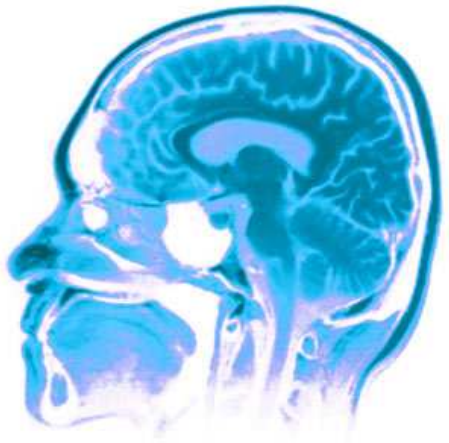
CPSC340



Probabilistic linear prediction

Nando de Freitas
October, 2012
University of British Columbia

Outline of the lecture

In this lecture, we formulate the problem of linear prediction using probabilities. In doing so, we find out that the maximum likelihood estimate coincides with the least squares estimate. The goal of the lecture is for you to learn:

- ❑ Multivariate Gaussian distributions
- ❑ How to formulate the likelihood for linear regression
- ❑ Computing the maximum likelihood estimates for linear regression.

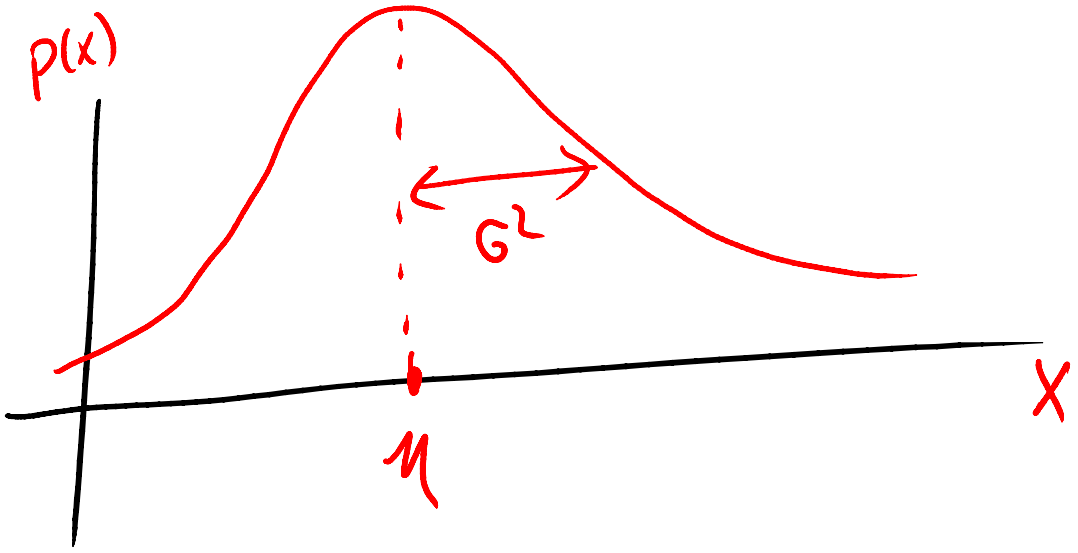
Univariate Gaussian distribution

The probability density function (pdf) of a Gaussian distribution is given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

$$x \sim \mathcal{N}(\mu, \sigma^2)$$

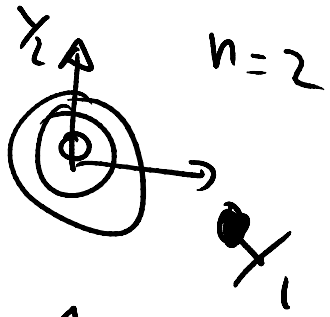
where μ is the mean or center of mass and σ^2 is the variance.



