

# Mathematics for Machine Learning

## — Probability & Distributions

Sum Rule, Product Rule, Bayes' Theorem & Summary Statistics

Joseph Chuang-Chieh Lin

Department of Computer Science & Engineering,  
National Taiwan Ocean University

Fall 2025

## Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Arnold J. Insel, Lawrence E. Spence, Stephen H. Friedberg: Linear Algebra, 4th Edition. Prentice Hall. 2013.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra, 12th Edition. Wiley. 2019.*
- We could partially refer to the monograph:  
*Francesco Orabona: A Modern Introduction to Online Learning.*  
<https://arxiv.org/abs/1912.13213>

# Outline

- 1 Sum & Product Rule
- 2 Bayes' Theorem
- 3 Means & Covariances
- 4 Sums & Transformations of Random Variables
- 5 Statistical Independence

# Outline

- 1 Sum & Product Rule
- 2 Bayes' Theorem
- 3 Means & Covariances
- 4 Sums & Transformations of Random Variables
- 5 Statistical Independence

## Sum Rule (1/2)

- $\mathbf{x}, \mathbf{y}$ : random variables (vectors).
- $p(\mathbf{x}, \mathbf{y})$ : joint distribution of  $\mathbf{x}, \mathbf{y}$ .
- $p(\mathbf{y} \mid \mathbf{x})$ : conditional probability of  $\mathbf{y}$  given  $\mathbf{x}$ .

### Sum Rule

$$p(\mathbf{x}) = \begin{cases} \sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{x}, \mathbf{y}) & \text{if } \mathbf{y} \text{ is discrete} \\ \int_{\mathcal{Y}} p(\mathbf{x}, \mathbf{y}) d\mathbf{y} & \text{if } \mathbf{y} \text{ is continuous} \end{cases}$$

where  $\mathcal{Y}$  stands for the states of the target space of random variable  $Y$ .

- **Marginalization property.**

## Sum Rule (2/2)

For  $\mathbf{x} = [x_1, \dots, x_D]^\top$ , the marginal

$$p(x_i) = \int p(x_1, \dots, x_D) d\mathbf{x}_{-i},$$

where “ $-i$ ” means all except  $i$ .

# Product Rule

## Product Rule

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

# Outline

- 1 Sum & Product Rule
- 2 Bayes' Theorem**
- 3 Means & Covariances
- 4 Sums & Transformations of Random Variables
- 5 Statistical Independence



# Bayes' Theorem

## Bayes' Theorem

$$\underbrace{p(\mathbf{x} | \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y} | \mathbf{x})}^{\text{likelihood}} \overbrace{p(\mathbf{x})}^{\text{prior}}}{\underbrace{p(\mathbf{y})}_{\text{evidence}}}.$$

- Prior: subjective prior knowledge (before observing data).
- Likelihood  $p(\mathbf{y} | \mathbf{x})$ : the probability of  $\mathbf{y}$  if we were to know the latent variable  $\mathbf{x}$ .
  - We call it “the likelihood of  $\mathbf{x}$ ”.
- Posterior  $p(\mathbf{x} | \mathbf{y})$ : the quantity that we know about  $\mathbf{x}$  after having observed  $\mathbf{y}$ .

# Marginal Likelihood/Evidence

$$p(\mathbf{y}) := \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}[p(\mathbf{y} | \mathbf{x})]$$

$$p(\mathbf{y}) := \int_{\mathbf{x} \in \mathcal{X}} p(\mathbf{y} | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{\mathbf{x}}[p(\mathbf{y} | \mathbf{x})].$$

# Outline

- 1 Sum & Product Rule
- 2 Bayes' Theorem
- 3 Means & Covariances**
- 4 Sums & Transformations of Random Variables
- 5 Statistical Independence

# Expected Value

## Expected value

The expected value of a function  $g : \mathbb{R} \rightarrow \mathbb{R}$  of a random variable  $X \sim p(x)$  is

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx,$$

or

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x).$$

# Expected Value

## Expected value

The expected value of a function  $g : \mathbb{R} \rightarrow \mathbb{R}$  of a random variable  $X \sim p(x)$  is

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx,$$

or

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)p(x).$$

## Multivariate $X = [X_1, \dots, X_D]^\top$

$$\mathbb{E}_X[g(\mathbf{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D,$$

where  $\mathbb{E}_{X_d}$ : taking the expectation w.r.t. the  $x_d$ .

