Computer Science

CSC535: Probabilistic Graphical Models

Probability Primer : Continuous Probability

Prof. Jason Pacheco

Administrative Items

Do not use Gradescope! Use D2L!

- If it’s not on D2L it won’t get graded and you’ll receive a zero
- This is the 3rd announcement

- HW1 Due Tonight @ 11:59 PM (**D2L**)

- HW2 Out Tonight (will announce on Piazza)

Outline

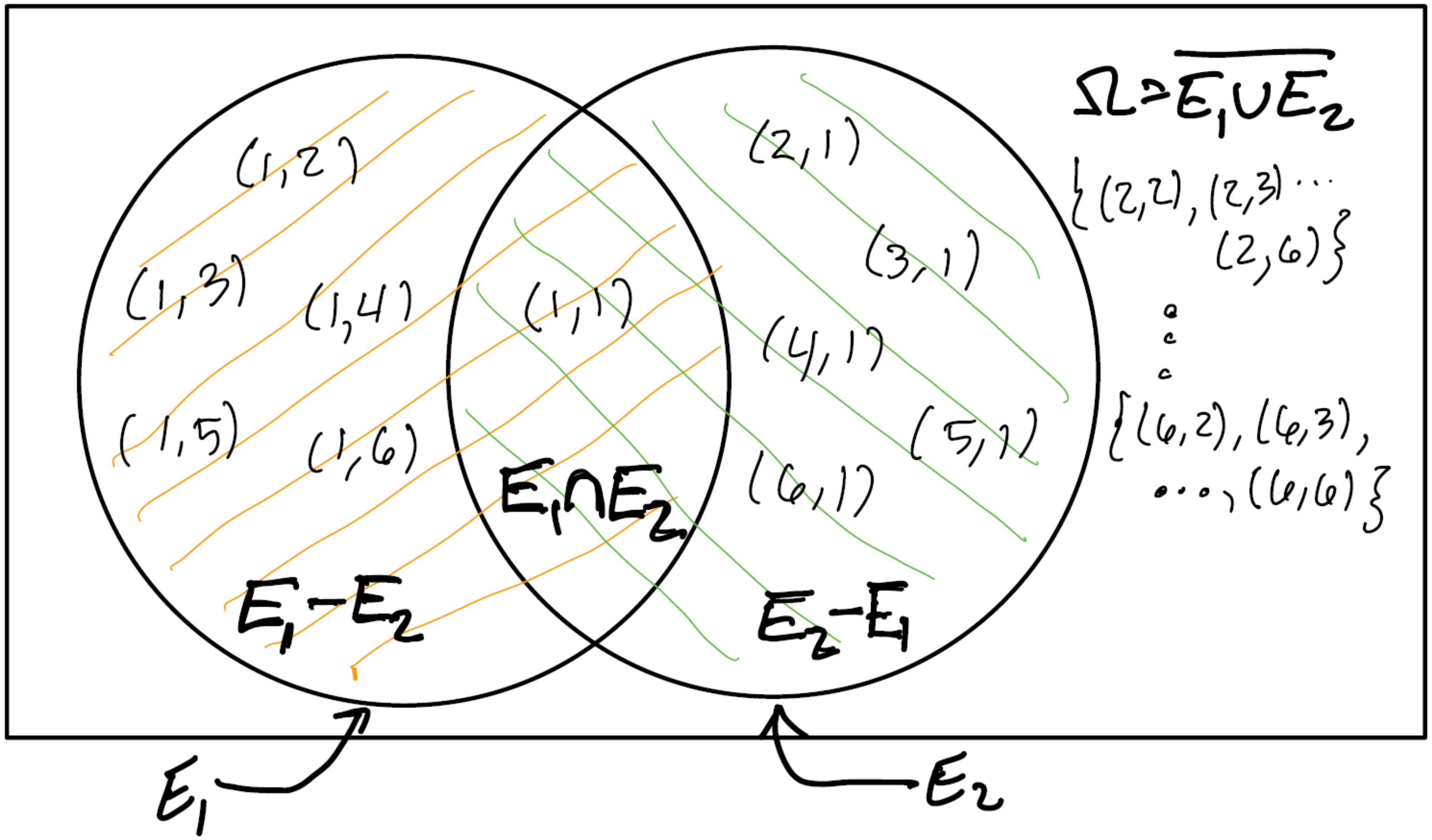
- Motivation and Definitions
- Fundamental Rules of Probability (recap)
- Useful Continuous Distributions / Densities

Outline

- Motivation and Definitions
- Fundamental Rules of Probability (recap)
- Useful Continuous Distributions / Densities