Multivariate Gaussian distribution

Let $\mathbf{y} \in \mathbb{R}^{n \times 1}$, then pdf of an n-dimensional Gaussian is given by

$$p(\mathbf{y}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})},$$



where

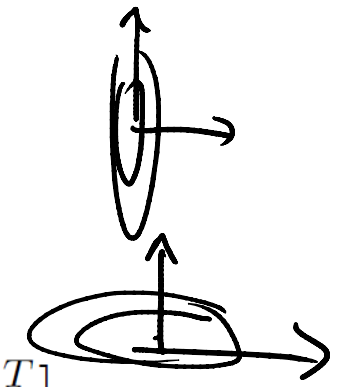
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \mathbb{E}(y_1) \\ \vdots \\ \mathbb{E}(y_n) \end{pmatrix}$$

$\Sigma = \Sigma^T$

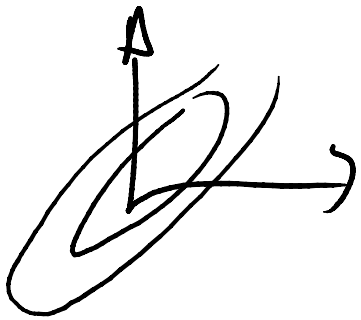
and

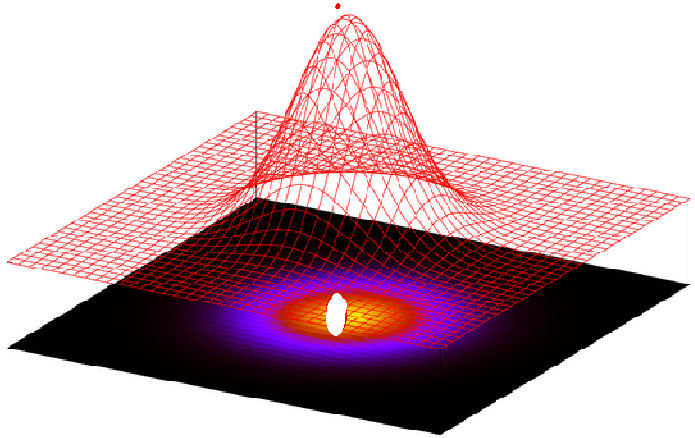
$\sigma_{12} = \sigma_{21}$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} \cdots \sigma_{1n} \\ \cdots \\ \sigma_{n1} \cdots \sigma_{nn} \end{pmatrix} = \mathbb{E}[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T]$$



$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$



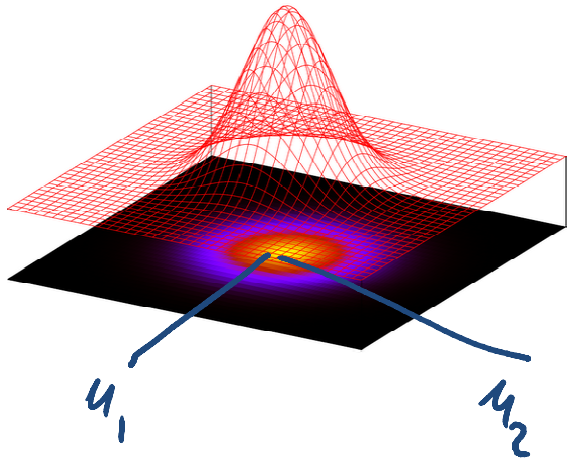


Bivariate Gaussian distribution example

Assume we have two univariate Gaussian variables (independent)

$$x_1 = \mathcal{N}(\mu_1, \sigma^2) \quad \textit{and} \quad x_2 = \mathcal{N}(\mu_2, \sigma^2)$$

Their joint distribution $\mathbf{p}(x_1, x_2)$ *is:*



$$P(x_1, x_2) = P(x_1)\,P(x_2)$$

$$= (2\pi\sigma^2)^{-1/2}(2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x-\mu_1)^2} e^{-\frac{1}{2\sigma^2}(x-\mu_2)^2}$$

