1 Gaussian Isocontours

(a) Consider a linear transformation $T(x) = Ux$, where $x \in \mathbb{R}^2$ and $U \in \mathbb{R}^{2\times 2}$, that takes a vector and rotates it by 45° counterclockwise. Find the matrix U. What does $T'(x) = U^\top x$ mean?

Solution: We want to find a matrix U such that $Ue_1 = \left[\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right]^\top$ and $Ue_2 = \left[-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right]^\top$, where e_i is the unit vector along the i-th coordinate axis. Hence $U = \begin{bmatrix} \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \end{bmatrix}$. Thus $T'(x) = U^\top x$ is the transformation that rotates x by $-45°$ counterclockwise; said differently, 45° clockwise.

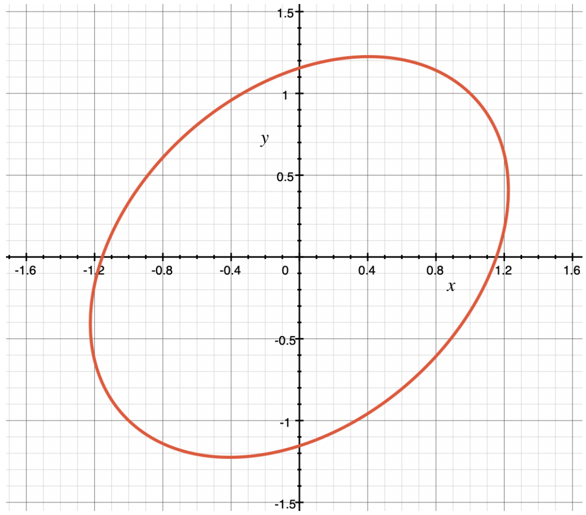
(b) Now let's verify some basic properties of this matrix U. What is the determinant of U? Does U^{-1} exist? If it does, can you express U^{-1} in terms of U? Usually, we refer to matrices that have the above properties as the Special Orthogonal group, or SO(2), as the matrix is 2 by 2. They are also called rotational matrices.

Solution: U has determinant 1, thus U is invertible. It has the special property that $U^{-1} = U^\top$, which implies that $UU^\top = U^\top U = I$. Matrices with this property are said to be *orthogonal* or *orthonormal*, meaning that all their columns have unit length and are mutually orthogonal. (And the same for their rows.)

(c) With the matrix U from part (a), we construct a new matrix $A = U\Lambda U^\top$ where $\Lambda = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$. What are the eigenvalues and eigenvectors of the matrix A? Now consider the quadratic function $Q(x) = x^\top A^{-1} x$. Draw the level set $Q(x) = 1$.

Solution: Notice that since U is an orthonormal matrix, $A = U\Lambda U^\top$ is the eigendecomposition of the matrix A. Hence, the eigenvectors and eigenvalues of A are $u_1 = \left[\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right]^\top$ with corresponding eigenvalue $\lambda_1 = 2$ and $u_2 = \left[-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right]^\top$ with corresponding eigenvalue $\lambda_2 = 1$.

Now let's draw the level set of the function $Q(x) = 1$. $Q(x) = x^\top A^{-1} x = x^\top U \Lambda^{-1} U^\top x = \frac{(x^\top u_1)^2}{\lambda_1} + \frac{(x^\top u_2)^2}{\lambda_2}$ where u_1 and u_2 are the first and second columns of the matrix U respectively. Hence we have the level set $\frac{(x^\top u_1)^2}{\lambda_1} + \frac{(x^\top u_2)^2}{\lambda_2} = 1$ which is an ellipse centered at the origin with the major axis given by $\sqrt{\lambda_1} u_1$ and the minor axis given by $\sqrt{\lambda_2} u_2$.



(d) Using the result from part (c), show that the isocontours of a multivariate Gaussian $X \sim N(\mu, \Sigma)$ where $\Sigma \succ 0$ are also ellipses. Isocontour means a set of X that has the same probability of being sampled. For now, you can assume that you can read off probabilities from continuous probability density functions.

Hint: Recall that the density of a multivariate Gaussian is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Solution: Notice that the multivariate Gaussian density has a quadratic term in the exponent. The isocontours for our density are $\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) = c$.

$$\begin{aligned}
\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right) &= c \\
\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) &= \ln \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} - \ln c \\
(x-\mu)^\top \Sigma^{-1}(x-\mu) &= -2\ln \sqrt{(2\pi)^d|\Sigma|} - 2\ln c
\end{aligned}$$

Now observe that the RHS of the equation is simply a constant and hence the isocontours will look like ellipses with major axes in the direction of the largest eigenvalue Σ and a minor axis in the direction of the smallest eigenvalue of Σ.

