Lecture 2: Foundation

Deep Generative Models

Sajjad Amini

Department of Electrical Engineering
Sharif University of Technology

Contents

Section 1

Color Codes

Color Coded Blocks

Definition Block

Result Block

Note Block

Example Block

Remember Block

Section 2

Notation

Notation

Scalars, Vctors and Matrices

Type	Non-random	Random
Scalar	x	X
Vector	$\boldsymbol{x}$	$\mathbb{X}$
Matrix	$\boldsymbol{X}$	$\mathbb{X}$
i-th element of a vector	x_i or $[\boldsymbol{x}]_i$	X_i or $[\mathbb{X}]_i$
(i,j)-th element of a matrix	x_{ij} or $[\boldsymbol{X}]_{ij}$	X_{ij} or $[\mathbb{X}]_{ij}$
i-th row of a matrix	$\boldsymbol{x}_{i:}$ or $[\boldsymbol{X}]_{i:}$	$\mathbb{X}_{i:}$ or $[\mathbb{X}]_{i:}$
j-th column of a matrix	$\boldsymbol{x}_{:j}$ or $[\boldsymbol{X}]_{:j}$	$\mathbb{X}_{:j}$ or $[\mathbb{X}]_{:j}$

*The element index appears at the end of subscript

i-th element of vector $\boldsymbol{x}_k$: $x_{k,i}$ or $[\boldsymbol{x}_k]_i$

Notation

Operators

- **Element-wise product:** Assume $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x} \in \mathbb{R}^D$, then:

$$\boldsymbol{z} = \boldsymbol{x} \odot \boldsymbol{y} \Leftrightarrow z_i = x_i \times y_i, i = 1, \ldots, D$$

- **Vectorization:** Assume $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, then:

$$\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X}) \Leftrightarrow \boldsymbol{x} = \begin{bmatrix} [\boldsymbol{X}]_{:1} \\ [\boldsymbol{X}]_{:2} \\ \ldots \\ [\boldsymbol{X}]_{:n} \end{bmatrix}$$

- **Trace:** The trace of a square matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(\boldsymbol{X})$, is the sum of diagonal elements in the matrix as:

$$\mathrm{tr}(\boldsymbol{X}) \triangleq \sum_{i=1}^{n} x_{ii}$$

Notation

Operators

- **Norm:** The ℓ_p norm for vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as:

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, p \geq 1 \Rightarrow \begin{cases} \ell_1 : \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \\ \ell_2 : \|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \\ \ell_\infty : \|\boldsymbol{x}\|_\infty = \max_i |x_i| \end{cases}$$

- **Transpose:** The transpose of matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, denoted by $\boldsymbol{X}^T$, is:

$$[\boldsymbol{X}^T]_{ji} = [\boldsymbol{X}]_{ij}, \begin{cases} i = 1, \ldots, n \\ j = 1, \ldots, m \end{cases}$$

Notation

Operators

- **Diag:**

$$\boldsymbol{x} \in \mathbb{R}^n \Rightarrow \boldsymbol{X} = \mathrm{diag}(\boldsymbol{x}) = \begin{bmatrix} [\boldsymbol{x}]_1 & 0 & \dots & 0 & 0 \\ 0 & [\boldsymbol{x}]_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & [\boldsymbol{x}]_{n-1} & 0 \\ 0 & 0 & \dots & 0 & [\boldsymbol{x}]_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$\boldsymbol{X} \in \mathbb{R}^{n \times n} \Rightarrow \boldsymbol{x} = \mathrm{diag}(\boldsymbol{X}) = \begin{bmatrix} [\boldsymbol{X}]_{11} \\ [\boldsymbol{X}]_{22} \\ \vdots \\ [\boldsymbol{X}]_{(n-1)(n-1)} \\ [\boldsymbol{X}]_{nn} \end{bmatrix}$$

Notation

Matrix Calculus

- **Gradient vector:** The gradient vector for $f : \mathbb{R}^n \to \mathbb{R}$ at point $\boldsymbol{x}$ is:

$$\frac{\partial f}{\partial \boldsymbol{x}} = \nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

- **Jacobian matrix:** The Jacobian matrix for $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ at point $\boldsymbol{x}$ is:

$$\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}^T} \triangleq \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla f_1(\boldsymbol{x})^T \\ \vdots \\ \nabla f_m(\boldsymbol{x})^T \end{bmatrix}$$

Notation

Definitions

- **LHS and RHS:** Left Hand Side (LHS) and Right Hand Side (RHS) refer to:

$$\overbrace{\boldsymbol{z}}^{\text{LHS}} = \overbrace{\boldsymbol{x} \odot \boldsymbol{y}}^{\text{RHS}}$$

