Philipp Lintl
12152498
lintl.philipp@gmail.com

Homework Assignment 4
Machine Learning 1, 18/19

2018-12-06

# 1 Mixture of Experts

This task explores MoEs, a model that softly partitions the input space and learns a supervised model for each area.
Consider that you have K experts available in order to model a specific dataset of N datapoints $\{\mathbf{x}_1, \mathbf{y}_1), \ldots, \mathbf{x}_N, \mathbf{y}_N)\}$, where $\mathbf{x}_n$ corresponds to a vector input in $\mathbb{R}^D$ and $\mathbf{y}_n$ corresponds to the particular label available for $\mathbf{x}_n$. Let $z_n$ correspond to a categorial random variable for datapoint n that denotes which of the K experts is active.
Further notation:

- $\mathbf{\Theta} \in \mathbb{R}^{D \times K}$ containing D-dimensional column vector of parameters for each expert.
- we assume each $y_i$ is continous random variable at $[0, \infty) \sim Exp(\lambda), \lambda > 0$

Given these assumptions, each expert $k \in K$ has the folowing linear predictive model:

$$\begin{aligned} p(y_n|\mathbf{x}_n, z_n, \mathbf{\Theta}) &= p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \\ &= \text{Exponential}(y_n|\lambda = exp(\theta_k^T \mathbf{x}_n)), \end{aligned}$$

where $\mathbf{z}_n$ corresponds to a 1-of-K vector representation of the categorial variable $z_n$ and

$$\text{Exponential}(y|\lambda) = \lambda exp(-\lambda y)\,, \text{for} y \leq 0.$$

A simple linear routing mechanism is the following:

$$p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) = \pi_{nk} = \frac{exp(\phi_k^T \mathbf{x}_n)}{\sum_j exp(\phi_j^T \mathbf{x}_n)},$$

where $\mathbf{\Phi}$ is a matrix in $\mathbb{R}^{D \times K}$ that contains all of the parameters of the routing function, i.e. $\mathbf{\Phi} = [\phi_1, \ldots, \phi_K]$.
As a priori we have no infromation about which of the experts is responsible for generating a particular prediction we have to marginalize over all possible experts in order to compute the likelihood of an observed point.

## 1.1

Write down the likelihood of the entire dataset, $p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi})$, and take its log under the i.i.d assumption.

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &\overset{iid}{=} \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \mathbf{\Theta}, \mathbf{\Phi}) \\ \text{(joint distribution)} \quad &= \prod_{n=1}^{N} \sum_{k=1}^{K} p(y_n, z_n = k|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n, \mathbf{\Phi}) \\ \text{(Bayes rule)} \quad &= \prod_{n=1}^{N} \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}), \end{aligned}$$

with $\pi_{nk} = p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})$ this yields:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) = \prod_{n=1}^{N} \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot \pi_{nk}$$

Taking the log over this yields the log-likelihood as follows:

$$\begin{aligned} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \ln \prod_{n=1}^{N} \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot \pi_{nk} \\ &= \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot \pi_{nk} \right) \\ \text{(filling in the exponential distribution)} \quad &= \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} exp(\theta_k^T \mathbf{x}_n) \cdot exp(-exp(\theta_k^T \mathbf{x}_n) \cdot y_n) \cdot \pi_{nk} \right) \end{aligned}$$

## 1.2

Write down the posterior probability $r_{ni}$ of expert i producing the label y for the datapoint n. We will also refer to this as the resposibility of expert i for datapoint n.

$$
\begin{aligned}
r_{ni} = p(z_n = i|y_n) &= \frac{p(z_n = i) \cdot p(y_n|z_n = i)}{p(y_n)} = \frac{p(z_n = i) \cdot p(y_n|z_n = i)}{\sum_{j=1}^{K} p(z_n = j) \cdot p(y_n|z_n = j)} \\
&= \frac{\pi_{ni} \cdot p(y_n|z_n = i)}{\sum_{j=1}^{K} \pi_{nj} \cdot p(y_n|z_n = j)},
\end{aligned}
$$

again filling in the exponential distribution yields:

$$
r_{ni} = \frac{\pi_{ni} \cdot exp(\theta_i^T \mathbf{x}_n) \cdot exp(-exp(\theta_i^T \mathbf{x}_n))}{\sum_{j=1}^{K} \pi_{nj} \cdot exp(\theta_j^T \mathbf{x}_n) \cdot exp(-exp(\theta_j^T \mathbf{x}_n))}
$$

