


DATA DRIVEN TECHNIQUES

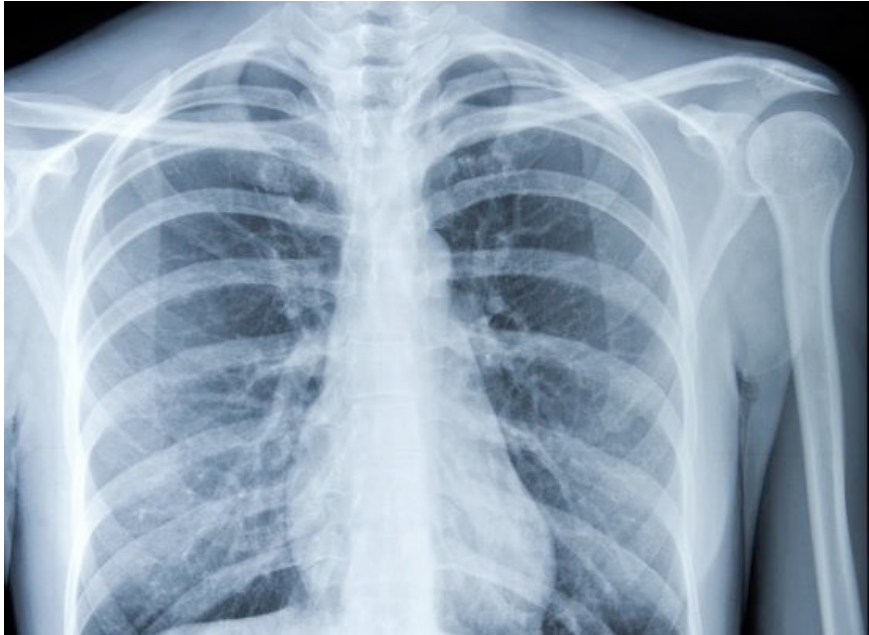
Lecture 1

General - Introduction to Probability Theory

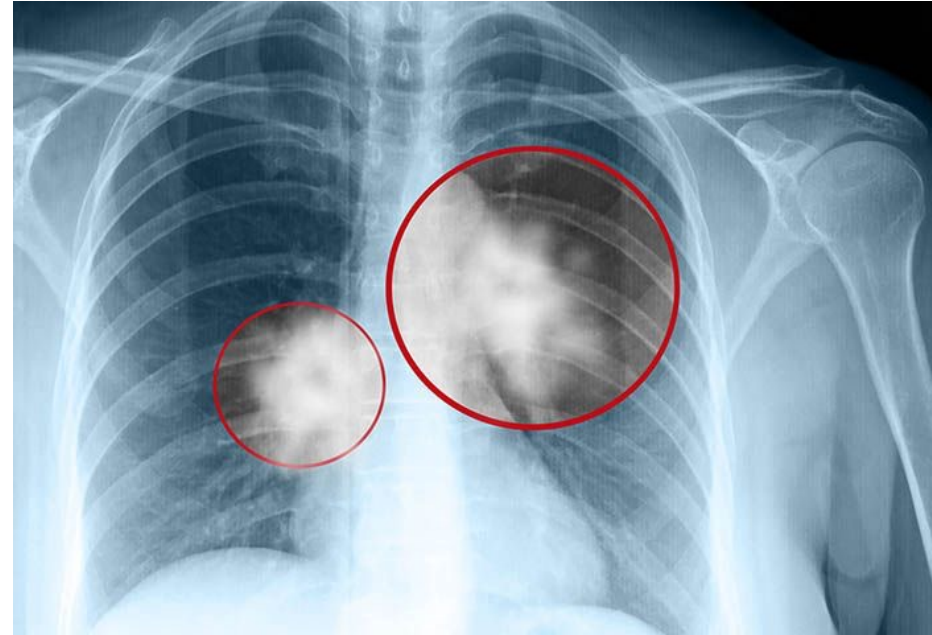


Applications of Neural Networks

Healthy



Problematic



Binary Classification

Design neural network that uses X-ray images to decide between healthy and cancerous lungs