Probability Space

Recall that we can think of outcomes of a random experiment as a space…



…this has a formal mathematical definition

Probability Space

1 A **sample space** Ω: *set of all possible outcomes* of the experiment.

2 An **event space** $\mathcal{F}$: Sets of allowable events $E \in \mathcal{F}$

3 A **probability function** $P : \mathcal{F} \to \mathbf{R}$ satisfying:

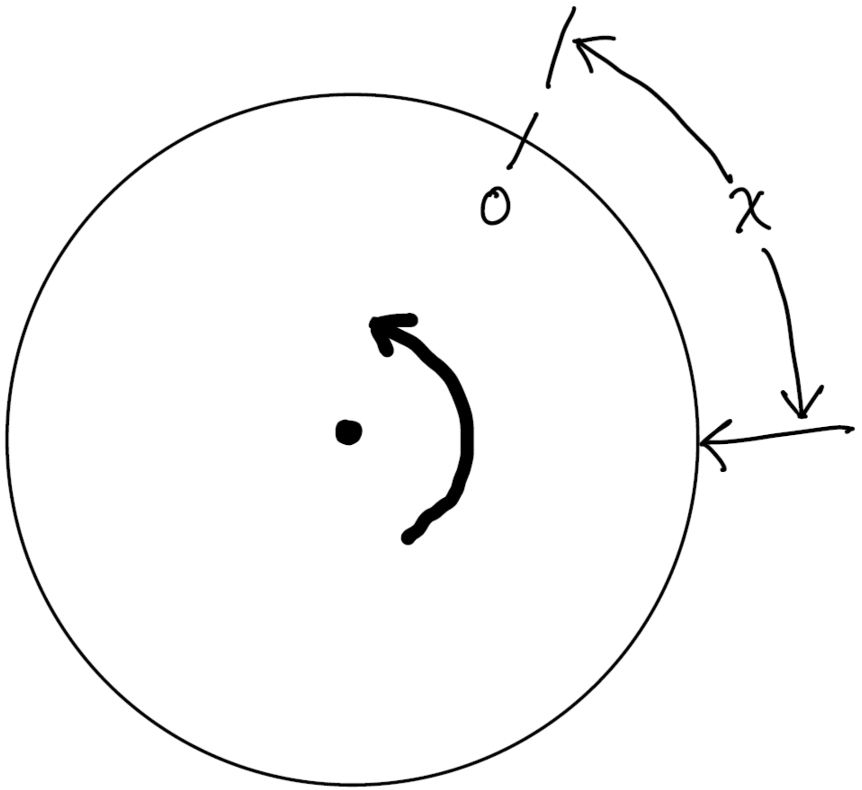
1. For any event E, $0 \leq P(E) \leq 1$
2. $P(\Omega) = 1 \text{ and } P(\emptyset) = 0$
3. For any *finite* or *countably infinite* sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$

$$P\left(\bigcup_{i \geq 1} E_i\right) = \sum_{i \geq 1} P(E_i)$$

Axioms of Probability

Continuous Probability

Experiment Spin continuous wheel and measure X displacement from 0



Question Assuming uniform probability, what is $p(X = x)$?

Continuous Probability

- Let $p(X = x) = \pi$ be the probability of any single outcome

- Let $S(k)$ be set of any k *distinct* points in $[0, 1)$ then,

$$P(x \in S(k)) = k\pi$$

- Since $0 < P(x \in S(k)) < 1$ by axioms of probability, $k\pi < 1$ for any k

- Therefore: $\pi = 0$ and $P(x \in S(k)) = p(X = x) = 0$

Continuous Probability

- We have a well-defined event that x takes a value in set $x \in S(k)$
- Clearly this event can happen… i.e. **it is possible**
- But we have shown it has zero probability of occurring,

$$P(x \in S(k)) = 0$$

- By the axioms of probability, the probability that it **doesn't happen** is,

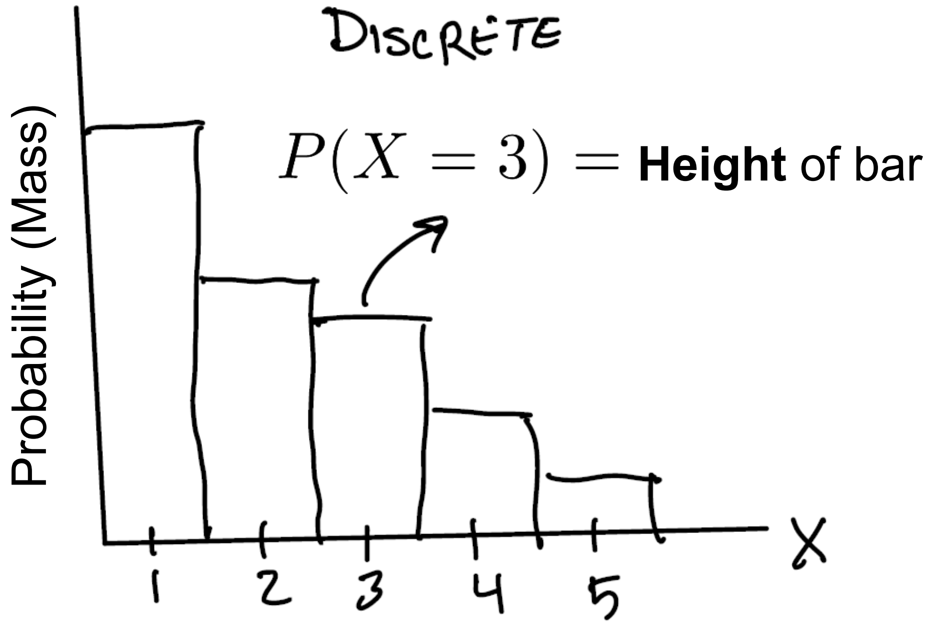
$$P(x \notin S(k)) = 1 - P(x \in S(k)) = 1$$

