Probability

Arthur J. Redfern
axr180074@utdallas.edu
Sep 10, 2018

Outline

- Motivation
- Probability spaces
- Random variables
- Random processes
- Information theory
- References

Motivation

Information

- Probability is the math that describes information
  - This course uses xNNs to extract information from data
  - This course uses xNNs to generate data from information

- Examples
  - Understanding machine learning as information extraction from training data to apply to the problem of information extraction from testing data
  - Understanding the flow of information through the network and implications of network design
  - Weight initialization as the application of known information
  - Error functions to quantify how well the information extraction process worked
  - Compressing filter coefficients and feature maps towards an information bound

Probability Spaces

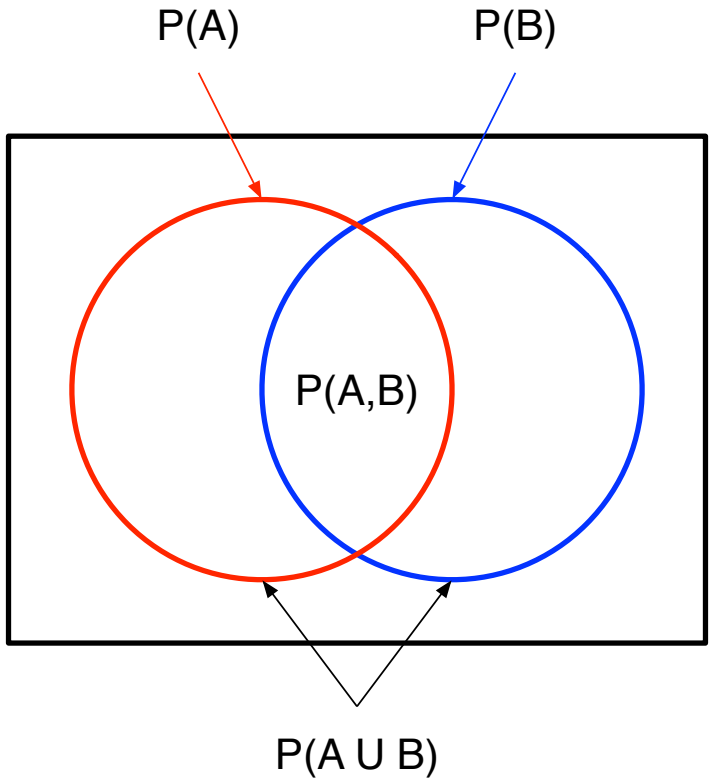
Probability Space Definition (S, E, P)

- A sample space S of all possible outcomes
  - Think: S is all possible outcomes of an experiment
  - Ex: flipping a coin 2x and recording heads (H) or tails (T) for each flip
  - S = {HH, HT, TH, TT}
- An event space E where each event is a set of 0 or more outcomes from the sample space
  - Think: E is all possible subsets of the sample space S (including nothing and everything)
  - E = {
    Ø, // null subset
    {HH}, {HT}, {TH}, {TT}, // all subsets of 1 outcome
    {HH, HT}, {HH, TH}, {HH, TT}, {HT, TH}, {HT, TT}, {TH, TT}, // all subsets of 2 outcomes
    {HH, HT, TH}, {HH, HT, TT}, {HH, TH, TT}, {HT, TH, TT} // all subsets of 3 outcomes
    {HH, HT, TH, TT} } // sample space subset
- A probability measure function P: E → [0, 1] that satisfies
  - $P(A) \in R$ and $P(A) \geq 0$ for all events $A \in E$
  - $P(S) = 1$
  - $P(\cup_i A_i) = \Sigma_i P(A_i)$ for mutually exclusive events A_i
  - Think: P is a function that assigns probabilities to subsets of the sample space

Events

- Notation
  - 1 event A, 2 events A and B
  - K events $\{A_0, ..., A_{K-1}\}$
- Single
  - The probability of an event occurring — $P(A) \in [0, 1]$
  - The probability of an event not occurring — $P(A^c) = 1 - P(A)$ — A^c denoting not A
- Joint
  - The probability of events A and B occurring — $P(A, B)$ — also written as $P(A \cap B)$
  - If $A \subseteq B$ — $P(A, B) = P(A)$
  - If A and B are independent — $P(A, B) = P(A)\,P(B)$
- Union
  - The probability of event A or B occurring — $P(A \cup B) = P(A) + P(B) - P(A, B)$
  - If A and B are mutually exclusive — $P(A \cup B) = P(A) + P(B)$

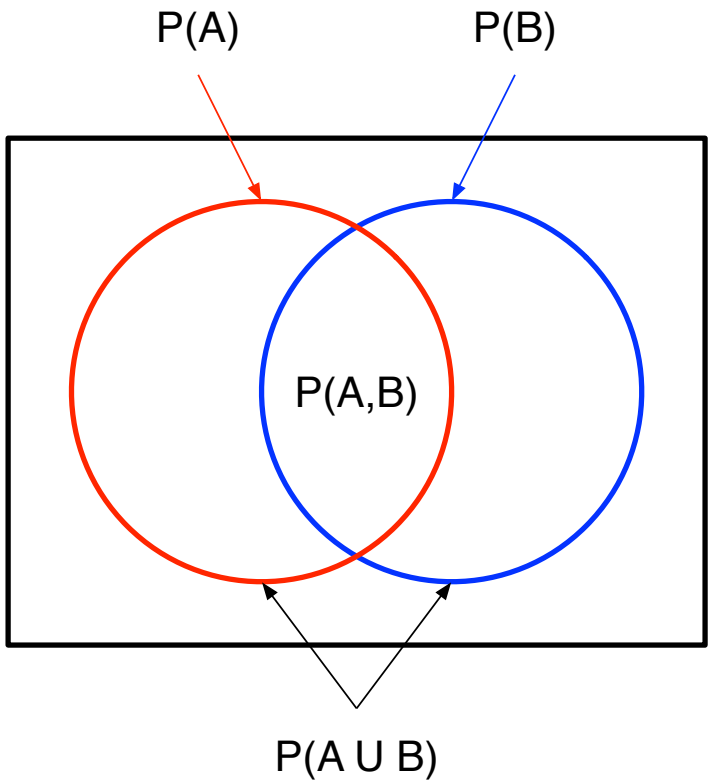
Events



Events

- Conditional
  - The probability of event A given event B
    If A and B are independent
    $P(A|B) = P(A, B) / P(B)$
    $P(A|B) = P(A)$
  - Bayes' theorem
    $P(A|B) = P(B|A)\, P(A) / P(B)$
  - Chain rule of probability
    $P(A_0, ..., A_{K-1}) = P(A_0|A_1, ..., A_{K-1})P(A_1, ..., A_{K-1})$
  - Can recursively apply to 2nd term on RHS

