sheet11

January 27, 2023

1 ML-Physics, Sheet 11

Leonard Benkendorff, Dorothea Schwärzel, Sebastian Preuß

[1]:
```
%matplotlib inline
import torch
import torch.nn as nn
from torch import Tensor
from tqdm.auto import tqdm
import numpy as np
import matplotlib.pyplot as plt
```

1.1 1 Positional Encoding

1.1.1 a)

Ex 1

Addition: $\text{score} = K^T Q = K^T K = \underbrace{(X+E)^T}_{n \times p} \underbrace{(X+E)}_{p \times n}$

where $E = \begin{pmatrix} e_{11} & \cdots & e_{1n} \\ \vdots & & \vdots \\ e_{p1} & \cdots & e_{pn} \end{pmatrix} \begin{array}{l} \leftarrow \cos \\ \leftarrow \sin \\ \leftarrow \cos \\ \leftarrow \sin \end{array} \Big\downarrow k$, $\xrightarrow{i}$

$$\rightsquigarrow \text{score}_{ij} = \sum_{k=1}^{p} (X+E)^T_{ik} (X+E)_{kj}$$

$$= \sum_{k=1}^{p} (x_{ki} + e_{ki})(x_{kj} + e_{kj})$$

$$= \sum_{k=1}^{p} x_{ki} x_{kj} + e_{ki} x_{kj} + x_{ki} e_{kj} + e_{ki} e_{kj}$$

Concatenation: $\text{score} = K^T Q = K^T K = \underbrace{\text{cat}(X,E)^T}_{n \times 2p} \underbrace{\text{cat}(X,E)}_{2p \times n}$

$$\text{cat}(X,E) = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{p1} & \cdots & x_{pn} \\ e_{11} & \cdots & e_{1n} \\ \vdots & & \vdots \\ e_{p1} & \cdots & e_{pn} \end{pmatrix} \in \mathbb{R}^{2p \times n}$$

$$\Rightarrow \text{score}_{ij} = \sum_{k=1}^{2p} (\text{cat}(X,E)^T)_{ik} \, \text{cat}(X,E)_{kj}$$

$$= \sum_{k=1}^{2p} \text{cat}(X,E)_{ki} \, \text{cat}(X,E)_{kj}$$

$$= \sum_{k=1}^{2p} \begin{cases} x_{ki} x_{kj} & , k \le p \\ e_{(k-p)i} e_{(k-p)j} & , k > p \end{cases} = \sum_{k=1}^{p} x_{ki} x_{kj} + e_{ki} e_{kj}$$

In the case of summation there are mixed terms in the entries of E and X in the entries of the score, for concatenation matrix entries of X and E do not mix.