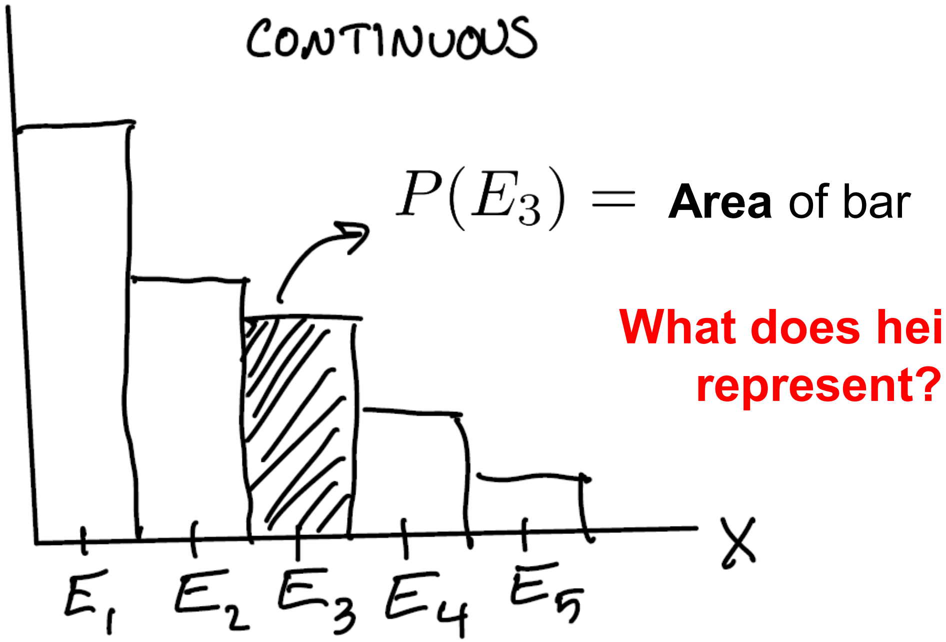
We seem to have a paradox!

Solution Rethink how we interpret probability in continuous setting

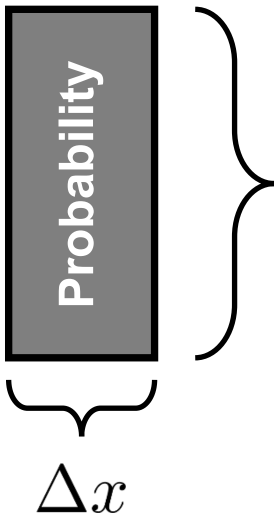
- Define events as *intervals* instead of discrete values
- Assign probability to those intervals

Continuous Probability





What does height represent?



Height $= \dfrac{\text{Probability}}{\Delta x}$

Height represents *probability per unit* in the x-direction

We call this a **probability density** (as opposed to probability mass)

Continuous Probability

- We denote the **probability density function** (PDF) as, $p(X)$
- An event E corresponds to an *interval* $a \leq X < b$
- The probability of an interval is given by the *area under the PDF,*

$$P(a \leq X < b) = \int_a^b p(X = x)\, dx$$

- Specific outcomes have zero probability $P(X = 0) = P(x \leq X < x) = 0$
- But may have nonzero *probability density* $p(X = x)$

Continuous Probability

Definition *The <u>cumulative distribution function</u> (CDF) of a real-valued continuous RV X is the function given by,*

$$F(x) = P(X \leq x)$$

Different ways to represent probability of interval, CDF is just a convention.

- Can easily measure probability of closed intervals,

$$P(a \leq X < b) = F(b) - F(a)$$

- If X is *absolutely continuous* (i.e. differentiable) then,

Fundamental Theorem of Calculus

$$f(x) = \frac{dF(x)}{dx} \quad \text{and} \quad F(t) = \int_{-\infty}^{t} f(x)\,dx$$

Where $f(x)$ is the *probability density function* (PDF)

- Typically use shorthand P for CDF and p for PDF instead of F and f

Outline

- Motivation and Definitions
- Fundamental Rules of Probability (recap)
- Useful Continuous Distributions / Densities

Continuous Probability

Most definitions for discrete RVs hold, replacing PMF with PDF/CDF…

Two RVs X & Y are **independent** if and only if,

$$p(x,y) = p(x)p(y) \qquad \text{or} \qquad P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

Conditionally independent given Z iff,

Shorthand: $P(x) = P(X \le x)$

$$p(x,y \mid z) = p(x \mid z)p(y \mid z) \qquad \text{or} \qquad P(x,y \mid z) = P(x \mid z)P(y \mid z)$$

Probability chain rule,

$$p(x,y) = p(x)p(y \mid x) \qquad \text{and} \qquad P(x,y) = P(x)P(y \mid x)$$

Continuous Probability

…and by replacing summation with integration…

Law of Total Probability for continuous distributions,

$$p(x) = \int_{\mathcal{Y}} p(x, y)\, dy$$

Expectation of a continuous random variable,

$$\mathbf{E}[X] = \int_{\mathcal{X}} x \cdot p(x)\, dx$$

Covariance of two continuous random variables X & Y,

$$\mathbf{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] = \int_{\mathcal{X}} \int_{\mathcal{Y}} (x - \mathbf{E}[X])(y - \mathbf{E}[Y]) p(x, y)\, dx dy$$

Continuous Probability

Caution *Some technical subtleties arise in continuous spaces…*

For **discrete** RVs X & Y, the conditional

P(Y=y)=0 means impossible

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

is **undefined** when P(Y=y) = 0 … no problem.

For **continuous** RVs we have,

$$P(X \leq x \mid Y = y) = \frac{P(X \leq x, Y = y)}{P(Y = y)}$$

but numerator and denominator are 0/0.

