Unsupervised Learning Methods
Problem Set IV –
MDS, Isomap, Laplacian-Eigenmaps, and T-SNE

Or Livne - 203972922
Daniel Levi - 302506712

Classical MDS
Consider the following inner product:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\phi} := \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle$$

for some suitable $\phi : \mathbb{R}^D \to \mathbb{R}^M$, where $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product, i.e. $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}^T \boldsymbol{b}$.
Consider:

1. The induced norm:
$$\|\boldsymbol{x}\|_{\phi} := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\phi}}$$
2. The induced metric:
$$d_{\phi}(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} - \boldsymbol{y}\|_{\phi}$$

Consider a training set $\{\boldsymbol{x}_i\}_{i=1}^{N}$ and let $\boldsymbol{D}_{\phi} \in \mathbb{R}^{N \times N}$ where $\boldsymbol{D}_{\phi}[i,j] = d_{\phi}^2(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

1.1
Show that
$$-\frac{1}{2}\boldsymbol{J}\boldsymbol{D}_{\phi}\boldsymbol{J} = \boldsymbol{J}\boldsymbol{K}_{\phi}\boldsymbol{J}$$
where:

1. $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T \in \mathbb{R}^{N \times N}$
2. $\boldsymbol{K}_{\phi} := \boldsymbol{\Phi}^T\boldsymbol{\Phi}$, $\boldsymbol{\Phi}$ is given by:
$$\boldsymbol{\Phi} = \begin{bmatrix} | & | & & | \\ \boldsymbol{\phi}_1 & \boldsymbol{\phi}_2 & \cdots & \boldsymbol{\phi}_N \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{M \times N}$$
and $\boldsymbol{\phi}_i = \phi(\boldsymbol{x}_i)$.

Steps:

1. Show that ϕ must be linear, namely:
$$\phi(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha\phi(\boldsymbol{x}) + \beta\phi(\boldsymbol{y})$$
Hint: Consider $\langle \alpha\boldsymbol{x} + \beta\boldsymbol{y}, \boldsymbol{z} \rangle_{\phi}$ and recall that this is true for all $\boldsymbol{z}$.
2. Show that:
$$\begin{aligned} d_{\phi}^2(\boldsymbol{x}, \boldsymbol{y}) &= \|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})\|_2^2 \\ &= \|\phi(\boldsymbol{x})\|_2^2 - 2\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle + \|\phi(\boldsymbol{y})\|_2^2 \end{aligned}$$
3. Repeat\use the lecture notes to conclude that $-\frac{1}{2}\boldsymbol{J}\boldsymbol{D}_{\phi}\boldsymbol{J} = \boldsymbol{J}\boldsymbol{K}_{\phi}\boldsymbol{J}$.

1. MDS:

1.1
Considering $< \alpha x + \beta y, z>_{\phi} = < \phi(\alpha x + \beta y), \phi(z) >$

And on the other hand,
$< \alpha x + \beta y, z>_{\phi} =_{linearity\ in\ the\ first\ argument} < \alpha x, z>_{\phi} + < \beta y, z>_{\phi} =$
$< \alpha\phi(x), \phi(z)> + < \beta\phi(y), \phi(z)>$

Then, $< \alpha\phi(x), \phi(z)> + < \beta\phi(y), \phi(z)> = < \phi(\alpha x + \beta y), \phi(z) >$

$< \alpha\phi(x) + \beta\phi(y) - \phi(\alpha x + \beta y), \phi(z) > = 0,$ *true for any z, hence,* ϕ *is linear.*

$d^2_{\phi}(x, y) = ||x - y||^2_{\phi} =$

$< \phi(x - y), \phi(x - y)>= < \phi(x) - \phi(y), \phi(x) - \phi(y)>= ||\phi(x) - \phi(y)||^2_2 =$

$||\phi(x)||^2_2 - 2\phi(x)^T\phi(y) + ||\phi(y)||^2_2 =||\phi(x)||^2_2 - 2 < \phi(x)\phi(y) >+ ||\phi(y)||^2_2$

$D_{\phi}[i, j] = d^2_{\phi}(x_i, x_j) = \left\|\phi(x_i)\right\|^2_2 - 2 < \phi(x_i)\phi(x_j) >+ \left\|\phi(x_j)\right\|^2_2$

$D_{\phi} = P - 2\Phi^T\Phi + P^T$, *where* $P =$

$$\begin{bmatrix} ||\phi(x_1)||_2^2 \\ ||\phi(x_2)||_2^2 \\ \vdots \\ ||\phi(x_N)||_2^2 \end{bmatrix} \mathbf{1}_N^T$$

Hence, $- JD_{\phi}J =- J(P - 2\Phi^T\Phi + P^T)J,$ *from the lecture,* $PJ = JP^T = 0$

Therefore, $- JD_{\phi}J = 2\Phi^T\Phi, \ -\frac{1}{2}JD_{\phi}J = JK_{\phi}J$

Consider a training set $\{\boldsymbol{x}_i\}_{i=1}^N$ and let $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ where $\boldsymbol{D}[i, j] = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$.

