Random Variables & Probability Review

CS 534: Machine Learning

Slides adapted from Lee Cooper, Joydeep Ghosh

Machine Learning & Probability

- Probability: a mathematical framework for uncertainty
- Machine learning problems fit well into this framework
  - How uncertain is our prediction?
  - Modeling structure with uncertainty
  - Noise

Probability Theory

Set Theory

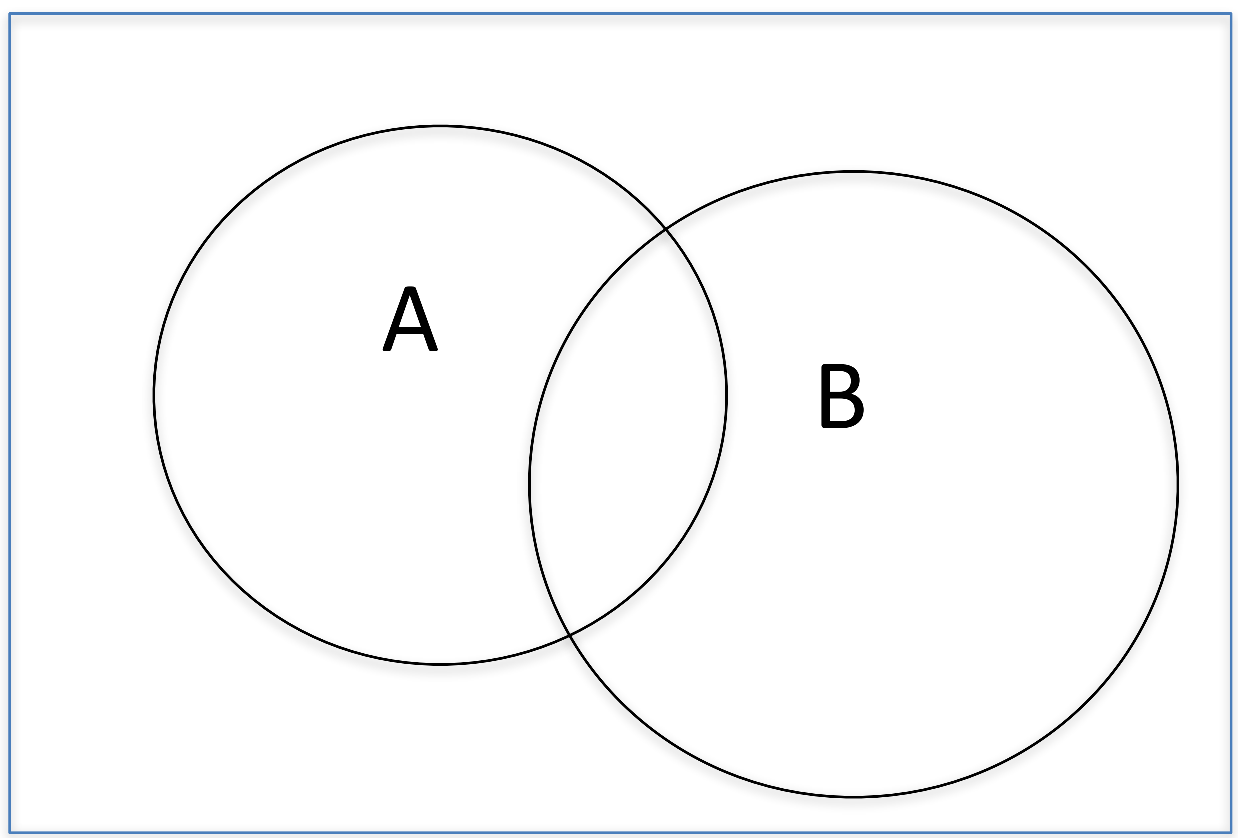
- Consider an experiment with an uncertain outcome:
  - A single outcome of this experiment is called an *event*
  - Collection of all possible outcomes is called a *space*
- *Probability* broadly describes the likelihood of events
- Set theory is used to reason about events and spaces, and to develop the fundamentals of probability theory

Set Theory: Example

- Experiment: flip a 2-sided coin in the air 2 times and record which side is facing up (heads (H) or tails (T))
- 4 possible events: heads two times (HH), tails two times (TT), heads then tails (HT), tails then heads (TH)
- Space is the set { HH, HT, TH, TT }

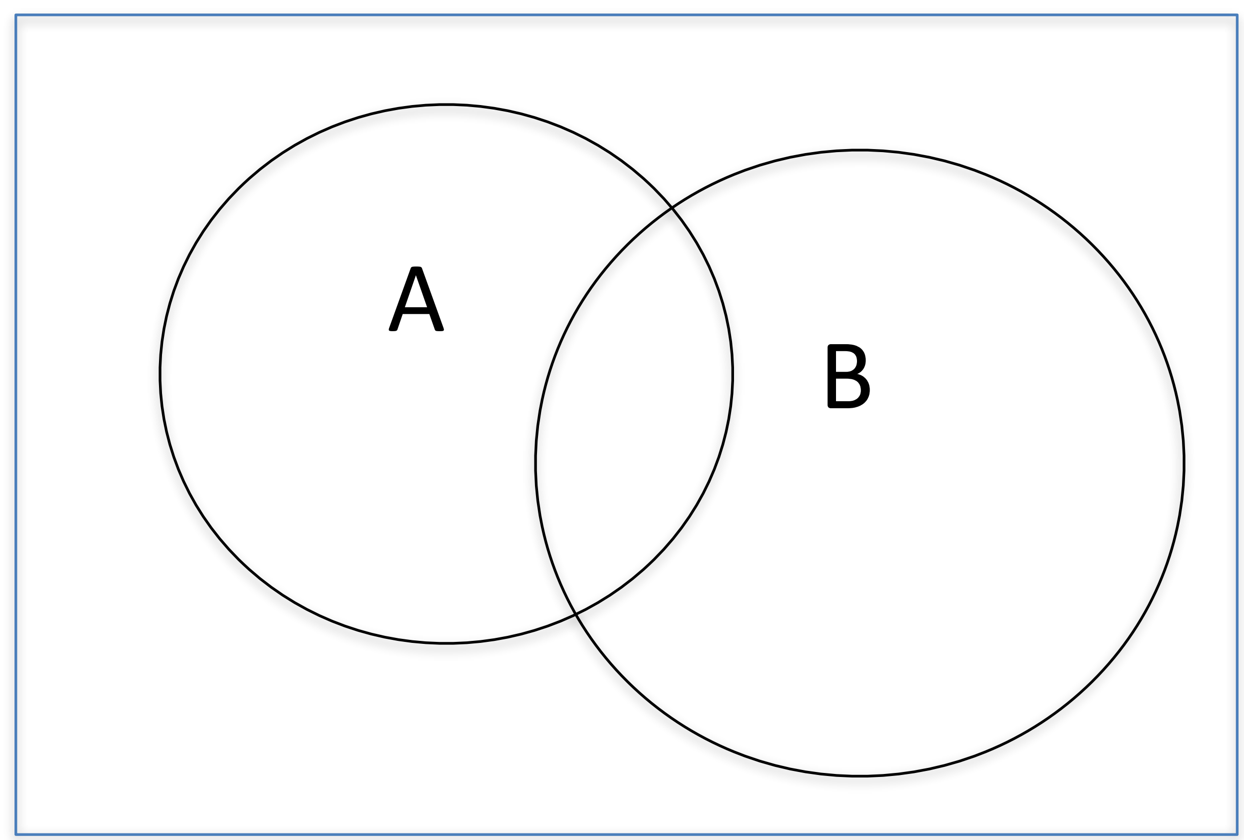
Set Theory: Visualization

- Convenient to visualize a set as a plane, and reason about the overlap or exclusivity of regions



