# Machine Learning - Notes

Dom Hutchinson

October 28, 2019

## Contents

## General

Lecturer - Carl Henrik Ek
Course Website - `http://carlhenrik.com/COMS30007/`
Course Repo - `https://github.com/carlhenrikek/COMS30007`
Course Subreddit - `https://www.reddit.com/r/coms30007/`

# 1   Introduction

## 1.1   Motivation

**Definition 1.1 -** *Deductive Reasoning*
A method of reasoning in which the premieses are viewed as supplying <u>all</u> the evidence for the truth of the conclusion.

**Definition 1.2 -** *Inductive Reasoning*
A method of reasoning in which the premieses are viewed as supplying <u>some</u> evidence for the truth of the conclusion, rather than all the evidence. This allows for the conclusion of the *Inductive Reasoning* to be false.

**Remark 1.1 -** *Free-Lunch Theorem*
There are infinite number of hypotheses that perfectly explain the data. Adding a data point removes an infinite number of possibilities, but still leaves infinite possibilities.

**Remark 1.2 -** *The Task of Machine Learning*
When proposing to use machine learning on a task, one should consider the following questions:

  i) How can we formulate beliefes ad assumptions mathematically?

  ii) How can we connect our assumptions with data?

  iii) How can we update our beliefs?

**Remark 1.3 -** *Useful Models are not always True*
Our goal is to understand realisations of a system. If we can then we can equate our model to the system. It is important to note thtat our model does not need to be perfectly true to be useful.

## 1.2   Probability Theory

**Definition 1.3 -** *Stochastic/Random Variable*
A variable whose value depends on outcomes of random phenomona.
*e.g.* $x \sim \mathcal{N}(0,1)$.

**Definition 1.4 -** *Probability Measure,$\mathbb{P}$*
A function with signature $\mathbb{P} : \mathcal{F} \to [0,1]$, where $\mathcal{F}$ is a sample space of rv $X$, and fulfils $\int_{-\infty}^{\infty} \mathbb{P}(x)dx = 1$.

**Definition 1.5 -** *Joint Probability Distribution*
A *Probability Measure* for multiple variables, $\mathbb{P} : X \times Y \to [0,1]$.
Let $n_{ij}$ be the number of outcomes where $X = x_i$ and $Y = y_j$ then

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{\sum_{i,j} n_{ij}}$$

**Definition 1.6 -** *Marginal Probability Distribution*
A *Probability Measure* for one variable when the sample space is over multiple variables.
Let $n_{ij}$ be the number of outcomes where $X = x_i$ and $Y = y_j$ then

$$\mathbb{P}(X = x_i) = \frac{\sum_j n_{ij}}{\sum_{i,j} n_{ij}}$$

**Definition 1.7 -** *Conditional Probability Distribution*
A *Probability Measure* for a variable, given another variable has a defined value. Let $n_{ij}$ be the number of outcomes where $X = x_i$ and $Y = y_j$ then

$$\mathbb{P}(Y = y_j | X = x_i) = \frac{n_{ij}}{\sum_j n_{ij}}$$

**Example 1.1 -** *Joint, Marginal & Conditional Probability*
The below image shows two marginals distributions in the bottom-left, $X$, & top-right, $Y$, their joint distribution in the top-left and a conditional in the bottom right $\mathbb{P}(Y | X = 1)$.
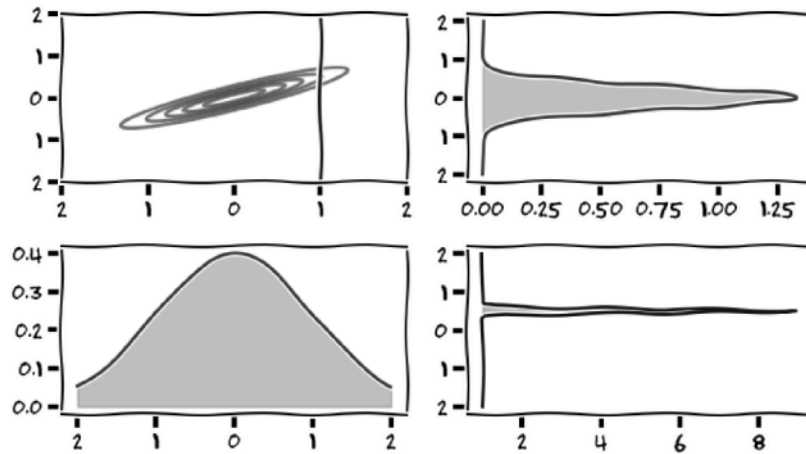


**Theorem 1.1 -** *Product Rule*
For random variables $X$ & $Y$

$$\mathbb{P}(X = x, Y = Y) = \mathbb{P}(Y = y | X = x)\mathbb{P}(X = x)$$

**Theorem 1.2 -** *Sum Rule*
For random variables $X$ & $Y$

$$\mathbb{P}(X = x) = \sum_j \mathbb{P}(X = x, Y = y_j)$$

**Theorem 1.3 -** *Baye's Theorem*
For random variables $X$ & $Y$

$$\mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(Y = y | X = x)\mathbb{P}(X = x)}{\mathbb{P}(Y = y)}$$

**Definition 1.8 -** *Elements of Bayes' Theorem*
The elements of *Bayes' Theory* can be broken down to explain parts of the model.

$$\underbrace{\mathbb{P}(\theta | Y)}_{\text{Posterior}} = \frac{\overbrace{\mathbb{P}(Y | \theta)}^{\text{Likelihood}} \overbrace{\mathbb{P}(\theta)}^{\text{Prior}}}{\underbrace{\mathbb{P}(Y)}_{\text{Evidence}}}$$

| | |
|---|---|
| Posterior | Which parameters of the model do I belive produce distributions have generated the data $Y$ |
| Likelihood | How likley is the data to come from the model specifely indexed by $\theta$ |
| Prior | What distribution do I think parameter $\theta$ has |
| Evidence | How likely do I think data $Y$ is for all models. |

*N.B.* The *Evidence* normalises this function.

