Exercise 1: Sparse Coding (20 + 20 P)

Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N examples. Let $\boldsymbol{s}_i \in \mathbb{R}^h$ be the source associated to example $\boldsymbol{x}_i$, and W be a matrix of size $d \times h$ that linearly reconstructs the examples from the sources. We wish to minimize the objective:

$$J = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i - W\boldsymbol{s}_i\|^2}_{\text{reconstruction}} + \lambda \cdot \underbrace{\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{s}_i\|_1}_{\text{sparsity}} + \underbrace{\eta \cdot \|W\|_F^2}_{\text{regularization}}$$

with respect to the weights W and the sources $\boldsymbol{s}_1, \dots, \boldsymbol{s}_N$. The objective consists of three terms: The reconstruction term is the standard mean square error, the sparsity term consists of a standard L_1 penalty on the sources, and the last regularization term prevents the sparsity term from becoming ineffective.

(a) *Show* that for fixed sources, the optimal matrix W is given in closed form as:

$$W = \Sigma_{XS}(\Sigma_{SS} + \eta I)^{-1}$$

where

$$\Sigma_{XS} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{s}_i^\top \quad \text{and} \quad \Sigma_{SS} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s}_i \boldsymbol{s}_i^\top.$$

$$J(W) = \frac{1}{N}\sum_i (x_i - Ws_i)^T I (x_i - Ws_i) + cst. + \eta \|W\|_F^2$$

$$= \frac{1}{N}\sum_i\sum_j -2x_{ij} w_j^T s_i + w_j^T s_i s_i^T w_j + cst + \eta \|W\|_F^2$$

$$\frac{\partial J}{\partial W} = \frac{1}{N}\sum_i -2(x_i - Ws_i)s_i^T + \eta 2W \overset{!}{=} 0$$

$$\Leftrightarrow W\Big(\underbrace{\frac{1}{N}\sum_i s_i s_i^T}_{\Sigma_{SS}} + \eta I\Big) = \underbrace{\frac{1}{N}\sum_i x_i s_i^T}_{\Sigma_{XS}}$$

$$\Leftrightarrow W = \Sigma_{XS}(\Sigma_{SS} + \eta I)^{-1}$$

(b) We now consider the optimization of sources. Due to the 1-norm in the sparsity term, we cannot find a closed form solution. However, we consider a local relaxation of the optimization problem where the 1-norm of the sparsity term is linearized as

$$\|\boldsymbol{s}_i\|_1 = \boldsymbol{q}_i^\top \boldsymbol{s}_i$$

with $\boldsymbol{q}_i \in \{-1, 0, 1\}^d$ a constant vector. This relaxation makes the objective function quadratic with $\boldsymbol{s}_i$.

Show that under this local relaxation, the solution of the optimization problem is given in closed form as:

$$\boldsymbol{s}_i = (W^\top W)^{-1}(W^\top \boldsymbol{x}_i - \lambda \cdot \boldsymbol{q}_i/2)$$

Although this solution is not the true minimum of J (e.g. it is not sparse), it can serve as the end-point of some line-search method for finding good source vectors $\boldsymbol{s}_i$.

$$\frac{\partial J}{\partial s_i} = \frac{1}{N}\cdot\left(-2W^T(x_i - Ws_i)\right) + \frac{1}{N}\lambda q_i \overset{!}{=} 0$$

$$\Leftrightarrow W^T W s_i = W^T x_i - \frac{\lambda}{2} q_i$$

$$\Leftrightarrow s_i = (W^T W)^{-1}(W^T x_i - \lambda \cdot q_i / 2)$$

Exercise 2: Auto-Encoders (20 P)

In this exercise, we would like to show an equivalence between linear autoencoders with tied weights (same parameters for the encoder and decoder) and PCA. We consider the special case of an autoencoder with a single hidden unit. In that case, the autoencoder consists of the two layers:

$$s_i = \boldsymbol{w}^\top \boldsymbol{x}_i \qquad \text{(encoder)}$$
$$\hat{\boldsymbol{x}}_i = \boldsymbol{w} \cdot s_i \qquad \text{(decoder)}$$

where $\boldsymbol{w} \in \mathbb{R}^d$. We consider a dataset $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ assumed to be centered (i.e. $\sum_i \boldsymbol{x}_i = 0$), and we define the objective that we would like to mininize to be the mean square error between the data and the reconstruction:

$$J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2$$

Furthermore, to make the objective closer to PCA, we can always rewrite the weight vector as $\boldsymbol{w} = \alpha \boldsymbol{u}$ where $\boldsymbol{u}$ is a unit vector (of norm 1) and α is some positive scalar, and search instead for the optimal parameters $\boldsymbol{u}$ and α.

(a) *Show* that the optimization problem can be equally rewritten as

$$\arg\min_{\alpha,\boldsymbol{u}} J(\boldsymbol{w}) = \arg\max_{\alpha,\boldsymbol{u}} \boldsymbol{u}^\top S \boldsymbol{u}$$

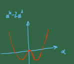
where $S = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top$, which is a common formulation of PCA.

$$J(w) = \frac{1}{N}\sum_i \|x_i - ww^T x_i\|^2 = \frac{1}{N}\sum_i \|x_i - \alpha^2 uu^T x_i\|^2$$

$$= \frac{1}{N}\sum_i \|x_i\|^2 - 2\alpha^2 x_i^T uu^T x_i + \alpha^4 x_i^T u \underbrace{u^T u}_{=1} u^T x_i$$

$$= \frac{1}{N}\sum_i (\alpha^4 - 2\alpha^2) \cdot u^T x_i \, x_i^T u + cst.$$

$$= \underbrace{(\alpha^4 - 2\alpha^2)}_{\min_\alpha(\cdot) = -1} \cdot \underbrace{u^T\Big[\frac{1}{N}\sum_i x_i x_i^T\Big]u}_{\geq 0 \ (S \text{ is p.s.d.})}$$



$$\Rightarrow \arg\min_{\alpha,u} J = \arg\max_{\alpha,u} u^T S u$$