Basic Mathematics, Fall 2020

Karen Keryan,
ASDS, YSU

November 24, 2020

Gradients of Matrices

	Scalar y Notation Type	Vector $\mathbf{y}$ (size m) Notation Type
Scalar x	$\frac{\partial y}{\partial x}$ scalar	$\frac{\partial \mathbf{y}}{\partial x}$ size-m col. vector
Vector $\mathbf{x}$ (size n)	$\frac{\partial y}{\partial \mathbf{x}}$ size-n row vector	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ $m \times n$ matrix
Matrix $\mathbf{X}$ (size $p \times q$)	$\frac{\partial y}{\partial \mathbf{X}}$ $p \times q$ matrix	$\frac{\partial \mathbf{y}}{\partial \mathbf{X}}$ $m \times (p \times q)$ tensor

Example

(Gradient of Scalars with respect to Matrices)

Let

$$y = y(\mathbf{X}) = \mathrm{tr}(\mathbf{X}), \text{ where } \mathbf{X} \in \mathbb{R}^{p \times p}.$$

Find the gradient $\dfrac{\partial y}{\partial \mathbf{X}}$.

Example

(Gradient of Vectors with respect to Matrices)
Let $\mathbf{v} \in \mathbb{R}^q$ *be a fixed vector and* $\mathbf{f} : \mathbb{R}^{p\times q} \to \mathbb{R}^p$ *be a function given by*

$$\mathbf{f}(\mathbf{X}) = \mathbf{X}\mathbf{v}, \text{ where } \mathbf{X} \in \mathbb{R}^{p\times q}.$$

Find the gradient $\dfrac{\partial \mathbf{y}}{\partial \mathbf{X}}$ *of the function* $\mathbf{y} = \mathbf{f}(\mathbf{X})$.

	Matrix $\mathbf{Y}$ (size $m \times k$) Notation Type
Scalar x	$\frac{\partial \mathbf{Y}}{\partial x}$ $m \times k$ matrix
Vector $\mathbf{x}$ (size n)	$\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}$ $(m \times k) \times n$ tensor
Matrix $\mathbf{X}$ (size $p \times q$)	$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ $(m \times k) \times (p \times q)$ tensor

Example

(Gradient of Matrices with respect to Matrices)
Let $\mathbf{f} : \mathbb{R}^{p \times q} \to \mathbb{R}^{q \times q}$ *be a function given by*

$$\mathbf{f}(\mathbf{X}) = \mathbf{X}^T \mathbf{X}, \ \textit{where } \mathbf{X} \in \mathbb{R}^{p \times q}.$$

Find the gradient $\dfrac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ *of the function* $\mathbf{Y} = \mathbf{f}(\mathbf{X})$.

Useful Identities for Computing Gradients

$$\frac{\partial}{\partial \boldsymbol{X}} f(\boldsymbol{X})^\top = \left(\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)^\top$$

$$\frac{\partial}{\partial \boldsymbol{X}} \mathrm{tr}(f(\boldsymbol{X})) = \mathrm{tr}\left(\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)$$

$$\frac{\partial}{\partial \boldsymbol{X}} \det(f(\boldsymbol{X})) = \det(f(\boldsymbol{X}))\mathrm{tr}\left(f^{-1}(\boldsymbol{X})\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right)$$

$$\frac{\partial}{\partial \boldsymbol{X}} f^{-1}(\boldsymbol{X}) = -f^{-1}(\boldsymbol{X})\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} f^{-1}(\boldsymbol{X})$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{X}^{-1}\boldsymbol{b}}{\partial \boldsymbol{X}} = -(\boldsymbol{X}^{-1})^\top \boldsymbol{a}\boldsymbol{b}^\top(\boldsymbol{X}^{-1})^\top$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^\top$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^\top$$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{X}\boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^\top$$

$$\frac{\partial \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^\top(\boldsymbol{B}+\boldsymbol{B}^\top)$$

$$\frac{\partial}{\partial \boldsymbol{s}}(\boldsymbol{x}-\boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W}(\boldsymbol{x}-\boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x}-\boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W}\boldsymbol{A} \quad \text{for symmetric } \boldsymbol{W}$$

https://explained.ai/matrix-calculus/index.html

Probability

Experiment, Outcomes and the Sample Space

- A **random (or probabilistic) Experiment** is a situation, where we are uncertain about the result.
- An **Outcome** is a possible result of an Experiment.
- The set of all Outcomes of an Experiment is called the **Sample Space** of that Experiment:

$$\Omega = \text{the Sample Space of the Experiment} =$$

$$= \text{the set of all outcomes of our Experiment}$$

Examples

- Our Experiment: we are tossing a (fair) coin.
- **Heads** is one of the outcomes.
- The Sample Space in this Example is:

$$\Omega = \text{Sample Space} = \{\text{Heads, Tails}\} = \{H, T\}$$

Examples

- Experiment: we are rolling a (fair) die.
- One of the outcomes is 3.
- The Sample Space in this Example is: $\{1, 2, 3, 4, 5, 6\}$

Examples

- Experiment: we are interested in the remaining lifetime (in years) of a person (for insurance reasons, say).
- One of the outcomes is 30.1.
- The Sample Space in this Example is: [0,150]

