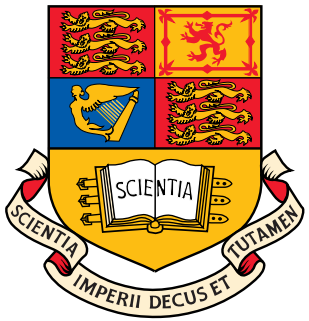
# Statistical Signal Processing & Inference
## BLUE and Maximum Likelihood Est.


**Danilo Mandic**

**room 813, ext: 46271**

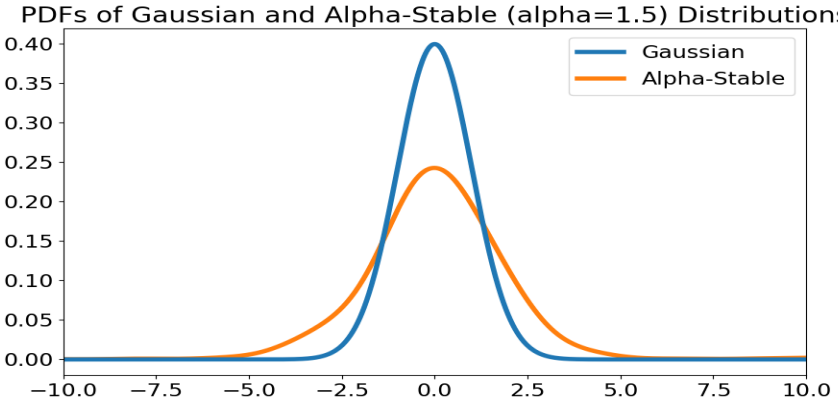


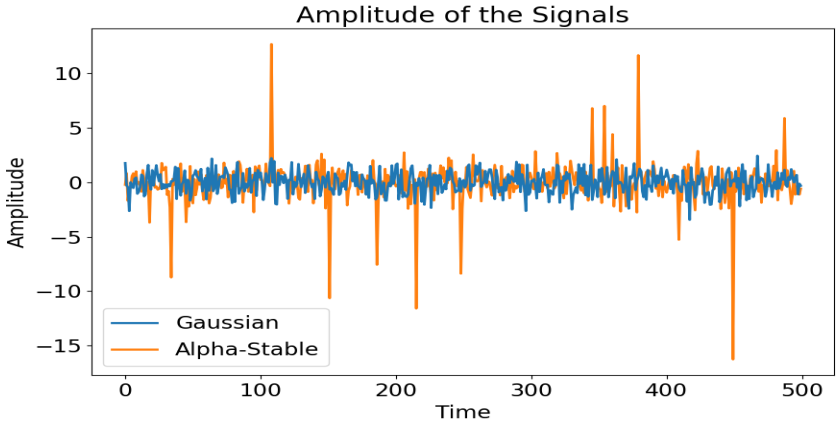

Department of Electrical and Electronic Engineering
Imperial College London, UK

d.mandic@imperial.ac.uk, URL: www.commsp.ee.ic.ac.uk/∼mandic

# Motivation for Best Linear Unbiased Estimator (BLUE) and Maximum Likelihood Estimation (MLE)

- In many applications, signals exhibit sharp spikes
- This results in heavy-tailed distributions (e.g. $\alpha$-stable distributions)
- There may not be a general form of pdf for such distributions




- If an **efficient estimator does not exist**, it is still of interest to be **able to find** an MVU estimator (assuming, of course, that it exists), as in BLUE
- To achieve this, we need the concept of **sufficient statistics** and the Rao–Blackwell–Lehmann–Scheffe theorem

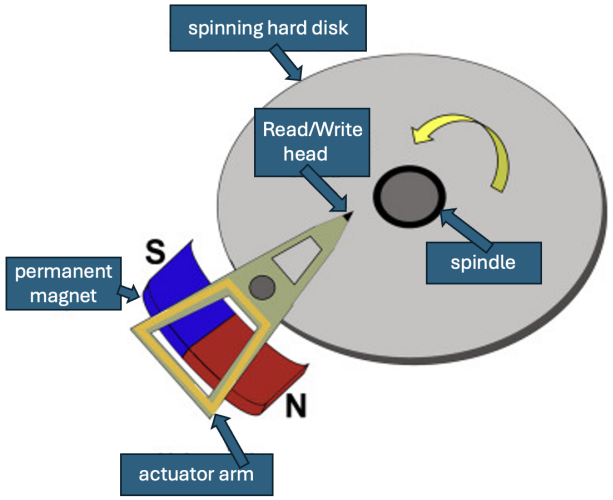
☞ The BLUE assumptions are also called the **Gauss–Markov assumptions**

- It is possible in many cases to determine an approximate MVU estimator (MVUE) by inspection of the PDF, using Maximum Likelihood Estimation

# Motivation for Maximum Likelihood Estimation (MLE)

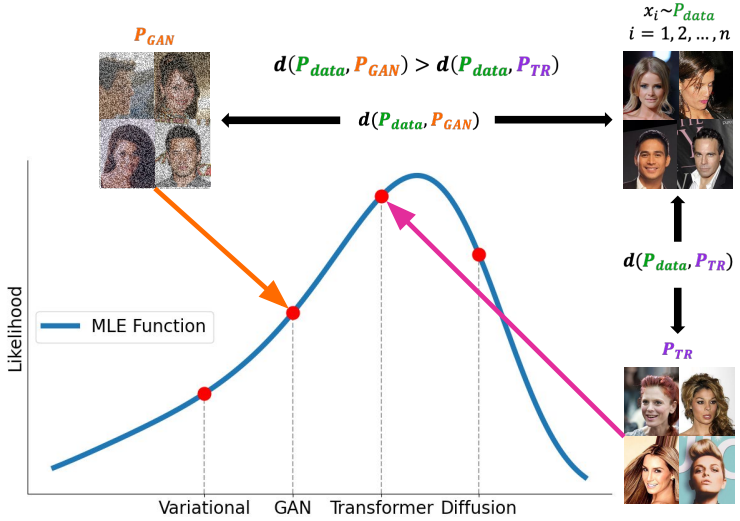
## What do you think these applications have in common?

### HDD microcontroller

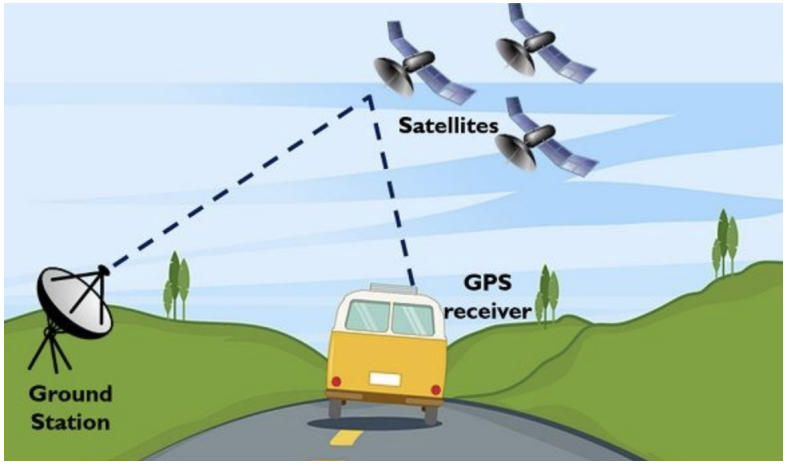



### Gesture (Swype) keyboard

$$P(S) = p(I|begin)\, p(would|I)\, p(like|would) \times$$
$$\times\, p(to|like)\, p(fly|to)\, p(London|fly) \cdots$$




### Generative AI (density estimation)




### Global Positioning System

# Overview

- It frequently occurs that the MVU estimator, even if it exists, cannot be found (mathematical tractability, violation of regularity conditions, ...)
- For instance, one typical case is that we may not know the pdf of the data, but we do know the 1st and 2nd moment (mean, variance, power). **In such cases pdf based methods cannot be applied**
- We therefore have to resort to **suboptimal solutions** $\looparrowright$ by imposing some constraints (domain knowledge) on the estimator and data model
- If the variance of a suboptimal estimator meets our system specifications, the use of such estimators is fully justified
- The best linear unbiased estimator (BLUE) $\looparrowright$ restricts the estimator to be **linear in the data** $\looparrowright$ finds a **linear estimator** that is **unbiased and has the minimum variance among such unbiased estimators**
- Alternatively, if the MVU estimator does not exist or BLUE is not applicable, we may resort to **Maximum Likelihood Estimation (MLE)**
- We first need to look at which data samples are pertinent to the estimation problem in hand $\rightsquigarrow$ the so called **sufficient statistics**

## The notion of a statistic

**Def:** Any real valued function, $T(\mathbf{x}) = f\big(x[0], x[1], \ldots, x[N-1]\big)$, of the observations in the sample space, $\{x\}$, is called a statistic. Importantly, there **should not be** any unknown parameter, $\theta$, in a statistic.

☞ The mean $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$, median, and $\max\{x[0], x[1], \ldots, x[N-1]\}$ are all statistics. However, $x[0] + \theta$ is not a statistic if $\theta$ is unknown.

Let us now reflect on the estimators we have considered so far:

$$\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1}(x[n] - \mu)^2 \quad \max\{x[0], \ldots, x[N-1]\}$$

☞ These estimators are a function of random observations, $\mathbf{x}$, and not of the unknown paramater, $\theta$ $\looparrowright$ each of these **estimators is a valid statistic.**

◦ Observe that the above estimators "compress" the available information, e.g. the sample mean takes $N$ datapoints in $\mathbf{x}$ and produces one sample, $\bar{x}$.

◦ In the best case, such compression is "loss-less", as it contains the same amount of information as that contained in the $N$ original observations, $\mathbf{x}$.

☞ We call such statistic a **sufficient statistic**, as it **summarises** (absorbs) all information about an unknown parameter, $\theta$, and reduces "data footprint".

# An insight into the 'sufficiency' of the data statistics

**Which data samples are pertinent to the est. problem? Q: ∃ a sufficient dataset?**

Consider two unbiased estimators of a DC level in WGN:

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n], \quad \text{var}(\hat{A}) = \frac{\sigma^2}{N} \qquad \& \qquad \tilde{A} = x[0], \quad \text{var}(\tilde{A}) = \sigma^2$$

☞ Although $\tilde{A}$ is **unbiased**, its variance is much larger than that of $\hat{A}$. This is is due to discarding samples $x$[1],...,$x$[N-1] that carry information about $A$.

**Consider now the following datasets:**

$$S_1 = \{x[0], x[1], \ldots, x[N-1]\} \quad S_2 = \{x[0]+x[1], x[2], \ldots, x[N-1]\} \quad S_3 = \{\sum_{n=0}^{N-1} x[n]\}$$

The original dataset, $S_1$, is **always sufficient** for finding $\hat{A}$, while $S_2$ and $S_3$ are also sufficient. In addition, $S_3$ is the **minimal sufficient statistic!**

**In a nutshell, sufficient statistics answer the questions:**

Q1: Can we find a transformation $T(\mathbf{x})$ of lower dimension that **contains all information** about $\theta$? (the data, $\mathbf{x} \in \mathbb{R}^{N\times 1}$, can be very long)

Q2: What is the lowest possible dimension of $T(\mathbf{x})$ which still contains all information about $\theta$? $\quad \looparrowright \quad$ **(minimal sufficient statistic)**

For example, for DC level in WGN, $T(\mathbf{x}) = \sum x[n]$ **(one-dimensional)**

## Intuition about a sufficient statistic

Denote by $x$ the video recording of your SSPI Lecture 4 (a dataset, $\mathbf{x}$), and by $y$ the notes you have taken about Lecture 4 (a statistic, $T(\mathbf{x})$).

The information needed to answer Assignment #3 in your Coursework is the unknown parameter $\theta$.

- Now, $y$ depends entirely on $x$ $\quad \looparrowright$ video contains all info in your notes.
- If you took sufficiently good notes, $T(\mathbf{x})$ will give you same information about $\theta$ as $\mathbf{x}$ does $\looparrowright$ conditional distrib. of Lecture 4, given your notes, $p(x|y;\theta)$, is independent of $\theta$ (the information related to Assignment #3).
- Here, conditional distribution simply means the probability distribution of the information in your notes, given the lecture $\looparrowright$ if the information is in the lecture, it is also in your notes.

☞ Once you have checked your notes, going back and listening to the lecture will not help you solve Assignment #3 (no additional information).

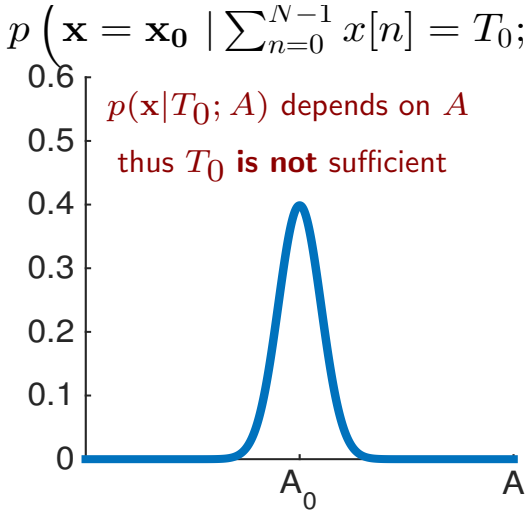
☞ A consideration of both the data set $\mathbf{x}$ and the statistic $T(\mathbf{x})$ does not give any more information about the distribution of $\theta$, than what is available based only on the statistic $T(\mathbf{x})$ $\looparrowright$ so, we can keep $T(\mathbf{x})$ and "throw away" $\mathbf{x}$ without losing any information. (see next Slide)

# Sufficient statistics: Putting it all together

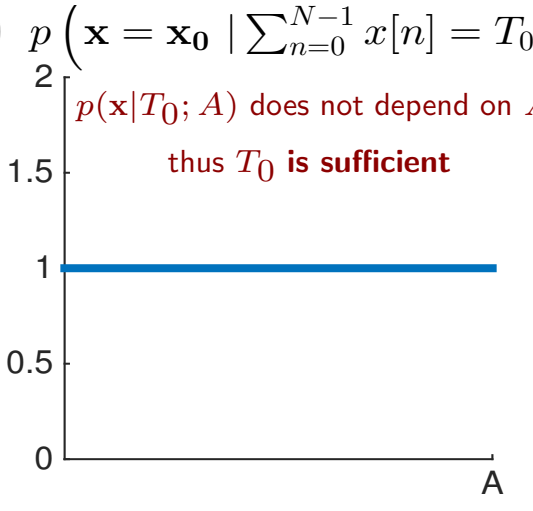
**In layman's terms, sufficiency is saying that $T(\mathbf{x})$ is informative enough**

**Aim:** A statistic $T(\mathbf{x})$ is sufficient if it allows us to estimate the unknown par. $\theta$ as well as when based on the entire data set $\mathbf{x}$. So, we no longer need to consider data, $\mathbf{x}$, after using it to calculate $T(\mathbf{x})$, it becomes redundant.

**Def:** A statistic $T(\mathbf{x}) = f\big(x[0], x[1], \ldots, x[N-1]\big)$ is a sufficient statistic, if for any value of $T_0$ the conditional distribution of $x[0], x[1], \ldots, x[N-1]$ given $T = T_0$, that is, $p(\mathbf{x} = \mathbf{x}_0 | T(\mathbf{x}) = T_0; \theta)$, does not depend on the unknown parameter $\theta$. In other words, after observing the statistic $T_0$, the data $\mathbf{x}$ will not give us any new information about $\theta$. (see also Appendix 1)




**Information present** in observations after $\mathrm{T}(\mathbf{x})$ observed



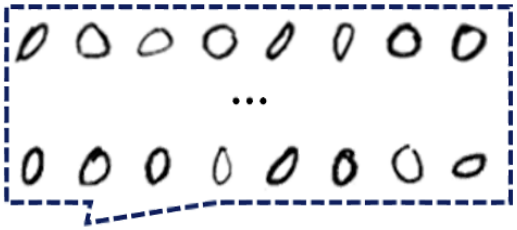

**No information** in observations after $\mathrm{T}(\mathbf{x})$ observed

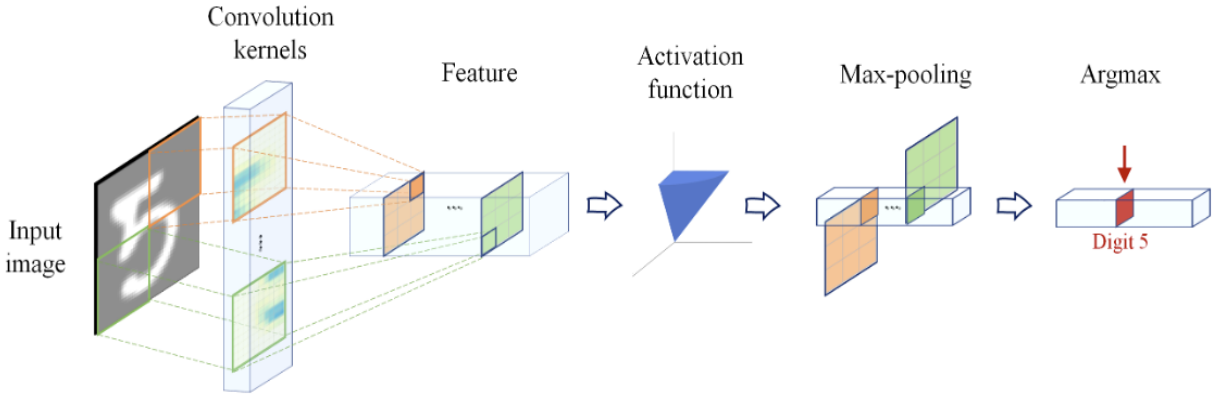
- Knowledge of $T_0$ changes the PDF to the conditional one $p(\mathbf{x}| \sum_{n=0}^{N-1} x[n] = T_0; A)$.
- **If a statistic is sufficient for estimating A, this conditional PDF should not depend on A** (as in the right hand panel).