Set Theory: Operations and Subsets

- Union, Intersection, Complement
- Subsets and proper subsets
- Empty/null set



Elements of Probability

- Sample space Ω : set of all outcomes of a random experiment
- Event space (set of events) $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ are subsets of sample space
- Probability measure: A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties
  - $P(A) \geq 0, \text{for all } A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - $\text{If } A_i \cap A_j = \emptyset \text{ when } i \neq j, \text{then } P(\cup_i A_i) = \sum_i P(A_i)$

Kolmogorov's Axioms
(Axioms of Probability)

Random Variables

- A random variable (RV) is a function that maps the space of events to numeric values $X : \Omega \to \mathbb{R}$
  - Simply put, assign a number to every outcome in Ω
  - Example: weight of a newborn child
- Represent quantities with some built-in uncertainty
- Textbook uses italicized capital letters to denote random variables

Random Variable Types

- Discrete random variable: X can take only a finite number of values
  - Example: Number of heads in a sequence of tosses
- Continuous random variable: X takes infinite number of possible values
  - Example: Amount of time for a radioactive particle to decay

Cumulative Distribution Function (CDF)

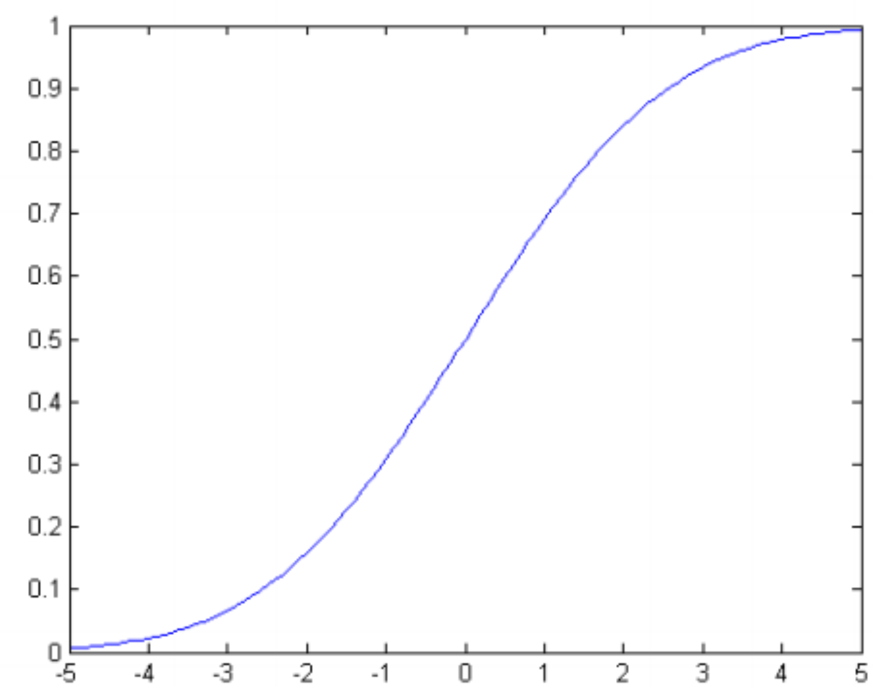
- CDF: function $F_X : \mathbb{R} \rightarrow [0, 1]$ that specifies a probability measure

$$F_X(x) \triangleq P(X \leq x)$$

- Used to calculate the probability of an event in $\mathcal{F}$

- Properties:

  - $0 \leq F_X(x) \leq 1$
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to \infty} F_X(x) = 1$
  - $x \leq y \implies F_X(x) \leq F_X(y)$

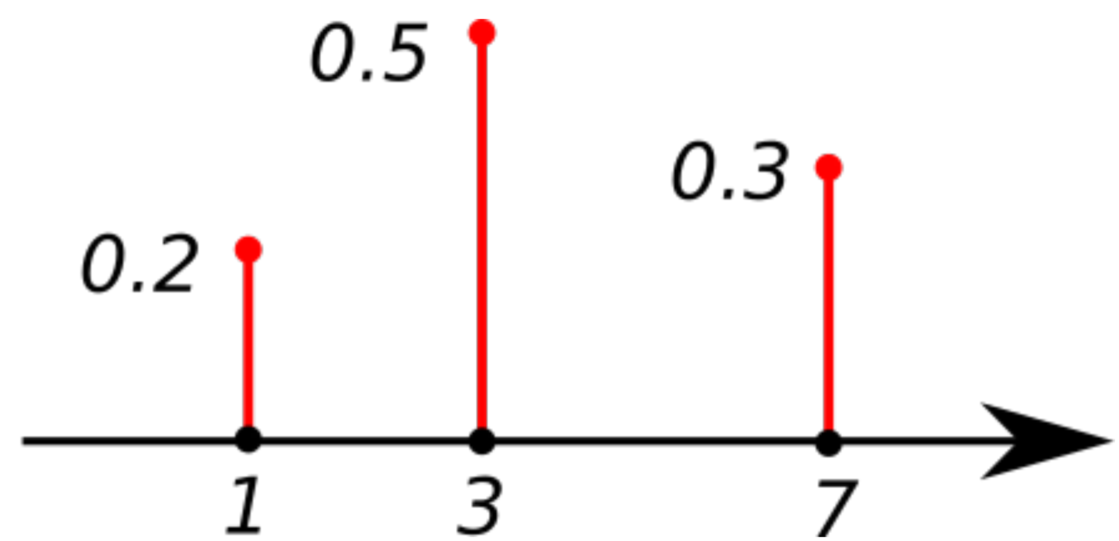


Probability Mass Function (PMF)

- Probability measure for discrete random variable
- PMF: function $p_X(x) : \Omega \to \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x)$$

