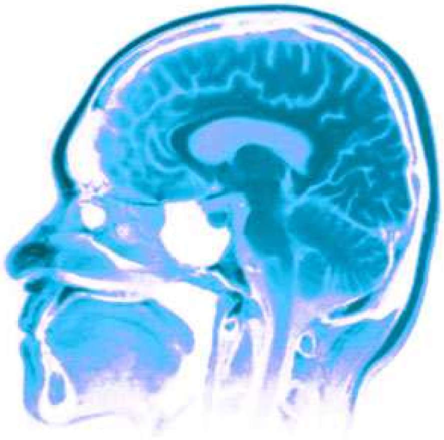
CPSC540



Probabilistic linear prediction and maximum likelihood

Nando de Freitas
January, 2013
University of British Columbia

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:

❑ Multivariate Gaussian distributions

❑ How to formulate the likelihood for linear regression

❑ Computing the maximum likelihood estimates for linear regression.

❑ Understand why maximum likelihood is used.

Univariate Gaussian distribution

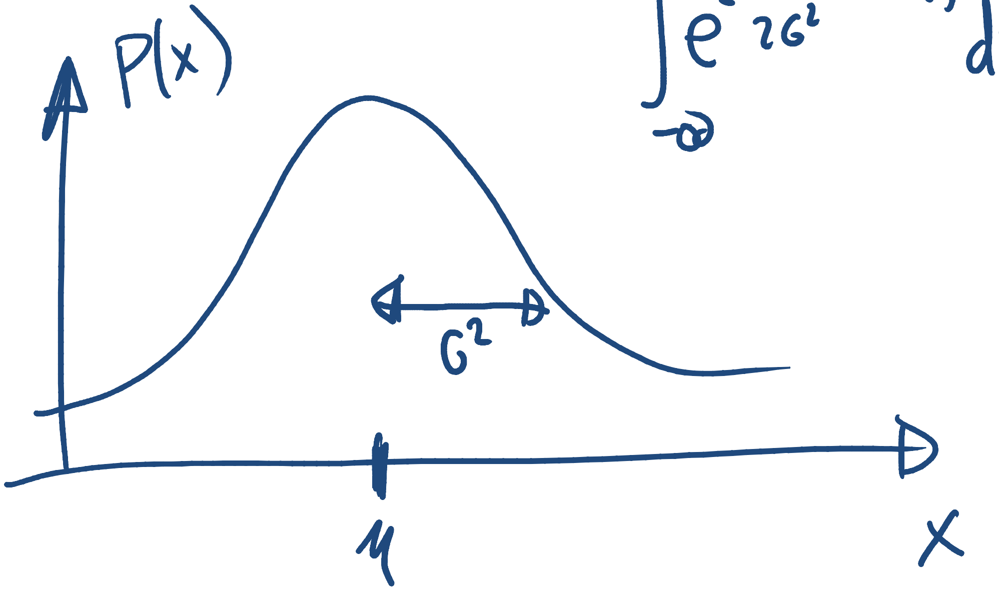
The probability density function (pdf) of a Gaussian distribution is given by

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

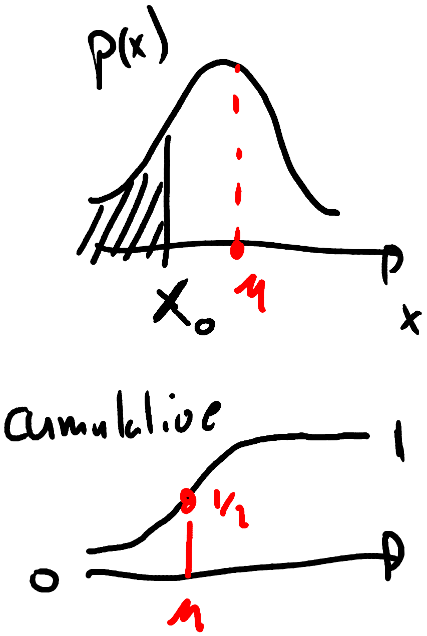
where μ is the mean or center of mass and σ^2 is the variance.