Events



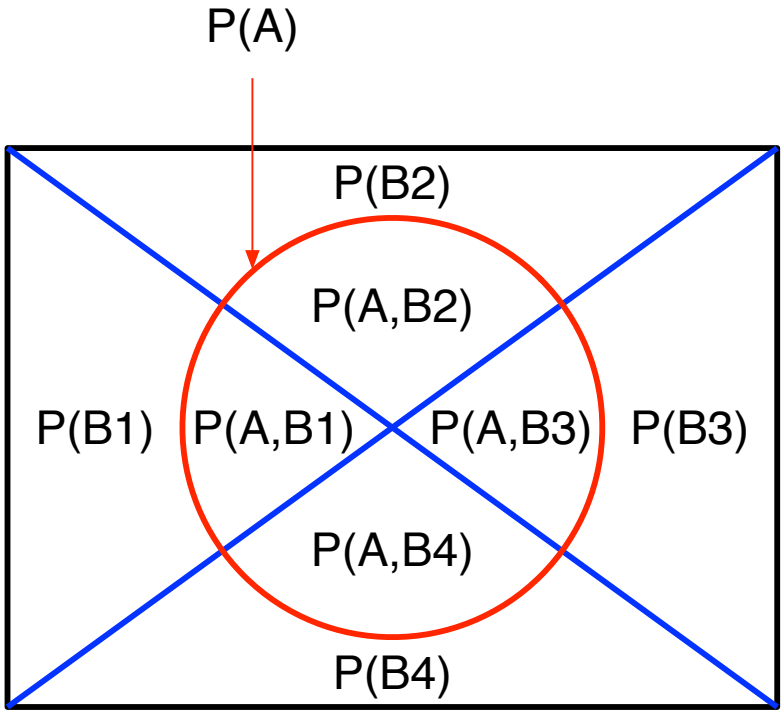
$$P(A \mid B) = P(A, B) / P(B)$$
$$P(B \mid A) = P(A, B) / P(A)$$

Events

- Law of total probability
  - Let $\{B_0, ..., B_{K-1}\}$ be a set of disjoint events whose union is the full event space
  - Let A be an event in the same event space
  - Marginal probability of A

$$P(A) = \Sigma_k P(A, B_k) = \Sigma_k P(A|B_k) P(B_k)$$

Events



$$P(A) = \text{sum}_k\ P(A, B_k)$$
$$= \text{sum}_k\ P(A \mid B_k)\ P(B_k)$$

Random Variables

Discrete

- A discrete random variable is a function X with a finite or countably infinite range that maps outcomes s from the sample space S to numbers $x \in R$

$$X(s) = x_k$$

- x_k is a realization of X
- Note that a random variable is not random and it's not a variable
  - The outcome of the experiment s is random
  - The mapping $X(s) = x_k$ by the random variable (function) to a real number is deterministic

Discrete

- A discrete random variable is described by it's probability mass function that specifies the probability that it takes on a specific value or it's cumulative distribution function that specifies the probability that it's value falls within an interval
- Probability mass function
  - Single
    $p_X(x_k) = P(X(s) = x_k)$
    where $\Sigma_k\ p_X(x_k) = 1$
  - Joint and conditional
    $p_{X,Y}(x_j, y_k) = p_{X|Y}(x_j|y_k)\ p_Y(y_k) = p_{Y|X}(y_k|x_j)\ p_X(x_j)$
  - Marginal
    $p_X(x_j) = \Sigma_k\ p_{X,Y}(x_j, y_k) = \Sigma_k\ p_{X|Y}(x_j|y_k)\ p_Y(y_k)$
  - Independent X and Y
    $p_{X,Y}(x_j, y_k) = p_X(x_j)\ p_Y(y_k)$
    $p_{X|Y}(x_j|y_k) = p_X(x_j)$
- Cumulative distribution function
  - Single
    $F_X(x_k) = P(X(s) \leq x_k) = \Sigma_{xj \leq xk}\ p_X(x_j)$

Continuous

- A continuous random variable is a function X with an uncountably infinite range that maps outcomes s from the sample space S to numbers $x \in R$

$$X(s) = x$$

- x is a realization of X
- Note that a random variable is (still) not random and it's (still) not a variable
  - The outcome of the experiment s is random
  - The mapping X(s) = x by the random variable (function) to a real number is deterministic

Continuous

- A continuous random variable is described by it's cumulative distribution function that specifies the probability that it's value falls within an interval
  - If the cumulative distribution function is absolutely continuous then it also has a probability density function (this set of slides will assume this is true so we don't have to use the word measure and weird looking integrals)
- Probability density function
  - Single: $\int_a^b p_X(x)\,dx = P(a \le X(s) \le b)$ where $\int p_X(x)\,dx = 1$
  - Joint and conditional: $p_{X,Y}(x, y) = p_{X|Y}(x|y)\,p_Y(y) = p_{Y|X}(y|x)\,p_X(x)$
  - Marginal: $p_X(x) = \int p_{X,Y}(x, y)\,dy = \int p_{X|Y}(x|y)\,p_Y(y)\,dy$
  - Independent X and Y: $p_{X,Y}(x, y) = p_X(x)\,p_Y(y)$; $p_{X|Y}(x|y) = p_X(x)$
- Cumulative distribution function
  - Single: $F_X(x) = \int^x p_X(t)\,dt$