1.2

Show that $\boldsymbol{v}^T\boldsymbol{D}\boldsymbol{v} < 0$ for any $\boldsymbol{v}$ such that $\langle \boldsymbol{v}, \mathbf{1}\rangle = 0$.

1.2

$< v, 1 >= 0,$ *and* $< v, 1 >= v^T 1_N = 1^T_N v =< 1, v >= 0$

Defining $\widehat{\Lambda}$ *as* $-$

$$\begin{bmatrix} ||x_1||_2^2 \\ ||x_2||_2^2 \\ \vdots \\ ||x_N||_2^2 \end{bmatrix}$$

$v^T Dv = v^T(\Lambda - 2X^TX + \Lambda^T)v = v^T\Lambda v - 2v^TX^TXv + v^T\Lambda^Tv =$

$< v, \Lambda v >- 2(Xv)^TXv +< \Lambda v, v >=< v\widehat{\Lambda}1_N^T v >- 2||Xv||^2_2 -< \widehat{\Lambda}_N^T v, v >$

$\hat{\Lambda}_N^T v = 0,\ hence,\ -2||Xv||_2^2 \leq 0$

MM (Majorization Minimization\Maximization)
Consider:

$$\boldsymbol{Y}[i,j] = \begin{cases} \boldsymbol{X}[i,j] & \boldsymbol{M}[i,j] = 1 \\ 0 & \boldsymbol{M}[i,j] = 0 \end{cases}$$

In other words:

$$\boldsymbol{Y} = \boldsymbol{M} \odot \boldsymbol{X}$$

where $\boldsymbol{M} \in \{0,1\}^{M \times N}$ is a binary mask matrix.
Given $\boldsymbol{Y} \in \mathbb{R}^{M \times N}$, the low-rank matrix completion objective is given by:

$$\begin{cases} \min_{\boldsymbol{X}} \|\boldsymbol{M} \odot (\boldsymbol{Y} - \boldsymbol{X})\|_F^2 \\ \text{s.t.} \\ \text{rank}(\boldsymbol{X}) \leq d \end{cases}$$

Consider the following function:

$$g(\boldsymbol{X}, \boldsymbol{Z}) := \|\boldsymbol{X} - \boldsymbol{Z} + \boldsymbol{M} \odot (\boldsymbol{Z} - \boldsymbol{Y})\|_F^2$$

1.3

Show that g surrogates the objective $f(\boldsymbol{X}) := \|\boldsymbol{M} \odot (\boldsymbol{Y} - \boldsymbol{X})\|_F^2$.
Hint:
Show that $g(\boldsymbol{X}, \boldsymbol{Z}) = \left\|\boldsymbol{M} \odot (\boldsymbol{X} - \boldsymbol{Y}) + \widetilde{\boldsymbol{M}} \odot (\boldsymbol{X} - \boldsymbol{Z})\right\|_F^2$ where $\widetilde{\boldsymbol{M}} := \boldsymbol{1}\boldsymbol{1}^T - \boldsymbol{M}$ is the complement of $\boldsymbol{M}$.

1.3

$g(X,Z) = ||X - Z + M \odot (Z - Y)||_F^2 =$

$\left|\left|11^T \odot X - 11^T \odot Z + M \odot Z - M \odot Y + M \odot Y + M \odot X - M \odot X)\right|\right|_F^2 =$

$\left|\left|M \odot (X - Y) - (11^T - M) \odot Z + (11^T - M) \odot X\right|\right|_F^2 =$

$||M \odot (X - Y) - \overline{M} \odot Z + \overline{M} \odot X||_F^2 =$

$||M \odot (X - Y) + \overline{M} \odot (X - Z)||_F^2$

$a.\ g(X,X) = ||M \odot (X - Y) + 0||_F^2 = f(X)$

$b.\ g(X,X) = ||M \odot (X - Y) + 0||_F^2 = \Sigma_{M[i,j]=1}\sqrt{(X - Y)_{i,j}^2}$

$Moreover,\ Z = \overline{X},\ g(X,\overline{X}) = ||M \odot (X - Y) + \overline{M} \odot (X - \overline{X})||_F^2 =$

$\Sigma_{M[i,j]=1}\sqrt{(X - Y)_{i,j}^2} + \Sigma_{M[i,j]=0}\sqrt{(X - \overline{X})_{i,j}^2}$

$Now,\ \overline{M}\ is\ the\ complement\ of\ M,\ it\ is\ guaranteed\ that\ -\ f(X) \leq g(X,\overline{X})$ for all X and $\overline{X}$.
Thus, g(X,Z) is surrogate.

