Maximum Likelihood

Nando de Freitas

UNIVERSITY OF OXFORD

Outline of the lecture

In this lecture, we formulate the problem of linear prediction using probabilities. We also introduce the maximum likelihood estimate and show that it coincides with the least squares estimate. The goal of the lecture is for you to learn:

❑ Gaussian distributions
❑ How to formulate the likelihood for linear regression
❑ Computing the maximum likelihood estimates for linear regression.
❑ Entropy and its relation to loss, probability and learning.

Univariate Gaussian distribution

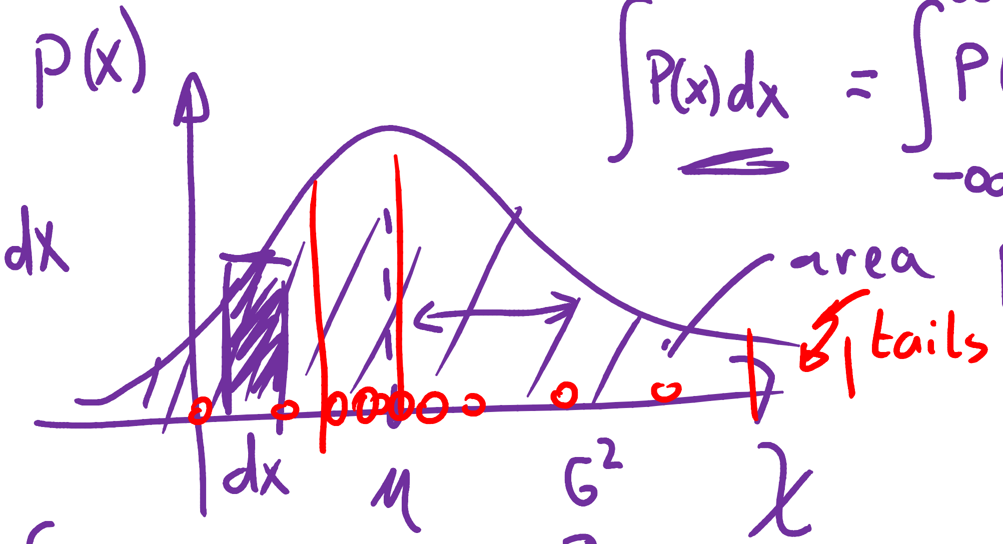
The probability density function (pdf) of a Gaussian distribution is given by

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.} \qquad x \sim \mathcal{N}(\mu, \sigma^2)$$