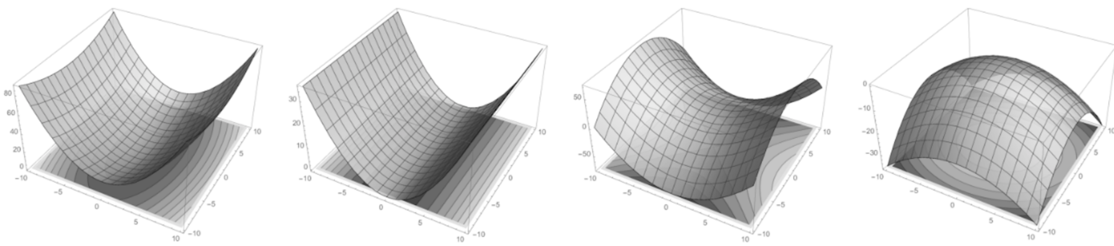
For the remainder of this problem, we will explore the shape of quadratic forms by examining the eigenstructure of the Hessian matrix. Recall that the Hessian $H \in \mathbb{R}^{d\times d}$ of a function $f : \mathbb{R}^d \to \mathbb{R}$ is the matrix of second derivatives $H_{i,j} = \frac{\partial f}{\partial x_i x_j}$ of the function. The eigenstructure of H contains information about the curvature of f.

(e) Suppose you are given the a quadratic function $Q(x) = \frac{1}{2}x^\top Ax + b^\top x$ where $x, b \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2\times 2}$ is a symmetric matrix. What is the Hessian of Q?

Solution:

$$\begin{aligned} \nabla_x f(x) &= \frac{1}{2}(A + A^\top)x + b = Ax + b \\ \nabla_x^2 f(x) &= \nabla_x Ax + b = A^\top = A \end{aligned}$$

(f) We will now think about how the eigenstructure of the Hessian matrix affects the shape of $Q(x)$. Recall that by the Spectral Theorem, A has two real eigenvalues. Match each of the following cases, to the appropriate plot of $Q(x)$. How does the magnitude of the eigenvalues affect your sketch? For each of these 4 cases, is there an unique local optimum?



(a) $\lambda_1(A), \lambda_2(A) > 0$

(b) $\lambda_1(A), \lambda_2(A) < 0$

(c) $\lambda_1(A) > 0, \lambda_2(A) < 0$

(d) $\lambda_1(A) > 0, \lambda_2(A) = 0$

Solution:

(a) Since both eigenvalues are strictly positive, $A > 0$ which means that $x^\top Ax > 0, \forall x \neq \mathbf{0}$. Thus the function has a unique global minimum at $x = \mathbf{0}$. Let's examine the curvature around the origin. Our goal is to look at all possible directions and understand what the curvature looks like for each one. Recall from Homework 2 that $\max\limits_{\|u\|=1} u^\top Au = \lambda_{\max}(A)$. Likewise, $\min\limits_{\|u\|=1} u^\top Au = \lambda_{\min}(A)$. As a result, if we move away from 0 in the direction of the eigenvector corresponding to $\lambda_{\max}(A)$, we will experience the steepest curvature. On the other hand, if we move in the direction of the eigenvector corresponding to $\lambda_{\min}(A)$, we will experience the least curvature. The quadratic should look something like the surface in the first figure from left. There is an unique local minimum.

(b) Figure 4. Unique local maximum.

(c) Figure 3. May nor may not exist.

(d) Figure 2. No unique local optimum exists.

2 Linear Discriminant Analysis

In this question, we will explore some of the mechanics of LDA and understand why it produces a linear decision boundary in the case where the covariance matrix is anisotropic.

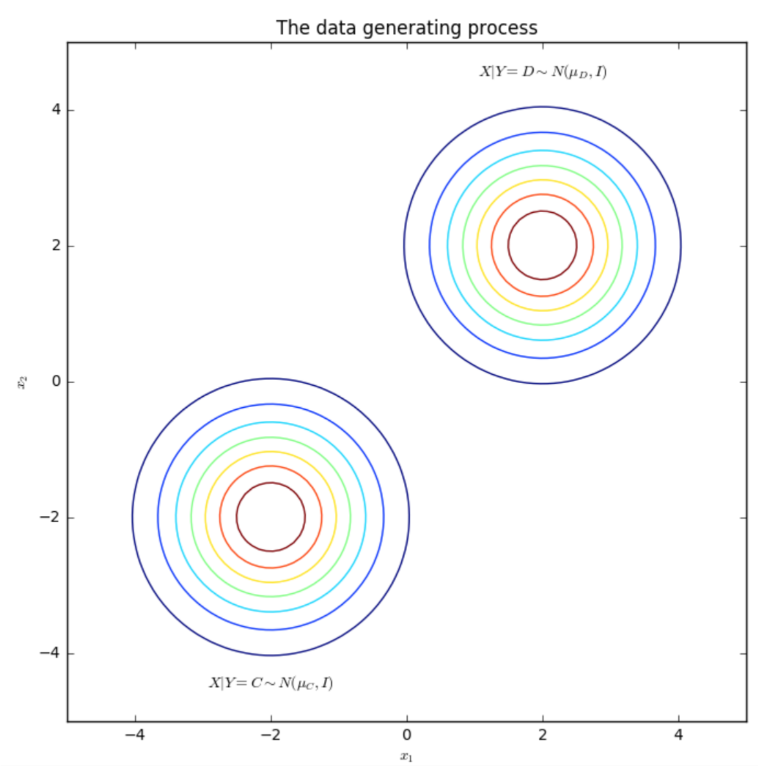
(a) Suppose $\Sigma = \mathrm{Var}(X)$ is the covariance matrix of random vector $X \in \mathbb{R}^d$. Prove that $\mathrm{Var}(AX) = A\Sigma A^\top$.