Metric MDS

The metric MDS objective is given by:

$$\min_{Z \in \mathbb{R}^{d \times N}} \|\boldsymbol{\Delta}_x - \boldsymbol{D}_z\|_F^2$$

where:

- $\boldsymbol{\Delta}_x[i,j] = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a given distance matrix.
- $\boldsymbol{D}_z[i,j] = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$.

Consider the surrogate function:

$$g\left(\boldsymbol{Z}, \tilde{\boldsymbol{Z}}\right) = \|\boldsymbol{\Delta}_x\|_F^2 + 2N\mathrm{Tr}\left\{\boldsymbol{Z}\boldsymbol{J}\boldsymbol{Z}^T\right\} - 4\left\langle \boldsymbol{Z}^T\tilde{\boldsymbol{Z}}, \boldsymbol{B}\right\rangle$$

where:

- $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T$ is the centering matrix.
- $\boldsymbol{B} = \boldsymbol{C} - \mathrm{diag}(\boldsymbol{C}\boldsymbol{1})$
- $\boldsymbol{C}[i,j] = \begin{cases} 0 & i = j \\ -\frac{\boldsymbol{\Delta}_x[i,j]}{\tilde{\boldsymbol{D}}_{\tilde{z}}[i,j]} & i \neq j \end{cases}$
- $\tilde{\boldsymbol{D}}_{\tilde{z}}[i,j] = \|\tilde{\boldsymbol{z}}_i - \tilde{\boldsymbol{z}}_j\|_2$

1.4

Show that:

1.

$$\boldsymbol{B}\boldsymbol{J} = \boldsymbol{B}$$

2.

$$g(\boldsymbol{Z}, \boldsymbol{Z}) = \|\boldsymbol{\Delta}_x - \boldsymbol{D}_z\|_F^2$$

Notes: (See lecture slides)

1. $\|\boldsymbol{\Delta}_x - \boldsymbol{D}_z\|_F^2 = \|\boldsymbol{\Delta}_x\|_F^2 + \|\boldsymbol{D}_z\|_F^2 - 2\langle \boldsymbol{\Delta}_x, \boldsymbol{D}_z\rangle$
2. $\|\boldsymbol{D}_z\|_F^2 = 2N\mathrm{Tr}\left\{\boldsymbol{Z}\boldsymbol{J}\boldsymbol{Z}^T\right\}$

Hint:

For $\tilde{\boldsymbol{Z}} = \boldsymbol{Z}$ we have:

$$\langle \boldsymbol{\Delta}_x, \boldsymbol{D}_z\rangle = -\left\langle \boldsymbol{C}, \boldsymbol{D}_z^{\circ 2}\right\rangle$$

where $\boldsymbol{D}_z^{\circ 2}[i,j] = \boldsymbol{p}\boldsymbol{1}^T - 2\boldsymbol{Z}^T\boldsymbol{Z} + \boldsymbol{1}\boldsymbol{p}^T$ and $\boldsymbol{p} = \begin{bmatrix} \|\boldsymbol{z}_1\|_2^2 \\ \vdots \\ \|\boldsymbol{z}_N\|_2^2 \end{bmatrix}$.

1.1) First let show that $BJ = B$

$$BJ = (C - diag(C1))J = (C - diag(C1))\left(I_N - \frac{1}{N}1_N1_N^T\right) =$$

$$= (C - diag(C1))I_N - (C - diag(C1))\frac{1}{N}1_N1_N^T$$

We will assign:

$$B = (C - diag(C1))I_N$$

$$B' = (C - diag(C1))\frac{1}{N}1_N1_N^T$$

$$C' = C - diag(C1)$$

Lets examine a general element ij:

$$(C'1_N1_N^T)_{ij} = \left(c'_{i1}, c'_{i2}, \ldots, c'_{iN}\right)(1\ 1\ .\ .\ .\ 1) = c'_{i1} + c'_{i2} + \ldots + c'_{iN} + c_{ii}$$

Note that

$$c'_{i1} + c'_{i2} + \ .... + \ c'_{iN} + c_{ii} - \sum_{j=1}^{N} c_{ij} = 0$$

Because of that:

$$B' = \frac{1}{N}\left(C - diag\left(C_1\right)\right)\left(1_N 1_N^T\right)$$

Where $C - diag\left(C_1\right) = 0 \quad B' = 0$

From the last point we can derive that:

$$BJ = B$$

Let show now that: $g(Z, Z) = || \Delta_x - D_z ||_F^2$

Let's examine $< \Delta_x , D_z >$

$$-< \Delta_x , D_z > \ = \ < C, D_z^{\circ 2} > \ = \ < C, p1_N^T - 2Z^TZ + 1_N p^T >$$

Similarly, to what we have saw in lecture:

$$p = diag\left(Z^TZ\right) = \quad < C, diag\left(Z^TZ\right)1_N^T - 2Z^TZ + 1_N diag^T\left(Z^TZ\right) >$$