- Properties:
  - $0 \leq p_X(x) \leq 1$
  - $\sum_{x \in \mathrm{Val}(X)} P_X(x) = 1$
  - $\sum_{x \in A} P_X(x) = P(X \in A)$



https://en.wikipedia.org/wiki/Probability_mass_function

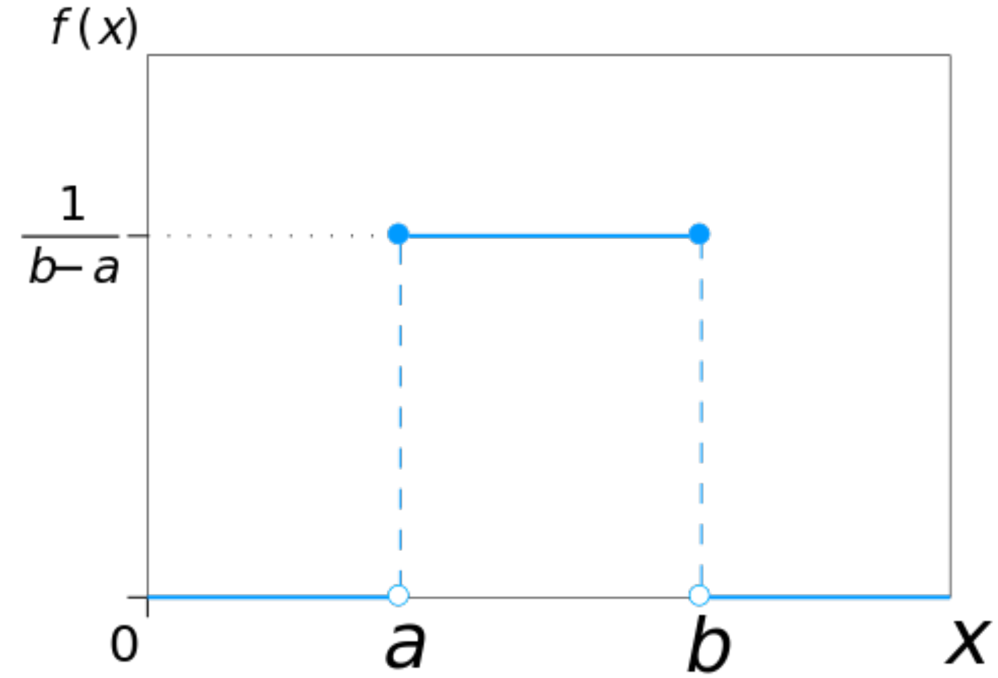
Probability Density Function (PDF)

- Probability measure for continuous random variable
- PDF is derivative of CDF

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

may not always exist if CDF is not differentiable

- Properties:
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $\int_{x \in A} f_X(x)dx = P(X \in A)$



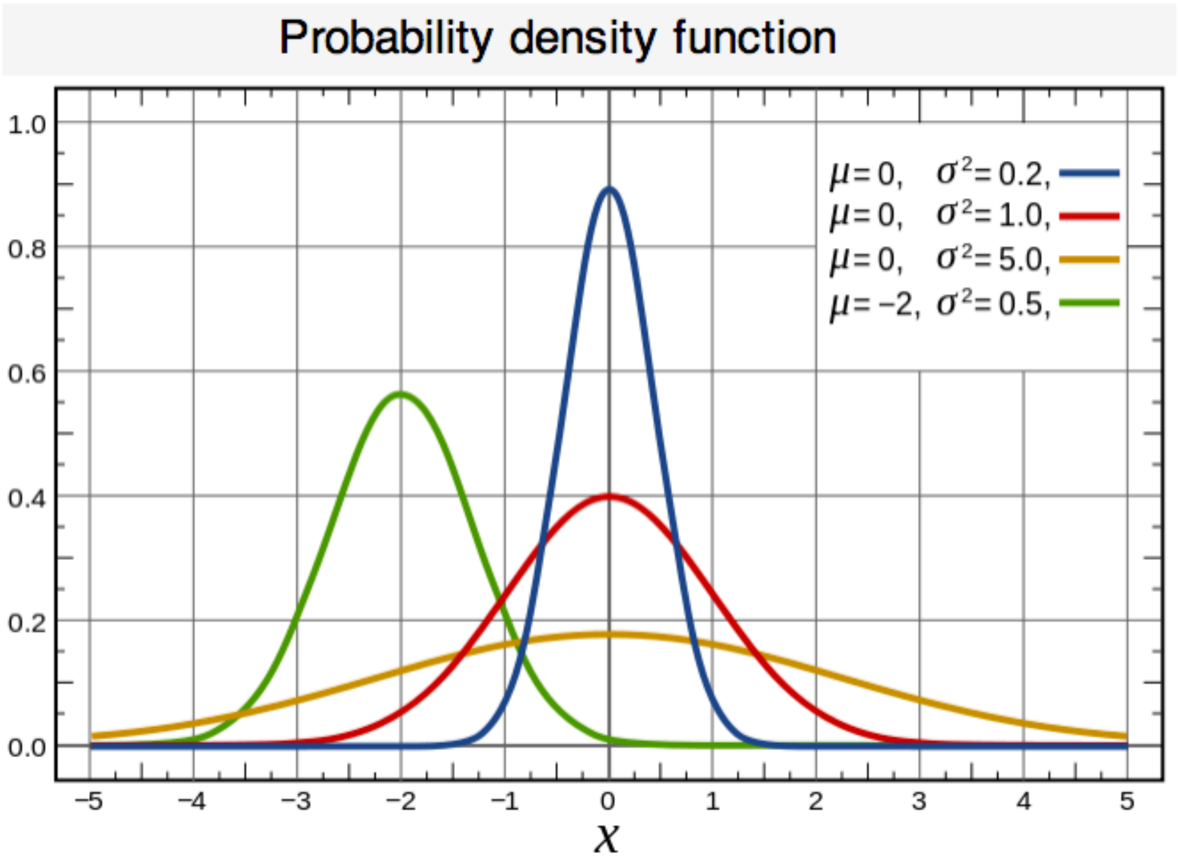
https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)