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \sqrt{2\pi\sigma^2}$$

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

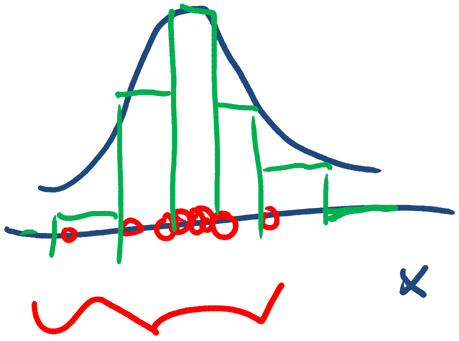


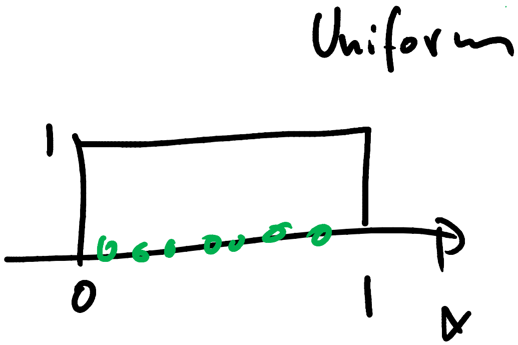
Sampling from a Gaussian distribution

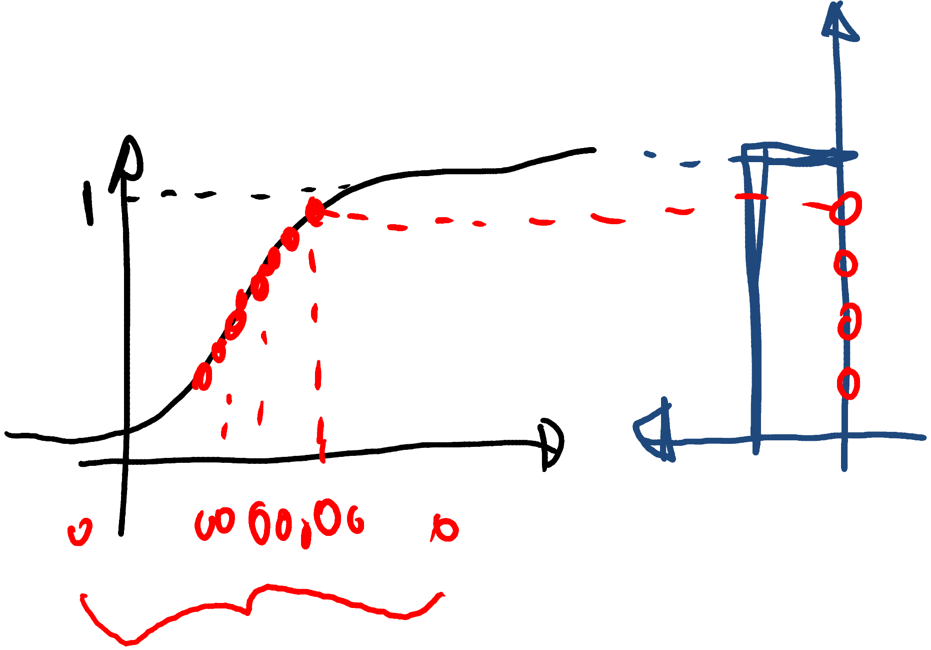


$$x \sim \mathcal{N}(\mu, \sigma^2)$$

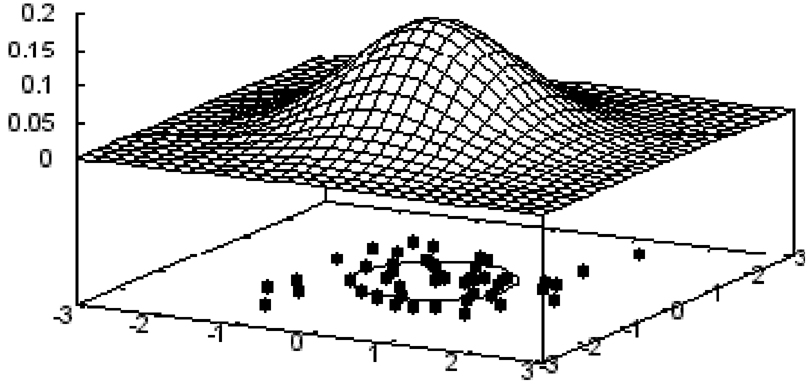
simulated
sample
generated

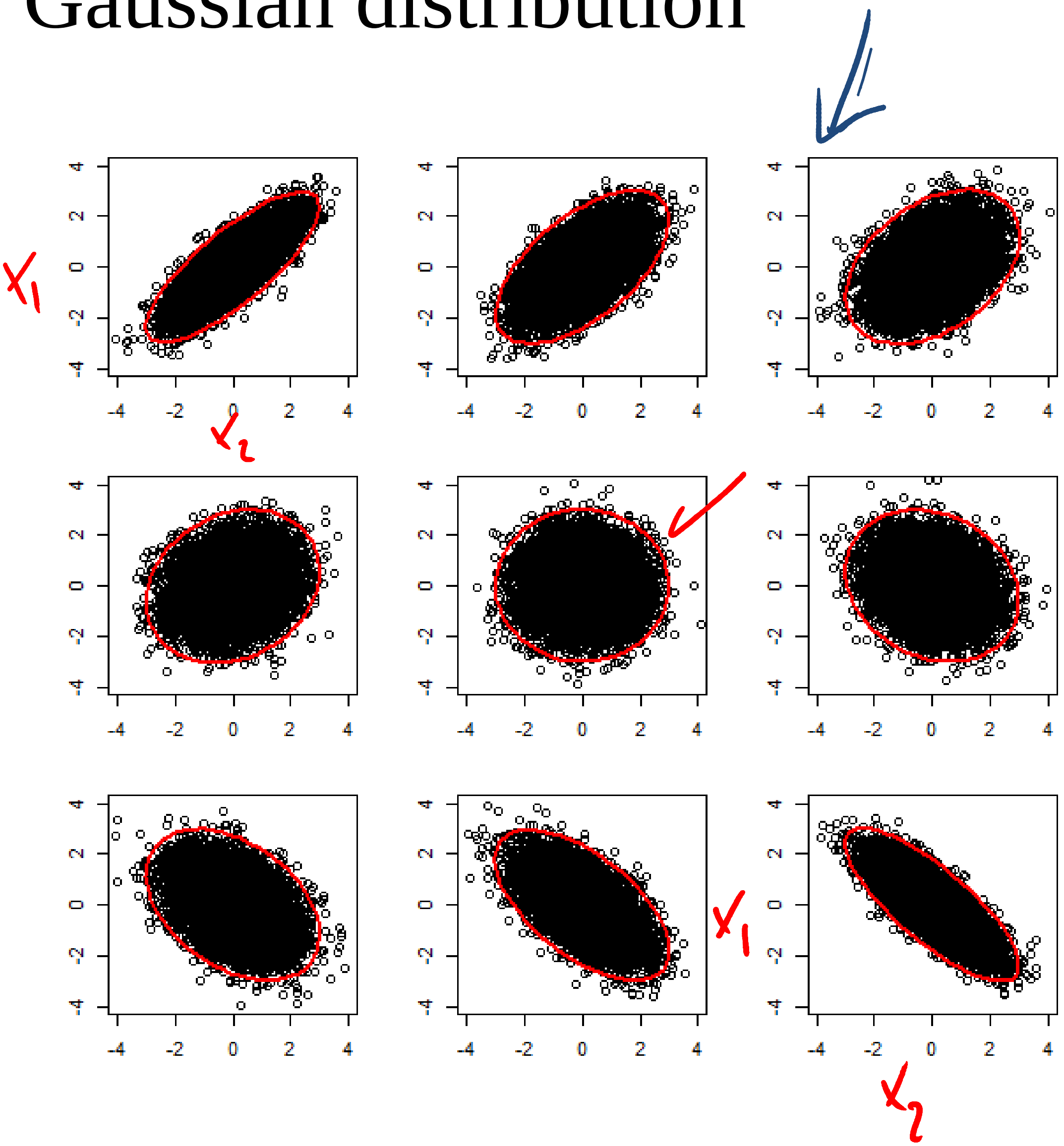






The bivariate Gaussian distribution





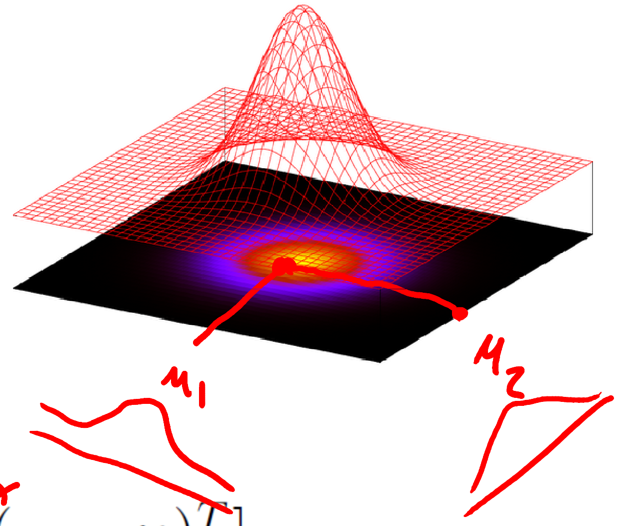
Multivariate Gaussian distribution

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

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

where

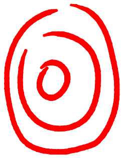
