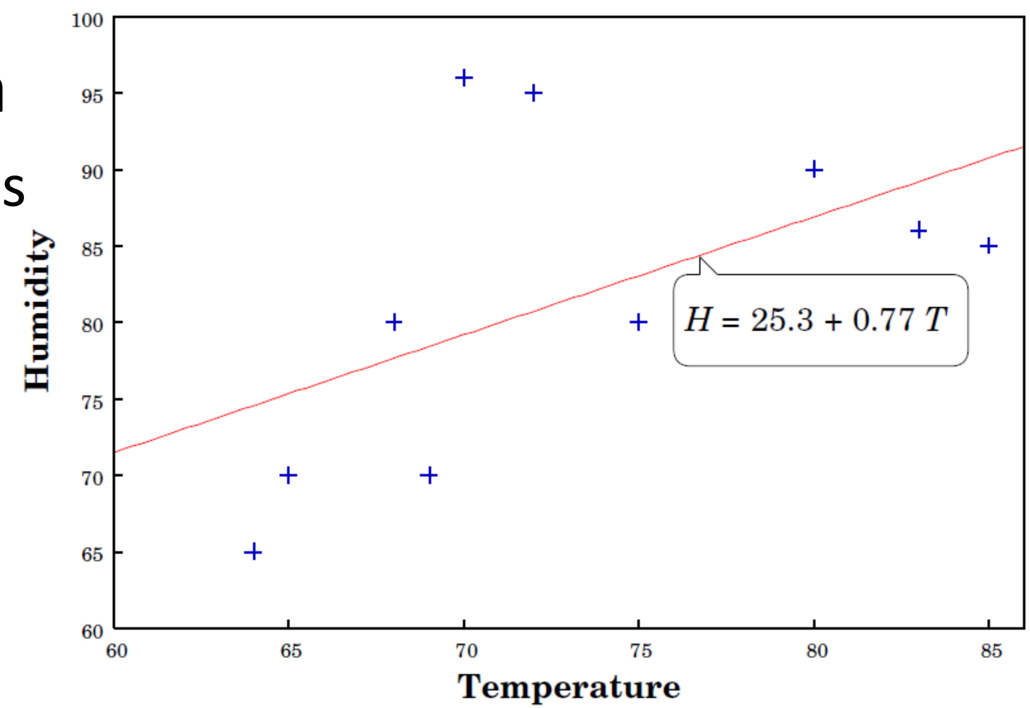
$$p(y_*|\mathbf{x}_*) = E_{p(\mathbf{w}|\mathbf{X}_i, \mathbf{y})}\left[\mathrm{Normal}(\mathbf{x}_*'\mathbf{w}, \sigma^2)\right]$$

Uncertainty

From small training sets, we rarely have complete confidence in any models learned. Can we quantify the uncertainty, and use it in making predictions?

