Solution 1

Solution 1 (a)

Define ℓ_1

The ℓ_1 or $||x||_1$ is defined as:

$$\ell_1 = ||x||_1 = \sum_{i=1}^{d} |x_i|$$

Compute ℓ_1

Let $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$

$$\begin{aligned} ||x||_1 &= \sum_{i=1}^{3} |x_i| \\ &= |x_1| + |x_2| + |x_3| \\ &= |1| + |-2| + |3| \\ &= 1 + 2 + 3 \\ &= 6 \end{aligned}$$

$\therefore ||x||_1 = 6$

Solution 1

Solution 1 (b)

Define ℓ_2

The ℓ_2 or $||x||_2$ is defined as:

$$\ell_2 = ||x||_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

Compute ℓ_2

Let $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$

$$\begin{aligned} ||x||_2 &= \sqrt{\sum_{i=1}^{3} x_i^2} \\ &= \sqrt{x_1^2 + x_2^2 + x_3^2} \\ &= \sqrt{1^2 + (-2)^2 + 3^2} \\ &= \sqrt{1 + 4 + 9} \\ &= \sqrt{14} \end{aligned}$$

$\therefore ||x||_2 = \sqrt{14}$

Solution 1

Solution 1 (c)

Define ℓ_∞

The ℓ_∞ or $||x||_\infty$ is defined as:

$$\ell_\infty = ||x||_\infty = \max_i |x_i|$$

Compute ℓ_∞

Let $x = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$

$$\begin{aligned} ||x||_\infty &= \max(\{|x_1|, |x_2|, |x_3|\}) \\ &= \max(\{|1|, |-2|, |3|\}) \\ &= \max(\{1, 2, 3\}) \\ &= 3 \end{aligned}$$

$\therefore ||x||_\infty = 3$

Solution 2

Solution 2 (a)

Define ℓ_2 distance

The ℓ_2 distance is defined as:

$$d(x, x')_{\ell_2} = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

Compute ℓ_2 distance

Let $x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$$\begin{aligned} d(x, x')_{\ell_2} &= \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + (x_3 - x'_3)^2 + (x_4 - x'_4)^2} \\ &= \sqrt{((-1) - 1)^2 + (1 - 1)^2 + ((-1) - 1)^2 + (1 - 1)^2} \\ &= \sqrt{(2)^2 + (0)^2 + (2)^2 + (0)^2} \\ &= \sqrt{4 + 0 + 4 + 0} \\ &= \sqrt{8} \end{aligned}$$

$\therefore d(x, x')_{\ell_2} = \sqrt{8}$

Solution 2

Solution 2 (b)

Define ℓ_1 distance

The ℓ_1 distance is defined as:

$$d(x, x')_{\ell_1} = \sum_{i=1}^{n} |x_i - x'_i|$$

Compute ℓ_1 distance

Let $x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$$\begin{aligned} d(x, x')_{\ell_1} &= |x_1 - x'_1| + |x_2 - x'_2| + |x_3 - x'_3| + |x_4 - x'_4| \\ &= |-1-1| + |1-1| + |-1-1| + |1-1| \\ &= |-2| + |0| + |-2| + |0| \\ &= 2 + 0 + 2 + 0 \\ &= 4 \end{aligned}$$

$\therefore d(x, x')_{\ell_1} = 4$

Solution 2

Solution 2 (c)

Define ℓ_∞ distance

The ℓ_∞ distance is defined as:

$$d(x, x')_{\ell_\infty} = \max_i |x_i - x'_i|$$

Compute ℓ_∞ distance

Let $x = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

$$\begin{aligned} d(x, x')_{\ell_\infty} &= \max\{|x_1 - x'_1|, |x_2 - x'_2|, |x_3 - x'_3|, |x_4 - x'_4|\} \\ &= \max\{|-1-1|, |1-1|, |-1-1|, |1-1|\} \\ &= \max\{|-2|, |0|, |-2|, |0|\} \\ &= \max\{2, 0, 2, 0\} \\ &= 2 \end{aligned}$$

$\therefore d(x, x')_{\ell_\infty} = 2$

Solution 3

Solution 3 (a)

(i) max(ℓ_1) given $\|x\|_\infty = 1$

1. Find implications given $\|x\|_\infty = 1$

   Given the constraint $\|x\|_\infty = \max_i |x_i| = 1$. It follows that $|x_i| \leq 1 : \forall i \in \{1, 2, ..., d\}$ where d is the dimension of vector x. The ℓ_1-norm, is the sum of the absolute values of the components(x_i) in the vector x or :

   $$\ell_1 = \sum_{i=1}^{d} |x_i| = |x_1| + |x_2| + \ldots + |x_d|$$

   Each term $|x_i|$ in the sum must be maximized in order to maximaize ℓ_1-norm. The given constraint($\|x\|_\infty = 1$) allows each $|x_i|$ to be at most 1. Hence, the maximum is achieved when $|x_i| = 1$ or $x_i = \pm 1$.

2. Find value of ℓ_1-norm

   $$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{d} |x_i| \\ &\leq \sum_{i=1}^{d} 1 \quad (\text{since } |x_i| \leq \|x\|_\infty = 1) \\ &\leq d \end{aligned}$$

$\therefore$ The vector $x = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ maximizes the norm, with a value of $\|x\|_1 = d$.

(ii): Maximize $\|x\|_2$ given $\|x\|_\infty = 1$

1. Apply part (i) results and solve for ℓ_2-norm

   Given the same constraint $\|x\|_\infty = 1$. From the results in part (i):

   $$\begin{aligned} \|x\|_2 &= \sqrt{\sum_{i=1}^{d} x_i^2} \\ &= \sqrt{\sum_{i=1}^{d} |x_i|} \quad (\text{since all} x_i \in \{-1, 1\} for finding max) \\ &\leq \sqrt{\ell_1} \quad (\text{definition of } \ell_1) \\ &\leq \sqrt{d} \end{aligned}$$

$\therefore$ The vector $x = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ maximizes the norm, with a value of $\|x\|_2 = \sqrt{d}$.

