Eric Elmoznino
eric.elmoznino@mail.utoronto.ca
CS281-F15

CSC412/2056 Assignment #1

Due: Thursday February 9, 11:59pm

Collaborators:

Problem 1 (Variance and covariance, 6 points)
Let X and Y be two continuous independent random variables.

(a) Starting from the definition of independence, show that the independence of X and Y implies that their covariance is zero.

(b) For a scalar constant a, show the following two properties, starting from the definition of expectation:

$$\mathbb{E}(X+aY) = \mathbb{E}(X) + a\mathbb{E}(Y)$$
$$\text{var}(X+aY) = \text{var}(X) + a^2\text{var}(Y)$$

(a)

$$\begin{aligned}
\text{cov}(X,Y) &= \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) \\
&= \mathbb{E}(XY - X\mathbb{E}(Y) - Y\mathbb{E}(X) + \mathbb{E}(X)\mathbb{E}(Y)) \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \iint_{-\infty}^{\infty} xyp(x,y)dxdy - \int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} yp(y)dy \\
&= \iint_{-\infty}^{\infty} xyp(x)p(y)dxdy - \iint_{-\infty}^{\infty} xyp(x)p(y)dxdy \\
&= 0
\end{aligned}$$

(b)

$$\begin{aligned}
\mathbb{E}(X+aY) &= \iint_{-\infty}^{\infty} (x+ay)p(x,y)dxdy \\
&= \iint_{-\infty}^{\infty} (xp(x,y) + ayp(x,y))dxdy \\
&= \iint_{-\infty}^{\infty} xp(x)p(y)dxdy + a\iint_{-\infty}^{\infty} yp(y)p(x)dydx \\
&= E(X)\int_{-\infty}^{\infty} p(y)dy + aE(Y)\int_{-\infty}^{\infty} p(x)dx \\
&= \mathbb{E}(x) + a\mathbb{E}(X+aY)
\end{aligned}$$

$$\begin{aligned}
\text{var}(X+aY) &= \mathbb{E}\left(((X+aY) - \mathbb{E}(X+aY))^2\right) \\
&= \mathbb{E}\left(X + aY - \mathbb{E}(X) - a\mathbb{E}(Y))^2\right) \\
&= \iint_{-\infty}^{\infty} ((x-\mathbb{E}(X)) + a(Y-\mathbb{E}(Y)))^2 p(x)p(y)dxdy \\
&= \iint_{-\infty}^{\infty} ((x-\mathbb{E}(X))^2 p(x)p(y)dxdy + a^2\iint_{-\infty}^{\infty} ((y-\mathbb{E}(Y))^2 p(x)p(y)dxdy \\
&\quad + 2a\iint_{-\infty}^{\infty} (x-\mathbb{E}(X))(y-\mathbb{E}(Y))p(x)p(y)dxdy \\
&= \text{var}(x)\int_{-\infty}^{\infty} p(y)dy + a^2\text{var}(y)\int_{-\infty}^{\infty} p(x)dx \\
&\quad + 2a\iint_{-\infty}^{\infty} xp(x)yp(y)dxdy + 2a\iint_{-\infty}^{\infty} \mathbb{E}(x)p(x)\mathbb{E}(y)p(y)dxdy \\
&\quad - 2a\iint_{-\infty}^{\infty} xp(x)\mathbb{E}(y)p(y)dxdy - 2a\iint_{-\infty}^{\infty} yp(y)\mathbb{E}(X)p(x)dxdy \\
&= \text{var}(x) + a^2\text{var}(y) + 2a\mathbb{E}(X)\mathbb{E}(Y) + 2a\mathbb{E}(X)\mathbb{E}(Y) - 2a\mathbb{E}(X)\mathbb{E}(Y) - 2a\mathbb{E}(X)\mathbb{E}(Y) \\
&= \text{var}(x) + a^2\text{var}(y)
\end{aligned}$$

Problem 2 (Densities, 5 points)
Answer the following questions:

(a) Can a probability density function (pdf) ever take values greater than 1?

(b) Let X be a univariate normally distributed random variable with mean 0 and variance $1/100$. What is the pdf of X?

(c) What is the value of this pdf at 0?

(d) What is the probability that $X = 0$?

(a) Yes, only the integral must be equal to 1. The pdf values correspond to relative likelihoods normalized over the domain of the random variable

(b)

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi\frac{1}{100}}} e^{\frac{-(x-0)^2}{2\frac{1}{100}})} \\ &= \frac{1}{\sqrt{\frac{\pi}{50}}} e^{-50x^2} \end{aligned}$$

(c)

$$p(X = 0) = \frac{1}{\sqrt{\frac{\pi}{50}}}$$

(d)

$$P(X = 0) = \int_0^0 p(x)dx = 0$$

Problem 3 (Calculus, 4 points)
Let $x, y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times m}$. Please answer the following questions, writing your answers in vector notation.

(a) What is the gradient with respect to x of $x^T y$?

(b) What is the gradient with respect to x of $x^T x$?

(c) What is the gradient with respect to x of $x^T A$?

(d) What is the gradient with respect to x of $x^T Ax$?