Regression Revisited

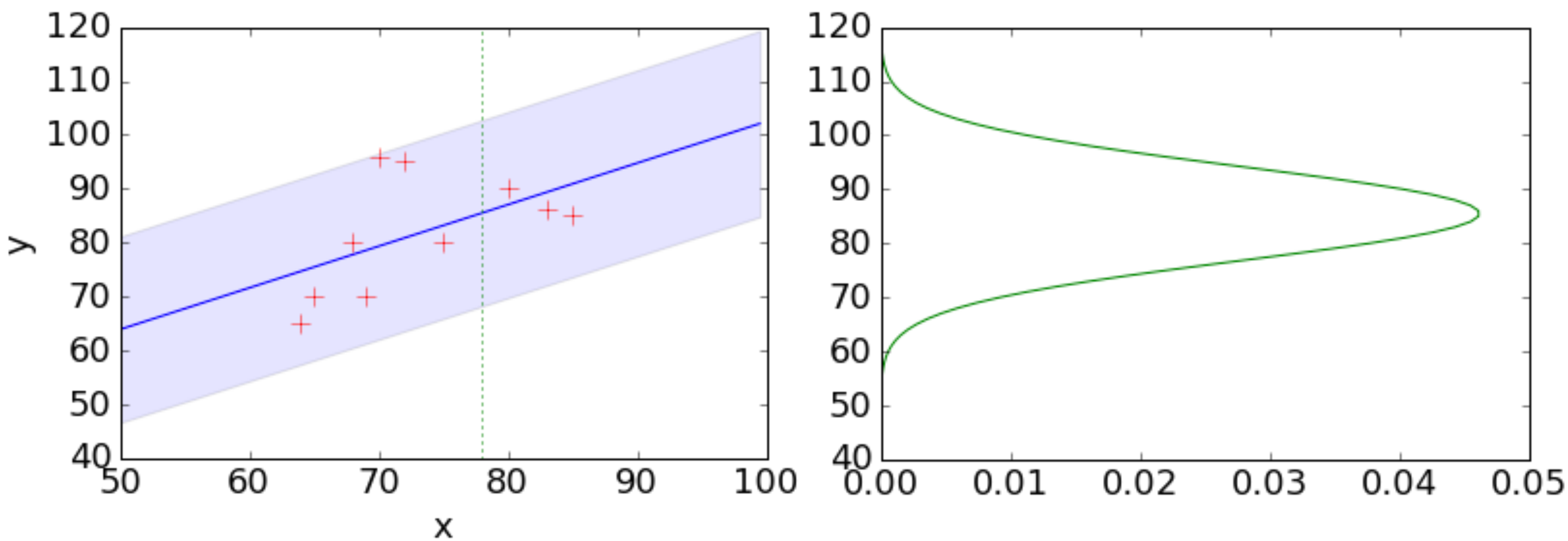
- Learn model from data
  - minimise error residuals by choosing weights $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- But… how confident are we
  - in $\hat{\mathbf{w}}$?
  - in the predictions?



Linear regression: $y = w_0 + w_1 x$
(here y = humidity, x = temperature)

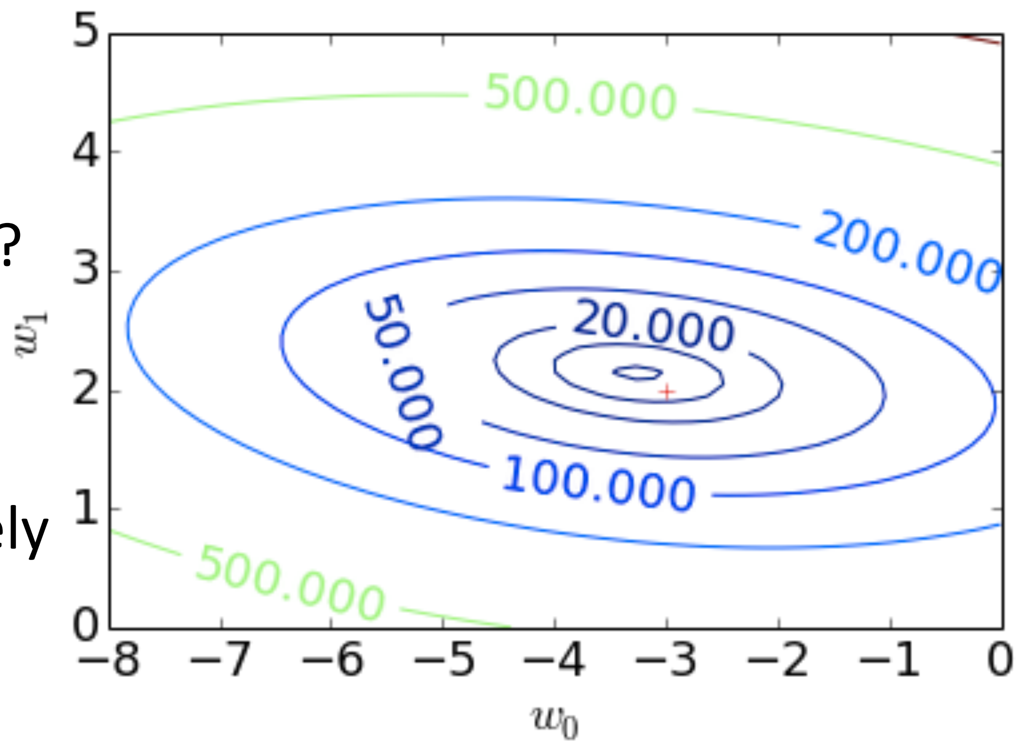
Prediction uncertainty

- Single prediction is of limited use due to uncertainty
  - single number uninformative - may be wildly off
  - might want to formulate decision from prediction, e.g., if Pr(y < 70)