Dataset CIFAR-10

airplane



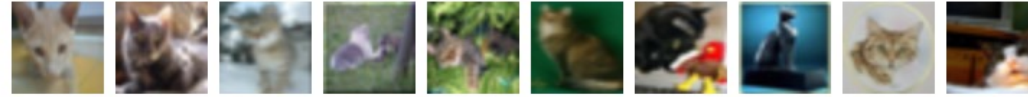
automobile



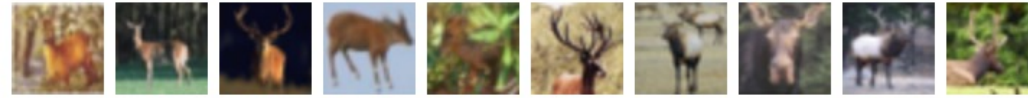
bird



cat



deer



dog



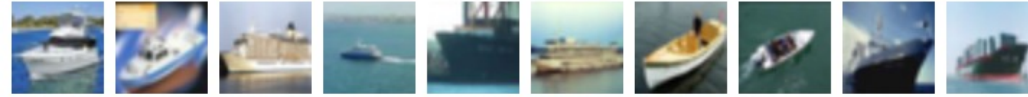
frog



horse



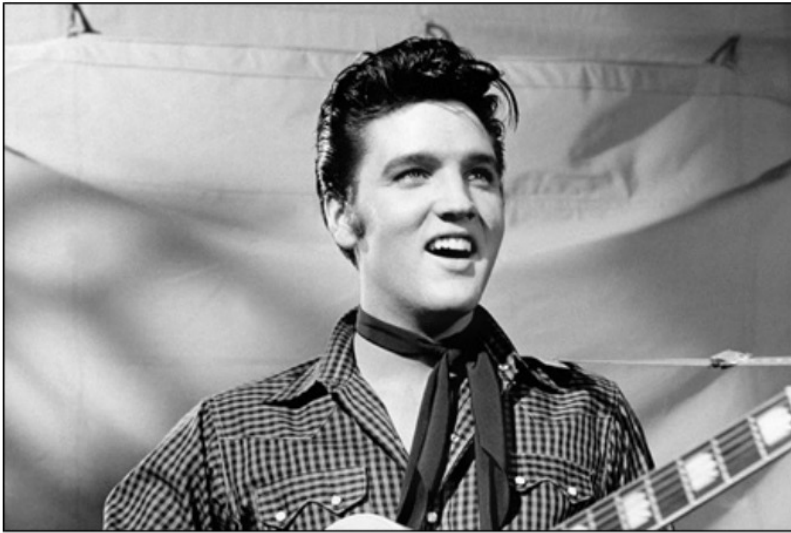
ship



truck



Design neural network that distinguishes between the ten categories



Coloring of gray-scale images

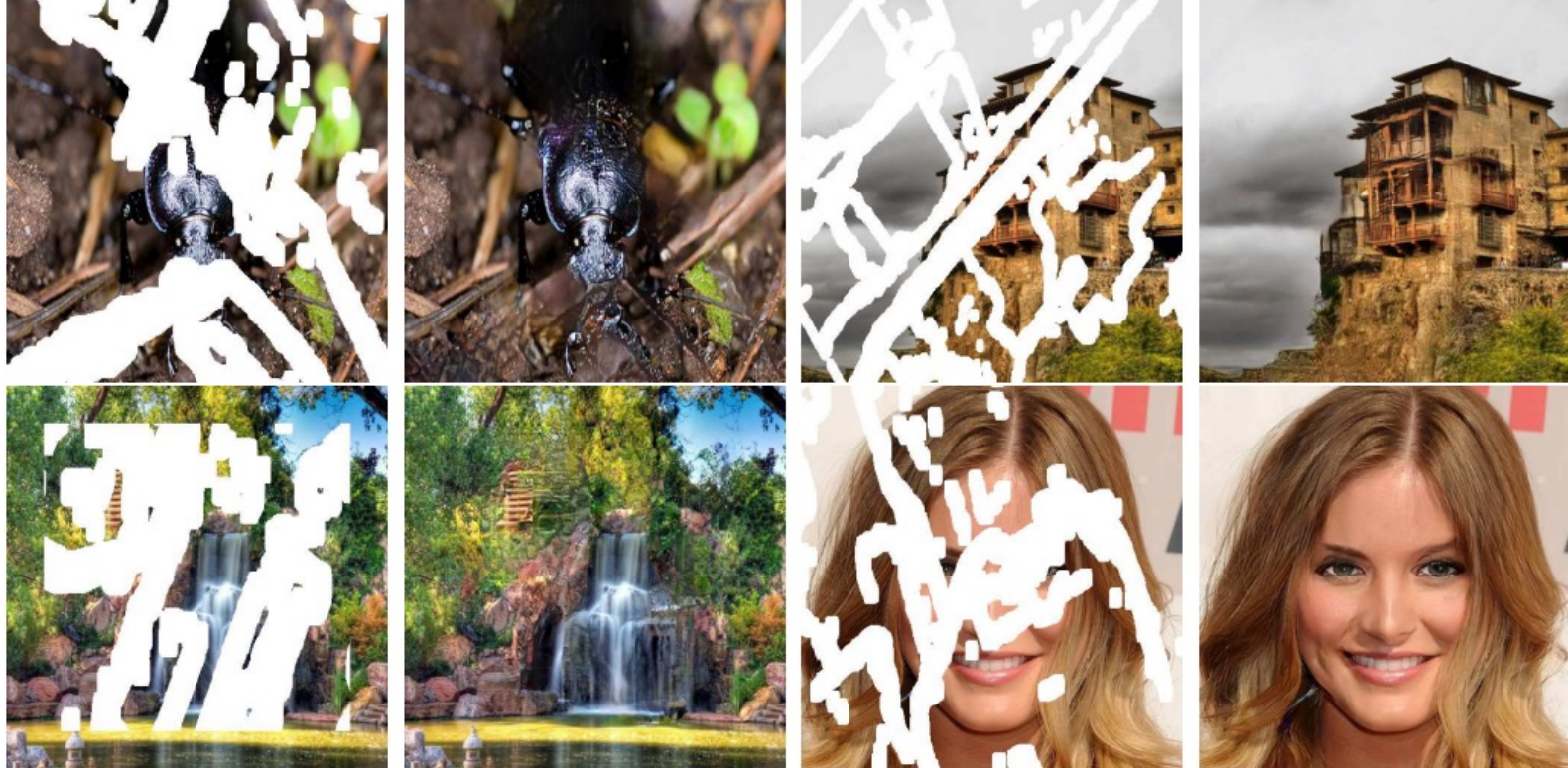


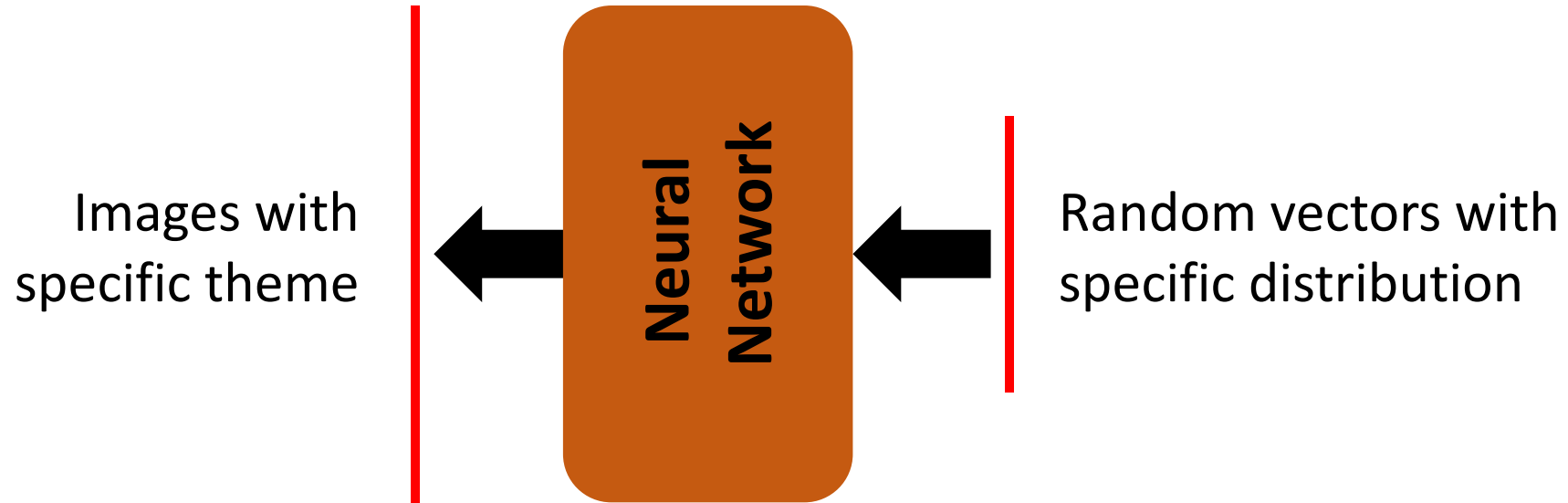
Image restoration (inpainting)



Increase image resolution (super-resolution)

Generative (Adversarial) Networks (GANs)

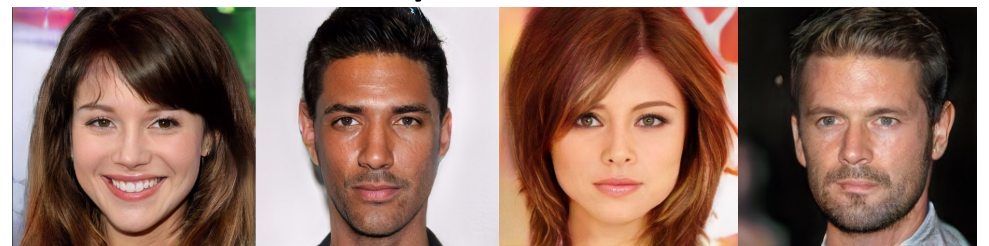
Neural networks that can generate realizations of random quantities that follow a specific (known or unknown) density



CelibA-HD



Synthetic



Syllabus

- Basic Probability theory
- Learning algorithms for equation solving and function optimization.
