Deep Learning

Homework 1 Part 1

Instructor: Mingsheng Long Jianyu Tang (2017012221)

Problem 1: Block 1

1-1

Solution
Since

$$y_i = \gamma \hat{x}_i + \beta$$

, we have

$$\frac{\partial y_i}{\partial \beta} = 1, \frac{\partial y_i}{\partial \gamma} = \hat{x}_i$$

. ■

1-2

Solution
Let $\boldsymbol{x}$ denote the input vector of the dropout layer, and let $\boldsymbol{y}$ denote the output vector. We have:

$$\boldsymbol{y}_i = \begin{cases} 0, & r_i < p \\ \boldsymbol{x}_i/(1-p), & r_i \geq p \end{cases}.$$

So the gradients are

$$\frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{x}_j} = 0, i \neq j$$
$$\frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{x}_i} = \begin{cases} 0, & r_i < p \\ 1/(1-p), & r_i \geq p \end{cases}.$$

Therefore, the gradients of the output of a dropout layer with respect to the input is

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = diag(\frac{\partial \boldsymbol{y}_1}{\partial \boldsymbol{x}_1}, ..., \frac{\partial \boldsymbol{y}_n}{\partial \boldsymbol{x}_n})$$

■

1-3

Solution
Let $\boldsymbol{x}$ denote the input vector of the softmax layer, and let $\boldsymbol{y}$ denote the output vector.

- When $i = j$,

$$\frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{x}_j} = \frac{\partial}{\partial \boldsymbol{x}_i} \frac{e^{\boldsymbol{x}_i}}{\sum_{k=1}^n e^{\boldsymbol{x}_k}} = e^{\boldsymbol{x}_i}(-\frac{e^{\boldsymbol{x}_i}}{(\sum_{k=1}^n e^{\boldsymbol{x}_k})^2}) + \frac{e^{\boldsymbol{x}_i}}{\sum_{k=1}^n e^{\boldsymbol{x}_k}} = \boldsymbol{y}_i - \boldsymbol{y}_i^2$$

- When $i \neq j$,

$$\frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{x}_j} = \frac{\partial}{\partial \boldsymbol{x}_j} \frac{e^{\boldsymbol{x}_i}}{\sum_{k=1}^n e^{\boldsymbol{x}_k}} = e^{\boldsymbol{x}_i}(-\frac{e^{\boldsymbol{x}_j}}{(\sum_{k=1}^n e^{\boldsymbol{x}_k})^2}) = -\frac{e^{\boldsymbol{x}_i} e^{\boldsymbol{x}_j}}{(\sum_{k=1}^n e^{\boldsymbol{x}_k})^2} = -\boldsymbol{y}_i \boldsymbol{y}_j$$

We have:

$$\frac{\partial \boldsymbol{y}_i}{\partial \boldsymbol{x}_j} = \begin{cases} \boldsymbol{y}_i - \boldsymbol{y}_i^2, & i = j \\ -\boldsymbol{y}_i \boldsymbol{y}_j, & i \neq j \end{cases}.$$

Therefore, the gradients of the output of a Softmax function with respect to the input of a Softmax function is

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{W}, where\ \boldsymbol{W}_{ij} = \begin{cases} \boldsymbol{y}_i - \boldsymbol{y}_i^2 & , i = j \\ -\boldsymbol{y}_i \boldsymbol{y}_j & , i \neq j \end{cases}$$

■

Problem 2: Block 2

2-1

Solution

Let's denote the output of a fully connected layer FC_i before activation as $\boldsymbol{z}_i$ and denote that after activation as $\boldsymbol{a}_i$.

After FC_{1A}:

$$\boldsymbol{z}_{1A} = \theta_{1A}\boldsymbol{x}$$
$$\boldsymbol{a}_{1A} = ReLU(\boldsymbol{z}_{1A})$$

After dropout layer:

$$\boldsymbol{d} = \boldsymbol{a}_{1A} \circ \boldsymbol{M}$$

, where $\circ$ is element-wise multiplication and $\boldsymbol{M}$ is the mask tensor.

