Solution 1

Solution 1 (a)

Step 1: Identify characteristics of the dataspace

We are given 10 dimensional vectors where each element can be any real number ($x_i \in \mathbb{R}$):

$\therefore$ we can express the dataspace χ as: $\chi = \mathbb{R}^{10}$

Solution 1 (b)

Step 1: Identify characteristics of the dataspace

We are given 3 dimensional vectors where each element is zero or one ($x_i \in [0, 1]$):

$\therefore$ we can express the dataspace χ as: $\chi = [0, 1]^3$

Solution 2

Solution 2 (a)

Step 1: Define Euclidean distance (ℓ_2)

$$\ell_2 = \|p - q\|_2 = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Step 2: Compute ℓ_2

Let $p = 1$ and $q = 10$

$$\begin{aligned}
\ell_2 &= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \\
\ell_2 &= \sqrt{\sum_{i=1}^{1}(1 - 10)^2} \\
\ell_2 &= \sqrt{(-9)^2} \\
\ell_2 &= 9
\end{aligned}$$

$\therefore \ell_2 = 9$

Solution 2 (b)

Step 1: Define Euclidean distance (ℓ_2)

$$\ell_2 = \|p - q\|_2 = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Step 2: Compute ℓ_2

Let $p = \begin{bmatrix} -1 \\ 12 \end{bmatrix}, q = \begin{bmatrix} 6 \\ -12 \end{bmatrix}$

$$\begin{aligned}
\ell_2 &= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \\
\ell_2 &= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \\
\ell_2 &= \sqrt{(-1 - 6)^2 + (12 - (-12))^2} \\
\ell_2 &= \sqrt{(-7)^2 + (24)^2} \\
\ell_2 &= \sqrt{625} \\
\ell_2 &= 25
\end{aligned}$$

$\therefore \ell_2 = 25$

Solution 2

Solution 2 (c)

Step 1: Define Euclidean distance (ℓ_2)

$$\ell_2 = \|p - q\|_2 = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Step 2: Compute ℓ_2

Let $p = \begin{bmatrix} 1 \\ 5 \\ -1 \end{bmatrix}, q = \begin{bmatrix} 5 \\ 2 \\ 11 \end{bmatrix}$

$$\begin{aligned}
\ell_2 &= \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \\
\ell_2 &= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2} \\
\ell_2 &= \sqrt{(1-5)^2 + (5-2)^2 + (-1-11)^2} \\
\ell_2 &= \sqrt{(-4)^2 + (3)^2 + (-12)^2} \\
\ell_2 &= \sqrt{169} \\
\ell_2 &= 13
\end{aligned}$$

$\therefore \ell_2 = 13$

Solution 3

Solution 3 (a)

Step 1: Normalize the vector x

Let $x = \begin{bmatrix} 10 \\ 15 \\ 25 \end{bmatrix}$

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 10 + 15 + 25 = 50$$

Now, divide each entry by the total sum:

$$p = \frac{1}{50} \cdot x = \frac{1}{50} \begin{bmatrix} 10 \\ 15 \\ 25 \end{bmatrix} = \begin{bmatrix} 10/50 \\ 15/50 \\ 25/50 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$$

$\therefore$ the result (p) of scaling vertor x is the following:

$$p = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$$

Solution 3 (b)

Step 1: Define dimension of the probability simplex

The dimension of vector p is 3 and $k = n - 1$ where k is the dimension of the probability simplex

$\therefore$ vector p lies in the probability simplex(Δ_2) for $k = 2$

Solution 4

Step 1: Understand what scaling means and the constraints of Δ_2

Scaling a vector means multiplying all entries by the same positive constant $c > 0$.
The probability simplex Δ_2 is defined as:

$$\Delta_2 = \left\{ x \in \mathbb{R}^3 : x_i \geq 0 \text{ for all } i = 1, 2, 3, \text{ and } \sum_{i=1}^{3} x_i = 1 \right\}$$

For a point to be scalable to Δ_2, after scaling it must satisfy:

- All components must be non-negative
- The sum of components must equal 1

Step 2: Find a point that violates the constraints

Since we need a 2-dimensional point, let's consider $x = \begin{bmatrix} a \\ b \end{bmatrix}$ where we interpret this as a 3-dimensional vector $\begin{bmatrix} a \\ b \\ 0 \end{bmatrix}$.
For any point with a negative component, scaling cannot make it non-negative.
Example: Let $x = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$
If we scale by any $c > 0$: $cx = \begin{bmatrix} c \\ -2c \end{bmatrix}$
The second component $-2c < 0$ for any $c > 0$, so this scaled vector cannot satisfy the non-negativity constraint of Δ_2.

$\therefore$ Final Answer Example: $x = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ (or any 2D point with at least one negative component)

Graduate Level Explanation The key insight is that scaling preserves the sign of each component - it cannot transform negative values to positive ones. Since the probability simplex requires all components to be non-negative, any vector containing negative components cannot be scaled to lie within it. This geometric constraint is fundamental in applications like topic modeling or mixture models, where negative probabilities are meaningless. The orthant restriction (non-negativity) combined with the sum constraint defines the simplex as a bounded convex polytope, and scaling operations represent rays from the origin that either intersect this polytope or miss it entirely.