**Definition 1.9 -** *Expectaction Value,* $\mathbb{E}$
The mean value a random variable will produce from a large number of samples.

| Continuous | Discrete |
|---|---|
| $\mathbb{E}(X) = \int_{-\infty}^{\infty} x\mathbb{P}(X)dx$ | $\mathbb{E}(X) = \sum_{-\infty}^{\infty} x\mathbb{P}(X)dx$ |
| $\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(x)\mathbb{P}(X)dx$ | $\mathbb{E}(f(X)) = \sum_{-\infty}^{\infty} f(x)\mathbb{P}(X)dx$ |

**Definition 1.10 -** *Variance*
Describes the amount of spread in the values a single random variable will produce.

$$\text{var}(X) = \mathbb{E}\left(x - \mathbb{E}(x))^2\right) = \mathbb{E}(X^2) - \left(\mathbb{E}(X)\right)^2$$

**Definition 1.11 -** *Covariance*
Describes the joint variability between two random variables.

$$\text{cov}(X, Y) = \mathbb{E}\left(\left(X - \mathbb{E}(X)\right)\left(Y - \mathbb{E}(Y)\right)\right)$$

**Definition 1.12 -** *Marginalisation*
The process of summing out the probability of one random variable using its joing probability with another rando variable.

$$\begin{array}{rcl} \text{Continuous} & \mathbb{P}(X = x) = & \int (X = x, Y = y)dy \\ \text{Discrete} & \mathbb{P}(X = x) = & \sum_i \mathbb{P}(X = x, Y = y_i) \end{array}$$

**Definition 1.13 -** *Likelihood Function*
Define $\mathbf{X} \sim f_n(\cdot; \theta^*)$ for some unknown $\theta^* \in \Theta$ and let $\mathbf{x}$ be an observation of $\mathbf{X}$.
A *Likelihood Function* is any function, $L(\cdot; \mathbf{x}) : \Theta \to [0, \infty)$, which is proportional to the PMF/PDF of the obeserved realisation $\mathbf{x}$.

$$L(\theta; \mathbf{x}) := Cf_b(\mathbf{x}; \theta) \ \forall \ C > 0$$

*N.B.* Sometimes this is called the *Observed* Likelihood Function since it is dependent on observed data.

**Definition 1.14 -** *Log-Likelihood Function*
Let $\mathbf{X} \sim f_n(\cdot; \theta^*)$ for some unknown $\theta^* \in \Theta$ and $\mathbf{x}$ be an observation of $\mathbf{X}$.
The *Log-Likelihood Function* is the natural log of a *Likelihood Function*

$$\ell(\theta; \mathbf{x}) := \ln f_n(\mathbf{x}; \theta) + C, \ C \in \mathbb{R}$$

**Definition 1.15 -** *Maximum Likelihood Estimation*
The *Maximum Likelihood Estimate* is an estimate for a parameter of a probability distribution which is the value which maximises the *Likelihood Function* (or the *Likelihood Function*).

$$\hat{\theta} := \text{argmax}_\theta L(\theta; \mathbf{x})$$

**Definition 1.16 -** *Central Limit Theorem*
The distribution of the sum (or mean) of a large number of independent, identically distributied random variables can be approximated to a normal distribution, regardless of the distributions of the random variables.

## 1.3    Conjugate Priors

**Definition 1.17 -** *Conjugate Prior*
If we have a *Likelihood Function*, $\mathbb{P}(X|\theta)$, with a known distribution (*e.g.* Normal) we can choose our *Prior*, $\mathbb{P}(\theta)$, to be from a distribution which is *Conjugate* to the distribution of the *Likelihood Function*.
These are defined in *tables*

**Remark 1.4 -** *Why use Conjugate Priors?*
If we have a *Conjugate Prior* then the *Posterior*, $\mathbb{P}(\theta|X)$, will be in the same distribution family as the Prior too. We can then work out the distribution of the *Posterior* by passing the parameters of the *Prior* through pre-derived functions

$$\begin{aligned} \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \\ \mathbb{P}(\theta|X) &\propto \mathbb{P}(X|\theta) \times \mathbb{P}(\theta) \end{aligned}$$

*N.B.* - https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions

**Example 1.2 -** *Conjugate Priors*
Consider a scenario where we are flipping a coin. We may have *Likelihood Function* $\theta^x(1-\theta)^{n-x}$.
If we choose our *Prior* to be $\theta^{a-1}(1-\theta)^{b-1}$ which is a *Beta Distribution*.
Then (after some maths) we find the *Posterior*

# 2    Distributions

**Definition 2.1 -** *Bernoulli Distribution*
Models an event with a binary outcome (0 or 1) with parameter $p$ st $\mathbb{P}(X = 1) = p$ Let $X \sim \text{Bernoulli}(p)$. Then

$$\begin{aligned} f_X(x) &= \begin{cases} p & , x = 1 \\ 1 - p & , x = 0 \\ 0 & \text{otherwise} \end{cases} \\ F_X(x) &= \begin{cases} 0 & , x < 0 \\ 1 - p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases} \\ \mathbb{E}(X) &= p \\ \text{Var}(X) &= p(1 - p) \end{aligned}$$

**Definition 2.2 -** $\beta$*-Distribution*
A $\beta$*-Distribution* is a continuous distribution over interval $[0, 1]$ which is parameterised by two positive *shape parameters*, $\alpha$ & $\beta$. A $\beta$*-Distribution* can be used to encode assumptions as a *Prior*.
Let $X \sim \beta(\alpha, \beta)$. Then

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

**Example 2.1 -** $\beta$*-Distribution*
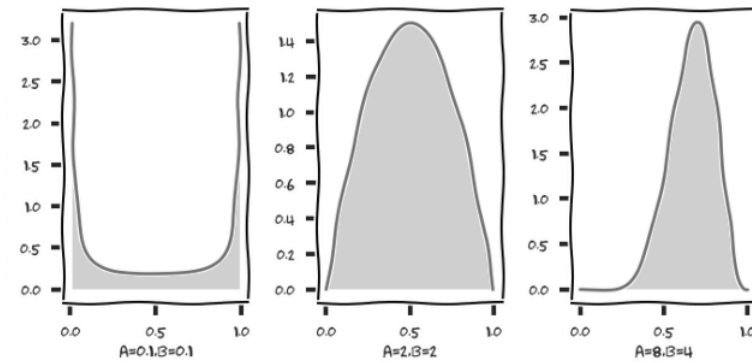
**Definition 2.3 -** *Direchlet Distribution*
Let $X \sim \text{Dir}(\boldsymbol{\alpha})$. Then

$$f_X(x) := \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \times \cdots \times \Gamma(\alpha_N)} \prod_{i=1}^{N} x_i^{\alpha_i - 1}$$

**Definition 2.4 -** *Exponential Distribution Family*
The *Exponential Distribution Family* is a set of probability distributions which fit the form.

$$\mathbb{P}(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x})g(\boldsymbol{\theta})e^{\boldsymbol{\theta}^T \mathbf{u}(\mathbf{x})}$$

With conjugate prior

$$\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\chi})^\nu e^{\nu \boldsymbol{\theta}^T \boldsymbol{\chi}}$$

**Definition 2.5 -** *Multivariate Normal Distribution*
Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

*N.B.* Also known as *Gaussian Distribution*.

# 3   Regression

**Definition 3.1 -** *Supervised Learning*
Learning the relationship $f(\cdot)$ between pairs of data $x_i$ and $y_i$ where $y_i = f(x_i)$.