# Expected Value (contd.)

## Mean

For  $\mathbf{x} \in \mathbb{R}^D$ ,

$$\mathbb{E}_X[\mathbf{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D,$$

where

- $\mathbb{E}_{X_d}[x_d] = \int_{\mathcal{X}} x_d p(x_d) dx_d$  if  $X$  is continuous ;
- $\mathbb{E}_{X_d}[x_d] = \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) dx_d$  if  $X$  is discrete.

# Linearity of Expectation

Let  $f(\mathbf{x}) = ag(\mathbf{x}) + bh(\mathbf{x})$  for  $a, b \in \mathbb{R}$  and  $\mathbf{x} \in \mathbb{R}^D$ .

$$\begin{aligned}\mathbb{E}_X[f(\mathbf{x})] &= \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= \int [ag(\mathbf{x}) + bh(\mathbf{x})]p(\mathbf{x})d\mathbf{x} \\ &= a \int g(\mathbf{x})p(\mathbf{x})d\mathbf{x} + b \int h(\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= a\mathbb{E}_X[g(\mathbf{x})] + b\mathbb{E}_X[h(\mathbf{x})].\end{aligned}$$

# Linearity of Expectation (Discrete Case)

Let  $f(\mathbf{x}) = ag(\mathbf{x}) + bh(\mathbf{x})$  for  $a, b \in \mathbb{R}$  and  $\mathbf{x} \in \mathcal{X}$ .

$$\begin{aligned}\mathbb{E}_{\mathcal{X}}[f(\mathbf{x})] &= \sum_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})p(\mathbf{x}) \\ &= \sum_{\mathbf{x} \in \mathcal{X}} [ag(\mathbf{x}) + bh(\mathbf{x})]p(\mathbf{x}) \\ &= a \sum_{\mathbf{x} \in \mathcal{X}} g(\mathbf{x})p(\mathbf{x}) + b \sum_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x})p(\mathbf{x}) \\ &= a\mathbb{E}_{\mathcal{X}}[g(\mathbf{x})] + b\mathbb{E}_{\mathcal{X}}[h(\mathbf{x})].\end{aligned}$$



# Linearity of Expectation (Matrix)

## Exercise

Let  $\mathbf{A} \in \mathbb{R}^{n \times n}$  and  $\mathbf{x} \in \mathbb{R}^{n \times 1}$ . Show that

$$\mathbb{E}_{\mathbf{x}}[\mathbf{A}(\mathbf{x})] = \mathbf{A} \mathbb{E}_{\mathbf{x}}[\mathbf{x}].$$

# Covariance

The (univariate) **covariance** between two univariate random variables  $X, Y \in \mathbb{R}$  is

$$\text{Cov}_{X,Y}[x, y] := \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])].$$

Omit the subscript.

$$\text{Cov}[x, y] := \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

# Covariance

The (univariate) **covariance** between two univariate random variables  $X, Y \in \mathbb{R}$  is

$$\text{Cov}_{X,Y}[x, y] := \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])].$$

Omit the subscript.

$$\text{Cov}[x, y] := \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].$$

Note that

$$\text{Cov}[x, x] := \mathbb{E}[x^2] - (\mathbb{E}[x])^2$$

is the **variance** and denoted by  $\mathbb{V}_X[x]$  and  $\sqrt{\text{Cov}[x, x]}$  denoted by  $\sigma(x)$  is called the **standard deviation**.

# Covariance of Multivariate R.V.'s

## Covariance (Multivariate)

Consider random variables  $X$  and  $Y$  with states  $\mathbf{x} \in \mathbb{R}^D$  and  $\mathbf{y} \in \mathbb{R}^E$ . The covariance between  $X$  and  $Y$ :

$$\text{Cov}[\mathbf{x}, \mathbf{y}] =$$

# Covariance of Multivariate R.V.'s

## Covariance (Multivariate)

Consider random variables  $X$  and  $Y$  with states  $\mathbf{x} \in \mathbb{R}^D$  and  $\mathbf{y} \in \mathbb{R}^E$ . The covariance between  $X$  and  $Y$ :

$$\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}\mathbf{y}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]^\top$$

# Covariance of Multivariate R.V.'s

## Covariance (Multivariate)

Consider random variables  $X$  and  $Y$  with states  $\mathbf{x} \in \mathbb{R}^D$  and  $\mathbf{y} \in \mathbb{R}^E$ . The covariance between  $X$  and  $Y$ :

$$\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[\mathbf{x}\mathbf{y}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]^\top = \text{Cov}[\mathbf{y}, \mathbf{x}]^\top \in \mathbb{R}^{D \times E}.$$