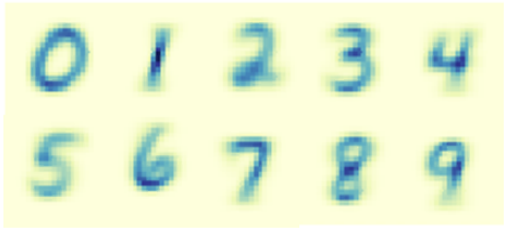
# Some intuition: Efficient use of training data in NNs

## More on "sufficient statistic": Kernel initialisation in Convolutional NNs (CNN)x

## Matched filter CNN for digit recognition

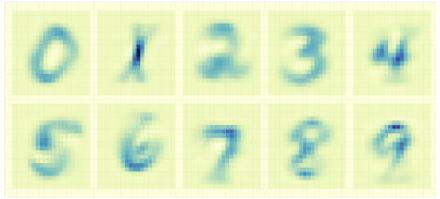




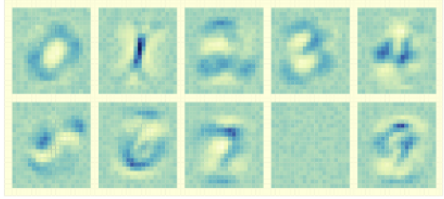



**Kernels after convergence**

**Trained kernels: MF Init.**



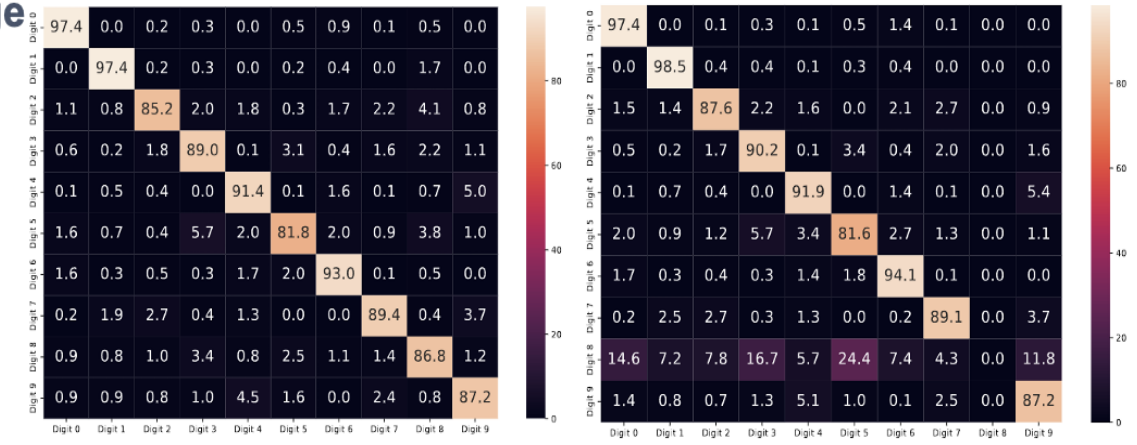
**Trained kernels: Random Init.**



**We pre-initialise with domain knowledge** (average digit over all available samples)

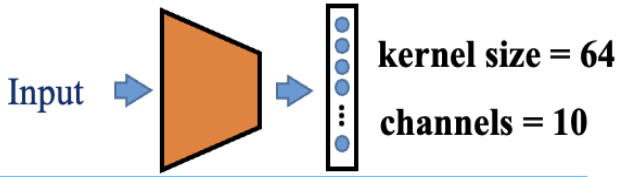
**Left:** Trained kernels of Matched Filter CNN, **Accuracy = 90%**

**Right:** Trained kernels after random initialization, **Accuracy = 82%**



# More intuition: CNNs for British sign language

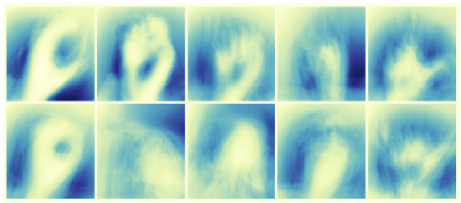
## Our own work: Domain-Informed CNNs



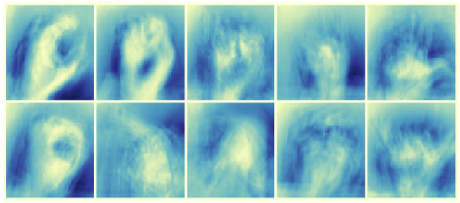

Sign Language

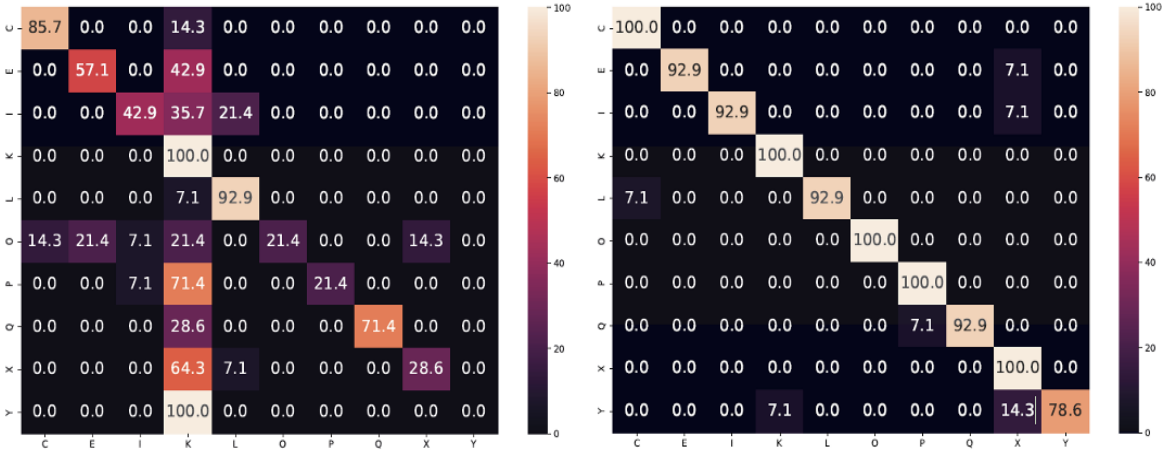
10 Letters: C, E, I, K, L, O, P, Q, X, Y


MF: Pre-initialisation



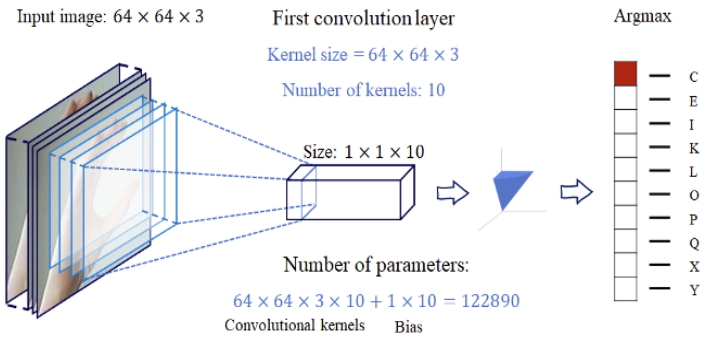
Training by MF pre-init.









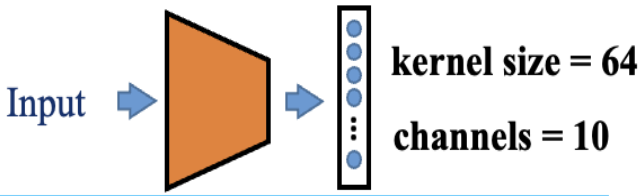



**No training just MF:**
Acc. = 52%

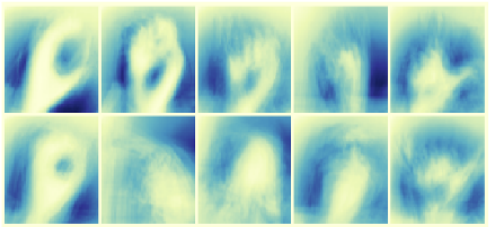
**After training from MF:**
Acc. = 95%

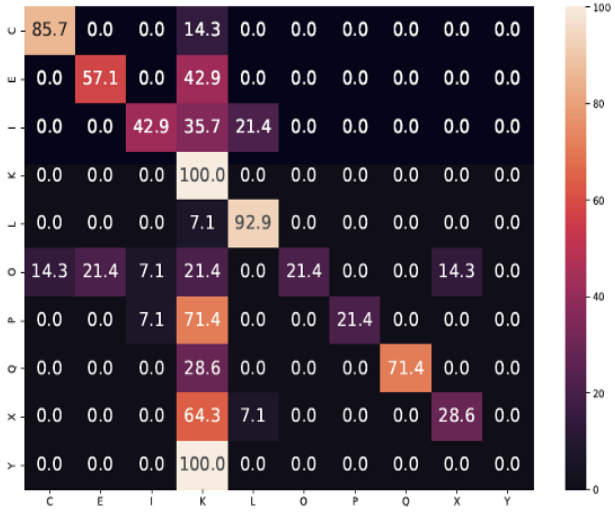
# Efficient use of training data: Initialisation of CNNs through domain knowledge

## CNNs as matched filters: Advantages



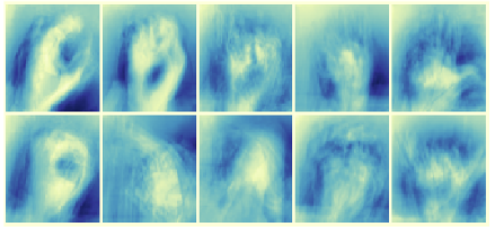

Pre-initialisation

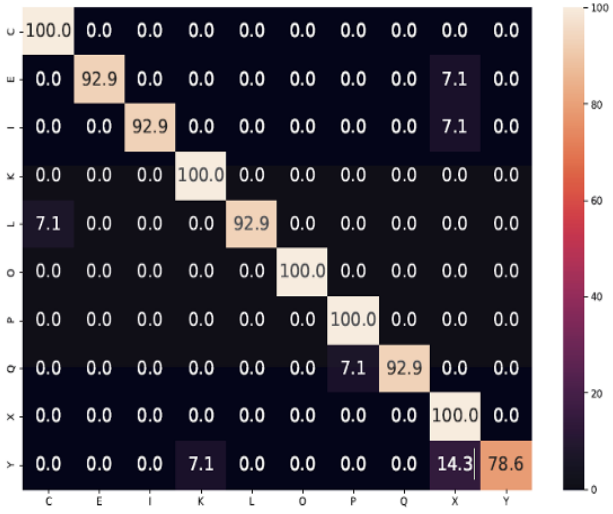





Acc. = 52%

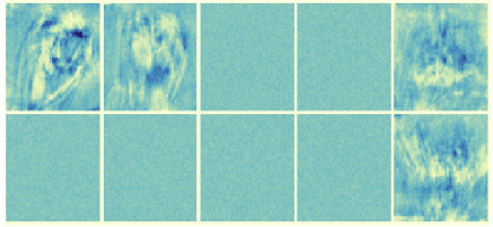
Training by pre-initialisation

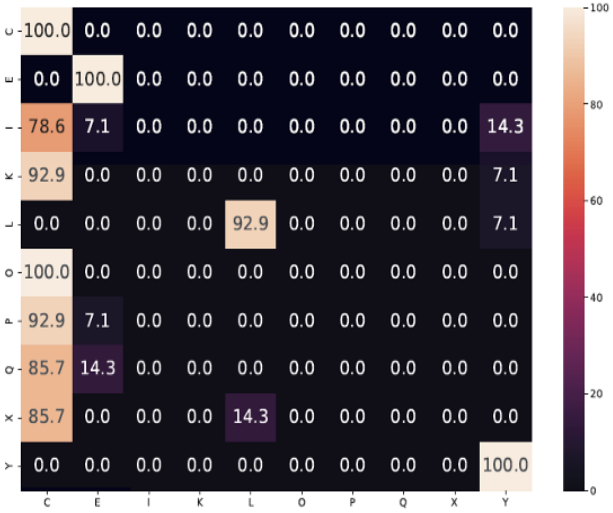





Acc. = 95%

Training by random initialisation






Acc. = 39%

# Sufficient statistics: Neyman-Fisher factorisation

**Recall: The Gaussian** $p(\mathbf{x};A)=\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\}$

Finding the conditional distribution $p(\mathbf{x}|\sum_{n=0}^{N-1}x[n]=T_0;A)$ can be extremely difficult. An intuitive way to deal with this is through the factorisation of $p(\mathbf{x}|T(\mathbf{x});A)$.

**Th: Neyman-Fisher factorisation.** Consider a set of random samples, $\mathbf{x}$, with a PDF $p(\mathbf{x};\theta)$ which depends on the unknown parameter $\theta$. Then, the statistic $T(\mathbf{x})$ is sufficient for $\theta$ iff the PDF can be factored as

$$p(\mathbf{x};\theta)=g\big(T(\mathbf{x}),\theta\big)h(\mathbf{x}) \qquad = \text{g (parameters \& data)} \times \text{h (data only)}$$

where $g(\cdot)$ depends on $\mathbf{x}$ only through $T(\mathbf{x})$, and $h$ is a function of only $\mathbf{x}$.

☞ **For a DC level in WGN**

$$p(\mathbf{x};A)=\frac{1}{(2\pi\sigma^2)^{N/2}}\underbrace{\exp\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\}}_{h(\mathbf{x})}\underbrace{\exp\Big\{-\frac{1}{2\sigma^2}\big[NA^2-2A(\overbrace{\sum_{n=0}^{N-1}x[n]}^{T(\mathbf{x})})\big]\Big\}}_{g(T(\mathbf{x}),A)}$$

Therefore, the sufficient statistic is $T(\mathbf{x})=\sum_{n=0}^{N-1}x[n]$ (minimal & linear)

☞ Any 1-2-1 mapping of $T(\mathbf{x})$ is also a sufficient st. e.g. $T_1(\mathbf{x})=\frac{1}{N}\sum_{n=0}^{N-1}x[n]$ and $T_2(\mathbf{x})=(\bar{x})^3$ are sufficient, but $T_3(\mathbf{x})=(\bar{x})^2$ is not as $\bar{x}=\pm\sqrt{T_3(\mathbf{x})}$.

# More examples of sufficient statistics ↬ Estimating the power of white Gaussian noise

Consider a parametrised PDF for DC level estimation in WGN, given by

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2 \Big\}$$

where $A$=0 and the noise power, $\sigma^2$, is the unknown parameter.

☞ To find the sufficient statistic for the estimation of $\sigma^2$, we factorise

$$p(\mathbf{x}; \sigma^2) = \underbrace{\frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big\{ -\frac{1}{2\sigma^2} \overbrace{\sum_{n=0}^{N-1} x^2[n]}^{T(\mathbf{x})} \Big\}}_{g(T(\mathbf{x}), \sigma^2)} \times \underbrace{1}_{h(\mathbf{x})}$$

This gives the sufficient statistic for the estimation of the unknown $\sigma^2$ as $T(\mathbf{x}) = \sum_{n=0}^{N-1} x^2[n]$ which, of course, makes perfect physical sense.

**Homework:** Prove that $\sum_{n=0}^{N-1} x^2[n]$ is a sufficient statistic for $\sigma^2$ by using the definition that $p(\mathbf{x} \,|\, \sum_{n=0}^{N-1} x^2[n] = T_0; \sigma^2)$ does not depend on $\sigma^2$ (no information left in observation after $T_0$ is observed). (see Slide 8)

# How to find the MVU from a sufficient statistic?

Raw data $\mathbf{x} = [x[0], \ldots, x[N-1]]^T \in \mathbb{R}^{N\times 1} \not\hookrightarrow$ an $N$-dim. sufficient statistic

**Neyman-Fisher Th.:** For $T(\mathbf{x})$ to be a sufficient statistic, we need to be able to factor $p(\mathbf{x};\theta)$ as $\quad p(\mathbf{x};\theta) = g\big(T(\mathbf{x}),\theta\big)h(\mathbf{x})$

## Finding the MVU from sufficiency

We can do this in two ways:

1. Find **any unbiased estimator**, say $\tilde{\theta}$, of $\theta$ and determine

$$\hat{\theta} = E[\tilde{\theta}|T(\mathbf{x})]$$

(often mathematically intractable)

2. Find a $g(\cdot)$, s.t. $\hat{\theta} = g(T(\mathbf{x}))$ is an unbiased estimator of $\theta$

(preferable in practice)

**Then:** If $g(\cdot)$ is **unique**, we have a **complete** statistic and MVU est.

If $g(\cdot)$ is not unique $\rightarrow$ no MVUE

## Rao-Blackwell-Lehmann-Scheffe Th:

Assume that $\tilde{\theta}$ is an unbiased estimator of $\theta$ and $T(\mathbf{x})$ is a sufficient statistic for $\theta$. Then, the estimator $\hat{\theta} = E[\tilde{\theta}|T(\mathbf{x})]$ is:

- valid (not dependent on $\theta$)
- unbiased
- of $\leq$ variance than that of $\tilde{\theta}$

In addition, if the sufficient statistic is **complete**, then $\hat{\theta}$ is the MVU estimator.

**Def: Complete stat.** There is only one function of the statistic that is unbiased.

# Best Linear Unbiased Estimator: BLUE

**Motivation:** When the PDF of the data is **unknown**, or it **cannot be assessed**, the MVU estimator, even if it exists, cannot be found!

- In this case methods which rely on the PDF cannot be applied

**Remedy:** Resort to a sub-optimal estimator $\looparrowright$ check its variance and ascertain whether it meets the required specifications (and/or CRLB)

**Common sense approach:** Assume an estimator to be:

- **Linear in the data**, that is, $\hat{\theta}_{BLUE} = \sum_{n=0}^{N-1} a_n x[n]$, with $a_n$ as parameters,
- Among all such linear estimators, seek for an **unbiased** one,
- Then, **minimise the variance** of this unbiased estimator.

☞ Such an estimator is termed the Best Linear Unbiased Estimator (BLUE) which **requires only knowledge of the first two moments of the PDF**.

**We will see that if the data are Gaussian, the BLUE and MVUE are equivalent**

# The form and optimality of BLUE

Consider the data $\mathbf{x} = \big[x[0], x[1], \ldots, x[N-1]\big]^T$, for which the PDF $p(\mathbf{x};\theta)$ depends on the unknown parameter $\theta$.