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \mathbb{E}(y_1) \\ \vdots \\ \mathbb{E}(y_n) \end{pmatrix}$$



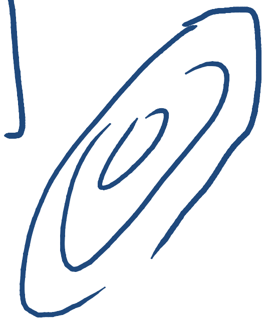
and

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

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$$

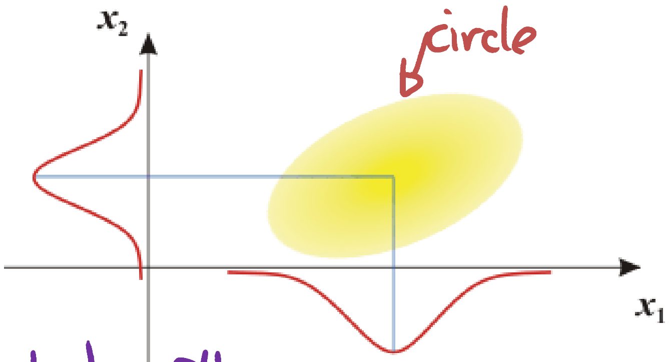


$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Bivariate Gaussian distribution example

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

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



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

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

$$|\Sigma| = \sigma^4$$

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