# Variance (Multivariate)

## Variance (Multivariate)

The variance of a random variables  $X$  with states  $\mathbf{x} \in \mathbb{R}^D$  and mean  $\boldsymbol{\mu} \in \mathbb{R}^D$  is

$$\mathbb{V}_X[\mathbf{x}] = \text{Cov}_X[\mathbf{x}, \mathbf{x}] = \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]$$

# Variance (Multivariate)

## Variance (Multivariate)

The variance of a random variables  $X$  with states  $\mathbf{x} \in \mathbb{R}^D$  and mean  $\boldsymbol{\mu} \in \mathbb{R}^D$  is

$$\mathbb{V}_X[\mathbf{x}] = \text{Cov}_X[\mathbf{x}, \mathbf{x}] = \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top$$



# Variance (Multivariate)

## Variance (Multivariate)

The variance of a random variables  $X$  with states  $\mathbf{x} \in \mathbb{R}^D$  and mean  $\boldsymbol{\mu} \in \mathbb{R}^D$  is

$$\begin{aligned} \mathbb{V}_X[\mathbf{x}] &= \text{Cov}_X[\mathbf{x}, \mathbf{x}] = \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \\ &= \begin{bmatrix} \text{Cov}[x_1, x_1] & \text{Cov}[x_1, x_2] & \cdots & \text{Cov}[x_1, x_D] \\ \text{Cov}[x_2, x_1] & \text{Cov}[x_2, x_2] & \cdots & \text{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[x_D, x_1] & \text{Cov}[x_D, x_2] & \cdots & \text{Cov}[x_D, x_D] \end{bmatrix}. \end{aligned}$$

# Variance (Multivariate)

## Variance (Multivariate)

The variance of a random variables  $X$  with states  $\mathbf{x} \in \mathbb{R}^D$  and mean  $\boldsymbol{\mu} \in \mathbb{R}^D$  is

$$\begin{aligned}\mathbb{V}_X[\mathbf{x}] &= \text{Cov}_X[\mathbf{x}, \mathbf{x}] = \mathbb{E}_X[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \\ &= \begin{bmatrix} \text{Cov}[x_1, x_1] & \text{Cov}[x_1, x_2] & \cdots & \text{Cov}[x_1, x_D] \\ \text{Cov}[x_2, x_1] & \text{Cov}[x_2, x_2] & \cdots & \text{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[x_D, x_1] & \text{Cov}[x_D, x_2] & \cdots & \text{Cov}[x_D, x_D] \end{bmatrix}.\end{aligned}$$

- The **covariance matrix** of the multivariate  $X$ .

# Correlation Coefficient

## Correlation

The correlation between two random variables  $X, Y$  is

$$\text{corr}[x, y] = \frac{\text{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1].$$

# Correlation Coefficient

## Correlation

The correlation between two random variables  $X, Y$  is

$$\text{corr}[x, y] = \frac{\text{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1].$$

- By the Cauchy–Schwarz inequality.

## Inner product of two random variables

**Assumption:**  $U, V$  are two real random variables,  $\mathbb{E}[U^2], \mathbb{E}[V^2] < \infty$ .

**Definition.**  $\langle U, V \rangle := \mathbb{E}[UV]$ .

- *Well-defined:*  $2|UV| \leq U^2 + V^2 \Rightarrow \mathbb{E}[UV] \leq |UV| < \infty$ .

## Inner product of two random variables

**Assumption:**  $U, V$  are two real random variables,  $\mathbb{E}[U^2], \mathbb{E}[V^2] < \infty$ .

**Definition.**  $\langle U, V \rangle := \mathbb{E}[UV]$ .

- *Well-defined:*  $2|UV| \leq U^2 + V^2 \Rightarrow \mathbb{E}[UV] \leq |UV| < \infty$ .
- *Linearity:*  $\langle aU + bW, V \rangle = \mathbb{E}[(aU + bW)V] = \mathbb{E}[aUV + bWV] = a\mathbb{E}[UV] + b\mathbb{E}[WV] = a\langle U, V \rangle + b\langle W, V \rangle$  (and similarly in the 2nd slot).

# Inner product of two random variables

**Assumption:**  $U, V$  are two real random variables,  $\mathbb{E}[U^2], \mathbb{E}[V^2] < \infty$ .

**Definition.**  $\langle U, V \rangle := \mathbb{E}[UV]$ .

- *Well-defined:*  $2|UV| \leq U^2 + V^2 \Rightarrow \mathbb{E}[UV] \leq |UV| < \infty$ .
- *Linearity:*  $\langle aU + bW, V \rangle = \mathbb{E}[(aU + bW)V] = \mathbb{E}[aUV + bWV] = a\mathbb{E}[UV] + b\mathbb{E}[WV] = a\langle U, V \rangle + b\langle W, V \rangle$  (and similarly in the 2nd slot).
- *Symmetry:*  $\langle U, V \rangle = \mathbb{E}[UV] = \mathbb{E}[VU] = \langle V, U \rangle$ .