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

$$\underline{\int P(x)dx} = \int_{-\infty}^{\infty} P(x)\,dx = 1$$

$$P(dx) = P(x)\,dx$$



area probability

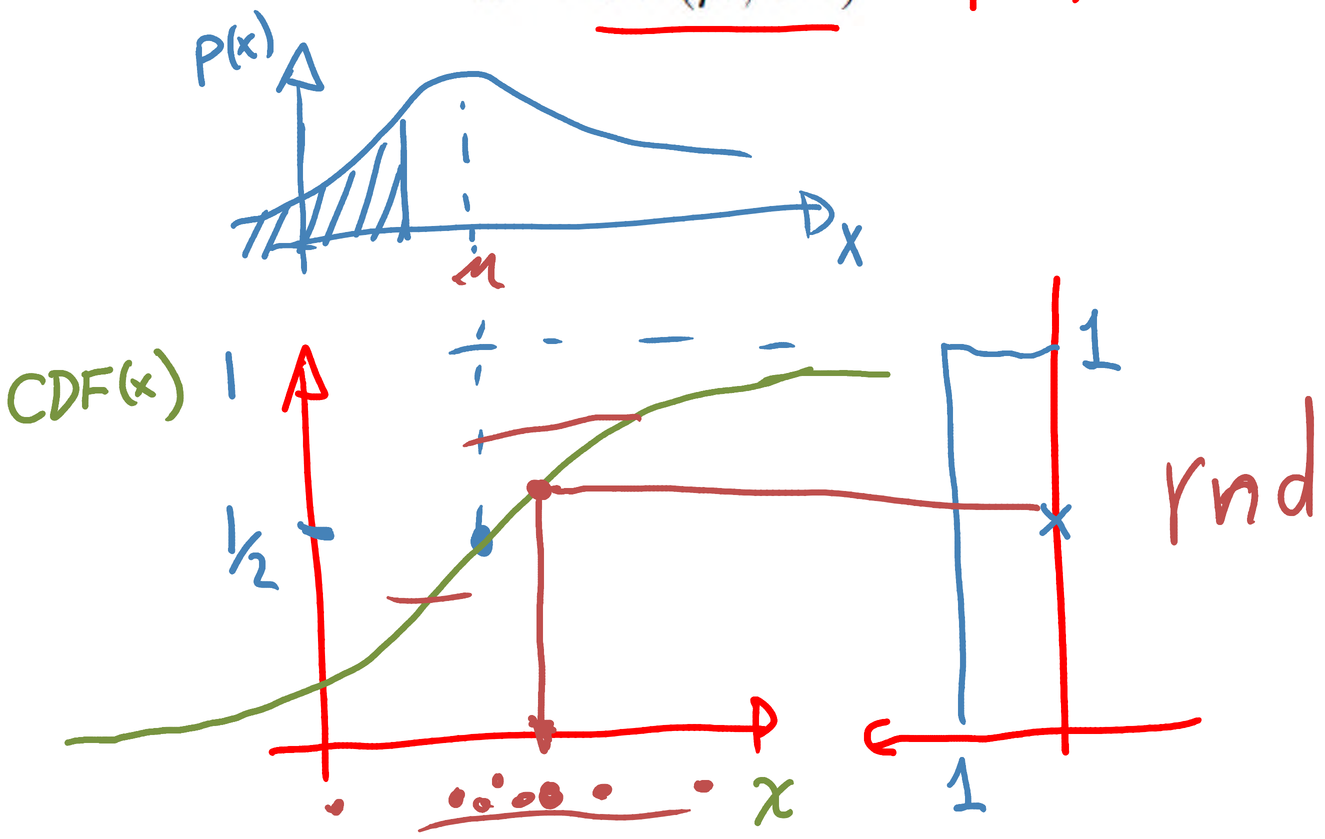
$$\int P(x)\,dx = \int P(dx)$$

parameters or sufficient statistics

Sampling from a Gaussian distribution

simulate / imagine / hallucinate / draw

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



Sample $x^{(i)} \sim p(x)$

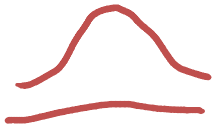
Covariance, correlation and multivariate Gaussians

The **covariance** between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

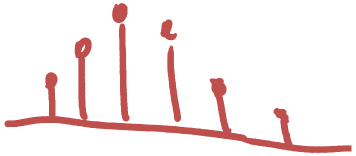
$$\operatorname{cov}[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y]$$

Expectation

$$\mathbb{E}(x) = \int x\, p(x)\, dx = \mu \approx \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$



$p(x) \longleftrightarrow$ histogram estimator



$$\int f(x)\,\delta(a)\,dx = f(a)$$

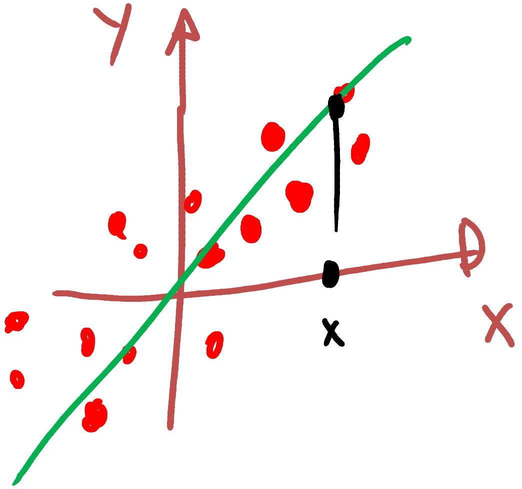
$$\mathbb{E}(x) \approx \int x\, \frac{1}{N}\sum_{i=1}^{N} \delta(x^{(i)})\, dx = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

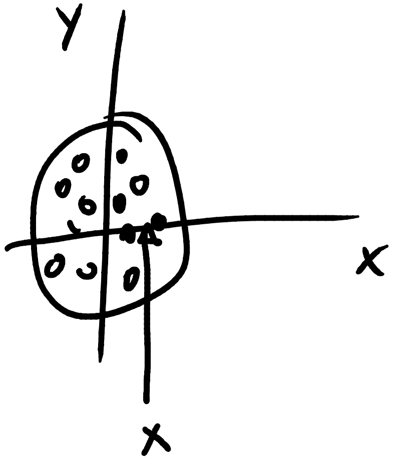
Covariance, correlation and multivariate Gaussians

