# Expectation maximization

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2017-2018

## Dataset

1. Observed dataset $\mathbf{X}$, including all observed elements $\mathbf{x}_1, \ldots, \mathbf{x}_n$
2. Complete dataset $(\mathbf{X}, \mathbf{Z})$, including the values of all random variables in the model (that is, also latent variables values)

Since $\mathbf{Z}$ is unknown, the knowledge about latent variables is only probabilistic: it is given by the distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\psi})$

## Dataset evidence

$$p(\mathbf{X}|\boldsymbol{\psi}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j q(\mathbf{x}_i|\theta_j)$$

$$\begin{aligned}
p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi}) &= \prod_{i=1}^{n} p(\mathbf{x}_i, z_i|\boldsymbol{\psi}) = \prod_{i=1}^{n} p(z_i|\boldsymbol{\psi}) p(\mathbf{x}_i|z_i, \boldsymbol{\psi}) = \\
&= \prod_{i=1}^{n} \prod_{j=1}^{K} \left(p(z_{ij}|\boldsymbol{\psi}) p(x_i|z_{ij}, \boldsymbol{\psi})\right)^{z_{ij}} = \prod_{i=1}^{n} \prod_{j=1}^{K} \pi_j^{z_{ij}} q(x_i|\theta^j)^{z_{ij}}
\end{aligned}$$

# Expectation Maximization

Dataset log-likelihood

$$l(\boldsymbol{\psi}|\mathbf{X}) = \log p(\mathbf{X}|\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \left( \sum_{j=1}^{K} \pi_j q(\mathbf{x}_i|\theta_j) \right)$$

$$l(\boldsymbol{\psi}|\mathbf{X},\mathbf{Z}) = \log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\psi}) = \log \prod_{i=1}^{n} \prod_{j=1}^{K} \pi_j^{z_{ij}} q(x_i|\theta^j)^{z_{ij}} =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} (\log \pi_j + \log q(x_i|\theta^j))$$

Maximization of the log-likelihood of $\mathbf{X}$

$$\operatorname*{argmax}_{\pi_i} l(\boldsymbol{\psi}|\mathbf{X}) \qquad \text{e} \qquad \operatorname*{argmax}_{\theta_i} l(\boldsymbol{\psi}|\mathbf{X})$$

usually hard to derive (solutions not closed-form)

Maximization of the log-likelihood of $(\mathbf{X}, \mathbf{Z})$

$$\underset{\pi_i}{\operatorname{argmax}}\, l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) \Rightarrow \begin{cases} 0 = \frac{\partial}{\partial \pi_i} \left( l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) + \lambda(1 - \sum_{j=1}^{K} \pi_j) \right) \\ 0 = \frac{\partial}{\partial \lambda} \left( l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) + \lambda(1 - \sum_{j=1}^{K} \pi_j) \right) \end{cases}$$

hence $\lambda = n$, $\pi_j = \frac{1}{n} \sum_{i=1}^{n} z_{ji}$

$$\underset{\theta_i}{\operatorname{argmax}}\, l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) \Rightarrow 0 = \frac{\partial}{\partial \theta_i} l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) = \sum_{j=1}^{n} z_{ji} \frac{1}{q(x_j|\theta_i)} \frac{\partial q(x_j|\theta_i)}{\partial \theta_i}$$

In many cases, closed form solutions.

