Exercises for the course
Machine Learning 1
Winter semester 2024/25

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 4

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.

(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.

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.

(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.

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.

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

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

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

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

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

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.

Exercise 5: Programming (30 P)

Download the programming files on ISIS and follow the instructions.

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.

a)

$$\mathcal{L}(\theta, \lambda) = J(\theta) + \lambda(\theta^T b)$$

$$= \sum_{k=1}^{N} \|\theta - x_k\|^2 + \lambda(\theta^T b)$$

$$= \sum_{k=1}^{N} \left[\theta^2 - 2\theta x_k + x_k^2\right] + \lambda \cdot \theta^T \cdot b$$

$$= N\theta^2 - 2 \cdot \sum_{k=1}^{N} \theta x_k + \sum_{k=1}^{N} x_k^2 + \lambda \theta^T \cdot b$$

$$\nabla_\theta \mathcal{L} = 2N\theta - 2\sum_{k=1}^{N} x_k + \lambda b \overset{!}{=} \vec{0} \quad \Leftrightarrow \quad \theta = \frac{2\sum_{k=1}^{N} x_k - \lambda b}{2N}$$

$$= m - \frac{\lambda b}{2N}$$

$$\theta^T b = 0$$

$$\Leftrightarrow \left(m - \frac{\lambda b}{2N}\right)^T b = 0$$

$$\Leftrightarrow \left(m^T - \frac{\lambda}{2N} b^T\right) b = 0$$

$$\Leftrightarrow \lambda\, b^T \cdot b = 2N m^T \cdot b$$

$$\Leftrightarrow \lambda = \frac{2N \cdot m^T \cdot b}{b^T \cdot b}$$

$$\Rightarrow \theta^* = m - \frac{\lambda b}{2N} = m - \frac{m^T \cdot b \cdot b}{b^T \cdot b} = m - \frac{m^T \cdot b \cdot b}{\|b\|^2} = m - (b^T m) b$$

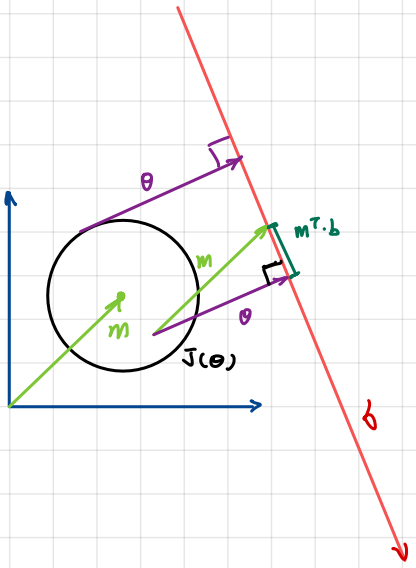
$$\frac{b^T\theta}{0} = b^T m - \frac{\lambda \|b\|}{2N} \quad \Rightarrow \lambda = \frac{2N b^T m}{\|b\|} = 2N b^T m$$

$$\operatorname{argmin}_\theta \sum_{k=1}^{N} \|\theta - x_k\|^2 = \operatorname{argmin}_\theta \sum_k \|\theta\|^2 - 2\theta^T x_k$$

$$= \operatorname{argmin}_\theta \sum_k \|\theta\|^2 - 2\theta^T m$$

$$= \operatorname{argmin}_\theta \sum_k \|\theta - m\|^2$$

$$= \operatorname{argmin}_\theta \tfrac{1}{2}\|\theta - m\|^2$$



θ is orthogonal to b by $\theta^T b = 0$.

Geometrically the optimal θ is obtained by substracting the projection of m onto the direction of b

The function $\theta \mapsto \frac{1}{2}\|\theta - m\|^2$ is strictly convex

↳ Found solution is global minimum.

-2

(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.

$$\mathcal{L}(\theta, \lambda) = J(\theta) + \lambda(\|\theta - c\|^2 - 1)$$

$$= \sum_{k=1}^{N} \|\theta - x_k\|^2 + \lambda(\|\theta - c\|^2 - 1)$$

$$= \sum_{k=1}^{N} (\theta - x_k)^T(\theta - x_k) + \lambda(\theta - c)^T(\theta - c) - \lambda$$

$$\nabla_\theta \mathcal{L} = \sum_{k=1}^{N} 2(\theta - x_k) + 2\lambda(\theta - c) = \vec{0}$$

$$\Leftrightarrow \quad 2\theta \cdot N - \sum_{k=1}^{N} 2x_k + 2\lambda\theta - 2\lambda c = \vec{0}$$

$$\Leftrightarrow \quad (2N + 2\lambda)\theta = \sum_{k=1}^{N} 2x_k + 2\lambda c$$

$$\Leftrightarrow \quad \theta = \frac{1}{2N + 2\lambda}\left(\sum_{k=1}^{N} 2x_k + 2\lambda c\right)$$

$$= \frac{1}{N+\lambda} \cdot (N \cdot m + \lambda c)$$

$$\|\theta - c\|^2 = 1 \quad \Leftrightarrow \quad (\theta - c)^T(\theta - c) = \frac{1}{(N+\lambda)^2}(N \cdot m + \lambda c - cN - c\lambda)^T(N \cdot m - cN) = 1$$