Section 3

Probability and Statistics

Subsection 1

Probability Mass/Density Function

Discrete Random Variable

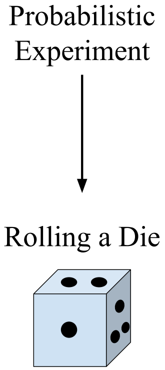


Figure: Probabilistic experiment: an experiment where the result is NOT certain a priori

Discrete Random Variable

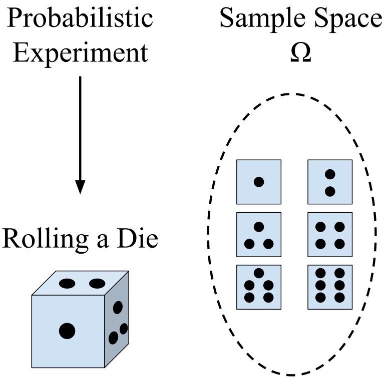


Figure: Sample space Ω: set of all possible outcomes

Discrete Random Variable

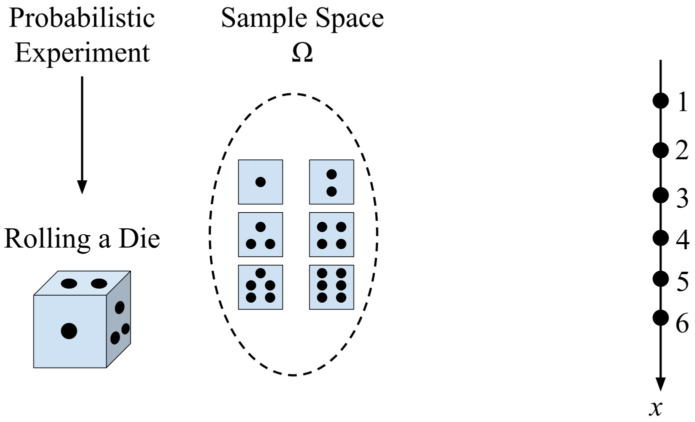


Figure: Numeric numbers x

Discrete Random Variable

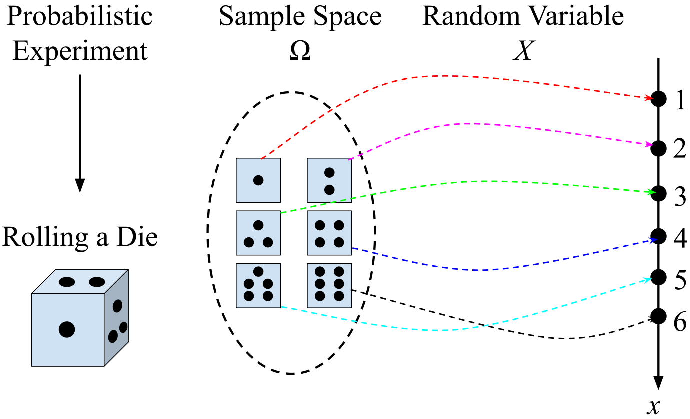


Figure: Random variable X: a function which maps every sample in Ω to a numeric number x

Event and Probability Mass Function

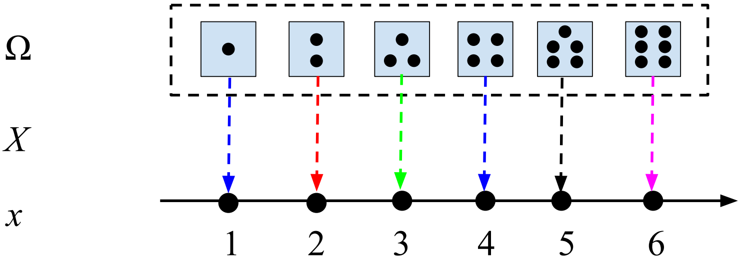
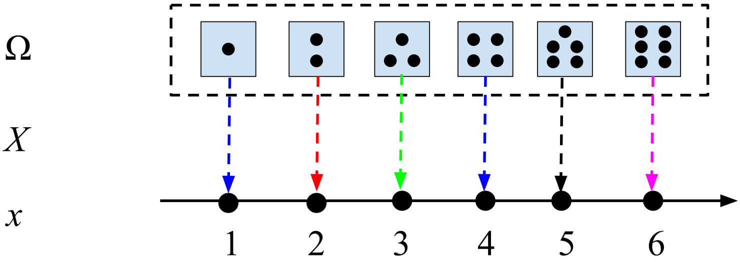


Figure: Sample space Ω, random variable X and numeric number x

Event and Probability Mass Function



$$\text{Event} \longrightarrow \text{“}X = x\text{”} \xrightarrow{\text{Equivalent}} \{w \in \Omega : X(w) = x\}$$