## 1.3

Take the derivative of the log-likelihood w.r.t. the parameters of each expert $\theta_i$ and the parameters ot the routing mechanism for each expert $\phi_i$. Do not substitute expressions for the probabilities but rather provide your answer in terms of $p(y_n|\mathbf{x}_n, z_n, \theta_i), p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})$. Make sure to express the derivatives in terms of the responsabilities of each expert $r_{ni}$. Hint: $\frac{\partial f(x)}{\partial x} = f(x)\frac{\partial \log f(x)}{\partial x}$, as that term will be present in the derivatives for both $\theta_i, \phi_i$.

i) w.r.t. $\theta_i$:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \frac{\partial}{\partial \theta_i} \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot \pi_{nk} \right) \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \theta_i} \ln \left( \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) \right) \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})} \\
&\quad \cdot \frac{\partial}{\partial \theta_i} p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = i|\mathbf{x}_n, \mathbf{\Phi})
\end{aligned}
$$

Using the hint given, this can be rewritten in a way, the responsibility of expert i can be plugged in:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \sum_{n=1}^{N} \underbrace{\frac{p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = i|\mathbf{x}_n, \mathbf{\Phi})}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})}}_{= r_{ni}} \\
&\quad \cdot \frac{\partial}{\partial \theta_i} \ln p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = i|\mathbf{x}_n, \mathbf{\Phi}) \\
&= \sum_{n=1}^{N} r_{ni} \cdot \frac{\partial}{\partial \theta_i} \ln p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = i|\mathbf{x}_n, \mathbf{\Phi})
\end{aligned}
$$

ii) w.r.t. $\phi_i$:

$$\begin{aligned}
\frac{\partial}{\partial \phi_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \frac{\partial}{\partial \phi_i} \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) \right) \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \phi_i} \ln \left( \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) \right) \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})} \\
&\quad \cdot \frac{\partial}{\partial \phi_i} \sum_{l=1}^{K} p(y_n|\mathbf{x}_n, \theta_l = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = l|\mathbf{x}_n, \mathbf{\Phi}) \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})} \\
&\quad \cdot \sum_{l=1}^{K} p(y_n|\mathbf{x}_n, \theta_l = \mathbf{\Theta}\mathbf{z}_n) \cdot \frac{\partial}{\partial \phi_i} p(z_n = l|\mathbf{x}_n, \mathbf{\Phi})
\end{aligned}$$

The latter derivative is looked at more closely:

$$\frac{\partial}{\partial \phi_i} p(z_n = l|\mathbf{x}_n, \mathbf{\Phi}) = \frac{\partial}{\partial \phi_i} \frac{exp(\phi_l^T \mathbf{x}_n)}{\sum_j exp(\phi_j^T \mathbf{x}_n)}$$

For $l = i$ using quotient rule:

$$\frac{\partial}{\partial \phi_i} \frac{exp(\phi_l^T \mathbf{x}_n)}{\sum_j exp(\phi_j^T \mathbf{x}_n)} = \frac{exp(\phi_i^T \mathbf{x}_n) \cdot \mathbf{x}_n^T}{\sum_j exp(\phi_k^T \mathbf{x}_n)} - \frac{\left(exp(\phi_i^T \mathbf{x}_n)\right)^2 \cdot \mathbf{x}_n^T}{\left(\sum_j exp(\phi_j^T \mathbf{x}_n)\right)^2} = (\pi_{ni} - \pi_{ni}^2) \cdot \mathbf{x}_n^T$$

The trick in the last brace if for later so the sums match up. For $l \neq i$ using quotient rule:

$$\frac{\partial}{\partial \phi_i} \frac{exp(\phi_l^T \mathbf{x}_n)}{\sum_j exp(\phi_l^T \mathbf{x}_n)} = -\frac{exp(\phi_l^T \mathbf{x}_n) \cdot exp(\phi_i^T \mathbf{x}_n) \cdot \mathbf{x}_n^T}{\left(\sum_j exp(\phi_l^T \mathbf{x}_n)\right)^2} = -\pi_{nl} \cdot \pi_{ni} \cdot \mathbf{x}_n^T$$