$$\Leftrightarrow \quad (Nm - cN)^T(Nm - cN) = (N+\lambda)^2$$

$$\Leftrightarrow \quad N^2(m - c)^T(m - c) = (N+\lambda)^2$$

$$\Leftrightarrow \quad (N+\lambda)^2 = N^2\|m - c\|^2$$

$$\Leftrightarrow \quad \lambda = -N \pm \sqrt{N^2\|m - c\|^2}$$

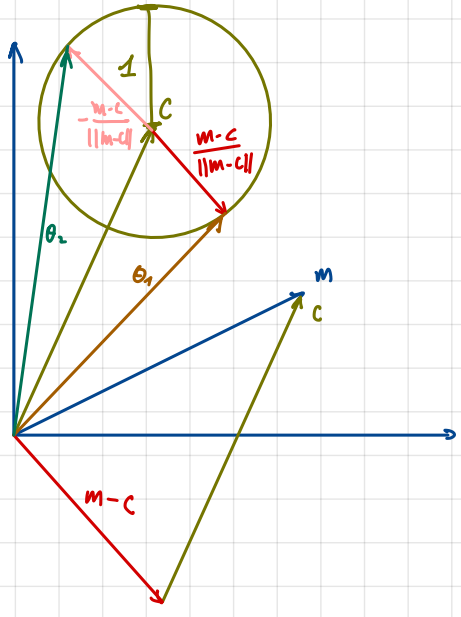
$$= -N \pm N \cdot \|m - c\|$$

$$\theta^* = \frac{1}{N+\lambda} \cdot (N \cdot m + \lambda c) = \frac{1}{N - N \pm N \cdot \|m - c\|} \cdot (m \cdot N - cN \pm cN \cdot \|m - c\|$$

$$= \frac{N}{\pm N \cdot \|m - c\|} \cdot (m - c \pm c \cdot \|m - c\|)$$

$$= \frac{1}{\pm \|m - c\|}(m - c \pm c\|m - c\|)$$

$$= \frac{m - c}{\pm\|m - c\|} + c$$



θ must lie on a circle of radius 1 centered at c by $\|\theta - c\|^2 - 1$.

The optimal θ can be obtained by moving 1 unit from c along the unit vector $m - c$ or the opposite direction $-(m - c)$.

To find the optimal solution, we observe that:

$$J\left(c + \frac{m-c}{\|m-c\|}\right) \leq J\left(c - \frac{m-c}{\|m-c\|}\right)$$

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.

$$\begin{aligned}
&\arg\max_u \tfrac{1}{N}\textstyle\sum_{k=1}^{N}\big[u^T x_k - \tfrac{1}{N}\big(\sum_{l=1}^{N} u^T x_l\big)\big]^2 \quad \text{with } \|u\|^2 = 1 \\
=\ &\arg\max_u \tfrac{1}{N}\textstyle\sum_{k=1}^{N}\big[u^T x_k - u^T \cdot m\big]^2 \\
=\ &\arg\max_u \tfrac{1}{N}\textstyle\sum_{k=1}^{N}\big[u^T (x_k - m)\big]^2 \\
=\ &\arg\max_u \tfrac{1}{N}\textstyle\sum_{k=1}^{N}\big[u^T (x_k - m)(x_k - m)^T u\big] \\
=\ &\arg\max_u \tfrac{1}{N}\big[u^T \textstyle\sum_{k=1}^{N}(x_k - m)(x_k - m)^T u\big] \\
=\ &\arg\max_u \tfrac{1}{N}\big[u^T S u\big] \\
\Leftrightarrow\ &\arg\max_u u^T S u \quad \text{with } \|u\|^2 = 1
\end{aligned}$$

(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.

$$\begin{aligned}
&\mathcal{L}(u, \lambda) = u^T S u - \lambda(\|u\|^2 - 1) \\
&\quad = u^T S u - \lambda u^T u + \lambda \\
&\nabla_u \mathcal{L} = 2Su - 2\lambda u = 0 \quad \Leftrightarrow Su = \lambda u \qquad (S \text{ is symmetric}) \\
&\arg\max_u \quad u^T S u = \lambda \cdot u^T \cdot u = \lambda \|u\| = \lambda
\end{aligned}$$

$\Rightarrow$ To find the maximum of $u^T S u$ we should choose the highest eigenvalue λ

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.

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

So $\sum_{i=1}^{d} S_{ii} \geq \lambda_1$ holds if $\sum_{i=2}^{d} \lambda_i \geq 0$

To prove $\sum_{i=2}^{d} \lambda_i \geq 0$, we could show the matrix S is semi-positive definite.