P(Y=y)=0 means improbable, but not impossible

Continuous Probability

Defining the conditional distribution as a limit fixes this…

$$P(X \leq x \mid Y = y) = \lim_{\delta \to 0} P(X \leq x \mid y \leq Y \leq y + \delta)$$

$$= \lim_{\delta \to 0} \frac{P(X \leq x, y \leq Y \leq y + \delta)}{P(y \leq Y \leq y + \delta)}$$

$$= \lim_{\delta \to 0} \frac{P(X \leq x, Y \leq y + \delta) - P(X \leq x, Y \leq y)}{P(Y \leq y + \delta) - P(Y \leq y)}$$

$$= \int_{-\infty}^{x} \lim_{\delta \to 0} \frac{\frac{\partial}{\partial x} P(u, y + \delta) - \frac{\partial}{\partial x} P(u, y)}{P(y + \delta) - P(y)} \, du$$

(Fundamental theorem of calculus)

(Assume interchange limit / integral)

$$= \int_{-\infty}^{x} \lim_{\delta \to 0} \frac{\left(\frac{\partial}{\partial x} P(u, y + \delta) - \frac{\partial}{\partial x} P(u, y)\right) / \delta}{\left(P(y + \delta) - P(y)\right) / \delta} \, du$$

(Multiply by $\frac{\delta}{\delta} = 1$)

$$= \int_{-\infty}^{x} \frac{\frac{\partial^2}{\partial x \partial y} P(u, y)}{\frac{\partial}{\partial y} P(y)} \, du \quad = \int_{-\infty}^{x} \frac{p(u, y)}{p(y)} \, du$$

(Definition of partial derivative)

(Definition PDF)

Definition The conditional PDF is given by,

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

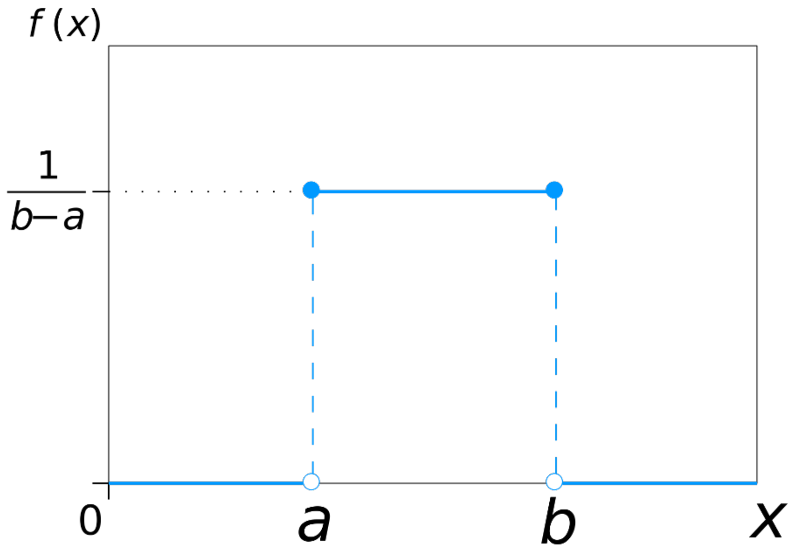
Outline

- Motivation and Definitions
- Fundamental Rules of Probability (recap)
- Useful Continuous Distributions / Densities

Useful Continuous Distributions

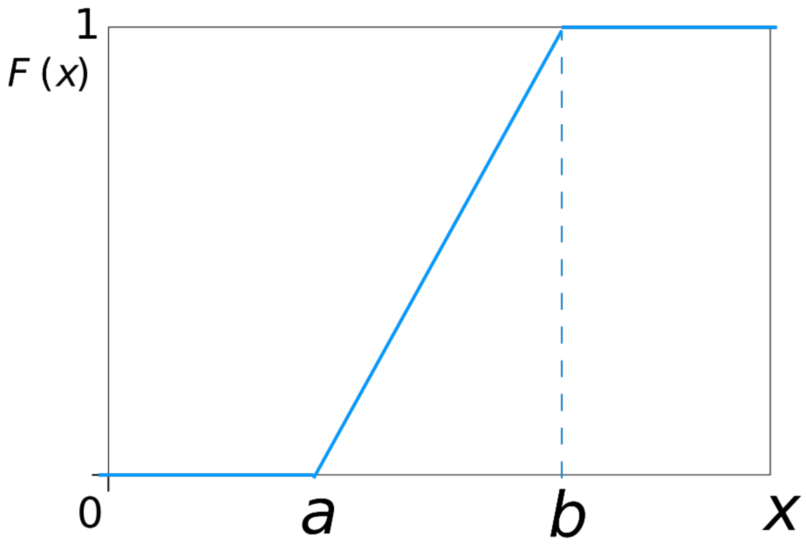
Uniform distribution on interval $[a, b]$,

$$p(x) = \begin{cases} 0 & \text{if } x \le a, \\ \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{if } b \le x \end{cases} \qquad P(X \le x) = \begin{cases} 0 & \text{if } x \le a, \\ \frac{x-a}{b-a} & \text{if } a \le x \le b, \\ 1 & \text{if } b \le x \end{cases}$$



Say that $X \sim U(a, b)$ whose moments are,

$$\mathbf{E}[X] = \frac{b+a}{2} \qquad \mathbf{Var}[X] = \frac{(b-a)^2}{12}$$



Suppose $X \sim U(0, 1)$ and we are told $X \le \frac{1}{2}$ what is the conditional distribution?