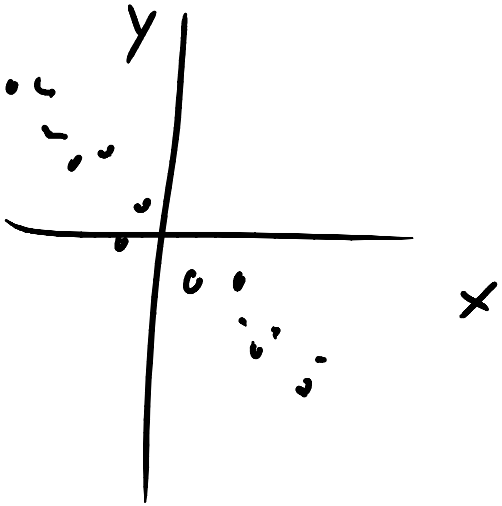
The **covariance** between two rv's X and Y measures the degree to which X and Y are (linearly) related. Covariance is defined as

$$\underline{\operatorname{cov}[X,Y]} \triangleq \mathbb{E}[(X-\underbrace{\mathbb{E}[X]}_{\mu_X})(Y-\underbrace{\mathbb{E}[Y]}_{\mu_Y})] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

$$\text{mean} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$







Covariance, correlation and multivariate Gaussians

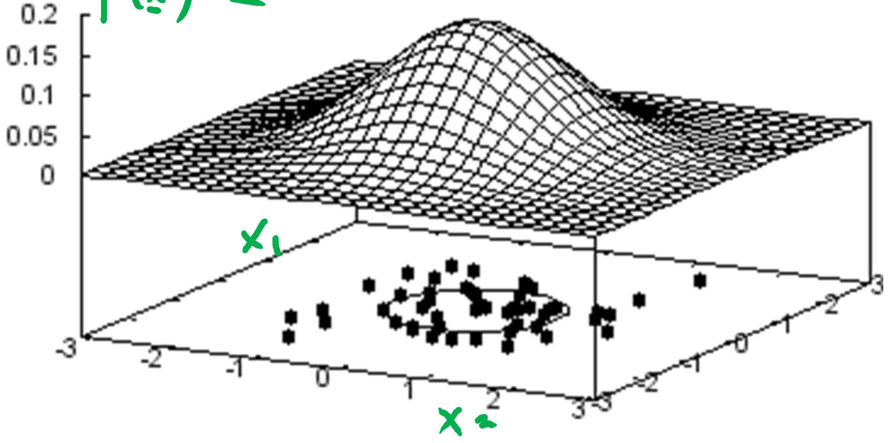
If $\mathbf{x}$ is a $\underline{d}$-dimensional random vector, its **covariance matrix** is defined to be the following symmetric, positive definite matrix:

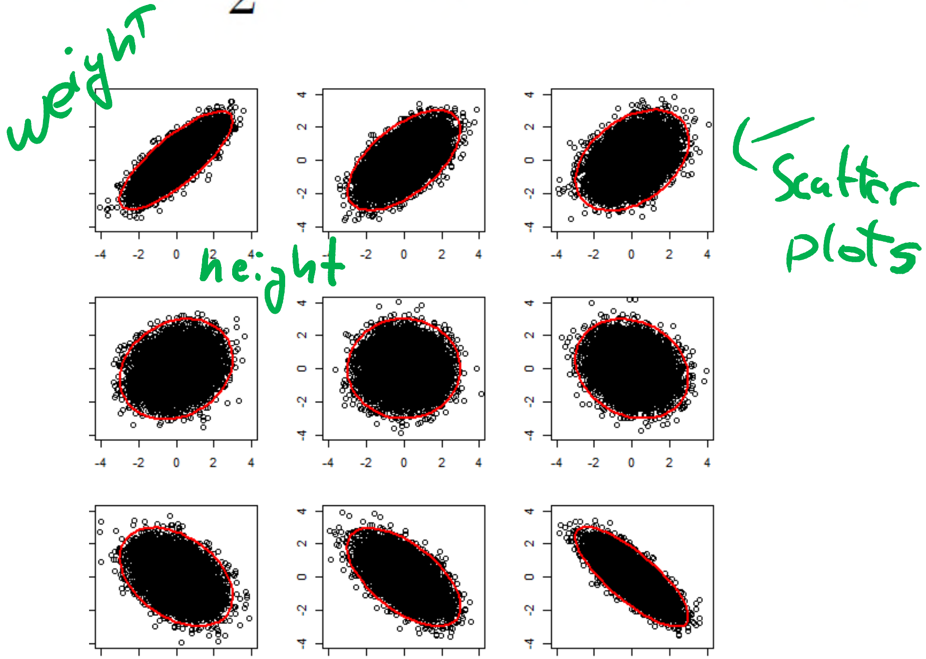
$$\mathrm{cov}\left[\mathbf{x}\right] \triangleq \mathbb{E}\left[(\overset{\downarrow}{\mathbf{x}} - \mathbb{E}\left[\mathbf{x}\right])(\overset{\downarrow}{\mathbf{x}} - \mathbb{E}\left[\mathbf{x}\right])^T\right] = \begin{pmatrix} \underline{\mathrm{var}\left[X_1\right]} & \underline{\mathrm{cov}\left[X_1, X_2\right]} & \cdots & \mathrm{cov}\left[X_1, X_d\right] \\ \mathrm{cov}\left[X_2, X_1\right] & \mathrm{var}\left[X_2\right] & \cdots & \mathrm{cov}\left[X_2, X_d\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}\left[X_d, X_1\right] & \mathrm{cov}\left[X_d, X_2\right] & \cdots & \mathrm{var}\left[X_d\right] \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \mathrm{cov}\left[\mathbf{X}\right]$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