Explanation for 5 year old Imagine you have a recipe with ingredients, but one of the "amounts" is negative - like "negative 2 cups of flour." No matter how much you shrink or grow the recipe (scaling), you'll still need a negative amount of flour, which doesn't make sense! The probability simplex is like a rule that says "all ingredients must be positive amounts," so any recipe with negative ingredients can never follow this rule, no matter how you scale it.

Solution 5

Step 1: Define Δ_3 and identify key properties

The probability simplex Δ_3 is defined as:

$$\Delta_3 = \left\{ x \in \mathbb{R}^4 : x_i \geq 0 \text{ for all } i = 1, 2, 3, 4, \text{ and } \sum_{i=1}^{4} x_i = 1 \right\}$$

This is a 3-dimensional simplex (tetrahedron) embedded in 4-dimensional space. However, due to the constraint $\sum x_i = 1$, we can visualize it in 3D by using three coordinates and letting the fourth be determined by the constraint.

The given points are the vertices of the simplex:

- $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ corresponds to $\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ in 4D
- $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ corresponds to $\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ in 4D
- $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ corresponds to $\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ in 4D

Step 2: Determine coordinates and sketch description

The fourth vertex (not shown) would be $\begin{bmatrix}0\\0\\0\\1\end{bmatrix}$, which corresponds to $\begin{bmatrix}0\\0\\0\end{bmatrix}$ in our 3D representation (since the fourth coordinate is $1-0-0-0=1$).
The most central point (centroid) of the simplex is:

$$\text{centroid} = \frac{1}{4}\left(\begin{bmatrix}1\\0\\0\\0\end{bmatrix}+\begin{bmatrix}0\\1\\0\\0\end{bmatrix}+\begin{bmatrix}0\\0\\1\\0\end{bmatrix}+\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\right)=\begin{bmatrix}1/4\\1/4\\1/4\\1/4\end{bmatrix}$$

In 3D coordinates, this is $\begin{bmatrix}1/4\\1/4\\1/4\end{bmatrix}$.

Sketch Description:

- Draw 3D coordinate axes labeled x_1, x_2, x_3
- The simplex is a triangular region (actually a tetrahedron) with vertices at:
  - $(1,0,0)$ on the x_1-axis
  - $(0,1,0)$ on the x_2-axis
  - $(0,0,1)$ on the x_3-axis
  - $(0,0,0)$ at the origin (representing the fourth vertex)
- Connect these four points to form a tetrahedron
- Mark the centroid at $(1/4,1/4,1/4)$

$\therefore$ Final Answer The most central point in Δ_3 has coordinates: $\begin{bmatrix}1/4\\1/4\\1/4\end{bmatrix}$ (or $\begin{bmatrix}0.25\\0.25\\0.25\end{bmatrix}$)

Note: The fourth coordinate is $1-1/4-1/4-1/4=1/4$

Graduate Level Explanation The 3-simplex Δ_3 is a regular tetrahedron when embedded properly in 3D space. Each vertex represents a pure probability distribution (all probability mass on one outcome), while interior points represent mixed distributions. The centroid represents the uniform distribution over 4 outcomes. This geometric structure is fundamental in Bayesian statistics, where it represents the space of all possible probability distributions over 4 categories. The simplex's convex hull property ensures that convex combinations of probability distributions remain valid probability distributions, making it a natural space for optimization in machine learning algorithms like EM or variational inference.

Explanation for 5 year old Imagine a pyramid made of triangles (like the ones in Egypt, but pointier)! Each corner of the pyramid represents putting all your marbles in one basket - like having 100% chocolate ice cream, or 100% vanilla, or 100% strawberry, or 100% mint. The very center of the pyramid is where you have equal amounts of all four flavors - 25% of each! Any point inside the pyramid represents some mix of the four flavors that adds up to 100%.

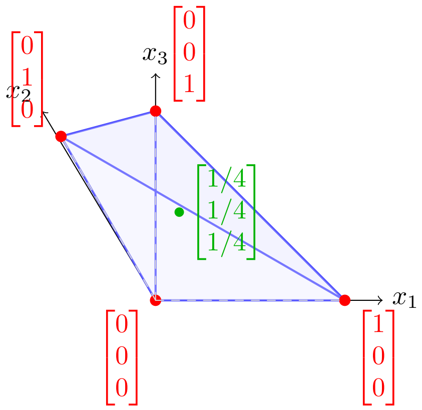

Figure 1: The probability simplex Δ_3 showing vertices and centroid

Solution 6

Step 1: Recall the formulas for ℓ_1 distance and KL divergence

IDK.. LOG IS BASE e NOT 2 The ℓ_1 distance (Manhattan distance) between two vectors u and v is:

$$\|u - v\|_1 = \sum_{i=1}^{n} |u_i - v_i|$$

The KL divergence between two probability distributions P and Q is:

$$KL(P\|Q) = \sum_{i=1}^{n} P_i \log\left(\frac{P_i}{Q_i}\right)$$

where we use the convention that $0\log(0/Q_i) = 0$ and $P_i \log(P_i/0) = \infty$ if $P_i > 0$.