Expected Value

- Expected value is a linear operator that maps functions of random variables to a probability weighted average of all events (shown here for a discrete random variable)

$$E[f(X(s))] = \Sigma_k \, p_X(x_k) \, f(x_k)$$

- Scalar examples
  - Mean $\mu_X = E[X(s)]$
  - Variance $\sigma_X^2 = E[(X(s) - \mu_X)^2]$
  - Standard deviation σ_X
  - nth order moment about the mean $E[(X(s) - \mu_X)^n]$
  - Covariance (units of X(s)*Y(s)) $cov(X(s), Y(s)) = E[(X(s) - \mu_X)(Y(s) - \mu_Y)] = E[X(s)\,Y(s)] - \mu_X \mu_Y$
  - Correlation ([-1, 1]) $corr(X(s), Y(s)) = cov(X(s), Y(s)) / (\mu_X \mu_Y)$
  - Independent X(s) and Y(s) $cov(X(s), Y(s)) = corr(X(s), Y(s)) = 0$

$cov(X(s), Y(s)) = corr(X(s), Y(s)) = 0$ does not imply independent X(s) and Y(s)

Expected Value

- Vector examples
  - Notation
  - Mean vector

$$\mathbf{x} = [X_0(s), ..., X_{K-1}(s)]^T$$

$$\boldsymbol{\mu}_\mathbf{x} = E[\mathbf{x}] = [E[X_0(s)], ..., E[X_{K-1}(s)]]^T$$

- Matrix examples
  - Covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}} = E[(\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{X})^T]$$

$$\boldsymbol{\Sigma}_{\mathbf{x},\mathbf{x}}(m, k) = E[(X_m(s) - \mu_{Xm})(X_k(s) - \mu_{Xk})]$$

Expected Value

- Linear regression
  - $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{e}$, $\mathbf{e}$ is a 0 mean vector random variable representing measurement error
  - $\mathbf{e} = \mathbf{y} - \mathbf{A}\,\mathbf{x}$
  - $\min_{\mathbf{x}} \mathbf{e}^T\,\mathbf{e} = (\mathbf{y} - \mathbf{A}\,\mathbf{x})^T\,(\mathbf{y} - \mathbf{A}\,\mathbf{x}) = \mathbf{x}^T\,\mathbf{A}^T\,\mathbf{A}\,\mathbf{x} - 2\,\mathbf{y}^T\,\mathbf{A}\,\mathbf{x} + \mathbf{y}^T\,\mathbf{y}$
  - $\mathbf{x}^{hat} = (\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{y}$
- Estimator mean
  - $E[\mathbf{x}^{hat}] = E[(\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{y}]$
    $= E[(\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,(\mathbf{A}\,\mathbf{x} + \mathbf{e})]$
    $= E[(\mathbf{A}^T\,\mathbf{A})^{-1}\,(\mathbf{A}^T\,\mathbf{A})\,\mathbf{x}] + E[(\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{e}]$
    $= E[\mathbf{x}] + (\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,E[\mathbf{e}]$
    $= \mathbf{x}$
- Estimator covariance
  - Substitute $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{e}$ into the $\mathbf{x}^{hat}$ formula to get $\mathbf{x}^{hat} = \mathbf{x} + (\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{e}$ or $\mathbf{x}^{hat} - \mathbf{x} = (\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{e}$
  - $E[(\mathbf{x}^{hat} - \mathbf{x})(\mathbf{x}^{hat} - \mathbf{x})^T] = E[((\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{e})((\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,\mathbf{e})^T] = (\mathbf{A}^T\,\mathbf{A})^{-1}\,\mathbf{A}^T\,E[\mathbf{e}\,\mathbf{e}^T]\,\mathbf{A}\,(\mathbf{A}^T\,\mathbf{A})^{-1} = \sigma_e^2\,(\mathbf{A}^T\,\mathbf{A})^{-1}$

Examples Of Discrete PMFs

- Bernoulli
  - $p_X(x_k)$ $= 1 - p,$ $x_k = 0, p \in [0, 1]$
    $= p,$ $x_k = 1$
    $= 0,$ elsewhere
  - Expectations
    - Mean $= p$
    - Variance $= p\ (1 - p)$
- Uniform
  - $p_X(x_k)$ $= 1 / N,$ $x_k \in \{a, a + 1, ..., b\}, b - a + 1 = N$
    $= 0,$ elsewhere
  - Expectations
    - Mean $= (a + b) / 2$
    - Variance $= (N^2 - 1) / 12$

Examples Of Continuous PDFs

- Uniform
  - $p_X(x)$ $= 1 / (b - a)$, $x \in [a, b]$, a and b finite
    $= 0$, elsewhere
  - Expectations
    - Mean $= (a + b) / 2$
    - Variance $= (b - a)^2 / 12$ Side note: this leads to a famous SNR formula for quantizers
- Gaussian (or normal)
  - $p_X(x)$ $= (1 / (2\pi\sigma^2)^{1/2}) \exp(-(x - \mu_x)^2 / 2\sigma_x^2)$
  - Expectations
    - Mean $= \mu_x$
    - Variance $= \sigma_x^2$

- xNN use: filter coefficient initialization
  - For initialization with a Gaussian distribution it's frequently truncated (limited)

Experiment

- A class generated discrete probability mass function

- Experiment
  - Think of a 2 digit number between 10 and 50
  - Both digits are odd
  - Both digits are different from each other

Experiment

- A class generated discrete probability mass function
- Experiment
  - Think of a 2 digit number between 10 and 50
  - Both digits are odd
  - Both digits are different from each other
- How many people thought of the number 37?

Normalization

- Purpose
  - Take a random variable with an arbitrary distribution and normalize it to 0 mean and unit variance
    - Note that other variations of normalization exist
  - This is used by batch norm layers in CNNs to improve training
  - Note: CNNs use the word norm and normalization a lot for different operations
    - Input data normalization (a variant of what is described here)
    - Normalization layer (operates across feature maps, famous in AlexNet, rarely used now)
    - Batch normalization layer (a variant of what is described here, used in many places to improve training, can frequently be absorbed into convolution for deployment)
    - Group normalization layer (similar purpose to batch normalization, different operation)
    - …
- Normalization
  - $Y(s) = (X(s) - \mu_x) / \sigma_x$
  - $E[Y(s)] = E[(X(s) - \mu_x) / \sigma_x] = (1 / \sigma_x)(E[X(s)] - \mu_x) = (1 / \sigma_x)(\mu_x - \mu_x) = 0$
  - $E[(Y(s))^2] = E[((X(s) - \mu_x) / \sigma_x)^2] = (1 / \sigma_x^2)\, E[(X(s) - \mu_x)^2] = (1 / \sigma_x^2)\, \sigma_x^2 = 1$

