5. Probability Theory Review

STA3142 Statistical Machine Learning

Kibok Lee

Assistant Professor of
Applied Statistics / Statistics and Data Science

Mar 14, 2024

** Slides adapted from CS229 @ Stanford*

연세대학교
YONSEI UNIVERSITY

Outline

Note: Basics as Recap

- This review assumes basic background in probability (events, sample space, probability axioms *etc.*) and focuses on concepts useful to machine learning in general.

Definitions, Axioms, and Corollaries

- Performing an **experiment** $\rightarrow$ **outcome**
- **Sample Space** (S): set of all possible outcomes of an experiment
- **Event** (E): a subset of S $(E \subseteq S)$
- **Probability** (**Bayesian** definition)

  A number between 0 and 1 to which we ascribe meaning

  i.e., our belief that an event E occurs
- **Frequentist** definition of probability

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}.$$

Axiom 1: $0 \leq P(E) \leq 1$.
Axiom 2: $P(S) = 1$.
Axiom 3: if E and F are mutually exclusive ($E \cap F = \emptyset$), then $P(E) + P(F) = P(E \cup F)$.
Corollary 1: $P(E^C) = 1 - P(E)$. $\quad$ ($= P(S) - P(E)$; E^C is a complement of E.)
Corollary 2: if $E \subseteq F$, then $P(E) \leq P(F)$.
Corollary 3: $P(E \cup F) = P(E) + P(F) - P(EF)$. (inclusion-exclusion principle)
General inclusion-exclusion principle:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(E_{i1} E_{i2} \cdots E_{ir}).$$

Equally Likely Outcomes: Define S as a sample space with equally likely outcomes. Then

$$P(E) = \frac{|E|}{|S|}.$$

Conditional Probability and Bayes' Rule

For any events A, B such that $P(B) \neq 0$, we define:

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

Let's apply conditional probability to obtain **Bayes' Rule!**

$$\begin{aligned} P(B|A) &= \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} \\ &= \boxed{\frac{P(B)P(A|B)}{P(A)}.} \end{aligned}$$

Conditioned Bayes' Rule: given events A, B, and C,

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}.$$

Law of Total Probability

Let $B_1, ..., B_n$ be n disjoint events whose union is the entire space sample space. Then, for any event A,

$$\begin{aligned} P(A) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i). \end{aligned}$$

We can write Bayes' Rule as:

$$\begin{aligned} P(B_k|A) &= \frac{P(B_k)P(A|B_k)}{P(A)} \\ &= \boxed{\frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}}. \end{aligned}$$

Law of Total Probability

Treasure chest **A** holds 100 gold coins. Treasure chest **B** holds 60 gold and 40 silver coins. Choose a treasure chest uniformly at random, and pick a coin from that chest uniformly at random. If the coin is gold, then what is the probability that you chose chest **A**?[1]

Solution:

$$
\begin{aligned}
P(A|G) &= \frac{P(A)P(G|A)}{P(A)P(G|A) + P(B)P(G|B)} \\
&= \frac{0.5 \times 1}{0.5 \times 1 + 0.5 \times 0.6} \\
&= \boxed{0.625.}
\end{aligned}
$$

[1] Question based on slides by Koochak & Irvin

Chain Rule

For any n events $A_1, \ldots, A_n$, the joint probability can be expressed as a product of conditionals:

$$\begin{aligned} &P(A_1 \cap A_2 \cap \ldots \cap A_n) \\ &\quad = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1)\ldots P(A_n|A_{n-1} \cap A_{n-2} \cap \ldots \cap A_1). \end{aligned}$$

Independence

Events A and B are independent if

$$P(AB) = P(A)P(B).$$

We denote this as $A \perp B$. From this, we know that if $A \perp B$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Implication: If two events are independent, observing one event does not change the probability that the other event occurs.

In general: events $A_1, ..., A_n$ are **mutually independent** if

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

for any subset $S \subseteq \{1, \cdots, n\}$.

Outline

Random Variables

- A **random variable** X is a variable that probabilistically takes on different values. It maps outcomes to real values.
- X takes on values in $\mathrm{Val}(X) \subseteq \mathbb{R}$ or Support $\mathrm{Sup}(X)$.
- $X = k$ is the **event** that random variable X takes on value k.