$$P(X \le x \mid X \le \tfrac{1}{2}) = U(0, \tfrac{1}{2})$$

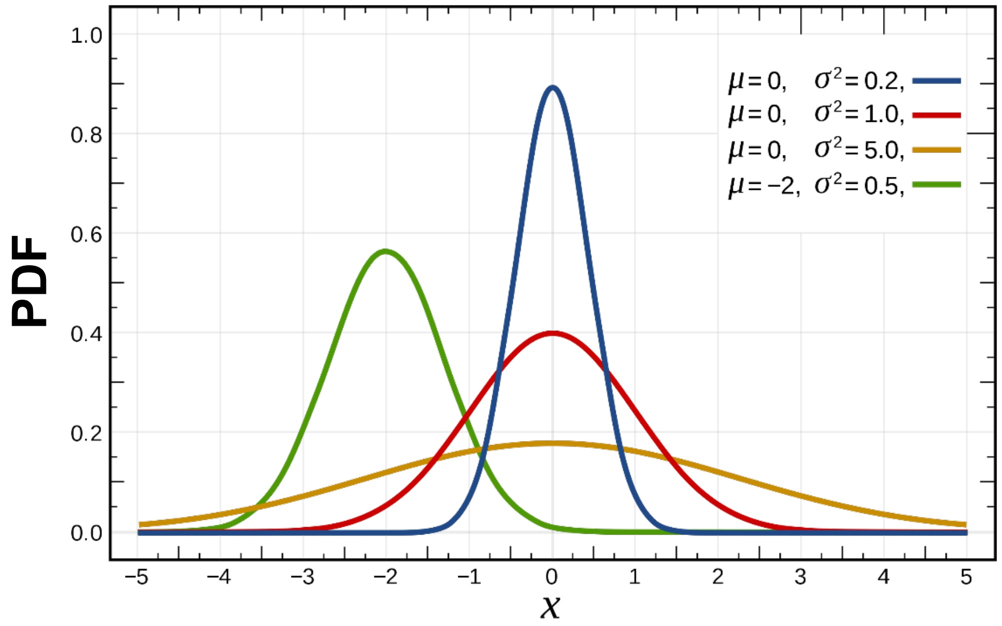
Holds generally: Uniform closed under conditioning

Useful Continuous Distributions

Gaussian (a.k.a. Normal) distribution with mean mean (location) μ and variance (scale) σ^2 parameters,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2}(x-\mu)^2/\sigma^2$$

We say $X \sim \mathcal{N}(\mu, \sigma^2)$.



Useful Properties

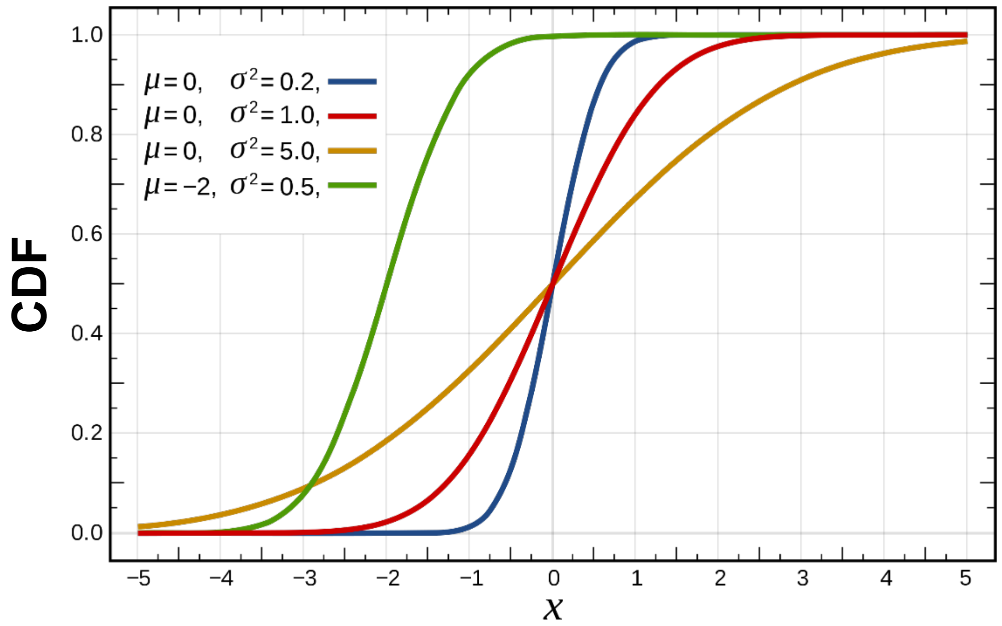
- Closed under additivity:

$$X \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

$$X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- Closed under linear functions (a and b constant):