**Remark 3.1 -** *Summary of Regression*

Linear Regression  We are limited to lines

unctions Regression  We can use non-linear functions, but it is hard to determine how many & what basis functions should be used. The prior is hard to interpret.

Kernel Regression  The complexity is defined by the data (Good) but there is no uncertainty in our estimate.

## 3.1   Linear Regression

**Definition 3.2 -** *Linear Regression*
*Linear Regression* is the process of taking a set of data points & producing a linear relationship between a dependent varaible & one of more explanatory variables.
Let $\mathbf{x} \in \mathbb{R}^n$ be a set of observed values from $n$ explanatory variables & $\mathbf{a} \in \mathbb{R}^n + 1$ be a set of parameters. Then we predict the value of the dependent variable to be

$$y(\mathbf{x}, \mathbf{a}) = a_0 + \sum_{i=0}^{n} a_{i+1} x_i$$

**Remark 3.2 -** *Limitation of Linear Regression*
The formula defined in **Definition 3.1** is a linear function of the coefficients defined by **a and**
the observed values of **x** this limits the relationships we can model between elements of **x**. The
model can be extended to avoid this using *Basis Functions*.

**Definition 3.3 -** *Linear Regression - Basis Functions*
We can extend *Linear Regression* to include *Basis Functions* so that relationships between
explanatory variables can be modelled.
Let $\mathbf{x} \in \mathbb{R}^n$ be a set of observed values from explanatory variables, $\mathbf{a} \in \mathbb{R}^m$ be a set of coefficients
(weightings), and $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^{m-1}$ be a set of basis functions. Then we can predict the dependent
variable to be

$$y(\mathbf{x}, \mathbf{a}) = a_0 + \sum_{i=1}^{m} a_i \phi_{i-1}(\mathbf{x})$$

**Remark 3.3 -** *Linear Regression - Basis Function*
To simplfy the equation used in **Definition 3.2** we can define $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^m$ with $\phi_0(\mathbf{x})$. Then

$$y(\mathbf{x}, \mathbf{a}) = \sum_{i=0}^{m} a_i \phi_i(\mathbf{x}) = \mathbf{a}^T \boldsymbol{\phi}(\mathbf{x})$$

**Proposition 3.1 -** *Noise*
We will often introduce the concept of *Noise* into a *Linear Regression* model. Typically we as-
sume noise to be modelled by a zero-mean Normal distribution with precision $\beta$, $\varepsilon \sim \text{Normal}(0, \beta^{-1})$,
so

$$t := y(\mathbf{x}, \mathbf{a}) = \mathbf{a}^T \boldsymbol{\phi}(\mathbf{x}) + \varepsilon$$

From this we can derive a likelihood

$$\mathbb{P}(t|\mathbf{x}, \mathbf{a}, \beta) \sim \text{Normal}(t|\mu = y(\mathbf{x}, \mathbf{a}), \sigma^2 = \beta^{-1}) = \text{Normal}(t|\mu = \mathbf{a}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2 = \beta^{-1})$$

If we have a series of sets of observations, $\mathbf{X} \in \mathbb{R}^{m \times n}$, then

$$\mathbb{P}(\mathbf{t}|\mathbf{X}, \mathbf{a}, \beta) \sim \prod_{i=1}^{m} \text{Normal}(t_i|\mu = \mathbf{a}^T \boldsymbol{\phi}(\mathbf{x}_i), \sigma^2 = \beta^{-1})$$

**Definition 3.4 -** *Maximum Likelihood Estimate*
A *Maximum Likelihood Estimate* is estimating the value of a parameter to be the most likely,
according to our *Likelihood Function*.

**Remark 3.4 -** *Finding Maximum Likelihood Estimate*
Suppose we have a defined *Likelihood Function* $\mathbb{P}(\mathbf{t}|\mathbf{X}, \mathbf{a}, \beta)$ and we want to find *Maximum
Likelihood Estimates* for parameters **a**. Then

   i) Define the *Likelihood Function*, $\mathbb{P}(\mathbf{t}|\mathbf{X}, \mathbf{a}, \beta)$.

   ii) Take the natural log, $\ln \mathbb{P}(\mathbf{t}|\mathbf{X}, \mathbf{a}, \beta)$.

   iii) Take the derivative wrt **a**, $\frac{\partial}{\partial \mathbf{a}} \ln \mathbb{P}(\mathbf{t}|\mathbf{X}, \mathbf{a}, \beta)$.

   iv) Set the derivatite to 0, $\frac{\partial}{\partial \mathbf{a}} \ln \mathbb{P}(\mathbf{t}|\mathbf{X}, \mathbf{a}, \beta) = 0$.

   v) Solve to find the stationary point of **a**.

   vi) Check this stationary point is a maximum, if it is then it is a *Maximum Likelihood Estimate*

**Example 3.1 -** *Maximum Likelihood Estimate*
Here I shall find the *Maximum Likelihood Estimate* for $\mathbf{a}$

$$
\begin{aligned}
\mathbb{P}(\mathbf{t}|\mathbf{X},\mathbf{a},\beta) &\sim \prod_{i=1}^{m} \text{Normal}(t_i|\mathbf{a}^T\boldsymbol{\phi}(\mathbf{x}_i),\beta^{-1}) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\beta^{-1}}} e^{-\frac{1}{2}\beta(t_i-\mathbf{a}^T\boldsymbol{\phi}(x_i))^2} \\
&= \left(\frac{\beta}{2\pi}\right)^{\frac{m}{2}} e^{-\frac{\beta}{2}\sum_{i=1}^{m}(t_i-\mathbf{a}^T\boldsymbol{\phi}(x_i))^2} \\
\implies \ln\mathbb{P}(\mathbf{t}|\mathbf{X},\mathbf{a},\beta) &= \tfrac{m}{2}(\underbrace{\ln(\beta)}_{\text{Noise Precision}} - \underbrace{\ln(2\pi)}_{\text{Constant}}) - \frac{\beta}{2}\underbrace{\sum_{i=1}^{m}(t_i-\mathbf{a}^T\boldsymbol{\phi}(x_i))}_{\text{Error}} \\
\implies \tfrac{\partial}{\partial\mathbf{a}}\ln\mathbb{P}(\mathbf{t}|\mathbf{X},\mathbf{a},\beta) &= \beta\sum_{i=1}^{2}(\mathbf{t}_i-\mathbf{a}^T\boldsymbol{\phi}(\mathbf{x}_i))\boldsymbol{\phi}(x_i)^T \\
\text{Setting} \quad 0 &= \tfrac{\partial}{\partial\mathbf{a}}\ln\mathbb{P}(\mathbf{t}|\mathbf{X},\mathbf{a},\beta) \\
\implies 0 &= \beta\sum_{i=1}^{2}(\mathbf{t}_i-\mathbf{a}^T\boldsymbol{\phi}(\mathbf{x}_i))\boldsymbol{\phi}(x_i)^T \\
&= \left(\sum_{i=1}^{m}\mathbf{t}_i\boldsymbol{\phi}(\mathbf{x}_i)^T\right) - \mathbf{a}^T\left(\sum_{i=1}^{m}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T\right) \\
\implies \mathbf{a}_{MLE} &= \left(\boldsymbol{\phi}(\mathbf{X})^T\boldsymbol{\phi}(\mathbf{X})\right)^{-1}\boldsymbol{\phi}(\mathbf{X})^T\mathbf{t}
\end{aligned}
$$