$P(\underline{x})$ $\underline{X} = (x_1 \; x_2)$

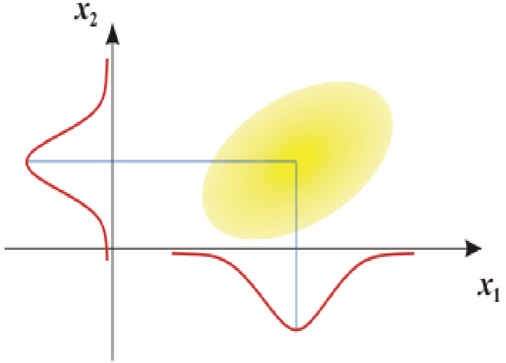




Bivariate Gaussian distribution example

Assume we have two ***independent*** *univariate Gaussian variables*

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



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

$$P(x_1 x_2) = P(x_2 | x_1) P(x_1)$$

$$= P(x_2) P(x_1)$$

$$= (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x_1-\mu_1)^2} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x_2-\mu_2)^2}$$

$$= |2\pi\Sigma|^{-1/2} e^{-\frac{1}{2}\left\{[(x_1-\mu_1)\ (x_2-\mu_2)]\begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}^{-1}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\right\}}$$

$$= |2\pi\Sigma|^{-1/2} e^{-1/2[\underline{x}-\underline{\mu}]^T\Sigma^{-1}[\underline{x}-\underline{\mu}]} \leftarrow$$