$$aX + b \sim \mathcal{N}(a\mu_x + b, a^2\sigma_x^2)$$

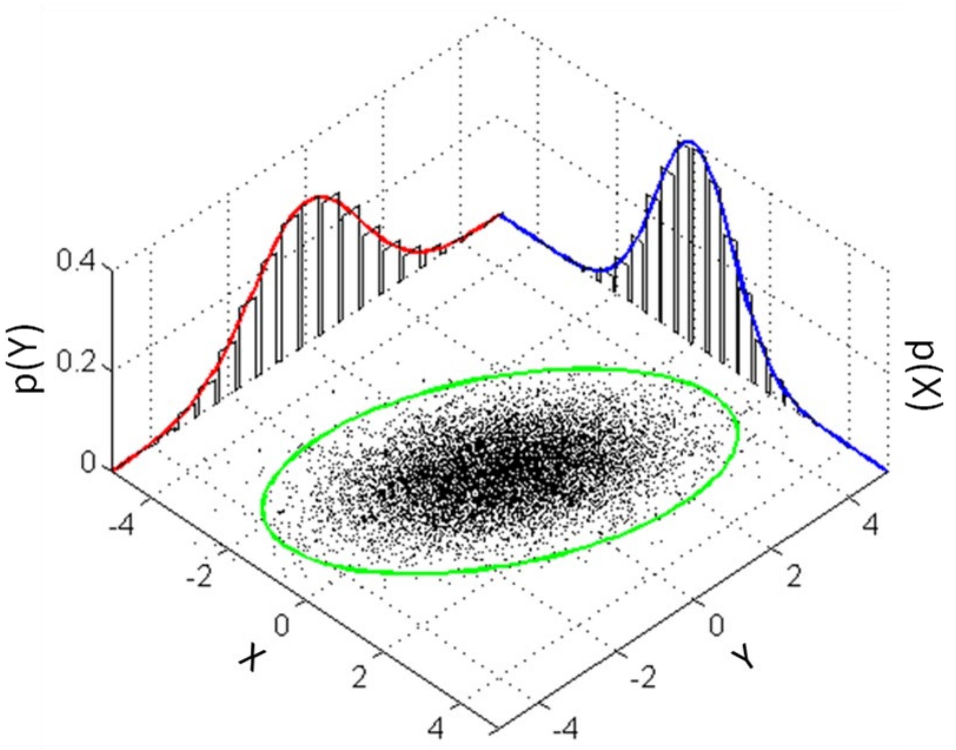


Useful Continuous Distributions

Multivariate Gaussian On RV $X \in \mathcal{R}^d$ with mean $\mu \in \mathcal{R}^d$ and positive semidefinite covariance matrix $\Sigma \in \mathcal{R}^{d \times d}$,

$$p(x) = |2\pi\Sigma|^{-1/2} \exp -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$$

Moments given by parameters directly.



Useful Properties

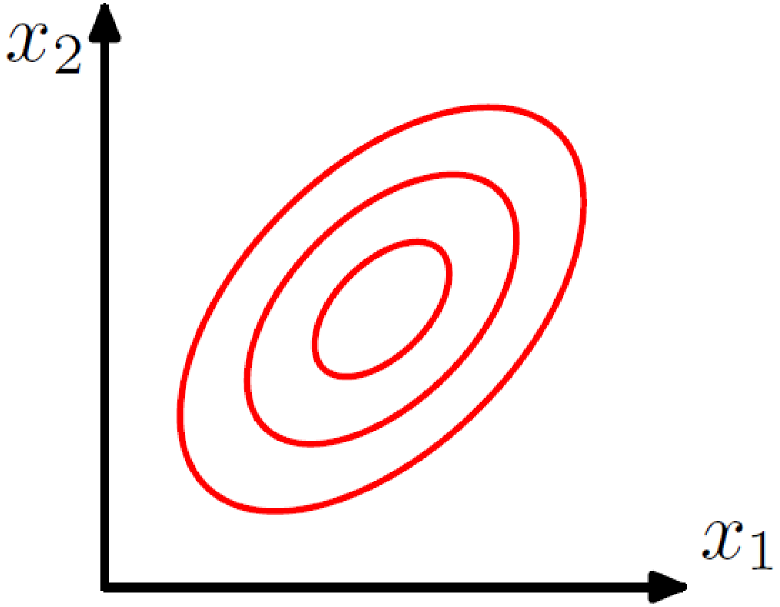
- Closed under additivity (same as univariate case)
- Closed under linear functions,

$$AX + b \sim \mathcal{N}(A\mu_x + b, A\Sigma A^T)$$

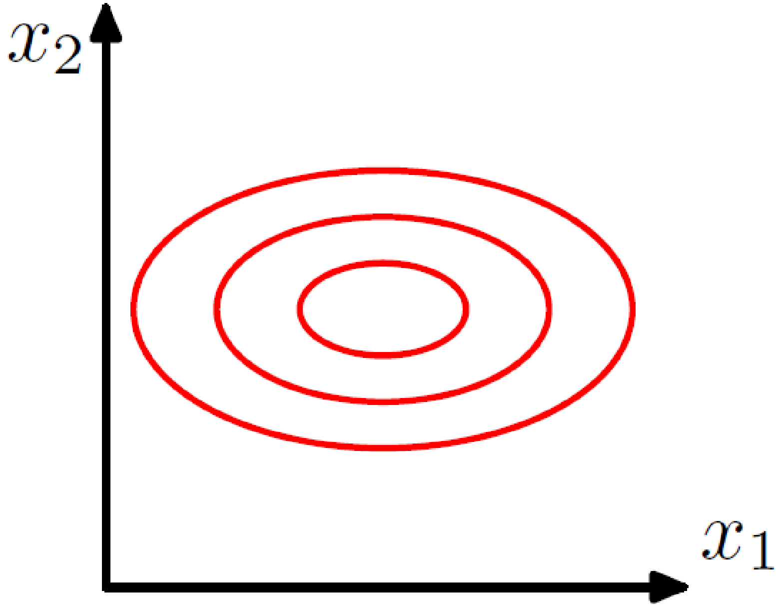
  Where $A \in \mathcal{R}^{m \times d}$ and $b \in \mathcal{R}^m$ (output dimensions may change)
- Closed under conditioning and marginalization