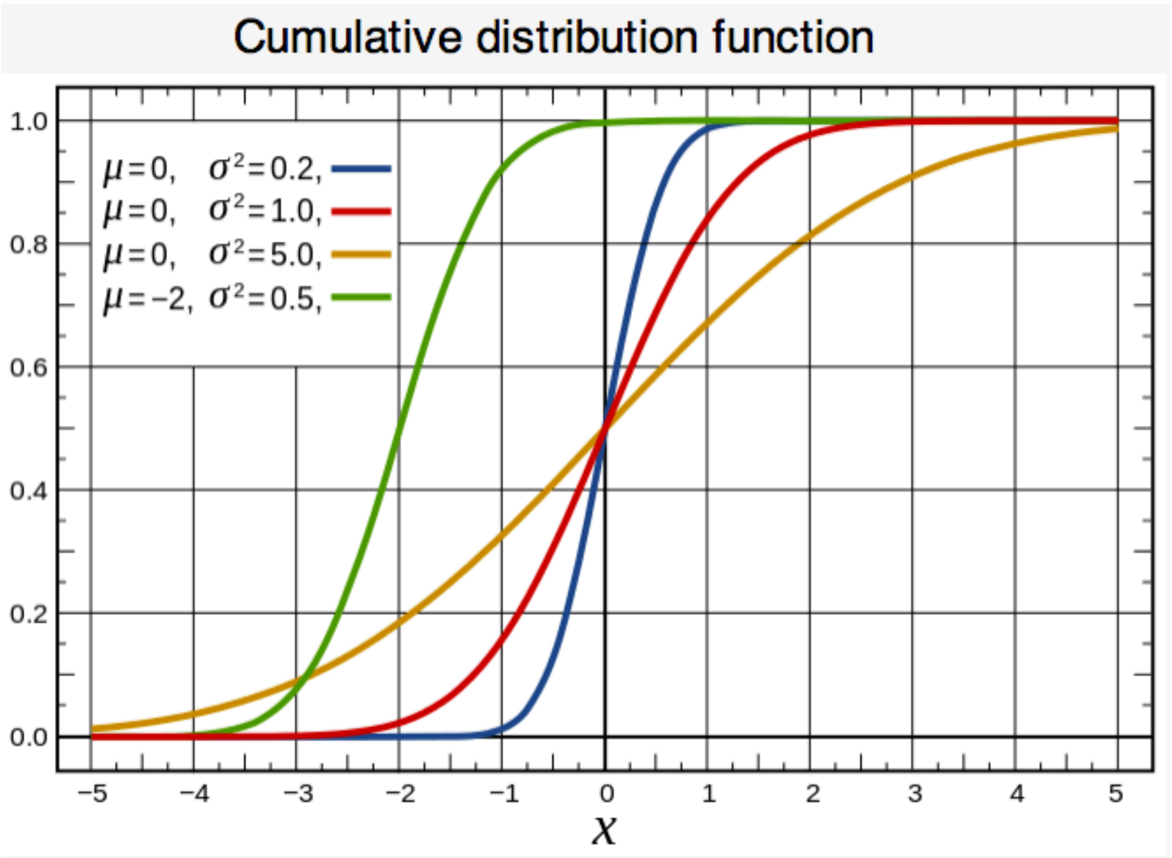
Example: Normal Distribution

mean

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

variance





https://en.wikipedia.org/wiki/Normal_distribution

PDFs vs. PMFs

	PDF	PMF
Values	Continuous valued RVs	Discrete-valued RVs
Representation	Function f(x)	Table
Probability	Calculated via integration	Calculated via summation
P(x = k)	0	Non-zero

Expectation: Mean and Variance

Expectation

- What is the expected value of a random variable?
- Expectation of g(X):

$$E[g(X)] \triangleq \sum_{x \in \mathrm{Val}(X)} g(x) p_X(X)$$

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(X)$$

- "Weighted average" of values that g(x) with weights given by pdf or pmf

Expectation: Properties

- Constant

$$E[a] = a, \;\; a \in \mathbb{R}$$

- Scalar

$$E[af(X)] = aE[f(X)], \;\; a \in \mathbb{R}$$

- Linearity

$$E[f(X) + g(X)] = E[f(X)] + E[g(X)]$$

Expectation: Common Forms

- Mean: expectation of random variable

$$E[X], \text{ where } g(x) = x$$

- Variance: measure of how concentrated the distribution of the random variable is around its mean

$$\mathrm{Var}[X] \triangleq E[(X - E[X])^2]$$

Common Distributions

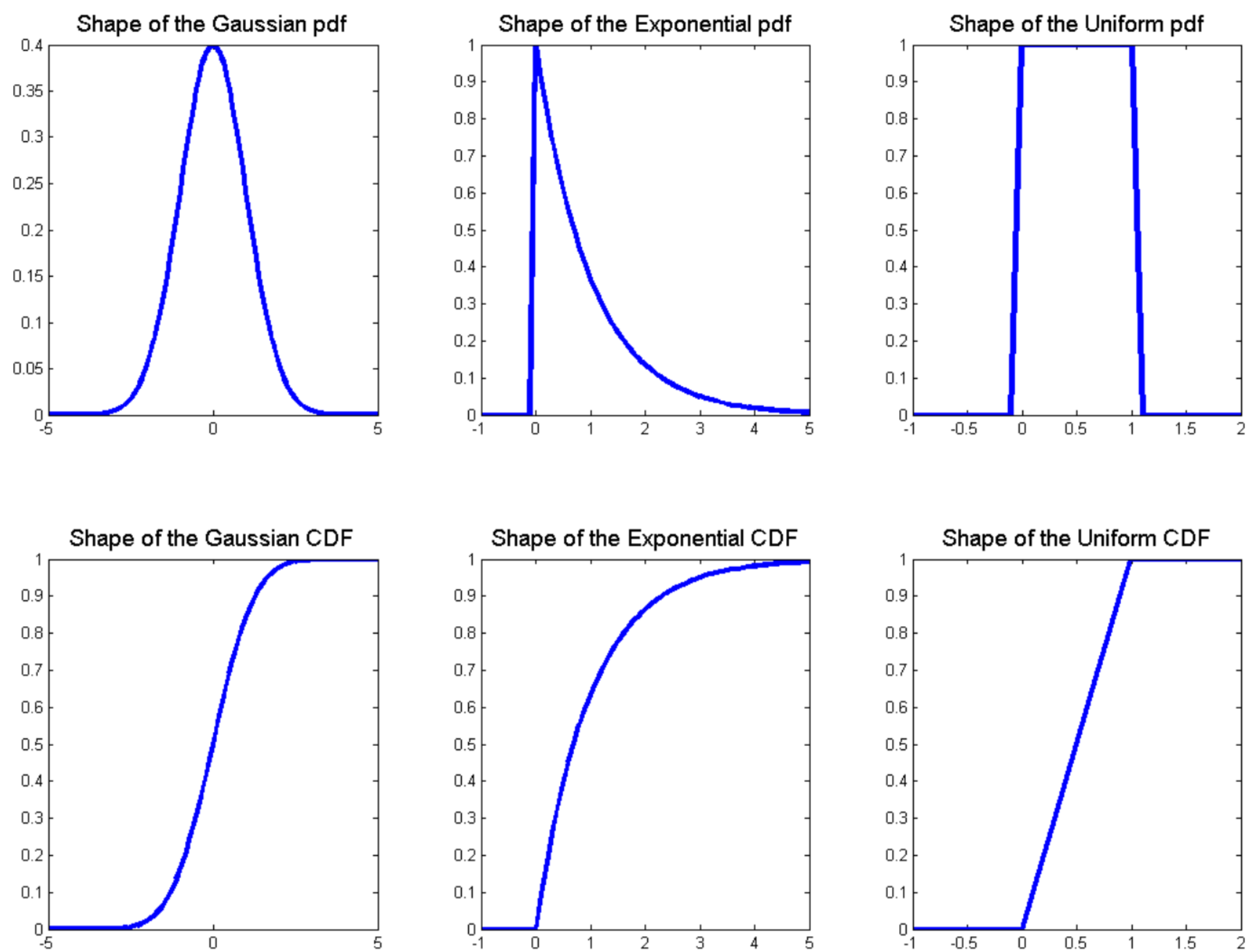
Discrete RV Distributions

- Bernoulli(p): coin flip with probability p of getting a heads ($p = 1$)
- Binomial(n,p): number of heads in n independent flips of a coin with probability p of a heads
- Geometric(p): number of flips of a coin until the first heads
- Poisson(λ): frequency of events or counts