From symmetric matrix W we know that:

$$< W, Y > \ = \ < W, Y^T >$$

Therefore:

$$< C, 2diag\left(Z^TZ\right)1_N^T - 2Z^TZ > \ =$$

$$= 2\left(< C, diag\left(Z^TZ\right)1_N^T > - < C, Z^TZ >\right)$$

$$= 2\left(< C1_N, diag\left(Z^TZ\right) > - < C, Z^TZ >\right)$$

From equation HW we know that: $< a, diag(X) > \ = \ < diag(a), X >$

$$= 2\left(< diag\left(C1_N\right), Z^TZ > - < C, Z^TZ >\right) = \ - 2 < B, Z^TZ >$$

Now we can open the original equation:

$$|| \Delta_x - D_z ||_F^2 = \left|\left|\Delta_x\right|\right|_F^2 + \left|\left|D_z\right|\right|_F^2 - 2 < \Delta_x, D_z > \ =$$

$$\left|\left|\Delta_x\right|\right|_F^2 + 2N * Tr\left\{ZJZ^T\right\} + 2 < C, D_z^{\circ 2} > \ =$$

$$\left|\left|\Delta_x\right|\right|_F^2 + 2N * Tr\left\{ZJZ^T\right\} - 4 < B, Z^TZ > \ = g(Z, Z)$$

2 Isomap

Let $G = (V, E, W)$ be a simple, undirected, and weighted graph, and assume no negative weights\edges. Let $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ be the shortest path distance matrix, where $N = |V|$.

2.1

Prove or disprove:
Necessarily exists an embedding $\left\{\boldsymbol{z}_i \in \mathbb{R}^d\right\}_{i=1}^N$ (for some $d \in \mathbb{N}$) such that (for all i, j):

$$\boldsymbol{D}[i, j] = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$$

2.1
Assuming that D, the shortest path distance matrix is computed based on graph G, where G is a graph constructed from training data $X \in R^{DxN}$.
Using MDS to solve the embeddings space Z, and for any training dataset $X \in R^{DxN}$ we can choose the embeddings vector size - d, to get the optimal solution Z* that minimizes the problem.

For the optimal solution Z*=X -
$D_x[i, j] = \left|\left|x_i - x_j\right|\right|_2 = \left|\left|z^*_i - z^*_j\right|\right|_2 = D_2^*[i, j]$

- Let $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ be the training set.
- Let $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^N$ be the representation obtained by Isomap (training encoding).
- Consider a new point $\boldsymbol{x}^\star$ where $\boldsymbol{x}^\star = \boldsymbol{x}_k$ for some $k \leq N$.
- Let $\boldsymbol{z}^\star$ be the out of sample encoding applied to $\boldsymbol{x}^\star$.

2.2

Prove of disprove:

$$\boldsymbol{z}^\star = \boldsymbol{z}_k$$

2.2

$x^* = x_k \Rightarrow D_{xx} = D_{xz}$

$\overline{K}_{xz} =- \frac{1}{2}J(D_{xz} - \frac{1}{N_x}D_{xx}1_{N_x}1^T{}_{N_z}) =- \frac{1}{2}J(D_{xx} - \frac{1}{N_x}D_{xx}1_{N_x}1^T{}_{N_x}) =- \frac{1}{2}JD_{xx}(1 - \frac{1}{N_x}1_{N_x}1^T{}_{N_x}) = - \frac{1}{2}D_{xx}J = \overline{K}_{xx}$

Hence, $Z_y = \sum_d^{-1} V_d{}^T\overline{K}_{xz} = \sum_d^{-1} V_d{}^T\overline{K}_{xx} = \sum_d^{-1} V_d{}^TV\Sigma^2V^T = \Sigma_d\, V^T = Z$

Z is the training encoding, hence, $z^* = z_k$

3 Laplacian Eigenmaps

- Consider $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=1}^N$.
- Let $G = (V, E, W)$ be a weighted graph with $V = \mathcal{X}$ and:

$$\boldsymbol{W}[i,j] = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right) & \boldsymbol{x}_i \in \mathcal{N}_j \text{ or } \boldsymbol{x}_j \in \mathcal{N}_i \\ 0 & \text{else} \end{cases}$$

- $e_{ij} \in E$ if $\boldsymbol{W}[i,j] \neq 0$.
- Let $\boldsymbol{Z} \in \mathbb{R}^{d\times N}$ and $\boldsymbol{D}_z \in \mathbb{R}^{N\times N}$ such that $\boldsymbol{D}_z[i,j] = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2$ where $\boldsymbol{z}_i$ is the ith column of $\boldsymbol{Z}$.