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

$$= (2\pi\sigma^2)^{-\frac{2}{2}} e^{-\frac{1}{2}\left[(x_1-\mu_1)^T(\sigma^2)^{-1}(x_1-\mu_1) + (x_2-\mu_2)^T(\sigma^2)^{-1}(x_2-\mu_2)\right]}$$

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

Sampling from a multivariate Gaussian distribution

$$x \sim N(\mu, \sigma^2) \qquad\qquad \underbrace{x \sim N(0, 1)}$$

$$|||$$

$$x \sim \mu + \sigma N(0, 1)$$

$$\underline{x} \sim N(\underline{\mu}, \underline{\underline{\Sigma}}) \quad , \quad \underline{\underline{\Sigma}} = \underline{\underline{B}}\,\underline{\underline{B}}^T \text{ (cholesky)}$$

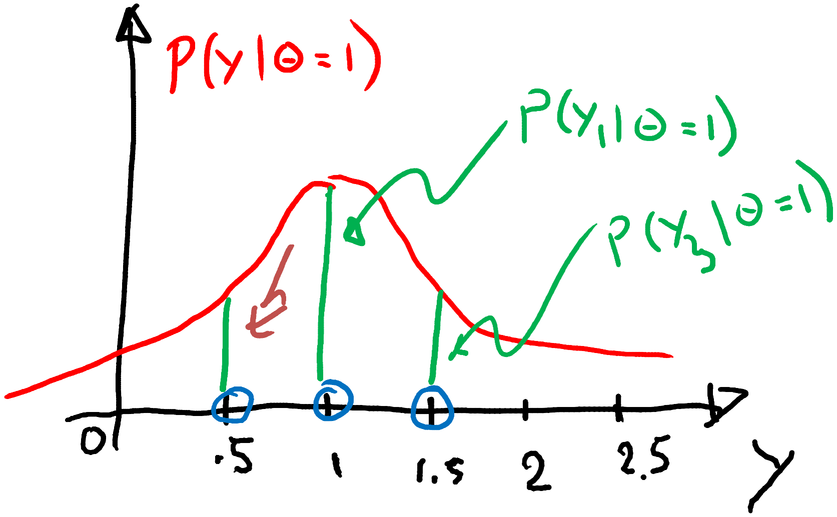
$$\underline{x} \sim \underline{\mu} + \underline{\underline{B}}\ \underline{N(0, I)}$$