Continuous RV Distributions

- Uniform(a, b): equal probability density between every value a and b on the real line
- Exponential(λ): decaying probability density over the nonnegative real numbers
- Normal(μ, σ^2): Gaussian distribution
  - Will be dealing with this 99% of the time
  - Interesting properties

Continuous RVs: PDF & CDF



http://cs229.stanford.edu/section/cs229-prob.pdf

Common RV Summary

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \le k \le n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a} \quad \forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x} \quad x \ge 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

http://cs229.stanford.edu/section/cs229-prob.pdf

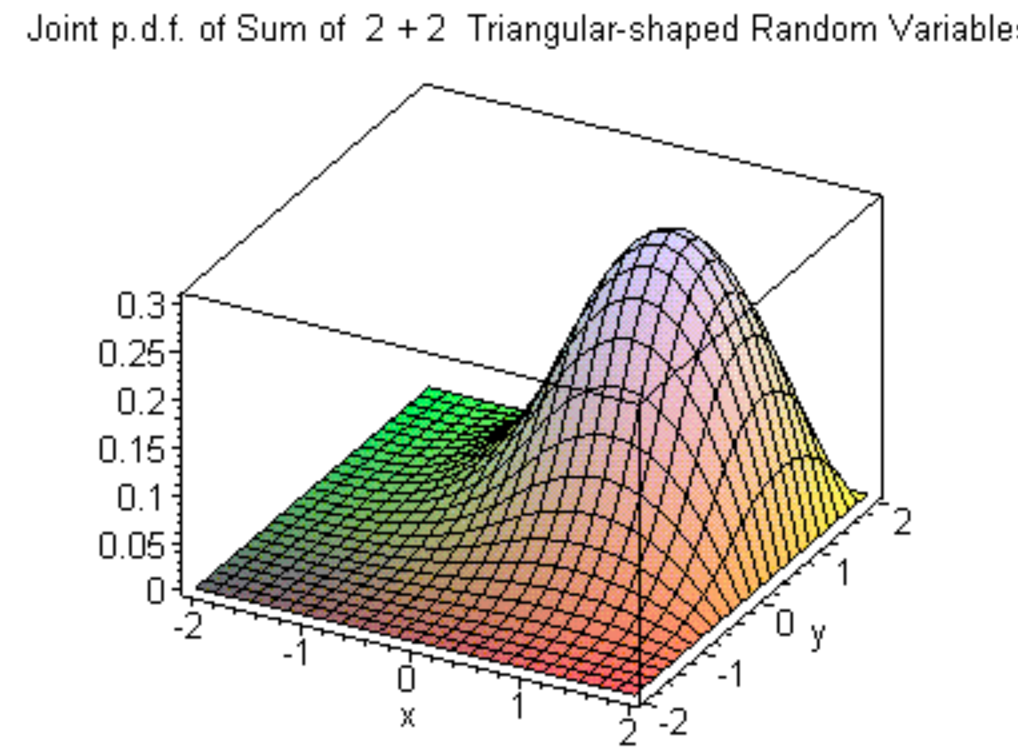
Multiple Random Variables

Multiple RVs & Machine Learning

- Most machine learning problems contain multiple random variables
- Values are not always independent
  - Example: Height (x) and weight (y) of newborn
- E[XY] = E[X] E[Y]?

Multiple RVs: Joint PDF

- The joint pdf of a collection of random variables completely captures their individual and collective properties (dependencies):



$$\Pr(X_1, \cdots, X_N \in D) = \int_D f_{X_1, \cdots, X_N}(x_1, \cdots, x_n) dx_1 \cdots dx_N$$

http://www.math.hope.edu/tanis/maa99/triang.html

Marginal Distributions

- Given a joint distribution, what is the distribution over each variable separately?
- "Integrate" out the other RVs that are not of interest
- Marginal PDF

$$f_{X_i}(x_i) = \int \cdots \int f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

Conditional Distributions

- What if the value of a variable in a joint density is known?
  - E.g. if weight known, how does distribution of height change?
- Conditional PDF:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

Bayes Rule

- Bayes Rule:

$$f_X(x|Y=y) = \frac{f_Y(y|X=x)f_X(x)}{f_Y(y)}$$

- Why is this helpful? Estimation!
- Say you observe Y and want to guess X

$$Y = aX + e$$

Independence

- When are two values unrelated whatsoever?
- Independence if and only if:

$$f_{X_1,\cdots,X_N}(x_1,\cdots,x_N) = f_{X_1}(x_1)\cdots f_{X_N}(x_N)$$

- Corallary: If you can factor a PDF of N RVs as a product of N one-variable terms, then these RVs are independent

Covariance

- Covariance: relationship between two random variables

$$\text{Cov}[X, Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

- Covariance matrix describes pairwise covariance between RVs

$$\Sigma_{ij} = \text{Cov}[X_i, X_j]$$

Covariance Properties

- Positive semidefinite

$$\Sigma \geq 0$$

- Symmetric

$$\Sigma = \Sigma^\top$$

- Diagonalize

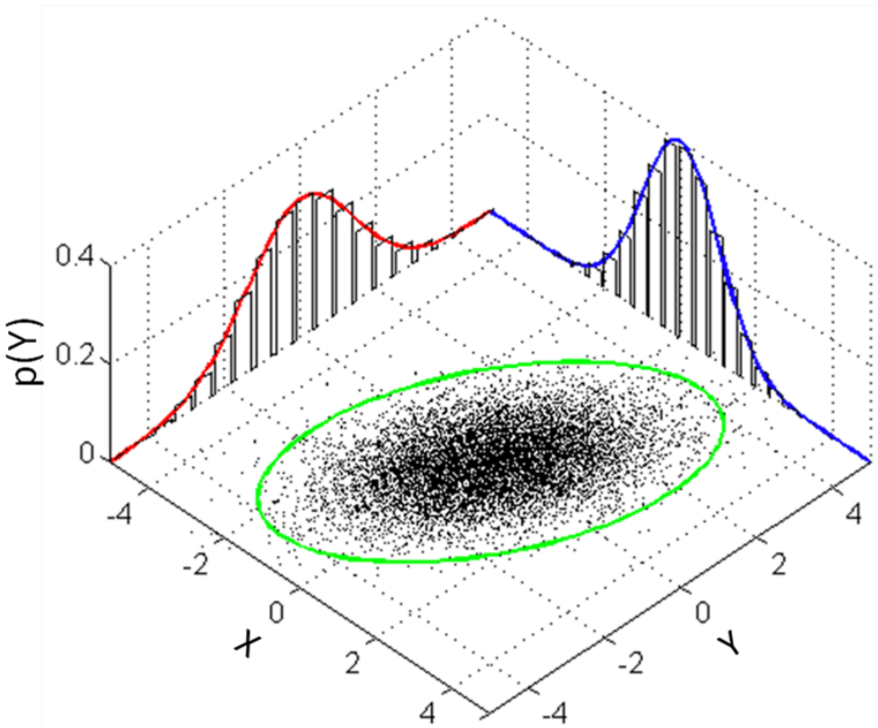
$$\Sigma = V \Lambda V^{-1} = V \Lambda V^\top$$