(a)

$$\begin{aligned} x^\mathsf{T} y &= \sum_{i=1}^m x_i y_i \\ \nabla(x^\mathsf{T} y) &= \left(\frac{\partial}{\partial x_1} \sum_{i=1}^m x_i y_i, ..., \frac{\partial}{\partial x_m} \sum_{i=1}^m x_i y_i \right) \\ &= (y_1, ..., y_m) \\ &= y \end{aligned}$$

(b)

$$\begin{aligned} x^\mathsf{T} x &= \sum_{i=1}^m x_i^2 \\ \nabla(x^\mathsf{T} x) &= \left(\frac{\partial}{\partial x_1} \sum_{i=1}^m x_i^2, ..., \frac{\partial}{\partial x_m} \sum_{i=1}^m x_i^2 \right) \\ &= (2x_1, ..., 2x_m) \\ &= 2x \end{aligned}$$

(c)

$$
\begin{aligned}
x^{\mathsf{T}} A &= \left(\sum_{i=1}^{m} x_i A_{i1}, ..., \sum_{i=1}^{m} x_i A_{im} \right) \\
\nabla(x^{\mathsf{T}} x) &= \left(\frac{\partial}{\partial x_1} \left(\sum_{i=1}^{m} x_i A_{i1}, ..., \sum_{i=1}^{m} x_i A_{im} \right), ..., \frac{\partial}{\partial x_m} \left(\sum_{i=1}^{m} x_i A_{i1}, ..., \sum_{i=1}^{m} x_i A_{im} \right) \right) \\
&= \left(\frac{\partial}{\partial x_1} (x_1 A_{11}, ..., x_1 A_{1m}), ..., \frac{\partial}{\partial x_m} (x_m A_{m1}, ..., x_m A_{mm}) \right) \\
&= ((A_{11}, ..., A_{1m}), ..., (A_{m1}, ..., A_{mm})) \\
&= A^{\mathsf{T}}
\end{aligned}
$$

(d)

$$
\begin{aligned}
x^{\mathsf{T}} A x &= \sum_{i=1}^{m} \sum_{j=1}^{m} x_i A_{ij} x_j \\
\nabla(x^{\mathsf{T}} A x) &= \left(\frac{\partial}{\partial x_1} \sum_{i=1}^{m} \sum_{j=1}^{m} x_i A_{ij} x_j, ..., \frac{\partial}{\partial x_m} \sum_{i=1}^{m} \sum_{j=1}^{m} x_i A_{ij} x_j \right) \\
&= (A_{11} x_1 + ... + A_{1m} x_m + A_{11} x_1 + ... + A_{m1} x_m, ..., \\
&\quad A_{m1} x_1 + ... + A_{mm} x_m + A_{1m} x_1 + ... + A_{mm} x_m) \\
&= (A_1 x + (A^{\mathsf{T}})_1 x, ..., A_m x + (A^{\mathsf{T}})_m x) \\
&= (A + A^{\mathsf{T}}) x
\end{aligned}
$$

Problem 4 (Linear Regression, 10pts)
Suppose that $X \in \mathbb{R}^{n \times m}$ with $n \geq m$ and $Y \in \mathbb{R}^n$, and that $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$. In this question you will derive the result that the maximum likelihood estimate $\hat{\beta}$ of β is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

(a) What are the expectation and covariance matrix of $\hat{\beta}$, for a given true value of β?

(b) Show that maximizing the likelihood is equivalent to minimizing the squared error $\sum_{i=1}^{n} (y_i - x_i \beta)^2$. [Hint: Use $\sum_{i=1}^{n} a_i^2 = a^T a$]

(c) Write the squared error in vector notation, (see above hint), expand the expression, and collect like terms. [Hint: Use $\beta^T x^T y = y^T x \beta$ (why?) and $x^T x$ is symmetric]

(d) Take the derivative of this expanded expression with respect to β to show the maximum likelihood estimate $\hat{\beta}$ as above. [Hint: Use results 3.c and 3.d for derivatives in vector notation.]

(a)

$$
\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}((X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} Y) \\
&= (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} \mathbb{E}(Y) \\
&= (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} X \beta \\
&= \beta
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{cov}(\hat{\beta},\hat{\beta}) &= \mathbb{E}\left((\hat{\beta}-\mathbb{E}(\hat{\beta}))(\hat{\beta}-\mathbb{E}(\hat{\beta}))^{\top}\right)\\
&= \mathbb{E}(\hat{\beta}\hat{\beta}^{\top}) - \mathbb{E}(\hat{\beta})\mathbb{E}(\hat{\beta}^{\top})\\
&= \mathbb{E}\left((X^{\top}X)^{-1}X^{\top}Y((X^{\top}X)^{-1}X^{\top}Y)^{\top}\right) - \beta\beta^{\top}\\
&= \mathbb{E}\left((X^{\top}X)^{-1}X^{\top}YY^{\top}X(X^{\top}X)^{-1}\right) - \beta\beta^{\top}\\
&= \mathbb{E}\left(X^{-1}(X^{\top})^{-1}X^{\top}YY^{\top}XX^{-1}(X^{\top})^{-1}\right) - \beta\beta^{\top}\\
&= \mathbb{E}\left(X^{-1}YY^{\top}(X^{\top})^{-1}\right) - \beta\beta^{\top}\\
&= X^{-1}\mathbb{E}(YY^{\top})(X^{\top})^{-1} - \beta\beta^{\top}\\
&= X^{-1}(\sigma^2 I + (X\beta)(X\beta)^{\top})(X^{\top})^{-1} - \beta\beta^{\top}\\
&= X^{-1}(\sigma^2 I + X\beta\beta^{\top}X^{\top})(X^{\top})^{-1} - \beta\beta^{\top}\\
&= \sigma^2 X^{-1}(X^{\top})^{-1} + X^{-1}X\beta\beta^{\top}X^{\top}(X^{\top})^{-1} - \beta\beta^{\top}\\
&= \sigma^2 X^{-1}(X^{\top})^{-1}\\
&= \sigma^2 (X^{\top}X)^{-1}
\end{aligned}
$$