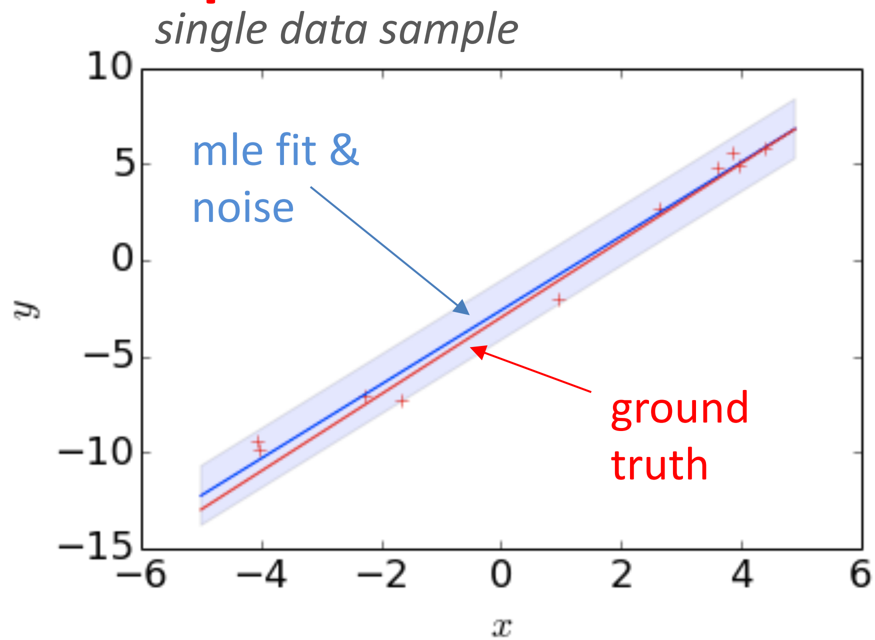
Confidence in MLE point estimate

- What does it mean to minimise objective?
  - … are other nearby solutions similarly good?
- Effect of data
  - lots of data relative to dimensionality, MLE likely to be a good estimate
  - otherwise unreliable
- MAP a *partial* solution, but still reliant on single point