**Hypothesis #1**

The maximization of the log-likelihood of the observed dataset is hard

$$l(\boldsymbol{\psi}|\mathbf{X}) = \log p(\mathbf{X}|\boldsymbol{\psi})$$

**Hypothesis #2**

The maximization of the log-likelihood of the complete dataset is easy

$$l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) = \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi})$$

**Hypothesis #3**

The posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\psi})$ is known

**Problem**

$$l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) = \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi})$$

cannot be computed, since $\mathbf{Z}$ is unknown

# Expectation Maximization: E-step

## Idea

Assume an estimate $\overline{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ is available: then, instead of $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi})$ we could consider its expected value wrt the distribution $p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\psi}})$ of $\mathbf{Z}$ conditioned on the observed data and on the estimate $\overline{\boldsymbol{\psi}}$

$$\mathcal{Q}(\boldsymbol{\psi}, \overline{\boldsymbol{\psi}}) = E_{p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\psi}})}[l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z})] = \sum_{\mathbf{Z}} l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z}) p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\psi}}) =$$
$$= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\psi}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi})$$

Where,

- $\overline{\boldsymbol{\psi}}$ and $\mathbf{X}$ are known
- $\mathcal{Q}(\boldsymbol{\psi}, \overline{\boldsymbol{\psi}})$ is a function of $\boldsymbol{\psi}$
- $\mathcal{Q}(\boldsymbol{\psi}, \overline{\boldsymbol{\psi}})$ is not dependent from $\mathbf{Z}$

## Note

- For each term, $p(\mathbf{Z}|\mathbf{X},\overline{\boldsymbol{\psi}})$ is known (hypothesis #3)
- For each $\mathbf{Z}$, $l(\boldsymbol{\psi}|\mathbf{X},\mathbf{Z}) = \log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\psi})$ can be easily maximized (hypothesis #2)

Hence,

$$\mathcal{Q}(\boldsymbol{\psi},\overline{\boldsymbol{\psi}}) = \sum_{\mathbf{Z}} c_{\mathbf{Z}} \log p(\mathbf{X},\mathbf{Z}|\boldsymbol{\psi})$$

is a linear combination of functions which can be easily maximized: as a consequence it can be easily maximized too

## Log-likelihood approximation

$\mathcal{Q}(\boldsymbol{\psi},\overline{\boldsymbol{\psi}})$ is an approximation of $l(\boldsymbol{\psi}|\mathbf{X},\mathbf{Z})$, which is simpler to maximize

### Idea

Since a function $\mathcal{Q}(\boldsymbol{\psi}, \overline{\boldsymbol{\psi}})$ approximating the log-likelihood $\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\psi}) = l(\boldsymbol{\psi}|\mathbf{X}, \mathbf{Z})$ of the complete dataset is available, and since it can be easily maximized, an estimate of the maximizing value $\boldsymbol{\psi}$ is derived

$$\hat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}}\ \mathcal{Q}(\boldsymbol{\psi}, \overline{\boldsymbol{\psi}})$$

### Iteration

The estimate $\hat{\boldsymbol{\psi}}$ is used for the next E-step.

The algorithm starts with an initial parameter estimate $\boldsymbol{\psi}_0$ and stops when some predefined convergence condition is verified (for example, $\overline{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\psi}}$ do not differ too much).

# Expectation Maximization: overall structure

## Structure

The algorithm works as follows

- Initialize $\boldsymbol{\psi}_{\text{old}} = \boldsymbol{\psi}_0$, with $\boldsymbol{\psi}_0$ an arbitrary estimate of $\boldsymbol{\psi}$
- while not ``stopping condition''
  - compute $\mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\psi}_{\text{old}})$ (E-step)
  - $\boldsymbol{\psi}_{\text{new}} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\psi}, \boldsymbol{\psi}_{\text{old}})$ (M-step)
  - $\boldsymbol{\psi}_{\text{old}} = \boldsymbol{\psi}_{\text{new}}$

## Property

It is possible to show that, at each step, the algorithm increases the log-likelihood of $\boldsymbol{\psi}$ on the observed dataset $\mathbf{X}$. It is a gradient-based algorithm, which converges toward a (local) maximum of the log-likelihood of wrt $\mathbf{X}$.

# Expectation Maximization: why does it work?

## General definition

Probabilistic model:

- observed variables $\mathbf{X}$
- latent variables $\mathbf{Z}$
- parameters $\boldsymbol{\theta}$

Joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ not observable.

