Exercises for Chapter 4, Part 2

4.2 Consider Sammon mapping of a dissimilarity matrix D^X.

a) For which values of q can Sammon mapping yield a q-dimensional representation of $Y \subset \mathbb{R}^q$ with zero error for Euclidean distances for *any* 4×4 dissimilarity matrix D^X?

   $n = 2$ will yield 2 points with a distance $d_{12} = d_{21}$ which can be mapped with zero error in $q = 1$ dimensions. For each additional point we may (or may not) need another dimension. So for $n = 4$ we arrive at $q = 3$ in the worst case. Any higher-dimensional representation with $q > 3$ can be, for example, realized by adding dimensions with constant values. $\Rightarrow q \geq 3$

b) Sketch a Shepard diagram for such a mapping.

   Zero error means that all points are on the main diagonal. $n = 4$ yields $n \cdot (n-1)/2 = 4 \cdot 3/2 = 6$ pairwise dissimilarities, some may be equal. So, the Shepard diagram has a maximum of 6 unique points, all on the positive main diagonal.

c) Explain why this does not work for $D^X = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 3 \\ 1 & 1 & 3 & 0 \end{pmatrix}$.

   We have $d^x_{34} > d^x_{31} + d^x_{14}$, so the triangle inequality does not hold. Hence, D^X is not Euclidean.

4.3 Consider an auto-encoder $X \to Y \to X'$, where $X, X' \in \mathbb{R}^2$, $Y \in \mathbb{R}$, with

$$y = f(x) = \tanh\left(\frac{x^{(1)} + x^{(2)}}{2}\right).$$

a) Find a suitable function $x' = g(y)$.

   $X' \in \mathbb{R}^2$, so g must yield a two–dimensional vector. g should compensate the nonlinearity tanh in f, so we may use $x' = (\operatorname{atanh} y, \operatorname{atanh} y)$

b) Calculate the average quadratic error of the transformation $g \circ f$ for the data set $X = \{(0,0), (0,1), (1,0), (1,1)\}$.

   $y_1 = \tanh\left(\frac{0+0}{2}\right) = \tanh 0 = 0$
   $x_1' = (\operatorname{atanh} 0, \operatorname{atanh} 0) = (0,0)$
   $y_2 = \tanh\left(\frac{0+1}{2}\right) = \tanh \frac{1}{2}$
   $x_2' = (\operatorname{atanh}\tanh\frac{1}{2}, \operatorname{atanh}\tanh\frac{1}{2}) = (\frac{1}{2}, \frac{1}{2})$
   $y_3 = \tanh\left(\frac{1+0}{2}\right) = \tanh \frac{1}{2}$
   $x_3' = (\operatorname{atanh}\tanh\frac{1}{2}, \operatorname{atanh}\tanh\frac{1}{2}) = (\frac{1}{2}, \frac{1}{2})$
   $y_4 = \tanh\left(\frac{1+1}{2}\right) = \tanh 1$
   $x_4' = (\operatorname{atanh}\tanh 1, \operatorname{atanh}\tanh 1) = (1,1)$
   $e = \frac{1}{4}\left(\|x_1 - x_1'\|^2 + \|x_2 - x_2'\|^2 + \|x_3 - x_3'\|^2 + \|x_4 - x_4'\|^2\right)$
   $= \frac{1}{4}\big((0-0)^2 + (0-0)^2 + (0-\frac{1}{2})^2 + (1-\frac{1}{2})^2 + (1-\frac{1}{2})^2 + (0-\frac{1}{2})^2 +$
   $(1-1)^2 + (1-1)^2\big) = \frac{1}{4} \cdot \frac{4}{4} = \frac{1}{4}$

c) Which other projection methods would for this data set X yield the same X'?

   X' can be obtained by linear projection of X to the main diagonal. This can be achieved, for example, by one–dimensional PCA, but only if the 45° line is enforced as the main axis. For this data set, PCA may yield a line at any angle α as main axis, since for any α the variance is the same:

$$v = \frac{1}{n-1}\sum_{k=1}^{n} \|(x_k - \bar{x})(\cos\alpha, \sin\alpha)^T\|^2 = \frac{4}{3}\left(\frac{1}{2^2}\cos^2\alpha + \frac{1}{2^2}\sin^2\alpha\right) = \frac{1}{3}$$