Law Of Large Numbers

- Let $X_0(s)$, $X_1(s)$, ... be a sequence of independent identically distributed random variables with $E[X_i(s)] = \mu_X$ and let the sample average be

$$X_{0:K-1}^{bar}(s) = (X_0(s) + ... + X_{K-1}(s))/K$$

- $X_{0:K-1}^{bar}(s)$ converges to μ_X as $K \rightarrow \infty$
  - In probability for the weak law (unlikely outcome probability reduces as $K \rightarrow \infty$)
  - Almost surely for the strong law (pointwise)
- Variants exist that replace the independence constraint with a variance constraint
- The law of large numbers allows the expected value of a random variable with a finite mean to be estimated from it's sample average
  - Note that the sample average is a random variable

Central Limit Theorem

- The central limit theorem describes the distribution of the sample average on the previous slide (a random variable) about μ as $K \rightarrow \infty$
- Let $\{X_0(s), ..., X_{K-1}(s)\}$ be a set of independent identically distributed random variables each with mean μ_X and finite variance σ_X^2, then

$$K^{1/2} \left(X_{0:K-1}{}^{bar}(s) - \mu_X\right) \rightarrow N(0, \sigma_X^2)$$

- $N(0, \sigma^2)$ is 0 mean σ^2 variance Gaussian distribution
  - So $X_{0:K-1}{}^{bar}(s)$ is "close" to $N(\mu_X, \sigma_X^2 / K)$
- Convergence is in distribution (the cdf converges as $K \rightarrow \infty$)
- Variants exist that replace the independent identically distributed condition

Central Limit Theorem

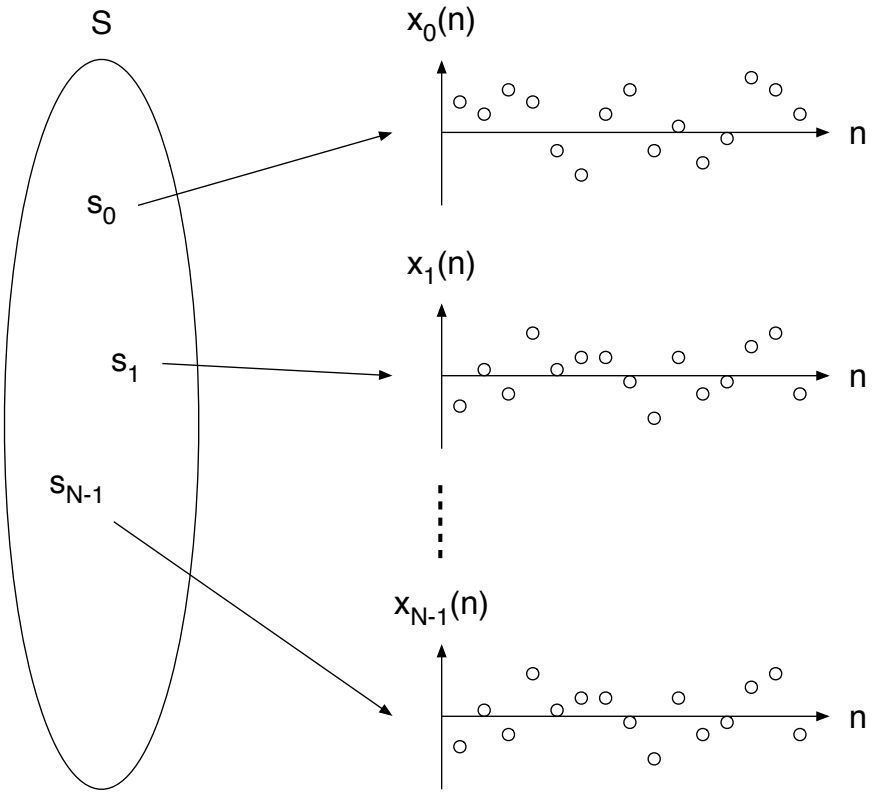
- A few places where the central limit theorem sort of sometimes comes up
  - Viewing the inner product in matrix vector or matrix matrix multiplication as a weighted sum of random variables
  - Viewing the DFT operation as a (rotated) sum of random variables
- Why this matters
  - Input can have ~ arbitrary distribution, maybe nicely bounded
  - But the output of the operation starts to look Gaussian
  - Gaussian random variables have long tails
  - With finite precision arithmetic this affects accuracy
- More on precision when CNN performance and implementation is discussed