Figure: Event definition

Event and Probability Mass Function

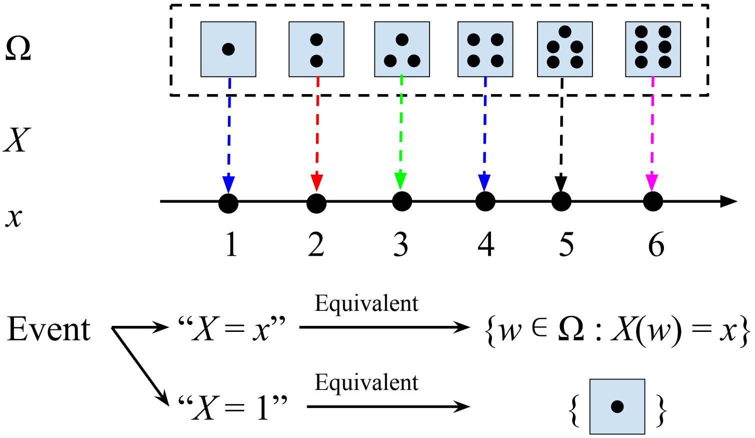


Figure: A typical event

Event and Probability Mass Function

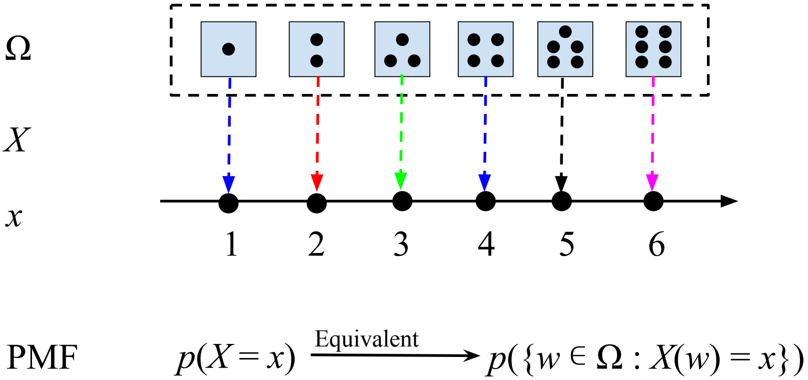


Figure: Probability mass function $p(X = x)$

Probability Mass Function

Properties

PMF assigns a mass to each event corresponding to the event probability, so it must satisfy the following properties:

$$p(X = x) \geq 0$$
$$\sum_x p(X = x) = 1$$

Sample Probability Mass Function

Bernoulli

Assume $x \in \{0, 1\}$, then random variable X is Bernoulli, dented by:

$$X \sim \mathrm{Ber}(\theta)$$

or

$$p_\theta(X) = \mathrm{Ber}(X|\theta)$$

And we have:

$$p_\theta(X = x) = \begin{cases} \theta & x = 1 \\ 1 - \theta & x = 0 \end{cases}$$

Note that θ must satisfy $0 \leq \theta \leq 1$.

Sample Probability Mass Function

Categorical

Assume $x \in \{1, 2, \dots, L\}$, then random variable X is Categorical, dented by:

$$X \sim \mathrm{Cat}(\boldsymbol{\theta})$$

or

$$p_\theta(X) = \mathrm{Cat}(X|\boldsymbol{\theta})$$

And we have:

$$p_\theta(X = l) = \theta_l$$

Note that θ must satisfy $0 \leq \theta \leq 1$ and $\sum_l \theta_l = 1$.

Extension to Random Vector

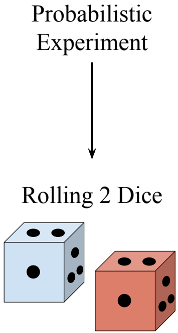


Figure: Rolling two dice experiment

Extension to Random Vector

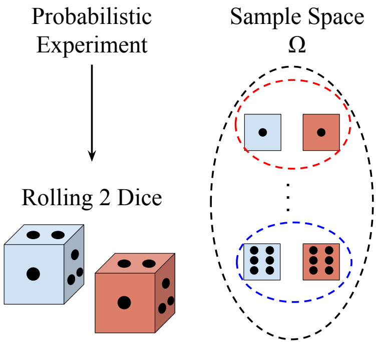


Figure: Sample space Ω

Extension to Random Vector

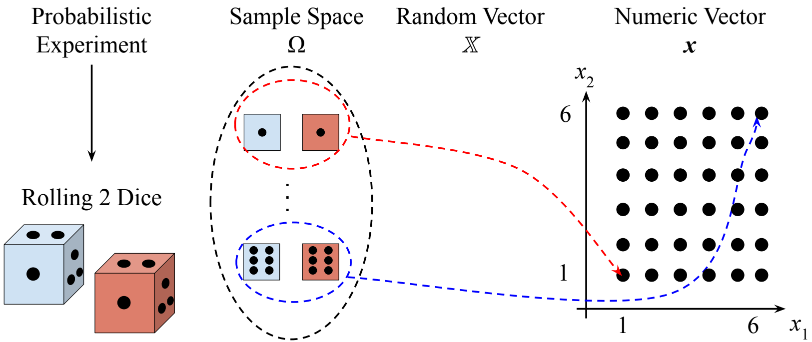


Figure: Random vector $\mathbb{X}$ and numeric vector $\boldsymbol{x}$

Probability Mass Function

Properties

PMF properties for a random variable can be easily extended to random vectors. Assume we have:

$$\mathbb{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

then we have:

$$p(\mathbb{X} = \boldsymbol{x}) = p(X_1 = x_1, X_2 = x_2) \geq 0$$
$$\sum_{\boldsymbol{x}} p(\mathbb{X} = \boldsymbol{x}) = \sum_{x_1} \sum_{x_2} p(X_1 = x_1, X_2 = x_2) = 1$$

To have PMF over only X_1 random variable, we can use *Marginalization* as:

$$p(X_1 = x_1) = \sum_{x_2} p(X_1 = x_1, X_2 = x_2)$$

Conditional Distribution

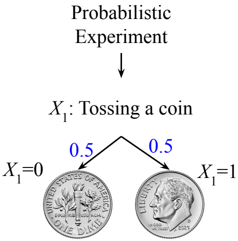


Figure: First random variable: Tossing a coin

Conditional Distribution

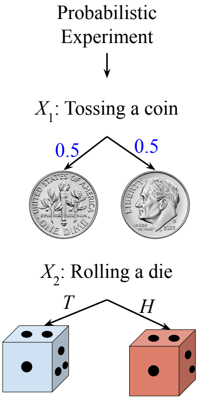


Figure: Second random variable: Rolling a die based on coin experiment

Conditional Distribution

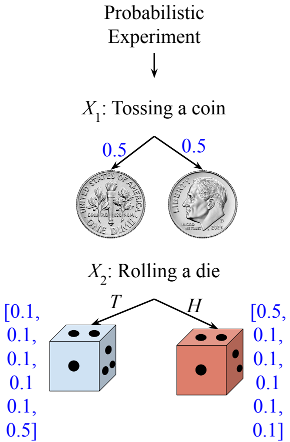


Figure: The distribution for each die (the dice are not fair)

Conditional Distribution

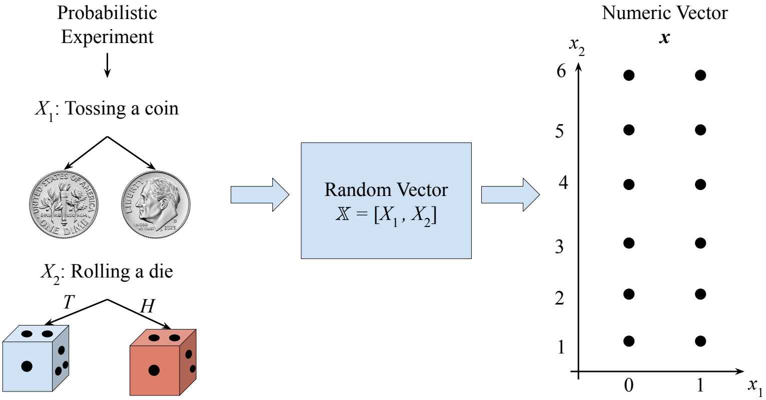


Figure: Two dimensional random variable $\mathbb{X}$ and corresponding numeric vectors

Conditional Distribution

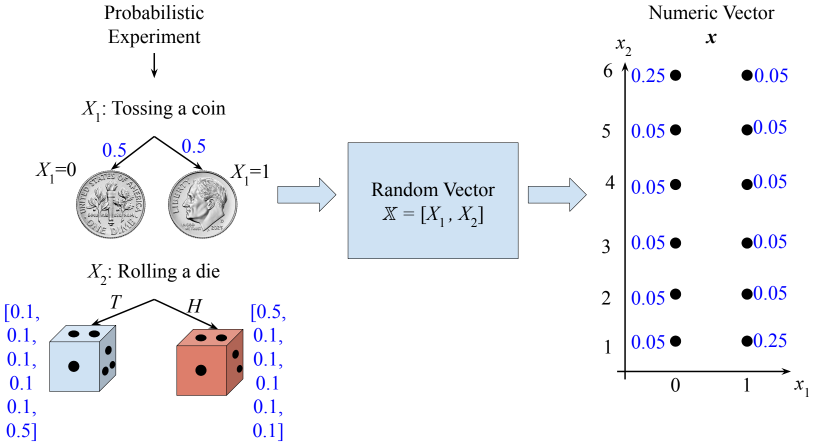
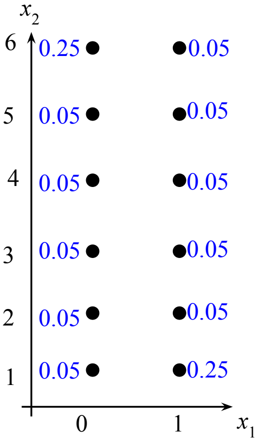