$e^A e^B = e^{A+B}$

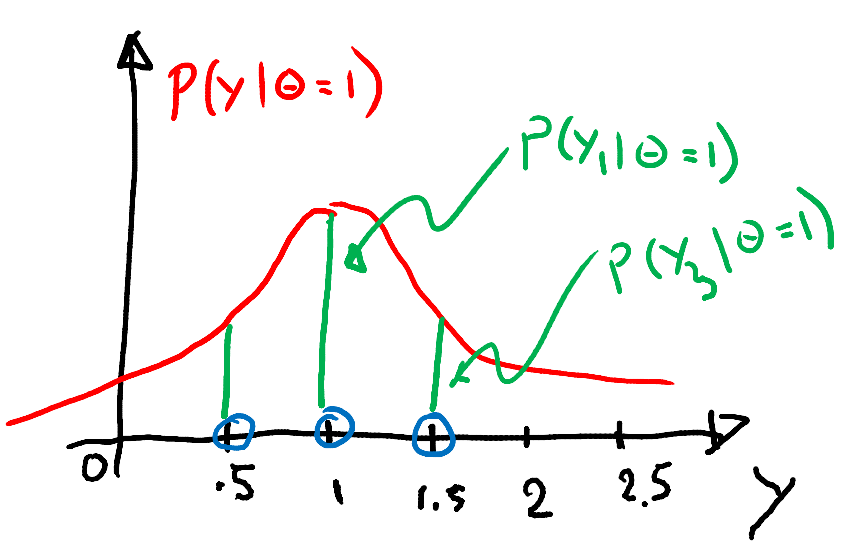
$$= (2\pi\sigma^2)^{-2/2} e^{-\frac{1}{2}\left[(x-\mu_1)(\sigma^2)^{-1}(x-\mu_1) + (x-\mu_2)(\sigma^2)^{-1}(x-\mu_2)\right]}$$

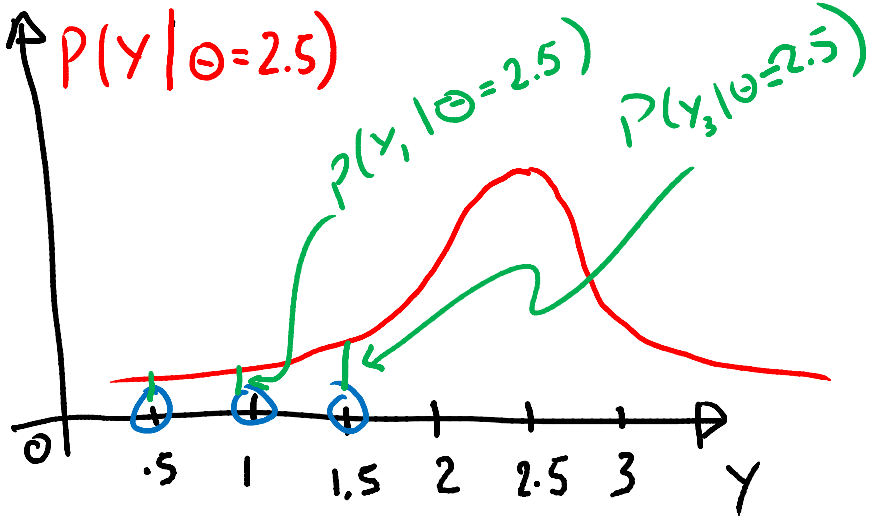
$$= (2\pi\sigma^2)^{-2/2} e^{-\frac{1}{2}\left\{\left[(x-\mu_1)\ (x-\mu_2)\right]\begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ x-\mu_2\end{bmatrix}\right\}}$$

Let us assume that we have $n=3$ *data points* $y_1 = 1, y_2 = 0.5, y_3 = 1.5,$ *which we assume independent and Gaussian distributed with unknown mean* θ *and variance* 1*. That is,*

$$y_i = \mathcal{N}(\theta, 1) = \theta + \mathcal{N}(0, 1)$$

with likelihood $P(y_1 y_2 y_3 | \theta) = P(y_1 | \theta) P(y_1 | \theta) P(y_3 | \theta)$ *. Consider the following two cases. Clearly the one on the left has higher likelihood (higher green bars). Finding the* θ *that maximizes the likelihood is equivalent to moving the Gaussian to the left and right until the product of 3 green bars (likelihood) is maximized.*





The likelihood for linear regression

Let us assume that each label y_i *is Gaussian distributed with mean* $x_i^T\theta$ *and variance* σ^2*, which in short we write as:*

