# CS229 Section: Probability Theory

**Drew Kaul**

Content adapted from past CS229 teams, CS109 content

April 14, 2023

# Outline


1. Basics
2. Random Variables
3. Expectation-Variance
4. Joint Distributions
5. Covariance
6. RV Conditionals
7. Random Vectors
8. Multivariate Gaussian

# Note: Basics as Recap

- This review assumes basic background in probability (events, sample space, probability axioms etc.) and focuses on concepts useful to CS229 and 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: $P(E^C) = 1 - P(E) \qquad (= P(S) - P(E))$

Corollary 2: $E \subseteq F$, 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:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{i_1 < \cdots < i_r} P(E_{i_1} E_{i_2} \ldots E_{i_r})$$

Equally Likely Outcomes: Define $S$ as a sample space with equally likely outcomes. Then $P(E) = \dfrac{|E|}{|S|}$

# Conditional Probability and Bayes' Rule

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

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

# Conditional Probability and Bayes' Rule

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

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

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

# Conditional Probability and Bayes' Rule

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

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

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

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

$$= \boxed{\frac{P(B)P(A \mid B)}{P(A)}}$$

# Conditional Probability and Bayes' Rule

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

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

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

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

$$= \boxed{\frac{P(B)P(A \mid B)}{P(A)}}$$

**Conditioned Bayes' Rule**: given events $A, B, C$,

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

# Law of Total Probability

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

# Law of Total Probability

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

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$
$$= \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

# Law of Total Probability

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

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$
$$= \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

We can then write Bayes' Rule as:

# Law of Total Probability

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

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$
$$= \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

We can then write Bayes' Rule as:

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{P(A)}$$
$$= \boxed{\frac{P(B_k)P(A \mid B_k)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}}$$

# 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:**

[1] Question based on slides by Koochak & Irvin

# 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:**

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

[1] Question based on slides by Koochak & Irvin

# Chain Rule

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

$$P(A_1 \cap A_2 \cap ... \cap A_n)$$
$$= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_2 \cap A_1)...P(A_n \mid A_{n-1} \cap A_{n-2} \cap ... \cap A_1)$$

# Independence

Events $A, B$ are independent if

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

We denote this as $A \perp B$.

# Independence

Events $A, B$ are independent if

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

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

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

# Independence

Events $A, B$ are independent if

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

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

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

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

# Independence

Events $A, B$ are independent if

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

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

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

**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 $S \subseteq \{1, ..., n\}$.

# Outline


1 Basics

2 Random Variables

3 Expectation-Variance

4 Joint Distributions

5 Covariance

6 RV Conditionals

7 Random Vectors

8 Multivariate Gaussian

# 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 $Val(X) \subseteq \mathbb{R}$ or Support $\text{Sup}(X)$
- $X = k$ is the **event** that random variable $X$ takes on value $k$

# 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 $Val(X) \subseteq \mathbb{R}$ or Support $\text{Sup}(X)$
- $X = k$ is the **event** that random variable $X$ takes on value $k$

**Discrete RVs:**

- $Val(X)$ is a set
- $P(X = k)$ can be nonzero

# 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 $Val(X) \subseteq \mathbb{R}$ or Support $\text{Sup}(X)$
- $X = k$ is the **event** that random variable $X$ takes on value $k$

**Discrete RVs:**

- $Val(X)$ is a set
- $P(X = k)$ can be nonzero

**Continuous RVs:**

- $Val(X)$ 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.

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

# 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, $\sum_{x \in Val(x)} p_X(x) = 1$.

# Cumulative Distribution Function (CDF)

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

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

# 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 fulfill 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 nondecreasing)

# 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 fulfill 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 nondecreasing)

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.

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

# 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


1 Basics

2 Random Variables

3 Expectation-Variance

4 Joint Distributions

5 Covariance

6 RV Conditionals

7 Random Vectors

8 Multivariate Gaussian

# Expectation

Let $g$ be an arbitrary real-valued function.

- If $X$ is a discrete RV with PMF $p_X$:

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

# Expectation

Let $g$ be an arbitrary real-valued function.

- If $X$ is a discrete RV with PMF $p_X$:

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

- If $X$ is a continuous RV with PDF $f_X$:

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

# Expectation

Let $g$ be an arbitrary real-valued function.