## The form of BLUE

The BLUE is restricted to have the form ($\mathbf{a} = \{a_n\}$)

$$\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T\mathbf{x}$$

↑ Constants to be determined

◦ We choose $\mathbf{a} = \{a_n\}$ which yield an unbiased estimate $E\{\hat{\theta}\} = \theta$

◦ Then, we perform $\min(var)$

↬ the BLUE estimator is that which is **unbiased** and has **the minimum variance.**

## Optimality of BLUE

Note, the BLUE **will be optimal only when the actual MVU estimator is linear!**

This is the case, for instance, when estimating the DC level in WGN

$$\hat{\theta} = \bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \qquad \left\{a_n = \frac{1}{N}\right\}$$

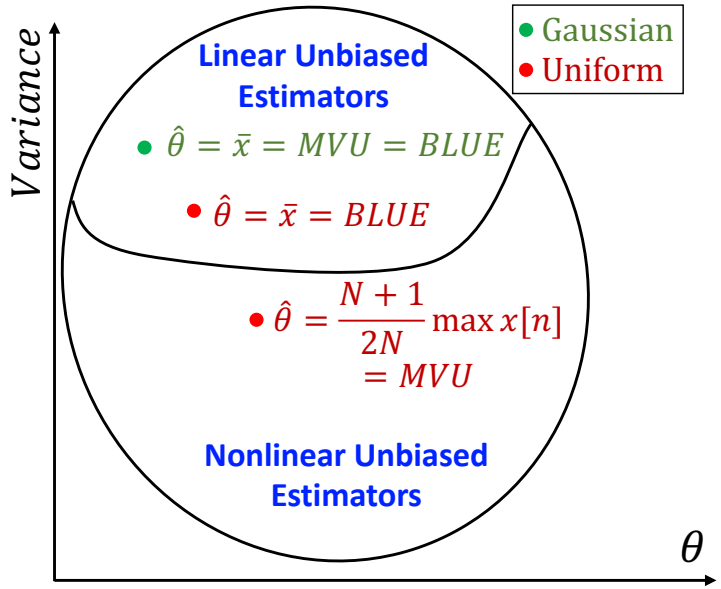
which is clearly **linear in the data.**

**Then, BLUE is an optimal MVU estimator giving** $a_n = 1/N$**.**

# The place of BLUE amongst other estimators

## We illustrate this on the estimation of a DC level in noise of different distributions

Consider the space of all unbiased estimators of DC level in noise:




- For white Gaussian noise, the MVU is **linear in the data** and is given by the sample mean $\bar{x}$.

- The MVU estimator for the mean $\theta = \frac{\beta}{2}$ of **uniform** noise, $x[n] \sim \mathcal{U}(0, \beta)$, is **nonlinear in the data**, and is given by

$$\text{mean: } \hat{\theta} = \frac{N+1}{2N} \max\{x[n]\}$$

$$\text{variance: } \text{var}(\hat{\theta}) = \frac{\beta^2}{4N(N+2)}$$

The **sample mean** estimator of uniform noise gives $\text{var}(\hat{\theta}) = \frac{\beta^2}{12N}$

So, sample mean is not an MVU estimator for uniform noise!

(see Problem 4.1 in your P&A sets)

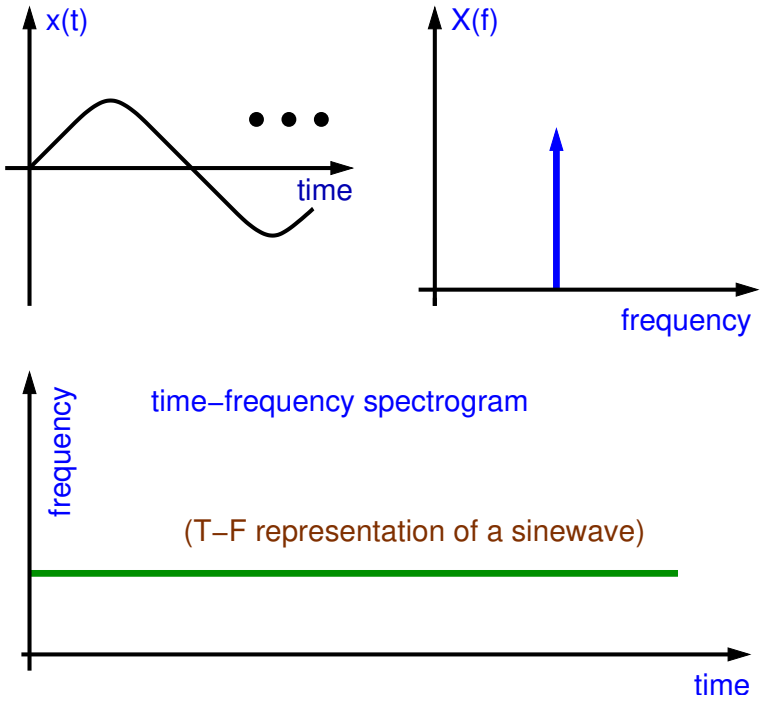
☞ The difference in performance between the BLUE and MVU estimators can, in general, be substantial, and can only be rigorously quantified through the underlying data generating pdf.

# Example 1: How useful is an estimator of DC level in noise?

(see Appendix 7)

In fact, very useful. It is up to us to provide a correct data representation.

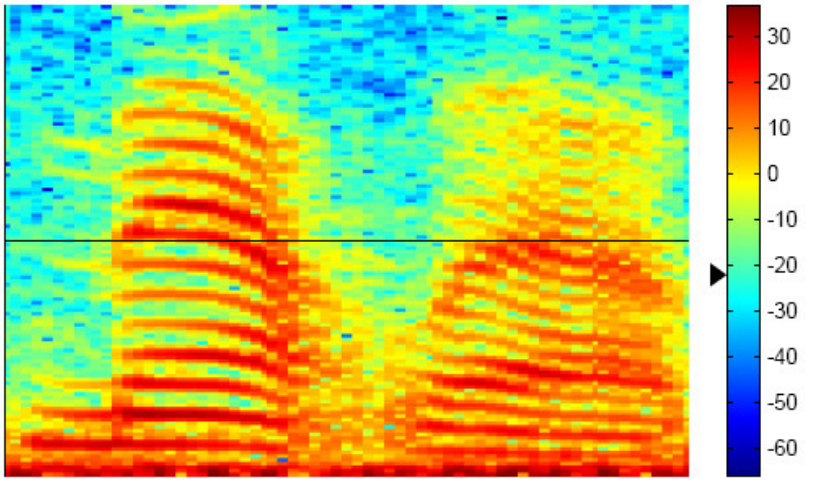
## Sinusoidal frequency estimation




- Sinusoid in time $\looparrowright$ DC level in time–frequency
- Chirp in time $\looparrowright$ ramp in T-F

## Practical example: Real-world speech

time-frequency representation




m —- aaaaa — ttt– lll – aaaaa—bb

**horizontal axis: time vertical: frequency**

This is a speech waveform of the utterance of word "matlab"

Observe DC-like harmonics for "a"

# Example 2: Problems with BLUE (are often surmountable!)

**Its direct form is inappropriate for nonlinear prob. ↬ population dynamics example**

Owing to the **linearity assumptions**, the BLUE estimator can be totally inappropriate for some estimation problems.

Power of WGN estimation

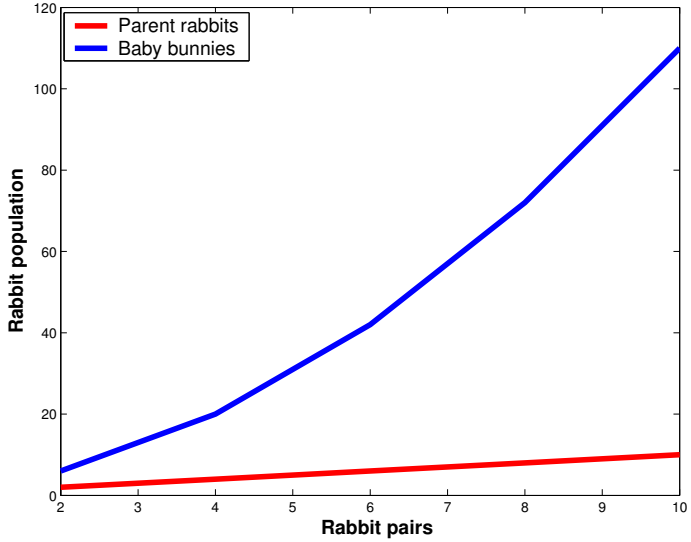
The MVU estimator $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1} x^2[n]$ is **nonlinear** in the data. Forcing the estimator to be linear, e.g. by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1} a_n x[n]$$

yields $E\{\hat{\sigma}^2\} = 0$, which is guaranteed to be biased!

A non-linear transformation of the data, i.e. $y[n] = x^2[n]$, could overcome this problem. (next Slide)

Example: Rabbit population

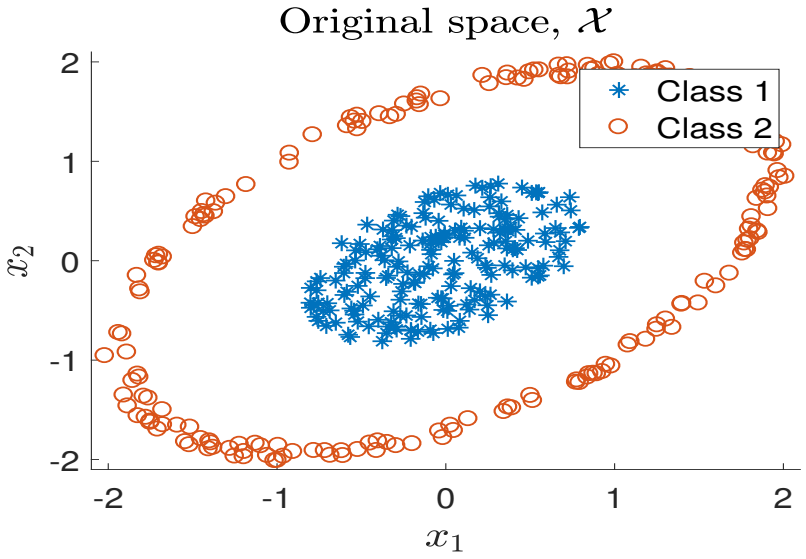


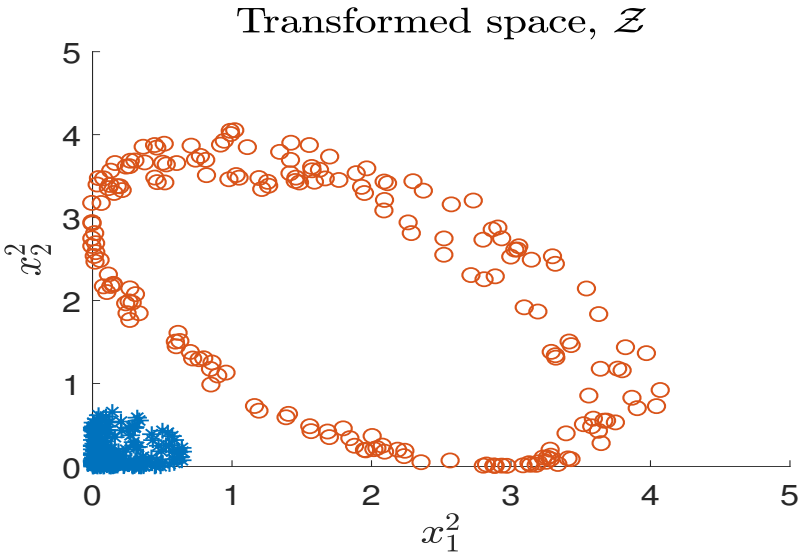

The time evolution of the rabbit population is nonlinear (exponential)

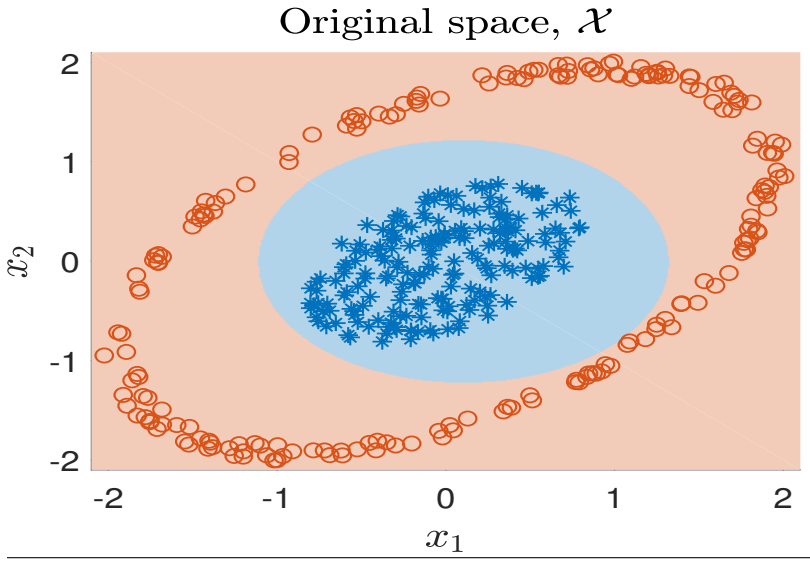
However, the number of parent pairs is linear in time!

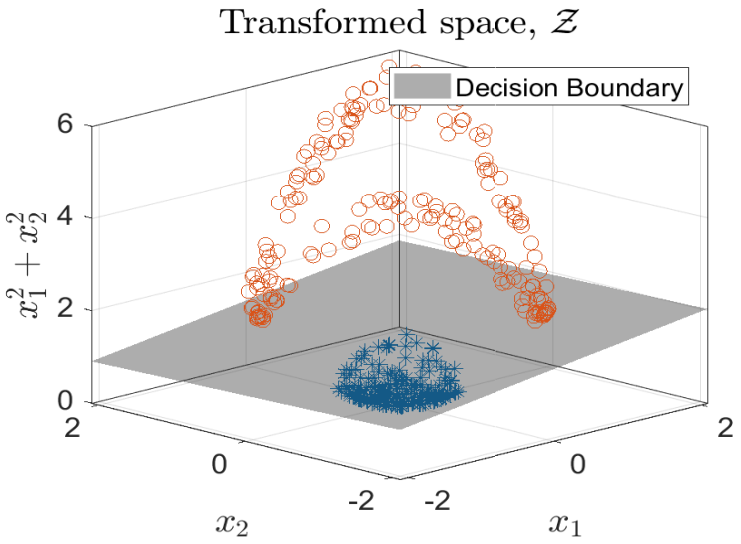
# Example 3: Nonlinear transformation of data often helps

## Left: Original data (nonlin. separab.) Right: Linear separab. after nonlin. tran.



# How to find BLUE?

## Recall: BLUE is linear in data $\hat{\theta} = \sum_{n=0}^{N-1} a_n x[n] = \mathbf{a}^T\mathbf{x}$

Consider a scalar linear observation $x[n] = \theta s[n] + w[n] \quad \Rightarrow \quad E\{x[n]\} = \theta s[n]$

and notice that $E\{\hat{\theta}\} = \theta \sum_{n=0}^{N-1} a_n s[n]$ $\qquad s[n] \rightsquigarrow$ scaled mean

### 1. Unbiased constraint

$$E\{\hat{\theta}\} = \sum_{n=0}^{N-1} a_n \underbrace{E\{x[n]\}}_{\theta s[n]} = \mathbf{a}^T\mathbf{s}\,\theta = \theta$$

$$\mathbf{a}^T\mathbf{s}\;\theta = \theta \quad \Rightarrow \quad \mathbf{a}^T\mathbf{s} = 1$$

unbiased constraint ↗

where the **scaled data vector**

$\mathbf{s} = [s[0],\ s[1], \ldots, s[N-1]]^T$

☞ In other words, to satisfy the unbiased constraint for the estimate $\hat{\theta}$, $E\{x[n]\}$ must be linear in $\theta$, or

$$E\{x[n]\} = s[n]\,\theta$$

### 2. Variance minimisation

$$\hat{\theta} = \mathbf{a}^T\mathbf{x}$$

$$\text{var}(\hat{\theta}) = E\{\hat{\theta}^2\} = E\{\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{a}\}$$

**BLUE optimisation task**

**Minimise:**

$$\text{var}(\hat{\theta}) = \mathbf{a}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{a} = \mathbf{a}^T\mathbf{C}\mathbf{a}$$

subject to the **unbiased constraint**

$$\mathbf{a}^T\mathbf{s} = 1$$

This is a constrained minimisation problem

## Some remarks on variance calculation

A closer look at the variance yields

$$var(\hat{\theta}) = E\Big\{\Big(\sum_{n=0}^{N-1} a_n x[n] - E\{\sum_{n=0}^{N-1} a_n x[n]\}\Big)^2\Big\} = E\left\{\left(\mathbf{a}^T\mathbf{x} - \mathbf{a}^T E\{\mathbf{x}\}\right)^2\right\}$$

With $\mathbf{a} \equiv [a_0, a_1, \ldots, a_{N-1}]^T$, $y^2 = y \times y^T$, and $(\mathbf{a}^T\mathbf{x})^T = \mathbf{x}^T\mathbf{a}$, we have

$$E\Big\{\underbrace{\mathbf{a}^T\big(\mathbf{x} - E\{\mathbf{x}\}\big)}_{y}\underbrace{\big(\mathbf{x} - E\{\mathbf{x}\}\big)^T\mathbf{a}}_{y^T}\Big\} \quad = \quad \mathbf{a}^T\mathbf{C}\mathbf{a} \qquad \text{like } var(aX) = a^2 var(X)$$

Also assume

$$E\{x[n]\} = s[n]\theta, \quad \text{easy to show from} \quad x[n] = E\{x[n]\} + \big[x[n] - E\{x[n]\}\big]$$

$$\text{by viewing} \quad w[n] = x[n] - E\{x[n]\}, \quad \text{we have} \quad x[n] = \theta s[n] + w[n]$$

☞ **BLUE is linear in the unknown parameter** $\theta$, which corresponds to the amplitude estimation of known signals in noise (to generalise this, a nonlinear transformation of the data is required). (see Slide 20)

# BLUE as a constrained optimisation paradigm

For Lagrange optimisation see Lecture 1 and Appendix 11

**Task:** minimize the variance subject to the unbiased constraint

$$\underbrace{\min\left\{\mathbf{a}^T\mathbf{C}\mathbf{a}\right\}}_{optimisation\ \ task} \quad \text{subject to} \quad \underbrace{\mathbf{a}^T\mathbf{s}=1}_{equality\ \ constraint}$$

**Method of Lagrange multipliers**

1. $$J = \mathbf{a}^T\mathbf{C}\mathbf{a} - \lambda(\mathbf{a}^T\mathbf{s} - 1)$$

2. Calculate
$$\frac{\partial J}{\partial \mathbf{a}} = 2\mathbf{C}\mathbf{a} - \lambda\mathbf{s}$$

3. Equate to zero and solve for $\mathbf{a}$
$$\mathbf{a} = \frac{\lambda}{2}\mathbf{C}^{-1}\mathbf{s}$$

Solve for the Lagrange multiplier $\lambda$

4. From the constraint equation
$$\mathbf{a}^T\mathbf{s} = \frac{\lambda}{2}\,\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s} = 1$$
$$\Rightarrow \quad \frac{\lambda}{2} = \frac{1}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$$

5. Replace into Step 3, with the constraint satisfied for
$$\mathbf{a}_{opt} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$$

**These are the coefficients of BLUE**

# Summary: BLUE

## Recall that $\theta = \mathbf{a}^T\mathbf{x}$ $\quad$ $\mathrm{var}(\hat{\theta}) = \mathbf{a}^T\mathbf{C}\mathbf{a}$

**BLUE of an unknown parameter:**

$$\hat{\theta} = \mathbf{a}_{opt}^T\mathbf{x} = \frac{\boldsymbol{s}^T\mathbf{C}^{-1}}{\boldsymbol{s}^T\mathbf{C}^{-1}\boldsymbol{s}}\,\boldsymbol{x} \qquad \text{where} \quad \mathbf{a}_{opt} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$$

**BLUE variance:**

$$\mathrm{var}\left(\hat{\theta}\right) = \mathbf{a}_{opt}^T\ \mathbf{C}\ \mathbf{a}_{opt} = \frac{1}{\mathbf{s}^T\ \mathbf{C}^{-1}\ \mathbf{s}}$$

☞ To determine the BLUE we only require knowledge of

$\mathbf{s}$ $\quad\rightarrowtail\quad$ the scaled mean

$\mathbf{C}$ $\quad\rightarrowtail\quad$ the covariance matrix $\quad$ ($\mathbf{C}^{-1}$ is called the "precision matrix", see also Slide 46 in Lecture 4)

**That is, for BLUE we only need to know the first two moments of the $PDF$**

**Notice that we do not need to know the functional relation of PDF**

# Example 4: Estimation of a DC level in unknown noise

**Notice that the PDF is unspecified and does not need to be known**

Consider the estimation of a DC level in white noise, which is of an unspecified PDF and with variance $\sigma^2$.