(b)

$$
MSE(\beta; X, Y) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 = (y - x\beta)^{\top}(y - x\beta)
$$

$$
\begin{aligned}
Ł(\beta; X, Y) &= \sum_{i=1}^{n}\log(p(y_i \mid x_i, \beta))\\
&= \sum_{i=1}^{n}\log\left(\frac{1}{2\pi\sigma^2}e^{\frac{-(y_i - x_i\beta)^2}{2\sigma^2}}\right)\\
&= \sum_{i=1}^{n}\log\left(\frac{1}{2\pi\sigma^2}\right) + \sum_{i=1}^{n}\frac{-(y_i - x_i\beta)^2}{2\sigma^2}\\
&= \sum_{i=1}^{n}\log\left(\frac{1}{2\pi\sigma^2}\right) + \frac{-1}{2\sigma^2}(y - x\beta)^{\top}(y - x\beta)
\end{aligned}
$$

$$
\begin{aligned}
\arg\max_{\beta} Ł(\beta; X, Y) &= \arg\max_{\beta}\left(\sum_{i=1}^{n}\log\left(\frac{1}{2\pi\sigma^2}\right) + \frac{-1}{2\sigma^2}(y - x\beta)^{\top}(y - x\beta)\right)\\
&= \arg\max_{\beta}\left(-(y - x\beta)^{\top}(y - x\beta)\right)\\
&= \arg\min_{\beta}\left((y - x\beta)^{\top}(y - x\beta)\right)\\
&= \arg\min_{\beta} MSE(\beta; X, Y)
\end{aligned}
$$

(c)

$$
\begin{aligned}
(y - x\beta)^{\top}(y - x\beta) &= (y^{\top} - (x\beta)^{\top})(y - x\beta)\\
&= (y^{\top} - \beta^{\top}x^{\top})(y - x\beta)\\
&= y^{\top}y - y^{\top}x\beta - \beta^{\top}x^{\top}y + \beta^{\top}x^{\top}x\beta\\
&= y^{\top}y - 2\beta^{\top}x^{\top}y + \beta^{\top}x^{\top}x\beta
\end{aligned}
$$

(d)

$$\begin{aligned}
\frac{\partial}{\partial \beta}(y - x\beta)^{\mathsf{T}}(y - x\beta) &= \frac{\partial}{\partial \beta}(y^{\mathsf{T}}y - 2y^{\mathsf{T}}x\beta + \beta^{\mathsf{T}}x^{\mathsf{T}}x\beta) \\
&= \frac{\partial}{\partial \beta}y^{\mathsf{T}}y - 2\frac{\partial}{\partial \beta}\beta^{\mathsf{T}}x^{\mathsf{T}}y + \frac{\partial}{\partial \beta}\beta^{\mathsf{T}}x^{\mathsf{T}}x\beta \\
&= -2x^{\mathsf{T}}y + (x^{\mathsf{T}}x + (x^{\mathsf{T}}x)^{\mathsf{T}})\beta \\
&= -2x^{\mathsf{T}}y + 2x^{\mathsf{T}}x\beta
\end{aligned}$$

Evaluate at 0 to find $\hat{\beta}$

$$\begin{aligned}
0 &= -2x^{\mathsf{T}}y + 2x^{\mathsf{T}}x\hat{\beta} \\
x^{\mathsf{T}}y &= x^{\mathsf{T}}x\hat{\beta} \\
\hat{\beta} &= (x^{\mathsf{T}}x)^{-1}x^{\mathsf{T}}y
\end{aligned}$$

Problem 5 (Ridge Regression, 10pts)
Suppose we place a normal prior on β. That is, we assume that $\beta \sim \mathcal{N}(0, \tau^2 I)$.

(a) Show that the MAP estimate of β given Y in this context is

$$\hat{\beta}_{MAP} = (X^T X + \lambda I)^{-1} X^T Y$$

where $\lambda = \sigma^2/\tau^2$.

Estimating β in this way is called *ridge regression* because the matrix λI looks like a "ridge". Ridge regression is a common form of *regularization* that is used to avoid the overfitting that happens when the sample size is close to the output dimension in linear regression.

(b) Show that ridge regression is equivalent to adding m additional rows to X where the j-th additional row has its j-th entry equal to $\sqrt{\lambda}$ and all other entries equal to zero, adding m corresponding additional entries to Y that are all 0, and and then computing the maximum likelihood estimate of β using the modified X and Y.