Analysis and fair comparison methods
- Neural Networks, basic property, categories, design (training)
 - Hypothesis testing, Decision making, Classification
Bayesian methods, Optimum schemes (Statistical theory)
 - Bayes consistent network design. Data-driven decisions for Markov processes
- Realization of random variables, Classical methods, Methods based on transformations, Generative networks, Adversarial (GANs) and non-adversarial design of generative networks. Probability density vs generative modeling for random data on manifolds
- Parameter estimation, Bayesian and non-Bayesian estimators (Statistical Theory)
Data-driven parameter estimation, Generative models for efficient solution of high-dimensional inverse problems

Syllabus (cont)

- Data-driven estimation of conditional expectations, application to stochastic optimization problems (optimal stopping, reinforcement learning)
- Clustering, K-means, Gaussian mixtures, Expectation/maximization
- Kernels and vector spaces, Mercer kernels, Nonlinear function approximation using kernels

Grading

- 4 homework assignments, each must be completed within a week. **Deadline is strict!**
- 1 final homework (as take-home exam), must be completed within 2 days
- Each homework has 2-4 problems. All problems are considered equivalent in difficulty (even if they are not). Each problem receives a grade between 0 and 10
The homework grade is **the average of the grades of the problems**
- Final grade is **the average of the grades of the 5 homeworks** ~~the average of the grades of the 5 homeworks~~

AXIOMATIC DEFINITION OF PROBABILITY

(10)

Probability Space = $\{\Omega, \mathcal{F}, \mathbb{P}\}$.