We know that

$$x[n] = A + w[n], \qquad n = 0, 1, \ldots, N-1$$

where $\{w[n]\}$ is **any white noise with known variance $\sigma^2$ (power)**.

In other words, $\{w[n]\}$ is **not necessarily Gaussian** $\not\hookrightarrow$ there may be some statistical dependence between samples (although they are uncorrelated)

**Task:** Estimate the DC level, $A$.

**Solution:** From the assumptions of BLUE, we have

$$E\{x[n]\} = s[n]A = A \qquad \text{and therefore} \qquad s[n] = 1$$

$$\text{so that} \qquad \mathbf{s} = \mathbf{1} = [\underbrace{1, \ldots, 1}_{\text{N elements}}]^T = \mathbf{1}_{N\times 1}$$

☞ **Follows from $E\{x[n]\}$ being linear in $\theta$** $\quad\Rightarrow\quad$ $E\{x[n]\} = s[n]\theta$.

# Example 4: DC level in white noise of unknown PDF, contd.

Recall that $\mathbf{a}_{opt} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}$, $var(\hat{\theta}) = \frac{1}{\mathbf{s}^T\,\mathbf{C}^{-1}\,\mathbf{s}}$, and $\hat{\theta} = \mathbf{a}_{opt}^T\mathbf{x}$

**For any uncorrelated white noise** $\{w\}$ with power $\sigma^2$

$$\mathbf{C} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I} \quad \Rightarrow \quad \mathbf{C}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma^2} \end{bmatrix} = \frac{1}{\sigma^2}\mathbf{I}$$

The BLUE for the estimation of DC level in noise then becomes (see Slide 24)

$$\hat{A} = \frac{\mathbf{1}^T\,\frac{1}{\sigma^2}\,\mathbf{I}}{\mathbf{1}^T\,\frac{1}{\sigma^2}\,\mathbf{I}\,\mathbf{1}}\,\mathbf{x} = \frac{\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}}\,\mathbf{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] = \bar{x}$$

($\mathbf{1}^T\mathbf{1} = N$)

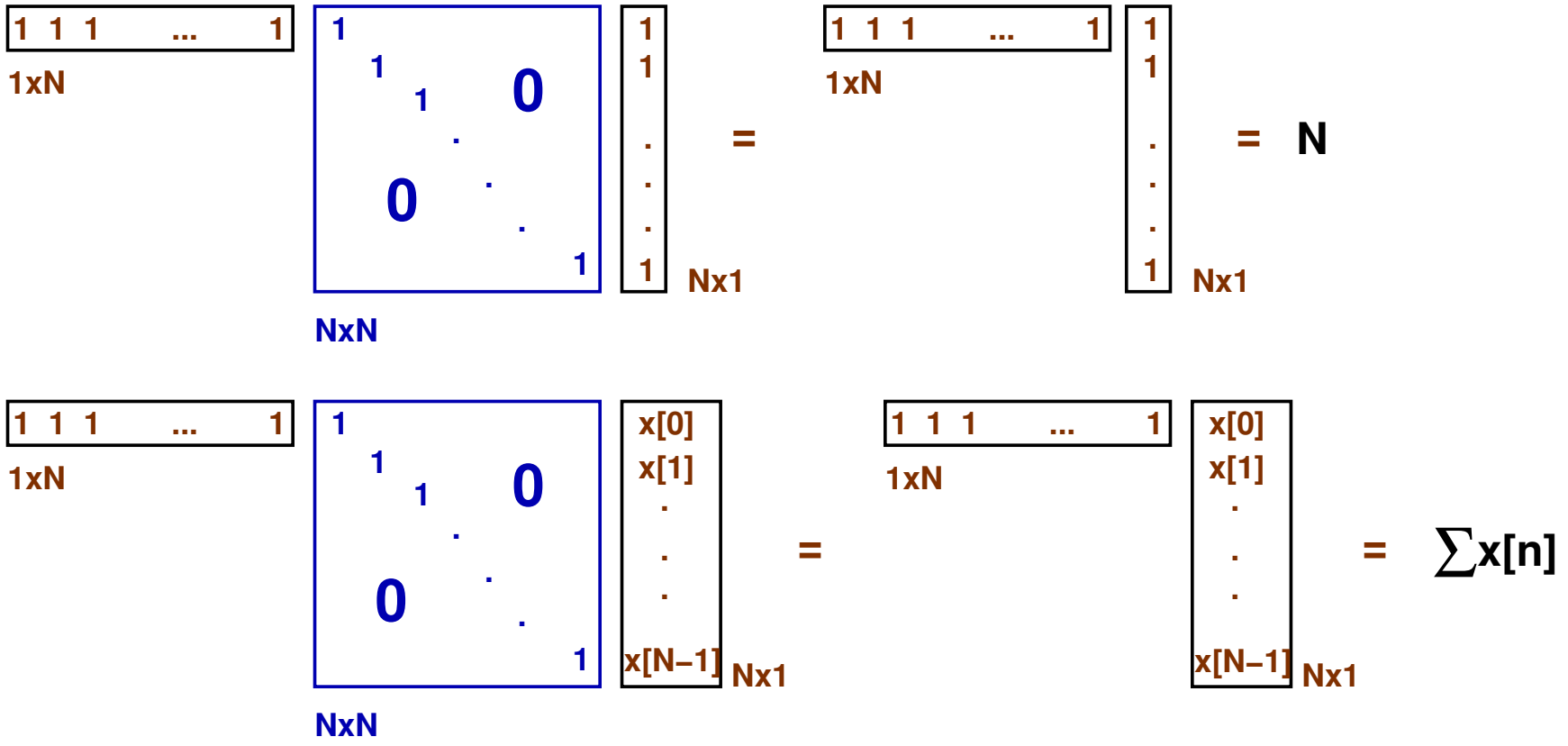
and has minimum variance (CRLB for a linear estimator) of

$$\mathrm{var}(\hat{A}) = \frac{1}{\mathbf{1}^T\frac{1}{\sigma^2}\,\mathbf{I}\,\mathbf{1}} = \frac{\sigma^2}{N}$$

- The sample mean is the BLUE, independent of the PDF of the data
- BLUE is also the MVU estimator if the noise $\{w\}$ is Gaussian

☞ **If the noise is not Gaussian (e.g. uniform) the CRLB and MVU estimator may not exist, but BLUE still exists!** (P&A sets and Slide 17)

# Some help with the quadratic forms of the type $\mathbf{a}^T\mathbf{A}\mathbf{a}$
## We shall analyse the expressions $1^T\mathbf{I}\,1$ and $1^T\mathbf{I}\,\mathbf{x}$




It is useful to visualise any type of vector–matrix expression.

☞ It is now obvious that e.g. the **scalar** $\mathbf{a}^T\mathbf{A}\mathbf{a}$ is 'quadratic' in $\mathbf{a}$.

This is easily proven by considering $\mathbf{x}^T\mathbf{I}\,\mathbf{x}$ in the diagrams above.

# Example 5: DC Level in non-iid but uncorrelated zero mean noise with var$(w[n]) = \sigma_n^2$ (de–emphasising bad samples)

**Notice that now the noise variance depends on the sample number!**

**As before, s = 1.**

The covariance matrix of the noise

$$\mathbf{C} = \begin{bmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^2 \end{bmatrix}$$

and thus

$$\mathbf{C}^{-1} = \begin{bmatrix} \sigma_0^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_1^{-2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{N-1}^{-2} \end{bmatrix}$$

The BLUE solution:

$$\hat{A} = \frac{\mathbf{1}^T \mathbf{C}^{-1}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \mathbf{x} = \frac{\sum_{n=0}^{N-1} \frac{x[n]}{\sigma_n^2}}{\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}}$$

- The term $\sum_{n=0}^{N-1} \frac{1}{\sigma_n^2}$ ensures that the estimator is unbiased
- BLUE assigns greater weights to samples with smaller variances
- Notice that

$$var(\hat{A}) = \frac{1}{\sum_{n=0}^{N-1} 1/\sigma_n^2}$$

## BLUE: Extension to the vector parameter (see Appendix 6)

**System model:** $\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n], i = 1, \ldots, p \Rightarrow \hat{\boldsymbol{\theta}} = \mathbf{Ax}$

For every $\theta_i \in \boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^T$ we have ($a_{in} \leftrightarrow$ weighting coefficients)

Scalar BLUE: $\hat{\theta}_i = \sum_{n=0}^{N-1} a_{in} x[n] = \mathbf{a}_i^T \mathbf{x} \quad i = 1, 2, \ldots, p \quad \overset{\text{vector BLUE}}{\longrightarrow} \quad \hat{\boldsymbol{\theta}} = \mathbf{Ax}$

**Now: Unbiased constraint**

Scalar BLUE: $E\{\hat{\theta}_i\} = \sum_{n=0}^{N-1} a_{in} E\{x[n]\} = \theta_i \quad \overset{\text{vector BLUE}}{\longrightarrow} \quad E\{\hat{\boldsymbol{\theta}}\} = \mathbf{A}E\{\mathbf{x}\} = \boldsymbol{\theta}$

Scalar BLUE: $E\{x[n]\} = s[n]\theta \quad \overset{\text{vector BLUE}}{\longrightarrow} \quad E\{\mathbf{x}\} = \mathbf{H}\,\boldsymbol{\theta} \;\rightarrow\; E\{\boldsymbol{\theta}\} = \mathbf{AH}\boldsymbol{\theta}$

where the coefficients $\mathbf{A} = [a_{in}]_{p \times N}$ and $\mathbf{H}$ is a vector/matrix form of $\{s[n]\}$ terms

☞ **Unbiased constraint:** $\mathbf{AH} = \mathbf{I}$ and we wish to minimise: $\text{var}(\hat{\theta}_i) = \mathbf{a}_i^T \mathbf{C} \mathbf{a}_i$

**The vector BLUE becomes:**

$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}$ with the covariance $\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1}$

**If the data are truly Gaussian**, as in

$$\mathbf{x} = \mathbf{H}\,\boldsymbol{\theta} + \mathbf{w} \quad \text{with} \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$$

**then the vector BLUE is also the Minimum Variance Unbiased estimator.**

## The Gauss – Markov Theorem

Consider the observed data in the form of a general linear model

$$\mathbf{x} = \mathbf{H}\,\boldsymbol{\theta} + \mathbf{w}$$

with $\mathbf{w}$ having zero mean and covariance $\mathbf{C}$, **otherwise an arbitrary PDF.**

**Then, the vector BLUE of $\theta$ can be found as**

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\,\mathbf{C}^{-1}\,\mathbf{H}\right)^{-1}\,\mathbf{H}^T\,\mathbf{C}^{-1}\,\mathbf{x}$$

and for every $\hat{\theta}_i \in \hat{\boldsymbol{\theta}}$, the minimum variance of $\hat{\theta}_i$ is

$$var(\hat{\theta}_i) = \left[\left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\right]_{ii}$$

with the covariance matrix of $\hat{\boldsymbol{\theta}}$ given by

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{H}^T\,\mathbf{C}^{-1}\,\mathbf{H}\right)^{-1}$$

# Example 6: Sinusoidal phase estim. (DSB, PSK, QAM)

## Motivation for Maximum Likelihood Estimation (MLE)

**Signal model:** $x[n] = A\cos(2\pi f_0 n + \Phi) + w[n] \qquad w \sim \mathcal{N}(0, \sigma^2)$

**Signal to noise ratio (SNR):** $SNR = \frac{P_{signal}}{P_{noise}} = \frac{A^2}{2\sigma^2}$

**Parametrised pdf:** $p(\mathbf{x}; \Phi) = \frac{1}{(2\pi\sigma^2)^{N/2}}\, e^{-\frac{\sum_{n=0}^{N-1}\left(x[n] - A\cos(2\pi f_0 n + \Phi)\right)^2}{2\sigma^2}}$

**Regularity condition within CRLB:** $\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta} = I(\theta)\big[g(\mathbf{x}) - \theta\big]$

**In our case:** (see Example 8, slide 40)

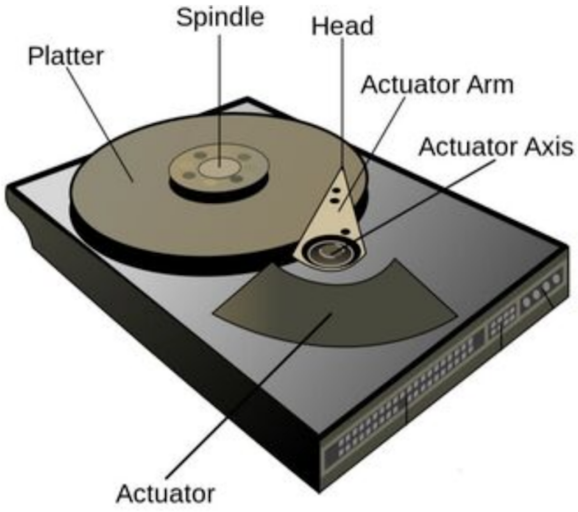
$$\frac{\partial \ln p(\mathbf{x}; \Phi)}{\partial \Phi} = -\frac{A}{\sigma^2}\sum_{n=0}^{N-1}\left(x[n]\sin(2\pi f_0 n + \Phi) - \frac{A}{2}\sin(4\pi f_0 n + 2\Phi)\right)^2$$

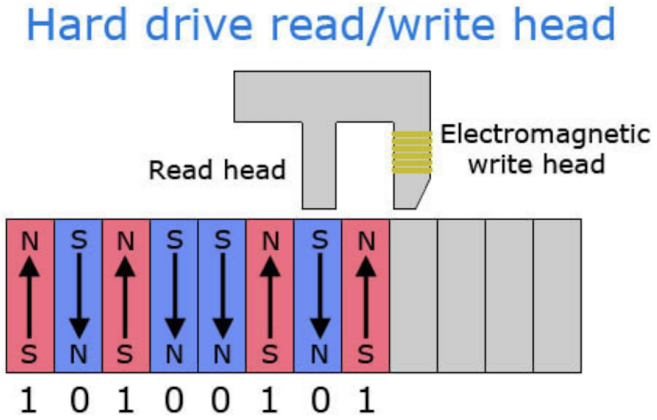
☞ **We cannot arrive at the above regularity condition, and an efficient estimator for sinusoidal phase estimation does not exist**

**Remedy:** Using MLE, we can still obtain an $\approx$ CRLB for freq. far from 0 and 1/2

**Approximate CRLB:** $var(\Phi) \geq \frac{1}{N \times SNR}$ (see Example 8)

# Maximum Likelihood Estimation (MLE): A familiar example from HDDs in our computers




- The spindle spins the HDD platter
- The actuator arm moves across the magnetic medium
- R/W head changes the polarity to either the North or South pole
- This is very much like our usual binary coding of 0's and 1's

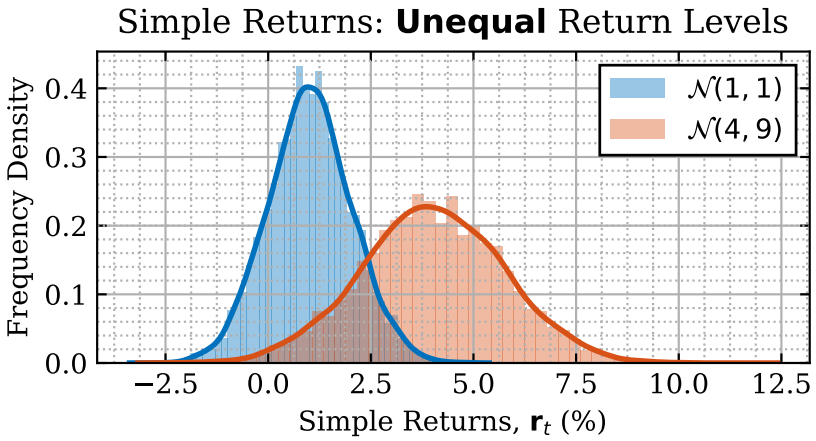
One method for recovering digital data from a weak analog magnetic signal is called the partial-response maximum-likelihood (PRML)

☞ A more advanced, and currently used method, is the correlation-sensitive maximum likelihood sequence detector (CS-MLSD) (Kavcic and Moura)

# Towards Maximum Likelihood Estimation (MLE)

## Effects of parametrisation on the shape of a PDF: An example from finance

Recall that $p_X(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ (parametrised by $\mu$ and $\sigma^2$)