Eigenvalues and eigenvectors — “natural” system for data

Example: 2D Normal Distributions

- Normal (Gaussian) distribution

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



- PDF

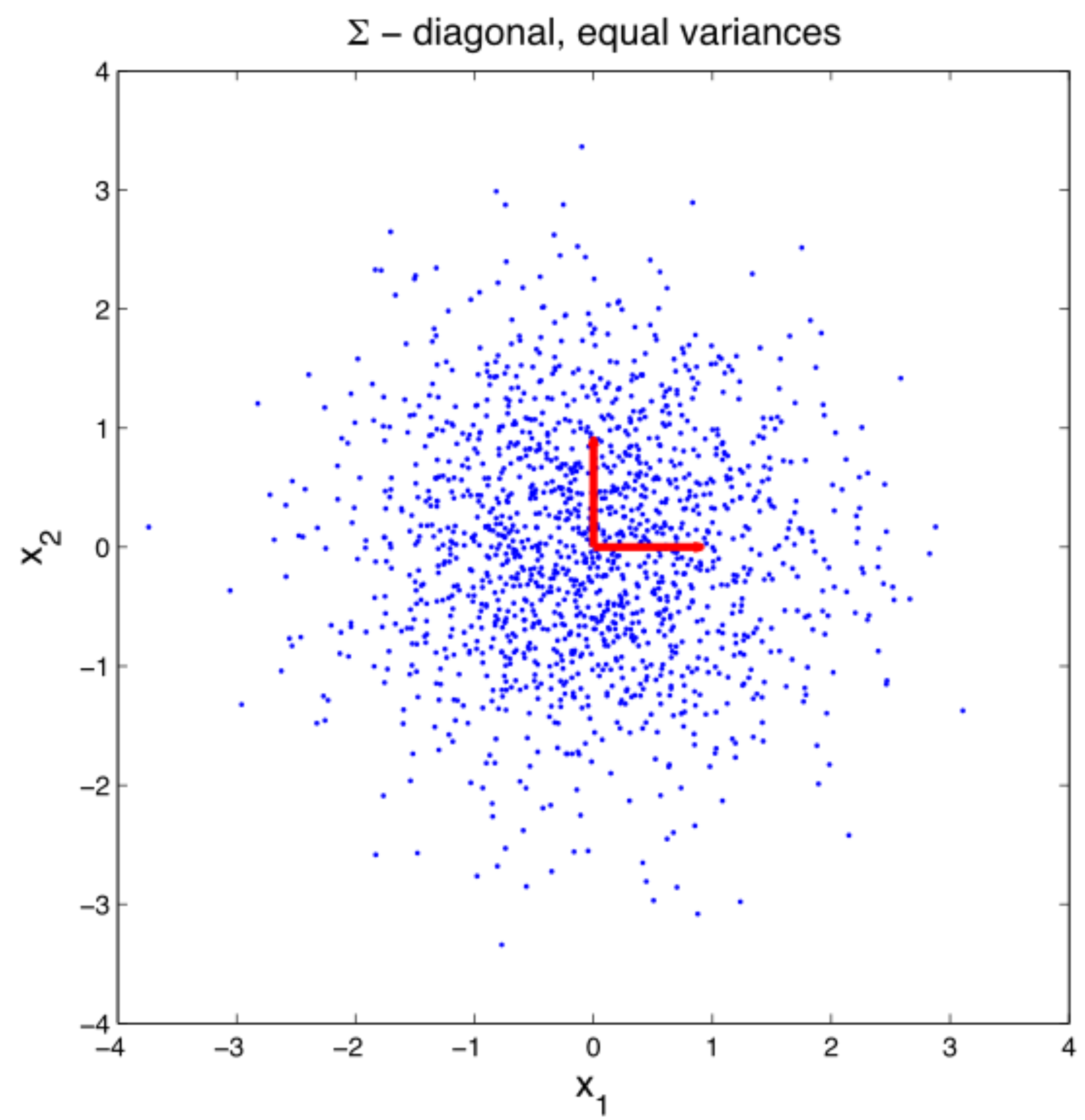
$$f_{\mathbf{X}}(x_1, \cdot, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)\right)$$

https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Example: 2D Normal Distribution

- Independent + equal variances

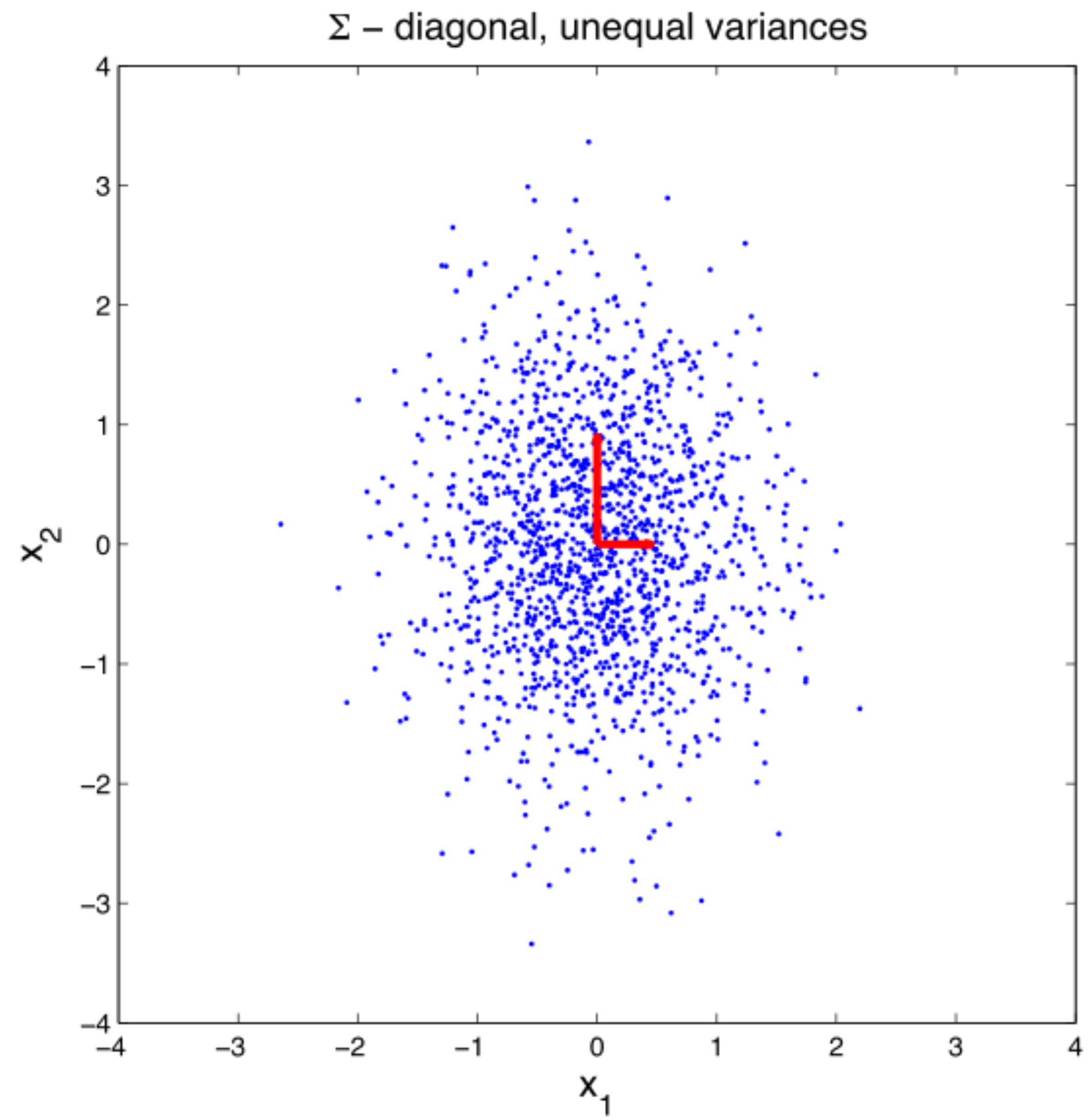
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = V \Lambda V^\top$$