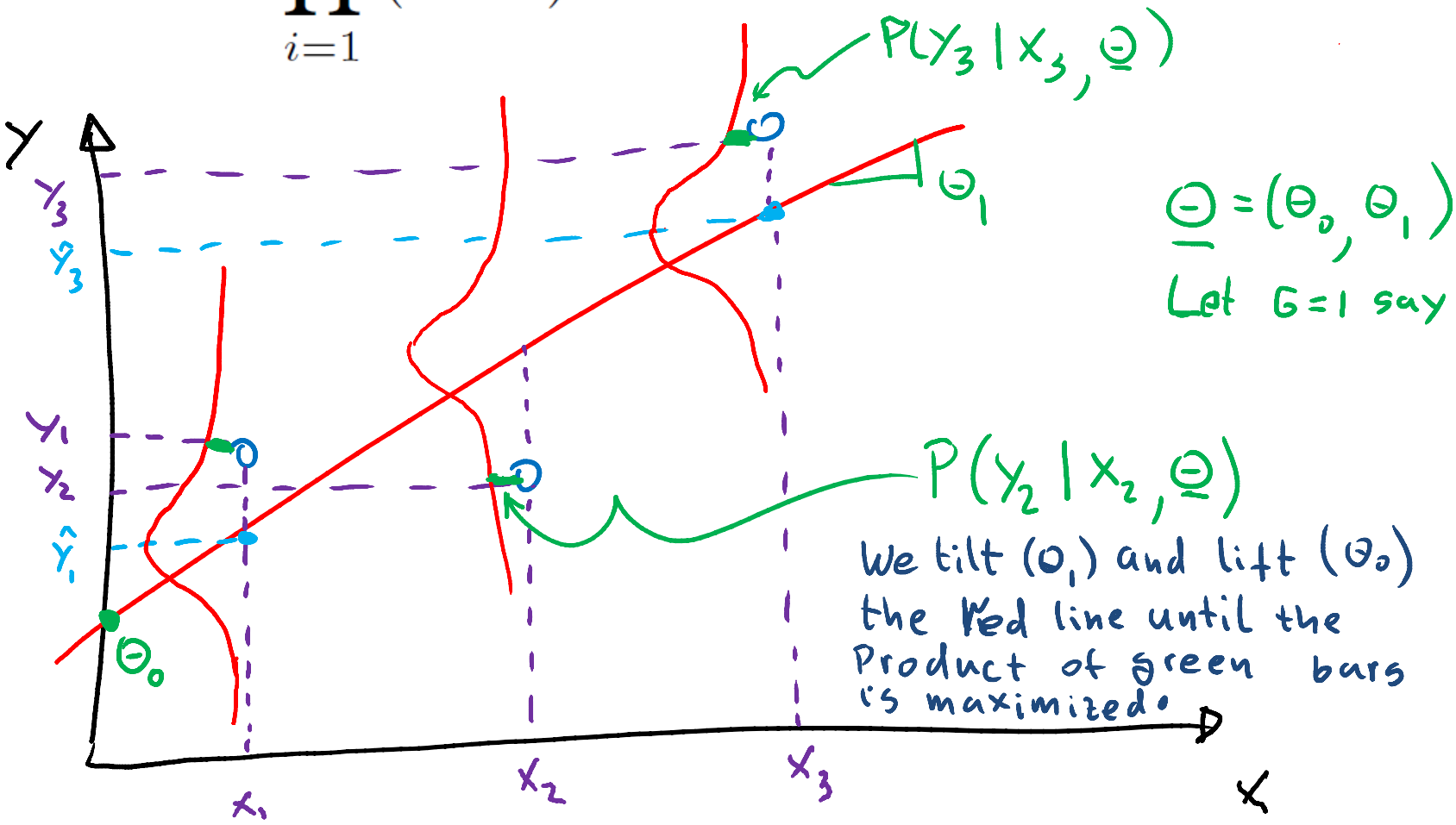
$$y_i = \mathcal{N}(x_i^T\theta, \sigma^2) = x_i^T\theta + \mathcal{N}(0, \sigma^2)$$

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) &= \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \sigma). \\ &= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2} \\ &= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2} \\ &= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})} \end{aligned}$$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2$$

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2}$$

$$= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2}$$



Maximum likelihood

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is obtained by taking the derivative of the log-likelihood, $\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma)$. The goal is to maximize the likelihood of seeing the training data $\mathbf{y}$ by modifying the parameters $(\boldsymbol{\theta}, \sigma)$.