**Theorem 3.1 -** *Variance of Posterior*
Let $\alpha$ be the parameter of the prior, $\beta$ be the parameter for the likelihood and $\mathbf{X}$ be the observed values from the predictor variables.

$$
\begin{aligned}
s_n &= (I\alpha + \beta\mathbf{X}^T\mathbf{X})^{-1} \\
&= \left(I\alpha + \beta\begin{pmatrix} \sum_{i=1}^{n}1 & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{pmatrix}\right)^{-1} \\
&= \begin{pmatrix} \beta n + \alpha & \beta\sum_{i=1}^{n}x_i \\ \beta\sum_{i=1}^{n}x_i & \alpha+\beta\sum_{i=1}^{n}x_i^2 \end{pmatrix}^{-1} \\
&= \frac{1}{(\beta n+\alpha)\left(\alpha+\beta\sum_{i=1}^{n}x_i^2\right) - \left(\beta\sum_{i=1}^{n}x_i\right)^2}\begin{pmatrix} \alpha+\beta\sum_{i=1}^{n}x_i^2 & -\beta\sum_{i=1}^{n}x_i \\ -\beta\sum_{i=1}^{n}x_i & \beta n+\alpha \end{pmatrix}
\end{aligned}
$$

Assume data is centred so $\sum_{i=1}^{n}x_i = 0$

$$
\begin{aligned}
&= \frac{1}{(\beta n+\alpha)\left(\alpha+\beta\sum_{i=1}^{n}x_i^2\right)}\begin{pmatrix} \alpha+\beta\sum_{i=1}^{n}x_i^2 & 0 \\ 0 & \beta n+\alpha \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{1}{\beta n+\alpha} & 0 \\ 0 & \dfrac{1}{\alpha+\beta\sum_{i=1}^{n}x_i^2} \end{pmatrix}
\end{aligned}
$$

**Theorem 3.2 -** *Mean of Posterior* Let $\alpha$ be the parameter of the prior, $\beta$ be the parameter for the likelihood, $\mathbf{X}$ be the observed values from the predictor variables and $\mathbf{t}$ be the observed

values for the dependent variable.

$$
\begin{aligned}
m_n &= (\alpha I + \beta \mathbf{X}^T \mathbf{X})^{-1} \beta \mathbf{X}^T \mathbf{t} \\
&= s_n \beta \mathbf{X}^T \mathbf{t} \\
&= \beta s_n \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \\
&= \beta s_n \begin{pmatrix} \sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} t_i x_i \end{pmatrix}
\end{aligned}
$$

Assume data is centred so $\sum_{i=1}^{n} x_i = 0$

$$
\begin{aligned}
&= \beta \begin{pmatrix} \dfrac{1}{\beta n + \alpha} & 0 \\ 0 & \dfrac{1}{\alpha + \beta \sum_{i=1}^{n} x_i^2} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} t_i x_i \end{pmatrix} \\
&= \begin{pmatrix} \dfrac{\beta \sum_{i=1}^{n} t_i}{\beta n + \alpha} \\ \dfrac{\beta \sum_{i=1}^{n} t_i x_i}{\alpha + \beta \sum_{i=1}^{n} x_i^2} \end{pmatrix}
\end{aligned}
$$

**Proposition 3.2 -** *Prediction*
Suppose we are given as inputs to the model: $\mathbf{X}$ observations from the predictor variables, $\mathbf{t}$ observations from the dependent variable; $\alpha$, parameter for the prior; and $\beta$, parameter for the likelihood.
If we now want to predict the value $\hat{t}$ at position $\hat{\mathbf{x}}$ we want to solve

$$
\mathbb{P}(\hat{t}|\hat{\mathbf{x}}, \mathbf{t}, \mathbf{X}, \alpha, \beta) = \int \mathbb{P}(\hat{t}|\hat{\mathbf{x}}, \mathbf{a}, \beta) \mathbb{P}(\mathbf{a}|\mathbf{t}|\mathbf{X}, \alpha, \beta) d\mathbf{a}
$$

where $\mathbf{a}$ is the coefficient for weighting each parameter.

## 3.2   Dual Linear Regression

**Definition 3.5 -** *Kernel*
A *Kernel* is a function that defines an inner-produce in some space.
Let $\mathbf{x}$ be a vector in the orignal space & $\phi(\cdot)$ map from the original space to the kernel space. Then the *Kernel Function* is defined as

$$
k(\mathbf{x}_i, \mathbf{x}_j) := \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)
$$

**Remark 3.5 -** *Usefulness of Kernels*
It is generally easier to define the inner-product of a space than to define a space & kernels allow us to never have to realise a space. This allows us to work with infinite dimensional spaces.
*N.B.* The space defined by the *Kernel* is called the *Induced Space*.

**Definition 3.6 -** *Kernel Regression*
*Kernel Regression* is the act of performing a linear regression in an *Induced Space*.
Let $\mathbf{X}$ & $\mathbf{t}$ be training data, $\lambda$ be a parameter for noise, $\hat{\mathbf{x}}$ be an unseen data point which we wish to predicit a value $\hat{y}$ for. Then

$$
\hat{y}(\hat{\mathbf{x}}) = k(\hat{\mathbf{x}}, \mathbf{x})(k(\mathbf{X}, \mathbf{X}) + \lambda I)^{-1} \mathbf{t}
$$

**Remark 3.6 -** *Usefulness of Kernel Regression*
Note that the problem is linear in the *Induced Space* but not in the original space, thus allowing us to learn non-linear functions using lines.

**Remark 3.7 -** *Changing Basis*
Some data is clearly non-linear & it may be helpful to transform it into another basis where linear regression is possible.
*e.g.* If the data appearas to fit a circle in a cartesian basis, it can be translated into polar co-ordinates which should be linear.
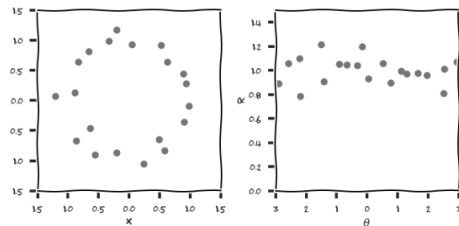


**Proposition 3.3 -** *Changing Basis*
Let $\mathbf{a}$ be weightings for observed parameters and $\mathbf{x}$ be a set of observed parameters.
Suppose we want to change the basis of $\mathbf{x}$, if we have a function $\phi : \mathcal{X} \to \mathcal{Z}$ which can do this then we predict $y$ as
$$t = \mathbf{a}^T \phi(\mathbf{x}) = \mathbf{a}^T \mathbf{z}$$

**Definition 3.7 -** *Dual Linear Regression*
Standard *Linear Regression* is defined as a linear combination of columns. *Dual Linear Regression* is a linear combination of the inner product of a new data point with each of the training data, this allows it to consider combinations of data points.

