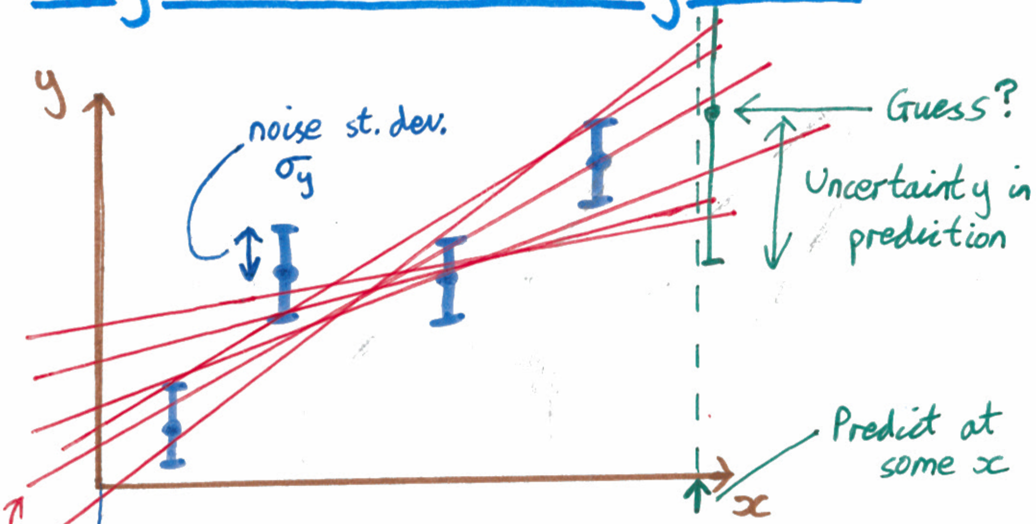
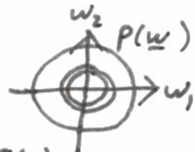
Bayesian Linear Regression

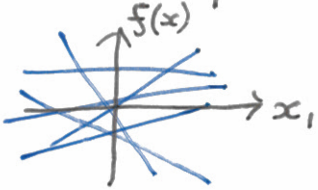


Some different plausible lines $(\underbrace{w_1 x + w_2}_{\underline{w}^T \underline{x},\ \text{with } x_2 = 1})$

Prior (Example) $p(\underline{w}) = N(\underline{w}; \underline{0}, \sigma_w^2 \mathbb{I})$



$\Rightarrow$ Broad range functions plausible before see data



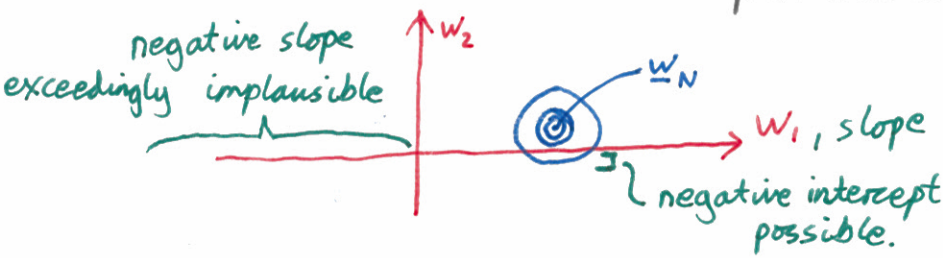
Posterior

"data"

$$p(\underline{w} \mid D) \propto p(\underline{w}) \underbrace{p(\underline{y} \mid X, \underline{w})}_{\text{likelihood}}$$

$$= N(\underline{w}; \underline{w}_N, V_N)$$

(for Gaussian prior and noise)



3 cards

W | B   B | B   W | W

1   2   3

Picked a card uniformly at random

The way up is random

Observed $x_1 = B$ (black face)

Q) $P(x_2 = W \mid x_1 = B)$?

$\uparrow$ other side of same card

A) $1/3$ B) $1/2$ C) $2/3$ D) Other

Z) Don't know?

What not to do

$$P(x_2 = W \mid x_1 = B) = \frac{\boxed{P(x_1 = B \mid x_2 = W)}\, P(x_2 = W)}{P(x_1 = B)}$$

Don't know

First step: write down model

Picked a card, c:

$$P(c) = \begin{cases} 1/3 & c=1 \quad W/B \\ 1/3 & c=2 \quad B/B \\ 1/3 & c=3 \quad W/W \end{cases}$$

Observed face 1:

$$P(x_1 = B \mid c) = \begin{cases} 1/2 & c=1 \\ 1 & c=2 \\ 0 & c=3 \end{cases}$$

Inference

$$P(c \mid x_1 = B) \propto p(x_1 = B \mid c)\, p(c)$$

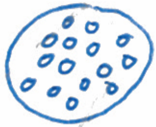
$$\propto \begin{cases} 1/2 & c=1 \\ 1 & c=2 \\ 0 & c=3 \end{cases}$$

$$= \begin{cases} 1/3 & c=1 \\ 2/3 & c=2 \end{cases}$$

Aside : another example

6 sided Dice

10-side Dice



100-sided Dice

Pick random Dice

Roll it → get an 8

Making a Prediction

$$P(x_2 = W \mid x_1 = B) = \sum_{c \in \{1,2,3\}} P(x_2 = W, c \mid x_1 = B)$$

(Sum Rule)

$$= \sum_c P(x_2 = W \mid c, x_1 = B)\, P(c \mid x_1 = B)$$

(Product Rule)

$$= 1/3$$

Prediction for Linear Regression

What's

Training data $\{X, y\}$

$$p(y \mid \underline{x}, D) = \int p(y, \underline{w} \mid \underline{x}, D)\, d\underline{w}$$

(sum rule)

$$= \int \underbrace{p(y \mid \underline{w}, \underline{x}, \cancel{D})}_{N(y;\, \underline{w}^T \underline{x},\, \sigma_y^2)}\, p(\underline{w} \mid D, \cancel{\underline{x}})\, d\underline{w}$$

(Product Rule)

noise var.

For this standard model.

$$p(y \mid \underline{x}, D) = \int \underbrace{p(y, \underline{w} \mid \underline{x}, D)}_{\text{Joint Gaussian on } y,\ \underline{w}} \, d\underline{w}$$

... several lines later ...

$$= \int N\left(\begin{bmatrix} \underline{w} \\ y \end{bmatrix} ; \begin{bmatrix} \underline{w}_N \\ m \end{bmatrix}, \begin{bmatrix} V_N & \Sigma_{\underline{w},y} \\ \Sigma_{y,\underline{w}} & r^2 \end{bmatrix} \right) d$$

$$= N(y ; \underbrace{m, r^2}_{\text{Lots of work to find these.}})$$

Yuck!

Instead...

$$y = f(\underline{x}) + \nu, \qquad \nu \sim N(0, \sigma_y^2)$$

What do we believe about the function:

$$f = \underline{w}^T \underline{x} = \underline{x}^T \underline{w}$$

$\left\{ \begin{array}{l} \text{Beliefs about } \underline{w} \\ p(\underline{w} \mid D) = N(\underline{w}; \underline{w}_N, V_N) \end{array} \right.$

Beliefs about f:

$$p(f \mid \underline{x}, D) = N(f; \underline{x}^T \underline{w}_N, \underline{x}^T V_N \underline{x})$$

Beliefs about prediction, y:

$$p(y \mid \underline{x}, D) = N(y; \underline{x}^T \underline{w}_N, \underline{x}^T V_N \underline{x} + \sigma_y^2)$$