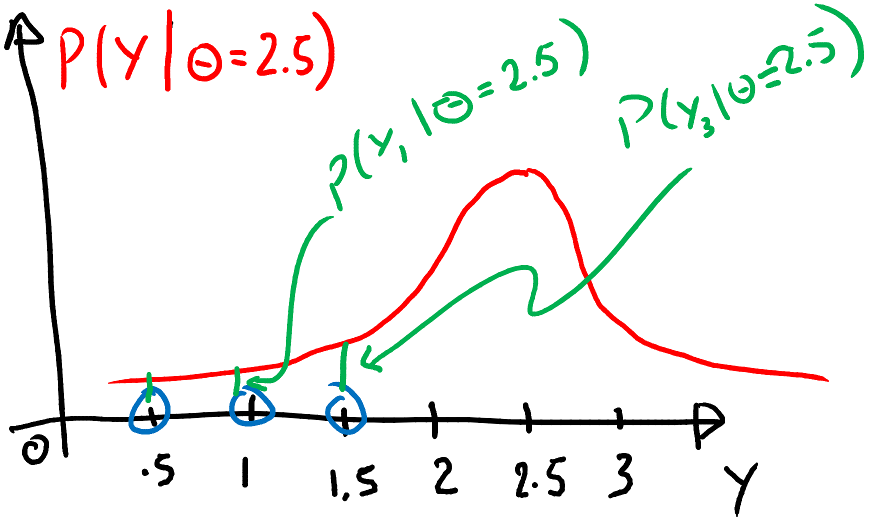
We have **n=3** *data points* y_1 **= 1,** y_2 **= 0.5,** y_3 **= 1.5,** *which are independent and Gaussian with* **unknown** *mean* θ *and variance* 1*:*

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

with likelihood $\mathbf{P}(y_1 y_2 y_3|\theta) = \mathbf{P}(y_1|\theta)\ \mathbf{P}(y_1|\theta)\ \mathbf{P}(y_3|\theta)$ **.** *Consider two guesses of* θ*, 1 and 2.5. Which has higher likelihood?*

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:*

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

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

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

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

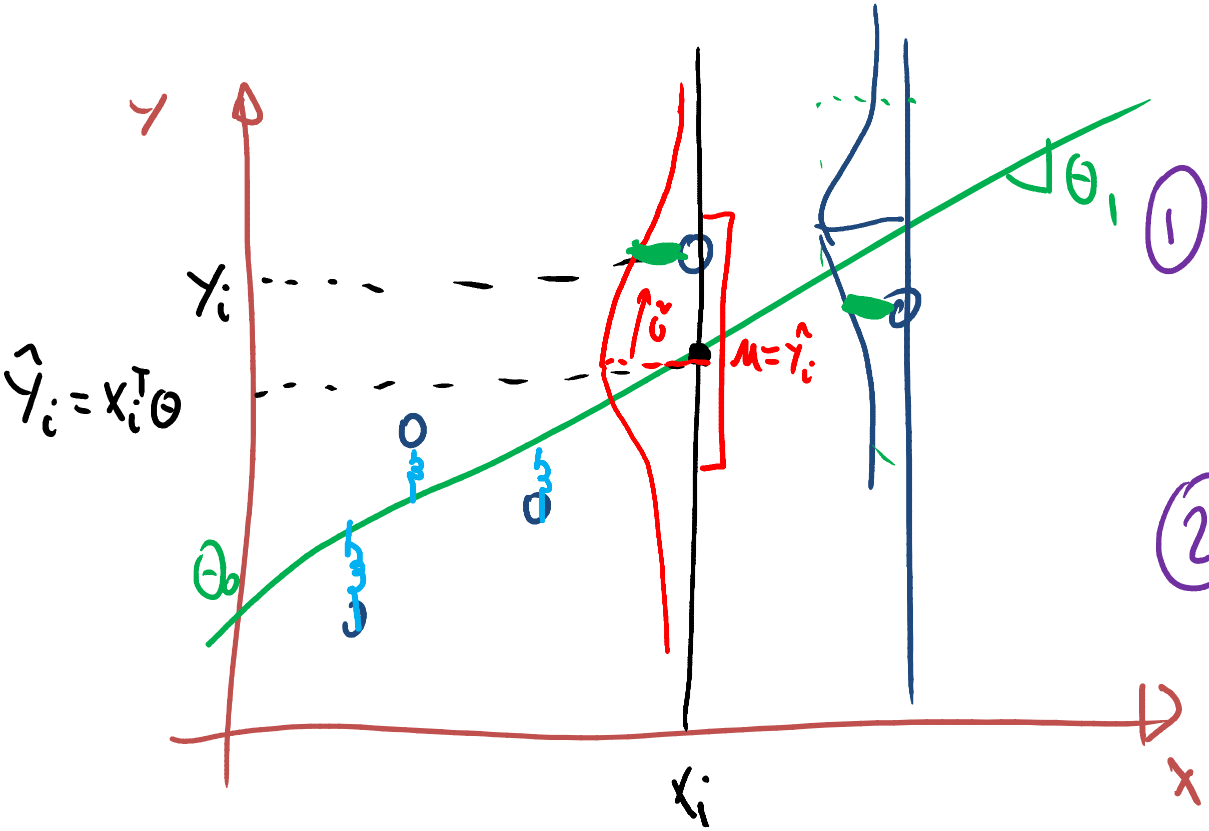
$$= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}$$