But first, we must show S is symmetric:

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

$$S^T = \sum_{k=1}^{N} \left[(x_k - m)(x_k - m)^T \right]^T$$

$$= \sum_{k=1}^{N} \left[(x_k - m)^{TT} (x_k - m)^T \right]$$

$$= \sum_{k=1}^{N} \left[(x_k - m)(x_k - m)^T \right] = S$$

$\Rightarrow S$ is symmetric

And S is semi-positive definite if $\vec{x}^T S \vec{x} \geq 0$ for all $\vec{x} \in \mathbb{R}^d$

$$\Rightarrow \sum_{k=1}^{N} \vec{x}^T (x_k - m)(x_k - m)^T \vec{x} = \sum_{k=1}^{N} x^T (x_k - m) \vec{x}^T (x_k - m)$$

$$= \sum_{k=1}^{N} x^T \vec{x} (x_k - m)^T (x_k - m)$$

$$= \sum_{k=1}^{N} \|\vec{x}\|^2 \cdot \|x_k - m\|^2 \geq 0$$

$\Rightarrow S$ is semi-positive definite

$$\Rightarrow \sum_{i=2}^{d} \lambda_i \geq 0$$

$$\Rightarrow \sum_{i=1}^{d} S_{ii} \geq \lambda_1$$

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

the upper bound is tight if $\sum_{i=1}^{d} S_{ii} = \lambda_1$

But $\sum_{i=1}^{d} S_{ii} = \sum_{i=1}^{d} \lambda_i = \lambda_1 + \sum_{i=2}^{d} \lambda_i \geq \lambda_1$

The equal sign holds only if $\sum_{i=2}^{d} \lambda_i = 0$

So the condition is $\sum_{i=2}^{d} \lambda_i = 0$

$\lambda_2 = \cdots = \lambda_d = 0$

→ data lies on a one-dimensional subspace

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

λ_1 is the largest eigenvalue of S

So $\lambda \vec{x} = S\vec{x} \Rightarrow \lambda_1 \vec{x}^T \vec{x} = \vec{x}^T S \vec{x}$

$\lambda_1 = \max_{u:\|u\|=1} u^T S u \geq \max_{u \in \{e_1, \dots e_d\}} u^T S u$

$= \max_{i=1}^{d} e_i^T S e_i$

$= \max_{i=1}^{d} S_{ii}$

Now we assume $\|\vec{x}\| = \vec{x}^T \cdot \vec{x} = 1$

$\Rightarrow \lambda_1 = \max \vec{x}^T S \vec{x}$

We choose a unit vector $\vec{e_i} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$

$\Rightarrow \lambda_1 = \max \vec{x}^T S \vec{x} \geq \max \vec{e_i}^T S \vec{e_i} = \max S_{ii}$

$\Rightarrow \lambda_1 \geq \max S_{ii}$

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

-5

the lower bound is tight if $\lambda_1 = \max S_{ii}$

$\sum_{i=1}^{d} S_{ii} = \sum_{i=1}^{d} \lambda_i = \lambda_1 + \sum_{i=2}^{d} \lambda_i = S_{11} + \sum_{i=2}^{d} S_{ii}$ (assume $\max S_{ii} = S_{11}$)

if $\lambda_1 = \max S_{ii} = S_{11} \Rightarrow \sum_{i=2}^{d} \lambda_i = \sum_{i=2}^{d} S_{ii}$

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

where $e_1, \dots, e_d$ are the canonical unit vector

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

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

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.

from lecture 2:

$$S w^{(t)} = S\Big(\sum_{i=1}^{d} \langle w^{(t)}, u_i \rangle u_i\Big) = \sum_{i=1}^{d} \langle w^{(t)}, u_i \rangle S u_i$$

$$= \sum_{i=1}^{d} \langle w^{(t)}, u_i \rangle \lambda_i u_i$$

$$w^{(t+1)} = \frac{S w^{(t)}}{\| S w^{(t)} \|} = \frac{\sum_{i=1}^{d} \langle w^{(t)}, u_i \rangle \lambda_i u_i}{\left\| \sum_{i=1}^{d} \langle w^{(t)}, u_i \rangle \lambda_i u_i \right\|}$$

$$w^{(t+1)\top} u_1 = \frac{\langle w^{(t)}, u_1 \rangle \lambda_1 u_1^\top}{\| S w^{(t)} \|} u_1 = \frac{\langle w^{(t)}, u_1 \rangle}{\| S w^{(t)} \|} \lambda_1$$

$$w^{(t+1)\top} u_k = \frac{\langle w^{(t)}, u_k \rangle \lambda_k}{\| S w^{(t)} \|} u_k^\top \cdot u_k = \frac{\langle w^{(t)}, u_k \rangle}{\| S w^{(t)} \|} \lambda_k$$

$$\mathcal{E}_k(w^{(t+1)}) = \left| \frac{w^{(t+1)\top} u_k}{w^{(t+1)\top} u_1} \right| = \frac{\langle w^{(t)}, u_k \rangle \lambda_k}{\langle w^{(t)}, u_1 \rangle \lambda_1}$$

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

$$= \frac{\lambda_k}{\lambda_1} \cdot \mathcal{E}_k(w^{(t)})$$

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

$$= \left(\frac{\lambda_k}{\lambda_1}\right)^{1+t} \mathcal{E}_k(w^{(0)})$$

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