Ω : A set with ANY elements.

\mathcal{F} : A set with elements that are SUBSETS of Ω
Has special structure.

\mathbb{P} : Mapping from \mathcal{F} to the interval $[0,1]$
 \mathbb{P} "measures" the significance of each element in \mathcal{F}
and assigns a number between 0 & 1 (probability)
Has special structure.

Ω : Sample Space — Set with any elements

(11)

\mathcal{G} : Has special structure known as σ -algebra

1) $\Omega, \emptyset \in \mathcal{G}$

2) If $A \in \mathcal{G}$ then $A^c \in \mathcal{G}$

3) If $A_1, A_2 \in \mathcal{G}$ then $A_1 \cup A_2 \in \mathcal{G}$

We can prove
 $A_1 \cap A_2 \in \mathcal{G}$

4) If $A_1, A_2, \dots \in \mathcal{G}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{G}$

The elements of \mathcal{G} are called events and they are the subsets of Ω that we like to "measure" — i.e. assign probability

P : Probability function.

Defined ONLY for $A \in \mathcal{G}$ it maps A to $[0, 1]$

$$P: \mathcal{G} \rightarrow [0, 1] \quad P(A) \in [0, 1].$$

$$1) P(\Omega) = 1, P(\emptyset) = 0.$$

$$2) A_1, A_2 \in \mathcal{G}, A_1 \cap A_2 = \emptyset, P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Can prove A_1, \dots, A_N , with $A_i \cap A_j = \emptyset$ when $i \neq j$ then

$$P\left(\bigcup_{i=1}^N A_i\right) = \sum_{i=1}^N P(A_i)$$
 Not possible to prove when $N = \infty$

$$3) A_1, A_2, \dots \in \mathcal{G}, A_i \cap A_j = \emptyset, P\left(\bigcup_{m=1}^{\infty} A_m\right) = \sum_{m=1}^{\infty} P(A_m)$$



Use ANR number and type of
of measurement devices

→ Apply Physics to compute
the result of the next throw.

Deterministic description

Suppose we apply the analysis and predict "Tail" --- But the result
is "Head"

Extremely complicated phenomenon to describe analytically and
very sensitive to measurement errors.

Probability Theory describes it as: $P(\text{Head}) = \frac{1}{2}$, $P(\text{Tail}) = \frac{1}{2}$

Is it possible with such simplistic descriptions of
complicated phenomena to experience any gain?

YES!!

Consider the problem of predicting whether the Wall St stock market will increase or decrease and we bet on our prediction.

- If I have a deterministic description which is accurate I can predict exactly what will happen the next day. I can become rich in a single day!
- If I have a probabilistic mechanism with $P(\text{correct}) = 0.51$ and $P(\text{wrong}) = 0.49$, then I can still become rich only it takes longer.

Note that if I don't know anything then
 $P(\text{correct}) = P(\text{wrong}) = 0.5$.

RANDOM VARIABLES

(15)

"Nature" can select elements $\theta \in \Theta$

For each $\theta \in \Theta$, Nature is giving us a real number $X(\theta)$

$X: \Theta \rightarrow \mathbb{R}$ is a function from the sample space Θ to the Reals \mathbb{R} .

ARE WE INTERESTED IN ALL FUNCTIONS? NO

Nature sends \downarrow m.y. selection \downarrow

$$A_x = \{ \theta : X(\theta) \leq x \}$$

Comparison \rightarrow simplest operation between two reals.

Interested in "measuring" $A_x, \forall x$. For this to be possible we need $A_x \in \mathcal{G}, \forall x$.