After FC_{2A}:

$$\hat{y_a} = \boldsymbol{z}_{2A} = \theta_{2A}\boldsymbol{d}$$

After FC_{1B}:

$$\boldsymbol{z}_{1B} = \theta_{1B}\boldsymbol{x}$$
$$\boldsymbol{a}_{1B} = \mathrm{ReLU}(\boldsymbol{z}_{1B})$$

After batchnorm layer:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} a_{1Bi}$$
$$\sigma_B = \frac{1}{m}\sum_{i=1}^{m}(a_{1Bi} - \mu_B)^2$$

where a_{1Bi} the ith output in the batch. Then we get the output denoted as $\boldsymbol{c}$

$$\hat{x} = \frac{a_{1B} - \mu_B}{\sqrt{\mu_B^2 + \epsilon}}$$
$$\boldsymbol{c} = \gamma\hat{x} + \beta$$

Merge the output of two branches:

$$\boldsymbol{s} = \boldsymbol{c} + \boldsymbol{z}_{2A}$$

After FC_{2B}:

$$\boldsymbol{z}_{2B} = \theta_{2B}\boldsymbol{s}$$
$$\hat{y_b} = \mathrm{Softmax}(\boldsymbol{z}_{2B})$$

■

2-2

Solution
With the loss function $\boldsymbol{L}$ given, we first consider its gradients with respect to $\hat{y}_{bi}^{(j)}$:

$$\begin{aligned}\frac{\partial \boldsymbol{L}}{\partial \hat{y}_{bi}^{(j)}} &= \frac{\partial}{\partial \hat{y}_{bi}^{(j)}} \frac{1}{m} \sum_{i=1}^{m} [\frac{1}{2} ||(\hat{y}_{ai} - y_{ai})||_2^2 - \sum_{k=1}^{n_{y_b}} y_{bi}^{(k)} log(\hat{y}_{bi}^{(k)})] \\ &= -\frac{1}{m} \frac{\partial}{\partial \hat{y}_{bi}^{(j)}} \sum_{k=1}^{n_{y_b}} y_{bi}^{(k)} log(\hat{y}_{bi}^{(k)}) \\ &= -\frac{1}{m} \frac{\partial}{\partial \hat{y}_{bi}^{(j)}} y_{bi}^{(j)} log(\hat{y}_{bi}^{(j)}) \\ &= -\frac{1}{m} \frac{y_{bi}^{(j)}}{\hat{y}_{bi}^{(j)}}.\end{aligned}$$

According to the result about the gradients of Softmax in Block 1, we have the gradients of $\hat{y}_{bi}^{(j)}$ with respect to $\boldsymbol{z}_{2Bi}^{(k)}$:

$$\frac{\partial \hat{y}_{bi}^{(j)}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} = \begin{cases} \hat{y}_{bi}^{(j)}(1 - \hat{y}_{bi}^{(j)}), & k = j \\ -\hat{y}_{bi}^{(j)} \hat{y}_{bi}^{(k)}, & k \neq j \end{cases}.$$

Then apply the chain rule and get:

$$\begin{aligned}\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} &= \sum_{j=1}^{n_{y_b}} \frac{\partial \boldsymbol{L}}{\partial \hat{y}_{bi}^{(j)}} \frac{\partial \hat{y}_{bi}^{(j)}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} \\ &= -\frac{1}{m} \sum_{j=1}^{n_{y_b}} \frac{y_{bi}^{(j)}}{\hat{y}_{bi}^{(j)}} \frac{\partial \hat{y}_{bi}^{(j)}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} \\ &= \frac{1}{m} \sum_{j=1}^{n_{y_b}} y_{bi}^{(j)} \hat{y}_{bi}^{(k)} - \frac{1}{m} y_{bi}^{(k)} \\ &= \frac{1}{m} (\hat{y}_{bi}^{(k)} - y_{bi}^{(k)}).\end{aligned}$$

Here we reach the gradients of the loss function with respect to the parameters in $\boldsymbol{FC}_{2B}$, namely $\theta_{2b}^{(kj)}$:

$$\begin{aligned}\frac{\partial \boldsymbol{L}}{\partial \theta_{2b}^{(kj)}} &= \sum_{i=1}^{m} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} \frac{\partial \boldsymbol{z}_{2Bi}^{(k)}}{\partial \theta_{2b}^{(kj)}} \\ &= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi}^{(k)} - y_{bi}^{(k)}) \boldsymbol{s}^{(j)}.\end{aligned}$$