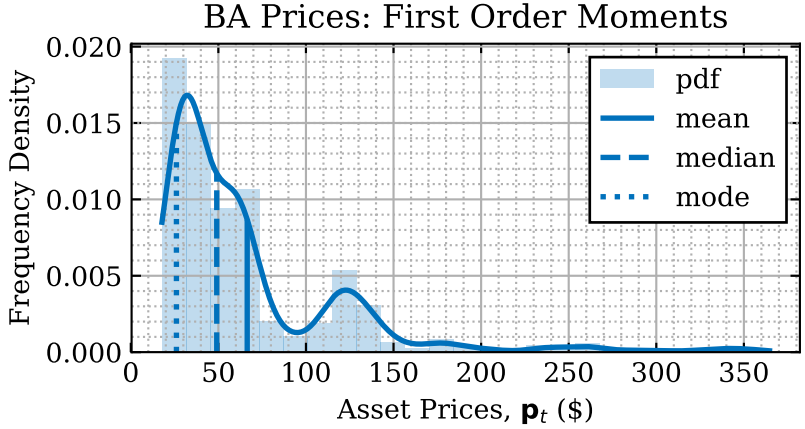


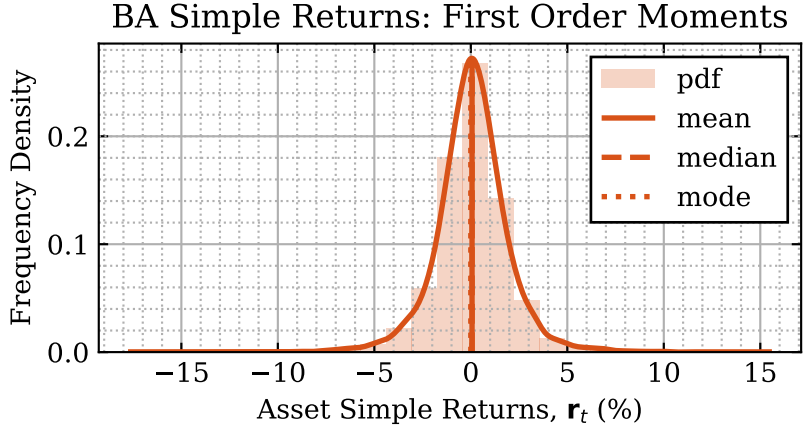

**Effects of parametrisation:**

$$p_{blue}(x;1,1) = \frac{1}{1\times\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2\times 1}}$$

$$p_{red}(x;4,9) = \frac{1}{3\times\sqrt{2\pi}} e^{-\frac{(x-4)^2}{2\times 9}}$$

☞ Virtue of imposing a distribution: $return(t) = price(t)/price(t-1)$



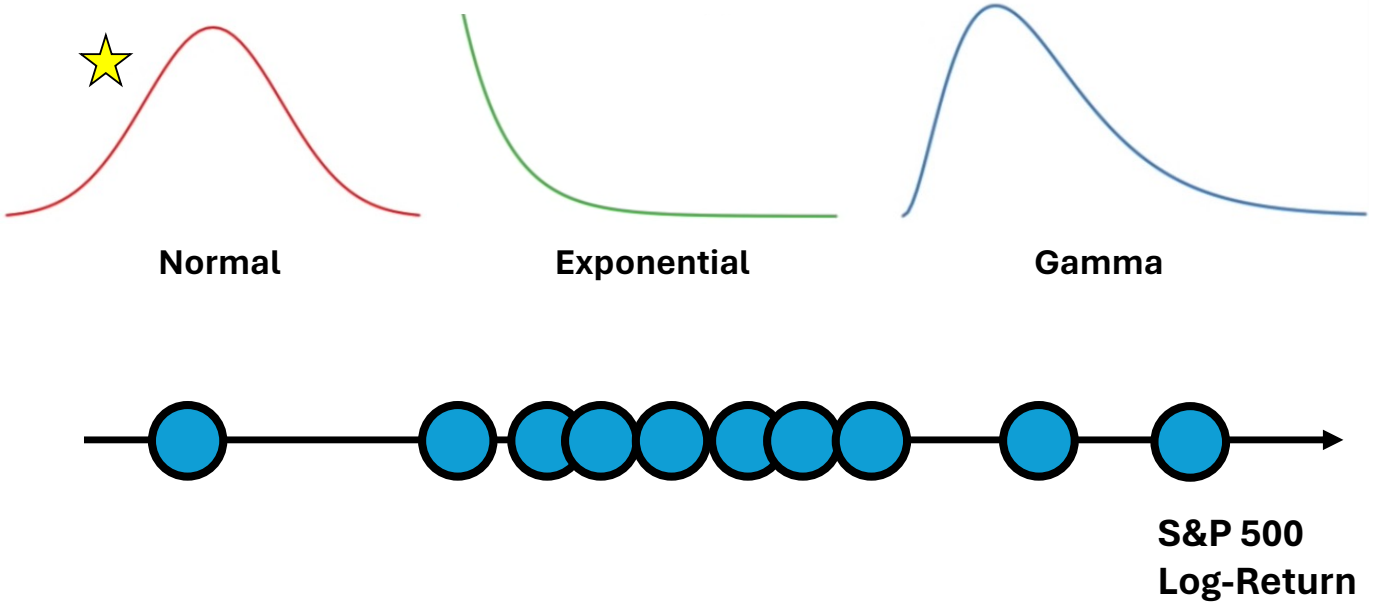



# MLE: Which distribution to assume for a given data?

## Let us consider an example from financial modelling

Consider the S&P 500 stock market index, which tracks performance of 500 largest companies listed on stock exchanges in the US. It serves as a benchmark in quantitative finance, as it indicates "market movement".

Observe log-returns of S&P 500: $\log \text{return} = \log\left[price(t)/price(t-1)\right]$

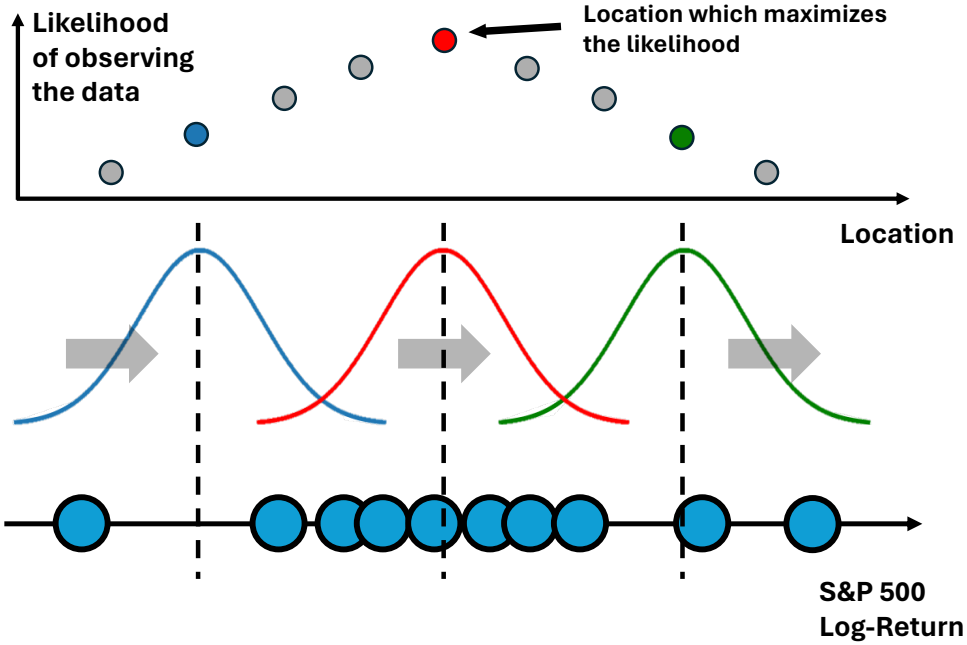



**Q:** Which of the three distribution is most natural to impose on the data?

☞ So, it is all about Maximising the Likelihood of obtaining the observed data!

# Intuition: Finding the mean and variance of the assumed distribution $\hookrightarrow$ parametrising the MLE

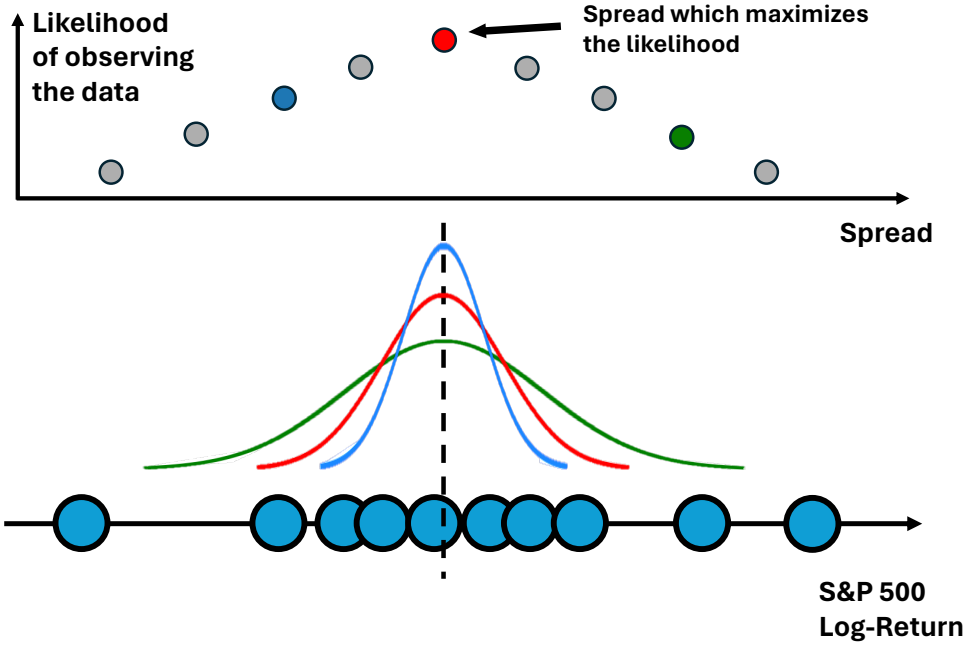
Finding the mean (location) of distribution




**Q:** Which of these functions best approximates the mean?

**A:** "Read-out" the value of likelihood function for data mean.

Finding the variance (spread)




**Q: Which of these functions best approximates the variance?**

**A:** Employ a similar procedure as for fitting the mean.

☞ **In other words, we desire to find $\hat{\theta}_{mle}$ so that $p(\mathbf{x}; \hat{\theta}_{mle})$ is largest!**

# Putting it all together: MLE as an alternative to MVU

**Effectively, we treat $\theta$ as a variable, not as a parameter,** $\hat{\theta} = arg\max_{\theta} p(\mathbf{x};\theta)$

## Rationale for Maximum Likelihood Estimation (MLE):

- The MVU estimator often does not exist or it cannot be found
- BLUE may not be applicable, that is, $\mathbf{x} \neq \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$

☞ However, if we assume a likely pdf of the data, MLE can always be found

- This yields an estimator which is generally a function of $\mathbf{x}$, while maximisation is performed over the allowable range of $\boldsymbol{\theta}$.

**Def:** The probability of observing the data, $\mathbf{x}$, given the model parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^T$, is called the **likelihood function**, $L$, and is given by

$$L(\theta_1, \theta_2, \ldots, \theta_p; \mathbf{x} = \text{observed}) = L(\boldsymbol{\theta}; \mathbf{x} = \text{fixed})$$

Unknown parameters, $\boldsymbol{\theta}$ ↗ ↖ known, observed data $\mathbf{x}$

☞ While pdf $p(\mathbf{x};\boldsymbol{\theta})$ gives the probability of occurrence of different possible values of $\mathbf{x}$, the likelihood function, $L$, is a function of the parameters $\boldsymbol{\theta}$ only, with the observed (known) data $\mathbf{x}$ held as a fixed constant!

☞ **Mathematically,** $L(\boldsymbol{\theta};\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$, so a more intuitive form of MLE is

$$\boldsymbol{\theta}_{mle} = arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathbf{x}) = arg\max_{\boldsymbol{\theta}} p_{model}(\mathbf{x};\boldsymbol{\theta}) = arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta})$$

# Principle of Maximum Likelihood Estimation (MLE)

**The unknown parameters, $\theta$, may be deterministic or random variables.**

**Principle of ML estimation:** We aim to determine **a set of parameters**, $\boldsymbol{\theta}$, from **a set of data**, $\mathbf{x}$, such that their values would yield **the highest probability of obtaining the observed data, $\mathbf{x}$.**

☞ NB: Data are "probable" and parameters are "likely" $\looparrowright$ two equivalent statements: "likelihood of the parameters" and "probability of the data".

No *a priori* distribution assumed $\looparrowright$ MLE  *A priori* distribution assumed $\looparrowright$ Bayesian

**MLE assumptions (the i.i.d. assumption):** With $L(\boldsymbol{\theta};\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$, it is often more convenient to consider the log–likelihood, $l(\boldsymbol{\theta};\mathbf{x})$, given by

$$l(\boldsymbol{\theta};\mathbf{x}) = \log L(\boldsymbol{\theta};\mathbf{x}) = \log L(\theta_1,\ldots,\theta_p; x[0],\ldots,x[N-1]) \overset{\text{i.i.d.}}{=} \prod_{n=0}^{N-1} p_{data}(x[n];\boldsymbol{\theta})$$

- The function $L(\theta; data) = p(data;\theta)$ does integrate to $1$ when integrated over data (property of PDF), but **does not integrate to 1 when integrated over the parameters**, $\boldsymbol{\theta}$ (property of likelihood fn.)
- So, $p(\mathbf{x};\theta)$ is a probability over the data, $\mathbf{x}$, and a likelihood function (not probability) over the parameters, $\boldsymbol{\theta}$.

# Example 7: MLE of a DC level in noise

Consider a DC level in WGN, where $w[n] \sim \mathcal{N}(0, \sigma^2)$

$$x[n] = A + w[n] \qquad n=0,1,\ldots,N\text{-}1$$

↑

A is to be estimated

**Step 1:** Start from the PDF

$$p(\mathbf{x}; A) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]$$

**Step 2:** Take the derivative of the log-likelihood function

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)$$

**Step 3:** Set the result to zero to yield the MLE (in general, no optimality)

$$\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

☞ **Clearly this is an MVU estimator which yields the CRLB (efficient)**

# Maximum Likelihood Estimation: Brief summary

**MLE: Makes the data you did observe the most likely data you have observed**

**MVU vs MLE:** MLE is a particular estimator which comes with a "recipe" for its calculation. MVU property relates to the properties of any estimator (unbiased, minimum variance). So, MLE could be an MVU estimator, depending on the chosen model and problem in hand.

- If an efficient estimator does exist (which satisfies the CRLB), the maximum likelihood procedure will produce it (see Example 7)
- When an efficient estimator does not exist, the MLE has the desirable property that it yields "an asymptotically efficient" estimator (Example 8)

If $\theta$ is the parameter to be estimated from a random observation $\mathbf{x}$, then the MLE $\hat{\theta}_{mle} = \arg\max_\theta p(\mathbf{x};\theta)$ is the value of $\theta$ that maximises $p(\mathbf{x};\theta)$

☞ **Conditional MLE:** Supervised Machine Learning employs conditioning between data labels $\mathbf{y}$ and input data $\mathbf{x}$, with $p_{model}$ dictated by a chosen architecture. Such conditional max. likelihood function is $L(\boldsymbol{\theta};\mathbf{y}|\mathbf{x})$, and

$$\hat{\boldsymbol{\theta}}_{mle} = arg\max_{\boldsymbol{\theta}} \sum_{n=0}^{N-1} \log p_{model}(\mathbf{y}^n|\mathbf{x}^n;\boldsymbol{\theta}) \qquad \text{(Lecture 7, Appendix 10, P\&A sets)}$$

☞ MLE is a **"turn-the-crank"** method which is optimal for large enough data. It may be computationally complex and require numerical methods.

# Example 8: MLE sinusoidal phase estimator (*cf.* Ex. 6)

**Recall the Neyman-Fisher factorisation:** $p(\mathbf{x};\theta) = g\big(T(\mathbf{x}),\theta\big)\,h(\mathbf{x})$

**MLE of sinusoidal phase.** No single sufficient statistic exists for this case. The sufficient statistics are: (see Slides 6, 7 and 8, and Appendix 3)

$$T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n) \qquad T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)$$

**The observed data:**

$$x[n] = A\cos(2\pi f_0 n + \Phi) + w[n] \qquad n = 0, 1, ..., N-1 \qquad w[n] \sim \mathcal{N}(0, \sigma^2)$$

**Task:** Find the MLE estimator of $\Phi$ by maximising

$$p(\mathbf{x};\Phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\big(x[n] - A\cos(2\pi f_0 n + \Phi)\big)^2\Big]$$

or, equivalently, minimise

$$J(\Phi) = \sum_{n=0}^{N-1}\big(x[n] - A\cos(2\pi f_0 n + \Phi)\big)^2$$

# Example 8: MLE sinusoidal phase estimator (*cf.* Ex. 6)

**Recall the Neyman-Fisher factorisation:** $p(\mathbf{x};\theta) = g\big(T(\mathbf{x}),\theta\big)\,h(\mathbf{x})$

For the minimum, differentiate w.r.t. the unknown parameter $\Phi$ to yield

$$\frac{\partial J(\Phi)}{\partial \Phi} = -2\sum_{n=0}^{N-1}\Big(x[n] - A\cos(2\pi f_0 n + \Phi)\Big)A\sin(2\pi f_0 n + \Phi)$$

and set the result to zero, to give

$$(SP1) \qquad \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n + \hat{\Phi}) = A\sum_{n=0}^{N-1}\underbrace{\sin(2\pi f_0 n + \hat{\Phi})\cos(2\pi f_0 n + \hat{\Phi})}_{\textit{inner product of sine and cosine}}$$

Recall that (use $sin(2a) = 2sin(a)cos(a)$, see also Example 9 in Lecture 4)

$$(SP2) \qquad \frac{1}{N}\sum_{n=0}^{N-1}\sin(2\pi f_0 n + \hat{\Phi})\cos(2\pi f_0 n + \hat{\Phi}) = \frac{1}{2N}\sum_{n=0}^{N-1}\sin(4\pi f_0 n + 2\hat{\Phi}) \approx 0$$

that is, it vanishes provided $f_0$ is not near 0 or $\frac{1}{2}$, and for a large enough $N$.