$$\Sigma = \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}$$

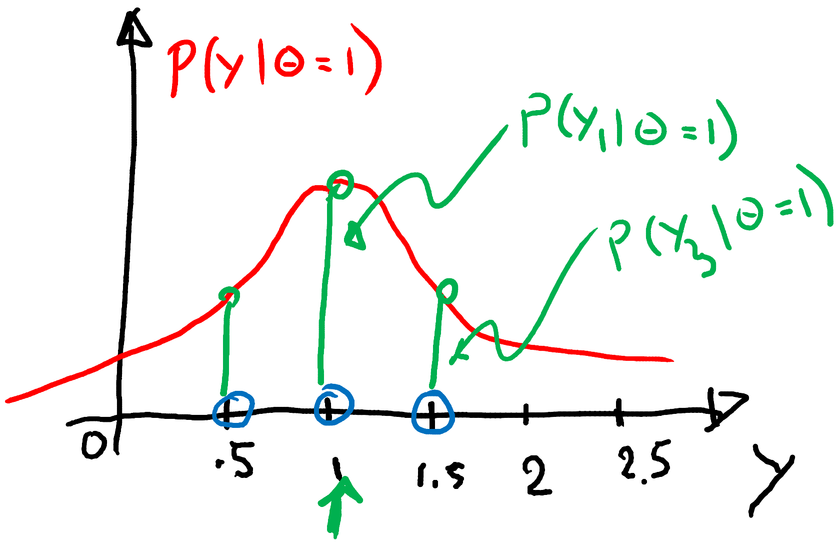
$$|\Sigma| = (\sigma^2)^2$$

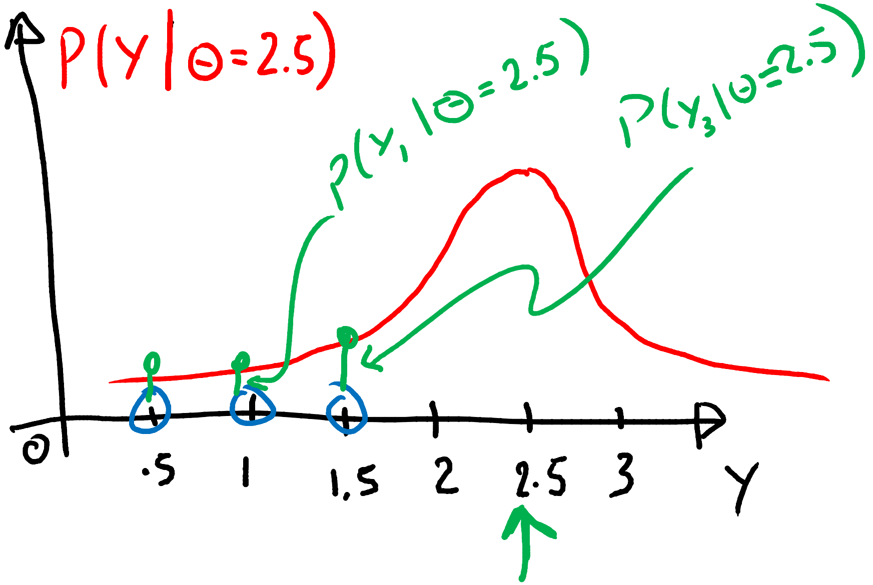
$$|a\Sigma| = a^2|\Sigma|$$

*We have **n=3** data points* y_1 **= 1**, y_2 **= 0.5**, y_3 **= 1.5**, *which are independent and Gaussian with* **unknown** *mean* θ *and variance* 1:

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

with **likelihood** $P(y_1 y_2 y_3|\theta) = P(y_1|\theta)\ P(y_2|\theta)\ P(y_3|\theta)$ **.** *Consider two guesses of* θ, *1 and 2.5. Which has higher likelihood (probability of generating the three observations)?*



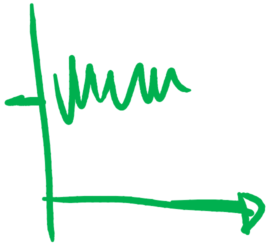


Finding the θ *that maximizes the likelihood is equivalent to moving the Gaussian until the product of 3 green bars (likelihood) is maximized.*

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:*

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



$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma) = \prod_{i=1}^{n} p(\underline{y_i|\mathbf{x}_i, \boldsymbol{\theta}, \sigma).}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1 & & x_{1d} \\ & \ddots & \vdots \\ & & x_{nd} \end{bmatrix}$$

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

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

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

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

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

$$\text{Prob}(\text{data} \overset{\text{given}}{|} \text{parameters}) = \frac{1}{Z} e^{-\text{Loss}(\text{data}, \text{parameters})}$$

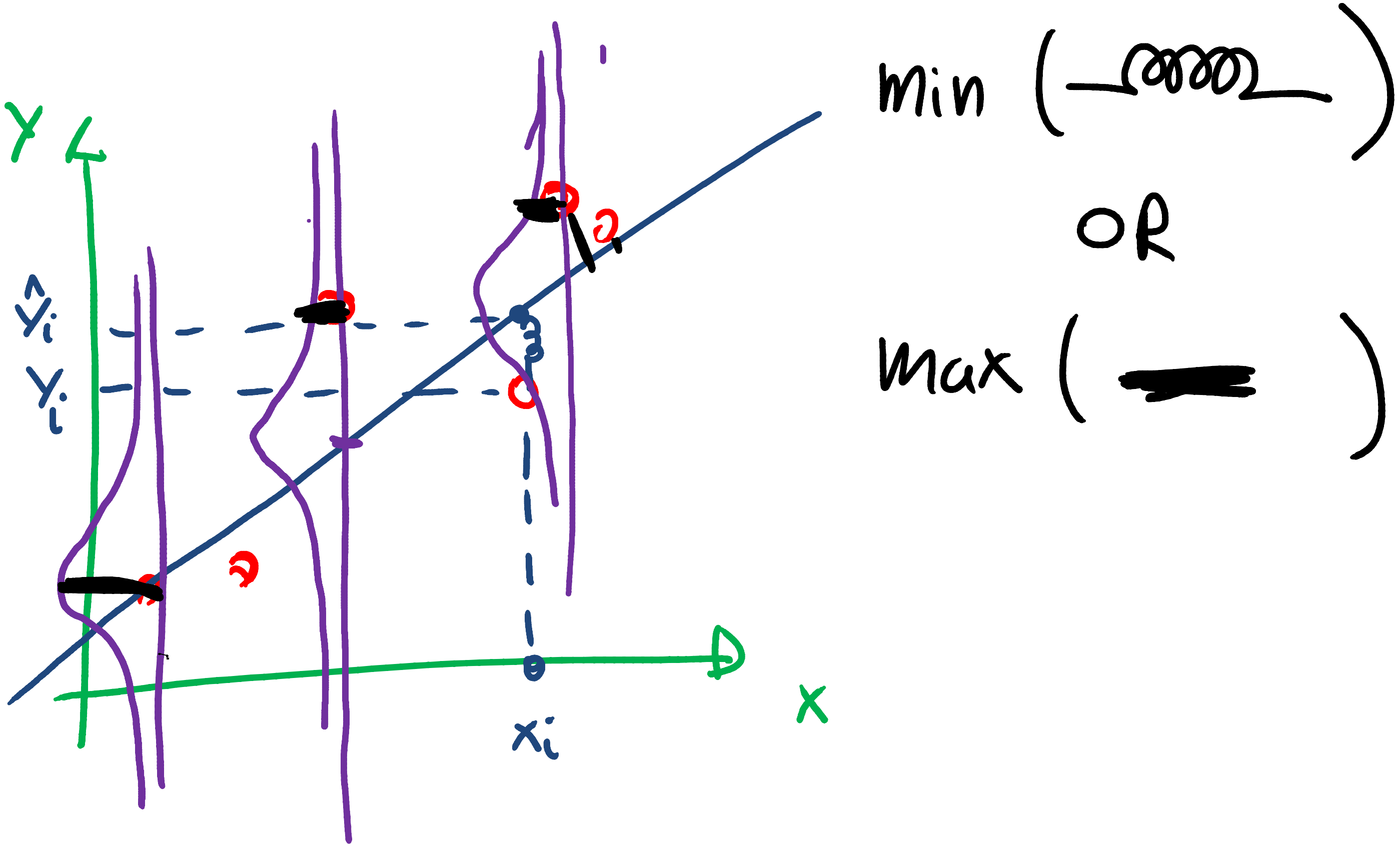
$$\hat{y}(\mathbf{x}_i) = \theta_1 + x_i \theta_2$$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \theta_1 - x_i \theta_2)^2$$