Random Processes

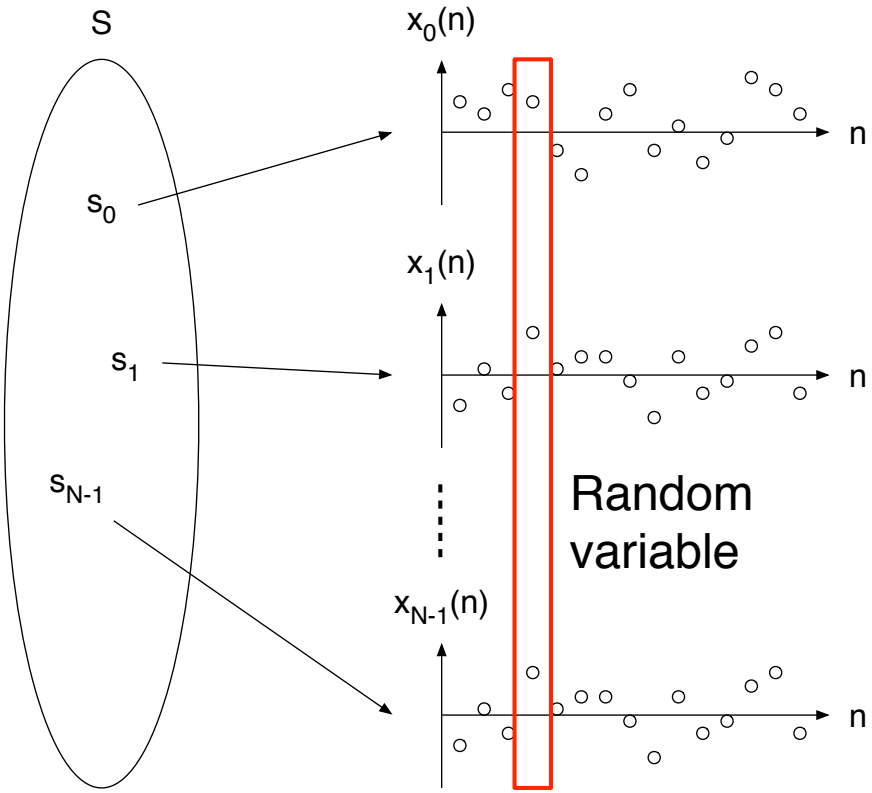
Definition

- A random process X(s, n) maps events s from the sample space S to functions x(n) where the domain of the function is the index set and the range of the function is the state space
  - X(s, n) is a random variable at a fixed n
    - By considering all times n this leads to the observation that a random process can be considered a collection of random variables $\{X(s, n_0), ..., X(s, n_{N-1})\}$
  - X(s, n) is a deterministic function of n for a fixed s
    - This is referred to as a realization of the random process
    - The set of all possible functions is referred to as the ensemble
  - X(s, n) is a number for a fixed s at a fixed n
- Names
  - If n refers to time then X(s, n) is called a random process
  - If n has multiple dimensions like width and height of an image then X(s, n) is called a random field

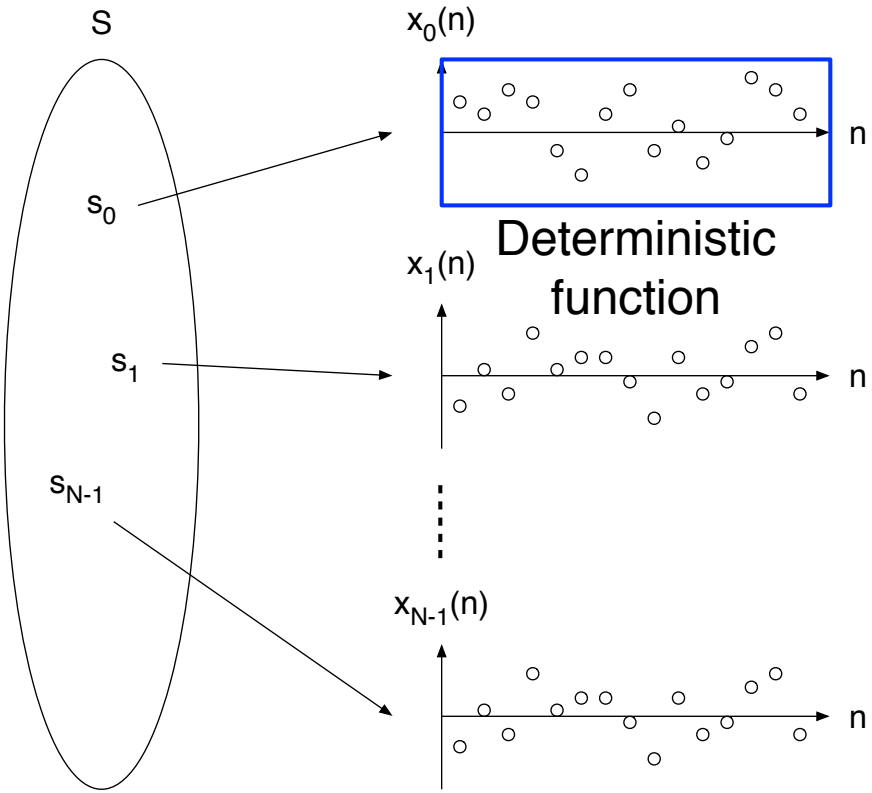
Definition



Definition



Definition



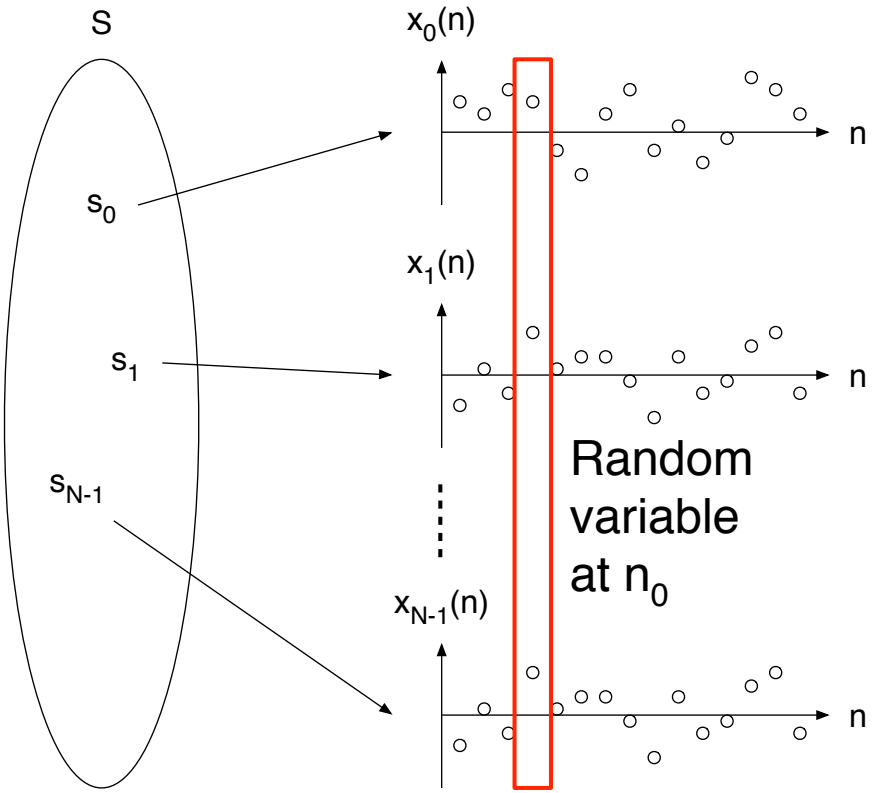
Stationarity

- Non stationary
  - Using the view of a random process as a collection of random variables, a random process is defined by is joint CDF $F_{X\_0,\ldots,X\_\{N-1\}}(x_{n\_0}, \ldots, x_{n\_(N-1)})$ which in general is a function of n_k
  - Informally, a non stationary random process has a CDF that changes with n (and doesn't fit neatly into 1 of the less restrictive stationary categorizations)
- (Strictly) stationary
  - Random processes X(s, n) for which the joint CDF does not change with time