## Inner product of two random variables

**Assumption:**  $U, V$  are two real random variables,  $\mathbb{E}[U^2], \mathbb{E}[V^2] < \infty$ .

**Definition.**  $\langle U, V \rangle := \mathbb{E}[UV]$ .

- *Well-defined:*  $2|UV| \leq U^2 + V^2 \Rightarrow \mathbb{E}[UV] \leq |UV| < \infty$ .
- *Linearity:*  $\langle aU + bW, V \rangle = \mathbb{E}[(aU + bW)V] = \mathbb{E}[aUV + bWV] = a\mathbb{E}[UV] + b\mathbb{E}[WV] = a\langle U, V \rangle + b\langle W, V \rangle$  (and similarly in the 2nd slot).
- *Symmetry:*  $\langle U, V \rangle = \mathbb{E}[UV] = \mathbb{E}[VU] = \langle V, U \rangle$ .
- *Positive-definite:*  $\langle U, U \rangle = \mathbb{E}[U^2] \geq 0$ ; if  $\mathbb{E}[U^2] = 0$ , then  $U = 0$  a.s. (else  $\exists \varepsilon > 0 : \Pr[|U| \geq \varepsilon] > 0 \Rightarrow \mathbb{E}[U^2] \geq \varepsilon^2 \Pr[|U| \geq \varepsilon] > 0$ ).



## Inner product of two random variables

**Assumption:**  $U, V$  are two real random variables,  $\mathbb{E}[U^2], \mathbb{E}[V^2] < \infty$ .

**Definition.**  $\langle U, V \rangle := \mathbb{E}[UV]$ .

- *Well-defined:*  $2|UV| \leq U^2 + V^2 \Rightarrow \mathbb{E}[UV] \leq |UV| < \infty$ .
- *Linearity:*  $\langle aU + bW, V \rangle = \mathbb{E}[(aU + bW)V] = \mathbb{E}[aUV + bWV] = a\mathbb{E}[UV] + b\mathbb{E}[WV] = a\langle U, V \rangle + b\langle W, V \rangle$  (and similarly in the 2nd slot).
- *Symmetry:*  $\langle U, V \rangle = \mathbb{E}[UV] = \mathbb{E}[VU] = \langle V, U \rangle$ .
- *Positive-definite:*  $\langle U, U \rangle = \mathbb{E}[U^2] \geq 0$ ; if  $\mathbb{E}[U^2] = 0$ , then  $U = 0$  a.s. (else  $\exists \varepsilon > 0 : \Pr[|U| \geq \varepsilon] > 0 \Rightarrow \mathbb{E}[U^2] \geq \varepsilon^2 \Pr[|U| \geq \varepsilon] > 0$ ).

Hence  $\mathbb{E}[UV]$  as an inner product is valid; thus, by Cauchy–Schwarz

$$|\mathbb{E}[UV]| = |\langle U, V \rangle| \leq \sqrt{\mathbb{E}[U^2]} \sqrt{\mathbb{E}[V^2]}.$$

## Correlation is in $[-1, 1]$ by Cauchy–Schwarz

- Let  $X, Y$  satisfy  $0 < \sigma_X^2 = \mathbb{V}(X) < \infty$  and  $0 < \sigma_Y^2 = \mathbb{V}(Y) < \infty$ .
- Set  $\mu_X = \mathbb{E}[X]$ ,  $\mu_Y = \mathbb{E}[Y]$ ,  $U = X - \mu_X$  and  $V = Y - \mu_Y$ .

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\mathbb{E}[(X - \mu_X)^2]} \sqrt{\mathbb{E}[(Y - \mu_Y)^2]}} = \frac{\langle U, V \rangle}{\|U\|_2 \|V\|_2}.$$

- By the Cauchy–Schwarz inequality,

$$-1 \leq \rho_{XY} \leq 1.$$

**Note:** If  $\sigma_X = 0$  or  $\sigma_Y = 0$ , correlation is undefined.

# Empirical Means & Covariances

In machine learning, we need to learn from empirical observations of data.