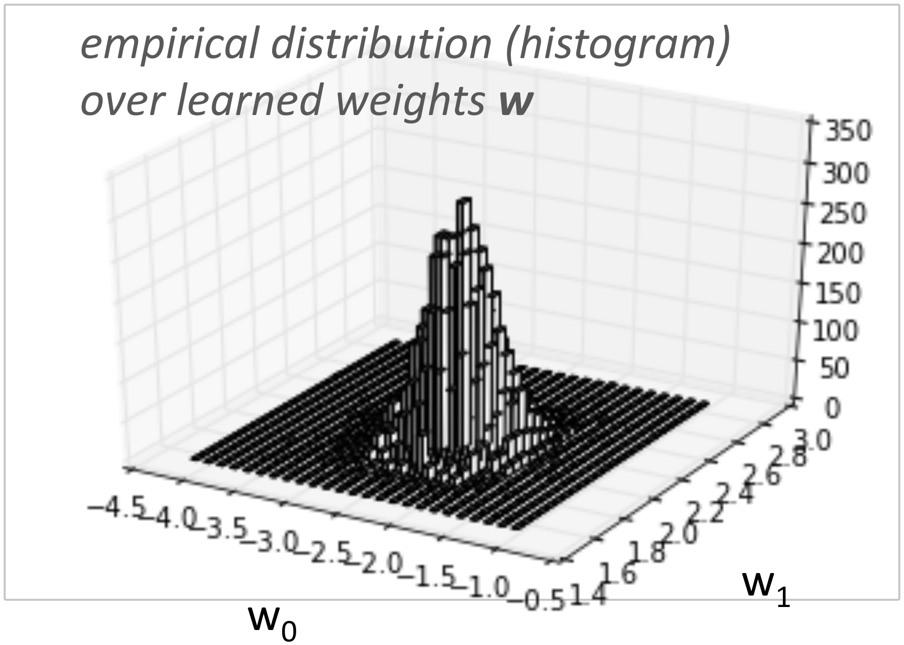


Effect of Training Sample on MLE

- Modelling y = 2x – 3
  - draw 1000s of training sets of 10 instances
  - small added noise



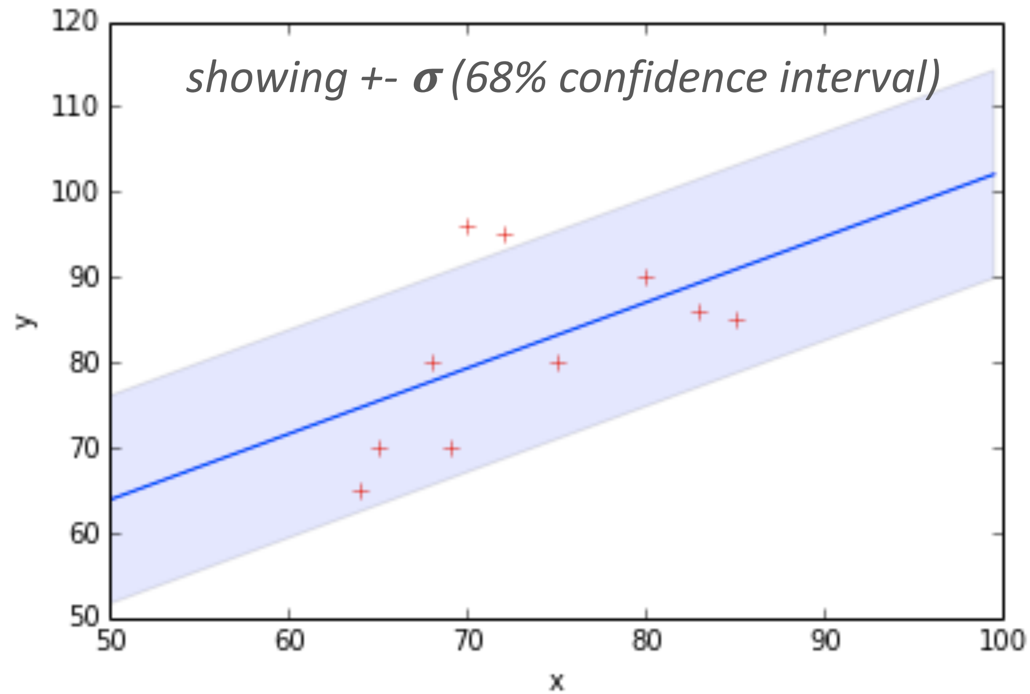
- Fit weights each time using MLE
  - observe variability in weights
  - peak at (2, -3)



Aside: Learning the noise rate

- Can also learn noise parameter, σ^2
  - express NLL as function of σ^2; differentiate; set to 0; solve
  - results in $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\hat{\mathbf{w}})^2$
- Quantifies the quality of the fit
  - allows smarter decision making, e.g., P(y < 60)

N.b., we compute better error bounds later on



Do we trust point estimate $\hat{\mathbf{w}}$?

