Lecture 3
Math & Probability Background

ch. 1-2 of *Machine Vision* by Wesley E. Snyder & Hairong Qi

Spring 2024

16-725 (CMU RI) : BioE 2630 (Pitt)

Dr. John Galeotti

General notes about the book

- The book is an overview of many concepts
- Top quality design requires:
  - Reading the cited literature
  - Reading more literature
  - Experimentation & validation

Two themes

- Consistency
  - A conceptual tool implemented in many/most algorithms
  - Often must fuse information from many local measurements and prior knowledge to make global conclusions about the image
- Optimization
  - Mathematical mechanism
  - The “workhorse” of machine vision

Image Processing Topics

- Enhancement
- Coding
  - Compression
- Restoration
  - “Fix” an image
  - Requires model of image degradation
- Reconstruction

Machine Vision Topics

Original Image → Feature Extraction → Classification & Further Analysis

CNN: Convolutional Neural Network

FCN: Fully Connected (Neural) Network

…or "another" CNN, e.g. U-Net decoder section

- AKA:
  - Computer vision
  - Image analysis
  - Image understanding
- Pattern recognition:

Our Focus

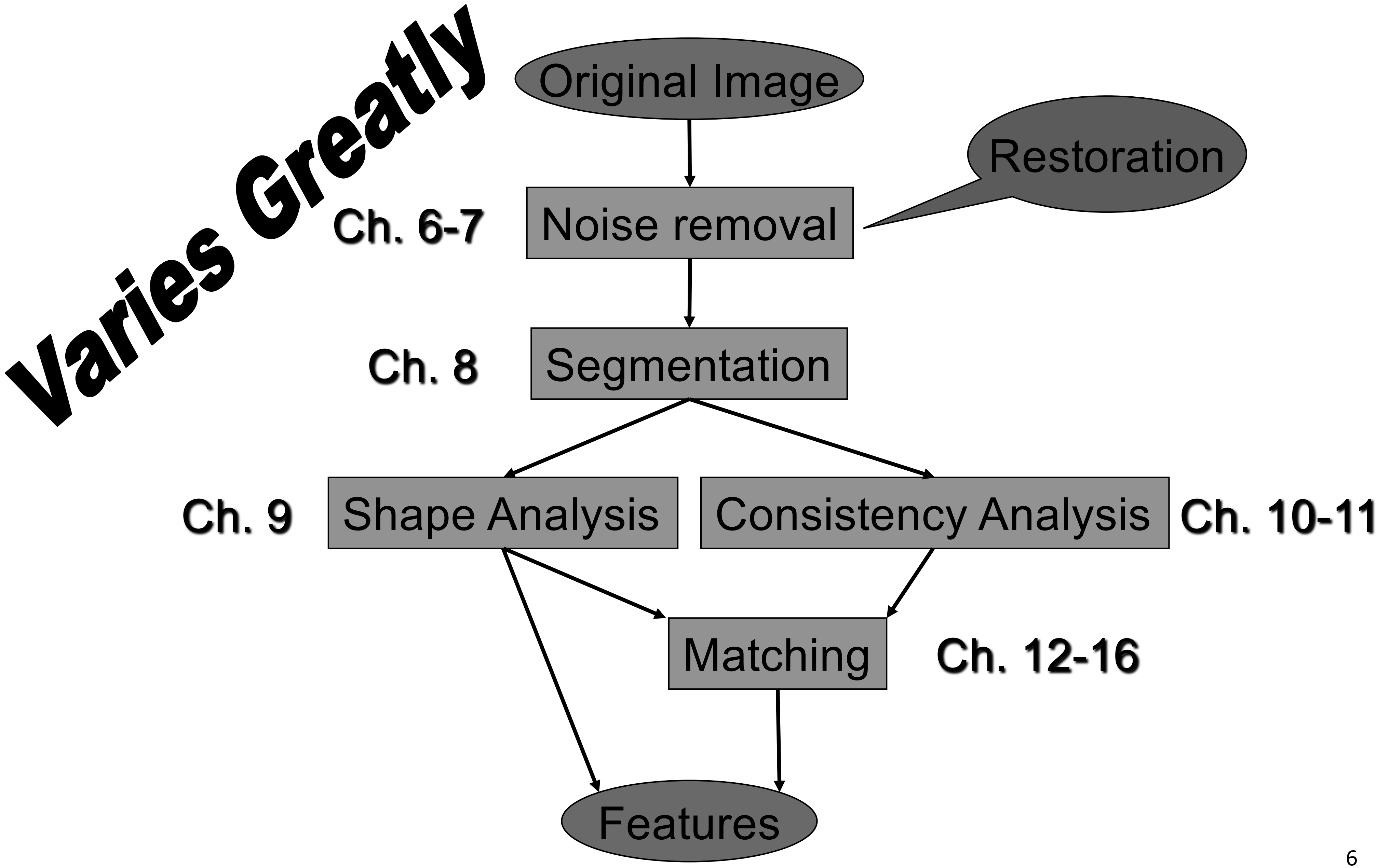
1. Measurement of features

   Features characterize the image, or some part of it

2. Pattern classification

   Requires knowledge about the possible classes

Feature measurement



Probability

- Probability of an event a occurring:
  - $Pr(a)$
- Independence
  - $Pr(a)$ does not depend on the outcome of event b, and vice-versa
- Joint probability
  - $Pr(a,b)$ = Prob. of both a and b occurring
- Conditional probability
  - $Pr(a|b)$ = Prob. of a if we already know the outcome of event b
  - Read "probability of a given b"

Probability for continuously-valued functions

- Probability distribution function:

$$P(x) = Pr(z<x)$$

- Probability density function:

$$p(x) = \frac{d}{dx} P(x)$$

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

Linear algebra

$$\mathbf{v} = \left[x_1 \ x_2 \ x_3 \right]^{\mathbf{T}} \qquad \mathbf{a}^{\mathbf{T}}\mathbf{b} = \sum_i a_i b_i \qquad \left|\mathbf{x}\right| = \sqrt{\mathbf{x}^{\mathbf{T}}\mathbf{x}}$$

- Unit vector: $|\boldsymbol{x}| = 1$