Example: 2D Normal Distribution

- Independent + unequal variances

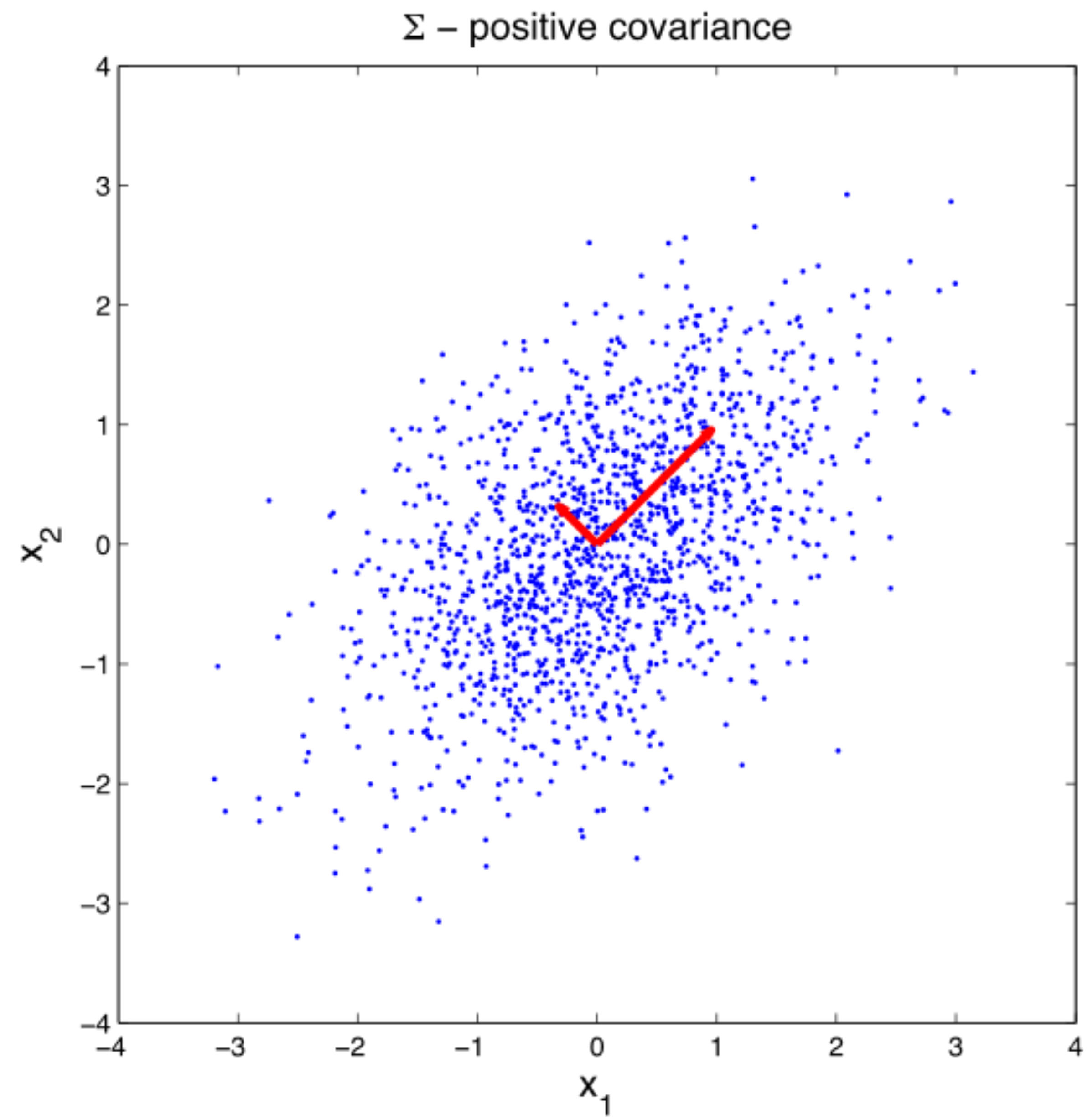
$$\Sigma = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix} = V \Lambda V^\top$$