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

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

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

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



① Minimize COST

Sum(⌇)

② Maximize PROB.

Product(▬)

Maximum likelihood

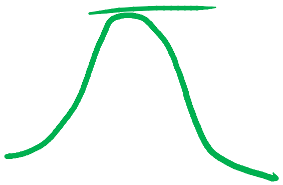
The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is obtained by taking the derivative of the log-likelihood, $\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma)$. The goal is to maximize the likelihood of seeing the training data $\mathbf{y}$ by modifying the parameters $(\boldsymbol{\theta},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

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

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

The ML estimate of θ *is:*

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



$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 - \frac{1}{2\sigma^2}\left[0 - 2X^Ty + 2X^TX\theta\right]$$

equating to zero

$$\hat{\theta}_{ML} = (X^TX)^{-1}X^Ty$$

The ML estimate of σ is:

$$\frac{\partial}{\partial \sigma} \ell(\sigma)$$

$\vdots$ $\underbrace{\text{DO THIS}}$!

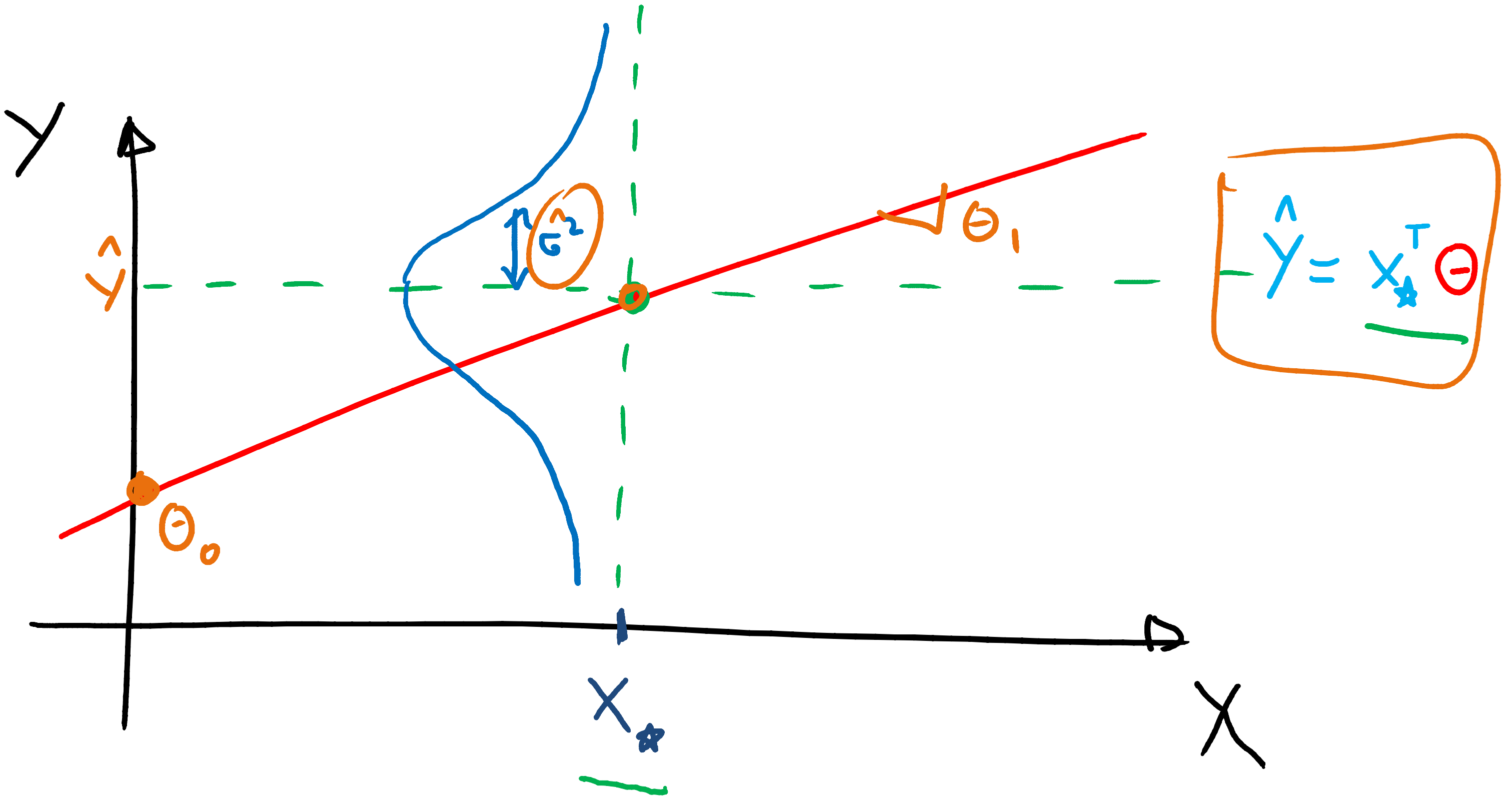
$$\sigma^2 = \frac{1}{n} (y - X\theta)^T (y - X\theta) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - x_i\theta)^2}$$