Solution 3

Solution 3 (b)

(i): Maximize $\|x\|_1$ given $\|x\|_2 = 1$

Given the constraint $\|x\|_2 = 1$, or $\sum_{i=1}^{d} x_i^2 = 1$. We seek to maximize $\|x\|_1 = \sum_{i=1}^{d} |x_i|$. By the Cauchy-Schwarz inequality, for vectors $u = (|x_1|, \dots, |x_d|)$ and $v = (1, \dots, 1)$, we have $(\sum |x_i|)^2 \leq (\sum |x_i|^2)(\sum 1^2)$. This is precisely $(\|x\|_1)^2 \leq (\|x\|_2^2)(d)$.

$$\begin{aligned}
(\|x\|_1)^2 &= \left(\sum_{i=1}^{d} |x_i|\right)^2 \\
&\leq \left(\sum_{i=1}^{d} x_i^2\right)\left(\sum_{i=1}^{d} 1^2\right) \quad \text{(Cauchy-Schwarz Inequality)} \\
&\leq (1)(d) \\
\implies \|x\|_1 &\leq \sqrt{d}
\end{aligned}$$

Equality holds when the vectors are linearly dependent, meaning $|x_1| = |x_2| = \cdots = |x_d| = c$. The constraint $\sum x_i^2 = 1$ implies $d \cdot c^2 = 1$, so $c = 1/\sqrt{d}$.

$\therefore$ The vector $x = \begin{bmatrix} \pm 1/\sqrt{d} \\ \vdots \\ \pm 1/\sqrt{d} \end{bmatrix}$ maximizes the norm, with a value of $\|x\|_1 = \sqrt{d}$.

(ii): Maximize $\|x\|_\infty$ given $\|x\|_2 = 1$

Given the constraint $\sum_{i=1}^{d} x_i^2 = 1$, we seek to maximize $\|x\|_\infty = \max_i |x_i|$. Let $|x_k|$ be the component with the maximum absolute value. From the constraint, we can write $x_k^2 + \sum_{i \neq k} x_i^2 = 1$. Since the sum of squares is non-negative, it must be that $x_k^2 \leq 1$, which implies $|x_k| = \|x\|_\infty \leq 1$.

$$\begin{aligned}
\|x\|_\infty^2 &= (\max_i |x_i|)^2 \\
&= \max_i (x_i^2) \\
&\leq \sum_{i=1}^{d} x_i^2 \quad \text{(as all terms are non-negative)} \\
&\leq 1 \\
\implies \|x\|_\infty &\leq 1
\end{aligned}$$

This maximum value of 1 is achieved when one component has a magnitude of 1, which forces all other components to be zero to satisfy the constraint.

$\therefore$ Any standard basis vector e_k (e.g., $x = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$) maximizes the norm, giving $\|x\|_\infty = 1$.

Graduate Level Explanation This exercise demonstrates the geometric relationship between the unit balls of different ℓ_p norms. The task of maximizing one norm subject to a constraint on another is equivalent to finding the point on the surface of one unit ball that is "farthest" from the origin as measured by the other norm's metric. The results confirm the well-known norm inequalities for $x \in \mathbb{R}^d$:

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2 \leq d\|x\|_\infty$$

Our findings show that these bounds are tight:

- When constrained to the ℓ_∞ unit ball (a hypercube), the points that maximize the ℓ_1 and ℓ_2 norms are the corners of the hypercube, such as the vector of all ones. At these points, the inequalities $\|x\|_1 \leq d\|x\|_\infty$ and $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$ become equalities.
- When constrained to the ℓ_2 unit ball (a hypersphere), the point that maximizes the ℓ_1 norm is where the energy is distributed equally among all components (e.g., $(1/\sqrt{d}, \ldots, 1/\sqrt{d})$). This point makes the inequality $\|x\|_1 \leq \sqrt{d}\|x\|_2$ an equality. Conversely, the ℓ_∞ norm is maximized when all energy is concentrated in a single component (e.g., a basis vector like $(1, 0, \ldots, 0)$), which corresponds to the points where the hypersphere intersects the coordinate axes and makes the inequality $\|x\|_\infty \leq \|x\|_2$ an equality.

Explanation for 5 year old Imagine a big, flat playground. We're going to draw two different "play zones" you have to stay inside.

1. A perfect **square** play zone.
2. A perfect **circle** play zone.

Now, we want to find the spot inside your zone that is the "farthest" from the very center of the playground. But we have different ways to measure "farthest"!

- **Walking Distance:** How many steps you take along the grid lines (like city blocks) to get back to the center.
- **Flying Distance:** The straight line distance if you could fly like a bird.

Here is what we find:

- If you are in the **SQUARE** zone: The farthest place you can be, for *both* walking and flying, is always at one of the four **corners** of the square!
- If you are in the **CIRCLE** zone: It gets tricky!
  - To get the biggest **Walking Distance**, you should stand exactly in the middle of the curvy edge, halfway between North and East. You have to walk a medium amount in two directions.
  - To get the biggest "single-step" distance (the longest part of your walk), you should stand right at the top of the circle (the "North Pole"). Here, you put all your effort into one big step North and took zero steps East or West.

So, the "best" spot depends on both the shape of your play zone and how you decide to measure the distance!

Solution 4

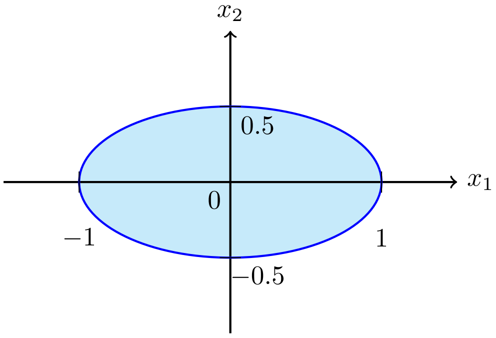
Find Unit Ball Equation

Given $||x||_w \leq 1$ dimension $d = 2$ and the weight vector $w = (w_1, w_2) = (1, 4)$. We can solve for the equation of this boundary by looking at the max value of the weighted-norm ($||x||_w = 1$).

$$\begin{aligned}
\|x\|_w &= 1 \\
\sqrt{\sum_{i=1}^{d} w_i x_i^2} &= 1 \quad \text{(Definition of norm for } d = 2\text{)} \\
\sqrt{w_1 x_1^2 + w_2 x_2^2} &= 1 \quad \text{(Expanding summation)} \\
\sqrt{1 \cdot x_1^2 + 4 \cdot x_2^2} &= 1 \quad \text{(Substituting givens)} \\
x_1^2 + 4x_2^2 &= 1 \quad \text{(Squaring both sides)} \\
\frac{x_1^2}{1^2} + \frac{x_2^2}{(1/2)^2} &= 1 \quad \text{(Arange into equation of ellipse)}
\end{aligned}$$

Hence, the unit ball is an ellipse centered at the origin from the result above.

