# Sheet 3 Theory

## Exercise 1: Lagrange Multipliers

Let $x_1, \dots, x_n \in \mathbb{R}^d$ be a dataset of $n$ samples. We consider the objective function

$$J(\theta) = \sum_{k=1}^{n} \|\theta - x_k\|^2$$

to be minimized with respect to the parameter $\theta \in \mathbb{R}^d$. It can be shown that in absence of constraints for $\theta$, the parameter $\theta^*$ that minimizes this objective is given by the empirical mean

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

However this is not necessarily the case when the parameter $\theta$ is constrained

---

General Lagrange Multiplier

We want to maximize (minimize) a multivariable fct (objective function e.g. $J(\theta)$) $f(x,y,\dots)$ to the constraint that another multivariable function equals a constant $g(x,y,\dots) = c$

1. Step: Lagrangian function $\mathcal{L}$:

$$\mathcal{L}(x,y,\dots,\lambda) = f(x,y,\dots) - \lambda\big(g(x,y,\dots) - c\big)$$ !

2. Step: Set the gradient of $\mathcal{L}$ equal to the zero vector:

$$\nabla \mathcal{L}(x,y,\dots,\lambda) = \vec{0} = \begin{bmatrix} \frac{\partial \mathcal{L}(x,y,\dots,\lambda)}{\partial x} \\ \frac{\partial \mathcal{L}(x,y,\dots,\lambda)}{\partial y} \\ \vdots \\ \frac{\partial \mathcal{L}(x,y,\dots,\lambda)}{\partial \lambda} \end{bmatrix}$$

3. Step: Consider each solution, which will look something like $(x_0, y_0, \dots, \lambda_0)$. Plug each one into $f$

Whichever one gives the greatest (or smallest) value is the maximum (or minimum) point you are seeking.

---

a.) Using the method of Lagrange multipliers, find the parameter $\theta$ that minimize $J(\theta)$ subject to the constraint $\theta^T b = 0$ where $b \in \mathbb{R}^d$.
Give a geometrical interpretation to your solution

$$J(\theta) = \sum_{k=1}^{n} \|\theta - x_k\|^2$$
$$= \sum_{k=1}^{n} \|\theta\|^2 - 2\theta^T x_k + \|x_k\|^2$$
$$= n \cdot \theta^T\theta - 2\theta \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^T x_k$$

$$\mathcal{L}(\theta,\lambda) = n \cdot \theta^T\theta - 2\theta \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^T x_k - \lambda(\theta^T b - 0)$$

Foil rule with $u, v \in \mathbb{R}^n$

$$\|u-v\|^2 = (u-v)^T(u-v)$$
$$= u^Tu - u^Tv - v^Tu + v^Tv$$
$$= \|u\|^2 - 2u^Tv + \|v\|^2$$

with $u^Tv = v^Tu = \sum_{i=1}^{n} u_i v_i$

$\nabla \mathcal{L}(\theta, \lambda) = \vec{0}$

important derivations (b → const, x → vector, B → matrix)

$x^T B \to B$
$x^T b \to b$
$x^T x \to 2x$
$x^T B x \to 2Bx$

$$\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta} = 2n\theta - 2\underbrace{\sum_{k=1}^{n} x_k}_{n\bar{x}} - \lambda b = 0$$

nach $\theta$ umstellen

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

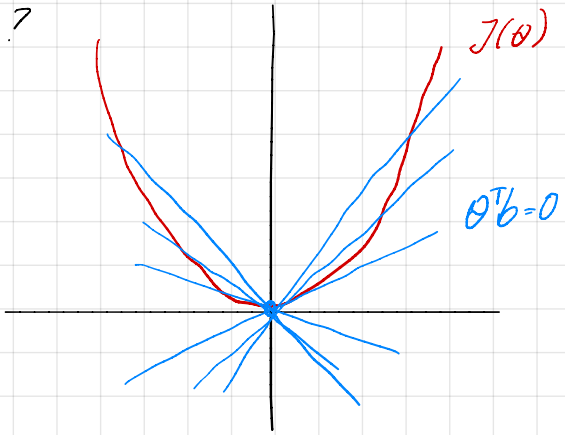
$$\theta^* = \frac{2n\bar{x} + \lambda b}{2n} = \bar{x} + \frac{\lambda}{2n} b$$

$$\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \lambda} = -\theta^T b = 0 = b^T\theta = b^T \cdot \left(\frac{2n\bar{x} + \lambda b}{2n}\right) = 0$$

$$\to \lambda = -\frac{2n\bar{x}b^T}{b^T b}$$

Plug $\lambda$ in $\theta^*$: $\theta^* = \bar{x} + \frac{1}{2n}\left(-\frac{2n\bar{x}b^T}{b^T b} b\right) = \bar{x} - \bar{x}\cdot b^T (b^T b)^{-1} b$

Geometrical interpretation?