- How *stable* is learning?
  - $\hat{\mathbf{w}}$ highly sensitive to noise
  - how much uncertainty in parameter estimate?
  - more *informative* if NLL objective highly peaked
- Formalised as *Fisher Information matrix*
  - E[2nd deriv of NLL]

$$\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}$$

  - measures *curvature of objective* about $\hat{\mathbf{w}}$

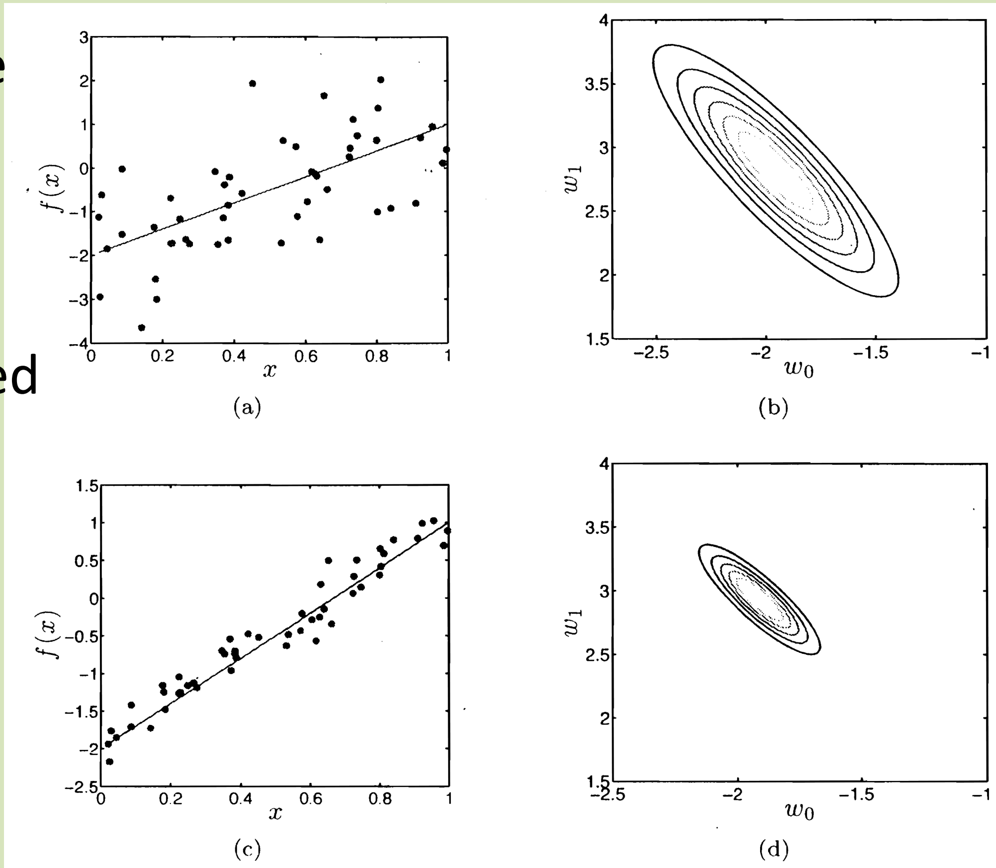


Figure: Rogers and Girolami p81

The Bayesian View

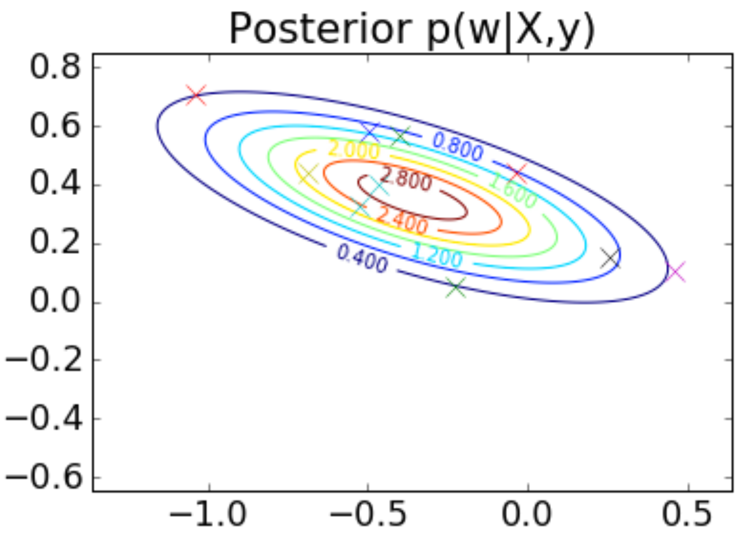
Retain and model all unknowns (e.g., uncertainty over parameters) and use this information when making inferences.