## Empirical Mean & Covariance

The **empirical mean** vector: arithmetic average of the observations for each variable:

$$\bar{\mathbf{x}} := \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i,$$

for  $\mathbf{x}_i \in \mathbb{R}^D$ . The **empirical covariance** matrix is a  $D \times D$  matrix

$$\Sigma := \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

# Empirical Means & Covariances

In machine learning, we need to learn from empirical observations of data.

## Empirical Mean & Covariance

The **empirical mean** vector: arithmetic average of the observations for each variable:

$$\bar{\mathbf{x}} := \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i,$$

for  $\mathbf{x}_i \in \mathbb{R}^D$ . The **empirical covariance** matrix is a  $D \times D$  matrix

$$\Sigma := \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

- $\Sigma$  is symmetric, positive semidefinite.

## Computing the Empirical Variance (1D Example)

Approaches:

- 1 By definition  $\Rightarrow \mathbb{V}_X[x] := \mathbb{E}_X[(x - \mu)^2]$ .
  - Two-pass; numerically stable.
- 2  $\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$ .
  - One-pass; more efficient but numerically unstable.
- 3 Averaging pairwise differences between all pairs of observations.

$$\frac{1}{N^2} \sum_{i,j=1}^N (x_i - x_j)^2 = 2 \left[ \frac{1}{N} \sum_{i=1}^N x_i^2 - \left( \frac{1}{N} \sum_{i=1}^N x_i \right)^2 \right].$$

# Computing the Empirical Variance (1D Example)

Approaches:

- 1 By definition  $\Rightarrow \mathbb{V}_X[x] := \mathbb{E}_X[(x - \mu)^2]$ .
  - Two-pass; numerically stable.
- 2  $\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$ .
  - One-pass; more efficient but numerically unstable.
- 3 Averaging pairwise differences between all pairs of observations.

$$\frac{1}{N^2} \sum_{i,j=1}^N (x_i - x_j)^2 = 2 \left[ \frac{1}{N} \sum_{i=1}^N x_i^2 - \left( \frac{1}{N} \sum_{i=1}^N x_i \right)^2 \right].$$

- Twice of the 2nd approach (left-hand side:  $O(N^2)$ ).
- Interesting perspective to compute the left-hand side target.

# Welford's Online Algorithm [1962]

- **Input:** Stream of observations  $x_1, x_2, \dots$
  - **Output:** (population) variances  $\sigma^2$ , and unbiased variance  $s^2$ .
  - ① **Initialization:**  $n \leftarrow 0, \mu \leftarrow 0, M_2 \leftarrow 0$ ;
  - ② for each  $x_i$  in stream,  $i = 1, 2, \dots$ 
    - ①  $n \leftarrow n + 1$ ;
    - ②  $\delta \leftarrow x - \mu, \mu \leftarrow \mu + \delta/n$ ; /\* empirical mean update \*/
    - ③  $\delta_2 \leftarrow x - \mu, M_2 \leftarrow M_2 + \delta \cdot \delta_2$  /\*  $M_2 = \sum_{i=1}^n (x_i - \mu_n)^2$  \*/
  - ③ **population variance:**  $\sigma^2 \leftarrow M_2/n$  (valid for  $n \geq 1$ );
  - ④ **unbiased variance:**  $s^2 \leftarrow M_2/(n-1)$  (valid for  $n \geq 2$ )
- Each increment  $\delta$  and  $\delta_2$  are on the scale of the deviation or variance, not on the scale of  $x$  and  $x^2$ .

## Accuracy of Welford's Online Algorithm (Mean)

**Setup.** For a stream  $x_1, x_2, \dots$ , maintain  $\mu_n :=$  mean after  $n$ ,  $M_2^{(n)} = \sum_{i=1}^n (x_i - \mu_n)^2$ . Given  $(\mu_{n-1}, M_2^{(n-1)})$  and new  $x_n$ ,

$$\delta = x_n - \mu_{n-1}, \quad \mu_n = \mu_{n-1} + \frac{\delta}{n}, \quad \delta_2 = x_n - \mu_n, \quad M_2^{(n)} = M_2^{(n-1)} + \delta \delta_2.$$

**Claim (Mean exactness).**  $\mu_n = \frac{1}{n} \sum_{i=1}^n x_i$ .

### Proof

Base  $n = 1$ :  $\mu_1 = x_1$ . For the step,

$$\mu_n = \mu_{n-1} + \frac{x_n - \mu_{n-1}}{n} = \frac{(n-1)\mu_{n-1} + x_n}{n} = \frac{\sum_{i=1}^{n-1} x_i + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i.$$



## Accuracy of Welford's Online Algorithm (2nd Moment)

**Claim.**  $M_2^{(n)} = \sum_{i=1}^n (x_i - \mu_n)^2$  is preserved by  $M_2^{(n)} = M_2^{(n-1)} + \delta \delta_2$  with  $\delta = x_n - \mu_{n-1}$  and  $\delta_2 = x_n - \mu_n$ .