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

$$g(\theta) = \|\theta - c\|^2 - 1 = \theta^T\theta - 2\theta^T c + c^T c - 1$$

$$\mathcal{L}(\theta,\lambda) = n\theta^T\theta - 2\theta\sum_{k=1}^{n} x_k + \sum \lambda x x^T - \lambda(\theta^T\theta - 2\theta^T c + c^T c - 1)$$

$$\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta} = 2n\theta - 2n\bar{x} - \lambda(2\theta - 2c) = 0$$

$$\Leftrightarrow 2n\theta - \lambda 2\theta = 2n\bar{x} - 2\lambda c$$

$$\Leftrightarrow \theta(n - \lambda) = n\bar{x} - \lambda c \longrightarrow \theta^* = \frac{n\bar{x} - \lambda c}{n - \lambda}$$

Vorzeichenfehler ...

$$\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \lambda} = \theta^T\theta - 2\theta^T c + c^T c - 1 = (\theta - c)^T(\theta - c) = \left(\frac{n\bar{x} - \lambda c}{n - \lambda} - c\right)^T \left(\frac{n\bar{x} - \lambda c}{n - \lambda} - c\right)$$

$$\Leftrightarrow \left(\frac{n\bar{x} - nc}{n - \lambda}\right)^T \left(\frac{n\bar{x} - nc}{n - \lambda}\right) = \frac{(n\bar{x} - nc)^T(n\bar{x} - nc)}{(n - \lambda)^2} = 1$$

$$\Leftrightarrow n^2 \underbrace{(\bar{x} - c)^T(\bar{x} - c)}_{\|\bar{x} - c\|^2} = (n - \lambda)^2 \longrightarrow \lambda = n \mp \sqrt{n^2\|\bar{x} - c\|^2}$$

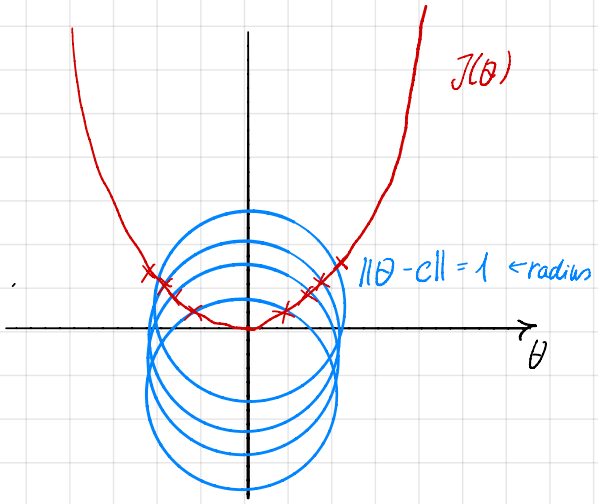
$\lambda_1 = n - n\|\bar{x} - c\|$
$\lambda_2 = n + n\|\bar{x} - c\|$

Which one minimize?

$$\theta_1^* = \frac{\lambda\bar{x} - \lambda(1 - \|\bar{x} - c\|)c}{\lambda - \lambda(1 - \|\bar{x} - c\|)} = \frac{\bar{x} - c + c\|\bar{x} - c\|}{-\|\bar{x} - c\|} = \frac{\bar{x} - c}{\|\bar{x} - c\|} - c \leftarrow \text{minimize}$$

$$\theta_2^* = \frac{\bar{x} - (1 + \|\bar{x} - c\|)c}{\|\bar{x} - c\|} = \frac{\bar{x} - c}{\|\bar{x} - c\|} + c$$

Geometrical interpretation ?




## Exercise 2 Bounds on Eigenvalues

We consider a dataset $x_1, \dots, x_n$, $x_n \in \mathbb{R}^d$. The empirical mean $m$, and the scatter matrix $S$ are given by

$$m = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{and} \quad S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^T$$

Let $\lambda_1$ be the largest eigenvalue of the matrix $S$. The eigenvalue $\lambda_1$ qualifies the amount of variation in the data on the first principal component. Because computation of the full scatter matrix and respective eigenvalues can be slow, it can be useful to relate them to the diagonal elements of the scatter matrix $\{S_{ii}\}$ that can be computed in linear time.

a.) Show that $\underbrace{\sum_{i=1}^{d} S_{ii}}_{\text{trace}}$ is an upper bound to the eigenvalue $\lambda_1$

We know that a trace of a matrix equals the sum of its eigenvalues

$$tr(S) = \sum_{i=1}^{d} S_{ii} = \sum_{i=1}^{d} \lambda_i = \lambda_1 + \sum_{i=2} \lambda_i$$

$$\rightarrow \lambda_1 = \sum_{i=1}^{d} S_{ii} - \sum_{i=2} \lambda_i$$

→ We want to proof that $\sum_{i=1}^{d} S_{ii}$ is an upper bound

$$\sum_{i=1}^{d} S_{ii} - \sum_{i=2} \lambda_i \leq \sum_{i=1}^{d} S_{ii} \quad \rightarrow \quad -\sum_{i=2} \lambda_i \leq 1$$