Discrete RVs:

- $\mathrm{Val}(X)$ is a set.
- $P(X = k)$ can be nonzero.

Continuous RVs:

- $\mathrm{Val}(X)$ is a range.
- $P(X = k) = 0$ for all k. $P(a \leq X \leq b)$ can be nonzero.

Probability Mass Function (PMF)

Given a **discrete** RV X, a PMF maps values of X to probabilities.

$$p_X(x) := p(x) := P(X = x).$$

For a valid PMF, $\sum_{x \in \mathrm{Val}(x)} p_X(x) = 1$.

Cumulative Distribution Function (CDF)

A CDF maps a continuous RV to a probability (*i.e.*, $\mathbb{R} \to [0, 1]$)

$$F_X(a) := F(a) := P(X \leq a).$$

A CDF must satisfy the following:

- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.
- If $a \leq b$, then $F_X(a) \leq F_X(b)$. (*i.e.*, CDF must be non-decreasing.)

Also note: $P(a \leq X \leq b) = F_X(b) - F_X(a)$.

Probability Density Function (PDF)

PDF of a continuous RV is simply the derivative of the CDF.

$$f_X(x) := f(x) := \frac{dF_X(x)}{dx}.$$

Thus,

$$P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx.$$

A valid PDF must be such that

- for all real numbers x, $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

Outline

Expectation

Let g be an arbitrary real-valued function.

- If X is a discrete RV with PMF p_X:

$$\mathbb{E}[g(X)] := \sum_{x \in \mathrm{Val}(X)} g(x) p_X(x).$$

- If X is a continuous RV with PDF f_X:

$$\mathbb{E}[g(X)] := \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Intuitively, expectation is a weighted average of the values of $g(x)$, weighted by the probability of x.

Properties of Expectation

For any constant $a \in \mathbb{R}$ and arbitrary real function f:

- $\mathbb{E}[a] = a$.
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$.

Linearity of Expectation. Given n real-valued functions $f_1(X), \ldots, f_n(X)$,

$$\mathbb{E}[\sum_{i=1}^{n} f_i(X)] = \sum_{i=1}^{n} \mathbb{E}[f_i(X)].$$

Law of Total Expectation. Given two RVs X, Y:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Note: $\mathbb{E}[X|Y] = \sum_{x \in \mathrm{Val}(x)} x p_{X|Y}(x|y)$ is a function of Y; for more information, see Appendix.

Example of Law of Total Expectation

El Goog sources two batteries, A and B, for its phone. A phone with battery A runs on average 12 hours on a single charge, but only 8 hours on average with battery B. El Goog puts battery A in 80% of its phones and battery B in the rest. If you buy a phone from El Goog, how many hours do you expect it to run on a single charge?

Solution: Let L be the time your phone runs on a single charge. We know the following:

- $p_X(A) = 0.8$, $p_X(B) = 0.2$,
- $\mathbb{E}[L|A] = 12$, $\mathbb{E}[L|B] = 8$.

Then, by Law of Total Expectation,

$$\begin{aligned}\mathbb{E}[L] = \mathbb{E}[\mathbb{E}[L|X]] &= \sum_{X \in \{A,B\}} \mathbb{E}[L|X] p_X(X) \\ &= \mathbb{E}[L|A] p_X(A) + \mathbb{E}[L|B] p_X(B) \\ &= 12 \times 0.8 + 8 \times 0.2 = \boxed{11.2}.\end{aligned}$$

Variance

The **variance** of a RV X measures how concentrated the distribution of X is around its mean.

$$\begin{aligned}\mathrm{Var}(X) &:= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

Interpretation: $\mathrm{Var}(X)$ is the expected deviation of X from $\mathbb{E}[X]$.
Properties: For any constant $a \in \mathbb{R}$ and real-valued function $f(X)$,

- $\mathrm{Var}[a] = 0$.
- $\mathrm{Var}[af(X)] = a^2 \, \mathrm{Var}[f(X)]$.