Sketch of the Unit Ball



$\therefore$ The unit ball is an ellipse defined by $x_1^2 + 4x_2^2 \leq 1$, with a semi-major axis of length 1 and a semi-minor axis of length 1/2.

Solution 5

Distance Table Discription

We are given a set of points $\mathcal{X} = \{A, B, C, D\}$ and a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by the following distance table. We will determine if d is a metric by checking the required axioms.

$d(x, y)$	A	B	C	D
A	0	2	1	5
B	2	0	3	4
C	1	3	0	2
D	5	4	2	0

Axiom 1: Non-negativity and Identity of Indiscernibles

This axiom requires that $d(x, y) \geq 0$ for all $x, y \in \mathcal{X}$, and that $d(x, y) = 0$ if and only if $x = y$.

- **Non-negativity:** All entries in the table are non-negative, so this condition holds.
- **Identity:** The diagonal entries are all zero, so $d(x, x) = 0$. All off-diagonal entries are strictly positive, so $d(x, y) > 0$ when $x \neq y$.

Hence, d satisifies the first axiom (non-negativity & identity of indiscernibles).

Axiom 2: Symmetry

This axiom requires that $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$. Using the distance table above we can look up the folowiing:

- $d(A, B) = 2 = d(B, A)$
- $d(A, C) = 1 = d(C, A)$
- $d(A, D) = 5 = d(D, A)$
- $d(B, C) = 3 = d(C, B)$
- $d(B, D) = 4 = d(D, B)$
- $d(C, D) = 2 = d(D, C)$

Hence, d satisifies the second axiom (symmetry).

Axiom 3: The Triangle Inequality

This axiom requires that for any three points $x, y, z \in \mathcal{X}$, the inequality $d(x, z) \leq d(x, y) + d(y, z)$ must hold.
Let $x = A$, $y = C$, and $z = D$.

$$\begin{aligned} d(x, z) &\leq d(x, y) + d(y, z) \\ d(A, D) &\leq d(A, C) + d(C, D) \\ 5 &\leq 1 + 2 \\ 5 &\leq 3 \quad (\textbf{False}) \end{aligned}$$

Hence, the triangle inequality does not hold, and d fails the third axiom.

$\therefore$ The function d is not a metric because it fails the triangle inequality.

Solution 6

Solution 6 (a)

The Largest Possible Value of $K(p,q)$

The KL divergence becomes unbounded and approaches infinity if there is an outcome $x \in \mathcal{X}$ that is possible under the true distribution p (i.e., $p(x) > 0$) but is considered impossible by the approximating distribution q (i.e., $q(x) = 0$). This situation violates the condition of absolute continuity. The term $p(x) \log \frac{p(x)}{q(x)}$ becomes infinite, driving the entire sum to infinity. We demonstrate this with two deterministic but opposing distributions.
Let $p = (1,0)$ and $q = (0,1)$. This means $p(x_1) = 1, p(x_2) = 0$ and $q(x_1) = 0, q(x_2) = 1$.

$$\begin{aligned} K(p,q) &= p(x_1) \log \frac{p(x_1)}{q(x_1)} + p(x_2) \log \frac{p(x_2)}{q(x_2)} \\ &= 1 \cdot \log \frac{1}{0} + 0 \cdot \log \frac{0}{1} \\ &= \infty + 0 \\ &= \infty \end{aligned}$$

$\therefore$ The largest possible value of the KL divergence is ∞.

Solution 6

Solution 6 (b)

(b): Proof of Non-Symmetry

To prove that the KL divergence is not a symmetric function, we must provide a concrete example where $K(p, q) \neq K(q, p)$. We choose one distribution to be uniform and the other to be deterministic. This highlights the asymmetry in how KL divergence penalizes impossible events.
Let $p = (1/2, 1/2)$ and $q = (1, 0)$. First, we compute $K(p, q)$.

$$\begin{aligned} K(p,q) &= p(x_1)\log\frac{p(x_1)}{q(x_1)} + p(x_2)\log\frac{p(x_2)}{q(x_2)} \\ &= \frac{1}{2}\log\frac{1/2}{1} + \frac{1}{2}\log\frac{1/2}{0} \\ &= \frac{1}{2}\log\frac{1}{2} + \infty \\ &= \infty \end{aligned}$$

Next, we compute the reverse divergence, $K(q, p)$.

$$\begin{aligned} K(q,p) &= q(x_1)\log\frac{q(x_1)}{p(x_1)} + q(x_2)\log\frac{q(x_2)}{p(x_2)} \\ &= 1\cdot\log\frac{1}{1/2} + 0\cdot\log\frac{0}{1/2} \\ &= \log(2) + 0 \\ &= \log(2) \end{aligned}$$

$\therefore$ Since $K(p, q) = \infty$ and $K(q, p) = \log(2)$, we have shown that $K(p, q) \neq K(q, p)$, and thus KL divergence is not symmetric.

Graduate Level Explanation The Kullback-Leibler divergence is a measure of **relative entropy**, not a distance metric. Its failure to be a metric is fundamental to its interpretation. The non-symmetry shown above is a key property. $K(p, q)$ measures the information lost when q is used to approximate p; this is an inherently directional concept.

The unboundedness of $K(p, q)$ is a direct consequence of the violation of **absolute continuity**. The divergence is finite if and only if the support of p is a subset of the support of q (denoted $\text{supp}(p) \subseteq \text{supp}(q)$ or $p \ll q$). If this condition fails, there exists an event x for which $p(x) > 0$ but $q(x) = 0$. The model q assigns zero probability to an event that can actually occur, leading to an infinite "surprise" or error. This property is crucial in variational inference and Bayesian model comparison, where an infinite divergence signals a catastrophic failure of the approximating distribution to cover the posterior's support. KL divergence is a member of the broader classes of *f-divergences* and *Bregman divergences*, none of which are required to be symmetric.