Covariance

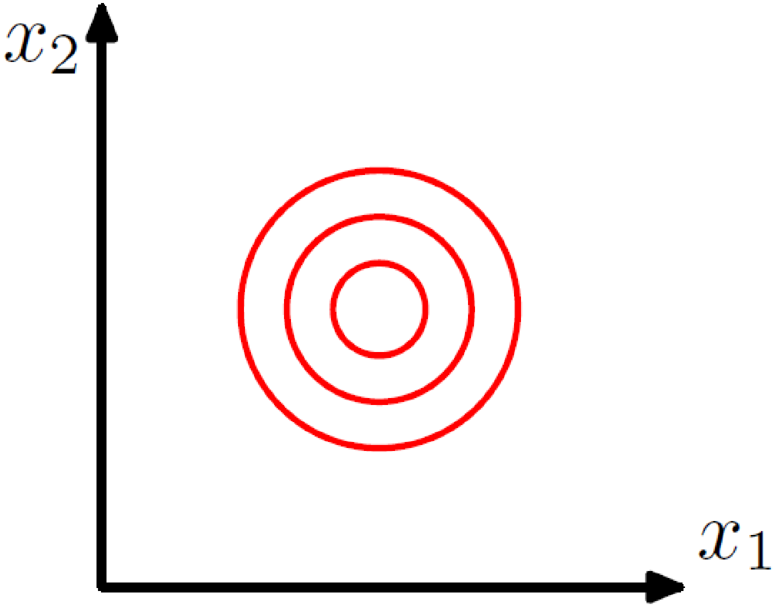
Captures correlation between random variables…can be viewed as set of ellipses…



Positive Correlation



Uncorrelated



Uncorrelated and same variance (isotropic / spherical)

Covariance Matrix

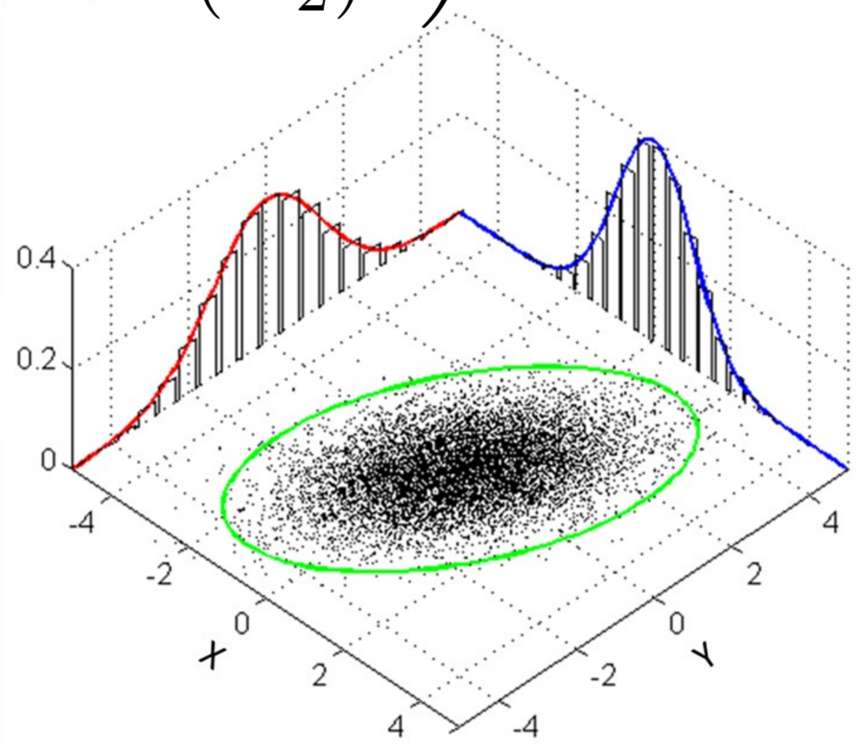
$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$

Covariance Matrix

Marginal variance of just the RV X_1

$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$

i.e. How "spread out" is the distribution in the X_1 dimension…



Covariance Matrix

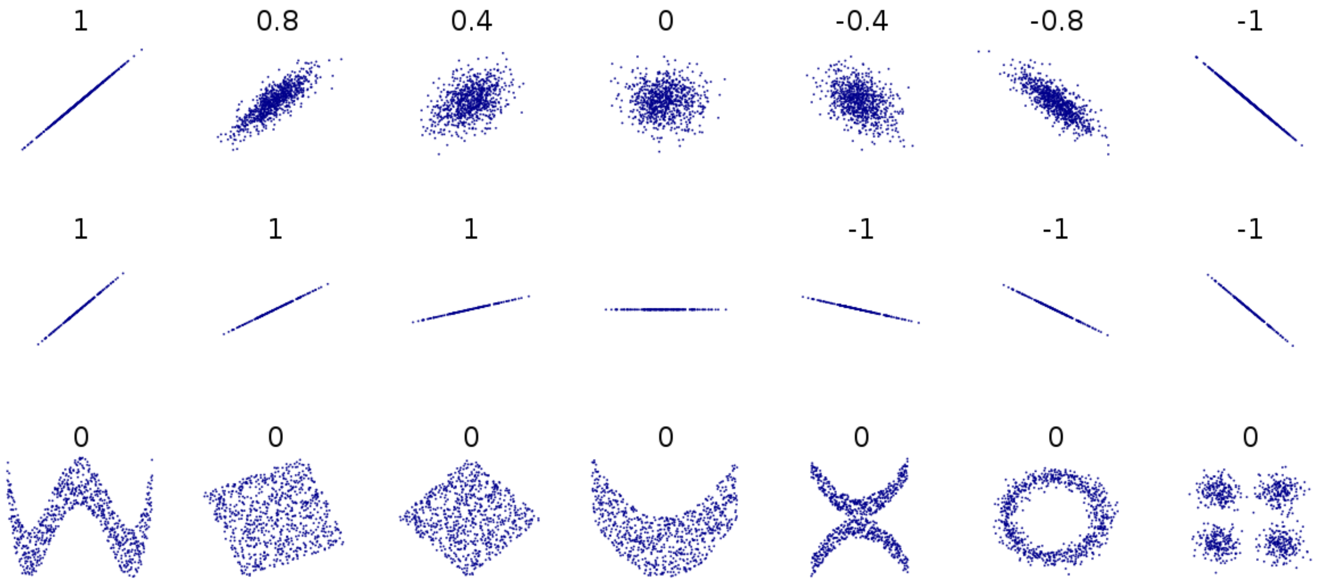
Correlation between X_1 and X_2

$$\Sigma = \mathrm{Cov}(X) = \begin{pmatrix} \mathrm{Var}(X_1) & \rho\sigma_{X_1}\sigma_{X_2} \\ \rho\sigma_{X_1}\sigma_{X_2} & \mathrm{Var}(X_2) \end{pmatrix}$$