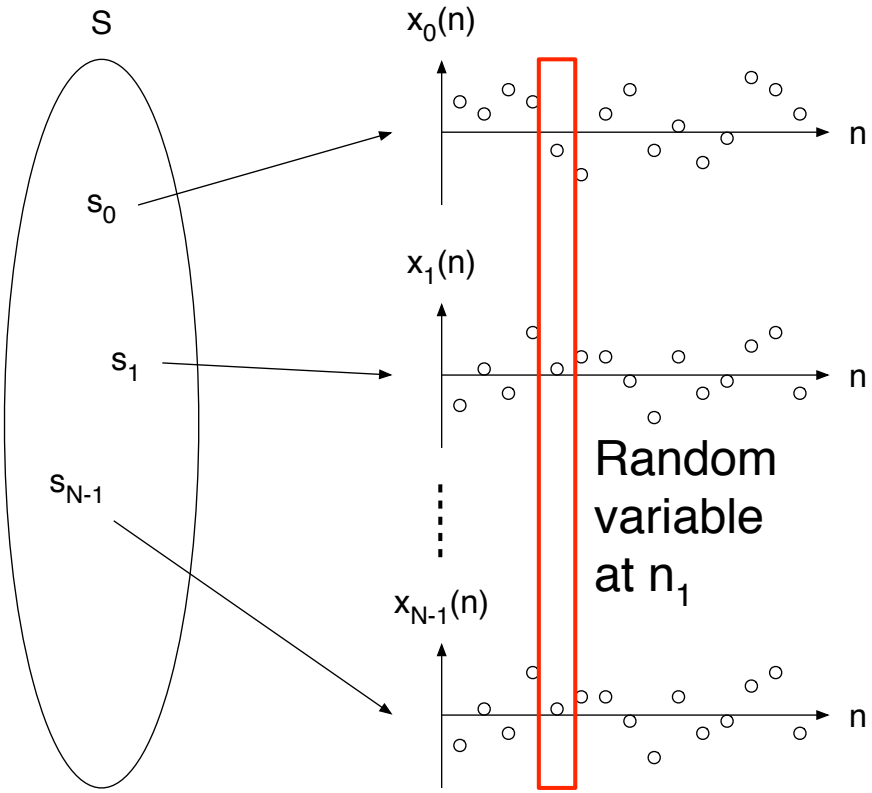
$$F_{X\_0,\ldots,X\_\{N-1\}}(x_{n\_0+\tau}, \ldots, x_{n\_(K-1)+\tau}) = F_{X\_0,\ldots,X\_\{N-1\}}(x_{n\_0}, \ldots, x_{n\_(K-1)}) \text{ for all K, n and } \tau$$

- Weakly (wide sense or second order) stationary
  - Random processes X(s, n) for which the mean and auto covariance do not change with time
  - Autocorrelation only depends on time difference $\tau = n_1 - n_2$
- Other types of stationarity exist (e.g., cyclostationary)