## Proof

Assume  $M_2^{(n-1)} = \sum_{i=1}^{n-1} (x_i - \mu_{n-1})^2$ . Then

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu_n)^2 &= \sum_{i=1}^{n-1} [(x_i - \mu_{n-1}) + (\mu_{n-1} - \mu_n)]^2 + (x_n - \mu_n)^2 \\ &= \underbrace{\sum_{i=1}^{n-1} (x_i - \mu_{n-1})^2}_{M_2^{(n-1)}} + (n-1)(\mu_{n-1} - \mu_n)^2 + (x_n - \mu_n)^2, \end{aligned}$$

## Accuracy of Welford's Online Algorithm (2nd Moment) Contd.

Since  $\sum_{i=1}^{n-1} (x_i - \mu_{n-1}) = 0$ . With  $\mu_n - \mu_{n-1} = \delta/n$  and  $\delta_2 = x_n - \mu_n = x_n - (\mu_{n-1} + \delta/n) = \delta(1 - 1/n)$ ,

$$\begin{aligned}(n-1)(\mu_{n-1} - \mu_n)^2 + (x_n - \mu_n)^2 &= \frac{(n-1)\delta^2}{n^2} + \frac{(n-1)^2\delta^2}{n^2} \\ &= \frac{(n-1)\delta^2}{n} \\ &= \delta \delta_2.\end{aligned}$$

Therefore  $\sum_{i=1}^n (x_i - \mu_n)^2 = M_2^{(n-1)} + \delta \delta_2 = M_2^{(n)}$ .

**Consequences.** Population variance:  $\sigma^2 = M_2^{(n)}/n$  (for  $n \geq 1$ );  
unbiased sample variance:  $s^2 = M_2^{(n)}/(n-1)$  (for  $n \geq 2$ ).

# Outline

- 1 Sum & Product Rule
- 2 Bayes' Theorem
- 3 Means & Covariances
- 4 Sums & Transformations of Random Variables**
- 5 Statistical Independence

# Basic Rules

## Simple Rules & Exercise

Consider two random variables  $X, Y$  with states  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ . Then,

$$\mathbb{E}[\mathbf{x} \pm \mathbf{y}] = \mathbb{E}[\mathbf{x}] \pm \mathbb{E}[\mathbf{y}]$$

$$\mathbb{V}[\mathbf{x} \pm \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] \pm \text{Cov}[\mathbf{x}, \mathbf{y}] \pm \text{Cov}[\mathbf{y}, \mathbf{x}] \quad (\text{Exercise}).$$

- **Note:** For a constant vector  $\mathbf{b} \in \mathbb{R}^D$ ,  $\mathbb{V}(\mathbf{x} \pm \mathbf{b}) = \mathbb{V}[\mathbf{x}]$  because  $\mathbb{V}[\mathbf{b}] = \mathbb{E}[\mathbf{b}\mathbf{b}^\top] - \mathbb{E}[\mathbf{b}]\mathbb{E}[\mathbf{b}]^\top = \mathbf{b}\mathbf{b}^\top - \mathbf{b}\mathbf{b}^\top = \mathbf{0}$  and

$$\text{Cov}(\mathbf{x}, \mathbf{b})$$

# Basic Rules

## Simple Rules & Exercise

Consider two random variables  $X, Y$  with states  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ . Then,

$$\mathbb{E}[\mathbf{x} \pm \mathbf{y}] = \mathbb{E}[\mathbf{x}] \pm \mathbb{E}[\mathbf{y}]$$

$$\mathbb{V}[\mathbf{x} \pm \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] \pm \text{Cov}[\mathbf{x}, \mathbf{y}] \pm \text{Cov}[\mathbf{y}, \mathbf{x}] \quad (\text{Exercise}).$$

- **Note:** For a constant vector  $\mathbf{b} \in \mathbb{R}^D$ ,  $\mathbb{V}(\mathbf{x} \pm \mathbf{b}) = \mathbb{V}[\mathbf{x}]$  because  $\mathbb{V}[\mathbf{b}] = \mathbb{E}[\mathbf{b}\mathbf{b}^\top] - \mathbb{E}[\mathbf{b}]\mathbb{E}[\mathbf{b}]^\top = \mathbf{b}\mathbf{b}^\top - \mathbf{b}\mathbf{b}^\top = \mathbf{0}$  and

$$\text{Cov}(\mathbf{x}, \mathbf{b}) = \mathbb{E}[\mathbf{x}\mathbf{b}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{b}]^\top$$