Events Examples

- Experiment: Rolling a die
- Sample Space $= \Omega = \{1, 2, 3, 4, 5, 6\}$
- Some Events:
  - The Result is Odd $= \{1, 3, 5\}$
  - The Result is larger than 2 $= \{3, 4, 5, 6\}$
  - Any Result $= \Omega$
  - No Result $= \varnothing$

Events Examples

- Experiment: Waiting Time (in minutes) for the Metro train
- An example of an outcome: 3.24.
- Sample Space $= \Omega = [0, 20]$
- It is not interesting to have the probability of one outcome: say, what is the probability that the waiting time will be 3.24312456231? **Exactly**, I mean. The answer is 0.
- So in this case we are interested in events' probabilities rather than in particular outcome probability.
- Some Events:
  - The WT is larger than $3 = (3, 20]$
  - The WT is between 2 and 5, included $= [2, 5]$
  - The WT is anything $= \Omega$
  - No Result $= \varnothing$

Probability (Measure) Definition

Probability Measure Definition

A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **Probability Measure** on $(\Omega, \mathcal{F})$, if it satisfies the following axioms:

P1. For any $A \in \mathcal{F}$,
$$\mathbb{P}(A) \geq 0;$$

P2. $\mathbb{P}(\Omega) = 1;$

P3. For any sequence of pairwise mutually exclusive (disjoint) events $A_n \in \mathcal{F}$, i.e., for any sequence $A_n \in \mathcal{F}$ with $A_i \cap A_j = \varnothing$ for $i \neq j$, we have
$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Probability Measure is very similar (and shares the properties of) any other Measure -

- Cardinality (no. of elements),
- Length (in 1D),
- Area (in 2D),
- Volume (in 3D and moreD).

The difference is only that the Probability of the Sample Space is 1, $\mathbb{P}(\Omega) = 1$.

Properties of the Probability Measure

1. $\mathbb{P}(\varnothing) = 0$;
2. if $A, B \in \mathcal{F}$ are mutually exclusive events, i.e., if $A \cap B = \varnothing$, then
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B);$$
3. for any event $A \in \mathcal{F}$,
$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A);$$
Here $\overline{A} = A^c = \Omega \setminus A$.

Properties of the Probability Measure

4. If $A_1, A_2, ..., A_n \in \mathcal{F}$ are pairwise disjoint (mutually exclusive), i.e., if $A_i \cap A_j = \varnothing$ for $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i);$$

5. for any events $A, B \in \mathcal{F}$ (not necessarily disjoint),

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B);$$

Discrete = Finite or Countably Infinite

To give Probability Models, Probability Spaces, we need to give:

- The Sample Space Ω;
- The set of Events $\mathcal{F}$;
- The Probability Measure $\mathbb{P}$.

Classical Probability Models: Finite Sample Spaces

Assume the Sample Space Ω is finite:

- Our Sample Space is $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$.
- Every subset of Ω is an Event, i.e., $\mathcal{F} = 2^\Omega$, the power set of Ω.
- We take any real numbers $p_1, p_2, ..., p_n$ with

  $$p_1 \geq 0, ..., p_n \geq 0, \quad p_1 + p_2 + ... + p_n = 1,$$

  and define

  $$\mathbb{P}(\{\omega_1\}) = p_1, \quad \mathbb{P}(\{\omega_2\}) = p_2, \quad ..., \quad \mathbb{P}(\{\omega_n\}) = p_n.$$

We write this in a more convenient table form:

Outcome	ω_1	ω_2	...	ω_n
$\mathbb{P}(\{\omega_k\})$	p_1	p_2	...	p_n

We are not done yet! We define, for any event A,

$$\mathbb{P}(A) = \sum_{\omega_i \in A} p_i,$$

and also add $\mathbb{P}(\varnothing) = 0$.

Definition: Conditional Probability

Assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a Probability Space and A, B are two events such that $\mathbb{P}(B) \neq 0$. The conditional probability of A given B (or the probability of A under the condition of B) is defined to be

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

Example

Suppose that somebody rolls two fair dice. Compute the probability that the value of the first one is 2, given the information that their sum is no greater than 5.

The Chain Rule (The multiplication rule)

Assume $B \subset \Omega$ is a fixed event and $\mathbb{P}(B) \neq 0$. Then

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

More general

$$\mathbb{P}(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n)$$

$$= \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1 \cap E_2) \ldots \mathbb{P}(E_n|E_1 \cap \ldots \cap E_{n-1})$$

Example

Suppose that an urn contains 6 red and 3 white balls. We draw 2 random balls from the urn without replacement. What is the probability that both drawn balls are red?
What is the probability of all drawn balls are red, if we draw 3 balls?

The Chain Rule (The multiplication rule)

Assume $B \subset \Omega$ is a fixed event and $\mathbb{P}(B) \neq 0$. Then

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

More general

$$\mathbb{P}(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n)$$

$$= \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1 \cap E_2) \ldots \mathbb{P}(E_n|E_1 \cap \ldots \cap E_{n-1})$$

Example

Suppose that an urn contains 6 red and 3 white balls. We draw 2 random balls from the urn without replacement. What is the probability that both drawn balls are red?
What is the probability of all drawn balls are red, if we draw 3 balls? Hint: Let A_i denote the event that the ith ball drawn is red.