Solution:

$$\begin{aligned}
\mathrm{Var}(AX) &= \mathrm{E}[(AX - A\,\mathrm{E}[X])(AX - A\,\mathrm{E}[X])^\top] = \mathrm{E}[A(X - \mathrm{E}[X])(X - \mathrm{E}[X])^\top A^\top]] \\
&= A\,\mathrm{E}[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^\top]A^\top \\
&= A\Sigma A^\top.
\end{aligned}$$

(b) Suppose you have a binary classification problem with $x \in \mathbb{R}^2$, and you already know the data generating process.

- The two classes have identical priors $P(Y = C) = P(Y = D) = \frac{1}{2}$.
- The class-conditional densities are $(X|Y = C) \sim N(\mu_C, I)$ and $(X|Y = D) \sim N(\mu_D, I)$ where $\mu_C = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$, $\mu_D = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.



We recognize this problem as a special case of LDA where the two classes have an equal prior probability and the common covariance matrix is simply the identity. Recall the optimal Bayes optimal decision boundary from the last discussion worksheet, which is the perpendicular bisector of the line connecting μ_0 and μ_1.

Now we will use this intuition to explain why the decision boundary also has to be linear when the class-conditional densities have a more general covariance matrix $\Sigma \geq 0$.

Assume that we are given the same setup as in the previous part, but this time the covariance matrix is some known $\Sigma \geq 0$ instead of the identity matrix. Find a linear transformation such that the class-conditional distributions are isotropic Gaussians in the transformed space. What is the decision boundary in the transformed space? What does that boundary correspond to in the original space?

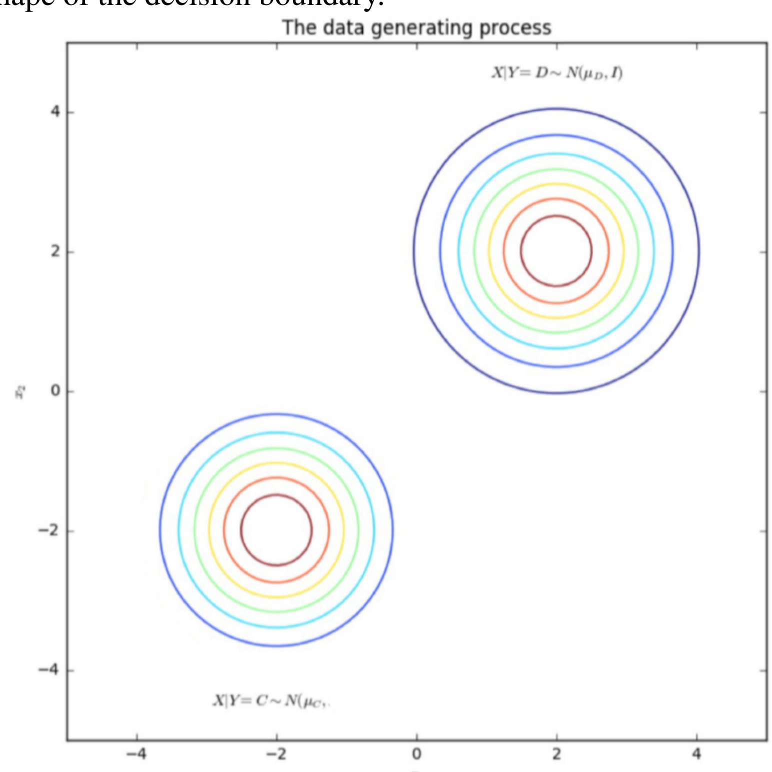
Hint: The result you proved in Problem 1 may be useful.

Solution: $T(x) = \Sigma^{-1/2}x$ and $\bar{x} = T(x)$ for all x in the original space.

The decision boundary has to be some plane in the transformed space of the form $\bar{w}^\top \bar{x} = 0$. (The decision boundary goes through the origin because of the specific values μ_C and μ_D that we were given.) In that case, the boundary in the original space will be $w^\top x = 0$ for $w = \Sigma^{-1/2}\bar{w}$ which is itself an affine decision function. This works because in the original space, the decision function would then be:

$w^\top x = (\Sigma^{-1/2}\bar{w})^\top x = \bar{w}^\top \Sigma^{-1/2} x = \bar{w}^\top \bar{x}$

(c) Consider the case below where $(X|Y = C) \sim N(\mu_C, \alpha I)$. Characterize alpha and roughly indicate the shape of the decision boundary.



Solution: We have $0 < \alpha < 1$. The shape of the decision boundary is a quadratic opening away from μ_D.