**Remark 3.8 -** *Intution of Dual Regression*
*Dual Regression* can be considered as describing an unseen data points as a combination of seen ones. *i.e.* Has the shape of a rhino, fur of a tiger, ...

**Proposition 3.4 -** *Dual Linear Regression Steps*
To perform a *Dual Linear Regression* perform the following

    i) Formulate Posterior, $\mathbb{P}(\theta|\mathbf{X})$;

    ii) Find stationary point of posterior;

    iii) Re-write the coefficients $\mathbf{a}$ in terms of the data;

    iv) Perform Kernel regression.

**Remark 3.9 -** *Useful Kernels*
Not all functions can be used as *Kernels*. Some that can, and can be useful,

    i) Kernelised Euclidean Distance $\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 = k(\mathbf{x}_i, \mathbf{x}_i) - 2k(\mathbf{x}_i, \mathbf{x}_j) + k(\mathbf{x}_j, \mathbf{x}_j)$

    ii) Exponented Quadratic $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 e^{-\frac{1}{2\ell^2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}$.

**Remark 3.10 -** *Limitations of Linear/Dual Regression*

    i) No uncertainty in our observed outputs;

    ii) No uncertainity in our mapping;

    iii) We have to make assumptions over the space of functions

## 3.3   Gaussian Processes

**Remark 3.11 -** *Motivation*
Here we want to introduce uncertainty into our observed outputs & mappings. This means that instead of outputting a discrete value we return a probability distribution. Now we can consider a few more features for our observations, such as how much does observing a value at $\mathbf{x}_0$ tell us about an observation at $\mathbf{x}_1$.
N.B. In the image before we have an observation for $x = -2$ and three marginals for $x = -2, 3, 6$ which our observation has a decreasing affec on, as distance increases.
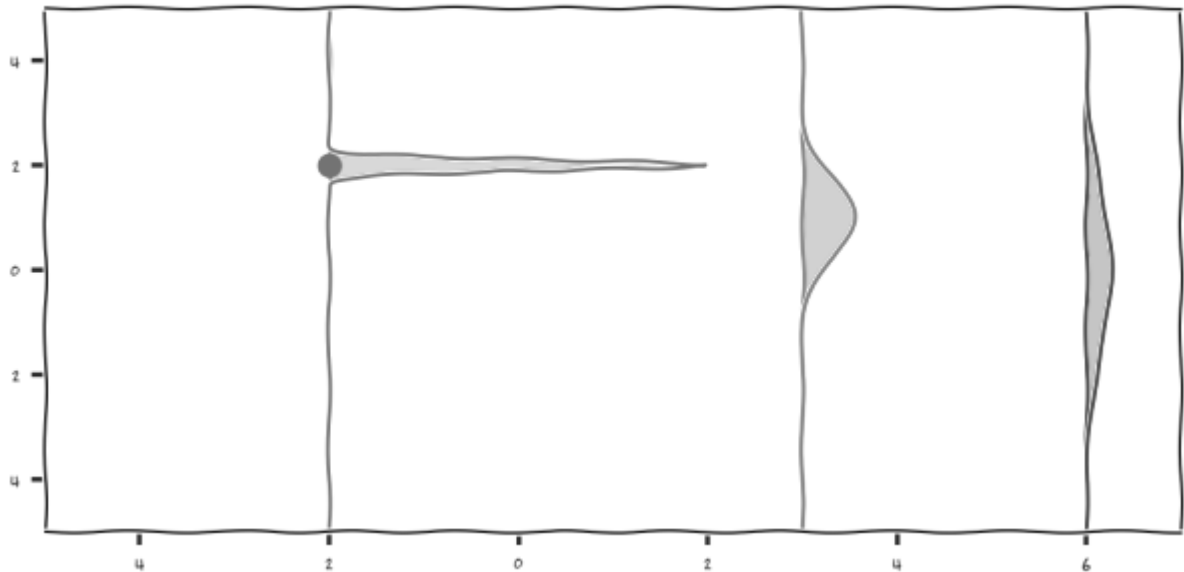


**Definition 3.8 -** *Gaussian Process*
A *Gaussina Process* is a generalisation of random variables into an infinite number of *Gaussian Distributions*. The specific process is defined by a mean function $\mu(\cdot)$ and a co-variance function function $k(\cdot)$.

$$\mathbb{P}(f_1, f_2, \ldots | \mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x})\right) = \text{Normal}\left(\begin{pmatrix} \mu(x_1) \\ \mu(x_2) \\ \vdots \end{pmatrix}, \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots \\ k(x_2, x_1) & k(x_2, x_2) & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}\right)$$

N.B. *Gaussian Processes* is non-parameteric.
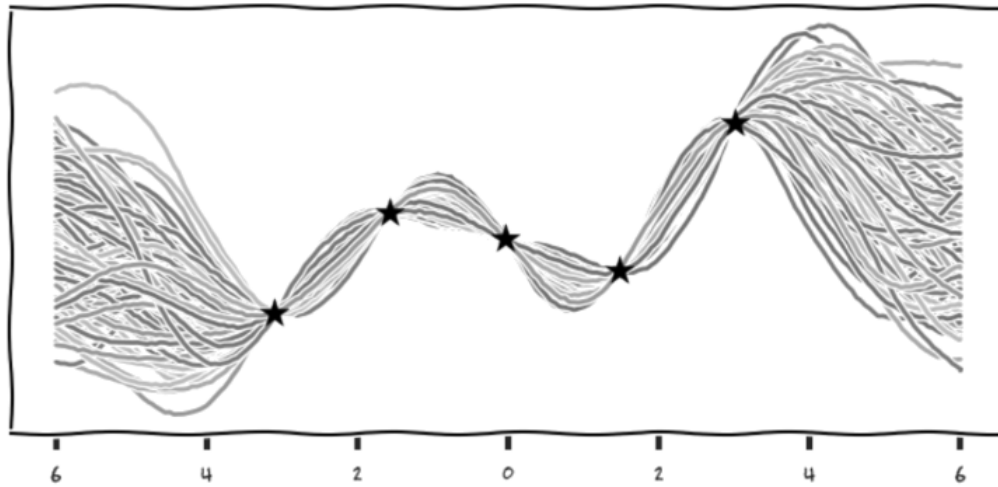
**Remark 3.12 -** *Covariance Function*
The *Covaraince Function* of a *Gaussian Process* defines how much an observation at $\mathbf{x}_0$ affects our prediction for $\mathbf{x}_1$. The greater the covariance values (Not on the main diagonal) the more an observation tells us. We can define the *Covariance Functions* to vary with distance & other factors.

$$\text{Very little effect} = \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}. \quad \text{A lot of effect} = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$$

**Remark 3.13 -** *Sampling from a Gaussian Process*
When we take a sample from a *Gaussian Process* we are given a function which fits the distributions defined the *Gaussian Process*.