3.1

Show that:

$$\frac{1}{2}\langle \boldsymbol{W}, \boldsymbol{D}_z \rangle = \mathrm{Tr}\left\{\boldsymbol{Z}\boldsymbol{L}\boldsymbol{Z}^T\right\}$$

where:

- $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the graph-Laplacian.
- $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{W}\mathbf{1})$ is the degree matrix.

3.1)

Let's examine $Tr\left(ZLZ^T\right)$

$$Tr\left(ZLZ^T\right) = \sum_{i=1}^{N} \left(ZLZ^T\right)_{ii} = \sum_{i=1}^{N} z_i L z_i^T = (I)$$

Notice that z_i is vector in R^N and saw in lecture

For $v \in R^N$, $vLv^T = 0.5 < W, D_v > and\ therefore$:

$$(I) = 0.5 \sum_{i=1}^{N} < W, D_{z_i} > \quad = 0.5 < W, D_z >$$

Assume that G has two connected components, i.e. $V = V_1 \cup V_2$ such that:

$$\left\{ e_{ij} \middle| i \in V_1, j \in V_2 \right\} = \emptyset$$

3.2

Show that the graph-Laplacian $\boldsymbol{L}$ has two **orthogonal** eigenvectors corresponding to the zero eigenvalue. That is, exist $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^N$ such that:

1. $\boldsymbol{L}\boldsymbol{u}_1 = \boldsymbol{L}\boldsymbol{u}_2 = \boldsymbol{0}$
2. $\langle \boldsymbol{u}_1, \boldsymbol{u}_2 \rangle = 0$

3.2

Assuming that there are N points such that $|V| = N$, *and* $|V_1| = N_1$, $|V_2| = N_2$

D is diagonal, V_1 *and* V_2 *are connected components, thus,* $W_{ij} = 0$, *for every* $i \in V_1, j \in V_2$

Forming W as having first N_1 *elements from* V_1, *we get a block of size* $N_1 x N_1$ *at the top left corner that describes the weights of* V_1 *edges. Similarly, we get a block of size* $N_2 x N_2$ *at the bottom right corner of W that describes the weights of* V_2 *edges*.

The rest of W elements are 0. *D is diagonal, hence, L has the same properties mentioned above. Let's note the blocks on L as* L_1, L_2.

Now, we can find a u_1 *that is constructed from* N_1 1*s and* N_2 0*s, and similarly,* u_2 *that is constructed by* N_1 0*s, and* N_2 1*s.*

$\Rightarrow < u_1, u_2 >= 0$, *moreover,* $L1 = 0$, *thus*

for each $0 \le i \le N_1$ $\sum_{j=0}^{N-1} L_{ij} = \sum_{j=0}^{N_1-1} L_{ij} u_1[f] = 0$, *the multiplication for each of the first* N_1 *rows with* u_1 *is* 0, *the multiplication for each of the last* N_2 *rows is* 0 *by definition.*

for each $N_i \le i \le N$ $\sum_{j=0}^{N-1} L_{ij} = \sum_{j=N_1}^{N-1} L_{ij} u_2[f] = 0$, *the multiplication for each of the first* N_1 *rows with* u_1 *is* 0, *the multiplication for each of the first* N_1 *rows is* 0 *by definition.*

Thus, u_1, u_2 *are eigenvectors with* 0 *eigenvalues.*

4 t-SNE

The t-SNE objective is given by:

$$\min_{\mathbf{Z}\in\mathbb{R}^{d\times N}} \underbrace{D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q})}_{:=f(\mathbf{Z})} = \min_{\mathbf{Z}\in\mathbb{R}^{d\times N}} \sum_{i=1}^{N}\sum_{j=1}^{N} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right)$$

See the definitions of $\mathbf{P}$ and $\mathbf{Q}$ in the lecture notes (do not get confused by the SNE definitions). The goal of this question is to compute the gradient of the objective:

$$\nabla f(\mathbf{Z}) = ?$$

Let us break this task into several **smaller** steps.

4.1

Show that $f(\mathbf{Z}) = D_{\mathrm{KL}}(\mathbf{P}||\mathbf{Q})$ can be written as:

$$f(\mathbf{Z}) = C - \langle \mathbf{P}, \log[\mathbf{Q}]\rangle$$

where C is some constant (the entropy of $\mathbf{P}$).