Using these two properties for the entire sum over k yields:

$$\begin{aligned}
\frac{\partial}{\partial \phi_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \sum_{n=1}^{N} \frac{\sum_{l=1}^{K} p(z_n = l|\mathbf{x}_n, \mathbf{\Phi}) \cdot -\pi_{nl} \pi_{ni} \mathbf{x}_n^T}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot \pi_{nk}} \\
&\quad + \frac{p(z_n = i|\mathbf{x}_n, \mathbf{\Phi}) \cdot \pi_{ni} \mathbf{x}_n^T}{\sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})} \\
&= \sum_{n=1}^{N} (-\pi_{ni} + r_{ni}) \cdot \mathbf{x}_n^T
\end{aligned}$$

## 1.4

Replace the expressions for each of the respective probability distributions and compute the final derivations for $\theta_i, \phi_i$.

i) for $\theta_i$:
Using the definitions of $r_{ni}$ from before and that the product in the derivative on the right can be spliited due to the log and then deriving for $\theta_i$ of term $\ln p(z_n = i|\mathbf{x}_n, \mathbf{\Phi})$ will be 0, only leaving the other term:

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \sum_{n=1}^{N} r_{ni} \cdot \frac{\partial}{\partial \theta_i} \ln p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) \\
&= \sum_{n=1}^{N} r_{ni} \cdot \frac{1}{p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n)} \cdot \frac{\partial}{\partial \theta_i} p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n),
\end{aligned}$$

At this step, for the sake of oversight, the parts are looked at separately and then combined for the final version. So,

$$\frac{\partial}{\partial \theta_i} p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) = \frac{\partial}{\partial \theta_i} \cdot exp(\theta_i^T \mathbf{x}_n) \cdot exp(-exp(\theta_i^T \mathbf{x}_n) y_n)$$

$$\text{(product rule)} \quad = \left(\frac{\partial}{\partial \theta_i} exp(\theta_i^T \mathbf{x}_n)\right) \cdot exp(-exp(\theta_i^T \mathbf{x}_n) y_n) + exp(\theta_i^T \mathbf{x}_n) \cdot \left(\frac{\partial}{\partial \theta_i} exp(-exp(\theta_i^T \mathbf{x}_n) y_n)\right)$$

Even more in detail:

$$\frac{\partial}{\partial \theta_i} exp(\theta_i^T \mathbf{x}_n) = exp(\theta_i^T \mathbf{x}_n) \cdot \mathbf{x}_n^T$$

And the other part:

$$\frac{\partial}{\partial \theta_i} exp(-exp(\theta_i^T \mathbf{x}_n) y_n) = exp(-exp(\theta_i^T \mathbf{x}_n) \cdot y_n) \cdot -y_n \cdot exp(\theta_i^T \mathbf{x}_n) \cdot \mathbf{x_n}^T$$

This can now be plugged into the product rule:

$$\begin{aligned}\frac{\partial}{\partial \theta_i} p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) = &\left(exp(\theta_i^T \mathbf{x}_n) \cdot \mathbf{x}_n^T\right) \cdot exp(-exp(\theta_i^T \mathbf{x}_n) \cdot y_n) + \\ &exp(\theta_i^T \mathbf{x}_n) \cdot \left(exp(-exp(\theta_i^T \mathbf{x}_n) y_n) \cdot -y_n \cdot exp(\theta_i^T \mathbf{x}_n) \cdot \mathbf{x_n}^T\right) \\ &= \left[1 - y_n \cdot exp(\theta_i^T \mathbf{x}_n\right] \cdot exp(\theta_i^T \mathbf{x}_n) \cdot exp(-exp(\theta_i^T \mathbf{x}_n) \cdot y_n) \cdot \mathbf{x}_n^T\end{aligned}$$