Figure: The PMF over random vector $\mathbb{X}$

Unconditional Event



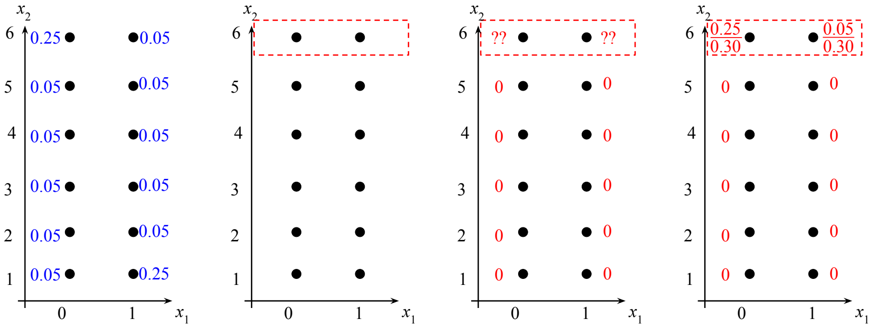
Unconditional Probability

Assume we are interested in calculating the probability for event $X_1 = 0$, then using marginalization we have:

$$p(X_1 = 0) = \sum_{x_2=1}^{6} p(X_1 = 1, X_2 = x_2)$$

$$=0.05 + 0.05 + 0.05 + 0.05 + 0.05 + 0.25 = 0.5$$

Conditional Event



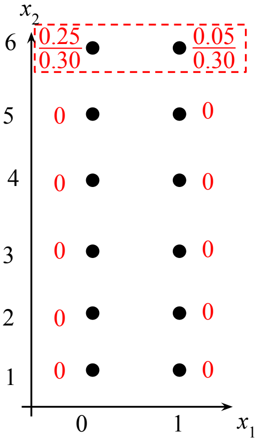
(a) $p(X_1, X_2)$

(b) Event $X_2 = 6$

(c) Impossible events

(d) $p(X_1, X_2 | X_2 = 6)$

Conditional Event



Conditional Probability

Assume we are interested in calculating the probability for event $X_1 = 1$ conditioned on the fact that $X_2 = 6$, then using marginalization we have:

$$
\begin{aligned}
& p(X_1 = 0 | X_2 = 6) \\
&= \sum_{x_2=1}^{6} p(X_1 = 0, X_2 = x_2 | X_2 = 6) \\
&= 0 + 0 + 0 + 0 + 0 + \frac{0.25}{0.30} \simeq 0.83
\end{aligned}
$$

Conditional Probability Mass Function

Conditional PMF

Conditional PMF is a principled way of updating a PMF given the information that some events happened. Conditional PMF is defined as:

$$p(X_1 = x_1 | X_2 = x_2) = \frac{p(X_1 = x_1, X_2 = x_2)}{p(X_2 = x_2)}$$

Extension to Continuous Random Variable [1]

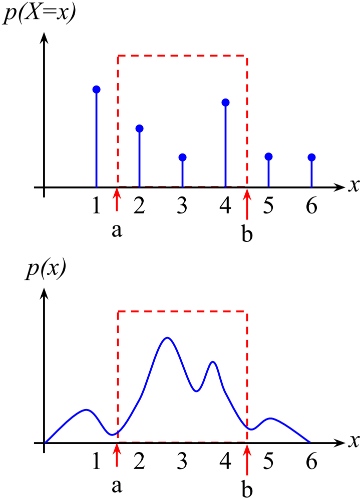


Figure: Probability Density Function

Properties

$$P(a \leq X \leq b) = \sum_{x:a\leq x\leq b} p(X=x)$$

where $p(X=x) \geq 0$ and $\sum_x p(X=x) = 1$

Properties

$$P(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$$

where

$$p(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

Sample Probability Density Function

Gaussian

Gaussian random variable is an example of a continuous random variable denoted by

$$X \sim \mathcal{N}(\mu, \sigma^2) \text{ or } p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

μ is the *mean* and σ^2 is the *variance* of random variable. The probability density function (PDF) is defined as:

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

Sample Probability Density Function

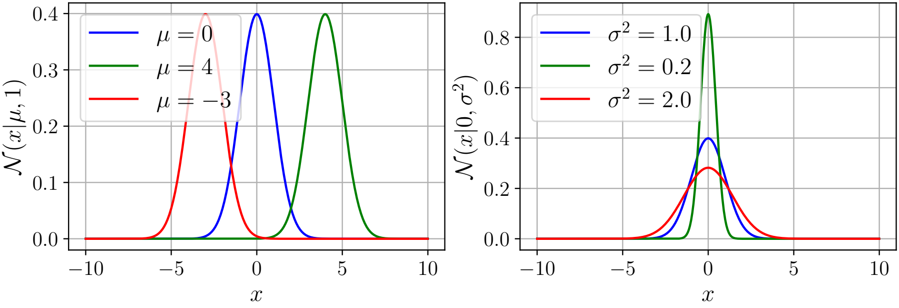


Figure: Mean μ effect (left) and Variance σ^2 (right) on the Gaussian PDF

Notation

PMF and PDF Notation

Throughout this course, we use the following notation:

Definition	Distribution	Numeric Probability
PMF/PDF	$p(X)$ or $p(\mathbb{X})$	$p(x)$ or $p(\boldsymbol{x})$
Conditional PMF/PDF	$p(X\|y)$ or $p(\mathbb{X}\|\boldsymbol{y})$	$p(x\|y)$ or $p(\boldsymbol{x}\|\boldsymbol{y})$
Model PMF/PDF	$p_\theta(\mathbb{X})$ or $p_\theta(\mathbb{X}\|y)$	$p_\theta(\boldsymbol{x})$ or $p_\theta(\boldsymbol{x}\|\boldsymbol{y})$
Data PMF/PDF	$p_{\text{data}}(\mathbb{X})$ or $p_{\text{data}}(\mathbb{X}\|y)$	$p_{\text{data}}(\boldsymbol{x})$ or $p_{\text{data}}(\boldsymbol{x}\|\boldsymbol{y})$

Subsection 2

Expectation

Expectation

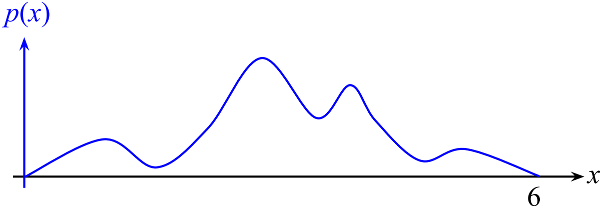


Figure: Probability density function for the time one can walk in 6 hours

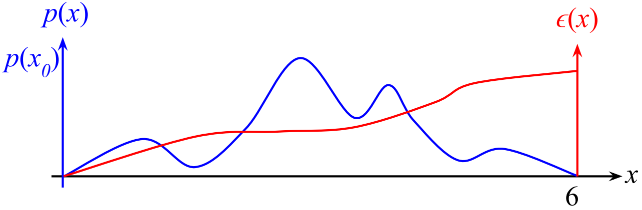


Figure: PDF with energy consumption function $\epsilon(x)$

Expectation

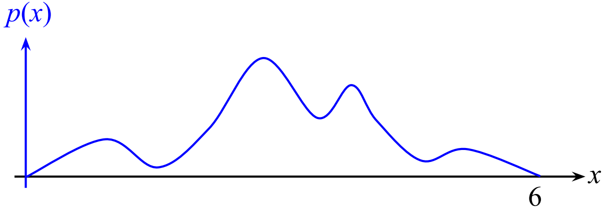


Figure: Probability density function for the time one can walk in 6 hours

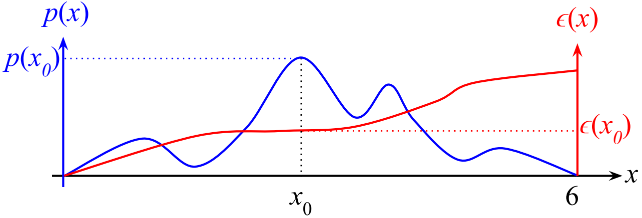


Figure: A sample high probable point x_0

Expectation

In the general case, expectation can be interpreted as the average for the function $\epsilon(x)$ in a large number of *independent* repetitions of the experiment. This value is determined by:

$$\mathbb{E}_{x\sim p(X)}\left[\epsilon(x)\right]$$

and is calculated as:

$$\mathbb{E}_{x\sim p(X)}\left[\epsilon(x)\right] = \int_{-\infty}^{\infty} p(x)\epsilon(x)\mathrm{d}x$$

Equivalently in the case of discrete a random variable, we have:

$$\mathbb{E}_{x\sim p(X)}\left[\epsilon(x)\right] = \sum_{x} p(x)\epsilon(x)$$

Sample Expectations

Sample Expectations

Consider $\mathbb{E}_{x \sim p(X)}\big[f(x)\big]$, then:

- If $f(x) = x$, then the resulting expectation is *mean* and denoted by μ.
- If $f(x) = (x - \mu)^2$, then the resulting expectation is *variance* and denoted by σ^2.
- If $f(x) = x^n$, then the resulting expectation is *n-th raw moment* and denoted by μ'_n.