A Bayesian View

- Could we reason over *all* parameters that are consistent with the data?
  - weights with a better fit to the training data should be more probable than others
  - make predictions with all these weights, *scaled by their probability*
- This is the idea underlying **Bayesian** inference

Uncertainty over parameters

- Many reasonable solutions to objective
  - why select just one?
- Reason under ***all*** possible parameter values
  - weighted by their *posterior probability*
- More robust predictions
  - less sensitive to overfitting, particularly with small training sets
  - can give rise to more expressive model class (Bayesian logistic regression becomes non-linear!)



Frequentist vs Bayesian divide

- **Frequentist:** learning using *point estimates*, regularisation, p-values …
  - backed by complex theory relying on strong assumptions
  - mostly simpler algorithms, characterises much practical machine learning research
- **Bayesian:** maintain *uncertainty*, marginalise (sum) out unknowns during inference
  - nicer theory with fewer assumptions
  - often more complex algorithms, but not always
  - when possible, results in more elegant models

Bayesian Regression

Application of Bayesian inference to linear regression, using Normal prior over $\boldsymbol{w}$

Revisiting Linear Regression

I_D = D x D identity matrix

- Recall probabilistic formulation of linear regression

$$y \sim \mathrm{Normal}(\mathbf{x}'\mathbf{w}, \sigma^2)$$

- Motivated by Bayes rule

$$\mathbf{w} \sim \mathrm{Normal}(\mathbf{0}, \gamma^2 \mathbf{I}_D)$$

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

$$\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \max_{\mathbf{w}} p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})$$

- Gives rise to the penalised RSS objective

point estimate taken here, avoids computing marginal likelihood term

Bayesian Linear Regression

- Rewind one step, consider full posterior

$$p(\mathbf{w}|\mathbf{X},\mathbf{y},\sigma^2) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

$$= \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w})}{\int p(\mathbf{y},|\mathbf{X},\mathbf{w},\sigma^2)p(\mathbf{w})d\mathbf{w}}$$

Here we assume noise var. known

- Can we compute the denominator (marginal likelihood or evidence)?
  * if so, we can use the full posterior, not just its mode

Bayesian Linear Regression (cont)

- We have two Normal distributions
  * normal likelihood x normal prior
- Their product is also a Normal distribution
  * ***conjugate prior:*** *when product of likelihood x prior results in the same distribution as the prior*
  * *evidence* can be computed easily using the normalising constant of the Normal distribution

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \text{Normal}(\mathbf{w}|\mathbf{0}, \gamma^2\mathbf{I}_D)\text{Normal}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_N) \\
&\propto \text{Normal}(\mathbf{w}|\mathbf{w}_N, \mathbf{V}_N)
\end{aligned}
$$

closed form solution for posterior!

Bayesian Linear Regression (cont)

$$p(\mathbf{w}|\mathbf{X},\mathbf{y},\sigma^2) \propto \mathrm{Normal}(\mathbf{w}|\mathbf{0},\gamma^2\mathbf{I}_D)\mathrm{Normal}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N)$$
$$\propto \mathrm{Normal}(\mathbf{w}|\mathbf{w}_N,\mathbf{V}_N)$$

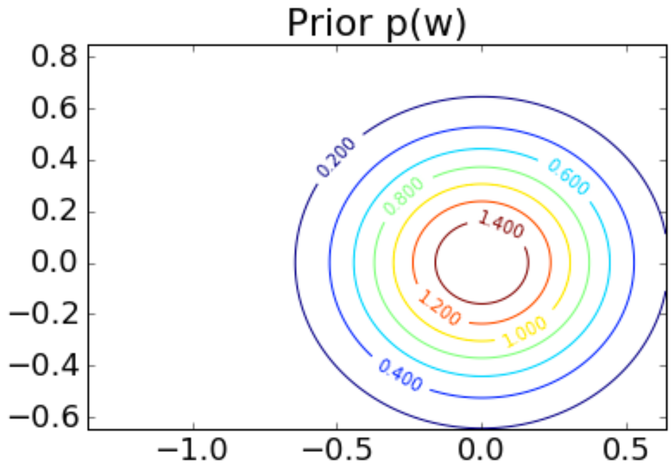
where

$$\mathbf{w}_N = \frac{1}{\sigma^2}\mathbf{V}_N\mathbf{X}'\mathbf{y}$$
$$\mathbf{V}_N = \sigma^2(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{\gamma^2}\mathbf{I}_D)^{-1}$$

