Exercise 1: Lagrange Multipliers (10 + 10 P)

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N data points. We consider the objective function

$$J(\boldsymbol{\theta}) = \sum_{k=1}^{N} ||\boldsymbol{\theta} - \boldsymbol{x}_k||^2$$

to be minimized with respect to the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. In absence of constraints, the parameter $\boldsymbol{\theta}$ that minimizes this objective is given by the empirical mean $\boldsymbol{m} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k$. However, this is generally not the case when the parameter $\boldsymbol{\theta}$ is constrained.

(a) Using the method of Lagrange multipliers, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to the constraint $\boldsymbol{\theta}^\top \boldsymbol{b} = 0$, with $\boldsymbol{b}$ some unit vector in $\mathbb{R}^d$. Give a geometrical interpretation to your solution.

$$\begin{aligned}
\operatorname*{argmin}_\theta J(\theta) &= \operatorname*{argmin}_\theta \sum_{k=1}^{N} \|\theta\|^2 - 2\theta^T x_k + \|x_k\|^2 = \operatorname*{argmin}_\theta N\cdot\|\theta\|^2 - 2\theta^T m \cdot N \\
&= \operatorname*{argmin}_\theta \|\theta\|^2 - 2\theta^T m = \operatorname*{argmin}_\theta \tfrac{1}{2}\left[\|\theta\|^2 - 2\theta^T \tilde{x} + \|m\|^2\right] \\
&= \operatorname*{argmin}_\theta \tfrac{1}{2}[\theta - m]^2
\end{aligned}$$

$$\mathcal{L}(\theta) = J(\theta) + \lambda \cdot g(x) = \tfrac{1}{2}[\theta - m]^2 + \lambda \cdot \theta^T b$$

$$\nabla_\theta \mathcal{L}(\theta) = \theta - m + \lambda \cdot b = 0 \quad \Rightarrow \quad \theta b^T - b^T m + \lambda b^T b = 0 \quad \Rightarrow \quad \lambda = b^T m$$

① 先得到 λ 和别的参数的关系

② 把 λ 代回得到 θ 的结果

$$\Rightarrow \quad \theta - m + b^T m = 0$$

$$\Rightarrow \quad \theta = m - b^T m\, b$$

(b) Using the same method, *find* the parameter $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ subject to $||\boldsymbol{\theta} - \boldsymbol{c}||^2 = 1$, where $\boldsymbol{c}$ is a vector in $\mathbb{R}^d$ different from $\boldsymbol{m}$. Give a geometrical interpretation to your solution.

$$\mathcal{L}(\theta) = \tfrac{1}{2}[\theta - m]^2 + \tfrac{1}{2}\lambda \cdot \left(\|\theta - c\|^2 - 2\right)$$

$$\nabla_\theta \mathcal{L}(\theta) = \theta - m + (\theta - c)\cdot\lambda = 0 \quad \Rightarrow \quad (1+\lambda)(\theta - c) = m - c$$

$$\Rightarrow \quad (1+\lambda)^2 \|\theta - c\|^2 = \|m - c\|^2$$

$$\Rightarrow \quad |1+\lambda| = \pm \|m - c\|$$

$$\Rightarrow \quad (\theta - c) = \pm \frac{m - c}{\|m - c\|}$$

$$\Rightarrow \quad \theta = c \pm \frac{m - c}{\|m - c\|}$$

Exercise 2: Principal Component Analysis (10 + 10 P)

We consider a dataset $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$. Principal component analysis searches for a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ such that projecting the data on that vector produces a distribution with maximum variance. Such vector can be found by solving the optimization problem:

$$\arg\max_{\boldsymbol{u}} \frac{1}{N} \sum_{k=1}^{N} \Big[\boldsymbol{u}^\top \boldsymbol{x}_k - \frac{1}{N} \Big(\sum_{l=1}^{N} \boldsymbol{u}^\top \boldsymbol{x}_l \Big) \Big]^2 \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

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

$$\arg\max_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{S} \boldsymbol{u} \quad \text{with} \quad \|\boldsymbol{u}\|^2 = 1$$

where $\boldsymbol{S} = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^\top$ is the scatter matrix, and $\boldsymbol{m} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{x}_k$ is the empirical mean.

$$\arg\max_u \frac{1}{N} \sum_{k=1}^{N} \Big[u^T x_k - \frac{1}{N} \Big(\sum_{l=1}^{N} u^T x_l \Big) \Big]^2$$

$$= \arg\max_u \frac{1}{N} \sum_{k=1}^{N} \Big[u^T \Big(x_k - \frac{1}{N} \sum_{l=1}^{N} x_l \Big) \Big]^2$$

$$= \arg\max_u \frac{1}{N} \sum_{k=1}^{N} \big[u^T (x_k - m) \big]^2$$

$$= \arg\max_u u^T \sum_{k=1}^{N} (x_k - m)^2 u$$

$$= \arg\max_u u^T \sum_{k=1}^{N} (x_k - m)(x_k - m)^T u$$

(b) *Show* using the method of Lagrange multipliers that the problem above can be reformulated as solving the eigenvalue problem

$$\boldsymbol{S}\boldsymbol{u} = \lambda \boldsymbol{u}$$

and retaining the eigenvector $\boldsymbol{u}$ associated to the highest eigenvalue λ.