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})}$$

$$\ell(\theta, \sigma^2) = \log P(y \mid X, \theta, \sigma)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta)$$

The ML estimate of θ is: (assume σ known)

$$\frac{\partial}{\partial\theta}\ell(\theta,\sigma^2) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\theta}\left[y-x\theta\right]^T\left[y-x\theta\right]$$

$$= -\frac{1}{2\sigma^2}\frac{\partial}{\partial\theta}\left(y^Ty - 2y^Tx\theta + 2x^Tx\theta\right)$$

(just as before)

$$= -\frac{1}{\sigma^2}\left(-y^Tx\theta + x^Tx\theta\right)$$

Equating to zero, yields

$$x^Tx\theta = x^Ty$$

$$\therefore \quad \hat{\theta}_{ML} = (x^Tx)^{-1}x^Ty$$

(same as Least Squares)

The ML estimate of σ is: (assume θ known)

$$\frac{\partial}{\partial \sigma} \ell(\theta, \sigma^2) = \frac{\partial}{\partial \sigma}\left(-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y-x\theta)^T(y-x\theta)\right)$$

$$= \left(-\frac{n}{2}\right)\frac{2\sigma}{\sigma^2} - \frac{1}{2}(-2)\frac{1}{\sigma^3}(y-x\theta)^T(y-x\theta)$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}(y-x\theta)^T(y-x\theta)$$

Equating to zero, yields

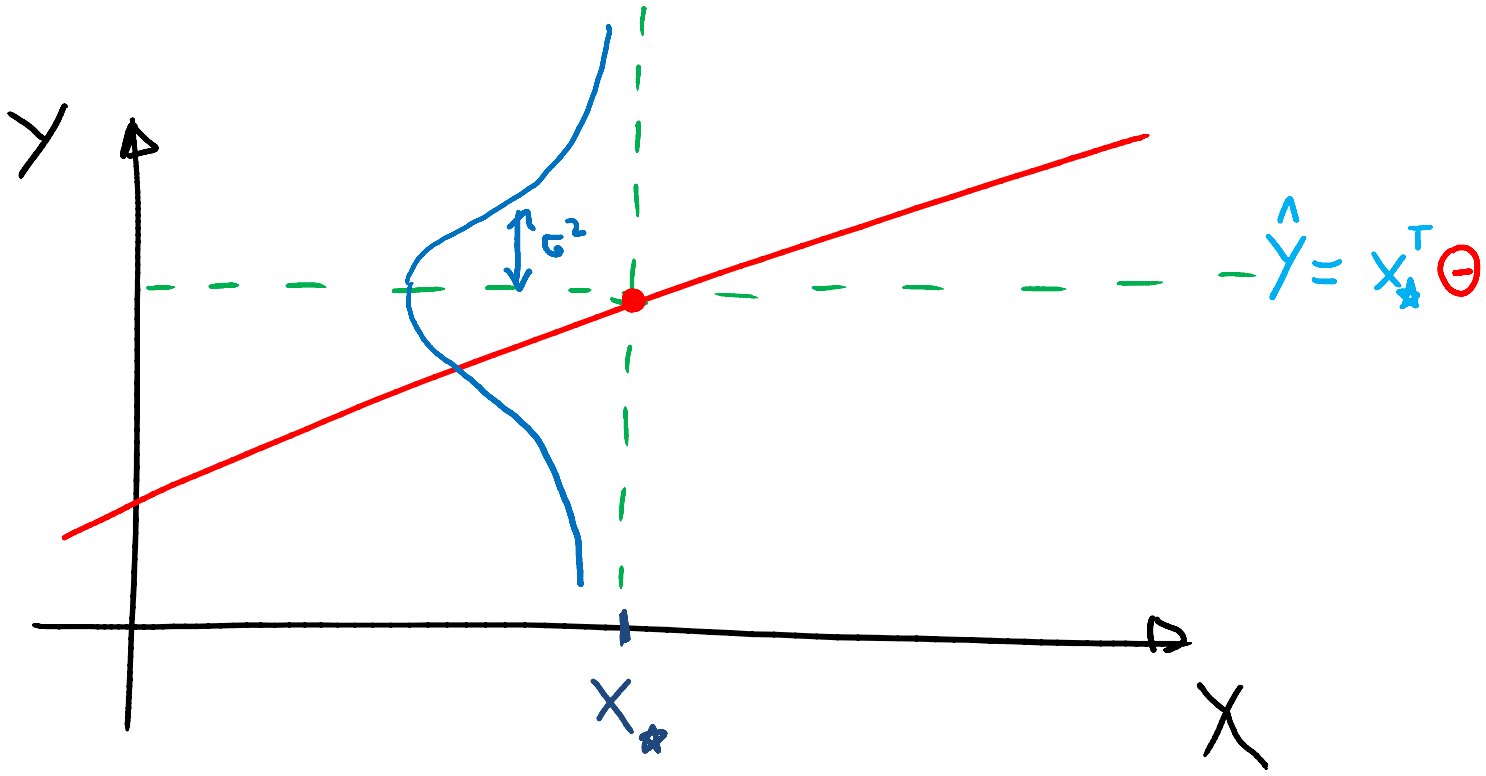
$$\sigma^2 = \frac{1}{n}(y-x\theta)^T(y-x\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\theta)^2$$