- Orthogonal vectors: $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y} = 0$
- Orthonormal: orthogonal unit vectors
- Inner product of continuous functions

$$\left\langle f(x), g(x) \right\rangle = \int_a^b f(x) g(x)\, dx$$

  - Orthogonality & orthonormality apply here too

Linear independence

- No one vector is a linear combination of the others
  - $\boldsymbol{x}_j \neq \Sigma a_i \boldsymbol{x}_i$ for any a_i across all $i \neq j$
- Any linearly independent set of d vectors $\{\boldsymbol{x}_{i=1\ldots d}\}$ is a basis set that spans the space $\Re^d$
  - Any other vector in $\Re^\mathrm{d}$ may be written as a linear combination of $\{\boldsymbol{x}_i\}$
- Often convenient to use orthonormal basis sets
- Projection: if $\boldsymbol{y}=\Sigma a_\mathrm{i} \boldsymbol{x}_i$ then $a_i=\boldsymbol{y}^\mathrm{T}\boldsymbol{x}_i$

Linear transforms

- = a matrix, denoted e.g. A
- Quadratic form:

$$\mathbf{x}^{\mathrm{T}} A \mathbf{x}$$

$$\frac{d}{d\mathbf{x}}\left(\mathbf{x}^{\mathrm{T}} A \mathbf{x}\right) = \left(A + A^{\mathrm{T}}\right)\mathbf{x}$$

- Positive definite:
  - Applies to A if

$$\mathbf{x}^{\mathrm{T}} A \mathbf{x} > 0 \;\; \forall \mathbf{x} \in \Re^{d}, \mathbf{x} \neq 0$$

More derivatives

- Of a scalar function of $\boldsymbol{x}$:
  - Called the gradient
  - Really important!

$$\frac{df}{d\mathbf{x}} = \left[\frac{\partial f}{\partial x_1} \ \frac{\partial f}{\partial x_2} \cdots \frac{\partial f}{\partial x_d} \right]^{\mathbf{T}}$$

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_d} \end{bmatrix}$$

- Of a vector function of $\boldsymbol{x}$
  - Called the Jacobian

$$\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

- Hessian = matrix of 2nd derivatives of a scalar function

Misc. linear algebra

- Derivative operators
- Eigenvalues & eigenvectors
  - Translates “most important vectors”
    - Of a linear transform (e.g., the matrix $\boldsymbol{A}$)
  - Characteristic equation: $A\mathbf{x} = \lambda\mathbf{x} \quad \lambda \in \Re$
  - A maps $\mathbf{x}$ onto itself with only a change in length
  - λ is an eigenvalue
  - $\mathbf{x}$ is its corresponding eigenvector