Example Distributions

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 $k=0,1,\ldots,n$	np	$np(1-p)$
Geometric(p)	$p(1-p)^{k-1}$ for $k=1,2,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k=0,1,\ldots$	λ	λ
Uniform(a,b)	$\frac{1}{b-a}$ for all $x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian(μ,σ^2)	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for all $x\in(-\infty,\infty)$	μ	σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$ for all $x\geq 0, \lambda\geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Outline

Joint and Marginal Distributions

- **Joint PMF** for discrete RV's X, Y:

$$p_{XY}(x, y) = P(X = x, Y = y).$$

Note that $\sum_{x \in \mathrm{Val}(X)} \sum_{y \in \mathrm{Val}(Y)} p_{XY}(x, y) = 1$.

- **Marginal PMF** of X, given joint PMF of X, Y:

$$p_X(x) = \sum_y p_{XY}(x, y).$$

Joint and Marginal Distributions

- **Joint PDF** for continuous X, Y:

$$f_{XY}(x,y) = \frac{\delta^2 F_{XY}(x,y)}{\delta_x \delta_y}.$$

  Note that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) dx dy = 1$.

- **Marginal PDF** of X, given joint PDF of X, Y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y) dy.$$

Joint and Marginal Distributions for Multiple RVs

- **Joint PMF** for discrete RV's $X_1, \ldots, X_n$:

$$p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

Note that $\sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} p(x_1, \ldots, x_n) = 1$.

- **Marginal PMF** of X_1, given joint PMF of $X_1, \ldots, X_n$:

$$p_{X_1}(x_1) = \sum_{x_2} \cdots \sum_{x_n} p(x_1, \ldots, x_n).$$

Joint and Marginal Distributions for Multiple RVs

- **Joint PDF** for continuous RV's $X_1, \ldots, X_n$:

$$f(x_1, \ldots, x_n) = \frac{\delta^n F(x_1, \ldots, x_n)}{\delta x_1 \delta x_2 \ldots \delta x_n}.$$

Note that $\int_{x_1} \int_{x_2} \cdots \int_{x_n} f(x_1, \ldots, x_n) dx_1 dx_2 \ldots dx_n = 1$.

- **Marginal PDF** of X_1, given joint PDF of $X_1, \ldots, X_n$:

$$f_{X_1}(x_1) = \int_{x_2} \cdots \int_{x_n} f(x_1, \ldots, x_n) dx_1 dx_2 \cdots dx_n.$$

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \to \mathbb{R}$ of X, Y,

- for discrete X, Y:

$$\mathbb{E}[g(X,Y)] := \sum_{x \in \mathrm{Val}(X)} \sum_{y \in \mathrm{Val}(Y)} g(x,y) p_{XY}(x,y).$$

- for continuous X, Y:

$$\mathbb{E}[g(X,Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dx dy.$$

These definitions can be extended to multiple random variables in the same way as in the previous slide. For example, for n continuous RV's $X_1, \ldots, X_n$ and function $g : \mathbb{R}^n \to \mathbb{R}$:

$$\mathbb{E}[g(X)] = \int \int \cdots \int g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

Outline

Covariance

Intuitively: measures how one RV's value tends to move with another RV's value.
For RV's X, Y:

$$\begin{aligned}\mathrm{Cov}[X,Y] &:= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

- If $\mathrm{Cov}[X,Y] < 0$, then X and Y are negatively correlated.
- If $\mathrm{Cov}[X,Y] > 0$, then X and Y are positively correlated.
- If $\mathrm{Cov}[X,Y] = 0$, then X and Y are uncorrelated.

Properties Involving Covariance

- If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Thus,

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

  This is unidirectional! $\mathrm{Cov}[X, Y] = 0$ **does not imply** $X \perp Y$.