**Example 3.2 -** *Sampling from a Gaussian Process*

**Proposition 3.5 -** *Gaussian Process - Posterior, No Noise*
Let $\mathbf{f}, \mathbf{X}$ be training data, $f^*, \mathbf{x}^*$ be training data and $k$ be the co-variance function. We have

$$\begin{pmatrix} \mathbf{f} \\ f^* \end{pmatrix} \quad \sim \quad \text{Normal}\left( \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} k(\mathbf{X}, \mathbf{X}) & k(\mathbf{X}, \mathbf{x}^*) \\ k(\mathbf{x}^*, \mathbf{X}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{pmatrix} \right)$$

$$\mathbb{P}(f^*|\mathbf{x}^*, \mathbf{X}, \mathbf{f}) \quad \sim \quad \text{Normal}\left( k(\mathbf{x}^*, \mathbf{X})^T k(\mathbf{x}^*, \mathbf{X})^{-1}\mathbf{f}, \ k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})^T k(\mathbf{X}, \mathbf{X})^{-1}k(\mathbf{X}, \mathbf{x}^*) \right)$$
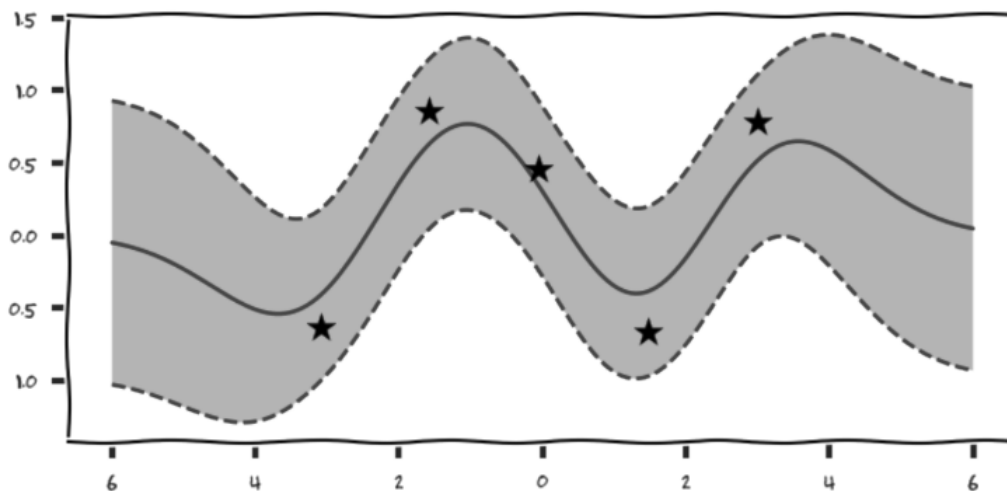
**Proposition 3.6 -** *Gaussian Process - Posterior, Noise*
Let $\mathbf{f}, \mathbf{X}$ be training data, $f^*, \mathbf{x}^*$ be training data and $k$ be the co-variance function.
Define $\mathbf{y}_i = f_i + \varepsilon$ where $\varepsilon \sim \text{Normal}(0, \sigma^2 I)$ We have

$$\begin{pmatrix} \mathbf{y} \\ f^* \end{pmatrix} \quad \sim \quad \text{Normal}\left( \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} k(\mathbf{X}, \mathbf{X}) + \sigma^2 I & k(\mathbf{X}, \mathbf{x}^*) \\ k(\mathbf{x}^*, \mathbf{X}) & k(\mathbf{x}^*, \mathbf{x}^*) \end{pmatrix} \right)$$

$$\mathbb{P}(f^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \sigma^2) \quad \sim \quad \text{Normal}\left( k(\mathbf{x}^*, \mathbf{x})^T (k(\mathbf{x}, \mathbf{x}) + \sigma^2 I)^{-1}\mathbf{y}, \ k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{x})^T (k(\mathbf{x}, \mathbf{x}) + \sigma^2 I)^{-1}k(\mathbf{x}, \mathbf{x}^*) \right)$$

**Example 3.3 -** *Noisy Gaussian Process*



# 4   Unsupervised Learning

**Definition 4.1 -** *Unsupervised Learning*
In *Unsupervised Learning* we are given only the ouput data & are tasked with deriving the
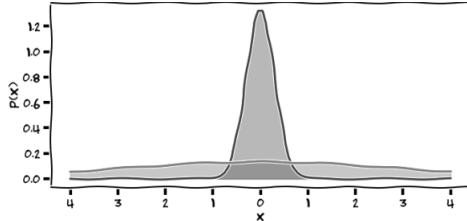
underlying properties of each event.

*i.e.* Given $y = f(x)$ recover both $f(\cdot)$ & $x$.

# 5  Bayesian Optimisation

**Remark 5.1 -** *Importance of Uncertainty*
Note that we can have $\hat{x} = argmax_x p(x) = \text{argmax}_x q(x)$ but $p(\hat{x}) \neq q(\hat{x})$. Due to this we need to introduce uncertainty about the value of $x$ so that we can weight out outcomes accordingly.



**Definition 5.1 -** *Optimisation*
*Optimisation* is the process of finding the best outcome for a problem.

**Remark 5.2 -** *Optimisation*
Classically *Optimisation* is seen as $\hat{x} = \text{argmin}_x f(x)$ however we typically have an objective function that we do not know explicitly, but are able to test. Since testing in real life situations (*e.g.* Medical Trials) are expensive we want to minimise the number required to achieve a good level of certainty.

**Definition 5.2 -** *Global Optimisation*
*Global Optimisation* is the set of techniques for finding $x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$ where $\mathcal{X}$ is a bounded domain (reducing the amount of testing required) and $f$ is not known explicitly. It is possible that evaluations of $f$ are noisy.

**Proposition 5.1 -** *Bayesian Optimisation*

   i) Choose a prior over the space of possible objective functions, $f$.

   ii) Combine the prior & likelihood to get a posterior over the space.

   iii) Use the posterior to choose a set of evaluations according to a strategyive .

   iv) Add new data, update posterior & re-evaluate.

   v) Repeat until budget is gone

**Proposition 5.2 -** *Naïve Strategies for Global Optimisation*
Below are some naïve *Global Optimisation* stategies

   i) Implicity knowledge, ask a SME;

   ii) Grid Search, test the domain at regular intervals; and,

   iii) Random Sampling.