4.1) let's look at $f(z)$:

$$f(z) = D_{KL}(P||Q) = \sum_{i=1}^{N}\sum_{j=1}^{N} p_{j|i} lo\left(\frac{p_{j|i}}{q_{j|i}}\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} p_{j|i}\left(log\left(p_{j|i}\right) - log\left(q_{j|i}\right)\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} p_{j|i} log\left(p_{j|i}\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} p_{j|i} log\left(q_{j|i}\right)$$

$C - < P,\ log[Q] >$, where c is the entropy of P

4.2

Show that:

1.

$$\boldsymbol{B} = \mathbf{1}^T (\boldsymbol{S} - \boldsymbol{I}) \mathbf{1} \in \mathbb{R}$$

2.

$$\boldsymbol{Q} = B^{-1} (\boldsymbol{S} - \boldsymbol{I}) \in \mathbb{R}^{N \times N}$$

- Reminder:

$$\boldsymbol{Q}[i,j] = \frac{1}{B} \begin{cases} 0 & i = j \\ (1 + \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2)^{-1} & i \neq j \end{cases}$$

- Let $\boldsymbol{D}_z \in \mathbb{R}^{N \times N}$ such that $\boldsymbol{D}_z[i,j] = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2$.
- Let $\boldsymbol{S} = (\mathbf{1}\mathbf{1}^T + \boldsymbol{D}_z)^{\circ -1} \in \mathbb{R}^{N \times N}$, that is:

$$\boldsymbol{S}[i,j] = (1 + \boldsymbol{D}_z[i,j])^{-1}$$

4.2.1)

- in the lecture we seen the formula for B

$$B = \sum_{i=1}^{N} \sum_{i!=j} (1 + ||z_i - z_j)||)^{-1}$$

$$= \sum_{i=1}^{N} \sum_{i!=j} (1 + ||z_i - z_j)||)^{-1} \pm \sum_{i=1}^{N} \sum_{i==j} (1 + ||z_i - z_j)||)^{-1}$$

$$\sum_{i=1}^{N} \sum_{j=1}^{N} (1 + ||z_i - z_j)||)^{-1} - N$$

$$= 1^T S11 - 1^T I1 = 1^T (S - I)1$$

- notice that:
  - $1 \in R^{Nx1}$
  - $S \in R^{NxN}$
  - $B \in R^{1x1}$

4.2.2)

- Let look about W, where W$\in R^{NxN}$

$$W = S - I$$

- Let look on 2 cases
  - $i = j$:

$$W[i,j] = S - I[i,i] = S[i,i] - I[i,i]$$
$$= (1 + ||z_i - z_j)||)^{-1} - 1 = 1 - 1 = 0$$

    - therefore, we can infer that:

$$W = BQ \rightarrow Q = \frac{1}{B} W = \frac{1}{B}(S - I)$$

    - because W$\in R^{NxN}$ multiplying by scalar does not change its dimension, so Q$\in R^{NxN}$.

  - $i! = j$

$$W[i,j] = S - I[i,j] = S[i,j] - I[i,j]$$
$$(1 + ||z_i - z_j)||)^{-1} - 0 = (1 + ||z_i - z_j)||)^{-1}$$

4.3

Show that:

$$-\langle \boldsymbol{P}, \log[\boldsymbol{Q}]\rangle = \log(B) + \left\langle \boldsymbol{P}, \log\left[\mathbf{1}\mathbf{1}^T + \boldsymbol{D}_z\right]\right\rangle$$

Hints:

- $\boldsymbol{P}[i,i] = ?$
- $\mathbf{1}^T \boldsymbol{P} \mathbf{1} = ?$

4.3)

- Let look at $-< P,\ log[Q] >$:

$$-\ (< P, \log log\ [Q]\ >\ =\ -< P, \log log\left[B^{-1}(S - I)\right]\ >)$$
$$=-\ (< P, \log log\ [(S - I)] - \log log\left[B11^T\right]\ >)$$
$$=\ -\ (< P, \log log\ [(S - I)] >-< P, \log log\left[B11^T\right]\ >)$$
$$-\ (< P, \log log\ [(S - I)] >-\ \sum_{i=1}^{N}\sum_{j=1}^{N} p_{j|i} lo(B)\ )$$
$$=-\ (< P, \log log\ [(S - I)] >-\ log(B)\sum_{i=1}^{N}\sum_{j=1}^{N} p_{j|i})$$
$$=-\ (< P, \log log\ [(S - I)] >-\ \log log\ (B)\ 1)$$
$$=-< P, \log log\ [(S - I)] >+\ \log log\ (B)\ \ = (*)$$

- Now let examine $P_{i,i}xlog[S - I]_{i,j}$, for 2 cases:
  - $i! = j$

$$P_{i,j}xlog[S - I]_{i,j} = P_{i,j}xlog[S]_{i,j}$$

- We can conclude that:

$$< P, \log log\ [S - I]\ >\ =\ < P, log[S] >$$

- Now let return to the (*)

$$(*) =\ -< P, \log log\ [(S - I)] >+\ \log log\ (B)$$
$$-< P, \log log\ [(S)] >+\ \log log\ (B)$$
$$-< P, \log log\left[\left(11^T + D_z\right)^{\circ-1}\right] >+\ \log log\ (B)$$

- Know let use the following rule:

$$\log log\left(a^b\right)\ = blog(a)$$

- Then we yield:

$$=< P, \log log\left[\left(11^T + D_z\right)^{\circ-1}\right] >+\ \log log\ (B)$$