Now going back to the first form of the entire derivative and plugging the intermediate result in results in:

$$\begin{aligned}\frac{\partial}{\partial \theta_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \sum_{n=1}^{N} r_{ni} \cdot \frac{1}{p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n)} \cdot \frac{\partial}{\partial \theta_i} p(y_n|\mathbf{x}_n, \theta_i = \mathbf{\Theta}\mathbf{z}_n) \\ &= \sum_{n=1}^{N} r_{ni} \cdot \frac{1}{exp(\theta_i^T \mathbf{x}_n) \cdot exp(-exp(\theta_i^T \mathbf{x}_n) y_n)} \cdot \left[1 - y_n \cdot exp(\theta_i^T \mathbf{x}_n)\right] \\ &\cdot exp(\theta_i^T \mathbf{x}_n) \cdot exp(-exp(\theta_i^T \mathbf{x}_n) \cdot y_n) \cdot \mathbf{x}_n^T \\ &= \sum_{n=1}^{N} r_{ni} \cdot \left[1 - y_n \cdot exp(\theta_i^T \mathbf{x}_n)\right] \cdot \mathbf{x}_n^T\end{aligned}$$

ii) Now the same happens to the derivative wrt to $\phi$:

$$\begin{aligned}\frac{\partial}{\partial \phi_i} \ln p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi}) &= \sum_{n=1}^{N} (-\pi_{ni} + r_{ni}) \cdot \mathbf{x}_n^T \\ &= \sum_{n=1}^{N} \left(-\frac{exp(\phi_i^T \mathbf{x}_n)}{\sum_j exp(\phi_j^T \mathbf{x}_n)} + r_{ni}\right) \cdot \mathbf{x}_n^T\end{aligned}$$

## 1.5

Write down an iterative algorithm that maximizes the log-probability of the data by jointly optimizing the $\mathbf{\Theta}$ and the $\mathbf{\Phi}$ parameters. Make use of an approriate convergence criterion.
Similar to the practical, this happens in the form of a variation of EM. As we did not solve the derivations for ML estimates, I suppose approximating the optimum by SGA (as we take positive log-likelihood) seems reasonable, particularly like this:

- Randomly initialize $\mathbf{\Theta}, \mathbf{\Phi}$
- Initialize learning rate $\eta$ for gradient ascent
- repeat until convergence (after each run compute $L = \ln p(y_n|\mathbf{x}_n, \mathbf{\Theta}, \mathbf{\Phi})$, when $\Delta L < \epsilon_1$, stop the algorithm)

– E-Step: calculate $r_{nk} \forall n, k$ :

$$r_{nk} = \begin{cases} 1 & \text{if argmax}_k \; p(z_n = k | y_n, \mathbf{x}_n, \mathbf{\Phi}, \mathbf{\Theta}) \\ 0 & \text{otherwise} \end{cases}$$

– M-Step: run until other convergence criterion (difference in $\ln p(y_n|\mathbf{x}_n, \mathbf{\Theta}, \mathbf{\Phi})$ from $\tau \quad to \quad \tau + 1 < \epsilon_2$) is met

$$\begin{aligned} gradient_{theta}^{(\tau)} &= \frac{\partial}{\partial \mathbf{\Theta}} \ln p(y_n|\mathbf{x}_n, \mathbf{\Theta}, \mathbf{\Phi}) \\ gradient_{phi}^{(\tau)} &= \frac{\partial}{\partial \mathbf{\Phi}} \ln p(y_n|\mathbf{x}_n, \mathbf{\Theta}, \mathbf{\Phi}) \\ \mathbf{\Theta}^{(\tau+1)} &= \mathbf{\Theta}^{(\tau)} + \eta \cdot \; gradient_{theta}^{(\tau)} \\ \mathbf{\Phi}^{(\tau+1)} &= \mathbf{\Phi}^{(\tau)} + \eta \cdot \; gradient_{phi}^{(\tau)} \end{aligned}$$

## 1.6

Assume that an oracle is available that provides you an extra set of M points that also has the information about which expert t should be employed.
This means that additional to the N datapoints $\{\mathbf{x}_1, \mathbf{y}_1), \ldots, \mathbf{x}_N, \mathbf{y}_N)\}$, we have new observations $\mathbf{X}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_M\}$ and new targets $\mathbf{Y}' = \{y'_1, \ldots, y'_M\}$. Special to these datappints are the expert labels $\mathbf{T} = \{t_1, \ldots, t_M\}$, whereas $t_i = k$ indicates that expert k should be employed for datapoint i.
1.) Write down the likelihood of this extended dataset and take its log.
The likelihood of this new dataset (only M datapoints) looks as follows: Beforehand,

$$p(z_m = k|\mathbf{x}_m, \mathbf{\Phi}) = \begin{cases} 1 & \text{if } t_m = k \\ 0 & else \end{cases}$$