**Remark 5.3 -** *Exploration v Exploitation*
When considering strategies for *Bayesian Optimisation* we need to consider how we approach *Exploration* (Testing new areas) & *Explotation* (Investigating areas which seem good). A *Acquisition Function* can be defined which returns the expected gain in information if a sample was to be taken at $x$.
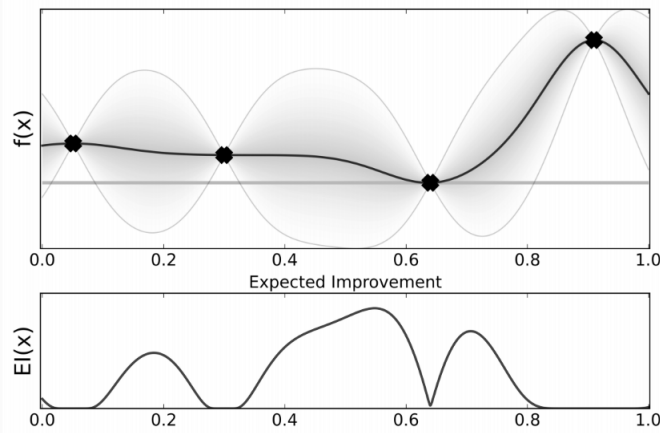
**Definition 5.3 -** *Expected Improvement*
*Expected Improvement* is an *Acquistion Function*.
Let $\mathbf{x}$ be a point in the domain, $\theta$ be parameters of the distribution and $X$ be data we already know. Then

$$EI(\mathbf{x};\theta,X) := \int \max(0, y_{best} - y)\mathbb{P}(y|\mathbf{x},\theta,X)dy$$

We should then sample at $\mathbf{x}$ where $EI(\mathbf{x}) \geq EI(\mathbf{x}') \ \forall \ x' \in \mathcal{X}$ since this point offers the greatest information gain.
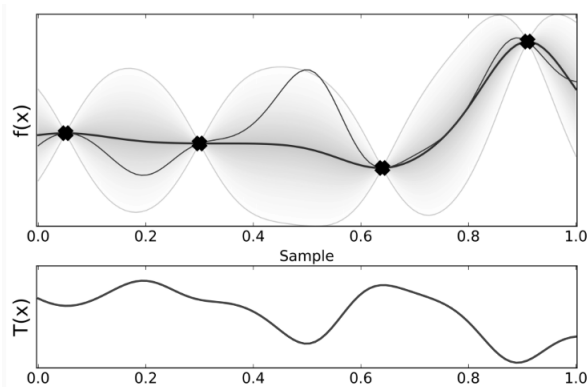*N.B.* $y_{best}$ is the best value for $y$ we have found so far. *i.e.* The value we are trying to improve.



**Definition 5.4 -** *Thomson Sampling*
*Thomson Sampling* is an *Acquistion Function*.
Let $\mathbf{x}$ be a point in the domain, $\theta$ be parameters of the distribution and $X$ be data we already know. Then

$$T(\mathbf{x};\theta,X) := \mathbb{P}(y|\mathbf{x},\mathbf{x},\theta,X)$$

# 0   Appendix

## 0.1   Definitions

**Definition 0.1 -** *Memory-Based Methods*
*Memory-Based Methods* for classification store the entire training set in order to make predictions for future data points. *e.g.* Nearest-Neighbours.
*N.B.* These generally require a distance measure to be defined.

## 0.2   Proofs

**Proof 0.1 -** *Deriving Gaussian Marginal Distribution*
*NOTE - This is dense as fuck & uses quite a bit of bullshit.*

Let $\mathbf{X} \sim \mathcal{N}\left( \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}^{-1} \right)$.

$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ can be considered as two parts of the mean vector $\boldsymbol{\mu}$.
Let $\mathbf{x}$ be a realisation of $\mathbf{X}$ where $\mathbf{x} := (\mathbf{x}_1, \mathbf{x}_2)$ with $\mathbf{x}_1$ & $\mathbf{x}_2$ representing the same partition as $\boldsymbol{\mu}_1$ & $\boldsymbol{\mu}_2$ respecitvely.
Define $D := \dim(\mathbf{x})$, $D_1 := \dim(\mathbf{x}_1)$ & $D_2 := \dim(\mathbf{x}_2)$.

Here we want to get from $\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2)$ to $\mathbb{P}(\mathbf{x}_1)$.
Consider the exponent of the joint distribution

$$E = -\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T \Lambda_1 1(\mathbf{x}_1-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T \Lambda_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2) - \frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T \Lambda_{21}(\mathbf{x}_1-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^T \Lambda_{22}(\mathbf{x}_2-\boldsymbol{\mu}_2)$$

To produce the marginal for $x_1$ we want to isolate the terms involving $x_2$ so they are easy to remove.

$$
\begin{aligned}
E &= -\tfrac{1}{2}\Bigg[ \left( \mathbf{x}_2^T \Lambda_{22}\mathbf{x}_2 - 2\mathbf{x}_2^T \Lambda_2 2(\boldsymbol{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{x}_1 - \boldsymbol{\mu}_1)) \right) \\
&\quad - 2\mathbf{x}_1^T \Lambda_{12}\boldsymbol{\mu}_2 + 2\boldsymbol{\mu}_1^T \Lambda_{12}\boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T \Lambda_{22}\boldsymbol{\mu}_2 + \mathbf{x}_1^T \Lambda_{11}\mathbf{x}_1 \\
&\quad - 2\mathbf{x}_1^T \Lambda_1 1\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \Lambda_{11}\boldsymbol{\mu}_1 \Bigg] \\
&= \underbrace{-\frac{1}{2}\left( \mathbf{x}_2 - (\boldsymbol{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{x}_1 - \boldsymbol{\mu}_1)) \right)^T \Lambda_{22} \left( \mathbf{x}_2 - (\boldsymbol{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbf{x}_1 - \boldsymbol{\mu}_1)) \right)}_{E_1} \\
&\quad + \underbrace{\frac{1}{2}\left( \mathbf{x}_1^T \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mathbf{x}_1 - 2\mathbf{x}_1^T \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1 \right)}_{A} \\
&\quad - \underbrace{\frac{1}{2}\left( \mathbf{x}_1^T \Lambda_1 1\mathbf{x}_1 - 2\mathbf{x}_1^T \Lambda_{11}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1\Lambda_{11}\boldsymbol{\mu}_1 \right)}_{B}
\end{aligned}
$$

Note that $A$ & $B$ do not contain any $x_2$ terms.
Since the co-variance matrix is symmetric we have $\Lambda_{12} = \Lambda_{21}^T$ we have

$$\mathbf{x}_1^T \Lambda_{12}\boldsymbol{\mu}_2 = \mathbf{x}_1^T \Lambda_{21}^T \boldsymbol{\mu}_2 = (\Lambda_{21}\mathbf{x}_1)^T \boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^T \Lambda_{21}\mathbf{x}_1$$