as expected, it is the estimator of variance!

Making predictions

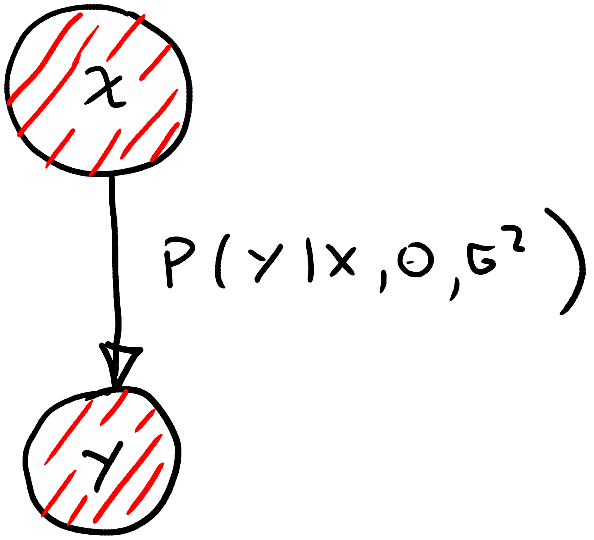
The ML plugin prediction, given the training data $D=(X, y)$*, for a new input* x_* *and known* σ^2 *is given by:*

$$P(y|x_*, D, \sigma^2) = \mathcal{N}(y|x_*^T \theta_{ML}, \sigma^2)$$

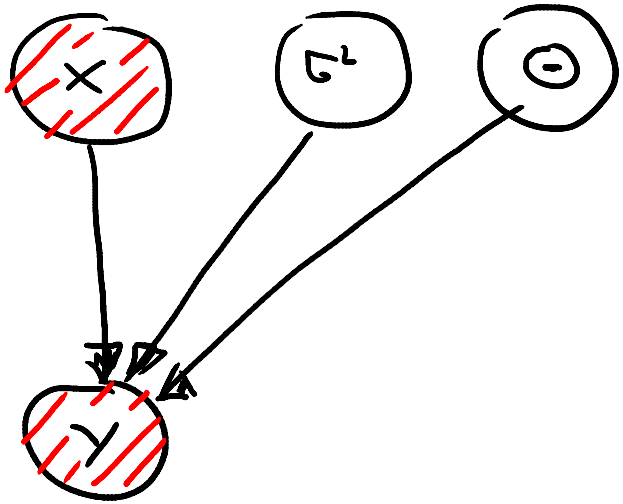


The DAG for linear regression

Frequentist model



Bayesian model



$P(Y|X,\sigma^2,\theta)\,P(\theta)\,P(\sigma^2)$

Next lecture

In the next lecture, we introduce ridge regression and the Bayesian learning approach for linear predictive models.