Making predictions

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

↳ $\hat{\theta}_{ML}$ $\hat{\sigma}_{ML}$

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



Frequentist learning and maximum likelihood

Frequentist learning *assumes that there exists a true model, say with parameters* θ_o .

The estimate (learned value) will be denoted $\hat{\theta}$.

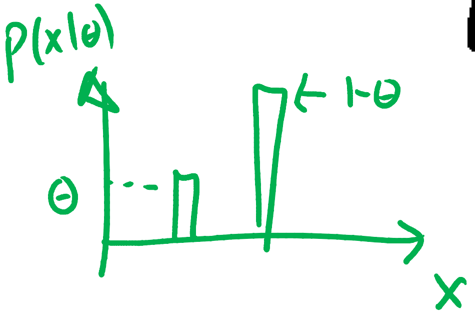
Given **n** *data,* $x_{1:n} = \{x_1, x_2, \ldots, x_n\}$*, we choose the value of* θ *that has more probability of generating the data. That is,*

$$\hat{\theta} = \arg\max_{\theta} \; p(x_{1:n} \mid \theta)$$

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:

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

Entropy

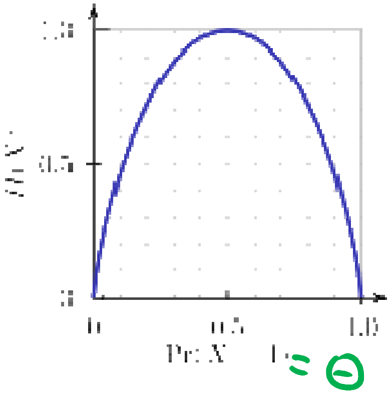
*In information theory, entropy **H** is a measure of the uncertainty associated with a random variable. It is defined as:*

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

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

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

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



MLE - properties

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

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \boxed{\arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta})} = \arg\max_{\theta} P(x_{1:n}|\theta) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}) = \arg\max_{\theta} \log P(x_{1:n}|\theta) \\
&= \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}) \boxed{- \frac{1}{N} \sum_{i=1}^{N} \log p(x_i|\boldsymbol{\theta}_0)} \checkmark \\
&= \arg\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{i=1}^{N} \log \frac{p(x_i|\boldsymbol{\theta})}{p(x_i|\boldsymbol{\theta}_0)} \\
&\xrightarrow{N\to\infty} \arg\min_{\boldsymbol{\theta}} \int \log \frac{p(x|\boldsymbol{\theta}_0)}{p(x|\boldsymbol{\theta})} p(x|\boldsymbol{\theta}_0)\,dx \cong \text{KL divergence}
\end{aligned}
$$

given

$x_i \sim P(x_i|\theta_0)$

LLN.