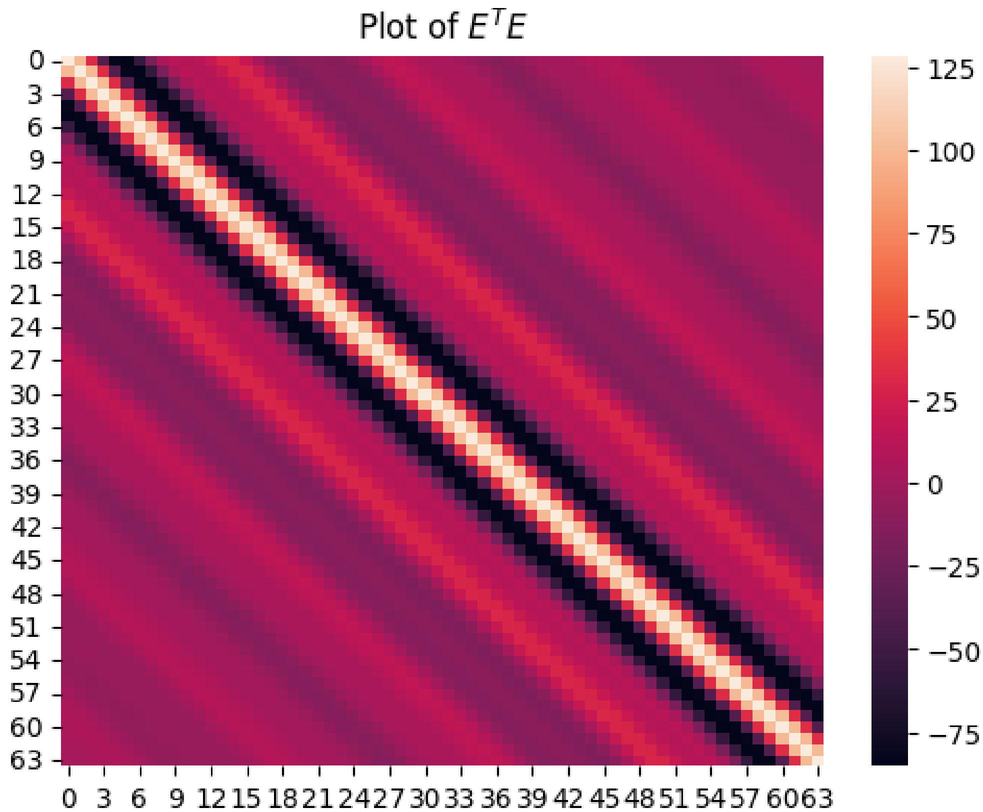
1.1.2 b)

$E \in \mathbb{R}^{p \times n}$

$E_{(2k),i} = \sin\left(i \cdot \exp\left(-\frac{2k \cdot \log(10000)}{5000}\right)\right)$

$E_{(2k+1),i} = \cos\left(i \cdot \exp\left(-\frac{2k \cdot \log(10000)}{5000}\right)\right)$

[4]:
```
import seaborn as sns

def get_E(n, p):
    i = np.arange(n)+1
    k = np.arange(p)+1

    E = np.zeros((p,n))
    ii, kk = np.meshgrid(i,k)

    E1 = np.sin(ii*np.exp(-2*kk*np.log(10000)/5000))
    E2 = np.cos(ii*np.exp(-2*kk*np.log(10000)/5000))
    E[k % 2 == 0,:] = E1[k % 2 == 0,:]
    E[k % 2 == 1,:] = E2[k % 2 == 1,:]
    return E

def get_score(E, X, mode='cat'):
    if mode=='add':
        Q = E + X
        return Q.T @ Q
    elif mode=='cat':
        Q = np.append(E, X, axis=0)
        print(Q.shape)
        return Q.T @ Q
    else:
        print('You did something wrong.')
        return

E = get_E(64,256)
#print(E.shape)
X = np.random.normal(0.0, np.std(E), size=(256,64))
plt.title('Plot of $E^TE$')
sns.heatmap(E.T @ E)
plt.show()
```