# Example 8: MLE sinusoidal phase estimator (*cf.* Ex. 6)

**Recall the Neyman-Fisher factorisation:** $p(\mathbf{x};\theta) = g\big(T(\mathbf{x}),\theta\big)\,h(\mathbf{x})$

Thus, the LHS of (SP1) when divided by N and set equal to zero will yield an approximation (see Appendix 2)

$$\frac{\partial J(\Phi)}{\partial \Phi} = 0 \qquad \approx \qquad \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n + \hat{\Phi}) \approx 0 \qquad \text{for large N}$$

Upon expanding $\sin(2\pi f_0 n + \hat{\Phi})$, we have (sin(a+b) = sin a cos b + cos a sin b)

$$\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)\cos\hat{\Phi} \;=\; -\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)\sin\hat{\Phi}$$

$$\text{so that the ML Estimator} \qquad \hat{\Phi} \;=\; -\arctan\frac{\displaystyle\sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)}{\displaystyle\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)}$$

☞ **The MLE $\hat{\Phi}$ is clearly a function of the two sufficient statistics,** which are $T_1(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n)$ $\qquad$ $T_2(\mathbf{x}) = \sum_{n=0}^{N-1} x[n]\sin(2\pi f_0 n)$

## Example 8: Sinusoidal phase $\looparrowright$ numerical results

The expected asymptotic PDF of the phase estimator: $\hat{\Phi}^{asy} \sim \mathcal{N}\big(\Phi, \mathcal{I}^{-1}(\Phi)\big)$

$\looparrowright$ so that the **asymptotic variance** $\mathrm{var}(\hat{\Phi}) = \dfrac{1}{\dfrac{NA^2}{2\sigma^2}} = \dfrac{1}{\eta N}$

where $\eta = \dfrac{P_{signal}}{P_{noise}} = \dfrac{A^2/2}{\sigma^2}$ $(SNR)$ is the **"signal-to-noise-ratio"**

**Below:** Simulation results with $A$=1, $f_0 = 0.08$, $\Phi = \pi/4$ and $\sigma^2 = 0.05$

| Data record length | Mean, E($\hat{\Phi}$) | $N_x\times$ variance, N var($\hat{\Phi}$) |
|---|---|---|
| 10 | 0.732 | 0.0978 |
| 40 | 0.746 | 0.108 |
| 60 | 0.774 | 0.110 |
| 80 | 0.789 | 0.0990 |
| **Theoretical asymptotic values** | $\Phi$=0.785 | $\frac{1}{\eta} = 0.1$ |

☞ For shorter data records the ML estimator is considerably biased. Part of this bias is due to the assumption (SP2) on Slide 41.

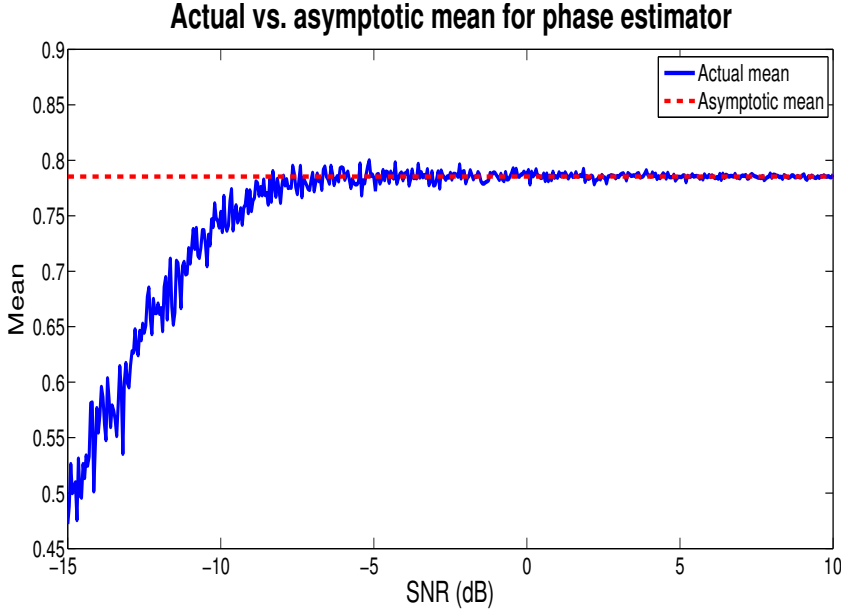
# Example 8: MLE of sinusoidal phase $\looparrowright$ asymptotic mean and variance (performance vs SNR for a fixed $N$)

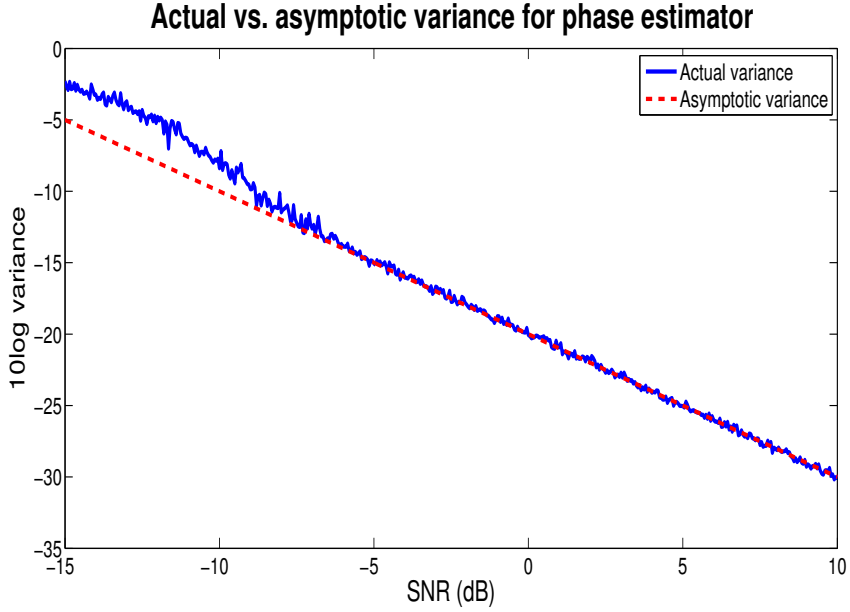
For a fixed data length of $N = 80$, SNR was varied from -15 dB to $+10$ dB

- The asymptotic variance (or CRLB) then becomes

$$10 \log_{10} \mathrm{var}(\hat{\Phi}) = 10 \log_{10} \frac{1}{N\eta} = -10 \log_{10} N - 10 \log_{10} \eta$$

- Mean and variance are also functions of SNR
- Asymptotic mean is attained for SNRs $>$ -10dB




**Observe that the minimum data length to attain CRLB also depends on SNR**

## Asymptotic properties of MLE

We can now formalise the asymptotic properties of $\hat{\theta}_{ML}^{asy}$ (see the previous slide).

**Theorem (asymptotic properties of MLE):** If $p(\mathbf{x};\theta)$ satisfies some "regularity" conditions, then the MLE is **asymptotically distributed** as

$$\hat{\theta}^{asy} \sim \mathcal{N}\left(\theta, \mathcal{I}^{-1}(\theta)\right)$$

where "regularity" refers to the existence of the derivative of the log–likelihood function (as well as Fisher information being non–zero), and $\mathcal{I}$ is the Fisher Information evaluated at the true value of the unknown parameter $\theta$.

☞ The Maximum Likelihood Estimator is therefore **asymptotically:**

- unbiased
- efficient (that is, it achieves the CRLB)

☞ For a small $N$, there is no guarantee how the MLE behaves

We can use **Monte Carlo simulations** to answer "*how large an $N$ do we need for an appropriate estimate ?*" (see Appendix 8 for more detail)

# MLE: Extension to vector parameter

☞ **A distinct advantage of the MLE** is that it can always find it for a given dataset numerically, as the MLE is a maximum of a known function.

◦ For instance, a grid search of $p(\mathbf{x};\boldsymbol{\theta})$ can be performed over a finite interval [a, b].

◦ If the grid search cannot be performed (e.g. infinite range of $\theta$) then we may resort to **iterative maximisation**, such as the Newton-Raphson method, the scoring approach, and the expectation-maximisation (EM) approach. MLE depends on a good initial guess of the underlying PDF.

◦ Since the likelihood function to be maximised **is not known a priori** and it changes for each dataset, we effectively maximise a **random function**.

☞ **Extension to the vector parameter** is straightforward: The MLE for a vector parameter $\boldsymbol{\theta}$ is the value that maximises the likelihood function $L(\boldsymbol{\theta};\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$ over the allowable domain of $\boldsymbol{\theta}$.

**Asymptotic properties:** If $\dfrac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ then $\hat{\boldsymbol{\theta}} \overset{\text{asy}}{\sim} \mathcal{N}\big(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\big)$

# Example 9: MLE of a DC level in WGN. Both the DC level $A$ and the noise variance (power) $\sigma^2$ are unknown

Consider the data $x[n] = A + w[n], \quad n = 0, 1, \ldots, N-1, \quad w[n]$ is zero–mean
The vector parameter $\boldsymbol{\theta} = [A, \sigma^2]^T$ is to be estimated ($var(w)$ is unknown too).

**Solution:** Assume $p(\mathbf{x}; \boldsymbol{\theta}) = p(\mathbf{x}; A, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\}$

$$\text{Now:} \qquad \frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)$$

$$\text{and:} \qquad \frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-A)^2$$

From first equation solve for $A$, from second equation solve for $\sigma^2$ to obtain

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \bar{x} \\ \frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\bar{x})^2 \end{bmatrix} \overset{N\to\infty}{\longrightarrow} \begin{bmatrix} A \\ \sigma^2 \end{bmatrix} \quad \text{asymptotic CRLB}$$

where $\bar{x} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$.

☞ Amazing, we only assumed a type the PDF, but not the mean or variance!

# Example 10: Sinusoidal parameter estimation with three unknown parameters $\hookrightarrow A$, $f_0$, and $\Phi$ (Example 7 in Lecture 4)

Now, $\boldsymbol{\theta} = [A, f_0, \Phi]^T$, and

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\Big[ -\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} \underbrace{\big(x[n] - A\cos(2\pi f_0 n + \Phi)\big)^2}_{\text{we need this as } (\mathbf{x}-\mathbf{H}\boldsymbol{\theta})^T(\mathbf{x}-\mathbf{H}\boldsymbol{\theta})} \Big]$$

For $A > 0,\ 0 < f_0 < \frac{1}{2}$, the MLE of $\boldsymbol{\theta} = [A, f_0, \Phi]^T$ is found by minimising

$$\begin{aligned} J(A, f_0, \Phi) &= \sum_{n=0}^{N-1} \big(x[n] - A\cos(2\pi f_0 n + \Phi)\big)^2 \\ &= \sum_{n=0}^{N-1} (x[n] - \underbrace{A\cos\Phi}_{\alpha_1} \cos 2\pi f_0 n + \underbrace{A\sin\Phi}_{-\alpha_2} \sin 2\pi f_0 n)^2 \end{aligned}$$

☞ **The function $J(A, f_0, \Phi)$ is "coupled" in $A$ and $\Phi$**, and thus hard to minimise. To this end, we may transform the multiplicative terms involving $A$ and $\Phi$ to new "linear terms" $\alpha_1 = A\cos\Phi, \quad \alpha_2 = A\sin\Phi$

with the inverse mapping $A = \sqrt{\alpha_1^2 + \alpha_2^2}$ & $\Phi = \tan^{-1}(\frac{-\alpha_2}{\alpha_1})$

# Example 10: Sinusoidal parameter estimation of three unknown parameters, cont.

(see Linear Models in Lecture 4)

For convenience of notation, we shall now introduce the vectors of sampled $\cos$ and $\sin$ terms (containing the unknown frequency $f_0$) in the form

$$\mathbf{c} = \left[1, \cos 2\pi f_0, \ldots, \cos 2\pi f_0(N-1)\right]^T \quad \mathbf{s} = \left[0, \sin 2\pi f_0, \ldots, \sin 2\pi f_0(N-1)\right]^T$$

to yield the function $J'(\alpha_1, \alpha_2, f_0)$ which is **quadratic in** $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^T$

$$J'(\alpha_1, \alpha_2, f_0) = \left(\mathbf{x} - \alpha_1\mathbf{c} - \alpha_2\mathbf{s}\right)^T\left(\mathbf{x} - \alpha_1\mathbf{c} - \alpha_2\mathbf{s}\right) = \left(\mathbf{x} - \mathbf{H}\boldsymbol{\alpha}\right)^T\left(\mathbf{x} - \mathbf{H}\boldsymbol{\alpha}\right) \quad (*)$$

☞ We arrive at a **linear estimator** of the vector parameter $\boldsymbol{\alpha} = [\alpha_1, \alpha_2]^T$, where $\mathbf{H} = [\mathbf{c}\,|\,\mathbf{s}]$ (see Example 9 in Lecture 4)

This function can be minimised over $\boldsymbol{\alpha}$, exactly as in the linear model (with $\mathbf{C} = \mathbf{I}$), to give (Slide 36, Lecture 4)

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\alpha}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} \quad \rightarrow \quad \text{insert into} \quad (*)$$

to yield $J'(\alpha_1, \alpha_2, f_0) = (\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\alpha}})^T(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\alpha}}) = \mathbf{x}^T\big(\mathbf{I} - \underbrace{\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T}_{\max\ this\ for\ \min J'}\big)\mathbf{x}$

# Example 10: Sinusoidal parameter estimation of three unknown parameters, cont. cont.

Hence, to find $\hat{f}_0$ we need to minimise $J'$ over $\hat{f}_0$ or, equivalently

$$\text{maximise } \mathbf{x}^T\mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}$$

☞ Using the definition of $\mathbf{H}$, the MLE for frequency $\hat{f}_0$ is **the value that maximises the power spectrum estimate** (see your P&A sets)

$$\begin{bmatrix} \mathbf{c^Tx} \\ \mathbf{s^Tx} \end{bmatrix}^T \begin{bmatrix} \mathbf{c^Tc} & \mathbf{c^Ts} \\ \mathbf{s^Tc} & \mathbf{s^Ts} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c^Tx} \\ \mathbf{s^Tx} \end{bmatrix} = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]e^{-\jmath 2\pi f_0 n}\right|^2 \quad \leftarrow \text{periodogram}$$

$\nwarrow \mathbf{x}^T\mathbf{H}$ $\qquad$ $\nwarrow (\mathbf{H}^T\mathbf{H})^{-1}$ $\qquad$ $\nwarrow \mathbf{H}^T\mathbf{x}$

Use this expression to find $\hat{f}_0$, and proceed to find $\hat{\boldsymbol{\alpha}}$ (Example 9, Lect. 4)

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \approx \frac{2}{N}\begin{bmatrix} \mathbf{c^Tx} \\ \mathbf{s^Tx} \end{bmatrix} = \begin{bmatrix} \frac{2}{N}\sum x[n]\cos 2\pi\hat{f}_0 n \\ \\ \frac{2}{N}\sum x[n]\sin 2\pi\hat{f}_0 n \end{bmatrix} \qquad \hat{\Phi} = -\arctan\frac{\sum_{n=0}^{N-1} x[n]\sin(2\pi\hat{f}_0 n)}{\sum_{n=0}^{N-1} x[n]\cos(2\pi\hat{f}_0 n)}$$

$$\text{and} \quad \hat{A} = \sqrt{\hat{\alpha}_1^2 + \hat{\alpha}_2^2} = \frac{2}{N}\left|\sum_{n=0}^{N-1} x[n]\exp(-j2\pi\hat{f}_0 n)\right|$$

# MLE for transformed parameters (invariance property)

## This invariance property of MLE is another big advantage of MLE

Following the above example, we can now state the **invariance property** of MLE (also valid for the scalar case).

**Theorem (invariance property of MLE):** The MLE of a vector parameter $\boldsymbol{\alpha} = f(\boldsymbol{\theta})$, where the pdf $p(\mathbf{x}; \boldsymbol{\theta})$ is parametrised by $\boldsymbol{\theta}$, is given by

$$\hat{\boldsymbol{\alpha}} = f(\hat{\boldsymbol{\theta}})$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

☞ Since MLE of $\hat{\boldsymbol{\theta}}$ is obtained by maximising $p(\mathbf{x}; \boldsymbol{\theta})$, if $f$ is a one-to-one function this is obvious, and the MLE of the transformed parameter is found by substituting the MLE of the original parameter into the transformation.

For example, if $x[n] = A + w[n], \quad w \in \mathcal{N}(0, \sigma^2)$, but we wish to find the MLE of $\alpha = exp(A)$.

◦ The resulting log–likelihood is still parametrised by $A$, and by using $ln\ \alpha = A$ as a transform, the resulting MLE is obtained as

$$\hat{\alpha} = exp(\hat{A})$$ (see also your P & A sets)

# Optimality of MLE for a linear model

We can now summarise the observations so far in the form of the optimality theorem for MLE.

**Theorem:** Assume that the observed data can be described by the general linear model

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

where $\mathbf{H}$ is a known $N \times p$ matrix with $N > p$ and of rank $p$ (tall matrix), $\boldsymbol{\theta}$ is a $p \times 1$ parameter vector to be estimated, and $\mathbf{w}$ is a noise vector with PDF $\mathcal{N}(\mathbf{0}, \mathbf{C})$. Then, the MLE of $\boldsymbol{\theta}$ takes the form

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$$

In addition, $\hat{\boldsymbol{\theta}}$ is also an **efficient estimator** in that it attains the CRLB. It is hence the MVU estimator, and the PDF of $\hat{\boldsymbol{\theta}}$ is Gaussian and is given by

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\right)$$

# Summary of notions from probabilistic modelling

The data, $\mathbf{x}$, are connected to all possible models, $\theta$, by a probability $P(\mathbf{x};\theta)$ or a probability density function $p(\mathbf{x};\theta)$. In other words, a pdf gives the probabilities of occurrence of different possible values.

**Sample space:** The sample space of a random variable represents all values that the random variable can take. For example, for the coin tossing experiment the sample space is: Heads and Tails.

**Parametric modelling:** Parametric models represent a set of density functions with one or more parameters. For different values of the parameters there will be different density functions. All of these density functions are referred to as parametric models.

**Probability density function:** Given a sample space, the PDF maps the random samples to their probabilities.

**Likelihood function:** It is the probability of observing the values in the sample space, if the true generating distribution was the model which uses the particular density function parameterised by $\theta$ (think Gen-AI).

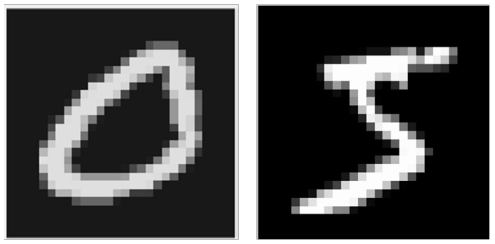
**Maximum Likelihood:** MLE aims to find the parameter values of a model which makes the observed data most probable

# Example 12: MLE in Generative Artificial Intelligence

We desire to learn a probab. distribution, $p_{data}(x)$, over data, $x$, such that:

**Generation:** If $p_{data}(x)$ is a distrib. of handwritten digit images, and we sample $x_{new} \sim p_{model}$, then $x_{new}$ should look like a digit (aka sampling)

**Density estimation:** The probability $p_{data}(x)$ should be high if a training sample, $x$, looks like a digit, and low otherwise (maximising likelihood)



Some $28 \times 28$-pixel images

L: Original, R: Generated

**Sufficient sample space for $p_{data}$.** A full ground truth space of all $28 \times 28$–pixel BW images has $28^2 = 784$ binary variables (BW pixels). This gives a total of $2^{784} = 10^{236}$ possible BW images.