Monte Carlo Estimation

Consider random variable $\mathbb{X}$ with distribution $p(\mathbb{X})$. The expectation of function $f(\boldsymbol{x})$ can be calculated as:

$$\mathbb{E}_{\boldsymbol{x}\sim p(\mathbb{X})}\left[f(\boldsymbol{x})\right] = \int p(\boldsymbol{x})f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

Now assume that instead of $p(\mathbb{X})$, we just have access to N independent samples of random variable $\mathbb{X}$ as $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. Then we define $f_N(\{\boldsymbol{x}_i\})$ as:

$$f_N(\{\boldsymbol{x}_i\}) = \frac{1}{N}\sum_n f(\boldsymbol{x}_n) \qquad \text{\# Monte Carlo Estimation (MCE)}$$

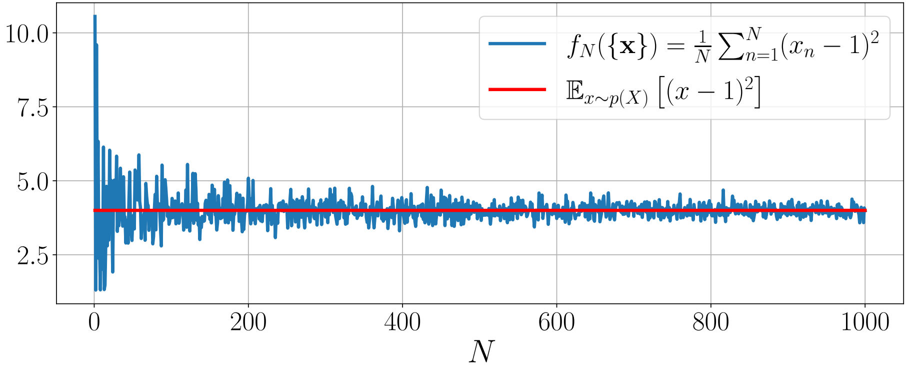
Then using the *weak law of large numbers*, for arbitrary small positive ϵ:

$$\lim_{n\to\infty} P\left(\left|f_N(\{\boldsymbol{x}_i\}) - \mathbb{E}_{\boldsymbol{x}\sim p(\mathbb{X})}\left[f(\boldsymbol{x})\right]\right| \geq \epsilon\right) = 0$$

Estimation Variance

Assume $X \sim N(1, 4)$, then:

- $\mathbb{E}_{x\sim p(X)}\left[(x-1)^2\right] = \sigma^2 = 4$
- The figure below shows the result of MCE for different values of n.



Subsection 3

Distance Metrics

Kullback-Leibler Divergence

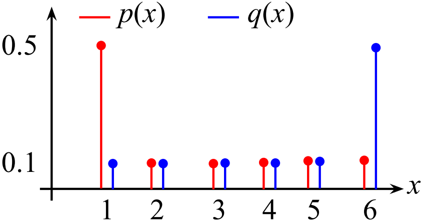
Kullback-Leibler Divergence

Kullback-Leibler divergence (KLD) is a metric to calculate the distance between two distributions. KLD for two distributions p and q defined over discrete random variable X is:

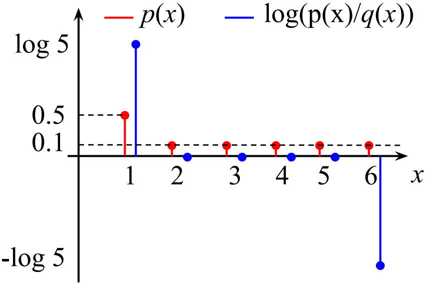
$$\mathrm{KL}\left(p(X)\|q(X)\right) \triangleq \mathbb{E}_{x\sim p(X)}\left[\log\frac{p(x)}{q(x)}\right] = \sum_x p(x)\log\frac{p(x)}{q(x)}$$

- $\mathrm{KL}\left(p(X)\|q(X)\right) \neq \mathrm{KL}\left(q(X)\|p(X)\right) \Rightarrow$ KLD is not symmetric
- $\mathrm{KL}\left(p(X)\|q(X)\right) \geq 0 \Rightarrow$ KLD is non-negative
- $\mathrm{KL}\left(p(X)\|q(X)\right) = 0 \Leftrightarrow p(x) = q(x)\ \forall x$
- KLD is not upper-bounded in general.
- KLD is not a distance metric.
    - It is not symmetric.
    - It does not satisfy the triangle inequality.