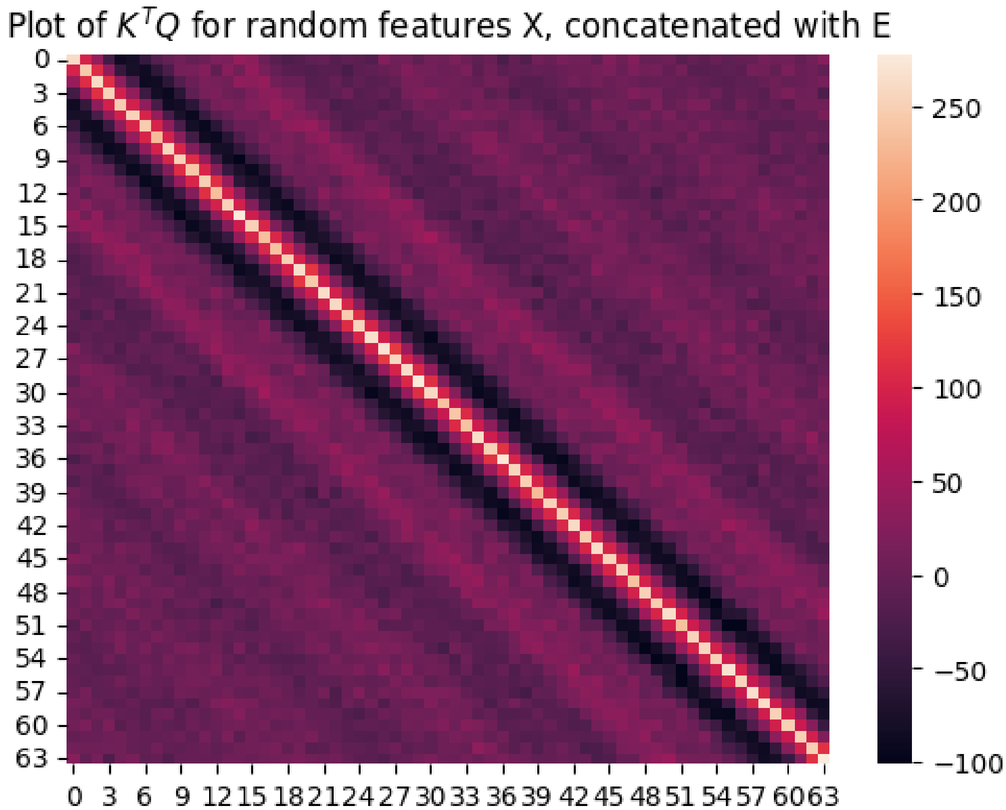
Our observation: We observe a strong score along the diagonal and decreasing, oscillating strenghts when moving away from the main diagonal.

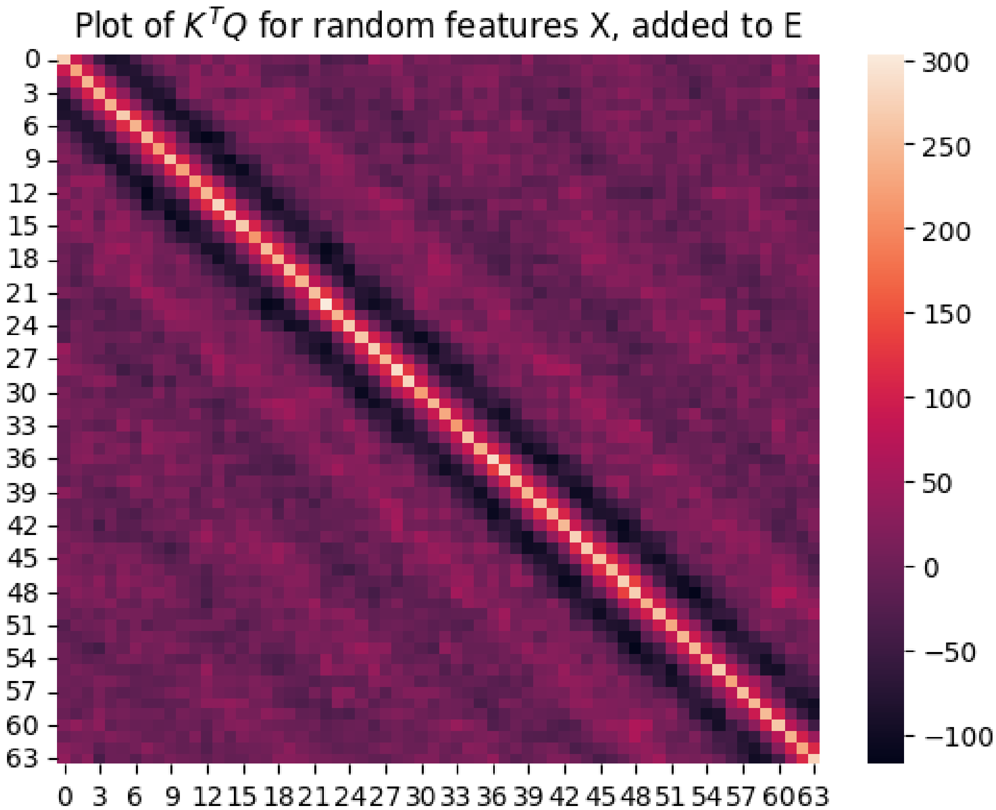
1.1.3 c)

We plot K^TQ for random features X for concatenation and addition with E.