# Basic Rules

## Simple Rules & Exercise

Consider two random variables  $X, Y$  with states  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ . Then,

$$\mathbb{E}[\mathbf{x} \pm \mathbf{y}] = \mathbb{E}[\mathbf{x}] \pm \mathbb{E}[\mathbf{y}]$$

$$\mathbb{V}[\mathbf{x} \pm \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] \pm \text{Cov}[\mathbf{x}, \mathbf{y}] \pm \text{Cov}[\mathbf{y}, \mathbf{x}] \quad (\text{Exercise}).$$

- **Note:** For a constant vector  $\mathbf{b} \in \mathbb{R}^D$ ,  $\mathbb{V}(\mathbf{x} \pm \mathbf{b}) = \mathbb{V}[\mathbf{x}]$  because  $\mathbb{V}[\mathbf{b}] = \mathbb{E}[\mathbf{b}\mathbf{b}^\top] - \mathbb{E}[\mathbf{b}]\mathbb{E}[\mathbf{b}]^\top = \mathbf{b}\mathbf{b}^\top - \mathbf{b}\mathbf{b}^\top = \mathbf{0}$  and

$$\text{Cov}(\mathbf{x}, \mathbf{b}) = \mathbb{E}[\mathbf{x}\mathbf{b}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{b}]^\top = \mathbb{E}[\mathbf{x}]\mathbf{b}^\top - \mathbb{E}[\mathbf{x}]\mathbf{b}^\top = \mathbf{0}.$$

# Affine Transformation of r.v.'s (1/2)

Consider  $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$  and let  $\mathbf{\Sigma} := \mathbb{V}_X[\mathbf{x}]$ .

$$\mathbb{E}_Y[\mathbf{y}] = \mathbb{E}_X[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\mathbb{E}_X[\mathbf{x}] + \mathbf{b}$$

$$\mathbb{V}_Y[\mathbf{y}] = \mathbb{V}_X[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbb{V}_X[\mathbf{A}\mathbf{x}] = \mathbf{A}\mathbb{V}_X[\mathbf{x}]\mathbf{A}^\top = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^\top.$$

$$\mathbb{V}_X[\mathbf{Ax}] = \mathbb{E}_X[(\mathbf{Ax})(\mathbf{Ax})^\top] - \mathbb{E}_X[\mathbf{Ax}](\mathbb{E}_X[\mathbf{Ax}])^\top$$



$$\begin{aligned}\mathbb{V}_X[\mathbf{Ax}] &= \mathbb{E}_X[(\mathbf{Ax})(\mathbf{Ax})^\top] - \mathbb{E}_X[\mathbf{Ax}](\mathbb{E}_X[\mathbf{Ax}])^\top \\ &= \mathbb{E}_X[\mathbf{Axx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top\end{aligned}$$

$$\begin{aligned}\mathbb{V}_X[\mathbf{A}\mathbf{x}] &= \mathbb{E}_X[(\mathbf{A}\mathbf{x})(\mathbf{A}\mathbf{x})^\top] - \mathbb{E}_X[\mathbf{A}\mathbf{x}](\mathbb{E}_X[\mathbf{A}\mathbf{x}])^\top \\ &= \mathbb{E}_X[\mathbf{A}\mathbf{x}\mathbf{x}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\ &= \mathbf{A}\mathbb{E}_X[\mathbf{x}\mathbf{x}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top\end{aligned}$$

$$\begin{aligned}\mathbb{V}_X[\mathbf{Ax}] &= \mathbb{E}_X[(\mathbf{Ax})(\mathbf{Ax})^\top] - \mathbb{E}_X[\mathbf{Ax}](\mathbb{E}_X[\mathbf{Ax}])^\top \\&= \mathbb{E}_X[\mathbf{Axx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbb{E}_X[\mathbf{Axx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top\end{aligned}$$

$$\begin{aligned}\mathbb{V}_X[\mathbf{Ax}] &= \mathbb{E}_X[(\mathbf{Ax})(\mathbf{Ax})^\top] - \mathbb{E}_X[\mathbf{Ax}](\mathbb{E}_X[\mathbf{Ax}])^\top \\&= \mathbb{E}_X[\mathbf{Axx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbb{E}_X[\mathbf{Axx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top\end{aligned}$$