(a)

$$\begin{aligned}
MAP(\beta; X, Y) &= \arg\max_{\beta} \sum_{i=1}^{n} \log(p(y_i \mid x_i, \beta)p(\beta)) \\
&= \arg\max_{\beta} \left(\sum_{i=1}^{n} \log(p(y_i \mid x_i, \beta)) + \sum_{i=1}^{n} \log(p(\beta)) \right) \\
&= \arg\max_{\beta} \left(-\frac{1}{2\sigma^2}(y - x\beta)^{\mathsf{T}}(y - x\beta) - \frac{1}{2\tau^2}\sum_{i=1}^{n} \beta_i^2 \right) \\
&= \arg\min_{\beta} \left(\frac{1}{2\sigma^2}(y - x\beta)^{\mathsf{T}}(y - x\beta) + \frac{1}{2\tau^2}\beta^{\mathsf{T}}\beta \right)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \beta}\left(\frac{1}{2\sigma^2}(y-x\beta)^\mathsf{T}(y-x\beta)+\frac{1}{2\tau^2}\beta^\mathsf{T}\beta\right) &= \frac{1}{2\sigma^2}\frac{\partial}{\partial \beta}(y-x\beta)^\mathsf{T}(y-x\beta)+\frac{1}{2\tau^2}\frac{\partial}{\partial \beta}\beta^\mathsf{T}\beta \\
&= \frac{1}{2\sigma^2}(-2x^\mathsf{T}y+2x^\mathsf{T}x\hat{\beta})+\frac{1}{2\tau^2}2\beta \\
&= \frac{1}{\sigma^2}(-x^\mathsf{T}y+x^\mathsf{T}x\hat{\beta})+\frac{1}{\tau^2}\beta
\end{aligned}
$$

Evaluate at 0 to find $\hat{\beta}$

$$
\begin{aligned}
0 &= \frac{1}{\sigma^2}(-x^\mathsf{T}y+x^\mathsf{T}x\hat{\beta})+\frac{1}{\tau^2}\hat{\beta} \\
\frac{1}{\sigma^2}x^\mathsf{T}y &= \frac{1}{\sigma^2}x^\mathsf{T}x\hat{\beta}+\frac{1}{\tau^2}I\hat{\beta} \\
x^\mathsf{T}y &= x^\mathsf{T}x\hat{\beta}+\frac{\sigma^2}{\tau^2}I\hat{\beta} \\
x^\mathsf{T}y &= (x^\mathsf{T}x+\frac{\sigma^2}{\tau^2}I)\hat{\beta} \\
\hat{\beta} &= (x^\mathsf{T}x+\frac{\sigma^2}{\tau^2}I)^{-1}x^\mathsf{T}y \\
\hat{\beta} &= (x^\mathsf{T}x+\lambda I)^{-1}x^\mathsf{T}y
\end{aligned}
$$

(b) Let $\hat{x}$ and $\hat{y}$ refer to their modified versions

$$\hat{\beta} = (\hat{x}^\mathsf{T}\hat{x})^{-1}\hat{x}^\mathsf{T}\hat{y}$$

$$
\begin{aligned}
(\hat{x}^\mathsf{T}\hat{y})_i &= \hat{x}_i^\mathsf{T}\hat{y} \\
&= \sum_{k=1}^{n+m}\hat{x}_{ik}^\mathsf{T}\hat{y}_k \\
&= \sum_{k=1}^{n}\hat{x}_{ik}^\mathsf{T}\hat{y}_k+\sum_{l=n+1}^{n+m}\hat{x}_{il}^\mathsf{T}\hat{y}_l \\
&= \sum_{k=1}^{n}\hat{x}_{ik}^\mathsf{T}\hat{y}_k+0 \quad \text{because for } l>n,\ \hat{y}_l=0 \\
&= (x^\mathsf{T}y)_i \\
&\therefore \hat{x}^\mathsf{T}\hat{y} = x^\mathsf{T}y
\end{aligned}
$$

$$
\begin{aligned}
(\hat{x}^\mathsf{T}\hat{x})_{ij} &= \sum_{k=1}^{n+m}\hat{x}_{ik}^\mathsf{T}\hat{x}_{kj} \\
&= \sum_{k=1}^{n}\hat{x}_{ik}^\mathsf{T}\hat{x}_{kj}+\sum_{l=n+1}^{n+m}\hat{x}_{il}^\mathsf{T}\hat{x}_{lj} \\
&= (x^\mathsf{T}x)_{ij}+\lambda I_{ij} \quad \text{because for } l>n,\ \hat{x}_{il}^\mathsf{T}\hat{x}_{lj}=\sqrt{\lambda}\sqrt{\lambda}=\lambda \text{ when } (l-n)=i=j \text{ and 0 otherwise} \\
&\therefore \hat{x}^\mathsf{T}\hat{x} = x^\mathsf{T}x+\lambda I
\end{aligned}
$$

$$\therefore \hat{\beta} = (\hat{x}^\mathsf{T}\hat{x})^{-1}\hat{x}^\mathsf{T}\hat{y} = (x^\mathsf{T}x+\lambda I)^{-1}x^\mathsf{T}y$$

Problem 6 (Gaussians in high dimensions, 10pts)
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

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}_D, \mathbb{I}_D),$$

where $\mathbf{0}_D$ indicates a column vector of D zeros and $\mathbb{I}_D$ is a $D \times D$ identity matrix.

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

2. In low-dimensions our intuition tells us that samples from the unit Gaussian will be near the origin. Draw 10000 samples from a $D = 1$ Gaussian and 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?

3. Draw 10000 samples from $D = \{1, 2, 3, 10, 100\}$ Gaussians and, on a single plot, show the normalized histograms for the 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).

4. From Wikipedia, if $\mathbf{x}_i$ 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 $k = \{1, 2, 3, 10, 100\}$.

5. Taking two samples from the D-dimensional unit Gaussian, $\mathbf{x}_a, \mathbf{x}_b \sim \mathcal{N}(\mathbf{0}_D, \mathbb{I}_D)$ how is $\mathbf{x}_a - \mathbf{x}_b$ distributed? Using the above result about χ-distribution, how is $||\mathbf{x}_a - \mathbf{x}_b||_2$ distributed? (Hint: start with a $\mathcal{X}$-distributed random variable and use the change of variables formula.) Plot the pdfs of this distribution for $k = \{1, 2, 3, 10, 100\}$. 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.