$$\begin{aligned} p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) &\overset{iid}{=} \prod_{m=1}^{M} p(y'_m|x'_m, \Phi, \Theta) = \prod_{m=1}^{M} \prod_{k=1}^{K} p(y'_m, z_m = k|x'_m, \Phi, \Theta) \\ &= \prod_{m=1}^{M} p(y'_m, z_m = t_m|x'_m, \Phi, \Theta) \\ &= \prod_{m=1}^{M} p(y'_m|x'_m, \theta_{t_m} = \mathbf{\Theta}\mathbf{z}_m) \cdot \underbrace{p(z_m = t_m|\mathbf{x}_m, \mathbf{\Phi})}_{\pi_{mt_m}}, \end{aligned}$$

having in mind that we already now the expert, $\pi_{mt_m} = 1$, leaving a likelihood of:

$$= \prod_{m=1}^{M} p(y'_m|x'_m, \theta_{t_m} = \mathbf{\Theta}\mathbf{z}_m) \cdot 1$$

Taking the log of this yields:

$$\ln p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) = \sum_{m=1}^{M} \ln p(y'_m|x'_m, \theta_{t_m} = \mathbf{\Theta}\mathbf{z}_m)$$

If we combine this with the original dataset of N datapoints, we simply multiply this term for the likelihood and add it for the loglikelihood, as we assume independence of the $(\mathbf{X}', Y')$ and $(\mathbf{X}, Y)$:

$$\begin{aligned} p(\mathbf{Y}', \mathbf{Y}|\mathbf{X}', \mathbf{X}, \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) &= p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) \cdot p(\mathbf{Y}|\mathbf{X}, \mathbf{\Phi}, \mathbf{\Theta}) \\ \Rightarrow \ln p(\mathbf{Y}', \mathbf{Y}|\mathbf{X}', \mathbf{X}, \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) &= \ln p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) + \ln p(\mathbf{Y}|\mathbf{X}, \mathbf{\Phi}, \mathbf{\Theta}) \\ &= \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} p(y_n|\mathbf{x}_n, \theta_k = \mathbf{\Theta}\mathbf{z}_n) \cdot \pi_{nk} \right) + \sum_{m=1}^{M} \ln p(y'_m|x'_m, \theta_{t_m} = \mathbf{\Theta}\mathbf{z}_m) \end{aligned}$$

2.) Take the derivatives of this new log-likelihood wrt the parameters $\phi_i, \theta_i$.

1)wrt to $\theta_i$:

First only the new term of the M datapoints is looked at:

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \ln p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) &= \frac{\partial}{\partial \theta_i} \sum_{m=1}^{M} \ln p(y'_m|x'_m, \theta_{t_m} = \mathbf{\Theta}\mathbf{z}_m) \\
&= \sum_{m=1}^{M} \frac{\partial}{\partial \theta_i} \ln p(y'_m|x'_m, \theta_{t_m} = \mathbf{\Theta}\mathbf{z}_m) \\
&= \sum_{m=1}^{M} \frac{\partial}{\partial \theta_i} \ln exp(\theta_{t_m}^T \mathbf{x}_n) \cdot exp(-exp(\theta_{t_m}^T \mathbf{x}_n) \cdot y_n) \\
&= \sum_{m=1}^{M} \frac{\partial}{\partial \theta_i} \left[\theta_i^T \mathbf{x}'_m - exp(\theta_i^T \mathbf{x}'_m) \cdot y'_m\right] \\
&= \sum_{m=1}^{M} \mathbb{I}(t_m = i) \left[\mathbf{x}'^T_m - exp(\theta_i^T \mathbf{x}'_m) \cdot y_m \cdot \mathbf{x}'^T_m\right]
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \frac{\partial}{\partial \theta_i} \ln p(\mathbf{Y}', \mathbf{Y}|\mathbf{X}', \mathbf{X}, \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) &= \frac{\partial}{\partial \theta_i} \ln p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta}) + \frac{\partial}{\partial \theta_i} \ln p(\mathbf{Y}|\mathbf{X}, \mathbf{\Phi}, \mathbf{\Theta}) \\
&= \sum_{n=1}^{N} r_{ni} \cdot \left[1 - y_n \cdot exp(\theta_i^T \mathbf{x}_n)\right] \cdot \mathbf{x}_n^T \\
&+ \sum_{m=1}^{M} \mathbb{I}(t_m = i) \left[\mathbf{x}'^T_m - exp(\theta_i^T \mathbf{x}'_m) \cdot y_m \cdot \mathbf{x}'^T_m\right]
\end{aligned}$$

2) wrt to $\phi_i$:

In my opinion, the derivative of the new datapoints according to $\mathbf{\Phi}$ is zero, as the log-likelihood $\ln p(\mathbf{Y}'|\mathbf{X}', \mathbf{T}, \mathbf{\Phi}, \mathbf{\Theta})$ is independent of $\Phi$. That way, it is the same as for the N datapoints derived before.