$$J(u) = u^T \sum_{k=0}^{N} (x_k - m)(x_k - m)^T u = u^T S u$$

$$g(u) = \|u\|^2 - 1$$

$$\mathcal{L}(u) = u^T S u - \lambda (\|u\|^2 - 1)$$

$$\nabla_u \mathcal{L}(u) = 2Su - 2\lambda u = 0 \quad \Rightarrow \quad Su = \lambda u$$

$$u^T S u = u^T \lambda u = u^T u \lambda = \lambda$$

$$\max u^T S u = \max \lambda$$

Exercise 3: Bounds on Eigenvalues (5 + 5 + 5 + 5 P)

Let λ_1 denote the largest eigenvalue of the matrix $\boldsymbol{S}$. The eigenvalue λ_1 quantifies the variance of the data when projected on the first principal component. Because its computation can be expensive, we study how the latter can be bounded with the diagonal elements of the matrix $\boldsymbol{S}$.

(a) *Show* that $\sum_{i=1}^{d} \boldsymbol{S}_{ii}$ is an upper bound to the eigenvalue λ_1.

$$\sum_{i=1}^{d} S_{ii} = \mathrm{Tr}(S) = \sum_{i=1}^{d} \lambda_i \geq \lambda_1$$

(b) *State* the conditions on the data for which the upper bound is tight.

$$\lambda_2, \dots, \lambda_d = 0$$

$$\Rightarrow \mathrm{Var}(u_i^T x) = 0 \qquad \forall i \in \{2, \dots, d\}$$

(c) *Show* that $\max_{i=1}^{d} \boldsymbol{S}_{ii}$ is a lower bound to the eigenvalue λ_1.

$$\lambda_1 = u_1^T S u_1 = \max_u u^T S u \geq \max_{i=1}^{d} e_i^T S e_i = \max_{i=1}^{d} S_{ii}$$

(d) *State* the conditions on the data for which the lower bound is tight.

$$u_1 \in \{e_1, \dots, e_d\}$$

Exercise 4: Iterative PCA (10 P)

When performing principal component analysis, computing the full eigendecomposition of the scatter matrix $\boldsymbol{S}$ is typically slow, and we are often only interested in the first principal components. An efficient procedure to find the first principal component is *power iteration*. It starts with a random unit vector $\boldsymbol{w}^{(0)} \in \mathbb{R}^d$, and iteratively applies the parameter update

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{S}\boldsymbol{w}^{(t)} / \left\|\boldsymbol{S}\boldsymbol{w}^{(t)}\right\|$$

until some convergence criterion is met. Here, we would like to show the exponential convergence of power iteration. For this, we look at the error terms

$$\mathcal{E}_k(\boldsymbol{w}) = \left|\frac{\boldsymbol{w}^\top \boldsymbol{u}_k}{\boldsymbol{w}^\top \boldsymbol{u}_1}\right| \qquad \text{with} \quad k = 2, \ldots, d,$$

and observe that they should all converge to zero as $\boldsymbol{w}$ approaches the eigenvector $\boldsymbol{u}_1$ and becomes orthogonal to other eigenvectors.

(a) Show that $\mathcal{E}_k(\boldsymbol{w}^{(T)}) = |\lambda_k/\lambda_1|^T \cdot \mathcal{E}_k(\boldsymbol{w}^{(0)})$, i.e. the convergence of the algorithm is exponential with the number of time steps T. *(Hint: to show this, it is useful to rewrite the scatter matrix in terms of eigenvalues and eigenvectors, i.e.* $\boldsymbol{S} = \sum_{i=1}^d \boldsymbol{u}_i \boldsymbol{u}_i^\top \lambda_i$*.)*

$$\mathcal{E}_k(w^{(T)}) = \left|\frac{w^{(T)T} u_k}{w^{(T)T} u_1}\right| = \left|\frac{S w^{(T-1)T} \cdot u_k}{S w^{(T-1)T} \cdot u_1}\right| = \left|\frac{w^{(t-1)T} \cdot \sum_{i=1}^{d} u_i u_i^T \lambda_i \cdot u_k}{w^{(t-1)T} \cdot \sum_{i=1}^{d} u_i u_i^T \lambda_i \cdot u_1}\right|$$

$$= \left|\frac{w^{(t-1)T} \sum_{i=1}^{d} u_i \, \delta_{i,k} \, \lambda_i}{w^{(t-1)T} \sum_{i=1}^{d} u_i \, \delta_{i,1} \, \lambda_1}\right|$$

$$= \left|\frac{w^{(t-1)T} \cdot u_k \cdot \lambda_k}{w^{(t-1)T} \cdot u_1 \cdot \lambda_1}\right|$$

$$= \mathcal{E}_k(w^{(T-1)}) \cdot \left|\frac{\lambda_k}{\lambda_1}\right|$$

$$\Rightarrow \mathcal{E}_k(w^{(T)}) = \mathcal{E}_k(w^{(0)}) \cdot \left|\frac{\lambda_k}{\lambda_1}\right|^T$$