Explanation for 5 year old Imagine a guessing game. The "Truth" knows the right answer (p), and you are making a guess (q). The KL divergence is a number that tells us how surprised the Truth is by your guess. A bigger number means more surprise!

- **Why it can be infinite:** The Truth is "The secret animal is a Dog" (p is 100% Dog). You guess, "I am 100% certain it's a Cat!" (q is 100% Cat). You said a Dog was impossible! When the Truth reveals it's a Dog, your surprise is infinite because you were completely wrong about something that was certain.
- **Why it's not the same backwards:** Let's look at two cases.
  1. The Truth is "The secret animal could be a Dog or a Cat, 50/50 chance" (p is 50/50). You guess, "It's definitely a Dog!" (q is 100% Dog). The Truth is infinitely surprised because you said a Cat was impossible, but it was possible.
  2. Now let's switch! The Truth is "It's definitely a Dog" (q is 100% Dog). You guess, "It could be a Dog or a Cat, 50/50 chance" (p is 50/50). Is the Truth surprised? A little bit! The Truth is surprised you weren't more confident, but not infinitely surprised, because you at least said a Dog was possible.

Since the surprise is **infinite** in the first game but just a **little bit** in the second game, the "surprise-o-meter" doesn't work the same forwards and backwards!

Solution 7

Define Jaccard Similarity

Jaccard similarity between two sets A and B is the following:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Compute Jaccard Similarity

Let $A = \{1,3,5,7,9\}$ and $B = \{2,3,5,7\}$

$$\begin{aligned} J(A,B) &= \frac{|A \cap B|}{|A \cup B|} \\ &= \frac{|\{1,3,5,7,9\} \cap \{2,3,5,7\}|}{|\{1,3,5,7,9\} \cup \{2,3,5,7\}|} \\ &= \frac{|\{3,5,7\}|}{|\{1,2,3,5,7,9\}|} \\ &= \frac{3}{6} \\ &= \frac{1}{2} \end{aligned}$$

$\therefore$ The Jaccard similarity $J(A,B)$ is $\frac{1}{2}$ or 50%.

Solution 8

Solution: Bigram Jaccard Similarity

We are asked to compute the Jaccard similarity based on word bigrams for the following two sentences, where x' is pre-processed to remove punctuation:
x: "Napoleon was born in 1769"
x': "Napoleon was born when"
The Jaccard similarity for the bigram sets $B(x)$ and $B(x')$ is defined as $J(B(x), B(x')) = \frac{|B(x) \cap B(x')|}{|B(x) \cup B(x')|}$.

Step 1: Derive Bigram Sets

First, we generate the sets of consecutive word pairs (bigrams) for each sentence.

- $B(x) = \{(\text{Napoleon, was}), (\text{was, born}), (\text{born, in}), (\text{in, 1769})\}$
- $B(x') = \{(\text{Napoleon, was}), (\text{was, born}), (\text{born, when})\}$

The sizes of the sets are $|B(x)| = 4$ and $|B(x')| = 3$.

Step 2: Compute Intersection and Union of Sets

Next, we find the intersection (common bigrams) and the union (all unique bigrams) of the two sets.

- Intersection: $B(x) \cap B(x') = \{(\text{Napoleon, was}), (\text{was, born})\}$
- Union: $B(x) \cup B(x') = \{(\text{Napoleon, was}), (\text{was, born}), (\text{born, in}), (\text{in, 1769}), (\text{born, when})\}$

The sizes are $|B(x) \cap B(x')| = 2$ and $|B(x) \cup B(x')| = 5$.

Step 3: Compute the Jaccard Similarity

Finally, we apply the Jaccard similarity formula using the sizes of the intersection and union.

$$\begin{aligned} J(B(x), B(x')) &= \frac{|B(x) \cap B(x')|}{|B(x) \cup B(x')|} \\ &= \frac{2}{5} \\ &= 0.4 \end{aligned}$$

$\therefore$ The bigram-based Jaccard similarity between the two sentences is 0.4

Graduate Level Explanation

Using Jaccard similarity over n-grams (in this case, bigrams) allows for a measure of text similarity that captures local syntactic and semantic structure, which is a significant improvement over simple bag-of-words models. Unlike methods that treat words as independent, n-grams preserve word order within a small window. This makes the metric sensitive to phrasal correspondence but robust to larger-scale sentence reordering. It is particularly effective for tasks like plagiarism detection, identifying near-duplicate documents, and record linkage, where detecting overlapping text chunks is more important than understanding deep semantic meaning. While it is computationally simpler than embedding-based methods like cosine similarity on BERT vectors, it effectively balances structural awareness with efficiency.

Explanation for 5 year old

Imagine you and your friend are building sentences with Lego bricks, where each brick is a word. Instead of just counting how many of the same color bricks you both used, we're going to look at how you connected them. We'll look at every pair of bricks stuck together. A pair of word-bricks is a "bigram".

1. First, we find all the two-brick connections that are exactly the same in both of your sentences. Let's say you find **2** matching connections.
2. Next, we count all the unique connections you both made. Maybe in total, there are **5** different connections.

To see how similar your sentences are, we just divide the number of matching connections by the total number of connections: 2 divided by 5. That gives a similarity score! It tells us how much of the sentences were built in the same way.

Solution 9

Solution 9 (a)

Define Cosine Similarity

The cosine similarity between two vectors x and y is defined as:

$$\cos(\theta) = \frac{x \cdot y}{||x||||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Compute Cosine Similarity

Let $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $x' = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

$$\begin{aligned}
\cos(\theta) &= \frac{x \cdot x'}{||x||||x'||} \\
&= \frac{\sum_{i=1}^{n} x_i x'_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} {x'}_i^2}} \\
&= \frac{(1)(3) + (2)(2) + (3)(1)}{\sqrt{1^2 + 2^2 + 3^2}\sqrt{3^2 + 2^2 + 1^2}} \\
&= \frac{3 + 4 + 3}{\sqrt{1 + 4 + 9}\sqrt{9 + 4 + 1}} \\
&= \frac{10}{\sqrt{14}\sqrt{14}} \\
&= \frac{10}{14} \\
&= \frac{5}{7}
\end{aligned}$$

$\therefore$ The cosine similarity between x and x' is $\frac{5}{7}$.