Function minimization

- Find the vector $\mathbf{x}$ which produces a minimum of some function $f(\mathbf{x})$
  - $\mathbf{x}$ is a parameter vector
  - $f(\mathbf{x})$ is a scalar function of $\mathbf{x}$
    - The "objective function"
- The minimum value of f is denoted:

$$\widehat{f}(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x})$$

- The minimizing value of $\mathbf{x}$ is denoted:

$$\hat{\mathbf{x}} = \mathrm{argmin}_{\mathbf{x}} f(\mathbf{x})$$

Numerical minimization

- Gradient descent
  - The derivative points away from the minimum
  - Take small steps, each one in the "down-hill" direction
- Local vs. global minima
- Combinatorial optimization:
  - Use simulated annealing
- Image optimization:
  - Use mean field annealing
- More recent improvements to gradient descent:
  - Momentum, changing step size
- Training CNN: <u>ADAM</u>: an enhanced version of Stochastic Gradient Descent (SGD) w/ Momentum

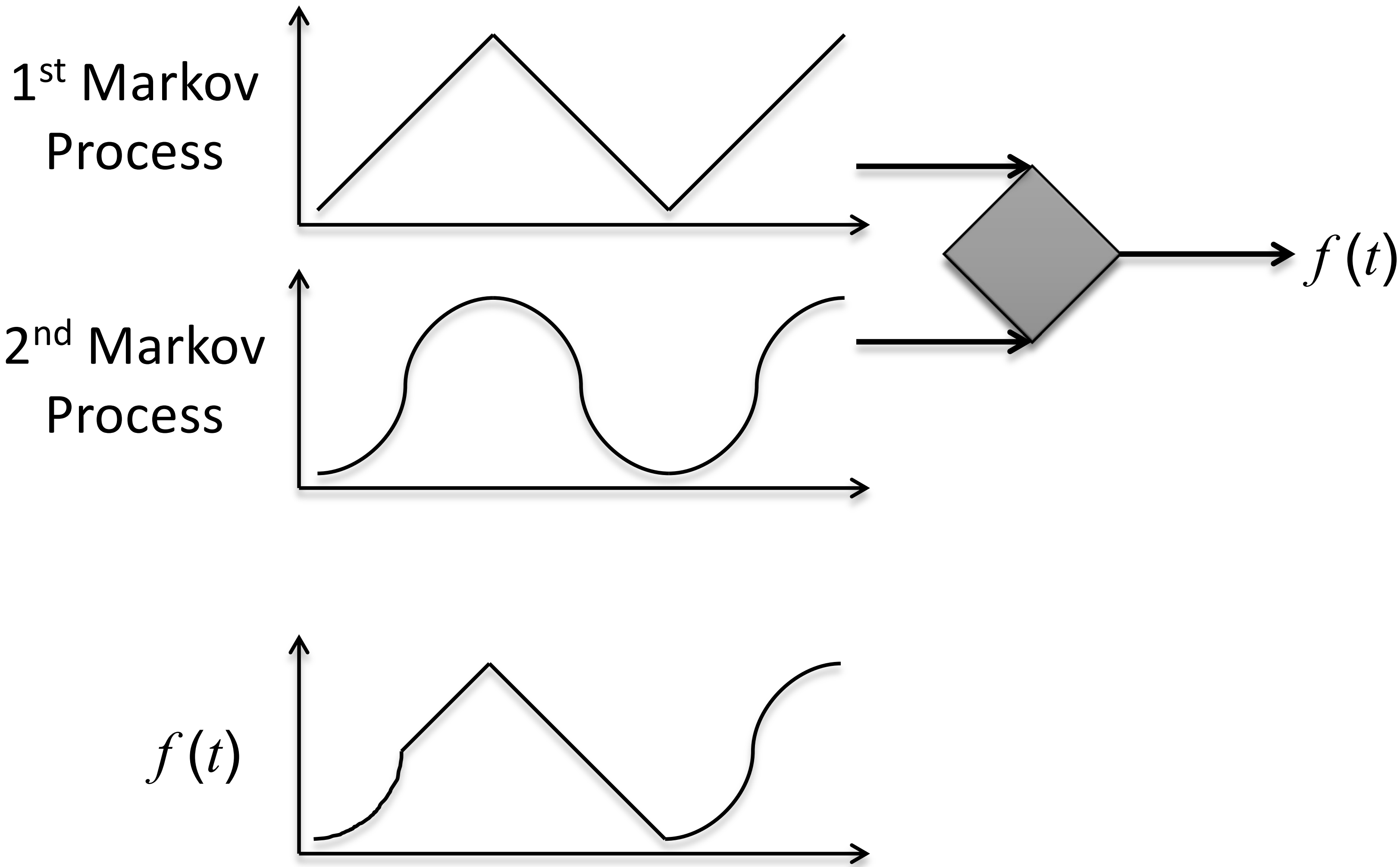
Markov models

- For temporal processes:
  - The probability of something happening is dependent on a thing that just recently happened.
- For spatial processes