Step 2: Calculate each distance and divergence

Part (i): $\|p-q\|_1$ where $p = \begin{bmatrix} 1/2 \\ 1/4 \\ 1/8 \\ 1/8 \end{bmatrix}$ and $q = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$

$$p - q = \begin{bmatrix} 1/2 - 1/4 \\ 1/4 - 1/4 \\ 1/8 - 1/4 \\ 1/8 - 1/4 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 0 \\ -1/8 \\ -1/8 \end{bmatrix}$$

$$\|p-q\|_1 = |1/4| + |0| + |-1/8| + |-1/8| = 1/4 + 0 + 1/8 + 1/8 = 1/4 + 1/4 = 1/2$$

Part (ii): $\|q-r\|_1$ where $q = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$ and $r = \begin{bmatrix} 1/2 \\ 0 \\ 1/4 \\ 1/4 \end{bmatrix}$

$$q - r = \begin{bmatrix} 1/4 - 1/2 \\ 1/4 - 0 \\ 1/4 - 1/4 \\ 1/4 - 1/4 \end{bmatrix} = \begin{bmatrix} -1/4 \\ 1/4 \\ 0 \\ 0 \end{bmatrix}$$

$$\|q-r\|_1 = |-1/4| + |1/4| + |0| + |0| = 1/4 + 1/4 + 0 + 0 = 1/2$$

Part (iii): $KL(p\|q)$

$$KL(p\|q) = \sum_{i=1}^{4} p_i \log\left(\frac{p_i}{q_i}\right)$$

$$= \frac{1}{2}\log\left(\frac{1/2}{1/4}\right) + \frac{1}{4}\log\left(\frac{1/4}{1/4}\right) + \frac{1}{8}\log\left(\frac{1/8}{1/4}\right) + \frac{1}{8}\log\left(\frac{1/8}{1/4}\right)$$

$$= \frac{1}{2}\log(2) + \frac{1}{4}\log(1) + \frac{1}{8}\log(1/2) + \frac{1}{8}\log(1/2)$$

$$= \frac{1}{2}\log(2) + 0 + \frac{1}{8}(-\log(2)) + \frac{1}{8}(-\log(2))$$

$$= \frac{1}{2}\log(2) - \frac{1}{4}\log(2) = \frac{1}{4}\log(2)$$

Part (iv): $KL(q\|r)$

$$KL(q\|r) = \sum_{i=1}^{4} q_i \log\left(\frac{q_i}{r_i}\right)$$

$$= \frac{1}{4}\log\left(\frac{1/4}{1/2}\right) + \frac{1}{4}\log\left(\frac{1/4}{0}\right) + \frac{1}{4}\log\left(\frac{1/4}{1/4}\right) + \frac{1}{4}\log\left(\frac{1/4}{1/4}\right)$$

Since $r_2 = 0$ and $q_2 = 1/4 > 0$, we have $\log(q_2/r_2) = \log(1/4/0) = +\infty$.
Therefore: $KL(q\|r) = +\infty$

$\therefore$ Final Answers

1. $\|p-q\|_1 = \frac{1}{2}$
2. $\|q-r\|_1 = \frac{1}{2}$
3. $KL(p\|q) = \frac{1}{4}\log(2) \approx 0.173$
4. $KL(q\|r) = +\infty$

Graduate Level Explanation The ℓ_1 distance is symmetric and satisfies the triangle inequality, making it a proper metric on the probability simplex. Interestingly, both pairs have the same ℓ_1 distance despite having different structures. The KL divergence, however, is asymmetric and not a true metric. $KL(p\|q)$ is finite because q has full support (no zero entries), but $KL(q\|r) = \infty$ because r has a zero entry where q has positive probability. This illustrates a fundamental property: KL divergence from a distribution with full support to one with restricted support is infinite, making it useful for detecting when probability mass is assigned to impossible events in the reference distribution.

Explanation for 5 year old The ℓ_1 distance is like counting how much you need to move marbles between jars to make them the same - it's always the same no matter which direction you go. But KL divergence is like asking "how surprised would I be?" If one jar is completely empty but you expected marbles there, you'd be infinitely surprised! That's why one answer is infinity - it's like expecting something that's impossible.

Solution 7

Part a: Dimensionality for each of the representations (raw pixel, HoG, VGG-last-fc, VGG-last-conv)

Feature Type	Dimensionality
Raw Pixel	3072
HoG	512
VGG-last-fc	4096
VGG-last-conv	512

Part b: Test accuracies for 1-nearest neighbor classification using the various representations (raw pixel, HoG, VGG-last-fc, VGG-last-conv, random-VGG-last-fc, random-VGG-last-conv).