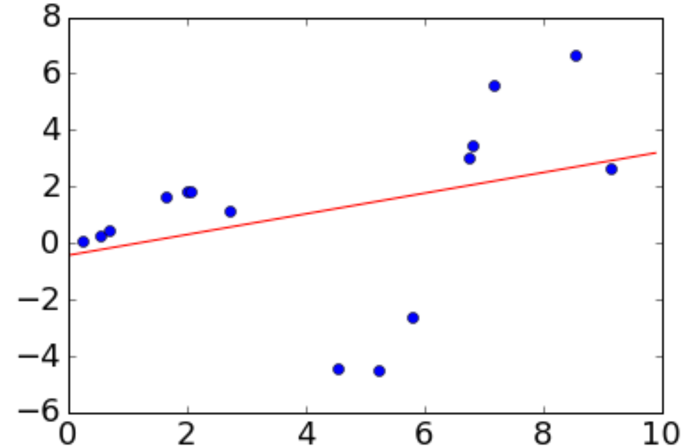
Note that mean (and mode) are the MAP solution from before

Advanced: verify by expressing product of two Normals, gathering exponents together and 'completing the square' to express as squared exponential (i.e., Normal distribution).

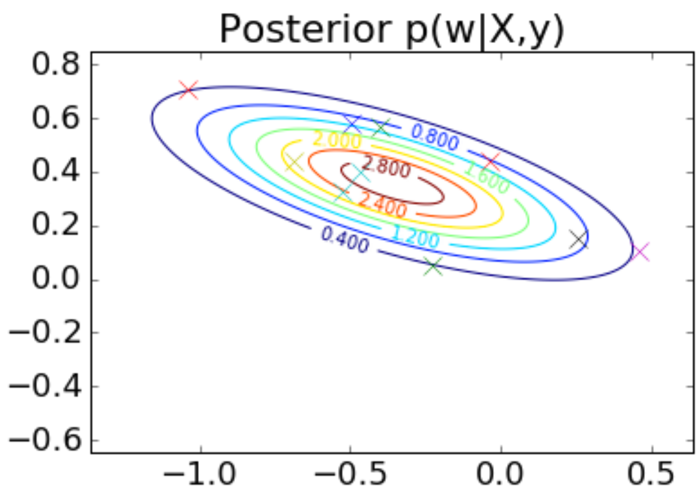
Bayesian Linear Regression example



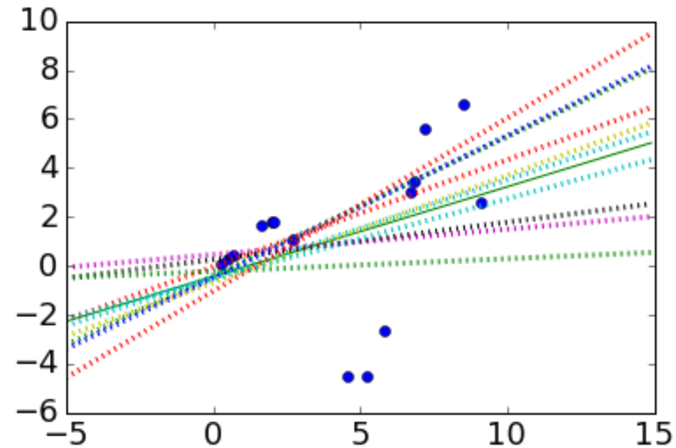
Step 1: select prior, here spherical about **0**