We shall not rewrite $A$ & $B$ as quadratic expressions

$$
\begin{aligned}
A &= \frac{1}{2}\left(\mathbf{x}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\mathbf{x}_1 - 2\mathbf{x}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\boldsymbol{\mu}_1\right) \\
&= \frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})(\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
B &= \frac{1}{2}\left(\mathbf{x}_1^T\Lambda_{11}\mathbf{x}_1 - 2\mathbf{x}_1^T\Lambda_{11}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1\Lambda_{11}\boldsymbol{\mu}_1\right) \\
&= \frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T\Lambda_{11}(\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
\implies A - B &= \frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T(\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21} - \Lambda_{11})(\mathbf{x}_1 - \boldsymbol{\mu}_1) \\
\text{Let } E_2 &:= A - B
\end{aligned}
$$

Now the exponent has been orgainised we can consider the whole gaussian expression.

$$
\begin{aligned}
\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) &= \frac{e^{E_1}e^{E_2}}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \\
\mathbb{P}(\mathbf{x}_1) &= \int \mathbb{P}(\mathbf{x}_1, \mathbf{x}_2)d\mathbf{x}_2 \\
&= \int \frac{e^{E_1}e^{E_2}}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}d\mathbf{x}_2 \\
&= \frac{e^{E_2}}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}\int e^{E_1}d\mathbf{x}_2 \quad \text{Since } E_2 \text{ is independent of } \mathbf{x}_2
\end{aligned}
$$

Now we consider $\int e^{E_1}d\mathbf{x}_2$.
Since we know a gaussian must intergrate to 1 over the whole domain we deduce that

$$
\begin{aligned}
\int \frac{1}{(2\pi)^{\frac{D_2}{2}}|\Lambda_{22}^{-1}|^{\frac{1}{2}}}e^{E_1}d_{\mathbf{x}_2} &= 1 \\
\implies \int e^{E_1}d\mathbf{x}_2 &= (2\pi)^{\frac{D_2}{2}}|\Lambda_{22}^{-1}|^{\frac{1}{2}}
\end{aligned}
$$

N.B. $\Lambda_{22}^{-1}$ is the variance of $\mathbf{x}_2$.
Using the result of this intergal we have

$$
\begin{aligned}
\mathbb{P}(\mathbf{x}_1) &= (2\pi)^{\frac{D_2}{2}}|\Lambda_{22}^{-1}|^{\frac{1}{2}}\frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}}e^{E_2} \\
&= \frac{e^{E_2}}{(2\pi)^{\frac{D-D_2}{2}}|\Lambda_{22}^{-1}|^{-\frac{1}{2}}|\Sigma|^{\frac{1}{2}}}
\end{aligned}
$$

The Schur complement of $\Lambda_{22}$ is $\Lambda_{22}^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.
Thus
$$
\begin{aligned}
|\Lambda_{22}^{-1}|^{-\frac{1}{2}}|\Sigma|^{\frac{1}{2}} &= |\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{-\frac{1}{2}}|\Sigma_{11}|^{\frac{1}{2}}|\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}|^{\frac{1}{2}} \\
&= |\Sigma_{11}|^{\frac{1}{2}}
\end{aligned}
$$

Now we have a full expression

$$
\begin{aligned}
\mathbb{P}(\mathbf{x}_1) &= \frac{e^{E_2}}{(2\pi)^{\frac{D-D_2}{2}}|\Lambda_{22}^{-1}|^{-\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{D_1}{2}}|\Sigma_{11}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu}_1)^T\Sigma_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)}
\end{aligned}
$$

■

**Proof 0.2 -** *Deriving Gaussian Conditional Distribution*

Let $\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$.

$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ can be considered as two parts of the mean vector $\boldsymbol{\mu}$.

Let $\mathbf{x}$ be a realisation of $\mathbf{X}$ where $\mathbf{x} := (\mathbf{x}_1, \mathbf{x}_2)$ with $\mathbf{x}_1$ & $\mathbf{x}_2$ representing the same partition as $\boldsymbol{\mu}_1$ & $\boldsymbol{\mu}_2$ respecitvely.

Define $D := \dim(\mathbf{x})$.

We want to find the distribution of $\mathbb{P}(\mathbf{x}_1|\mathbf{x}_2)$.

From the product rule we know that $\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) = \mathbb{P}(\mathbf{x}_1|\mathbf{x}_2)\mathbb{P}(\mathbf{x}_2)$ and we already know the joint & marginal distributions for a gaussian.

We have that

$$\mathbb{P}(\mathbf{x}_1, \mathbf{x}_2) \propto e^{-\frac{1}{2}\begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}}$$

We now want to factor the marginal distribution out of this expression.

$$\mathbb{P}(\mathbf{x}_2) \propto e^{-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)}$$

Lets look at the exponent of the joint distribution.

*N.B.* About to use a lot of Schur Complements

$$
\begin{aligned}
E &= -\frac{1}{2}\begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \\
&= -\frac{1}{2}\begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}^T \begin{pmatrix} I & 0 \\ \Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix}^T \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \\
&= -\frac{1}{2}\begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}^T \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & -(\Sigma/\Sigma_{22})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{21}\Sigma_{22}^{-1}(\Sigma/\Sigma_{22})^{-1} & \Sigma_{22}^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \\
&= -\frac{1}{2}\left[\mathbf{x}_1 - (\boldsymbol{\mu}_1 + \Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))\right]^T (\Sigma/\Sigma_{22})^{-1}\left[\mathbf{x}_1 - (\boldsymbol{\mu}_1 + \Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))\right] \\
&\underbrace{-\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)}_{E_2}
\end{aligned}
$$

Note that $E_2$ is exactly the exponent for the marginal distribution of $\mathbf{x}_2$ and thus what we want to factory out in order to get to the conditional distribution.

$$\mathbb{P}(\mathbf{x}_1|\mathbf{x}_2) \propto e^{-\frac{1}{2}\left[\mathbf{x}_1 - \underbrace{(\boldsymbol{\mu}_1 + \Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))}_{\text{mean}}\right]^T (\underbrace{\Sigma/\Sigma_{22}}_{\text{covariance}})^{-1}\left[\mathbf{x}_1 - (\boldsymbol{\mu}_1 + \Sigma_{21}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))\right]}$$

$\square$

## 0.3   Remarks

**Remark 0.1 -** *Worlds*

We can consider 3 different when answering an ml question.

    i) Deterministic, $x = 4$;

   ii) Point Estimate, $\text{argmax}_x p(x) = 4$;

  iii) Stochastic, $p(x) \sim \text{Normal}(4, 10^2)$.