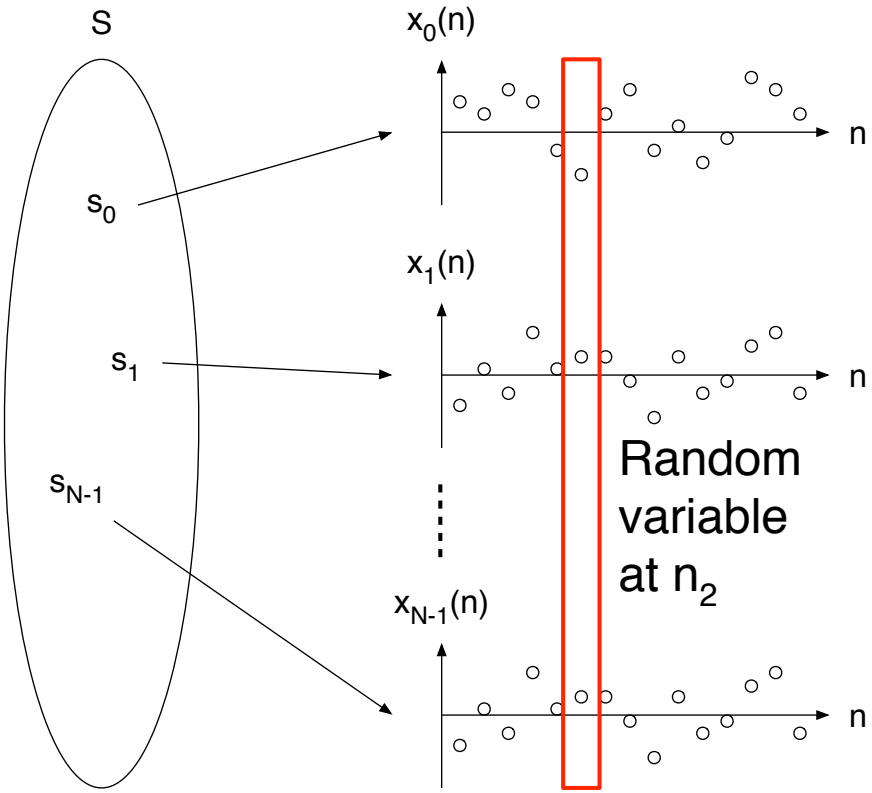
Stationarity



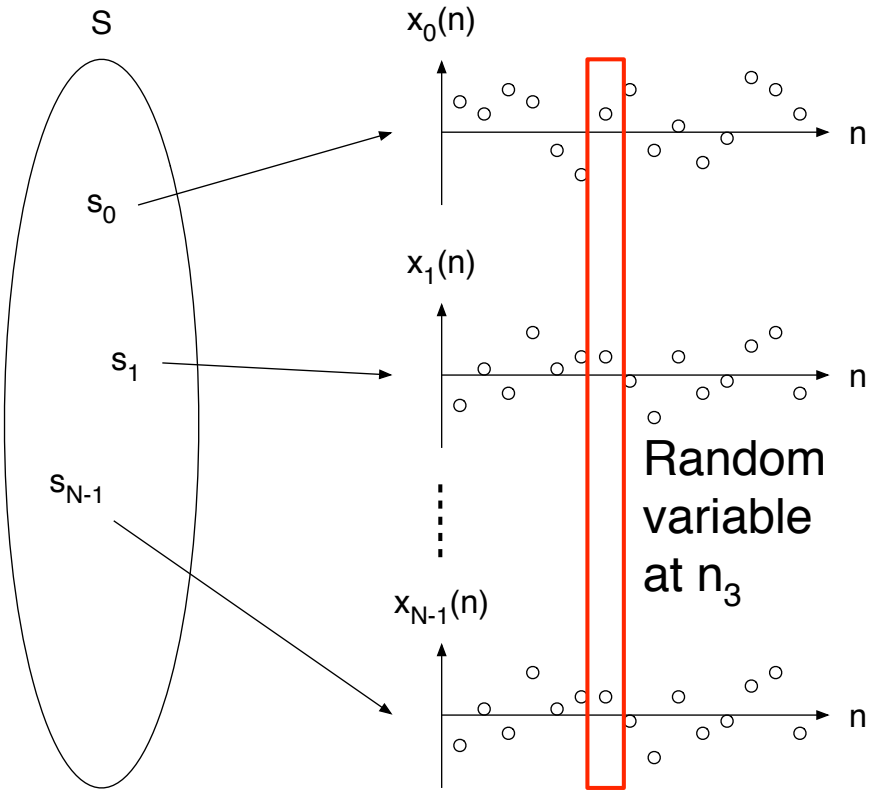
Stationarity



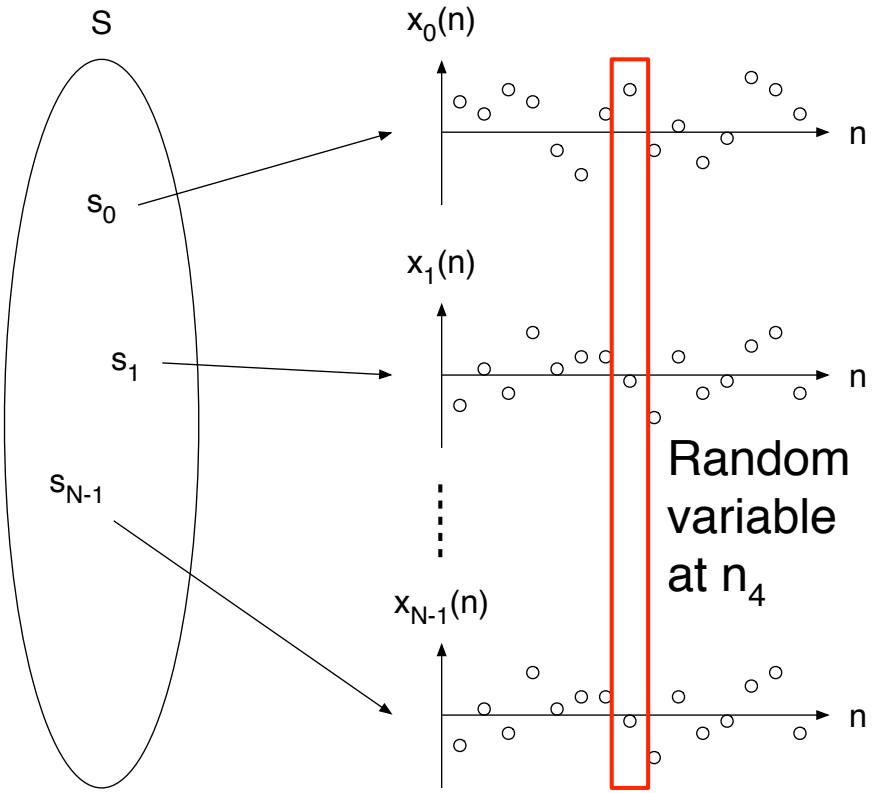
Stationarity



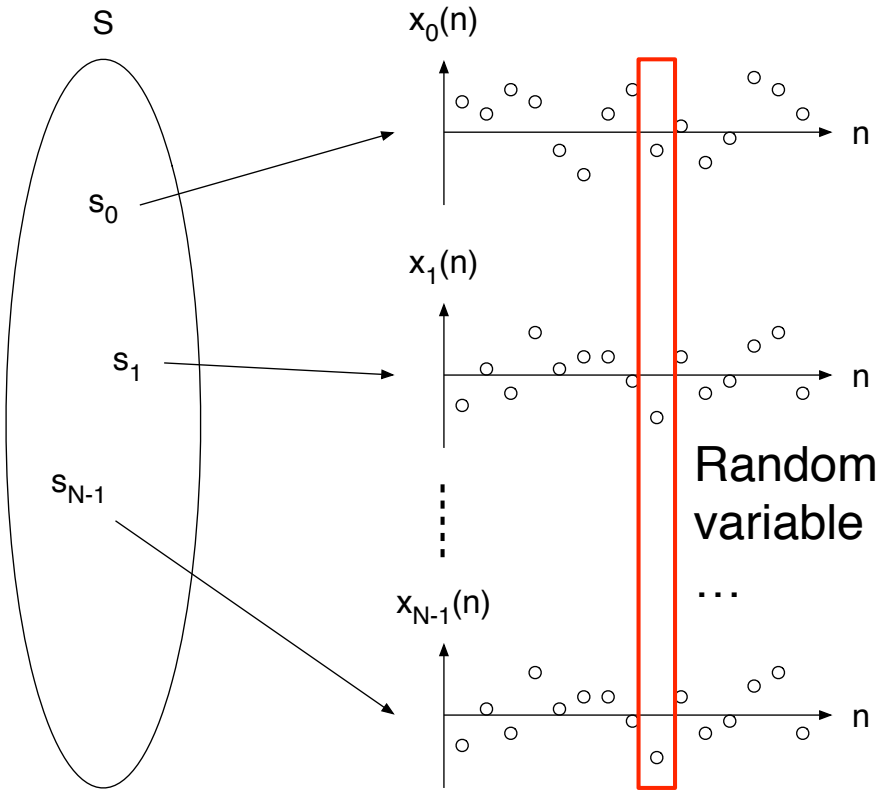
Stationarity



Stationarity



Stationarity



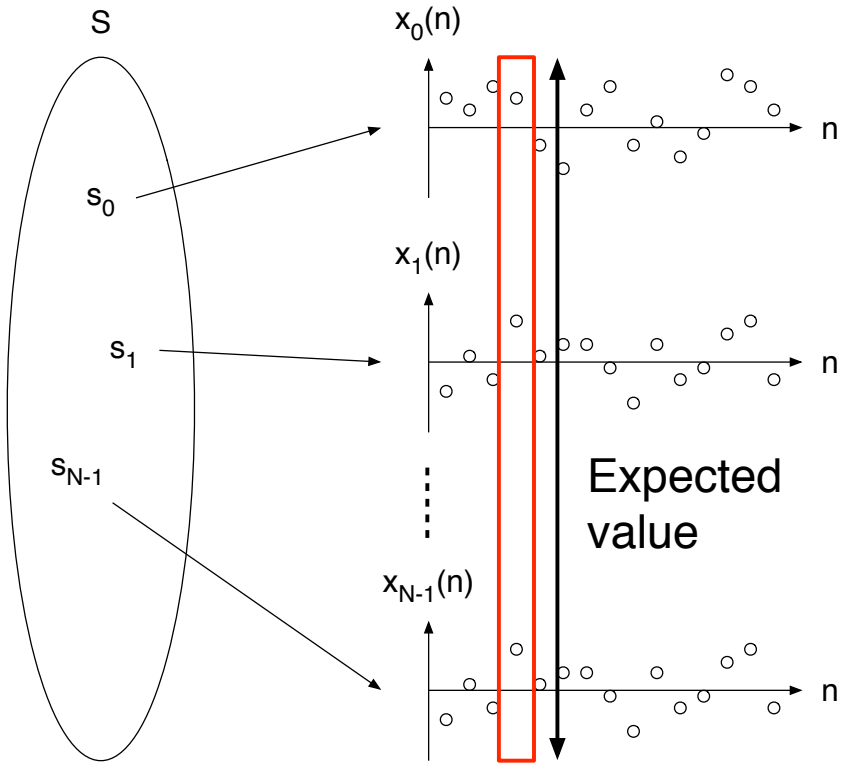
Expected Value

- The expected value of a random process is found by viewing the random process as a random variable at a fixed n and applying the expected value operator as before
  - Conceptually, it operates across many realizations s of a random process at a single n
  - Mean, variance and higher order moments are defined as in the case of a random variable
- Let $p_X(x_i, n) = P(X(s, n) = x_i)$ at a fixed n, then

$$E[f(X(s, n))] = \sum_i p_X(x_i, n) f(x_i)$$

- Which in general is a function of n

Expected Value

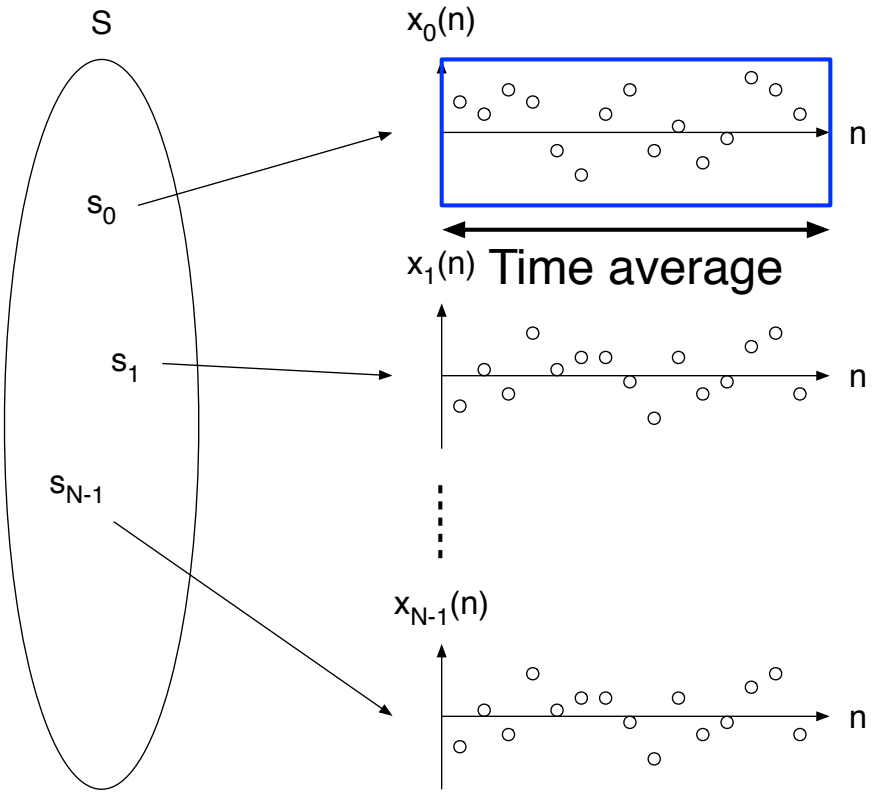


Time Average

- The time average of a random process is found by viewing the random process as a deterministic function for a fixed s and applying the time average operator
  - Conceptually, it operates across 1 realization s of a random process at many points n
  - Different time averages are defined similar to the expected value of a random variable