  - $i = j$

$$P_{i,i}xlog[S - I]_{ii} = 0 * \log log\ [S - I]_{i,i} = 0 = P_{i,i}xlog[S]_{i,i}$$

Let:

$$f(\boldsymbol{Z}) = C + \underbrace{\log(B)}_{(*)} + \underbrace{\langle \boldsymbol{P}, \log[\mathbf{1}\mathbf{1}^T + \boldsymbol{D}_z] \rangle}_{(**)}$$

4.4

Show that:

1.

$$\nabla_{\boldsymbol{Z}} \underbrace{\langle \boldsymbol{P}, \log[\mathbf{1}\mathbf{1}^T + \boldsymbol{D}_z] \rangle}_{(**)} [\boldsymbol{H}] = \langle \boldsymbol{S} \circ \boldsymbol{P}, \nabla \boldsymbol{D}_z [\boldsymbol{H}] \rangle$$

2.

$$\nabla_{\boldsymbol{Z}} \underbrace{\log(B)}_{(*)} [\boldsymbol{H}] = -\langle \boldsymbol{S} \circ \boldsymbol{Q}, \nabla \boldsymbol{D}_z [\boldsymbol{H}] \rangle$$

Hints:

- $\nabla \boldsymbol{S}[\boldsymbol{H}] = \nabla (\mathbf{1}\mathbf{1}^T + \boldsymbol{D}_z)^{\circ -1} [\boldsymbol{H}] = -(\mathbf{1}\mathbf{1}^T + \boldsymbol{D}_z)^{\circ -2} \circ \nabla (\boldsymbol{D}_z)[\boldsymbol{H}] = -\boldsymbol{S} \circ \boldsymbol{S} \circ \nabla(\boldsymbol{D}_z)[\boldsymbol{H}]$
- $\boldsymbol{Q} = B^{-1}(\boldsymbol{S} - \boldsymbol{I})$

4.4.0.1) help prove for 4.4.1

- If D is diagonal matrix, and we want to calculate the product $< D, D_Z > \ = f(z) =?$

$$L(z) < D, D_Z > \ = Tr\left(D^T D_z\right) = \sum_1^N D_{ii} D_z[i,i] = \sum_1^N D_{ii} D_z[i,i] = \sum_1^N D_{ii} 0 \ = \ 0$$

- Moreover, we know from that that $\nabla < D, D_Z > \ = 0$

4.4.0.2) Let express $\nabla L(z)$

$$\nabla L(z) = \ \nabla < D,\ diag\left(Z^T Z\right) 1^T - 2Z^T Z + 1(diag^T(Z^T Z)) >$$

- ○ We saw in last homework that $Z \in R^{dXN}$, $D \in \ R^{NXN}$

$$\nabla < D,\ diag\left(Z^T Z\right) 1^T - 2Z^T Z + 1\left(diag^T\left(Z^T Z\right)\right) > \ = \ -2 < D - diag(D1),\ Z^T Z >$$

- Let calculate the gradient of $Z^T Z$:

$$\nabla_g (Z)[H] = \ \nabla\left(Z^T Z\right)[H] = \frac{Z^T Z + tH^T Z + tZ^T H + t^2 H^T H - Z^T Z}{t}$$

$$H^T Z + Z^T H + tH^T H \ = \ tH^T H) + \ H^T Z + Z^T H \ = H^T Z + Z^T H$$

- $\nabla < D,\ diag\left(Z^T Z\right) 1^T - 2Z^T Z + 1\left(diag^T\left(Z^T Z\right)\right) > \ = \ -2 < D - diag(D1),\ Z^T Z >$

$$\nabla L(z)[H] = - 2(< 0,\ Z^T Z > + < D - diag(D1), \nabla(Z^T Z)[H] >)$$

$$= \ -2\left(< D - diag(D1),\ H^T Z + Z^T H >\right)$$

$$= \ -2(< D - diag(D1),\ H^T Z > + < D - diag(D1), Z^T H >)$$

- For 2 NXN matrices A, B we know that:

$$< A, B > \ = \ < A^T,\ B^T >$$

- Now let's use this fact

$$-2\left(< D^T - diag^T(D1),\ Z^T H > + < D - diag(D1), Z^T H >\right)$$

- Because D is symmetric $D = D^T,\ and\ diag^T(D1) = diag(D1)$

$$= - 4 < D - diag(D1), Z^T H > \ = 4Z < diag(D1) - D, H >$$

- Therefore $\nabla L(z) =$

$$\nabla L(z) = 4Z(diag(D1) - D)$$

4.4.0.2) help prove for 4.4.2

- If D is diagonal matrix, and we want to calculate the product $< D, D_Z > \ = ?$

$$< D, D_Z > \ = Tr\left(D^T D_z\right) = \sum_1^N D_{ii} D_z[i, i] = \sum_1^N D_{ii} D_z[i, i] = \sum_1^N D_{ii} 0 = 0$$