Recall, correlation is given by:

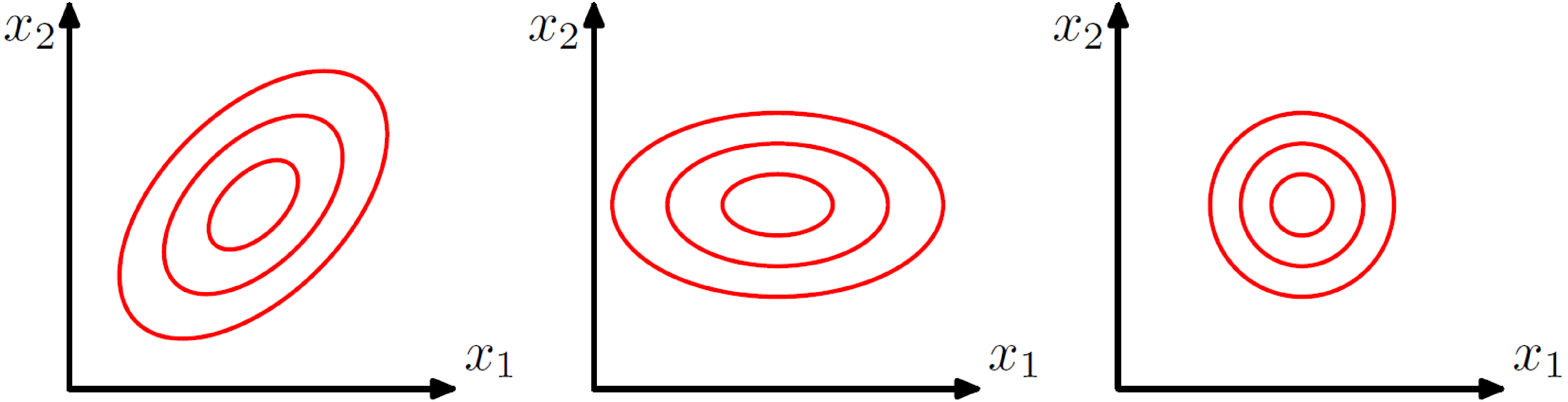
$$\rho = \frac{\mathbf{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

Captures *linear dependence of RVs*



Covariance

Captures correlation between random variables…can be viewed as set of ellipses…



Positive Correlation

$\rho > 0$

Full matrix Σ

Uncorrelated

$$\Sigma = \begin{pmatrix} \sigma^2_{X_1} & 0 \\ 0 & \sigma^2_{X_2} \end{pmatrix}$$

Isotropic / Spherical

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \sigma^2 I$$

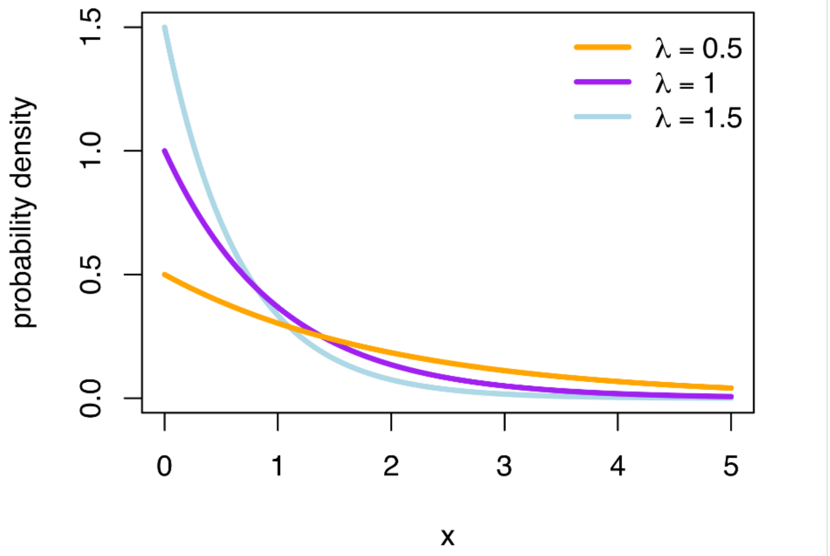
Useful Continuous Distributions

Exponential distribution with scale λ,

$$p(x) = \lambda e^{-\lambda x} \qquad P(x) = 1 - e^{-\lambda x}$$

for X>0. Moments given by,

$$\mathbf{E}[X] = \frac{1}{\lambda} \qquad \mathbf{Var}[X] = \frac{2}{\lambda^2}$$



Useful properties

- **Closed under conditioning** If $X \sim \text{Exponential}(\lambda)$ then,

$$P(X \geq s + t \mid X \geq s) = P(X \geq s) = e^{-\lambda s}$$

- **Minimum** Let $X_1, X_2, \ldots, X_N$ be i.i.d. exponentially distributed with scale parameters $\lambda_1, \lambda_2, \ldots, \lambda_N$ then,

$$P(\min(X_1, X_2, \ldots, X_N)) = \text{Exponential}(\textstyle\sum_i \lambda_i)$$