This is true because I can measure ONLY what is in \mathcal{G}

The functions from Θ to \mathbb{R} that satisfy

$\{ \theta : X(\theta) \leq x \} \in \mathcal{G}, \forall x$ are called Measurable or Random Variables

Σ

Random variables are functions

(16)

1) from $\Theta \rightarrow \mathbb{R}$.

2) $\{\theta = \chi(\theta) \in \mathbb{R}\} \in \mathcal{F}$, ~~if~~

It can be shown that the general form of a random variable is

$$\chi(\theta) = \sum_i c_i \mathbb{1}_{A_i}(\theta)$$

where $c_i \in \mathbb{R}$
and $A_i \in \mathcal{F}$

$$\mathbb{1}_A(\theta) = \begin{cases} 1 & \theta \in A \\ 0 & \theta \notin A \end{cases}$$

indicator of set A .

EXAMPLE OF A FUNCTION THAT IS NOT RANDOM VARIABLE

$$\Theta = [0, 1], \quad \mathcal{F} = \{\emptyset, [0, 1], [0, 0.5], [0.5, 1]\}$$

$$\chi(\theta) = 2 \mathbb{1}_{[0, 0.5]}(\theta) + 3 \mathbb{1}_{[0.5, 1]}(\theta)$$

$$\{\theta : \chi(\theta) = 2.5\} = [0, 0.5] \notin \mathcal{F}$$

I cannot give probability

Therefore NOT a random variable

If $X(\theta)$ is a r.v then

(1)

$\{\theta: X(\theta) \leq x\} \in \mathcal{F}$. I can assign probability $P(X(\theta) \leq x) = F_X(x)$
Cumulative Distribution function

a) $F_X(x)$ increasing b) $F_X(x) \rightarrow 0$ for $x \rightarrow -\infty$

c) $F_X(x) \rightarrow 1$ for $x \rightarrow +\infty$

$f_X(x) = \frac{dF_X(x)}{dx} \geq 0$ probability density function

$$\int f_X(x) dx = 1$$

EXAMPLES

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

Gaussian $N(\mu, \sigma^2)$
mean μ variance σ^2

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

Uniform on $[\alpha, \beta]$

It is very important to be able to assign probability to the event $\textcircled{2}$
 $\chi(\theta) = x$. When $f_\chi(x)$ is continuous we know $\mathbb{P}(\chi(\theta) = x) \approx 0$

To be able to avoid this trivial information we consider a DIFFERENTIAL
INTERVAL $(x-dx, x]$ where dx is the differential we use in integration

$$\mathbb{P}(x-dx \leq \chi(\theta) \leq x) = \mathbb{P}(\chi(\theta) \in dx) = f_\chi(x) dx$$

Very helpful for going from properties of events to properties of densities.

- A random variable is completely known if we know $f_\chi(x)$.

VERY DIFFICULT INFORMATION TO HAVE IN PRACTICE (Most of the time)
not necessary

MEAN / STOCHASTIC AVERAGE & VARIANCE

$\chi(\theta)$ is a function.

Interested in defining a deterministic representative
number of $\chi(\theta)$

We call mean value of $\chi(\theta)$, denote

$$\bar{X} = \mathbb{E}[\chi(\theta)] = \int x f_\chi(x) dx$$

the probability by
which $\chi(\theta)$ takes the value x ,

We average all the values weighted by their corresponding probability

If $x(\theta_1), \dots, x(\theta_N)$ N realizations, then

$$\bar{x} = \mathbb{E}[x(\theta)] \approx \frac{x(\theta_1) + x(\theta_2) + \dots + x(\theta_N)}{N}$$

LAW OF LARGE
NUMBERS (LLN)

\approx becomes $=$ as $N \rightarrow \infty$ (under mild conditions)

From a random variable $x(\theta)$ I can generate another random variable $y(\theta)$: Select $G(x)$ deterministic function and define

$$y(\theta) = G(x(\theta)).$$

If $x(\theta_1), x(\theta_2), \dots, x(\theta_N)$ realizations of $x(\theta)$ then