$$\begin{aligned}\mathbb{V}_X[\mathbf{Ax}] &= \mathbb{E}_X[(\mathbf{Ax})(\mathbf{Ax})^\top] - \mathbb{E}_X[\mathbf{Ax}](\mathbb{E}_X[\mathbf{Ax}])^\top \\&= \mathbb{E}_X[\mathbf{Axx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbb{E}_X[\mathbf{Axx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top] \mathbf{A}^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top\end{aligned}$$

$$\begin{aligned}\mathbb{V}_X[\mathbf{Ax}] &= \mathbb{E}_X[(\mathbf{Ax})(\mathbf{Ax})^\top] - \mathbb{E}_X[\mathbf{Ax}](\mathbb{E}_X[\mathbf{Ax}])^\top \\&= \mathbb{E}_X[\mathbf{Axx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top \mathbf{A}^\top] - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbb{E}_X[\mathbf{Axx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}(\mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top])^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{E}_X[\mathbf{xx}^\top] \mathbf{A}^\top - \mathbf{A}\mathbb{E}_X[\mathbf{x}]\mathbb{E}_X[\mathbf{x}]^\top \mathbf{A}^\top \\&= \mathbf{A}\mathbb{V}_X[\mathbf{x}] \mathbf{A}^\top.\end{aligned}$$

## Affine Transformation of r.v.'s (2/2)

Furthermore, let  $\boldsymbol{\mu} := \mathbb{E}_X[\mathbf{x}]$  and  $\boldsymbol{\Sigma} := \mathbb{V}_X[\mathbf{x}]$ .

$$\begin{aligned}\text{Cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}[\mathbf{x}(\mathbf{Ax} + \mathbf{b})^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{Ax} + \mathbf{b}]^\top \\ &= \boldsymbol{\mu}\mathbf{b}^\top + \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{A}^\top - \boldsymbol{\mu}\mathbf{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\mathbf{A}^\top \\ &= (\mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top)\mathbf{A}^\top \\ &= \boldsymbol{\Sigma}\mathbf{A}^\top.\end{aligned}$$

# Outline

- 1 Sum & Product Rule
- 2 Bayes' Theorem
- 3 Means & Covariances
- 4 Sums & Transformations of Random Variables
- 5 Statistical Independence**



## (Statistically) Independent

Two random variables  $X, Y$  are statistically independent if and only if

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}).$$

If  $X, Y$  are independent, then

- $p(\mathbf{y} \mid \mathbf{x}) = p(\mathbf{y})$ .
- $p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x})$ .
- $\mathbb{V}_{X,Y}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_X[\mathbf{x}] + \mathbb{V}_Y[\mathbf{y}]$ .
- $\text{Cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ .

## Remark

Note that  $\text{Cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$  does NOT necessarily imply that  $X$  and  $Y$  are independent.

## Remark

Note that  $\text{Cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$  does NOT necessarily imply that  $X$  and  $Y$  are independent.

- Consider a random variable  $X$  with  $\mathbb{E}_X[x] = 0$  and also  $\mathbb{E}_X[x^3] = 0$ .

## Remark

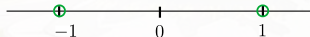
Note that  $\text{Cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$  does NOT necessarily imply that  $X$  and  $Y$  are independent.

- Consider a random variable  $X$  with  $\mathbb{E}_X[x] = 0$  and also  $\mathbb{E}_X[x^3] = 0$ .
- Let  $y = x^2$ . Hence,  $Y$  is **dependent on  $X$** .

## Remark

Note that  $\text{Cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$  does NOT necessarily imply that  $X$  and  $Y$  are independent.

- Consider a random variable  $X$  with  $\mathbb{E}_X[x] = 0$  and also  $\mathbb{E}_X[x^3] = 0$ .
- Let  $y = x^2$ . Hence,  $Y$  is **dependent on  $X$** .
- $\text{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x^3] = 0$ .



## Conditional Independence

Two random variables  $X, Y$  are conditionally independent given  $Z$  if and only if

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z}).$$

for all  $\mathbf{z} \in \mathcal{Z}$ .

By the product rule, we can have

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{y}, \mathbf{z})p(\mathbf{y} \mid \mathbf{z}).$$

Thus,

$$p(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}).$$

## Heads Up

If  $X, Y$  are independent, then  $\mathbb{V}_{X,Y}[\mathbf{x} + \mathbf{y}] = \mathbb{V}_X[\mathbf{x}] + \mathbb{V}_Y[\mathbf{y}]$ .

$$\therefore \text{Cov}_{X,Y}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$$

# Discussions