Solution 9

Solution 9 (b)

(b): Characterization of Zero Similarity

The cosine similarity between two non-zero vectors x and x' is zero if and only if the numerator of the formula is zero.

$$\cos(\theta) = 0 \iff x \cdot x' = 0$$

In a Euclidean vector space, a dot product of zero signifies that the two vectors are orthogonal (perpendicular) to each other. The angle θ between them is 90° or $\pi/2$ radians. Geometrically, they form a right angle.

$\therefore$ A cosine similarity of zero means the vectors are orthogonal to each other.

Solution 9

Solution 9 (c)

Question: Geometric Region

Sketch the region of all vectors $y \in \mathbb{R}^2$ whose cosine similarity with the vector $x = [1, 2]^T$ is at least 0.9.

Step 1: Define the Geometric Constraint

The cosine similarity between two non-zero vectors x and y in $\mathbb{R}^n$ is defined as the cosine of the angle θ between them. The region R is the set of all points $y \in \mathbb{R}^2$ that satisfy the inequality:

$$\cos(\theta) = \frac{x \cdot y}{\|x\|_2 \|y\|_2} \geq 0.9$$

Since this constraint depends only on the angle θ between the vectors and not on the magnitude of y, the region is unbounded and extends radially from the origin. The boundary of this region is defined by the equality $\cos(\theta) = 0.9$.

Step 2: Calculate Angular Boundaries

The maximum allowable angle, θ_{max}, between any vector y in the region and the reference vector x is given by the boundary condition. The reference vector $x = [1, 2]^T$ itself forms an angle ϕ with the positive x_1-axis.

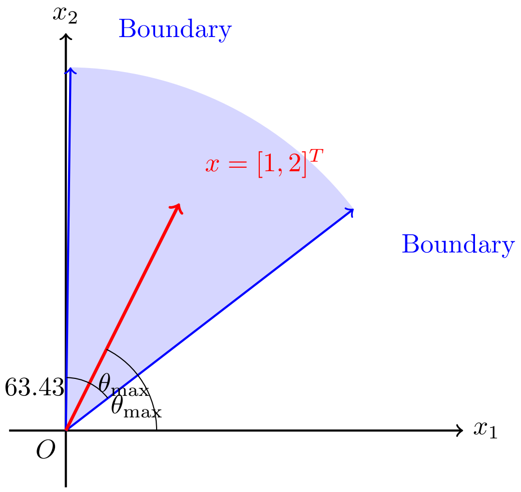
$$\begin{aligned} \text{Boundary Condition:} \quad & \cos(\theta_{\text{max}}) = 0.9 \\ & \implies \theta_{\text{max}} = \arccos(0.9) \approx 25.84^\circ \end{aligned}$$

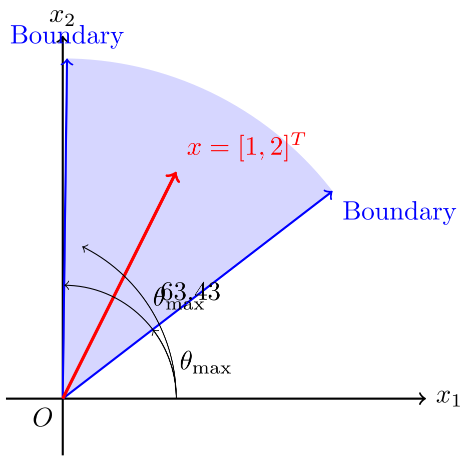
$$\text{Reference Vector Angle:} \quad \phi = \arctan\left(\frac{x_2}{x_1}\right) = \arctan\left(\frac{2}{1}\right) \approx 63.43^\circ$$

The region is therefore an angular sector (a cone in $\mathbb{R}^2$) centered at angle $\phi \approx 63.43^\circ$, with an angular radius of $\theta_{\text{max}} \approx 25.84^\circ$. The cone is bounded by the angles $\phi - \theta_{\text{max}} \approx 37.59^\circ$ and $\phi + \theta_{\text{max}} \approx 89.27^\circ$.

Step 3: Sketch of the Region

The shaded region below represents the cone containing all vectors y satisfying the condition.





(c): Sketch of a High-Similarity Region

We want to sketch the region in $\mathbb{R}^2$ where any vector x' has a cosine similarity of at least 0.9 with the vector $x = [1, 2]^T$. The condition is $\cos(\theta) \geq 0.9$, which implies that the angle θ between x and x' must satisfy $|\theta| \leq \arccos(0.9) \approx 25.84°$.

Description of the Sketch:

- Draw the standard Cartesian axes, x_1 and x_2.
- Draw the reference vector x as an arrow from the origin $(0, 0)$ to the point $(1, 2)$.
- The region of high similarity is a double-sided cone (a pair of sectors in 2D) centered on the line passing through the origin and the point $(1, 2)$.
- The cone opens with an angle of $2 \times 25.84°$. Any vector x' that originates at $(0, 0)$ and whose tip lies within this cone satisfies the condition.

$\therefore$ The region is a cone centered on the vector x with an angular radius of $\arccos(0.9)$.

Graduate Level Explanation

Cosine similarity is a measure of orientation, not magnitude. By normalizing the vectors to unit length via the ℓ_2-norm, it isolates the angle between them, making it an effective measure of similarity in applications where vector length is a confounding variable. For instance, in Natural Language Processing, documents are often represented as high-dimensional TF-IDF vectors. A longer document might have a larger Euclidean distance from a shorter one, even if they discuss the same topic. Cosine similarity correctly identifies them as similar by disregarding their ”length” (i.e., word count) and focusing solely on their ”direction” (i.e., topic, as defined by the distribution of word frequencies). This property is also critical in recommender systems, where it can measure the similarity of user preferences irrespective of the total number of items they have rated.

Explanation for 5 year old

Imagine you and your friend are each pointing with an arrow. Cosine similarity doesn't care how long your arrows are, only which way they're pointing!