3.) What is the difference between the derivatives computed here and the derivatives computed previously? Is the overall model a linear or a non-linear one?

The difference is that the derivative according to $\mathbf{\Theta}$ for the M datapoints does not depend on the responsibiity of the expert, meaning, that they are the same except $r_{ni}$ does not appear anymore. Regarding the derivative wrt $\mathbf{\Phi}$, it stays the same as for the N datapoints we had before.
Regarding linearity in parameters, it can be stated that both individually are linear in $\mathbf{\Theta}$ and the first one also in $\mathbf{\Phi}$, as we predict the parameter of the exponential distribution with a linear model. Combining the two however leads to a non linear case, as we multiply the parameters $\mathbf{Theta}$, leading to a non linear case.

# 2 PCA

Suppose we have a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ iof D-dimensional vectirs, which have a zero mean for each domension. Assume, we perform a complete eigenvalue decomposition of the empirical covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.

## 2.1

Initially, you are interested in only a single projection of your data such that the variance of this projection is maximized. Let $\mathbf{u}_i$ be the direction vector of a particular projection. Assume that $\mathbf{u}_i^T \mathbf{u}_i = 1$.

(a) What is the projection $z_{ni}$ of a given point $\mathbf{x}_n$ under the particular vector $\mathbf{u}_i$?

The projection of a given point $\mathbf{x}_n$ on any latent dimension i is given by the product of the principal component vector and the data vector itself:

$$z_{ni} = \mathbf{u}_i^T \cdot \mathbf{x}_n$$

(b) What is the empirical mean of the projection $z_i$ across all points $\mathbf{x}_n$?

For the empirical mean of the projection $z_i$ across all data points, we first need the empirical mean vector of all data points:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n,$$

which is a vector with the means of all dimensions in it. So the mean of the projection is:

$$\begin{aligned} \mu_{z_i} &= \mathbf{u}_i^T \cdot \bar{\mathbf{x}} \\ &= \frac{1}{N} \mathbf{u}_i^T \cdot \mathbf{0} \\ &= 0, \end{aligned}$$

as the data are normalized meaning that each mean is 0. So the mean of the projection in latent dimension i across all data points is 0.

(c) What is the empirical variance of the projection $z_i$? Provide your answer in terms of the empirical covariance matrix **S**.

The empirical variance of the projection $z_i$ is given by:

$$\begin{aligned} \sigma_{z_i}^2 &= \frac{1}{N} \sum_{n=1}^{N} (\mathbf{u}_i^T \mathbf{x}_n - \mathbf{u}_i^T \bar{\mathbf{x}})^2 \\ &= \frac{1}{N} \sum_{n=1}^{N} (\mathbf{u}_i^T (\mathbf{x}_n - \bar{\mathbf{x}}))^2 \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}_i^T (\mathbf{x}_n - \bar{\mathbf{x}}) \cdot (\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{u}_i \\ &= \mathbf{u}_i^T \left( \underbrace{\frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \bar{\mathbf{x}}) \cdot (\mathbf{x}_n - \bar{\mathbf{x}})^T}_{\mathbf{S}} \right) \mathbf{u}_i \\ &= \mathbf{u}_i^T \mathbf{S} \mathbf{u}_i \end{aligned}$$

(d) Replace S with its eigenvalue decomposition and simplify the aforementioned expression. What is the variance now?