6. In lecture we saw examples of interpolating between latent points to generate convincing data. Given two samples from a gaussian $\mathbf{x}_a, \mathbf{x}_b \sim \mathcal{N}(\mathbf{0}_D, \mathbb{I}_D)$ the linear interpolation between them x_α is defined as a function of $\alpha \in [0, 1]$

$$\text{lin\_interp}(\alpha, \mathbf{x}_a, \mathbf{x}_b) = \alpha \mathbf{x}_a + (1 - \alpha)\mathbf{x}_b$$

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?

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

$$\text{polar\_interp}(\alpha, \mathbf{x}_a, \mathbf{x}_b) = \sqrt{\alpha}\mathbf{x}_a + \sqrt{(1 - \alpha)}\mathbf{x}_b$$

This interpolates between two points while maintaining Euclidean norm. On the same plot from the previous question, plot the probabilitiy 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 intuative 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**.

8. (**Bonus 5pts**) In the previous two questions we compute the average loglikelihood of the linear and polar interpolations under the unit gaussian. Instead, consider the norm along the interpolation, $\sqrt{\mathbf{x}_\alpha^\top \mathbf{x}_\alpha}$. 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. $\mathcal{X}_D(\sqrt{\mathbf{x}_\alpha^\top \mathbf{x}_\alpha})$. 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)? Using your answer for questions 3 and 4, provide geometric intuition for the log-likelihood along the linear and polar interpolations. Use this to further justify your explanation for the suitability of polar v.s. linear interpolation.

1.

$$\begin{aligned} norm(x) &= \sqrt{\sum_{i=1}^{D} x_i^2} \\ &= \sqrt{x_1 x_1 + ... + x_D x_D} \\ &= \sqrt{x^\top x} \end{aligned}$$

2. Figure 1 confirms the intuition that samples drawn from a unit gaussian are near the origin.

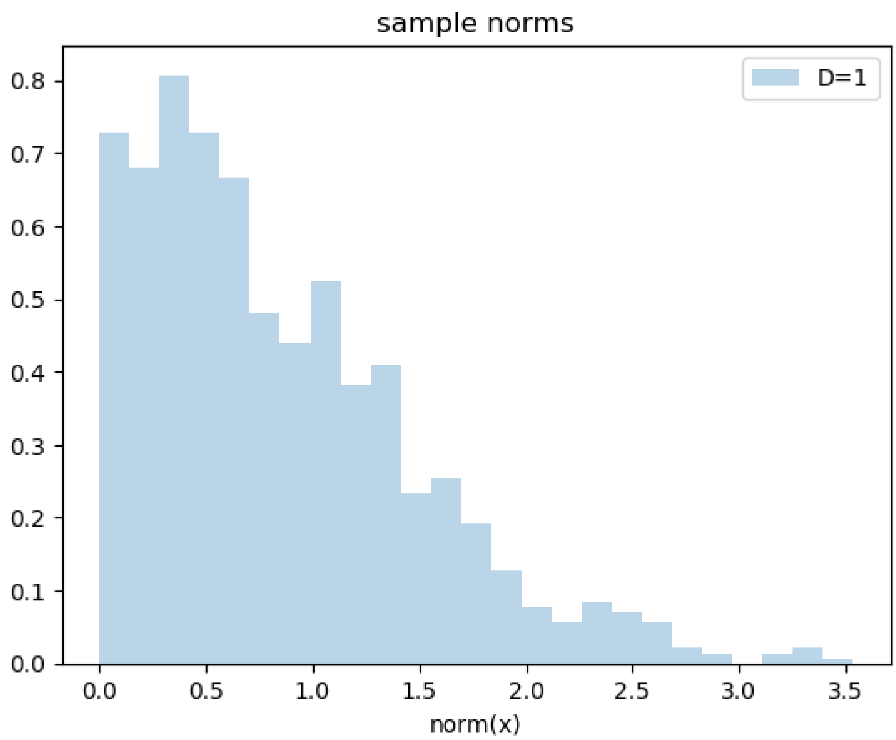

Figure 1: Normalized histogram of 1000 samples drawn from multivariate gaussian with dimension D.

3. Figure 2 shows that as the dimensionality of the gaussian increases, the expected distance of samples from the mean increases as well.

4. Figure 2 shows how the distance of samples drawn from a multivariate unit gaussian with mean at the origin is distributed according to the Chai distribution.

5. A difference between two Gaussian random variables is distributed according to a Gaussian with mean $\mu_a - \mu_b$ and covariance matrix $|\sigma_a^2 + \sigma_b^2|I$, so $\mathbf{x_a} - \mathbf{x_b}$ is given by:

$$\mathbf{x_a} - \mathbf{x_b} \sim \mathcal{N}(0, 2I)$$

If we call this new random variable $\mathbf{x}$, then because $\mathbf{x_i}$ is distributed according to a Gaussian, its norm is distributed according to a $\chi - distribution$:

$$norm(\mathbf{x}) = \sqrt{\frac{1}{2}\mathbf{x}^\top \mathbf{x}} = \frac{1}{\sqrt{2}}\sqrt{\mathbf{x}^\top \mathbf{x}} \sim \chi - distribution$$

If we call this random variable $\mathbf{z}$ and the random variable that we are interested in $\mathbf{y} = \sqrt{\mathbf{x}^\top \mathbf{x}}$, then we can find the