- **Variance of two variables:**

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X, Y].$$

  i.e., if $X \perp Y$, $Var[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

- **Special Case:**

$$\mathrm{Cov}[X, X] = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = \mathrm{Var}[X].$$

Outline

Conditional distributions for RVs

Works the same way with RV's as with events:

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}.$$

- For continuous X, Y:

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

- In general, for continuous $X_1, \ldots, X_n$:

$$f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots, x_n) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}{f_{X_2,\ldots,X_n}(x_2, \ldots, x_n)}.$$

Bayes' Rule for RVs

Also works the same ways for RV's as with events:

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in \mathrm{Val}(Y)} p_{X|Y}(x|y')p_Y(y')}.$$

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'}.$$

Chain Rule for RVs

Also works the same as with events:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_1) f(x_2|x_1) \cdots f(x_n|x_1, x_2, \ldots, x_{n-1}) \\ &= f(x_1) \prod_{i=2}^{n} f(x_i|x_1, x_2, \ldots, x_{i-1}). \end{aligned}$$

Independence for RVs

- For $X \perp Y$ to hold, it must that $F_{XY}(x, y) = F_X(x)F_Y(y)$ **for all values** of x, y.
- Since $f_{Y|X}(y|x) = f_Y(y)$ if $X \perp Y$, chain rule for mutually independent $X_1, \ldots, X_n$ is:

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i).$$

(very important assumption for a Naive Bayes classifier!)

Outline

Random Vectors

Given n RV's $X_1, \ldots, X_n$, we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}.$$

Note: all the notions of joint PDF/CDF will apply to X.
Given $g : \mathbb{R}^n \to \mathbb{R}^m$, we have:

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}, \quad \mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(x)] \\ \mathbb{E}[g_2(x)] \\ \vdots \\ \mathbb{E}[g_m(x)] \end{bmatrix}.$$

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose (i, j)-th entry contains the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} \mathrm{Cov}[X_1, X_1] & \cdots & \mathrm{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_n, X_1] & \cdots & \mathrm{Cov}[X_n, X_n] \end{bmatrix}.$$

Applying linearity of expectation and the fact that
$\mathrm{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$,

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top].$$

Properties:

- Σ is symmetric and PSD.
- If $X_i \perp X_j$ for all i, j, then $\Sigma = \mathrm{diag}(\mathrm{Var}[X_1], \ldots, \mathrm{Var}[X_n])$.

Outline

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{\det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

The univariate Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$, $X \in \mathbb{R}$ is just the special case of the multivariate Gaussian when $n = 1$.

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

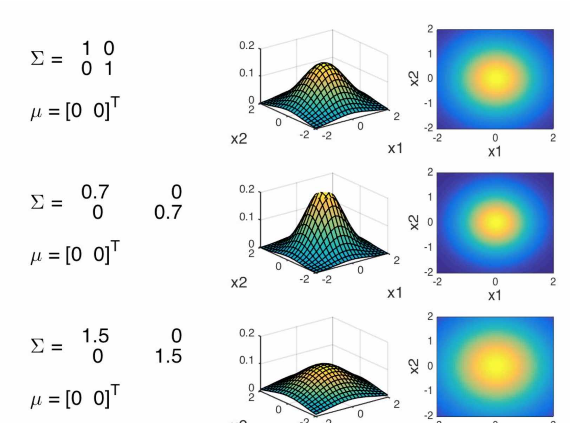
Note that if $\Sigma \in \mathbb{R}^{1\times 1}$, then $\Sigma = \mathrm{Var}[X_1] = \sigma^2$, and so $\Sigma^{-1} = \frac{1}{\sigma^2}$ and $\det(\Sigma)^{\frac{1}{2}} = \sigma$.

Some Nice Properties of MV Gaussians

- Marginals and conditionals of a joint Gaussian are Gaussian.
- A d-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2))$ is equivalent to a collection of d **independent** Gaussians $X_i \in \mathcal{N}(\mu_i, \sigma_i^2)$. This results in isocontours aligned with the coordinate axes.
- In general, the isocontours of a MV Gaussian are n-dimensional epllipsoids with principal axes in the directions of the eigenvectors of covariance matrix Σ (remember, Σ is PSD, so all n eigenvectors are non-negative). The axes' relative lengths depend on the eigenvalues of Σ.