$y(\theta_1) = G(x(\theta_1)), \dots, y(\theta_N) = G(x(\theta_N))$, realizations of $y(\theta)$

I can also define the mean of $y(\theta)$

$$\bar{y} = \mathbb{E}[y(\theta)] = \mathbb{E}[G(x(\theta))] = \int G(x) f_x(x) dx$$

Application of LLN

$$\bar{y} \approx \frac{y(\theta_1) + \dots + y(\theta_N)}{N} = \frac{G(x(\theta_1)) + \dots + G(x(\theta_N))}{N}.$$

If $G(x)$ is parametrized $G(x, \theta)$ parameters

then
$$U(\theta) = \int G(x, \theta) f_{\theta}(x) dx$$

$$\approx \frac{G(x(\theta_1), \theta) + \dots + G(x(\theta_n), \theta)}{n}$$

(5)

If $\chi(\theta)$ is a random variable with density $f_\chi(x)$ and mean value $\bar{\chi}$, then we like to define a measure of how much $\chi(\theta)$ fluctuates around its mean. We call variance the following mean square metric.

$$\sigma_\chi^2 = \int (x - \bar{\chi})^2 f_\chi(x) dx = \mathbb{E}[(\chi(\theta) - \bar{\chi})^2] \approx \frac{(\chi(\theta_1) - \bar{\chi})^2 + \dots + (\chi(\theta_n) - \bar{\chi})^2}{n}$$

Clearly the smaller the σ_χ^2 the more "deterministic" the random variable is. If in particular $\sigma_\chi^2 = 0$ then this means that $\chi(\theta) = \bar{\chi}$. In other words $\chi(\theta)$ is deterministic.

A deterministic quantity can be regarded as a random variable with variance equal to 0.

This suggests that deterministic quantities are special subset of random variables and we can treat them all at the same time.

(6)

It is possible Nature with every selection of $\theta \in \Theta$ to give us two real numbers $\theta \mapsto \chi_1(\theta), \chi_2(\theta)$

We are interested in "measuring" the set

$$\{\theta : \chi_1(\theta) \leq x_1 \text{ AND } \chi_2(\theta) \leq x_2\} = \underbrace{\{\theta : \chi_1(\theta) \leq x_1\}}_{\in \mathcal{G}} \cap \underbrace{\{\theta : \chi_2(\theta) \leq x_2\}}_{\in \mathcal{G}} \in \mathcal{G}$$

$$P(\chi_1(\theta) \leq x_1, \chi_2(\theta) \leq x_2) = F_{\chi_1, \chi_2}(x_1, x_2) \quad \text{Joint cumulative distribution}$$

$$F_{\chi_1}(x_1) = \lim_{x_2 \rightarrow \infty} F_{\chi_1, \chi_2}(x_1, x_2) \quad F_{\chi_2}(x_2) = \lim_{x_1 \rightarrow \infty} F_{\chi_1, \chi_2}(x_1, x_2) \quad \underline{\text{marginals}}$$

$$f_{\chi_1, \chi_2}(x_1, x_2) = \frac{\partial^2 F_{\chi_1, \chi_2}(x_1, x_2)}{\partial x_1 \partial x_2}$$

Joint probability density

$$\underline{f_{\chi_1}(x_1) = \int f_{\chi_1, \chi_2}(x_1, x_2) dx_2} \quad \leftarrow \text{Integrate out the unwanted variable} \quad \underline{\text{marginal}}$$

If $f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$ then $X_1(\theta), X_2(\theta)$ are called
INDEPENDENT

7

CORRELATION of $X_1(\theta), X_2(\theta)$

$$r_{X_1, X_2} = E[(X_1(\theta) - \bar{X}_1)(X_2(\theta) - \bar{X}_2)] = \iint (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