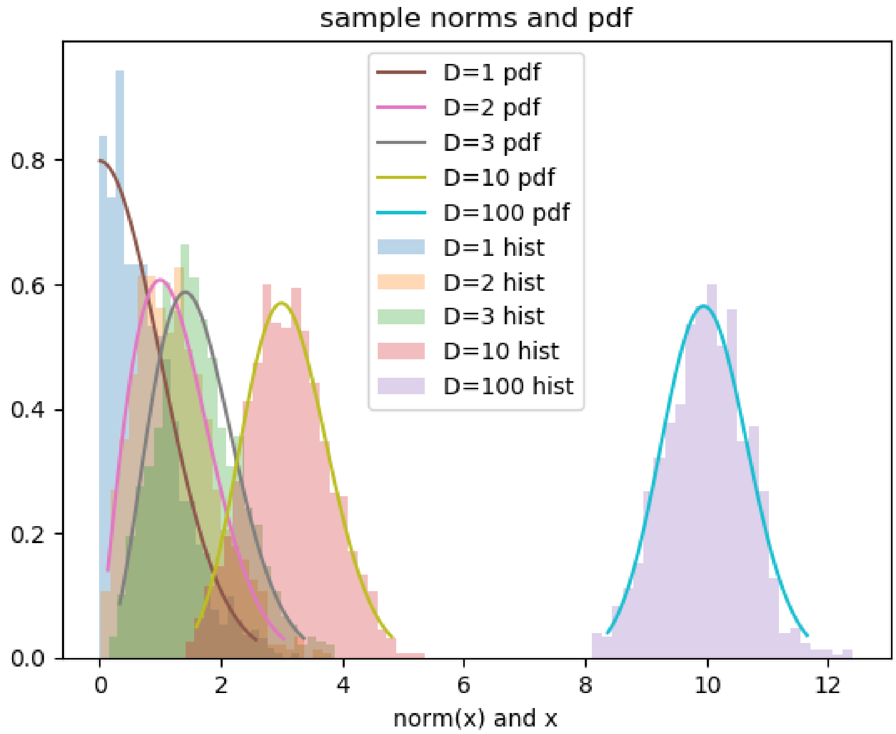

Figure 2: Normalized histogram of 1000 samples drawn from multivariate gaussian with dimensions D, as well as the pdf's of the Chai distribution with dimensions D.

distribution of $\mathbf{y}$ using a change of variables on the known distribution of $\mathbf{z}$. Namely:

$$\mathbf{y} = g(\mathbf{z}) = \sqrt{2}\mathbf{z}$$
$$\mathbf{z} = g^{-1}(\mathbf{y}) = \frac{1}{\sqrt{2}}\mathbf{y}$$
$$f_{\mathbf{y}}(\mathbf{y}) = \left|\frac{d}{d\mathbf{y}}(g^{-1}(\mathbf{y}))\right| f_{\mathbf{z}}(g^{-1}(\mathbf{y}))$$
$$f_{\mathbf{y}}(\mathbf{y}) = \frac{1}{\sqrt{2}} f_{\mathbf{z}}(\frac{1}{\sqrt{2}}\mathbf{y})$$

Figure 3 shows how this transformed $\chi - distribution$ fits the observations obtained from taking the norm of the difference between samples drawn from a unit Gaussian distribution.

6. Figure 4 shows that the log-likelihood under the unit Gaussian of points along a linear interpolation increases during the interpolation to a peak when $\alpha = 0.5$ before decreasing again. This means that points in the middle of the interpolation are more likely. Furthermore, the log-likelihood increase along the interpolation is greater as the dimension of the Gaussian increases. A higher log-likelihood for the interpolated points is generally not desired because we don't normally want to interpolate through more likely points, but rather through points of similar likelihoods to the two samples being interpolated between. Ideally, we want the likelihood of points along the interpolation to increase linearly (i.e. we want a linear interpolation of likelihood between two points). In addition, in cases where the vast majority of points have low likelihoods but a few points have high likelihoods, we do not want our interpolation to traverse through these uncommon but individually likely points. For these reasons, it is not a good idea to linearly interpolate between samples from a high dimensional Gaussian?

7. Figure 4 shows that, in contrast to linear interpolation, polar interpolation between two points sampled from a unit Gaussian preserves the log-likelihood of the points. In other words, the log-likelihood follows a roughly linear interpolation, even though the points along the interpolation do not. For this reason, it is more appropriate to interpolate between points sampled from a unit Gaussian using polar interpolation, especially for high dimensional Gaussians where linear interpolation results in large increases in the log-likelihood for interpolated points.
An intuitive explanation for why we want points along an interpolation to be no more likely than the two points being interpolated between can be obtained from an example. If we trained a generative model that could sample face images and we wanted to interpolate between two faces, we wouldn't want the interpolated samples to always