  - The probability of something being in a certain state is dependent on the state of something nearby.
  - Example: The value of a pixel is dependent on the values of its neighboring pixels.

Markov chain

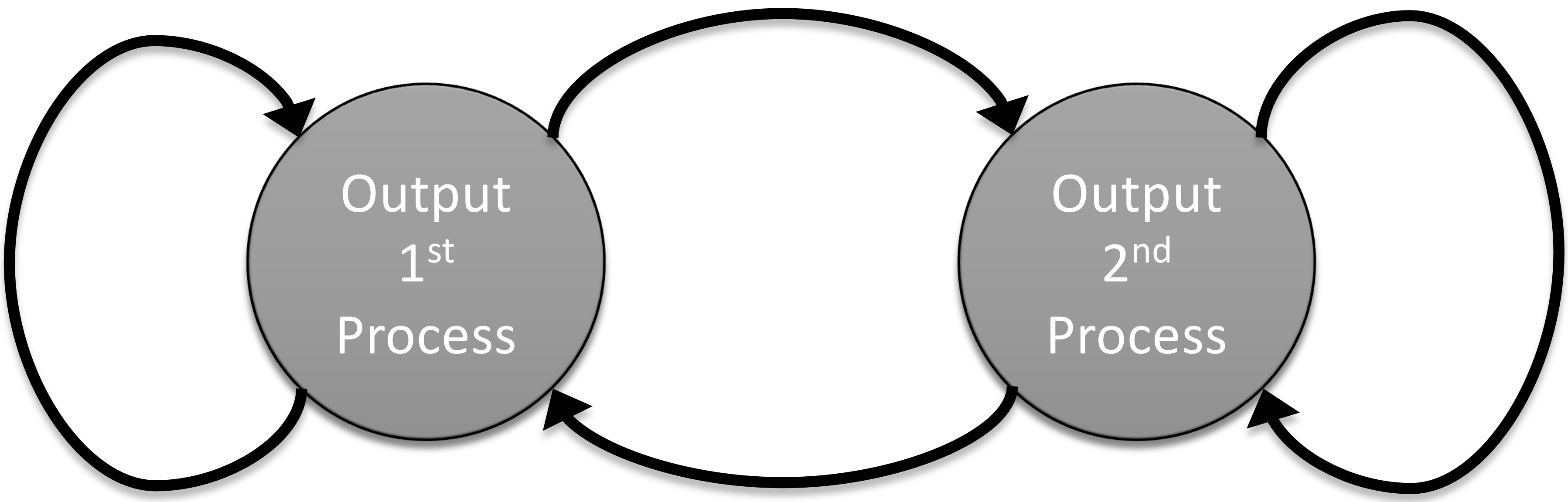
- Simplest Markov model
- Example: symbols transmitted one at a time
  - What is the probability that the next symbol will be w?
- For a “simple” (i.e. first order) Markov chain:
  - “The probability conditioned on all of history is identical to the probability conditioned on the last symbol received.”

Hidden Markov models (HMMs)



HMM switching

- Governed by a finite state machine (FSM)



The HMM Task

- Given only the output $f(t)$, determine:
  1. The most likely state sequence of the switching FSM
     - Use the Viterbi algorithm (much better than brute force)
     - Computational Complexity of:
       - Viterbi: (# state values)^2 * (# state changes)
       - Brute force: (# state values)$^{(\text{\# state changes})}$
  2. The parameters of each hidden Markov model
     - Use the iterative process in the book
     - Better, use someone else's debugged code that they've shared