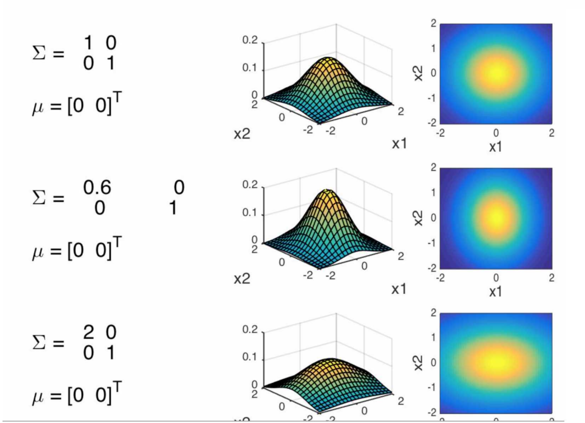
Visualizations MV Gaussians

Effect of changing variance:



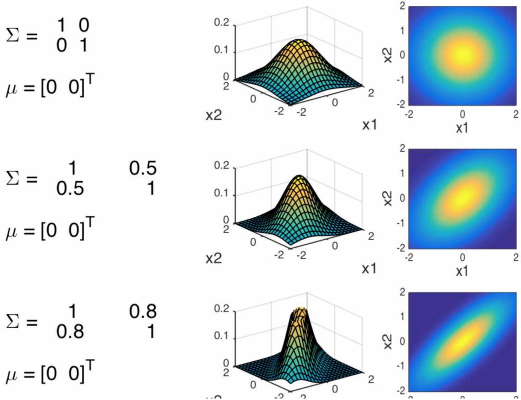
Visualizations MV Gaussians

If $\mathrm{Var}[X_1] \neq \mathrm{Var}[X_2]$:



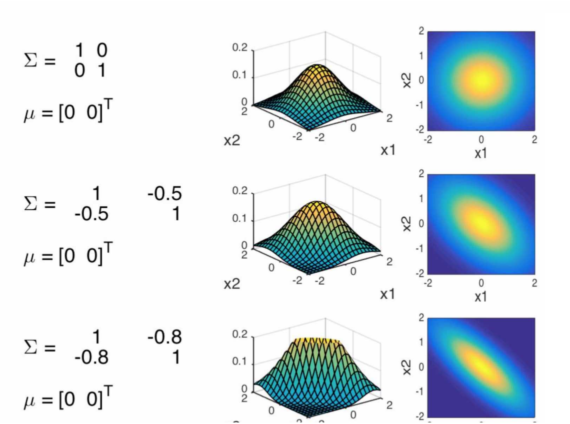
Visualizations MV Gaussians

If X_1 and X_2 are positively correlated:



Visualizations MV Gaussians

If X_1 and X_2 are negatively correlated:



Appendix: More on Total Expectation

Why is $\mathbb{E}[X|Y]$ a function of Y? Consider this following:

- $\mathbb{E}[X|Y=y]$ is a scalar that only depends on y.
- Thus, $\mathbb{E}[X|Y]$ is a random variable that only depends on Y. Specifically, $\mathbb{E}[X|Y]$ is a function of Y mapping $\mathrm{Val}(Y)$ to the real numbers.

An example: Consider a RV X such that

$$X = Y^2 + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0,1)$ is a standard Gaussian. Then,

- $\mathbb{E}[X|Y] = Y^2$,
- $\mathbb{E}[X|Y=y] = y^2$.

Appendix: More on Total Expectation

A derivation of Law of Total Expectation for discrete X, Y:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}\left[\sum_x xP(X=x|y)\right] = \sum_y \sum_x xP(X=x|Y)P(Y=y) \tag{1}$$

$$= \sum_y \sum_x xP(X=x, Y=y) \tag{2}$$

$$= \sum_x x \sum_y P(X=x, Y=y) \tag{3}$$

$$= \sum_x xP(X=x) = \boxed{\mathbb{E}[X]} \tag{4}$$

where (1) and (4) result from the definition of expectation, (2) results from the definition of cond. prob. and (3) results from marginalizing out Y.

Appendix: A proof of Conditioned Bayes' Rule

Repeatedly applying the definition of conditional probability, we have:

$$
\begin{aligned}
\frac{P(b|a,c)P(a|c)}{P(b|c)} &= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a|c)}{P(b|c)} \\
&= \frac{P(b,a,c)}{P(a,c)} \cdot \frac{P(a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b|c)P(c)} \\
&= \frac{P(b,a,c)}{P(b,c)} \\
&= P(a|b,c).
\end{aligned}
$$