↳ if we proof that all eigenvalues $\geq 0$ (positive) than it's true
→ Scatter matrix $S$ has to be positive semi-definite → that means
① symmetric matrix
② all eigenvalues $\geq 0$
→ $S$ is positive semi-definit if $z^T S z \geq 0$, for every non-zero column vector $z$

$$z^T S z = \sum_{k=1}^{n} z^T (x_k - m)(x_k - m)^T z = \sum_{k=1}^{n} z^T (x_k - m) \cdot z^T (x_k - m)$$

$$= \sum_{k=1}^{n} z^T z \, (x_k - m)^T (x_k - m) = \sum_{k=1}^{n} \|z\|^2 \|x_k - m\|^2 \overset{!}{\geq} 0 \quad \text{True ☺}$$

↳ $S$ is positive semi-definite $\Rightarrow \sum_{i=2} \lambda_i \geq 0$

$\Rightarrow \sum_{i=1}^{d} S_{ii}$ is an upper bound for $\lambda_1$

b.) State the condition on the data for which the upper bound is tight

The upper bound is tight, if the variance within the data set can be represented along one dimension.

That means that the variance of the dataset is represented by one single Eigenvalue with its corresponding Eigenvector

as equation $\sum_{i=1}^{d} S_{ii} = \sum_{i=1} \lambda_i = \lambda_1 + \underbrace{\sum_{i=2} \lambda_i}_{=0} \rightarrow \sum_{i=1}^{d} S_{ii} = \lambda_1$

↳ matrix $S$ is of rank 1, which means that all features are linearly dependent

c.) Show that $\max_{i=1}^{d} S_{ii}$ is a <u>lower bound</u> to the eigenvalue $\lambda_1$

Let $\max_{i=1}^{d} S_{ii} = S_{jj} \rightarrow S_{jj} \leq \lambda_1$ ? for any vector $v$, and $\|v\| = 1$

$v^T S_{jj} v \leq \lambda_1$

$a = v^T W$

$v^T W \Lambda W^T v \leq \lambda_1$ ) $S = W \Lambda W^T$ where $W = [\underbrace{w_1, \dots, w_d}_{\text{corresponding eigenvectors}}]$ with $W^T W = I$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$

$\Lambda = \sum_{i=1}^{d} \lambda_i$

$a^T \Lambda a = \lambda_1 a_1^2 + \dots + \lambda_d a_d^2 \leq \lambda_1 (a_1^2 + \dots + a_d^2) = \lambda_1 \|a\|^2$

$= \lambda_1 a^T a = \lambda_1 v^T \underbrace{W W^T}_{=1} v = \lambda_1 \underbrace{v^T v}_{=1} = \lambda_1 = \lambda_{max}$

Therefore if we let $a = w_i$ (eigenvector) $\rightarrow S_{jj} = w_i^T S_{jj} w_i \leq \lambda_1$

???

for any $i = 1, \dots, d$, which implies $\max_{i=1}^{d} S \leq \lambda_1$

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

The bound is tight if the maximum variance in the data is aligned along one particular dimension.

## Exercise 3 Iterative PCA

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

$w \leftarrow \frac{Sw}{\|Sw\|}$ until some convergence criterion is met.

a.) Show that application of the power iteration method is equivalent to defining the unconstrained objective → Ableitung

$$J(w) = \|Sw\| - \frac{1}{2} w^T S w$$

and performing the gradient ascent $v \leftarrow v + \gamma \frac{\partial J}{\partial v}$, where $v = S^{0.5} w$ is a reparameterization of $v$, for some learning rate $\gamma$. We assume the the matrix $S$ is invertible.

Solution: $v = S^{\frac{1}{2}} w \longrightarrow w = \frac{v}{S^{\frac{1}{2}}}$ We want $\frac{\partial J(w)}{\partial v}$

$\frac{\partial}{\partial x} \|ax\|_2 = \frac{(ax)' \cdot ax}{\|ax\|_2}$ $= a$

$$\frac{\partial J(w)}{\partial v} = \frac{\partial J(w)}{\partial w} \cdot \frac{\partial w}{\partial v} = \left(\frac{SSw}{\|Sw\|} - Sw\right) \cdot \frac{1}{S^{\frac{1}{2}}}$$

$$= \frac{Sw}{\|Sw\|} - S^{\frac{1}{2}} w \qquad w = \frac{v}{S^{\frac{1}{2}}}$$

$$= \frac{S^{\frac{1}{2}} v}{\|S^{\frac{1}{2}} v\|} - v \longrightarrow v \leftarrow v + \gamma \frac{S^{\frac{1}{2}} v}{\|S^{\frac{1}{2}} v\|} - v$$

$$\text{if } \gamma = 1 \longrightarrow v \leftarrow \frac{S^{\frac{1}{2}} v}{\|S^{\frac{1}{2}} v\|}$$

b.) Show that a necessary condition for $w$ to maximize the objective $J(w)$ is to be a unit vector $\|w\| = 1$

Maximize $\frac{\partial J(w)}{\partial w} = 0$ $\frac{SSw}{\|Sw\|} - Sw = 0$ $\quad | \cdot S^{-1}$

$$w = \frac{Sw}{\|Sw\|} \longrightarrow \|w\| = \left\| \frac{Sw}{\|Sw\|} \right\| = \frac{\|Sw\|}{\|Sw\|} = 1 \quad \smile$$