So we get an vectorial form of the gradients of the loss function with respect to θ_{2b}:

$$\frac{\partial \boldsymbol{L}}{\partial \theta_{2b}} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi}) \boldsymbol{s}^T.$$

Next we can solve the residual of the BN layer. Since

$$\boldsymbol{z}_{2B} = \theta_{2B} \boldsymbol{s} = \theta_{2B} \boldsymbol{c} + \theta_{2B} \boldsymbol{z}_{2A}$$

, we have:

$$\frac{\partial \boldsymbol{z}_{2B}}{\partial \boldsymbol{c}} = \theta_{2B}.$$

Thus the residual of the BN layer is:

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{c}_i^{(j)}} &= \sum_{k=1}^{n_{yb}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} \frac{\partial \boldsymbol{z}_{2Bi}^{(k)}}{\partial \boldsymbol{c}_i^{(j)}} \\
&= \frac{1}{m} \sum_{k=1}^{n_{yb}} (\hat{y}_{bi}^{(k)} - y_{bi}^{(k)}) \theta_{2b}^{(kj)} \\
&= \frac{1}{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,j)}.
\end{aligned}
$$

Using the result in Block 1 and we can get the gradients of the loss with respect to the BN layer's parameters γ and β:

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \beta} &= \sum_{i=1}^{m} \sum_{j=1}^{n_{yb}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{c}_i^{(j)}} \frac{\partial \boldsymbol{c}_i^{(j)}}{\partial \beta} \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_{yb}} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,j)} \\
&= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b} \boldsymbol{R}, \\
\frac{\partial \boldsymbol{L}}{\partial \gamma} &= \sum_{i=1}^{m} \sum_{j=1}^{n_{yb}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{c}_i^{(j)}} \frac{\partial \boldsymbol{c}_i^{(j)}}{\partial \gamma} \\
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_{yb}} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,j)} \hat{\boldsymbol{x}}_i^{(j)} \\
&= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b} \hat{\boldsymbol{x}}_i^T,
\end{aligned}
$$

where

$$
\boldsymbol{R} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n_{yb} \times 1}.
$$

Then we consider the residual of FC_{1B} layer:

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{1Bi}^{(j)}} &= \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{c}_i^{(j)}} \frac{\partial \boldsymbol{c}_i^{(j)}}{\partial \boldsymbol{z}_{1Bi}^{(j)}} \\
&= \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{c}_i^{(j)}} \frac{\partial \boldsymbol{c}_i^{(j)}}{\partial \boldsymbol{a}_{1Bi}^{(j)}} \frac{\partial \boldsymbol{a}_{1Bi}^{(j)}}{\partial \boldsymbol{z}_{1Bi}^{(j)}} \\
&= \frac{1}{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,j)} \frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \mathrm{sgn}(\boldsymbol{z}_{1Bi}^{(j)}).
\end{aligned}
$$

Therefore the gradients of the loss with respect to parameters in $\boldsymbol{FC}_{1B}$, namely $\theta_{1b}^{(kj)}$:

$$\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \theta_{1b}^{(kj)}} &= \sum_{i=1}^{m} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{1Bi}^{(k)}} \frac{\partial \boldsymbol{z}_{1Bi}^{(k)}}{\partial \theta_{1b}^{(kj)}} \\
&= \sum_{i=1}^{m} \frac{1}{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,k)} \frac{\gamma}{\sqrt{\sigma_B^2 + \epsilon}} \text{sgn}(\boldsymbol{z}_{1Bi}^{(k)}) \boldsymbol{x}_i^{(j)} \\
&= \frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,k)} \text{sgn}(\boldsymbol{z}_{1Bi}^{(k)}) \boldsymbol{x}_i^{(j)}.
\end{aligned}$$