Feature Type	1-NN test accuracy (%)
Raw Pixel	35.4
HoG	36.6
VGG-last-fc	92.1
VGG-last-conv	92.0
random VGG-last-fc	39.1
random VGG-last-conv	40.6

Solution 7

Part c: Raw Pixel correct/incorrect

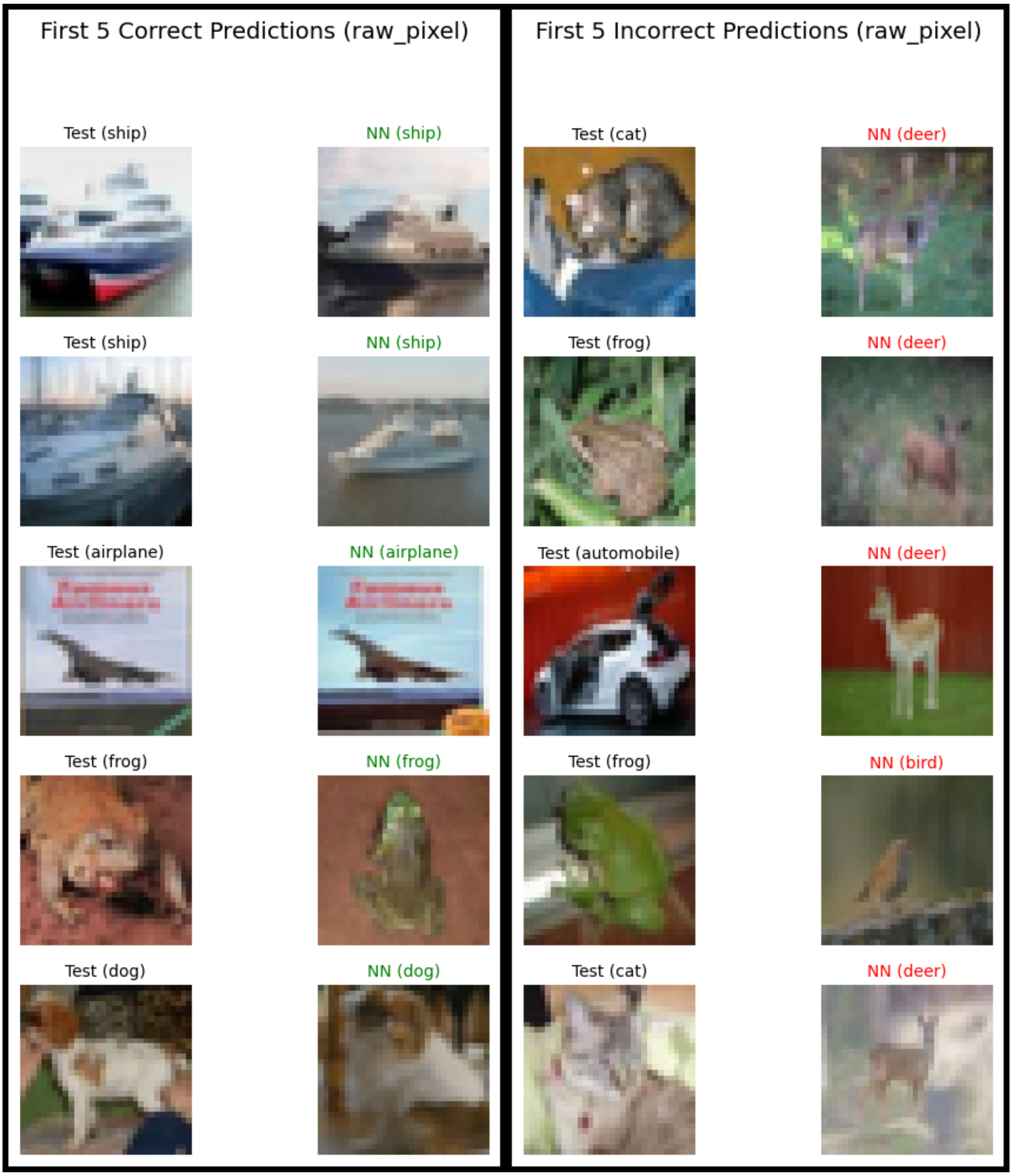

Figure 2: First five correct/incorrect images for Raw Pixel

Solution 7

Part c: HoG correct/incorrect

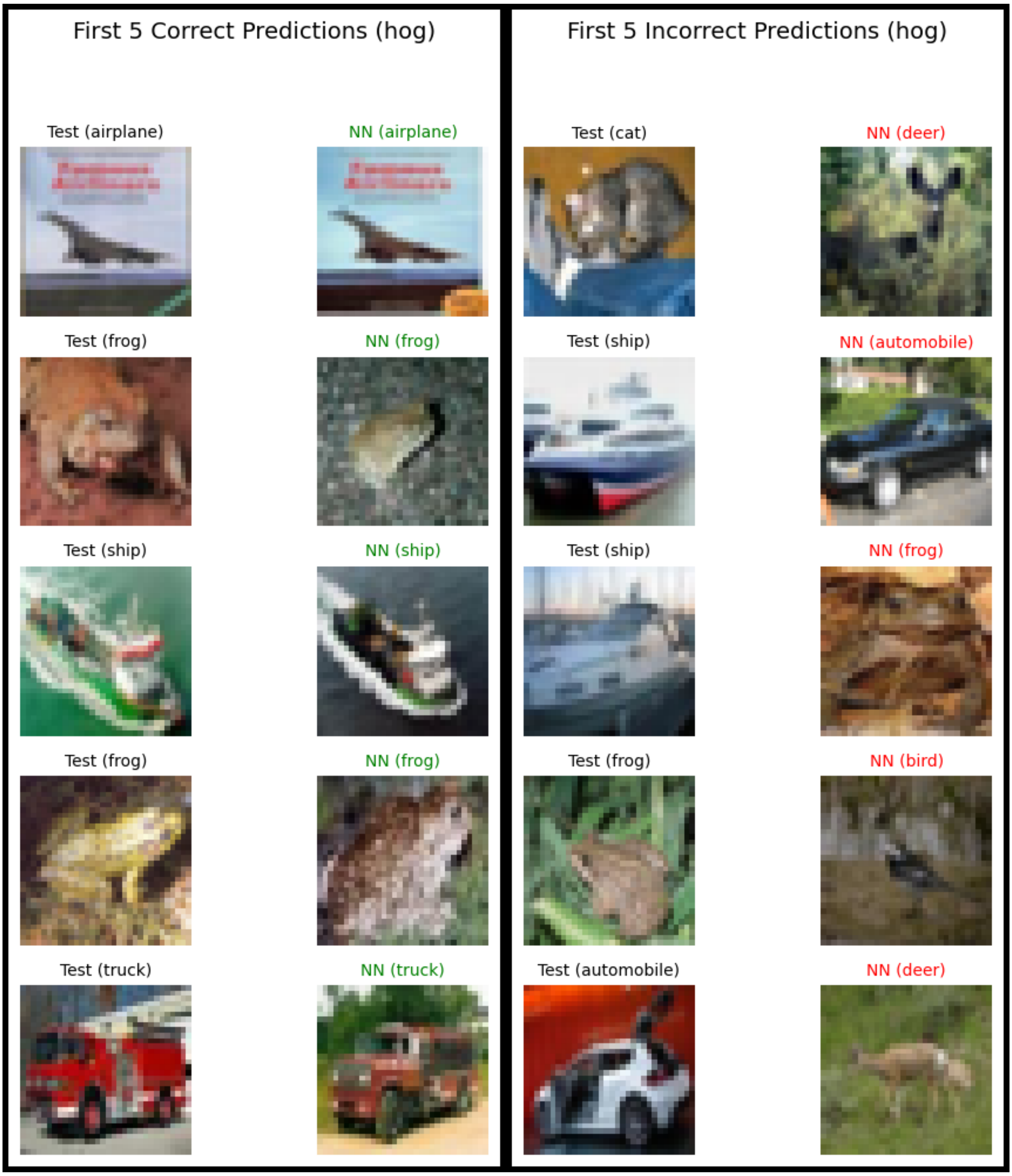

Figure 3: First five correct/incorrect images for HoG

Solution 7

Part c: VGG-last-fc correct/incorrect

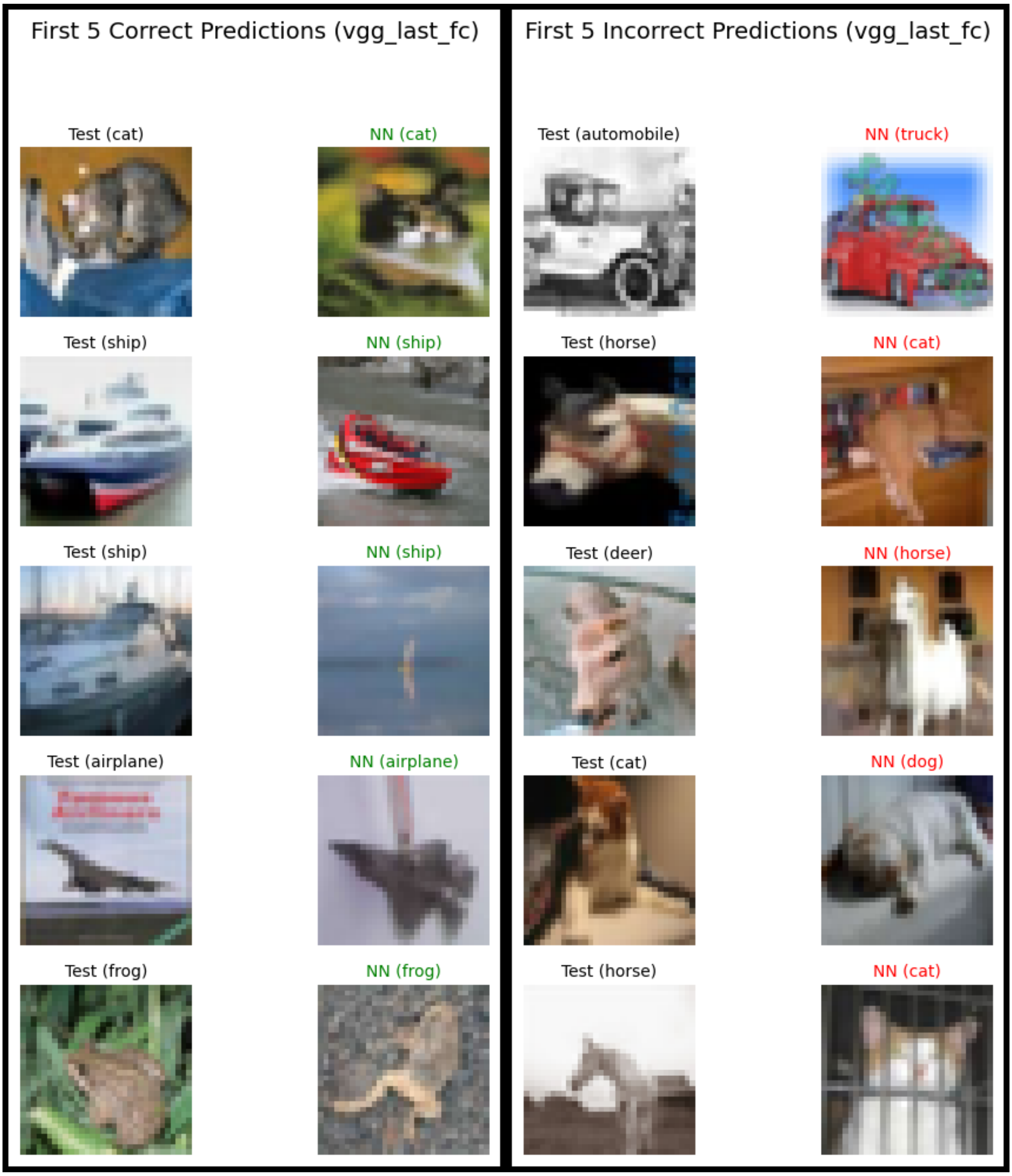

Figure 4: First five correct/incorrect images for VGG-last-fc

Solution 8

Word Vectors: 5 closest words

Target Word	Five Closest Words
communism	['fascism', 'capitalism', 'nazism', 'stalinism', 'socialism']
africa	['african', 'continent', 'south', 'africans', 'zimbabwe']
happy	['glad', 'pleased', 'always', 'everyone', 'sure']
sad	['sorry', 'tragic', 'happy', 'pathetic', 'awful']
upset	['upsetting', 'surprised', 'upsets', 'stunned', 'shocked']
computer	['computers', 'software', 'technology', 'laptop', 'computing']
cat	['cats', 'dog', 'pet', 'feline', 'dogs']