[5]:
```
plt.title('Plot of $K^TQ$ for random features X, concatenated with E')
sns.heatmap(get_score(E,X, mode='cat'))
plt.show()
plt.title('Plot of $K^TQ$ for random features X, added to E')
sns.heatmap(get_score(E,X, mode='add'))
plt.show()
```

```
(512, 64)
```


Plot of K^TQ for random features X, concatenated with E
250
200
150
100
50
0
−50
−100



1.1.4 d)

d) When observing the sum calculated in (a), we see that the scores between sum and aggregation only differ in the „cross terms" $e_{x_i} x_{x_j} + x_{x_i} e_{x_j}$ which mix the matrix entries of X and E. If X and E reside in different or minimally overlapping subspaces, these cross terms are small or cancel each other

1.2 2 Interpolating between Aggregation functions

1.2.1 a)

ML - Physics, sheet 11

Ex 2

a) $\text{soft(arg)max\_agg}(v, m, \lambda, \eta) = |\mathcal{N}(v)|^\eta \sum_{u \in \mathcal{N}(v)} \text{soft(arg)max}(m; \lambda)_u \cdot m_u$

In sheet 7, we proved $\lim_{\lambda \to \infty} \text{lse}(\sigma; \lambda) = \max(\sigma)$ and

$\text{soft(arg)max}(\sigma; \lambda) = \frac{\partial}{\partial \sigma} \text{lse}(\sigma; \lambda)$

Here we will use that the convergence of $\lim_{\lambda \to \infty} \text{lse}(\sigma; \lambda)$ is uniform (bc. log, sum, exp, derivation converge uniformly in λ).

For uniformly convergent functions f, it holds that

$$\lim_{\lambda \to \infty} \frac{\partial f(x; \lambda)}{\partial x} = \frac{\partial}{\partial x} \lim_{\lambda \to \infty} f(x; \lambda).$$

Therefore:

i) $$\lim_{\lambda \to \infty} \text{soft(arg)max\_agg}(v, m, \lambda, \eta) = |\mathcal{N}(v)|^\eta \sum_{u \in \mathcal{N}(v)} \lim_{\lambda \to \infty} \text{soft(arg)max}(m; \lambda)_u \cdot m_u$$

$$= |\mathcal{N}(v)|^\eta \sum_u \frac{\partial}{\partial m} \left[\lim_{\lambda \to \infty} \text{lse}(m; \lambda) \right] \cdot m_u$$

$$= |\mathcal{N}(v)|^\eta \sum_u \underbrace{\frac{\partial}{\partial m} \left[\max(m) \right]}_{= (*)} \cdot m_u = |\mathcal{N}(v)|^\eta \max(m)$$

$(*)$: $\frac{\partial}{\partial m} \max(m) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow$ index of max m $\Rightarrow \frac{\partial}{\partial m} (\max(m)) \cdot m = \max m$

$\Rightarrow$ one obtains the mean max aggregation.

For $\lim_{\lambda \to 0}$ soft(arg)max_agg, the same holds, hence we need to evaluate $\lim_{\lambda \to 0} \text{lse}(\sigma; \lambda)$

$$\lim_{\lambda \to 0} \text{lse}(\sigma; \lambda) = \lim_{\lambda \to 0} \frac{\overset{\to 0}{\log \sum_{i=0}^{K} \exp(\lambda \sigma_i)}}{\underset{\to 0}{\lambda}} \Rightarrow \text{L'Hospitals rule}$$