- If $X$ is a discrete RV with PMF $p_X$:

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

- 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 $f$:

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

# 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 $f_1(X), ..., f_n(X)$,

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

# 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 $f_1(X), ..., f_n(X)$,

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

**Law of Total Expectation**

Given two RVs $X, Y$:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

# 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 $f_1(X), ..., f_n(X)$,

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

**Law of Total Expectation**

Given two RVs $X, Y$:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

**N.B.** $\mathbb{E}[X \mid Y] = \sum_{x \in Val(x)} x p_{X|Y}(x|y)$ is a function of $Y$.

# 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?

# 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:

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

Then, by Law of Total Expectation,

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

# Variance

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

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

# Variance

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

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

**Interpretation:** $Var(X)$ is the expected deviation of $X$ from $\mathbb{E}[X]$.

# Variance

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

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

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

- $Var[a] = 0$
- $Var[af(X)] = a^2 Var[f(X)]$

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

# Outline


1 Basics

2 Random Variables

3 Expectation-Variance

4 Joint Distributions

5 Covariance

6 RV Conditionals

7 Random Vectors

8 Multivariate Gaussian

# Joint and Marginal Distributions

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

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

# Joint and Marginal Distributions

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

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

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

# Joint and Marginal Distributions

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

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

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

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

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

# 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 $X_1, ..., X_n$:

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

# Joint and Marginal Distributions for Multiple RVs

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

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

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

# Joint and Marginal Distributions for Multiple RVs

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

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

  Note that $\sum_{x_1} \sum_{x_2} ... \sum_{x_n} p(x_1, ..., x_n) = 1$
- **Marginal PMF** of $X_1$, given joint PMF of $X_1, ..., X_n$:

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

# Joint and Marginal Distributions for Multiple RVs

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

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

# Joint and Marginal Distributions for Multiple RVs

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

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

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

# Joint and Marginal Distributions for Multiple RVs

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

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

  Note that $\int_{x_1} \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_1 ... dx_n = 1$
- **Marginal PDF** of $X_1$, given joint PDF of $X_1, ..., X_n$:

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

# Expectation for multiple random variables

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

# 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 $X, Y$:

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

# 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 $X, Y$:

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

- for continuous $X, Y$:

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

# 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 $X, Y$:

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

- 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 $X_1, .., X_n$ and function $g : \mathbb{R}^n \to \mathbb{R}$:

# 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 $X, Y$:

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

- 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 $X_1, .., X_n$ and function $g : \mathbb{R}^n \to \mathbb{R}$:

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

# Outline


1 Basics

2 Random Variables

3 Expectation-Variance

4 Joint Distributions

5 Covariance

6 RV Conditionals

7 Random Vectors

8 Multivariate Gaussian

# Covariance

**Intuitively**: measures how much one RV's value tends to move with another RV's value.

# Covariance

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

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

# Covariance

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

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

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

## Properties Involving Covariance

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

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

## Properties Involving Covariance

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

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

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

## Properties Involving Covariance

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

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

  This is unidirectional! $Cov[X, Y] = 0$ **does not imply** $X \perp Y$
- **Variance of two variables:**

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

# Properties Involving Covariance

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

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

  This is unidirectional! $Cov[X, Y] = 0$ **does not imply** $X \perp Y$
- **Variance of two variables:**

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

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

## Properties Involving Covariance

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

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

  This is unidirectional! $Cov[X, Y] = 0$ **does not imply** $X \perp Y$
- **Variance of two variables:**

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

  i.e. if $X \perp Y$, $Var[X + Y] = Var[X] + Var[Y]$.
- **Special Case:**

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

# Outline


1. Basics
2. Random Variables
3. Expectation-Variance
4. Joint Distributions
5. Covariance
6. RV Conditionals
7. Random Vectors
8. Multivariate Gaussian

# 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 $X_1, ..., X_n$:

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

# Bayes' Rule for RVs

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

- For discrete $X, Y$:

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

- 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 way as with events:

$$f(x_1, x_2, ..., x_n) = f(x_1)f(x_2|x_1)...f(x_n|x_1, x_2, ..., x_{n-1})$$
$$= f(x_1)\prod_{i=2}^{n} f(x_i|x_1, ..., x_{i-1})$$

# Independence for RVs

- For $X \perp Y$ to hold, it must be that $F_{XY}(x, y) = F_X(x)F_Y(y)$ **FOR ALL VALUES** of $x, y$.

# Independence for RVs

- For $X \perp Y$ to hold, it must be 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 $X_1, ..., X_n$ is:

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

(very important assumption for a Naive Bayes classifier!)

# Outline


1. Basics
2. Random Variables
3. Expectation-Variance
4. Joint Distributions
5. Covariance
6. RV Conditionals
7. Random Vectors
8. Multivariate Gaussian

# Random Vectors

Given $n$ RV's $X_1, ..., X_n$, 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 $X$.