Observable distribution $p(\mathbf{X}|\boldsymbol{\Theta})$.

Objective: maximizing the likelihood of the observable distribution

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

## Hypotheses

- maximizing $p(\mathbf{X}|\boldsymbol{\theta})$ is hard
- maximizing $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is much simpler

# Expectation Maximization: why does it work?

## Decomposition

Let $q(\mathbf{Z})$ be any probability distribution on $\mathbf{Z}$, then

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta}) + KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))$$

where

$$\mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}$$

$$KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}$$

$\mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta})$ and $KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))$ are functionals of $q(\mathbf{Z})$ and functions of $\boldsymbol{\theta}$

## Kullback-Leibler divergence

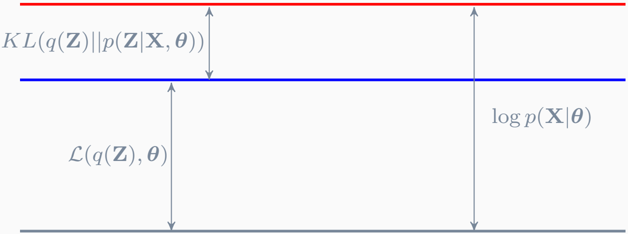
$KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))$ is the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.

By definition, $KL(q||p) \geq 0$, with $KL(q||p) = 0$ iff $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$

## Lower bound

$$KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})) \geq 0 \qquad \text{implies that} \qquad \mathcal{L}(q(\mathbf{Z}),\boldsymbol{\theta}) \leq \log p(\mathbf{X}|\boldsymbol{\theta})$$

Hence, $\mathcal{L}(q(\mathbf{Z}),\boldsymbol{\theta})$ is a lower bound of $\log p(\mathbf{X}|\boldsymbol{\theta})$



# Expectation Maximization: why does it work?

Let us find the probability distribution $q(\mathbf{Z})$ which results into the best (maximum) lower bound of $\log p(\mathbf{X}|\boldsymbol{\theta})$

## E-step

Let $\overline{\boldsymbol{\theta}}$ be the current estimate of $\boldsymbol{\theta}$. Then, as noticed above,

$$\log p(\mathbf{X}|\overline{\boldsymbol{\theta}}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\overline{\boldsymbol{\theta}})}{q(\mathbf{Z})} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})}{q(\mathbf{Z})}$$

In order to maximize the lower bound $\mathcal{L}(q(\mathbf{Z}), \overline{\boldsymbol{\theta}})$ wrt $q(\mathbf{Z})$, observe that, since $\log p(\mathbf{X}|\overline{\boldsymbol{\theta}})$ is independent from $\mathbf{Z}$, the maximum of $\mathcal{L}(q(\mathbf{Z}), \overline{\boldsymbol{\theta}})$ corresponds to the case $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$.

In such a case $KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})) = 0$ and

$$\begin{aligned}
\log p(\mathbf{X}|\overline{\boldsymbol{\theta}}) = \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \overline{\boldsymbol{\theta}}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\overline{\boldsymbol{\theta}})}{p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})} \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{X}, \mathbf{Z}|\overline{\boldsymbol{\theta}}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})
\end{aligned}$$

# Expectation Maximization: why does it work?

- The first term

$$\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{X}, \mathbf{Z}|\overline{\boldsymbol{\theta}})$$

  is the expected log-likelihood of $p(\mathbf{X}, \mathbf{Z}|\overline{\boldsymbol{\theta}})$ wrt $p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$, the posterior distribution deriving from the current estimation $\overline{\boldsymbol{\theta}}$ of parameters

- The second term

$$-\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$$

  is the entropy of such distribution $p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$

## M-step

Let $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$: then, if we consider the same decomposition of the log-likelihood $\log p(\mathbf{X}|\boldsymbol{\theta})$, with $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$, we have