- If you both point in the **exact same direction**, the similarity is **1**. Perfect match!
- If your arrows make a perfect ”L” shape (a right angle), the similarity is **0**. You're not pointing in similar directions at all.
- If you point in the **exact opposite direction** of your friend, the similarity is **-1**. You completely disagree!

It's just a score from -1 to 1 that tells you how much your arrows line up.

Solution 10

Solution 10 (a)

1 (a): Mean and Median (Theoretical)

The mean (Expected Value) $E[X]$ is calculated as the sum of each outcome weighted by its probability: $E[X] = \sum_i x_i P(X = x_i)$. The median is the value x_m for which the cumulative probability $P(X \leq x_m)$ first equals or exceeds 0.5.

$$\begin{aligned}
E[X] &= (1)\left(\frac{1}{3}\right) + (2)\left(\frac{1}{3}\right) + (3)\left(\frac{1}{12}\right) + (4)\left(\frac{1}{12}\right) + (5)\left(\frac{1}{12}\right) + (6)\left(\frac{1}{12}\right) \\
&= \frac{4}{12} + \frac{8}{12} + \frac{3}{12} + \frac{4}{12} + \frac{5}{12} + \frac{6}{12} \\
&= \frac{30}{12} = \frac{5}{2} = 2.5
\end{aligned}$$

For the median, we find the cumulative probabilities:
$P(X \leq 1) = 1/3 \approx 0.333$
$P(X \leq 2) = 1/3 + 1/3 = 2/3 \approx 0.667$
Since $P(X \leq 2)$ is the first cumulative probability to exceed 0.5, the median is 2.

$\therefore$ The theoretical mean is $E[X] = 2.5$ and the median is 2.

Solution 10

Solution 10 (b)

Variance is calculated as $\text{Var}(X) = E[X^2] - (E[X])^2$, where $E[X^2] = \sum_i x_i^2 P(X = x_i)$. The standard deviation is the square root of the variance, $\text{SD}(X) = \sqrt{\text{Var}(X)}$.

$$\begin{aligned} E[X^2] &= (1^2)\left(\frac{1}{3}\right) + (2^2)\left(\frac{1}{3}\right) + (3^2)\left(\frac{1}{12}\right) + (4^2)\left(\frac{1}{12}\right) + (5^2)\left(\frac{1}{12}\right) + (6^2)\left(\frac{1}{12}\right) \\ &= \frac{4}{12} + \frac{16}{12} + \frac{9}{12} + \frac{16}{12} + \frac{25}{12} + \frac{36}{12} \\ &= \frac{106}{12} = \frac{53}{6} \approx 8.833 \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E[X^2] - (E[X])^2 = \frac{53}{6} - \left(\frac{5}{2}\right)^2 \\ &= \frac{53}{6} - \frac{25}{4} = \frac{106}{12} - \frac{75}{12} = \frac{31}{12} \approx 2.5833 \end{aligned}$$

$$\text{SD}(X) = \sqrt{\frac{31}{12}} \approx 1.607$$

$\therefore$ The theoretical variance is $\text{Var}(X) = 31/12 \approx 2.5833$ and the standard deviation is $\text{SD}(X) \approx 1.607$.

Solution 10

Solution 10 (c)

1 (c): Empirical Probability Distribution

Given the set of $N = 10$ observations $D = \{2, 5, 1, 4, 2, 2, 5, 6, 1, 2\}$, the empirical probability $\hat{P}(X = x)$ for each outcome is its observed frequency divided by the total number of observations, N.
The counts for each outcome are:
Count(1) $= 2$
Count(2) $= 4$
Count(3) $= 0$
Count(4) $= 1$
Count(5) $= 2$
Count(6) $= 1$

$\therefore$ The empirical probabilities are: $\hat{P}(1) = 0.2$, $\hat{P}(2) = 0.4$, $\hat{P}(3) = 0.0$, $\hat{P}(4) = 0.1$, $\hat{P}(5) = 0.2$, $\hat{P}(6) = 0.1$.

Solution 10

Solution 10 (d)

1 (d): Mean, Median, Variance, and SD (Empirical)

The empirical mean is $\bar{x} = \frac{1}{N}\sum x_i$. The median is the middle value of the sorted data. The empirical variance is $s^2 = \frac{1}{N}\sum(x_i - \bar{x})^2$, and the standard deviation is $s = \sqrt{s^2}$.

$$\bar{x} = \frac{1}{10}(2+5+1+4+2+2+5+6+1+2) = \frac{30}{10} = 3.0$$

Sorted data: $\{1, 1, 2, 2, \mathbf{2}, \mathbf{2}, 4, 5, 5, 6\}$. The median is the average of the 5th and 6th values: $\frac{2+2}{2} = 2$.

$$\begin{aligned}
s^2 &= \frac{1}{10}\sum_{i=1}^{10}(x_i - 3.0)^2 \\
&= \frac{1}{10}\left[(1-3)^2 \times 2 + (2-3)^2 \times 4 + (4-3)^2 \times 1 + (5-3)^2 \times 2 + (6-3)^2 \times 1\right] \\
&= \frac{1}{10}\left[(-2)^2 \times 2 + (-1)^2 \times 4 + (1)^2 \times 1 + (2)^2 \times 2 + (3)^2 \times 1\right] \\
&= \frac{1}{10}\left[4 \times 2 + 1 \times 4 + 1 \times 1 + 4 \times 2 + 9 \times 1\right] \\
&= \frac{1}{10}\left[8 + 4 + 1 + 8 + 9\right] = \frac{30}{10} = 3.0 \\
\\
s &= \sqrt{3.0} \approx 1.732
\end{aligned}$$

$\therefore$ The empirical mean is $\bar{x} = 3.0$, median is 2, variance is $s^2 = 3.0$, and standard deviation is $s \approx 1.732$.

Graduate Level Explanation: Theoretical vs. Empirical

The theoretical distribution describes the true, underlying probability model of the random variable X. Its moments, such as the mean $\mu = E[X]$ and variance $\sigma^2 = \text{Var}(X)$, are fixed population parameters derived from this model. The empirical distribution, conversely, is constructed from a finite sample of observations. Its statistics, like the sample mean $\bar{x}$ and sample variance s^2, are estimates of the true parameters. The **Law of Large Numbers (LLN)** states that as the sample size N approaches infinity, the sample mean $\bar{x}$ converges in probability to the theoretical mean μ. Similarly, other empirical moments converge to their theoretical counterparts. The discrepancy between our calculated theoretical values ($\mu = 2.5, \sigma^2 \approx 2.58$) and empirical values ($\bar{x} = 3.0, s^2 = 3.0$) is expected due to random sampling variation in our small sample ($N = 10$).