Step 2: observe training data



Step 3: formulate posterior, from prior & likelihood

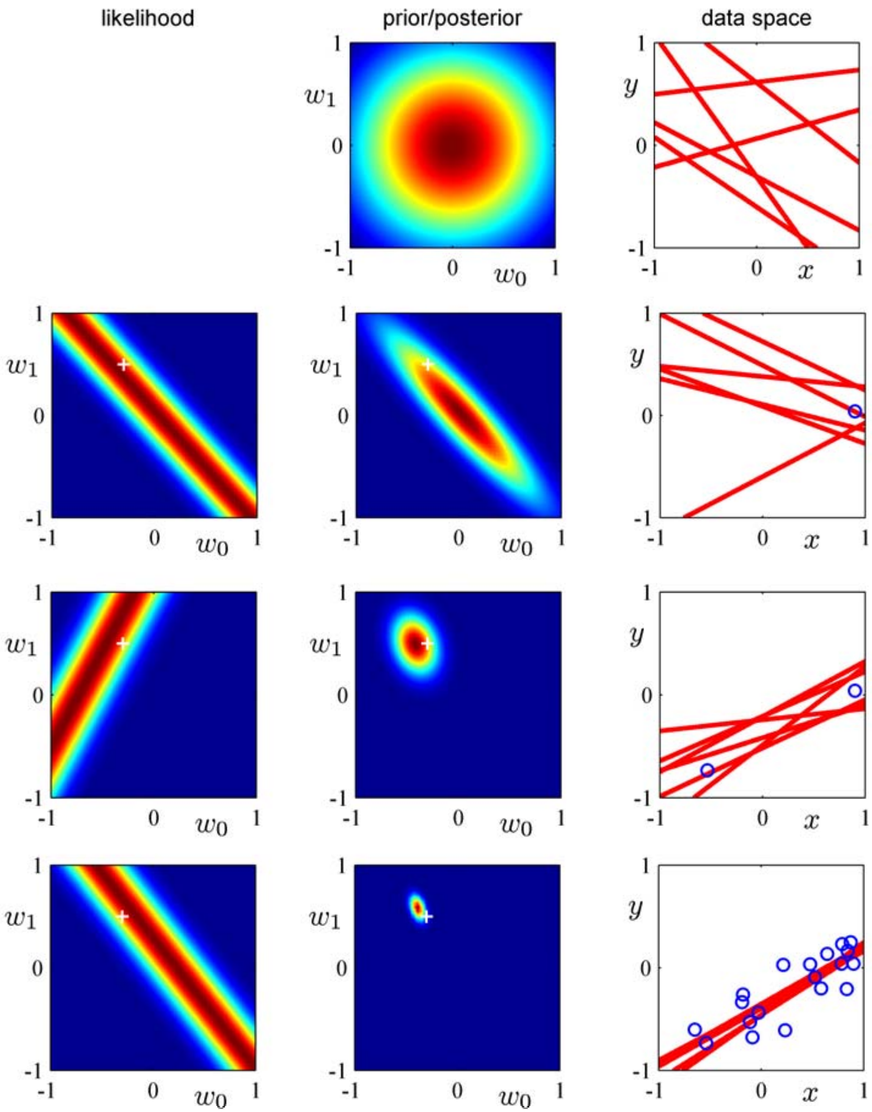


Samples from posterior

Sequential Bayesian Updating

- Can formulate $p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2)$ for given dataset
- What happens as we see more and more data?
  1. Start from prior $p(\mathbf{w})$
  2. See new labelled datapoint
  3. Compute posterior $p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2)$
  4. The ***posterior now takes role of prior***
     & repeat from step 2

Sequential Bayesian Updating



- Initially know little, many regression lines licensed
- Likelihood constrains possible weights such that regression is close to point
- Posterior becomes more refined/peaked as more data introduced
- Approaches a point mass about solution

Bishop Fig 3.7, p155

Summary

- Uncertainty not captured by point estimates (MLE, MAP)
- Bayesian approach preserves uncertainty
  - care about predictions NOT parameters
  - choose prior over parameters, then model posterior
- New concepts:
  - sequential Bayesian updating
  - conjugate prior (Normal-Normal)
- Still to come ... using posterior for Bayesian predictions on test