Realizations come in pairs

$$\theta_1: (X_1(\theta_1), X_2(\theta_1))$$

$$\theta_2: (X_1(\theta_2), X_2(\theta_2)) \quad r_{X_1, X_2} = \frac{(X_1(\theta_1) - \bar{X}_1)(X_2(\theta_1) - \bar{X}_2) + \dots + (X_1(\theta_N) - \bar{X}_1)(X_2(\theta_N) - \bar{X}_2)}{N}$$

$$\theta_N: (X_1(\theta_N), X_2(\theta_N))$$

When $r_{X_1, X_2} = 0$ then UNCORRELATED

INDEPENDENT \longrightarrow UNCORRELATED
 \longleftarrow not necessarily true

It is true when jointly Gaussian

With every θ Nature is giving us k real numbers

(8)

$\theta: x_1(\theta), x_2(\theta), \dots, x_n(\theta)$ or for $\theta: X(\theta)$ vector of length k

$x_1(\theta), \dots, x_k(\theta)$ must be random variables

$$\{\theta: x_1(\theta) \leq x_1, \dots, x_k(\theta) \leq x_k\} = \{\theta: X(\theta) \leq x\} \in \mathcal{G}$$

$P(X(\theta) \leq x) = F_X(x)$, Joint cumulative distribution of $x_1(\theta), \dots, x_k(\theta)$
or the cumulative distribution of the random vector $X(\theta)$

$$f_X(x) = \frac{\partial^k F_X(x)}{\partial x_1 \dots \partial x_k}$$

Joint probability density of $x_1(\theta), \dots, x_k(\theta)$
or probability density of vector $X(\theta)$

$$f_{x_i}(x_i) = \int \dots \int f_X(x) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_k$$

Integrate out unwanted variables

$$f_{x_i, x_j}(x_i, x_j) = \int \dots \int f_X(x) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_{j-1} dx_{j+1} \dots dx_k$$

$$x_1(\theta), \dots, x_n(\theta) \text{ INDEPENDENT : } f_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n f_{x_i}(x_i)$$

9

Mean $\bar{x} = E[x(\theta)] =$

$$\int x f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} = \begin{bmatrix} \int x_1 f_{x_1}(x_1) dx_1 \\ \vdots \\ \int x_n f_{x_n}(x_n) dx_n \end{bmatrix} \approx \frac{x(\theta_1) + \dots + x(\theta_N)}{N}$$

Covariance Matrix

$$\Sigma_{\mathbf{x}} = E[(x(\theta) - \bar{x})(x(\theta) - \bar{x})^T] = \int \dots \int (x - \bar{x})(x - \bar{x})^T f_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}$$

$$(\Sigma_{\mathbf{x}})_{ij} = E[(x_i(\theta) - \bar{x}_i)(x_j(\theta) - \bar{x}_j)] = r_{x_i x_j} \quad \text{correlation between } x_i(\theta), x_j(\theta)$$

Diagonal elements are variances

$$\Sigma_{\mathbf{x}} \approx \frac{(x(\theta_1) - \bar{x})(x(\theta_1) - \bar{x})^T + \dots + (x(\theta_N) - \bar{x})(x(\theta_N) - \bar{x})^T}{N} \quad \underline{\underline{LLN}}$$

GAUSSIAN RANDOM VECTOR

10

$$\mathcal{X} \sim \mathcal{N}(\mu, \Sigma)$$

μ : mean vector

Σ : Covariance matrix

$$-\frac{1}{2} (\mathcal{X} - \mu)^T \Sigma^{-1} (\mathcal{X} - \mu)$$

$$f_{\mathcal{X}}(\mathcal{X}) = \frac{e^{-\frac{1}{2} (\mathcal{X} - \mu)^T \Sigma^{-1} (\mathcal{X} - \mu)}}{\sqrt{(2\pi)^K |\Sigma|}}$$

$|\Sigma|$: determinant of Σ

K : length of \mathcal{X} .