Kullback-Leibler divergence



(a) Two distributions $p(X)$ and $q(X)$



Kullback-Leibler divergence

Kullback-Leibler divergence

KLD is similarly defined for consitnuous random variables as:

$$\mathrm{KL}\left(p(X)\|q(X)\right) \triangleq \mathbb{E}_{x\sim p(X)}\left[\log\frac{p(x)}{q(x)}\right] = \int_x p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x$$

Jensen-Shanon Divergence

Jensen-Shanon Divergence (JSD) is a symmetric and smoothed version of the KLD. To define JSD for two distributions p and q, first we should define new distribution $m(X)$ as:

$$m(X) \triangleq \frac{p(X) + q(X)}{2} \Leftrightarrow m(X = x) = \frac{p(X = x) + q(X = x)}{2} \ \forall x$$

Then JSD is defined as:

$$\begin{aligned}
\mathrm{JS}\left(p(X) \| q(X)\right) &\triangleq \frac{1}{2}\Bigg(\mathrm{KL}\left(p(X) \| m(X)\right) + \mathrm{KL}\left(q(X) \| m(X)\right)\Bigg) \\
&= \frac{1}{2}\left(\mathbb{E}_{x \sim p(X)}\left[\log \frac{p(x)}{m(x)}\right] + \mathbb{E}_{x \sim q(X)}\left[\log \frac{q(x)}{m(x)}\right]\right)
\end{aligned}$$

Jensen-Shanon Divergence

Jensen-Shanon Divergence

$$\mathrm{JS}\left(p(X)\|q(X)\right) \triangleq \frac{1}{2}\Bigg(\mathrm{KL}\left(p(X)\|m(X)\right) + \mathrm{KL}\left(q(X)\|m(X)\right)\Bigg)$$

- $\mathrm{JS}\left(p(X)\|q(X)\right) = \mathrm{JS}\left(q(X)\|p(X)\right) \Rightarrow$ JSD is symmetric
- $0 \leq \mathrm{JS}\left(p(X)\|q(X)\right) \leq 1 \Rightarrow$ JSD is non-negative and upper-bounded.
- $\mathrm{JS}\left(p(X)\|q(X)\right) = 0 \Leftrightarrow p(x) = q(x)\ \forall x$
- JSD is not a distance metric.
  - It does not satisfy the triangle inequality.

Square Root of JSD

The square root of JSD $\sqrt{\mathrm{JS}\left(p(X)\|q(X)\right)}$ is a distance metric.

Comparing Divergences

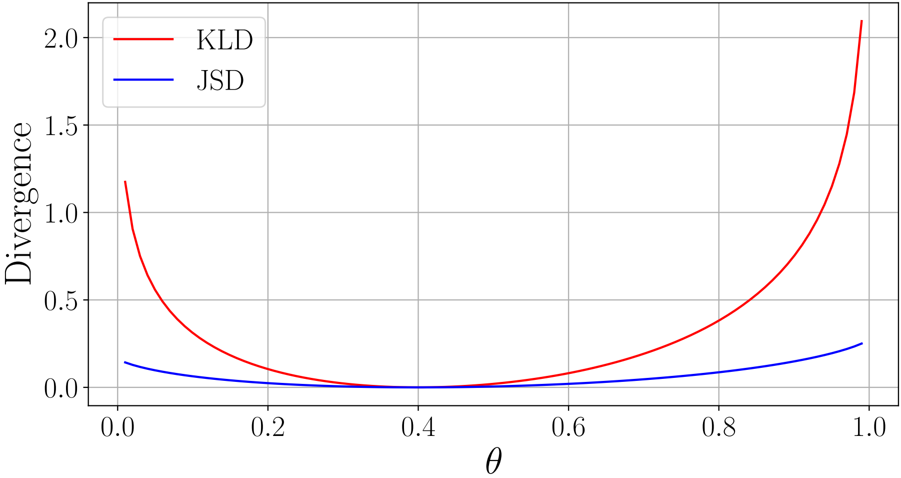


Figure: Distance between $p(X) = \text{Ber}(X|0.4)$ and $q(X) = \text{Ber}(X|\theta)$ as a function of θ

Section 4

Conclusions

Conclusions

Our Foundation

- Notation
- Probability and Statistics
  - Probability Mass/Density Function
  - expectation
  - Distance measurement between densities

List of Abbreviations

Complete	Abbreviation
Jensen-Shanon Divergenc	JSD
Kullback-Leibler divergence	KLD
Left Hand Side	LHS
Monte Carlo Estimation	MCE
Probability Density Function	PDF
Probability Mass Function	PMF
Right Hand Side	RHS

References I

John Tsitsiklis and Patrick Jaillet,
"Mit res.6-012 introduction to probability, spring 2018," Spring 2018.