Explanation for a 5-Year-Old

Imagine you have a magic cookie jar. The **plan** (the theoretical part) says that for every 12 cookies you pull out, you *should* get 4 chocolate chip, 4 oatmeal, 1 sugar, 1 peanut butter, 1 ginger, and 1 snickerdoodle. The plan's “average cookie” is a mix between a chocolate chip and an oatmeal cookie.

But then, you actually pull out just 10 cookies. This is your **handful** (the empirical part). In your handful, you got a lot of oatmeal cookies and no sugar cookies at all! What happened in your one small handful is a little different from the big plan for the whole jar. If you kept pulling out cookies all day (thousands of them!), your handful would start to look a lot more like the original plan.

Solution 11

Solution 11 (a)

Justification: The distribution of human height for a given population (e.g., adult males) is famously well-approximated by a symmetric, bell-shaped curve (a Normal or Gaussian distribution). In a perfectly symmetric distribution, the mean, median, and mode coincide. Therefore, the center of mass (mean) and the geometric center (median) are expected to be nearly identical.

$\therefore$ We expect $\text{Mean}(H) \approx \text{Median}(H)$. No significant difference.

Solution 11

Solution 11 (b)

Justification: The distribution of housing costs is almost always characterized by a strong right-skew. Most houses fall within a certain price range, there is a long tail of extremely expensive properties. These high-value outliers pull the mean significantly upward, while the median remains a more robust measure of the ”typical” house price, unaffected by these extreme values.

$\therefore$ We expect Mean(C) $>$ Median(C). A significant difference.

Solution 11

Solution 11 (c)

Justification: The distribution of GPAs is often left-skewed. This is due to a ”ceiling effect,” where a large number of students achieve high grades clustered near the maximum possible GPA (≈ 4.0), while fewer students have very low GPAs. This clustering at the high end pulls the median towards the right, while the lower-end scores pull the mean to the left.

$\therefore$ We expect $\text{Mean}(G) < \text{Median}(G)$, but the difference may not be as significant as for house cost or salary.

Solution 11

Solution 11 (d)

Justification: Similar to housing costs, salary data exhibits a pronounced right-skew. The majority of people earn modest. However, a small number of individuals (CEOs, top athletes, etc.) have high incomes. These outliers have strong influence on the mean, that pulls it far to the right of the median. The median salary is therefore a much more accurate representation of a typical worker's salary.

$\therefore$ We expect $\text{Mean}(S) > \text{Median}(S)$. A significant difference.

Solution 12

Define relationships

We know the following relationships are true:

$$Var[Z] = E[Z^2] - (E[Z])^2$$

$$\sigma^2 = \mathrm{Var}[Z]$$

$$E[Z] = \mu$$

Rearranging this formula allows us to solve for $E[Z^2]$:

$$E[Z^2] = Var[Z] + (E[Z])^2 = \sigma^2 + \mu^2$$

Compute $E[Z^2]$

Given: $\mu = E[Z] = -1$ and the standard deviation $SD[Z] = \sigma = 2$

$$\begin{aligned} E[Z^2] &= \sigma^2 + \mu^2 \\ &= (2)^2 + (-1)^2 \\ &= 4 + 1 \\ &= 5 \end{aligned}$$

$\therefore$ $E[Z^2]$ is 5.

Solution 13

Solution 13 (a)

Define X and Y Independence

Two discrete random variables X and Y are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for all possible pairs (x, y).

Compute Marginal Probability Mass Functions

We sum rows and columns of the joint PMF table to compute $P(X = x)$ and $P(Y = y)$.

$$\begin{aligned} P(X = 1) &= 0.10 + 0.20 + 0.05 = 0.35 \\ P(X = 2) &= 0.10 + 0.15 + 0.05 = 0.30 \\ P(X = 3) &= 0.10 + 0.15 + 0.10 = 0.35 \end{aligned}$$

$$\begin{aligned} P(Y = 1) &= 0.10 + 0.10 + 0.10 = 0.30 \\ P(Y = 2) &= 0.20 + 0.15 + 0.15 = 0.50 \\ P(Y = 3) &= 0.05 + 0.05 + 0.10 = 0.20 \end{aligned}$$

Test Independence Condition

The test below is the independence condition for the pair $(X = 1, Y = 1)$:

$$\begin{aligned} P(X = 1, Y = 1) &\overset{?}{=} P(X = 1)P(Y = 1) \\ 0.10 &\overset{?}{=} (0.35)(0.30) \\ 0.10 &\neq 0.105 \end{aligned}$$

$\therefore$ Since the joint probability $P(X = 1, Y = 1)$ does not equal the product of the marginal probabilities $P(X = 1)P(Y = 1)$, the random variables X and Y are NOT independent.

Solution 13

Solution 13 (b)

Define Covariance and Correlation of X and Y

The covariance is as follows:

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

The correlation is as follows

$$\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

Compute Expectations X Y and XY

$$\begin{aligned}
E[X] &= \sum_x xP(X=x) = 1(0.35) + 2(0.30) + 3(0.35) = 0.35 + 0.60 + 1.05 = 2.0 \\
E[Y] &= \sum_y yP(Y=y) = 1(0.30) + 2(0.50) + 3(0.20) = 0.30 + 1.00 + 0.60 = 1.9 \\
E[XY] &= \sum_{x,y} xyP(X=x, Y=y) \\
&= (1)(1)(0.10) + (1)(2)(0.20) + (1)(3)(0.05) \\
&\quad + (2)(1)(0.10) + (2)(2)(0.15) + (2)(3)(0.05) \\
&\quad + (3)(1)(0.10) + (3)(2)(0.15) + (3)(3)(0.10) \\
&= 0.10 + 0.40 + 0.15 + 0.20 + 0.60 + 0.30 + 0.30 + 0.90 + 0.90 \\
&= 3.85
\end{aligned}$$

