High-Dimensional Gaussians [20 pts]

In this question we will investigate how our intuition for samples from a Gaussian may break down in higher dimensions. Consider samples from a D-dimensional unit Gaussian

$x \sim \mathcal{N}(0_D, I_D)$ where~0_D indicates a column vector of~D zeros and~I_D is a $D \times D$ identity matrix.

Distance of Gaussian samples from origin

Starting with the definition of Euclidean norm, quickly show that the distance of x from the origin is $\sqrt{x^\intercal x}$

Distribution of distances of Gaussian samples from origin

In low-dimensions our intuition tells us that samples from the unit Gaussian will be near the origin.

1. Draw 10000 samples from a $D = 1$ Gaussian
2. Compute the distance of those samples from the origin.
3. Plot a normalized histogram for the distance of those samples from the origin.

Does this confirm your intuition that the samples will be near the origin?

```
• Enter cell code...
```

```
syntax: invalid interpolation syntax: "$ "
```

1. **top-level scope** @ *none:1*

Plot samples from distribution of distances

1. Draw a set of 10000 samples from $D = \{1, 2, 3, 10, 100\}$ Gaussians
2. Compute the distance of each sample from the origin
3. With all D dimensionality on a single plot, show the normalized histograms for the distribution of distance of those samples from the origin.

As the dimensionality of the Gaussian increases, what can you say about the expected distance of the samples from the Gaussian's mean (in this case, origin)?

Plot the χ-distribution

From Wikipedia, if x_i are k independent, normally distributed random variables with means μ_i and standard deviations σ_i then the statistic $Y = \sqrt{\sum_{i=1}^{k}(\frac{x_i-\mu_i}{\sigma_i})^2}$ is distributed according to the **χ-distribution**]

On the previous normalized histogram, plot the probability density function (pdf) of the χ-distribution for $k = \{1, 2, 3, 10, 100\}$.

```
• Enter cell code...
```

Distribution of distance between samples

Taking two samples from the D-dimensional unit Gaussian, $x_a, x_b \sim \mathcal{N}(0_D, I_D)$ how is $x_a - x_b$ distributed? Using the above result about χ-distribution, derive how $||x_a - x_b||_2$ is distributed.

(Hint: start with a $\mathcal{X}$-distributed random variable and use the **change of variables formula**.)

```
• Enter cell code...
```

Plot pdfs of distribution distances between samples

For for $D = \{1, 2, 3, 10, 100\}$. How does the distance between samples from a Gaussian behave as dimensionality increases? Confirm this by drawing two sets of 1000 samples from the D-dimensional unit Gaussian. On the plot of the χ-distribution pdfs, plot the normalized histogram of the distance between samples from the first and second set.

```
• Enter cell code...
```

Linear interpolation between samples

Given two samples from a gaussian $x_a, x_b \sim \mathcal{N}(0_D, I_D)$ the linear interpolation between them x_α is defined as a function of $\alpha \in [0, 1]$

$$\text{lin\_interp}(\alpha, x_a, x_b) = \alpha x_a + (1 - \alpha)x_b$$

For two sets of 1000 samples from the unit gaussian in D-dimensions, plot the average log-likelihood along the linear interpolations between the pairs of samples as a function of α.

(i.e. for each pair of samples compute the log-likelihood along a linear space of interpolated points between them, $\mathcal{N}(x_\alpha|0, I)$ for $\alpha \in [0, 1]$. Plot the average log-likelihood over all the interpolations.)

Do this for $D = \{1, 2, 3, 10, 100\}$, one plot per dimensionality. Comment on the log-likelihood under the unit Gaussian of points along the linear interpolation. Is a higher log-likelihood for the

interpolated points necessarily better? Given this, is it a good idea to linearly interpolate between samples from a high dimensional Gaussian?

Polar Interpolation Between Samples

Instead we can interpolate in polar coordinates: For $\alpha \in [0, 1]$ the polar interpolation is

$$\text{polar\_interp}(\alpha, x_a, x_b) = \sqrt{\alpha} x_a + \sqrt{(1-\alpha)} x_b$$

This interpolates between two points while maintaining Euclidean norm.

On the same plot from the previous question, plot the probability density of the polar interpolation between pairs of samples from two sets of 1000 samples from D-dimensional unit Gaussians for $D = \{1, 2, 3, 10, 100\}$.

Comment on the log-likelihood under the unit Gaussian of points along the polar interpolation. Give an intuitive explanation for why polar interpolation is more suitable than linear interpolation for high dimensional Gaussians. (For 6. and 7. you should have one plot for each D with two curves on each).

```
• Enter cell code...
```

Norm along interpolation

In the previous two questions we compute the average log-likelihood of the linear and polar interpolations under the unit gaussian. Instead, consider the norm along the interpolation, $\sqrt{x_\alpha^\intercal x_\alpha}$. As we saw previously, this is distributed according to the $\mathcal{X}$-distribution. Compute and plot the average log-likelihood of the norm along the two interpolations under the the $\mathcal{X}$-distribution for $D = \{1, 2, 3, 10, 100\}$, i.e. $\mathcal{X}_D(\sqrt{x_\alpha^\intercal x_\alpha})$. There should be one plot for each D, each with two curves corresponding to log-likelihood of linear and polar interpolations. How does the log-likelihood along the linear interpolation compare to the log-likelihood of the true samples (endpoints)?

```
• Enter cell code...
```