dollar	['currency', 'dollars', 'euro', 'multibillion', 'weaker']

Solution 1: Scalability on the Probability Simplex

Step 1: Define the Probability Simplex Δ_2

The probability simplex Δ_k is the set of all k-dimensional vectors with non-negative components that sum to 1. For $k = 2$, a vector $p = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$ is in Δ_2 if and only if it satisfies two conditions:

1. **Non-negativity:** $p_1 \geq 0$ and $p_2 \geq 0$.
2. **Sum-to-one:** $p_1 + p_2 = 1$.

The question asks if for any vector $p \in \mathbb{R}^2$ and scalar $c > 0$, the condition $c \cdot p \in \Delta_2$ implies that $p \in \Delta_2$.

Step 2: Analyze the Constraints under Scaling

Let $q = c \cdot p = \begin{bmatrix} cp_1 \\ cp_2 \end{bmatrix}$. We are given that $q \in \Delta_2$.

- **Non-negativity:** Since $c > 0$ and we are given $cp_1 \geq 0$ and $cp_2 \geq 0$, it must be that $p_1 \geq 0$ and $p_2 \geq 0$. This condition is satisfied for p.
- **Sum-to-one:** We are given that the components of q sum to 1: $cp_1 + cp_2 = 1$. Factoring out c, we get $c(p_1 + p_2) = 1$, which implies $p_1 + p_2 = \frac{1}{c}$.

For p to be in Δ_2, its components must sum to 1, i.e., $p_1 + p_2 = 1$. This only holds if $\frac{1}{c} = 1$, which means $c = 1$. Since the statement must hold for any $c > 0$, we can find a counterexample by choosing $c \neq 1$.

Step 3: Construct a Counterexample

Let $c = 2$. Choose a point $q \in \Delta_2$, for example, $q = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$. If $c \cdot p = q$, then $p = \frac{1}{c}q = \frac{1}{2}\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}$. Let's check if this p is in Δ_2:

- **Non-negativity:** $p_1 = 0.25 \geq 0$ and $p_2 = 0.25 \geq 0$. (Satisfied)
- **Sum-to-one:** $p_1 + p_2 = 0.25 + 0.25 = 0.5 \neq 1$. (Not satisfied)

Since p does not satisfy the sum-to-one constraint, $p \notin \Delta_2$. Thus, the statement is false.

$\therefore$ Final Answer
The statement is **false**. A counterexample is $p = \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}$ and $c = 2$. Here, $c \cdot p = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \in \Delta_2$, but $p \notin \Delta_2$ because its components sum to 0.5, not 1.