# Random Vectors

Given $n$ RV's $X_1, ..., X_n$, 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:

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

# Covariance Matrices

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

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

# Covariance Matrices

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

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

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

# Covariance Matrices

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

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

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

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

# Covariance Matrices

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

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

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

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

**Properties:**

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

# Outline


1 Basics

2 Random Variables

3 Expectation-Variance

4 Joint Distributions

5 Covariance

6 RV Conditionals

7 Random Vectors

8 Multivariate Gaussian

# Multivariate Gaussian

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

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

# Multivariate Gaussian

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

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

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

# Multivariate Gaussian

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

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

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

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

Notice that if $\Sigma \in \mathbb{R}^{1\times 1}$, then $\Sigma = Var[X_1] = \sigma^2$, 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

# Some Nice Properties of MV Gaussians

- Marginals and conditionals of a joint Gaussian are Gaussian
- A $d$-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$ 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.

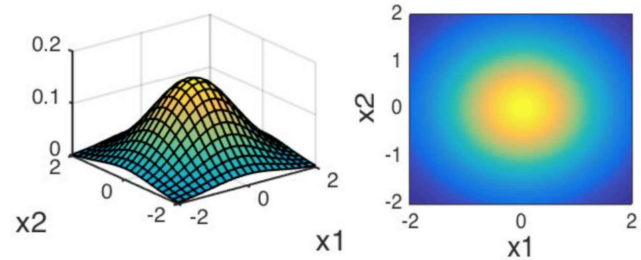
# Some Nice Properties of MV Gaussians

- Marginals and conditionals of a joint Gaussian are Gaussian
- A $d$-dimensional Gaussian $X \in \mathcal{N}(\mu, \Sigma = diag(\sigma_1^2, ..., \sigma_n^2))$ 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 ellipsoids with principal axes in the directions of the eigenvectors of covariance matrix $\Sigma$ (remember, $\Sigma$ is PSD, so all $n$ eigenvectors are non-negative). The axes' relative lengths depend on the eigenvalues of $\Sigma$.

# Visualizations of MV Gaussians

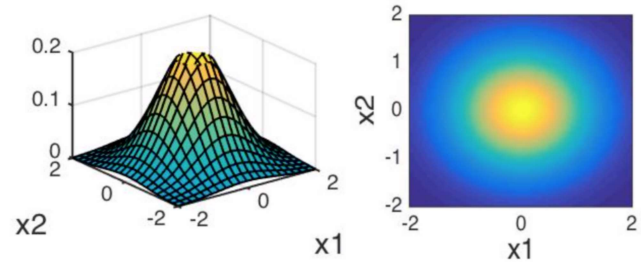
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



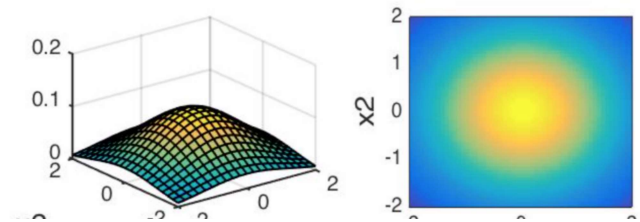

$$\Sigma = \begin{matrix} 0.7 & 0 \\ 0 & 0.7 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$




$$\Sigma = \begin{matrix} 1.5 & 0 \\ 0 & 1.5 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

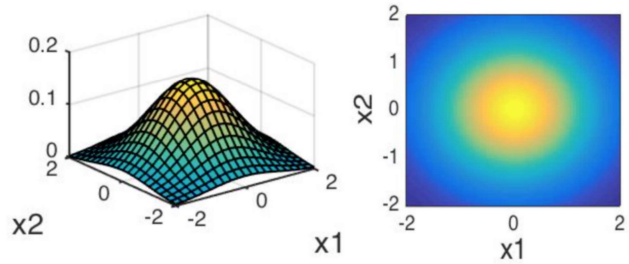



Effect of changing variance

# Visualizations of MV Gaussians

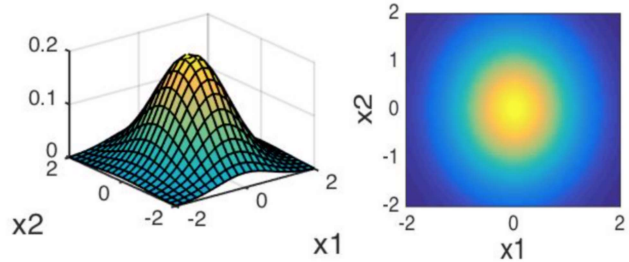
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



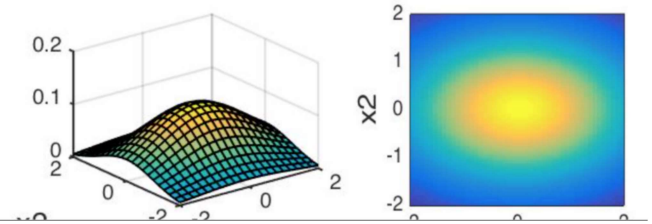

$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$




$$\Sigma = \begin{matrix} 2 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