Example: 2D Normal Distribution

- Positive covariance

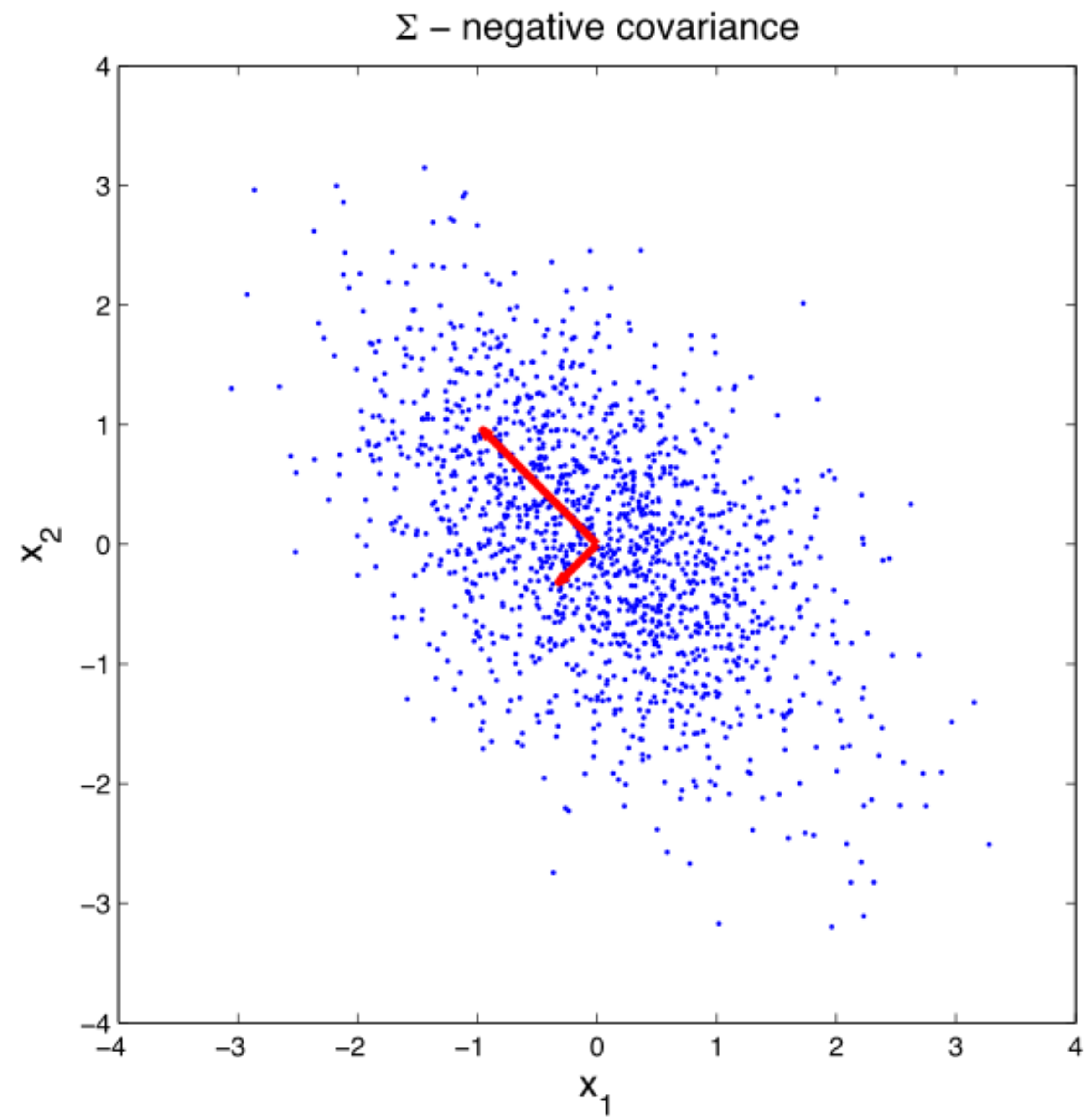
$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} = V \Lambda V^\top$$



Example: 2D Normal Distribution

- Negative covariance

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} = V \Lambda V^\top$$



Correlation

- Correlation: another measure of dependence

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

- Normalized covariance, bounded [-1, 1]
- Cov[X,Y] = 0 means X, Y are uncorrelated
- Independence implies uncorrelated
- Uncorrelated does not imply independent

Correlation Exercise

- The following are dependent, are they correlated?

$$X \sim U(-1, 1)$$
$$Y = X^2$$

Expectation with Multiple RVs

- Conditional expectation

$$E[X|Y=y] = \int x f_X(x|Y=y)dx$$

- Nested expectations

$$E[X] = E[E[X|Y]]$$

- Independence

$$E[XY] = E[X]E[Y]$$

Sum of Two RVs

- Sums of two normal RVs are also normally distributed

$$X \sim N(\mu_X, \sigma_X^2)$$
$$Y \sim N(\mu_Y, \sigma_Y^2)$$
$$Z = X + Y$$

- Independent case

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

- Dependent case

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$$

Central Limit Theorem

- Arithmetic mean of large numbers of independent and identically distributed RVs are approximately normally distributed

$$S_n = \sum_{i=1}^{n} X_i \rightarrow S_n \sim N(\mu, \frac{\sigma^2}{n})$$

Common Multivariate Distribution

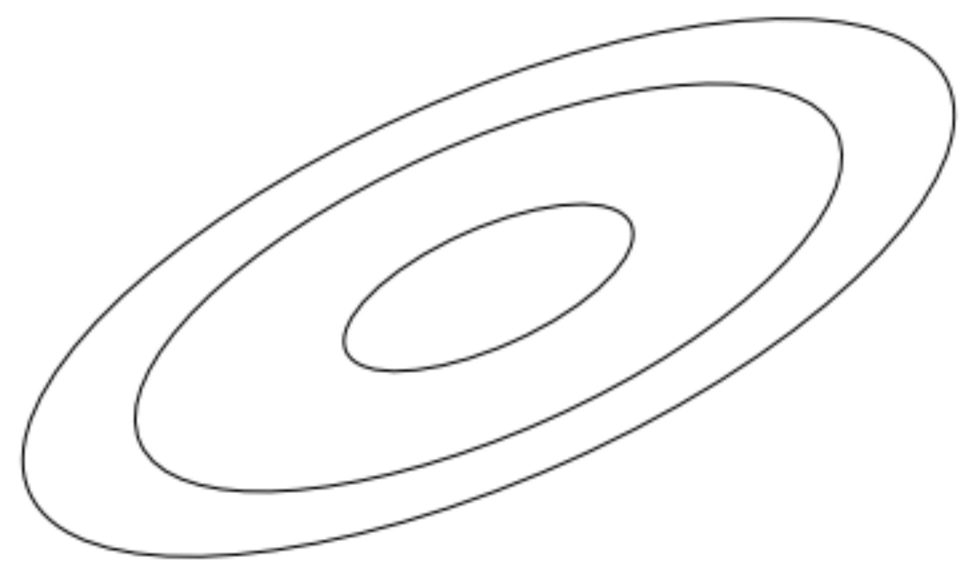
Multivariate Gaussian

- Extremely useful distribution
- Common for modeling “noise” in statistical algorithms
  - Central Limit Theorem of large number of small independent random perturbations
- Convenience for analytical manipulations because of simple closed form solutions

Multivariate Gaussian

$$f_{\mathbf{X}}(x_1, \cdot, x_k; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)\right)$$

$$X \sim N(\mu, \Sigma)$$



Gaussian Marginals / Conditionals

- Multivariate normal distribution

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{bmatrix} \right)$$

- Marginal distributions

$$\mathbf{x} \sim N(\mu_x, \Sigma_x)$$
$$\mathbf{y} \sim N(\mu_y, \Sigma_y)$$

- Conditional distributions:

$$\mathbf{x}|\mathbf{y} \sim N(\mu_x + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^\top)$$
$$\mathbf{y}|\mathbf{x} \sim N(\mu_y + \Sigma_{xy}^\top\Sigma_x^{-1}(x - \mu_x), \Sigma_y - \Sigma_{xy}^\top\Sigma_y^{-1}\Sigma_{xy})$$