Graduate Level Explanation
The probability simplex Δ_k is an affine subspace of $\mathbb{R}^k$, specifically the intersection of the hyperplane $\sum x_i = 1$ and the non-negative orthant $\mathbb{R}^k_+$. While the non-negative orthant is a convex cone (closed under non-negative scalar multiplication), the hyperplane $\sum x_i = 1$ is not a linear subspace as it does not contain the origin. Scaling a vector p by $c \neq 1$ moves it off this hyperplane, thus violating the sum-to-one constraint. The set of vectors whose scaled versions lie on the simplex forms a cone over the simplex, but these vectors are not, in general, on the simplex themselves.

Explanation for a 5 year old
Imagine a recipe for one special juice drink says you need 1 cup of ingredients in total. This "1 cup total" rule is very important. You find a bottle of juice that follows the rule. Your friend says, "I have a different bottle, and if I pour out half of it, it's exactly the same as your juice." Your friend's bottle might follow the non-negativity rule (it has juice in it), but it must have had 2 cups of ingredients to begin with. So, your friend's original bottle did not follow the "1 cup total" rule.

Solution 2: Sketching the Probability Simplex Δ_3

Step 1: Define the Geometry of Δ_3

The probability simplex Δ_3 is the set of points $p = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix}^\top$ in $\mathbb{R}^3$ satisfying:

1. $p_1 \geq 0, p_2 \geq 0, p_3 \geq 0$ (it lies in the first octant).
2. $p_1 + p_2 + p_3 = 1$ (it lies on a plane).

The intersection of the plane $p_1 + p_2 + p_3 = 1$ with the first octant forms a bounded, closed shape. To identify the shape, we find its vertices.

Step 2: Identify the Vertices and the Central Point

The vertices of the shape are the points where the plane intersects the coordinate axes.

- Intersection with p_1-axis ($p_2 = 0, p_3 = 0$): $p_1 = 1$. Vertex $v_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^\top$.
- Intersection with p_2-axis ($p_1 = 0, p_3 = 0$): $p_2 = 1$. Vertex $v_2 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\top$.
- Intersection with p_3-axis ($p_1 = 0, p_2 = 0$): $p_3 = 1$. Vertex $v_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\top$.

Connecting these three vertices in 3D space forms an equilateral triangle. The "most central" point of this triangle is its barycenter (or centroid), which is the average of the coordinates of its vertices.

Step 3: Calculate the Centroid

The coordinates of the centroid p_c are:

$$p_c = \frac{v_1 + v_2 + v_3}{3} = \frac{1}{3}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

This point corresponds to the uniform probability distribution over three outcomes.

$\therefore$ Final Answer

Sketch Description: The probability simplex Δ_3 is an equilateral triangle in 3D space whose vertices are at the standard basis vectors $\begin{bmatrix}1, 0, 0\end{bmatrix}^\top$, $\begin{bmatrix}0, 1, 0\end{bmatrix}^\top$, and $\begin{bmatrix}0, 0, 1\end{bmatrix}^\top$.

Most Central Point: The coordinates of the most central point (the barycenter) are $\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}^\top$.

Graduate Level Explanation

The standard k-simplex, Δ_k, is a $(k-1)$-dimensional convex polytope embedded in $\mathbb{R}^k$. For $k = 3$, this results in a 2-dimensional triangle. The vertices are the standard basis vectors e_1, e_2, e_3, representing deterministic probability distributions. The barycenter of the simplex, $(1/k, \ldots, 1/k)^\top$, corresponds to the uniform probability distribution. In information theory, this is the distribution with the maximum Shannon entropy, representing the state of maximum uncertainty.

Explanation for a 5 year old

Imagine a big glass cube. Now, imagine you slice it with a flat piece of glass. The slice starts at the number 1 on the 'x' line, goes to the number 1 on the 'y' line, and also to the number 1 on the 'z' line. The shape of this flat slice inside the corner of the cube is a perfect triangle. The very middle of that triangle is its balancing point. That special point is at $(1/3, 1/3, 1/3)$, which means it's an equal distance from all three number lines.

Solution 3: ℓ_1 Distance and KL Divergence

Step 1: Calculate the ℓ_1 Distance

The ℓ_1 distance between two vectors $p, q \in \mathbb{R}^n$ is given by $\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i|$. For $p = \begin{bmatrix}1/2 & 1/4 & 1/8 & 1/8\end{bmatrix}^\top$ and $q = \begin{bmatrix}1/4 & 1/4 & 1/4 & 1/4\end{bmatrix}^\top$:

$$\begin{aligned}
\|p - q\|_1 &= \left|\frac{1}{2} - \frac{1}{4}\right| + \left|\frac{1}{4} - \frac{1}{4}\right| + \left|\frac{1}{8} - \frac{1}{4}\right| + \left|\frac{1}{8} - \frac{1}{4}\right| \\
&= \left|\frac{2}{4} - \frac{1}{4}\right| + |0| + \left|\frac{1}{8} - \frac{2}{8}\right| + \left|\frac{1}{8} - \frac{2}{8}\right| \\
&= \frac{1}{4} + 0 + \left|-\frac{1}{8}\right| + \left|-\frac{1}{8}\right| \\
&= \frac{1}{4} + \frac{1}{8} + \frac{1}{8} = \frac{2}{8} + \frac{1}{8} + \frac{1}{8} = \frac{4}{8} = \frac{1}{2}
\end{aligned}$$