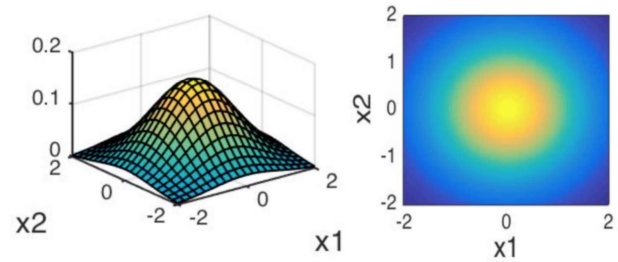



If $Var[X_1] \neq Var[X_2]$:

# Visualizations of MV Gaussians

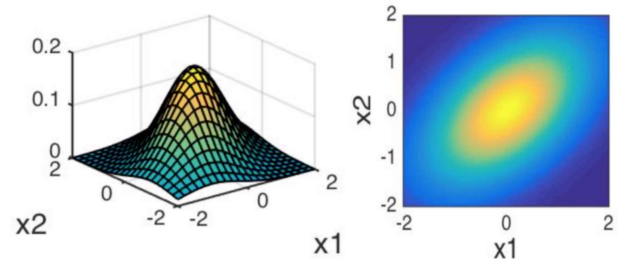
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



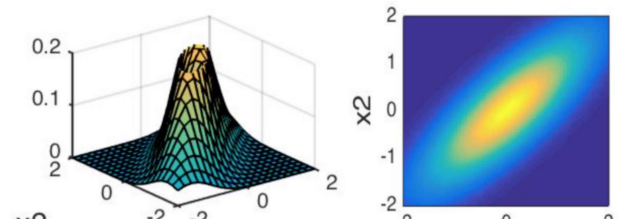

$$\Sigma = \begin{matrix} 1 & 0.5 \\ 0.5 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$




$$\Sigma = \begin{matrix} 1 & 0.8 \\ 0.8 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$

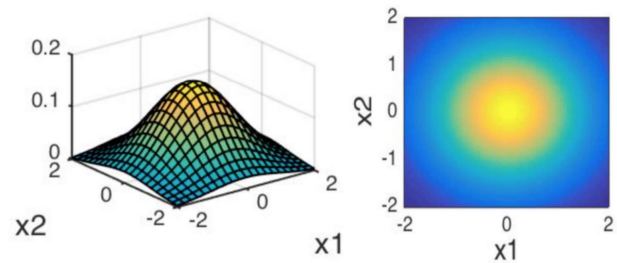



If $X_1$ and $X_2$ are positively correlated:

# Visualizations of MV Gaussians

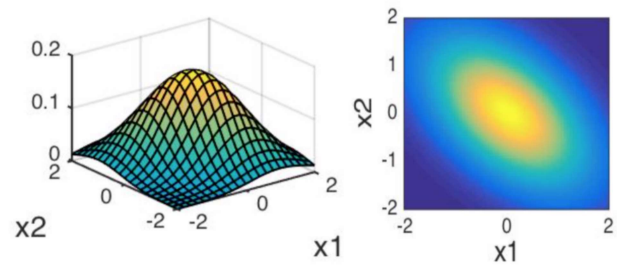
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$



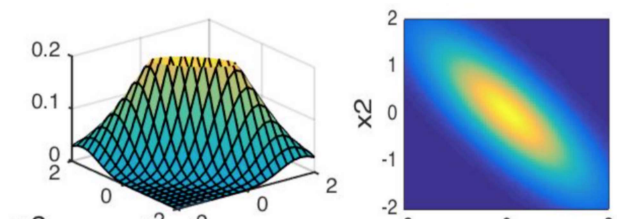

$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$




$$\Sigma = \begin{matrix} 1 & -0.8 \\ -0.8 & 1 \end{matrix}$$

$$\mu = [0 \;\; 0]^T$$




If $X_1$ and $X_2$ are negatively correlated:

# Thank you and good luck!

For further reference, consult the following CS229 handouts

- Probability Theory Review
- The MV Gaussian Distribution
- More on Gaussian Distribution

For a comprehensive treatment, see

- Sheldon Ross: **A First Course in Probability**

# Appendix: More on Total Expectation

Why is $\mathbb{E}[X|Y]$ a function of $Y$? Consider the 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 $Val(Y)$ to the real numbers.

An example: Consider RV $X$ such that

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

such that $\epsilon \sim \mathcal{N}(0,1)$ 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 $X, Y$:

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

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

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

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

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:

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