$$\hat{y}(\mathbf{x}_i) = \theta_1 + x_i\theta_2 = \hat{y}_i$$

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \theta_1 - x_i\theta_2)^2$$

$$P(y_i|x_i, \theta) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - \theta_1 - x_i\theta_2)^2}$$



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 $(\boldsymbol{\theta}, \sigma)$.

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

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta) \quad \underline{\text{Max}}$$

$$\text{neg } \ell\,\theta = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta) \quad \text{min}$$

The ML estimate of θ is:

$$\frac{\partial}{\partial \theta} \frac{1}{2\sigma^2} (y - X\theta)^T (y - X\theta)$$

$$\theta_{ML} = (X^T X)^{-1} X^T y$$

The ML estimate of σ is:

$$\frac{\partial}{\partial \sigma}\left[-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}(Y-X\theta)^T(Y-X\theta)\right]$$

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

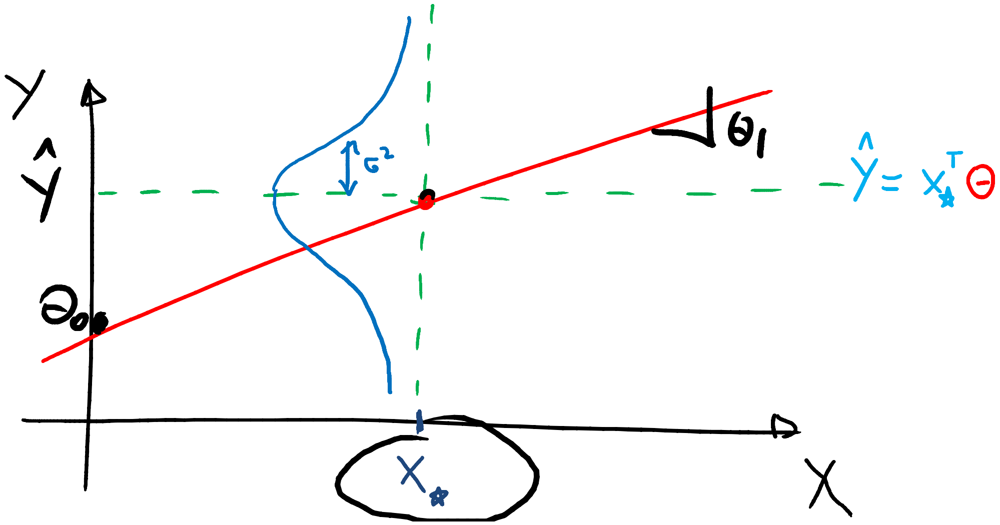
Equating $\frac{\partial \ell(\sigma)}{\partial \sigma}$ to zero:

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

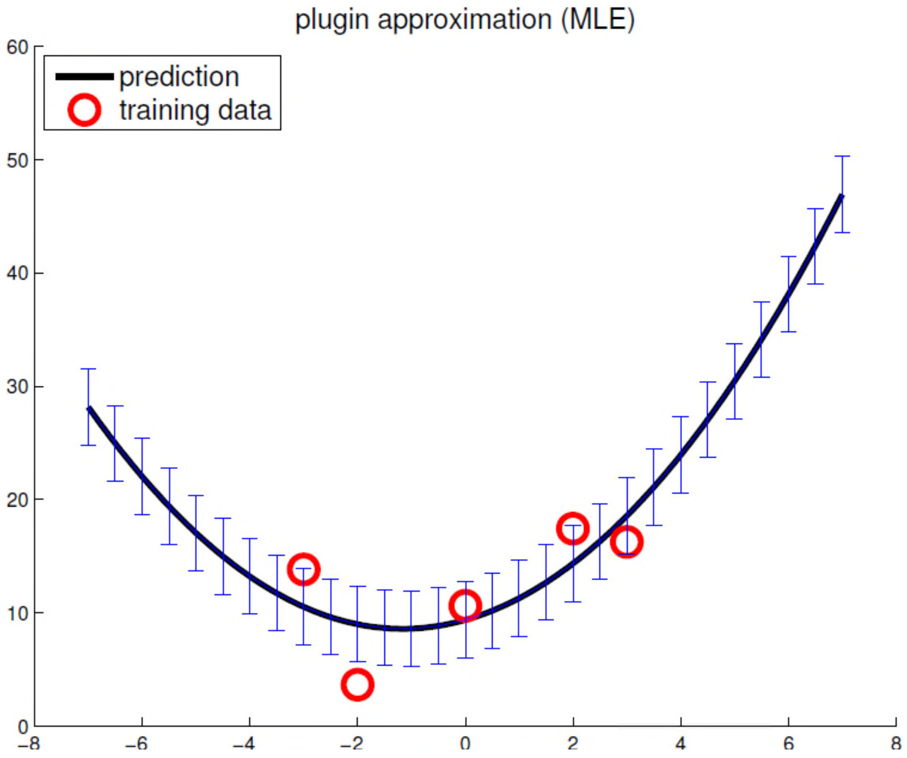
Making predictions

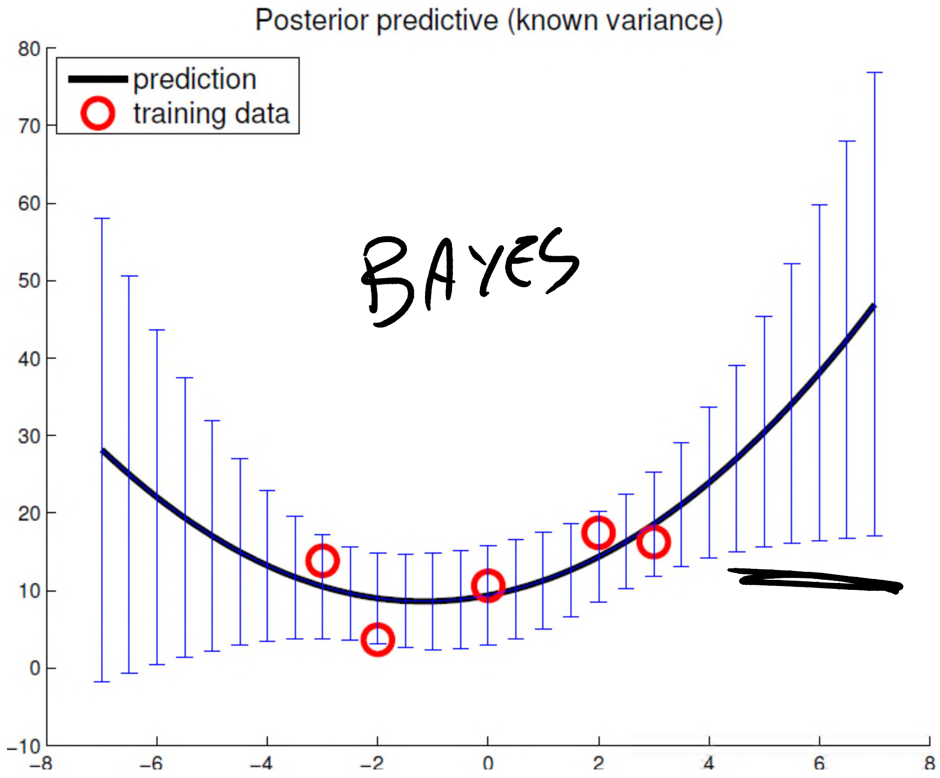
The ML plugin prediction, given the training data $\mathbf{D}=(\mathbf{X},\mathbf{y})$*, for a new input* $\mathbf{x}_*$ *and known* σ^2 *is given by:*