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \boldsymbol{\theta}) + KL(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))$$

Let us consider the maximization wrt $\boldsymbol{\theta}$ of the lower bound

$$\begin{aligned}\mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})\end{aligned}$$

Since

$$\sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})$$

is independent from $\boldsymbol{\theta}$, this is equivalent to maximize

$$\mathcal{Q}(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

### M-step

Let

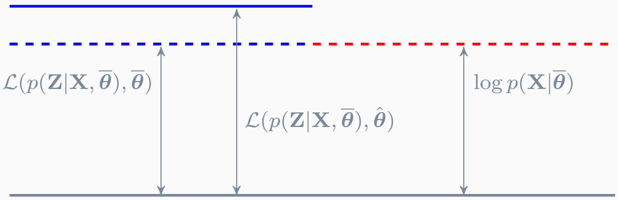
$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \boldsymbol{\theta})$$

then, by assumption, for any $\boldsymbol{\theta}$

$$\mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}) \geq \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \boldsymbol{\theta})$$

and, in particular,

$$\mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}) \geq \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \overline{\boldsymbol{\theta}}) = \log p(\mathbf{X}|\overline{\boldsymbol{\theta}})$$



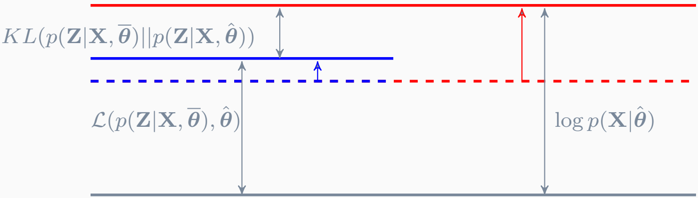
## M-step

By definition, we have

$$\log p(\mathbf{X}|\hat{\boldsymbol{\theta}}) = \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}) + KL(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})||p(\mathbf{Z}|\mathbf{X}, \hat{\boldsymbol{\theta}}))$$

Since in general $p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}) \neq p(\mathbf{Z}|\mathbf{X}, \hat{\boldsymbol{\theta}})$, we have $KL(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}})||p(\mathbf{Z}|\mathbf{X}, \hat{\boldsymbol{\theta}}) > 0$ and, as a consequence

$$\log p(\mathbf{X}|\hat{\boldsymbol{\theta}}) > \mathcal{L}(p(\mathbf{Z}|\mathbf{X}, \overline{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}})$$



**Conclusions**

After an E-step and an M-step, the estimated log-likelihood becomes larger.

In particular, it increases from

$$\log p(\mathbf{X}|\overline{\boldsymbol{\theta}}) = \mathcal{L}(p(\mathbf{Z}|\mathbf{X},\overline{\boldsymbol{\theta}}),\overline{\boldsymbol{\theta}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\overline{\boldsymbol{\theta}}) \log \frac{p(\mathbf{X},\mathbf{Z}|\overline{\boldsymbol{\theta}})}{p(\mathbf{Z}|\mathbf{X},\overline{\boldsymbol{\theta}})}$$

to

$$\begin{aligned}\log p(\mathbf{X}|\hat{\boldsymbol{\theta}}) &= \mathcal{L}(p(\mathbf{Z}|\mathbf{X},\hat{\boldsymbol{\theta}}),\hat{\boldsymbol{\theta}}) + KL(p(\mathbf{Z}|\mathbf{X},\overline{\boldsymbol{\theta}})||p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})) \\ &\geq \mathcal{L}(p(\mathbf{Z}|\mathbf{X},\hat{\boldsymbol{\theta}}),\hat{\boldsymbol{\theta}}) \\ &\geq \mathcal{L}(p(\mathbf{Z}|\mathbf{X},\hat{\boldsymbol{\theta}}),\overline{\boldsymbol{\theta}}) = \log p(\mathbf{X}|\overline{\boldsymbol{\theta}})\end{aligned}$$