Considerung $\mathbf{U}$ is a matrix with the principal component vectors as columns and the properties (all pcs are othonormal, meaning length one and orthogonality among each other) of principal components:

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Using $\mathbf{S} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$, the term then is:

$$\begin{aligned} \sigma_{z_i}^2 &= \mathbf{u}_i^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \mathbf{u}_i \\ &= \big[0, \ldots, 0, 1, 0, \ldots, 0\big] \cdot \Lambda \cdot \big[0, \ldots, 0, 1, 0, \ldots, 0\big]^T \\ &= \lambda_i, \end{aligned}$$

whereas $\mathbf{u}_i^T \cdot \mathbf{U} = \big[\mathbf{u}_i^T \mathbf{u}_1, \ldots, \mathbf{u}_i^T \mathbf{u}_i, \ldots, \mathbf{u}_i^T \mathbf{u}_D\big] = \big[0, \ldots, 0, 1, 0, \ldots, 0\big]$

So for this case, the empirical variance of the projection $z_i$ is equal to the i-th eigenvalue of the covariance matrix.

(e) Suppose that you are interesting in reducing the dimensionality from D to K, such that 99% of the variance is maintained. How can you select an appropriate K?

As the i-th eigenvalue of the covariance matrix displays the amount of variance the i-th principal component is capturing of the total variance within the data, we choose so many principal components, until the sum of the eigenvalues reaches 0.99. So the proportion of variance explained by the K principal components is 0.99. It can be seen as a proportion, as the sum over all pc's accounts for all appearing variance within the data. In formulas, considering $\lambda_i = \Lambda_{ii}$:

$$K^\star = min \left\{ k | \sum_{j=1}^{K} \lambda_j \geq 0.99 \right\}$$

In other words, it is the first k, for which the sum of the eigenvalues is larger or equal 0.99.

## 2.2

Consider the projections of your data along the K principal components. Prove that the dimensions of the projections are de-correlated (Hint: check the value of the empirical covariance w.r.t. dimension i and j).

The dimensions of the projections are decorrelated, if the covariance matrix of the projections is diagonal, i.e. no covariances among the principal components. Theoretically this must hold, as the variance maximation does not only need to guarantee $\mathbf{u}_1^T \mathbf{u}_1 = 1$ but also for all pc's higher than 1, for each dimension one additional constraint must hold. Namely, $\mathbf{u}_i^T \mathbf{u}_j = 0 \; \forall j \neq i$. Regarding the covariance matrix of the projections it can be seen, that it is:

$$\begin{aligned} \frac{1}{N} \sum_{n=1}^{N} \mathbf{z}_n \mathbf{z}_n^T &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{U}_K^T (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{U}_K \\ &= \mathbf{U}_K^T \mathbf{S} \mathbf{U}_K = \mathbf{U}_K^T \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \mathbf{U}_K = \Lambda_K, \end{aligned}$$

whereas $\mathbf{U}_K$ denotes a matrix with the first K principal components in its columns. And $\Lambda_K$ is a **diagonal** matrix $\in \mathbb{R}^{K \times K}$ having all eigenvalues on the diagonal. So the dimensions of the projections of all data points are decorrelated. Also, $\mathbf{z}_n$

## 2.3

Imagine that you want to de-correlate all of the dimensions but still want to enforce a mean of m and variance of $\tau$ across the D dimensions. How can you post-process your projections such that they satisfy these properties? Show that these hold by computing the empirical mean and variance across the dataset.

From before, we know that the means of the projections are 0 and the variances are the appropriate eigenvalues $\lambda$ of the covariance matrix. New, postprocessed projections $z_{ni}^\star$ then can be formulated as such by adding a mean term $m_i$ and the inverse of $\lambda$: with the wished variance $\tau_i$:

$$z_{ni}^\star = m_i + z_{ni} \cdot \sqrt{\frac{\tau_i}{\lambda_i}}$$

Then the mean:

$$\begin{aligned} \mathbb{E}(z_{ni}^\star) &= \mathbb{E}(m_i) + \mathbb{E}(z_{ni} \cdot \sqrt{\frac{\tau_i}{\lambda_i}}) \\ &= m_i + 0 = m_i, \end{aligned}$$

which was the wished enforced mean.
For the variance:

$$\begin{aligned} \mathbb{V}(z_{ni}^\star) &= \mathbb{V}(m_i) + \mathbb{V}(z_{ni}) \cdot \frac{\tau_i}{\lambda_i} \\ &= \frac{\tau_i}{\lambda_i} \cdot \lambda_i \\ &= \tau_i \end{aligned}$$

Using the variance rules and the derivations from before we arrived where we wanted.