$$= \lim_{\lambda \to 0} \frac{\frac{\partial}{\partial \lambda} \log \sum_i \exp(\lambda \sigma_i)}{\frac{\partial}{\partial \lambda} \lambda} = \lim_{\lambda \to 0} \frac{1}{\underbrace{\sum_i \exp(\lambda \sigma_i)}_{= 1}} \sum_i \overbrace{\exp(\lambda \sigma_i)}^{= 1} \sigma_i$$

$$= \frac{\sum \sigma_i}{K}$$

$$\Rightarrow \lim_{\lambda \to 0} \text{soft(arg)max\_agg}(v, m, \lambda, \eta) = |\mathcal{N}(v)|^\eta \sum_{u \in \mathcal{N}(v)} \frac{\partial}{\partial m} \lim_{\lambda \to 0} \left[\text{lse}(m; \lambda) \right] m_u$$

$= |\mathcal{N}(v)|^\eta \sum \frac{\partial}{\partial} \frac{\sum m_u}{\ldots} m_u = |\mathcal{N}(v)|^\eta \sum \frac{|\mathcal{N}(v)|}{\ldots} m_u = |\mathcal{N}(v)|^\eta \cdot \sum m_u$ [illegible]

1.2.2 b)

Ex 26

power_mean_agg$(v, \vec{m}, q, \eta) =: \mathrm{pma}(v, \vec{m}, q, \eta)$

$$= |\mathcal{N}(v)|^{\eta} \left(\frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} m_u^{q} \right)^{\frac{1}{q}}$$

We have $\mathrm{pma}(v, m, 1, \eta) = |\mathcal{N}(v)|^{\eta - 1} \sum_{u \in \mathcal{N}(v)} m_u$ i.e.

we retrieve the sum.

$$\lim_{q \to \infty} \mathrm{pma}(v, m, q, \eta) = \underbrace{|\mathcal{N}(v)|}_{=: \mathcal{N}}^{\eta} \left(\frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} m_u^{q} \right)^{\frac{1}{q}}$$

$$= \lim_{q \to \infty} \mathcal{N}^{\eta} \underbrace{\mathcal{N}^{-\frac{1}{q}}}_{\to 1} \left(\sum_u m_u^{q} \right)^{\frac{1}{q}}$$

$$= \mathcal{N}^{\eta} \lim_{q \to \infty} \left(\sum_u m_u^{q} \right)^{\frac{1}{q}} \overset{(*)}{=} \mathcal{N}^{\eta} \max_u m_u ,$$

i.e. we retrieve the max.

(*): To show $\lim_{q \to \infty} \left(\sum_u m_u^{q} \right)^{\frac{1}{q}} = \overbrace{\max m_u}^{=: m}$, we prove

$$\lim_{q \to \infty} \frac{\left(\sum_u m_u^{q} \right)^{\frac{1}{q}}}{m} = 1$$

$$\lim_{q \to \infty} \frac{\left(\sum_u m_u^{q} \right)^{\frac{1}{q}}}{m} = \lim_{q \to \infty} \Bigg(\underbrace{\sum_u \left(\frac{m_u}{m} \right)^{q}}_{=: S} \Bigg)^{\frac{1}{q}} = \lim_{q \to \infty} S^{\frac{1}{q}}$$

$= 1$

$1 \leq S \leq p$

1.2.3 c)

c) A GCN has layers of the type

$$h_u = \phi\Big(\bigoplus_{v \in \mathcal{N}(u)} \psi(x_u, X_v; \theta_t)\Big)$$

We can define $\overline{\psi} = \bigoplus_{v \in \mathcal{N}(u)} \text{agg}(\lambda, \eta)\, \psi(x_u, x_v, \theta_t)$

and learn the hyper-parameters of the generalised aggregation as well in a "new" layer $h_u = \phi(\overline{\psi}(x_u, v \in \mathcal{N}(u), \lambda, \eta, \theta_t))$.

Our observation: We observe that both score matrices look essentially the same. The scaling is a bit different, but the distribution is the same: A very strong score on the main diagonal and then a decaying score when moving away from this diagonal, which oscillates.

1.3 3 Observing Oversmoothing

We did not attempt to do exercise 3a,b