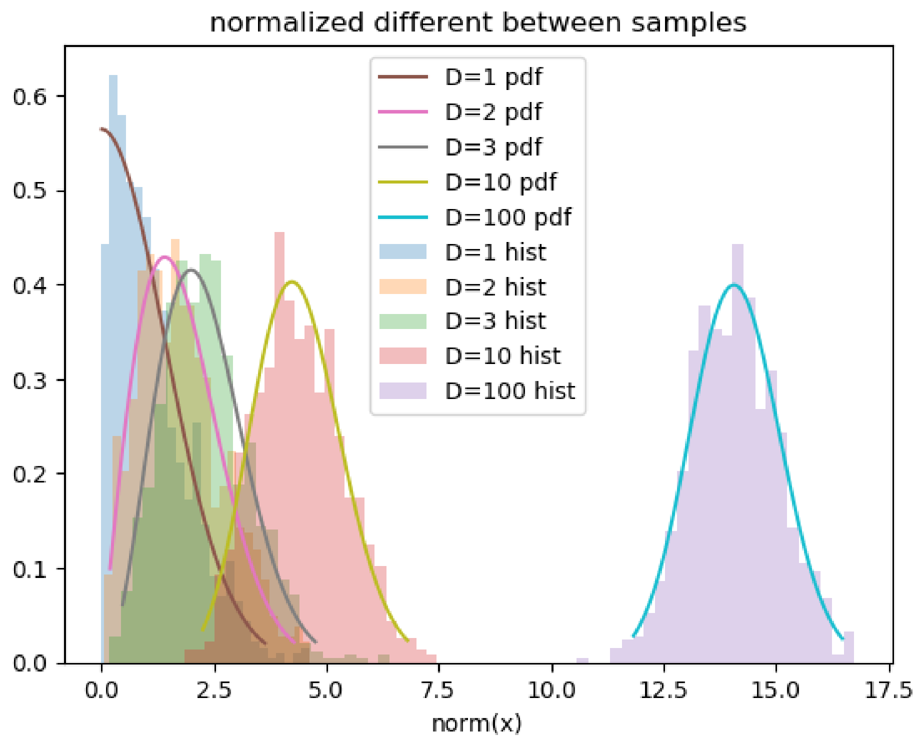

Figure 3: Normalized histogram of 1000 samples drawn from the difference between two samples of a multivariate gaussian with dimensions D, as well as the pdf's of the transformed Chai distribution that represents them with dimensions D.

transition through images with very globally common features, lighting, etc. Instead, we want the features of the face to transition smoothly between the two faces.

8. Figure 5 shows that the log-likelihood of the norm along the linear interpolation is less than that of the true samples (endpoints) for high dimensional Gaussians. In contrast, the log-likelihood of the norm along the polar interpolation remains roughly constant. A geometric explanation for this difference is that high points sampled from high dimensional Gaussians are far from the origin, but a linear interpolation between them traverses in a straight line and brings the interpolated points closer to the origin. Thus, the points during linear interpolation are less representative of the overall distribution. These interpolated points are individually more likely than the endpoints, as shown in part 6, but they correspond to unusual points in the distribution in terms of their proximity to the origin. In contrast, polar coordinates preserves the distance of the interpolated points to the origin, and so these points are more representative of the distribution. It is like traversing along the surface of a hypersphere. This provides additional intuition as to why polar interpolation is more suitable for high dimensional Gaussians, since we want the interpolated points to be representative of the distribution rather than be individually more likely but nonetheless unusual points near the origin.