- Moreover, we know from that that $\nabla < D, D_Z > \ = 0$

4.4.1)

$$\nabla < P, \log log\left[11^T + D_z\right] > [H] = \ ?$$

- Let's use the product rule first. And then the chain rule
  - product rule:

$$< P[H],\ log[11^T + D_z]] > \ + \ < P, \nabla(log[11^T + D_z])[H] >$$

$$= 0 + \ < P, \nabla(log[11^T + D_z])[H] >$$

  - chain rule:

$$< P,\ < \left[11^T + D_z\right]^{\circ -1}, \nabla D_z[H] \gg \ = \ < P,\ S \circ \nabla D_z[H] >$$

- now let use the first hint:

$$(P \circ S,\ \nabla(D_z)[H])$$

- missing part 1.4 …
- now let use the second hint + the knowledge that S, P are symmetric + 4.4.0.2:

$$4Z < diag(P \circ S)1\rightarrow - (P \circ S), H >$$

- $therefore\ we\ know\ that$:

$$\nabla < P, \log log\left[11^T + D_z\right] > [H]) = 4Z(diag(P \circ S)1\rightarrow - (P \circ S))$$

4.4.2)

$$\nabla \log log\ (B) \ = \log log\ \left(1^T(S - I)1\right) \ = \ ?$$

- Let's look on $\nabla B[H]$:

$$\nabla B[H] = \nabla\left(1^T(S - I)1\right)[H] = \nabla\left(1^T S1 - 1^T I1\right)[H]$$
$$= \nabla\left(1^T S1)[H] - \nabla(1^T I1\right)[H] = \nabla\left(1^T S1)[H] - (0\right) = \nabla\ (1^T S1)[H]$$

- $know\ by\ using\ multiple\ time\ the\ product\ rule$:

$$\left(1^T[H]S1 + 1^T(\nabla S1)[H]\right) = 0 - 1^T(\nabla S1[H]) = 1^T(\nabla S1[H])$$
$$= 1^T((\nabla S[H])1 + S(\nabla 1[H])) = 1^T((\nabla S[H])1 + 0) = 1^T\nabla S[H]1$$

- $now\ lets\ use\ the\ the\ first\ hint$:

$$- 1^T\left(S \circ S \circ \nabla\left(D_Z\right)[H]1\right) = \ - < 11^T,\ S \circ S \circ \nabla\left(D_Z\right)[H] >$$
$$= \ - < S\circ S,\ \nabla\left(D_Z\right)[G] >$$

- Let's do trick of adding and subtracting I from S:

$$- < (S - I + I)\circ S,\ \nabla\left(D_Z\right)[H] > \ = \ - < ((S - I)\circ S,\ \nabla\left(D_Z\right)[H] > \ - < (I\circ S,\ \nabla\left(D_Z\right)[H] >$$

- $- < (I\circ S,\ \nabla\left(D_Z\right)[H] > \ = 0\ follow(**)\ from$ 4.4.0.1

$$- < \left((S - I)\circ S,\ \nabla\left(D_Z\right)[H] > - \ 0\right) = \ - < ((S - I)\circ S,\ \nabla\left(D_Z\right)[H] >$$

- Finally, we can show the results of $\nabla \log log\ (B)[H] \ = \ ?$

$$\nabla \log log\ (B)\ [H] = \frac{1}{B}\nabla(B)[H] = - \frac{1}{B}\left(< (S - I)\circ S, \nabla\left(D_z\right)[H] >\right)$$
$$= \ < Q \circ S,\ \nabla\left(D_z\right)[H] >$$
$$= \ - 4Z(diag(Q\circ S1) - Q\circ S)$$

4.5

- Combine all previous results and write the gradient of the objective:

$$\nabla f(\boldsymbol{Z}) = ?$$

- Use $\boldsymbol{A} := (\boldsymbol{P} - \boldsymbol{Q}) \circ \boldsymbol{S}$ to simplify your answer.
- What can you say about the gradient $\nabla f(\boldsymbol{Z})$ when $\boldsymbol{P} = \boldsymbol{Q}$?

Hint: Use the lecture notes.

4.5)

$$\nabla f(Z) = \nabla(C - < P, \log log\, [Q]\ >) = \nabla\left(\log log\,(B)\ + < P, \log log \left[11^T + D_z\right]\ >\right)$$

$$= -4z(diag(Q \circ S1) - Q \circ S) + 4Z(diag(P \circ S1) - P \circ S)$$

$$4z(diag(P - Q) \circ S1 - (P - Q) \circ S1) = 4Z(diag(A1) - A)$$

- notice that in this case $P = Q$
  - gradient will be 0 because A will be 0
  - the same for diag(A1)