Compute Covariance

$$\text{Cov}[X,Y] = E[XY] - E[X]E[Y] = 3.85 - (2.0)(1.9) = 3.85 - 3.80 = 0.05$$

Compute Variances and Standard Deviations

$$\begin{aligned}
E[X^2] &= 1^2(0.35) + 2^2(0.30) + 3^2(0.35) = 0.35 + 1.20 + 3.15 = 4.7 \\
\text{Var}[X] &= E[X^2] - (E[X])^2 = 4.7 - 2.0^2 = 0.7 \\
\sigma_X &= \sqrt{0.7}
\end{aligned}$$

$$\begin{aligned}
E[Y^2] &= 1^2(0.30) + 2^2(0.50) + 3^2(0.20) = 0.30 + 2.00 + 1.80 = 4.1 \\
\text{Var}[Y] &= E[Y^2] - (E[Y])^2 = 4.1 - 1.9^2 = 4.1 - 3.61 = 0.49 \\
\sigma_Y &= \sqrt{0.49}
\end{aligned}$$

Compute Correlation Coefficient

$$\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\sigma_X \sigma_Y} = \frac{0.05}{\sqrt{0.7} \cdot \sqrt{0.49}} \approx 0.0854$$

$\therefore$ The covariance is $\text{Cov}[X,Y] = 0.05$, and the correlation coefficient is $\rho_{X,Y} \approx 0.0854$. This indicates a very weak positive linear relationship between X and Y.

Solution 14

Solution 14 (a)

The expectation of Y, $E[Y]$, via the linearity of expectation is defined as:

$$E[Y] = E[aX + b] = aE[X] + b$$

The covariance between X and Y is defined as:

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])]$$

Now,substitute the expression for $E[Y]$ into the expression for covariance:

$$\begin{aligned}
\text{Cov}[X, Y] &= E\left[(X - E[X])(Y - E[Y])\right] \\
&= E\left[(X - E[X])((aX + b) - (aE[X] + b))\right] \\
&= E\left[(X - E[X])(aX + b - aE[X] - b)\right] \\
&= E\left[(X - E[X])(aX - aE[X])\right] \\
&= E\left[(X - E[X])a(X - E[X])\right] \\
&= a \cdot E\left[(X - E[X])^2\right] \\
&= a \cdot \text{Var}[X]
\end{aligned}$$

$\therefore$ The covariance between X and Y ,in terms of X is the following:

$$\text{Cov}[X, Y] = a \cdot \text{Var}[X]$$

Solution 14

Solution 14 (b)

Part (b): Deriving the Correlation $\rho_{X,Y}$

The correlation coefficient $\rho_{X,Y}$ is defined as:

$$\rho_{X,Y} = \frac{\text{Cov}[X,Y]}{\text{SD}[X]\text{SD}[Y]}$$

The variance of Y is the following:

$$\text{Var}[Y] = \text{Var}[aX+b] = a^2\text{Var}[X]$$

First, we need find the standard deviation of Y:

$$\text{SD}[Y] = \sqrt{\text{Var}[Y]} = \sqrt{a^2\text{Var}[X]} = |a|\sqrt{\text{Var}[X]} = |a| \cdot \text{SD}[X]$$

Now, substitute the known quantities into the correlation formula:

$$\begin{aligned}\rho_{X,Y} &= \frac{\text{Cov}[X,Y]}{\text{SD}[X]\text{SD}[Y]} \\ &= \frac{a \cdot \text{Var}[X]}{\text{SD}[X] \cdot (|a| \cdot \text{SD}[X])} \\ &= \frac{a \cdot (\text{SD}[X])^2}{|a| \cdot (\text{SD}[X])^2} \\ &= \frac{a}{|a|}\end{aligned}$$

$\therefore$ The correlation between X and Y is $\rho_{X,Y} = \frac{a}{|a|}$, which is 1 if $a > 0$, -1 if $a < 0$, and undefined if $a = 0$.

Solution 15

Two random variables X and Y are uncorrelated if their covariance is zero. The covariance is defined as:

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

We first calculate the expected value of X, $E[X]$.

$$\begin{aligned}
E[X] &= \sum_{x\in\Omega_X} x \cdot P(X=x) \\
&= (-1)\cdot P(X=-1) + (0)\cdot P(X=0) + (1)\cdot P(X=1) \\
&= (-1)\cdot \frac{1}{3} + (0)\cdot\frac{1}{3} + (1)\cdot\frac{1}{3} \\
&= -\frac{1}{3} + 0 + \frac{1}{3} \\
&= 0
\end{aligned}$$

Since $E[X] = 0$, the covariance formula simplifies significantly:

$$\mathrm{Cov}[X,Y] = E[XY] - (0)\cdot E[Y] = E[XY]$$

Hence, for X and Y to be uncorrelated, we only need to find a function f such that:

$$E[Xf(X)] = 0$$

Next we expand the condition $E[Xf(X)] = 0$:

$$E[Xf(X)] = \sum_{x\in\Omega_X} xf(x)P(X=x) = \frac{1}{3}\left((-1)f(-1) + (0)f(0) + (1)f(1)\right) = 0$$

This implies that we need $-f(-1) + f(1) = 0$, or $f(-1) = f(1)$.

Let $f(x) = x^2$ and verify that $E[XY] = 0$.

$$\begin{aligned}
E[XY] &= E[X\cdot X^2] = E[X^3] \\
&= \sum_{x\in\Omega_X} x^3 \cdot P(X=x) \\
&= (-1)^3\cdot P(X=-1) + (0)^3\cdot P(X=0) + (1)^3\cdot P(X=1) \\
&= (-1)\cdot\frac{1}{3} + (0)\cdot\frac{1}{3} + (1)\cdot\frac{1}{3} \\
&= -\frac{1}{3} + 0 + \frac{1}{3} \\
&= 0
\end{aligned}$$

For $Y = f(X) = X^2$ we have $E[XY] = 0$ and $E[X] = 0$, and $\mathrm{Cov}[X,Y] = 0$. The variable $Y = X^2$ is determined by X, yet they are linearly uncorrelated.

$\therefore$ the function is $f(X) = X^2$