Step 2: Calculate the KL Divergence $K(p, q)$

The Kullback-Leibler (KL) divergence from q to p is $K(p, q) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}$.

$$\begin{aligned}
K(p, q) &= p_1 \ln\left(\frac{p_1}{q_1}\right) + p_2 \ln\left(\frac{p_2}{q_2}\right) + p_3 \ln\left(\frac{p_3}{q_3}\right) + p_4 \ln\left(\frac{p_4}{q_4}\right) \\
&= \frac{1}{2} \ln\left(\frac{1/2}{1/4}\right) + \frac{1}{4} \ln\left(\frac{1/4}{1/4}\right) + \frac{1}{8} \ln\left(\frac{1/8}{1/4}\right) + \frac{1}{8} \ln\left(\frac{1/8}{1/4}\right) \\
&= \frac{1}{2} \ln(2) + \frac{1}{4} \ln(1) + \frac{1}{8} \ln\left(\frac{1}{2}\right) + \frac{1}{8} \ln\left(\frac{1}{2}\right) \\
&= \frac{1}{2} \ln(2) + 0 - \frac{1}{8} \ln(2) - \frac{1}{8} \ln(2) \\
&= \left(\frac{1}{2} - \frac{1}{8} - \frac{1}{8}\right) \ln(2) = \left(\frac{4}{8} - \frac{2}{8}\right) \ln(2) = \frac{2}{8} \ln(2) = \frac{1}{4} \ln(2)
\end{aligned}$$

Step 3: Calculate the KL Divergence $K(q, p)$

The KL divergence from p to q is $K(q, p) = \sum_{i=1}^n q_i \ln \frac{q_i}{p_i}$.

$$\begin{aligned}
K(q, p) &= q_1 \ln\left(\frac{q_1}{p_1}\right) + q_2 \ln\left(\frac{q_2}{p_2}\right) + q_3 \ln\left(\frac{q_3}{p_3}\right) + q_4 \ln\left(\frac{q_4}{p_4}\right) \\
&= \frac{1}{4} \ln\left(\frac{1/4}{1/2}\right) + \frac{1}{4} \ln\left(\frac{1/4}{1/4}\right) + \frac{1}{4} \ln\left(\frac{1/4}{1/8}\right) + \frac{1}{4} \ln\left(\frac{1/4}{1/8}\right) \\
&= \frac{1}{4} \ln\left(\frac{1}{2}\right) + \frac{1}{4} \ln(1) + \frac{1}{4} \ln(2) + \frac{1}{4} \ln(2) \\
&= -\frac{1}{4} \ln(2) + 0 + \frac{1}{4} \ln(2) + \frac{1}{4} \ln(2) \\
&= \frac{1}{4} \ln(2)
\end{aligned}$$

$\therefore$ Final Answer
For the given probability distributions p and q:

- The ℓ_1 distance is $\|p - q\|_1 = \mathbf{\frac{1}{2}}$.
- The KL divergence from q to p is $K(p, q) = \mathbf{\frac{1}{4}} \ln(\mathbf{2})$.
- The KL divergence from p to q is $K(q, p) = \mathbf{\frac{1}{4}} \ln(\mathbf{2})$.

Graduate Level Explanation
The ℓ_1 distance is a true metric satisfying symmetry and the triangle inequality; on the probability simplex, it is equivalent to twice the total variation distance. The Kullback-Leibler divergence, conversely, is not a metric. It is asymmetric ($K(p,q) \neq K(q,p)$ in general, although they coincide in this specific case) and does not satisfy the triangle inequality. It is a Bregman divergence generated by the negative entropy function, and it quantifies the expected inefficiency (in terms of information) of using a code optimized for distribution q to encode data from the true distribution p. By Gibbs' inequality, $K(p,q) \geq 0$ with equality if and only if $p = q$.

Explanation for a 5 year old
L1 Distance: Imagine you have two towers built from 4 kinds of colored blocks. Tower P has 4 red, 2 blue, 1 green, 1 yellow. Tower Q has 2 red, 2 blue, 2 green, 2 yellow. The "distance" is how many blocks you have to move to make Tower P look exactly like Tower Q. You need to take 2 red blocks away and add 1 green and 1 yellow. That's 4 moves in total. Our math gives an answer of 1/2, which is like a grown-up way of counting this.
KL Divergence: This is like a guessing game. Your bag of marbles has the colors mixed like in Tower P. Your friend thinks the colors are mixed like in Tower Q. The KL number measures how "surprised" your friend will be, on average, each time they pull a marble from your bag. A bigger number means more surprise! It's not usually the same amount of surprise as if you pulled from their bag, but for these special towers, it happens to be the same.