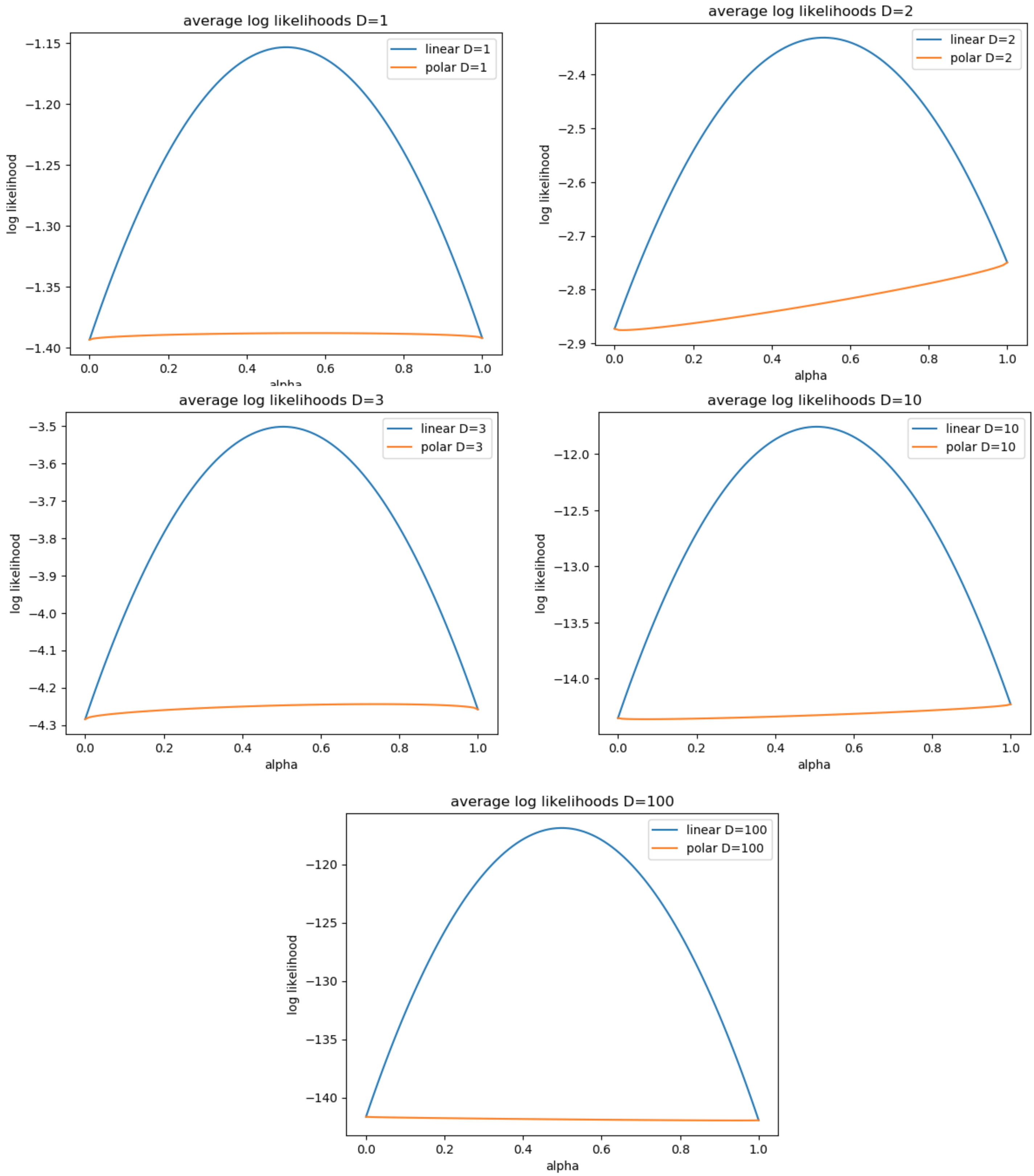

Figure 4: Log-likelihoods of points along an interpolation between two unit Gaussian random variables.

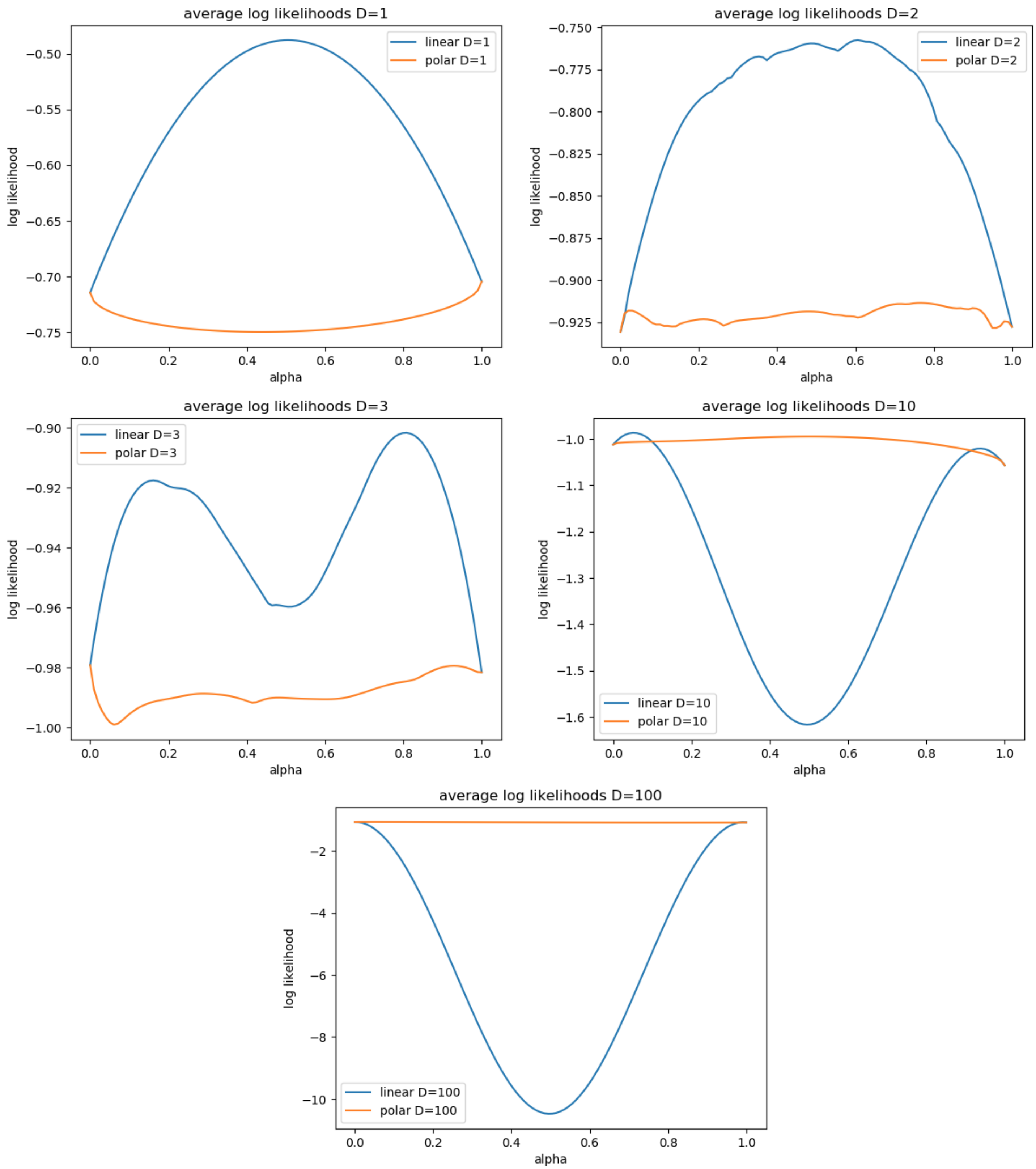

Figure 5: Log-likelihoods of norms of points along an interpolation between two unit Gaussian random variables according to the $\chi - distribution$.