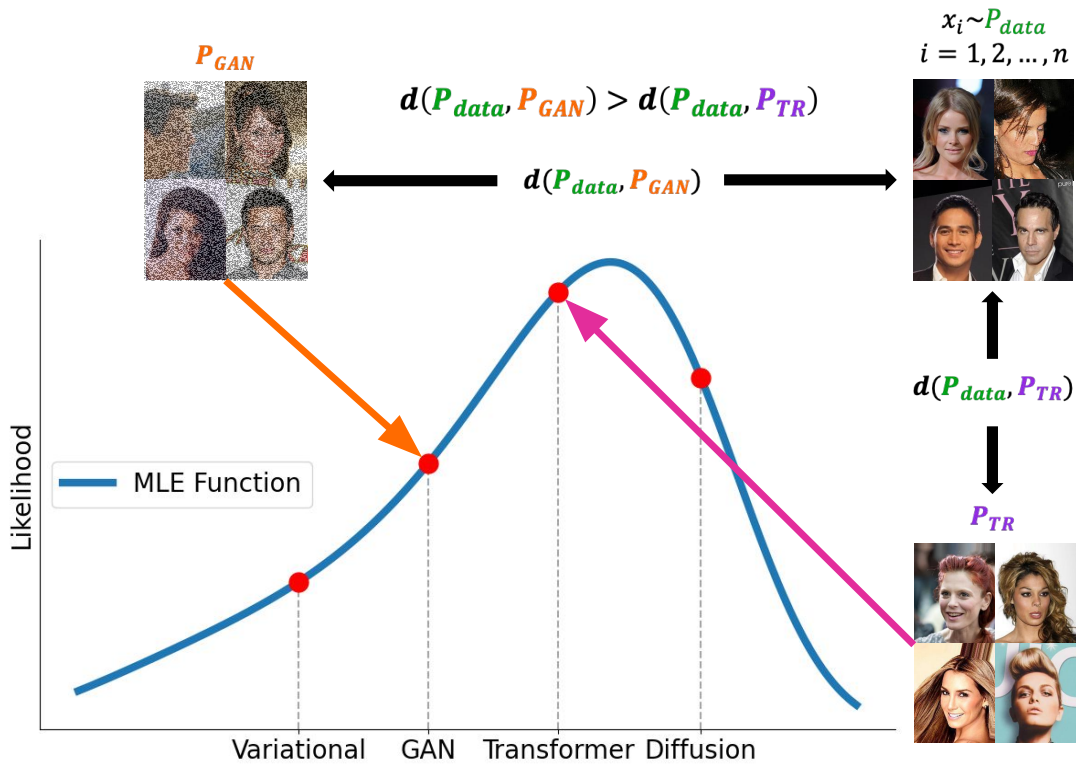
- Even a sample space of $10^9$ training data would give an extremely scarce coverage of ground truth.
- For a more realistic $1000 \times 1000 = 10^6$–pixel frames, we would have $2^{1,000,000}$ possible images.

☞ So we wish to learn $p_{model} = p_{\boldsymbol{\theta}} \approx p_{data}$, to construct the "best" fit to the available distribution $p_{data}$ from incomplete ground truth.

# Example 12a: MLE in Generative Artificial Intelligence

## (see also Appendix 9 and Appendix 10)

We often have a limited amount of samples of the dataset of interest, e.g. we do not know the true distribution of all male and female face images.




- Generative models aim to generate "new" data based on the available samples of a dataset of interest.
- Generated data should approximate the "true distribution" of unseen data, $p_{data}$, as best as possible in some statistical sense, e.g.

$$\min \text{distance}(p_{data}, p_{model}).$$

☞ We examine the likelihood of the model, given the dataset ($\equiv$ MLE).

☞ This boils down to **maximising the likelihood** that the generated data will have a similar distribution as the true data of interest.

# Example 12b: Density estimation as MLE

**The aim of learning is for $p_{model}(\mathbf{x};\boldsymbol{\theta})$ to become as close to $p_{data}(\mathbf{x})$ as possible**

Our context is **density estimation** $\looparrowright$ we desire to capture the data distribution $p_{data}(\mathbf{x})$, so as to enable either unconditional or conditional generation of new data from $\approx$ same distribution $p_{model}(\mathbf{x};\boldsymbol{\theta})$.

- MLE aims to pick a "good" model which incorporates domain knowledge (structure of the data), that is, a model with a **good inductive bias**.
- To measure "closeness" between the training data distribution $p_{data}(\mathbf{x})$ and model distribution $p_{model}(\mathbf{x};\boldsymbol{\theta})$ we use the KL divergence (Appendix 10).

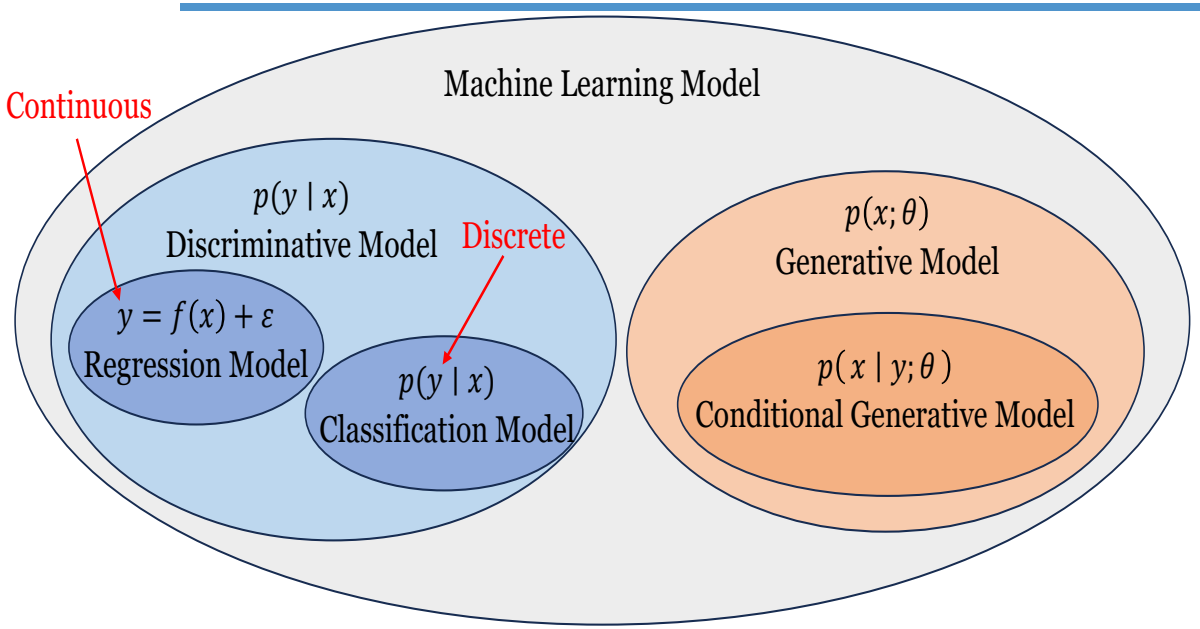
$$D_{KL}(p_{data}||p_{model}) = E_{\mathbf{x}\sim p_{data}}\left[\log \frac{p_{data}(\mathbf{x})}{p_{model}(\mathbf{x};\boldsymbol{\theta})}\right] = E_{\mathbf{x}\sim p_{data}}\left[\log p_{data}(\mathbf{x}) - \log p_{model}(\mathbf{x};\boldsymbol{\theta})\right]$$

Here, $E_{\mathbf{x}\sim p_{data}}$ is the expectation over all possible training data, which is a weighted average of all possible outcomes, with $p_{data}(\cdot)$ as "weights", i.e.

$$D_{KL}(p_{data}||p_{model}) = \sum p_{data}(\mathbf{x})\Big[\log p_{data}(\mathbf{x}) - \underbrace{\log p_{model}(\mathbf{x};\boldsymbol{\theta})}_{\text{max of this} = \text{min of } D_{KL}}\Big]$$

☞ $\arg\min_{p_\theta} D_{KL}(p_{data}||p_{model}) \equiv$ **Max. Likelihood Est.** $\arg\max_{p_\theta} \log p_{model}(\mathbf{x};\boldsymbol{\theta})$

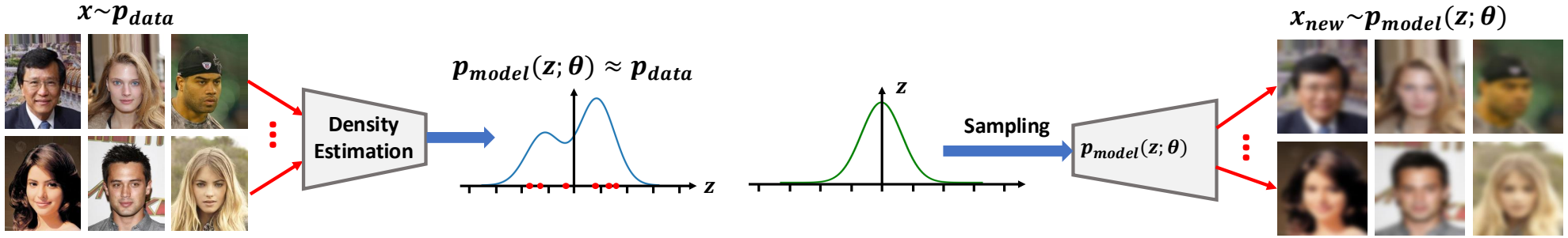
# Example 12c: Big picture of learning data distributions
## Most important general cases




- $p(y|x;\theta) \rightsquigarrow$ classification (discriminative model), e.g. logistic regression
- $p(y|x;\theta) \rightsquigarrow$ regression, prediction
- $p(x;\theta) \rightsquigarrow$ generative model (e.g. VAE, GANN)
- $p(x|y;\theta) \rightsquigarrow$ conditional generative model

☞ The difference between classification and prediction is that in classification $y$ takes discrete values (typically $0$ or $1$) while in prediction $y$ is continuous.

☞ Generative models learn a joint distribution over the entire dataset. They are typically used for **sampling applications** or **density estimation.**



# Summary: Maximum Likelihood Estimation (MLE)

**MLE:** 1) Assume a model, also known as a data generating process, for your observed data

2) For the assumed model, produce the likelihood funct. $L(\boldsymbol{\theta}; \mathbf{x}) = p(\mathbf{x}; \boldsymbol{\theta})$

3) Now, MLE becomes an optimisation problem

☞ **Differences between $p(\mathbf{x}|\theta)$ and $p(\mathbf{x}; \theta)$.** The "conditional" PDF $p(\mathbf{x}|\theta)$ is typically denoted with the semicolon ';' to indicate that $\boldsymbol{\theta}$ are not random variables but unknown parameters which "parametrise" the pdf.

☞ **Differences between the likelihood function, $L(\boldsymbol{\theta}; \mathbf{x})$, and the probability density function, $p(\mathbf{x}; \boldsymbol{\theta})$** are nuanced but important:

◦ A PDF gives the probability of observing your data, given the underlying parameters of the distribution, i.e. it maps samples to their probabilities.

◦ The likelihood function expresses the likelihood of parameter values, given your observed data. It assumes that the parameters are unknown.

MLE is grounded in probability theory; it provides a rigorous theoretical framework and underpins many probabilistic models in machine learning, such as generative AI, logistic regression, and Gaussian mixture models.

# Summary: BLUE vs MLE

**NB: The optimal MVU est. and CRLB may not exist or are impossible to find**

## Best Linear Unbiased Estimator

- It operates even when the *pdf* of data is unknown
- Restricts the estimates to be linear in the data (e.g. DC level in noise)
- Produces unbiased etimates
- Minimises the variance of such unbiased estimates
- Requires knowledge of only the mean and variance of the data, and not of the full *pdf*
- BLUE may be used more generally if the data model is linearised in an adequate way, for example, through T-F represent.

## Maximum Likelihood Estimator

- **Basic idea:** In the likelihood function, $\theta$ is regarded as a variable and not as a parameter!
- Can always be applied once the PDF is assumed; no restriction on the data model (*cf.* BLUE)
- Can be computationally complex

**Properties of MLE, as $N \to \infty$:**

- **Efficient** $\looparrowright$ attains the CRLB
- **Consistent:** Unbiased & var $\to 0$ i.e. converges to $\theta$ in probability.
- **Optimal** for the General Linear Model, **invariant to any transformation** of $\theta$, **asymptotic normality** of $\hat{\theta}_{MLE}$

# Appendix 1: "Sufficient" statistic for sequential estim.

**From Lecture 6 (new notation, $\hat{A}[N]$ = "estimate of $A$ at a time instant $N$")**

Consider the problem of LS estimation the DC level in noise, for which we have obtained

$$\hat{A}[N-1] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$$

If we now observe the new sample $x[N]$, then the new, enhanced, estimate

$$\hat{A}[N] = \frac{1}{N+1}\sum_{n=0}^{N} x[n] = \frac{1}{N+1}\Big(\sum_{n=0}^{N-1} x[n] + x[N]\Big)$$

$$\hat{A}[N] = \frac{N}{N+1}\hat{A}[N-1] + \frac{1}{N+1}x[N] \quad \rightsquigarrow \textbf{ a recursive estimate!}$$

The solution can be rewritten in a more physically insightful form, as

$$\hat{A}[N] = \hat{A}[N-1] + \frac{1}{N+1}\Big(x[N] - \hat{A}[N-1]\Big)$$

$$\textbf{new estimate} = \textbf{old estimate} + \underbrace{\textbf{gain} \times \textbf{error}}_{\text{correction}}$$

# Appendix 2: Sufficient statistic for the uniform distribution

Consider the random data $\mathbf{x} = \left[x_0, x_1, \ldots, x_{N-1}\right]^T$, which are uniformly distributed on the interval $[0, \theta]$, with the parameter $\theta$ unknown.

To find a sufficient statistic $T(\mathbf{x})$, we employ the i.i.d. assumption to yield

$$p\big(x_0, x_1, \ldots, x_{N-1}; \theta\big) = p(\mathbf{x}; \theta) = \theta^{-N}\, 1\big(x_n \leq \theta, n = 0, 1, \ldots, N-1\big) = \theta^{-N}\, 1(E)$$

where

the indicator function: $$1(E) = \begin{cases} 1, & \text{if the event E holds} \\ 0, & \text{if the event E does not hold} \end{cases}$$

The data $x_n \leq \theta, n = 0, \ldots, N-1$ iff $\max\{x_0, \ldots, x_{N-1}\} \leq \theta$ so that

$$p(\mathbf{x}; \theta) = \underbrace{\theta^{-N}\, 1\big(\max\{x_0, x_1, \ldots, x_{N-1}\} \leq \theta\big)}_{g(T(\mathbf{x}), \theta)} \times \underbrace{1}_{h(\mathbf{x})}$$

☞ By the Neyman-Fisher factorisation theorem, the sufficient statistic is

$$T(\mathbf{x}) = \max\left\{x[0], x[1], \ldots, x[N-1]\right\} = \max\{\mathbf{x}\}$$

☞ The sample mean, $\bar{x}$, **is not** a sufficient statistic for a uniform random var., as $1(\max\{\mathbf{x}\} \leq \theta)$ cannot be expressed as a function of just $\bar{x}$ and $\theta$.

# Appendix 3: Sufficient statistics for the estimation of the phase of a sinusoid

**Problem:** Estimate the phase of a sinusoid

$$x[n] = A\cos(2\pi f_0 n + \Phi) + w[n] \qquad w \sim \mathcal{N}(0, \sigma^2)$$

**Parametrised pdf:** $$p(\mathbf{x}; \Phi) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{\sum_{n=0}^{N-1}\left(x[n] - A\cos(2\pi f_0 n + \Phi)\right)^2}{2\sigma^2}}$$

The exponent may be expanded as

$$\sum_{n=0}^{N-1} x^2[n] - 2A\sum_{n=0}^{N-1} x[n]\cos(2\pi f_0 n + \Phi) + \sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \Phi)$$

$$= \sum_{n=0}^{N-1} x^2[n] - 2A\Big(\sum_{n=0}^{N-1}\cos 2\pi f_0 n\Big)\cos\Phi + 2A\Big(\sum_{n=0}^{N-1}\sin 2\pi f_0 n\Big)\sin\Phi + \sum_{n=0}^{N-1} A^2\cos^2(2\pi f_0 n + \Phi)$$

This pdf is not factorable as required by the Neyman-Fisher theorem. Hence, no single sufficient statistic exists. However, it can still be factorised as

# Appendix 3: Sufficient statistics for the estimation of the phase of a sinusoid

$$p(\mathbf{x};\Phi)=\underbrace{\frac{1}{(2\pi\sigma^2)^{N/2}}\exp\Big\{-\frac{1}{\sigma^2}\Big[\sum_{n=0}^{N-1}A^2\cos^2(2\pi f_0 n+\Phi)-2AT_1(\mathbf{x})\cos\Phi+2AT_2(\mathbf{x})\sin\Phi\Big]\Big\}}_{g(T_1(\mathbf{x}),T_2(\mathbf{x}),\Phi)}$$

$$\times\underbrace{\exp\Big\{-\frac{1}{2\sigma^2}\sum_{\mathrm{n}=0}^{\mathrm{N}-1}\mathrm{x}^2[\mathrm{n}]\Big\}}_{h(\mathbf{x})}$$

where

$$T_1(\mathbf{x})=\sum_{n-0}^{N-1}x[n]\cos 2\pi f_0 n \qquad T_2(\mathbf{x})=\sum_{n-0}^{N-1}x[n]\sin 2\pi f_0 n$$

☞ $T_1(\mathbf{x})$ and $T_2(\mathbf{x})$ are jointly sufficient statistics for the estimation of $\Phi$. However, no single sufficient statistic exists (we really desire a single sufficient statistic).

# Appendix 4: Motivation and Pro's and Con's of BLUE

**Motivation for BLUE:** Except for the Linear Model (Lecture 4), the optimal MVU estimator might:

- Not even exist,
- Be difficult or even impossible to find.

☞ BLUE is one such sub–optimal estimator.

**Idea behind BLUE:**

- Restrict the estimator to be **linear in data $\mathbf{x}$**,
- Restrict the estimate to be **unbiased**,
- Find the **best** among such unbiased estimates, **that is, the one with the minimum variance.**

**Advantages of BLUE:** It needs only the 1st and 2nd statistical moments (mean and variance).

**Disadvantages of BLUE:** 1) In general it is sub–optimal, and 2) It may be totally inappropriate for some problems (see the next slide).