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



Confidence in the predictions

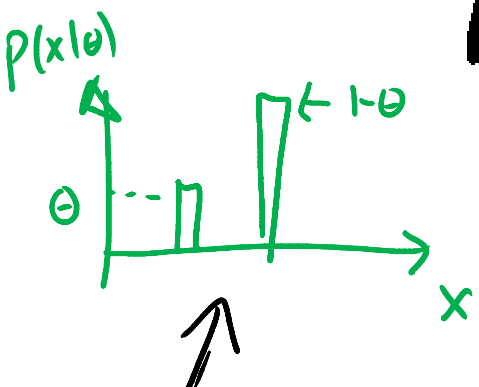




Bernoulli: a model for coins

*A **Bernoulli random variable r.v.** **X** takes values in {0,1}*

$$\underline{p(x|\theta)} = \begin{cases} \theta & \text{if } \; x=1 \\ 1-\theta & \text{if } \; x=0 \end{cases}$$



Where $\theta \in (0,1)$*. We can write this probability more succinctly as follows:*

$$P(x|\theta) = \underbrace{\theta^{x} (1-\theta)^{1-x}} = \begin{cases} \theta & x=1 \\ 1-\theta & x=0 \end{cases}$$

Entropy

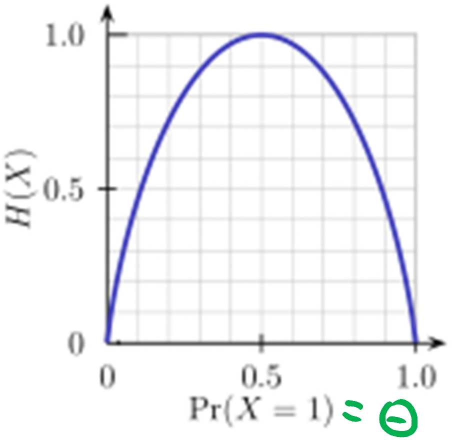
In information theory, entropy ***H*** *is a measure of the uncertainty associated with a random variable. It is defined as:*

$$\mathbf{H}(\mathbf{X}) = -\sum_{x} \underline{p(x|\theta)} \log \underline{p(x|\theta)}$$

<u>*Example:*</u> *For a Bernoulli variable* **X**, *the entropy is:*

$$H(x) = -\sum_{x=0}^{1} \underbrace{\theta^{x}(1-\theta)^{1-x}} \log \underbrace{\left[\theta^{x}(1-\theta)^{1-x}\right]}$$

$$= -\left[(1-\theta)\log(1-\theta) + \theta\log\theta\right] \leftarrow$$



Entropy of a Gaussian in D dimensions

$$h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2} \ln \left[(2\pi e)^D |\boldsymbol{\Sigma}| \right]$$

MLE - properties

For independent and identically distributed (i.i.d.) data from $p(x|\underline{\boldsymbol{\theta}_0})$, the MLE minimizes the **Kullback-Leibler divergence**:

true

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\underline{\boldsymbol{\theta}}} \prod_{i=1}^{N} \underline{p(x_i|\boldsymbol{\theta})} \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \underline{\log p}(x_i|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}) \overbrace{- \frac{1}{N} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}_0)} \\
&= \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i|\boldsymbol{\theta})}{p(x_i|\boldsymbol{\theta}_0)} \\
&\xrightarrow{N \to \infty} \arg\min_{\boldsymbol{\theta}} \int \log \frac{p(x|\boldsymbol{\theta}_0)}{p(\underline{x|\boldsymbol{\theta}})} p(x|\boldsymbol{\theta}_0) dx \Leftarrow KL \leftarrow \text{relative entropy}
\end{aligned}
$$

$X_i \sim P(x|\Theta_0)$

MLE - properties

$$\arg\min_{\boldsymbol{\theta}} \int \log \frac{p(x|\boldsymbol{\theta}_0)}{p(x|\boldsymbol{\theta})} p(x|\boldsymbol{\theta}_0)dx$$

$$= \underset{\theta}{\arg\min} \underbrace{\int P(x|\theta_0) \log \underline{P(x|\theta_0)} dx}_{\substack{\text{information} \\ \underline{\text{world}}}} \overset{\downarrow}{-} \underbrace{\int P(x|\theta_0) \log \underline{P(x|\theta)} dx}_{\substack{\text{in} \\ \underline{\text{model}}}}$$

Next lecture

In the next lecture, we introduce ridge regression, bases functions and look at the issue of controlling complexity.