  - The time average itself is a random variable as it depends on the chosen s

$$\langle f(X(s, n)) \rangle = 1/N \sum_n f(X(s, n))$$

Time Average



Ergodicity

- Ergodicity: when time averages converge to expectations
  - In some sense (e.g., mean square)
  - For some orders of moments for which the process is stationary

- Example: mean ergodic
  - $\langle X(s, n) \rangle$ converges in the mean and in the mean square sense to $E[X(s, n)]$
    - $\lim_{N \to \infty} E[((1/N \sum_n f(X(s, n))) - \mu_X)] = 0$
    - $\lim_{N \to \infty} E[((1/N \sum_n f(X(s, n))) - \mu_X)^2] = 0$

References

List

- Random
  - http://www.randomservices.org/random/index.html
- StatLect
  - https://www.statlect.com
- Lecture notes on probability, statistics and linear algebra
  - http://www.math.harvard.edu/~knill/teaching/math19b_2011/handouts/chapters1-19.pdf
- A mathematical theory of communication
  - http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
- Joint entropy, conditional entropy, relative entropy, and mutual information
  - http://octavia.zoology.washington.edu/teaching/429/lecturenotes/lecture3.pdf
- Visual information theory
  - http://colah.github.io/posts/2015-09-Visual-Information/
- Computational optimal transport
  - https://arxiv.org/abs/1803.00567