$E(x) = \int x\, p(x)\, dx$

$\uparrow N \to \infty$

$\hat{E}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$

True when $x_i \sim p(x)$

MLE - properties

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

$$= \arg\min_{\theta} \underbrace{\int P(x|\theta_0) \operatorname{Log} P(x|\theta_0) dx}_{\text{information world}} - \underbrace{\int P(x|\theta_0) \log P(x|\theta) dx}_{\text{in model}}$$

MLE - properties

Under smoothness and identifiability assumptions, the MLE is **consistent**:

$$\underline{\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0}$$

or equivalently,

$$|||$$

$$\text{plim}(\hat{\boldsymbol{\theta}}) = \theta_0$$

or equivalently,

$$|||$$

$$\lim_{N \to \infty} P(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0| > \alpha) \to 0$$

for every α.

MLE - properties

The MLE is **asymptotically normal**. That is, as $N \to \infty$, we have:

$$\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \Longrightarrow N(0, I^{-1})$$

All of Statistics

where I is the **Fisher Information matrix**.

It is asymptotically optimal or **efficient**. That is, asymptotically, it has the lowest variance among all well behaved estimators. In particular it attains a lower bound on the CLT variance known as the **Cramer-Rao lower bound**.

But what about issues like robustness and computation? Is MLE always the right option?

Bias and variance

Note that the estimator is a function of the data: $\boxed{\hat{\boldsymbol{\theta}} = g(\mathcal{D}).}$ $D = X_{1:n}$

Its **bias** is:

$$bias(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0 = \underline{\bar{\boldsymbol{\theta}}} - \underline{\boldsymbol{\theta}_0}$$

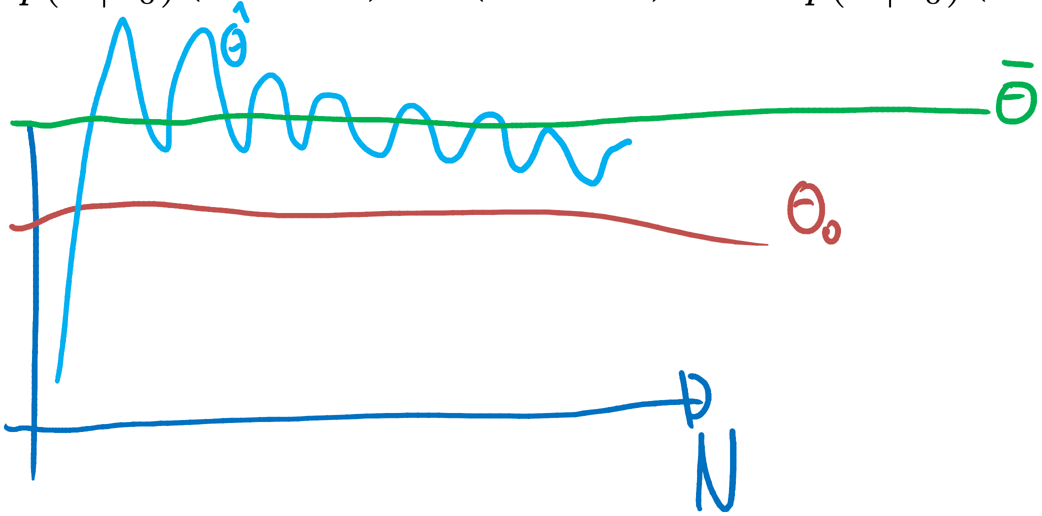
$$\bar{\theta} = \int \hat{\theta}\, p(D|\theta_0)\, dD$$

Its **variance** is:

$$\mathbb{V}(\hat{\boldsymbol{\theta}}) = \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\underline{\hat{\boldsymbol{\theta}}} - \underline{\bar{\boldsymbol{\theta}}})^2$$

Its **mean squared error** is:

$$\text{MSE} = \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{2} = (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^2 + \mathbb{E}_{p(\mathcal{D}|\boldsymbol{\theta}_0)}(\hat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^2$$



Next lecture

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