# Appendix 5: More on heavy tailed distributions

- The $\alpha$-stable distribution generalises the normal distribution.
- It was proposed as a distribution for asset returns and commodity prices by Mandelbrot in the early 1960s.

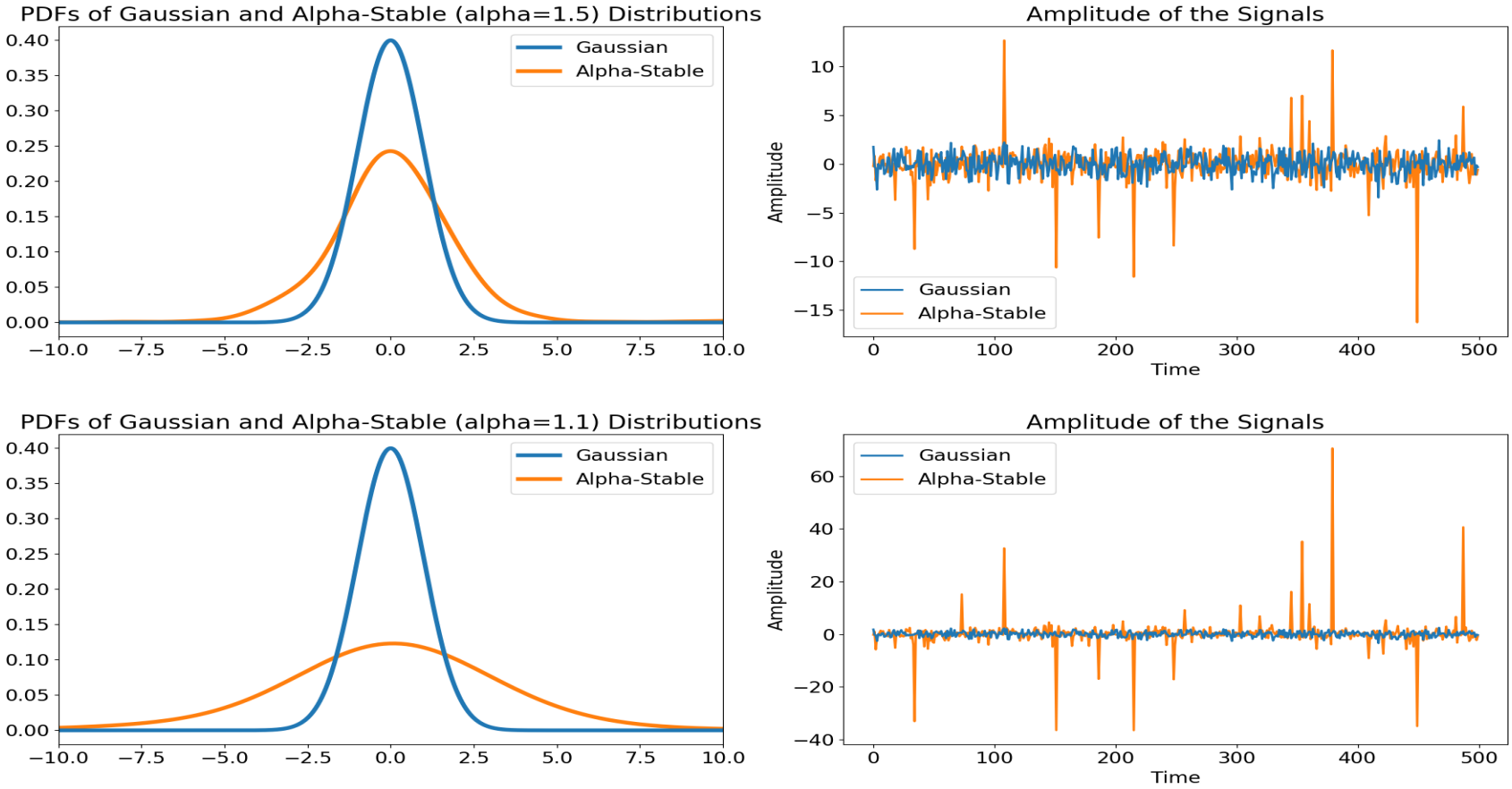



Illustration of $\alpha$-stable distributions based on synthetic data

# Appendix 6: Some observations about BLUE

- BLUE is applicable to amplitude estimation of known signals in noise, where to satisfy the unbiased constraint, $E\{x[n]\}$ must be linear in the unknown parameter $\theta$, or in other words, $E\{x[n]\} = s[n]\,\theta$.
- **Counter-example:** If $E\{x[n]\} = \cos\theta$, which is not linear in $\theta$, then from the unbiased assumption we have $\sum_{n=0}^{N-1} a_n \cos\theta = \theta$. Clearly, there are no $\{a_n\}$ that satisfy this condition.
- For the vector parameter BLUE, the unbiased constraint generalises from the scalar case as

$$E\{x[n]\} = s[n]\theta \quad \rightarrow \quad \mathbf{a}^T\mathbf{s} = 1 \qquad \Rightarrow \qquad E\{\mathbf{x}\} = \mathbf{H}\boldsymbol{\theta} \quad \rightarrow \quad \mathbf{AH} = \mathbf{I}$$

Since the **unbiased constraint yields:**

$$E\{\hat{\theta}_i\} = \sum_{n=0}^{N-1} a_{in} E\{x[n]\} = \theta_i \quad \Rightarrow \quad E\{\hat{\boldsymbol{\theta}}\} = \mathbf{A}E\{\mathbf{x}\} = \boldsymbol{\theta}$$

this is equivalent to $\mathbf{a}_i^T\mathbf{h}_j = \delta_{ij}$ (=0 for i≠j, = 1 for i=j)

# Appendix 7: Some BLUE–like "estimates"

## Composite faces $\looparrowright$ people face averages

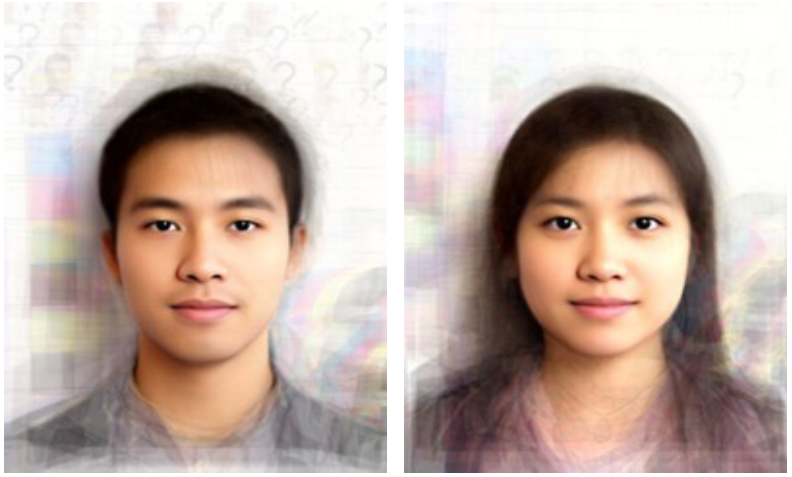
Can we estimate a "typical looking" person from a certain region, by taking a statistical average of a large ensemble of random faces photographed on the street?

Does the so-generated estimated average face exist in real life?

**Participans in Sydney, Australia, ranging from 0.83–93 years**

# Appendix 7: Some BLUE–like "estimates", contd.

Composite faces of Sydney



Composite faces of Hong Kong

Composite faces of London

Composite faces of Argentina

# Appendix 8: Monte Carlo (MC) simulations

**Use computer simulations to evaluate performance of any estimation method**

The MC simulations are illustrated here for a determin. sig. $s[n, \theta]$ in AWGN

## 1. Data collection

- Select a true parameter value, $\theta_{true}$ (usually performed over a range of values of $\theta$
- Generate a signal having $\theta_{true}$ as a parameter
- Generate WGN with unit variance and form the measurement $x = s + w$
- Choose $\sigma$ to obtain the desired SNR value and perform one MC simulation for one SNR value (usually you run many simulations over a range of SNR values)

## 2. Statistical evaluation

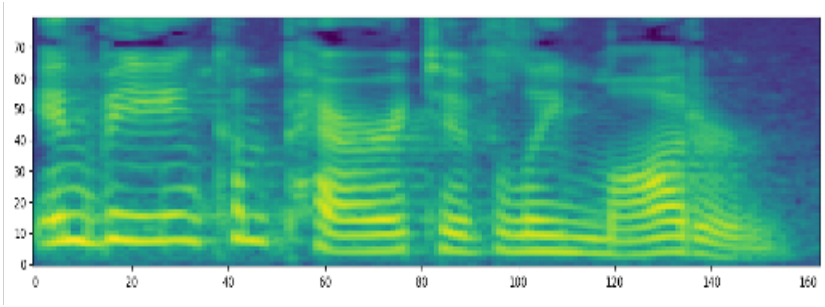
- Compute bias, $B = \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta_{true})$
- Compute RMS error, $RMS = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta_{true})^2}$
- Compute error variance, $var = \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - (\frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_M))^2$
- Plot histogram or scatter plot (if needed)

## 3. Explore (via plots)

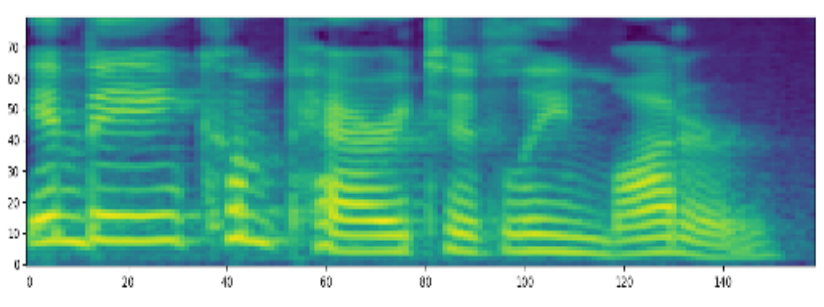
How bias, RMS, variance vary with the value of $\theta$, SNR, number of data points, $N$, etc. **Q:** Is bias =0, is RMS = CRLB$^{1/2}$, etc.

# Appendix 9: Generative AI for speech ↬ a diffusion model

Consider an 80-band mel-spectrogram of a sample of speech of a female speaker saying: "in being comparatively modern".



Original speech



Generated speech

The speech samples are generated with a score-based model, of which the training stage is maximising the data likelihood $p_{model}(\mathbf{x})$

Observe that the generated sample is very close to the original one

- Diffusion models are built upon Stochastic Differential Equation (SDE) functions
- They are solving an SDE process, which is usually either a variance preserving or exploding SDE function
- The performance of each SDE is different for a different task

# Appendix 10: Minimising KL–divergence and cross–entropy is equivalent to maximising the likelihood

The **KL-divergence** is a common loss function for training neural network (NN) based generative models, such as VAE, GAN, and diffusion models.

$$D_{KL}(p_{\boldsymbol{\theta}}||q_{\hat{\boldsymbol{\theta}}}) = E_{p_{\boldsymbol{\theta}}} log \frac{p_{\boldsymbol{\theta}}(x)}{q_{\hat{\boldsymbol{\theta}}}(x)}$$

where $p_{\boldsymbol{\theta}}$ is the probability distribution of the true labels and $q_{\hat{\boldsymbol{\theta}}}$ is the probability distribution of e.g. deep neural network (DNN) predictors. In other words, $p = p_{labels}$ and $q = q_{model}$.

- KL divergence measures the **dissimilarity** between $p_{\boldsymbol{\theta}}$ and $q_{\hat{\boldsymbol{\theta}}}$.
- To bring $q_{\hat{\boldsymbol{\theta}}}$ closer to the true $p_{\boldsymbol{\theta}}$, we minimize KL-divergence w.r.t. $\hat{\boldsymbol{\theta}}$.

$$\textbf{Goal}: \text{Minimize } D_{KL}(p_{\boldsymbol{\theta}}||q_{\hat{\boldsymbol{\theta}}}) = arg\min_{\hat{\boldsymbol{\theta}}} \sum_{\mathbf{x}} p_{\boldsymbol{\theta}}(x_i) log\left(\frac{p_{\boldsymbol{\theta}}(x_i)}{q_{\hat{\boldsymbol{\theta}}(x_i)}}\right)$$

$$= arg\min_{\hat{\boldsymbol{\theta}}} \sum_{\mathbf{x}} p_{\boldsymbol{\theta}}(x_i) log\left(p_{\boldsymbol{\theta}}(x_i) - q_{\hat{\boldsymbol{\theta}}}(x_i)\right) = arg\min_{\hat{\boldsymbol{\theta}}} - \sum_{\mathbf{x}} p_{\boldsymbol{\theta}}(x_i) log\left(q_{\hat{\boldsymbol{\theta}}}(x_i)\right)$$

# Appendix 10, contd.: Minimising KL–divergence and cross–entropy is equivalent to maximising the likelihood

**Cross-entropy**, $H(p_{\boldsymbol{\theta}}, q_{\hat{\boldsymbol{\theta}}})$, is another important *objective function* for training NNs, especially for classification purposes, and is given by

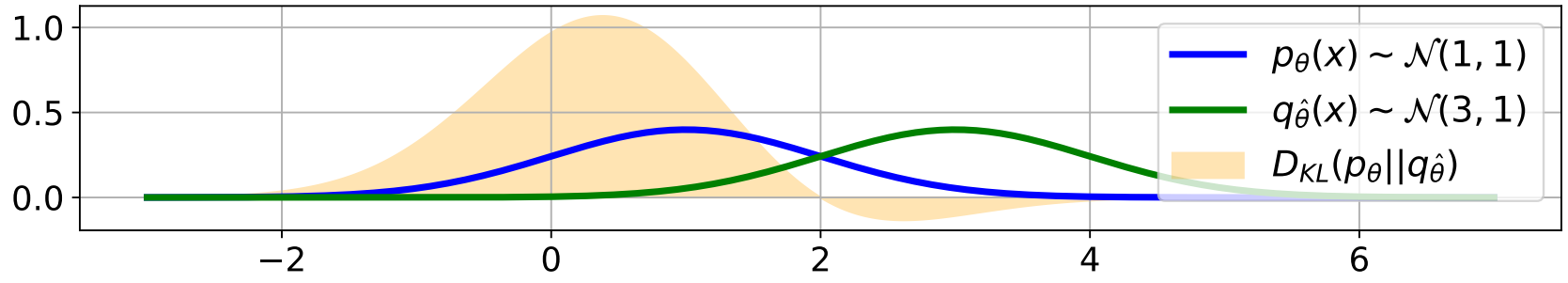
$$H(p_{\boldsymbol{\theta}}, q_{\hat{\boldsymbol{\theta}}}) = E_{p_{\boldsymbol{\theta}}} \log q_{\hat{\boldsymbol{\theta}}}(x) = H(p_{\boldsymbol{\theta}}) + D_{KL}(p_{\boldsymbol{\theta}} || q_{\hat{\boldsymbol{\theta}}})$$

**Goal:** minimize $H(p_{\boldsymbol{\theta}}, q_{\hat{\boldsymbol{\theta}}}) = arg\min_{\hat{\boldsymbol{\theta}}} - \sum_{\mathbf{x}} p_{\boldsymbol{\theta}}(x_i) log\left(q_{\hat{\boldsymbol{\theta}}}(x_i)\right)$

◦ In classification problems, the true distribution, $p_{\boldsymbol{\theta}}(y|x_i)$, is one-hot encoded as

$$p_{\boldsymbol{\theta}}(y|x_i) = \begin{cases} 1, & \text{if } y = y_i \\ 0, & \text{otherwise} \end{cases}$$

☞ The goal becomes $arg\min_{\hat{\boldsymbol{\theta}}} - \sum_{\mathbf{x}} log\left(q_{\hat{\boldsymbol{\theta}}}(y_i|x_i)\right)$, **which is precisely the objective of MLE** (min of the negative log-likelihood = max likelihood).



# Appendix 11: Constrained optimisation using Lagrange multipliers

Consider a two-dimensional problem:

$$\text{maximize} \qquad \underbrace{f(x,y)}_{function\ to\ max/min}$$

$$\text{subject to} \qquad \underbrace{g(x,y)=c}_{constraint}$$

☞ **We look for point(s) where curves $f$ & $g$ touch (but do not cross).**

In those points, the tangent lines for $f$ and $g$ are parallel $\Rightarrow$ so too are the gradients $\boldsymbol{\nabla_{x,y} f \parallel \lambda \nabla_{x,y} g}$, where $\lambda$ is a scaling constant.

Although the two gradient vectors are parallel they can have different magnitudes!

Therefore, we are looking for $\max$ or $\min$ points $(x,y)$ of $f(x,y)$ for which

$$\nabla_{x,y} f(x,y) = -\lambda \nabla_{x,y} g(x,y) \quad \text{where } \nabla_{x,y} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) \text{ and } \nabla_{x,y} g = \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}\right)$$

We can now combine these conditions into one equation as:

$$F(x,y,\lambda) = f(x,y) - \lambda\big(g(x,y) - c\big) \quad \text{and solve} \quad \nabla_{x,y,\lambda} F(x,y,\lambda) = \mathbf{0}$$

**Obviously,** $\nabla_\lambda F(x,y,\lambda) = 0 \quad \Leftrightarrow \quad g(x,y) = c$

# App. 11: Method of Lagrange multipliers in a nutshell

## max/min of a function $f(x,y,z)$ where $x,y,z$ are coupled

Since $x,y,z$ **are not independent** there exists a constraint $g(x,y,z)=c$

**Solution:** Form a new function

$F(x,y,z,\lambda)=f(x,y,z)-\lambda\big(g(x,y,z)-c\big)$ and calculate $F'_x,F'_y,F'_z,F'_\lambda$

Set $F'_x,F'_y,F'_z,F'_\lambda=0$ and solve for the unknown $x,y,z,\lambda$.

**Example 13: Economics** Two factories, A and B make TVs, at a cost

$f(x,y)=6x^2+12y^2$ where $x=\#TV\ in\ A$ & $y=\#TV\ in\ B$

**Task:** Minimise the cost of producing 90 TVs, by finding optimal numbers of TVs, $x$ and $y$, produced respectively at factories A and B.

**Solution:** The constraint $g(x,y)$ is given by (x+y=90), so that

$$\boldsymbol{F(x,y,\lambda)=6x^2+12y^2-\lambda(x+y-90)}$$

Then: $F'_x=12x-\lambda$, $F'_y=24y-\lambda$, $F'_\lambda=-x-y+90$, and we need to set $\nabla F=\mathbf{0}$ in order to find $\min/\max$.

☞ Upon setting $[F'_x,F'_y,F'_\lambda]=\mathbf{0}$ we find $x=60,y=30,\lambda=720$

# Notes:

○

# Notes:

○