So we get an vectorial form of the gradients of the loss function with respect to θ_{1b}:

$$\frac{\partial \boldsymbol{L}}{\partial \theta_{1b}} = \frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}} \sum_{i=1}^{m} ((\theta_{2b}^T (\hat{y}_{bi} - y_{bi})) \circ \text{sgn}(\boldsymbol{z}_{1Bi})) \boldsymbol{x}_i^T.$$

So far we have finished the gradients in branch B and we can solve for the gradients in Branch A. We first consider its gradients with respect to $\boldsymbol{z}_{2Ai}^{(j)}$:

$$\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Ai}^{(j)}} &= \frac{\partial}{\partial \boldsymbol{z}_{2Ai}^{(j)}} \frac{1}{m} \sum_{i=1}^{m} [\frac{1}{2} ||(\hat{y}_{ai} - y_{ai})||_2^2 - \sum_{k=1}^{n_{y_b}} y_{bi}^{(k)} log(\hat{y}_{bi}^{(k)})] \\
&= \frac{1}{2m} \frac{\partial}{\partial \boldsymbol{z}_{2Ai}^{(j)}} ||(\hat{y}_{ai} - y_{ai})||_2^2 - \frac{1}{m} \frac{\partial}{\partial \boldsymbol{z}_{2Ai}^{(j)}} \sum_{k=1}^{n_{y_b}} y_{bi}^{(k)} log(\hat{y}_{bi}^{(k)}) \\
&\text{note that } \boldsymbol{z}_{2Ai} = \hat{y}_{ai} \\
&= \frac{1}{2m} \frac{\partial}{\partial \boldsymbol{z}_{2Ai}^{(j)}} (\boldsymbol{z}_{2Ai}^{(j)} - y_{ai}^{(j)})^2 + \sum_{k=1}^{n_{yb}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Bi}^{(k)}} \frac{\partial \boldsymbol{z}_{2Bi}^{(k)}}{\partial \boldsymbol{z}_{2Ai}^{(j)}} \\
&= \frac{1}{m} (\boldsymbol{z}_{2Ai}^{(j)} - y_{ai}^{(j)}) + \frac{1}{m} \sum_{k=1}^{n_{yb}} (\hat{y}_{bi}^{(k)} - y_{bi}^{(k)}) \theta_{2b}^{(kj)} \\
&= \frac{1}{m} (\hat{y}_{ai}^{(j)} - y_{ai}^{(j)}) + \frac{1}{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,j)}.
\end{aligned}$$

Therefore the gradients of the loss with respect of θ_{2a} is:

$$\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \theta_{2a}^{(kj)}} &= \sum_{i=1}^{m} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Ai}^{(k)}} \frac{\partial \boldsymbol{z}_{2Ai}^{(k)}}{\partial \theta_{2a}^{(kj)}} \\
&= \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}_{ai}^{(k)} - y_{ai}^{(k)}) + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,k)}] \boldsymbol{d}_i^{(j)}.
\end{aligned}$$

In vectorial form:

$$\frac{\partial \boldsymbol{L}}{\partial \theta_{2a}} = \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}_{ai} - y_{ai}) + \theta_{2b}^T (\hat{y}_{bi} - y_{bi})] \boldsymbol{d}_i.$$

Next we consider the residual of the dropout layer:

$$\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{d}_i^{(j)}} &= \sum_{k=1}^{n_{ya}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{2Ai}^{(k)}} \frac{\partial \boldsymbol{z}_{2Ai}^{(k)}}{\partial \boldsymbol{d}_i^{(j)}} \\
&= \frac{1}{m} \sum_{k=1}^{n_{ya}} [(\hat{y}_{ai}^{(k)} - y_{ai}^{(k)}) + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}^{(:,k)}] \theta_{2a}^{(kj)} \\
&= \frac{1}{m} [(\hat{y}_{ai} - y_{ai})^T + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}] \theta_{2a}^{(:,j)}.
\end{aligned}$$

Then we look at the residual of $\boldsymbol{FC}_{1A}$:

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{1Ai}^{(j)}} &= \sum_{k=1}^{n_{1a}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{d}_i^{(k)}} \frac{\partial \boldsymbol{d}_i^{(k)}}{\partial \boldsymbol{z}_{1Ai}^{(j)}} \\
&= \sum_{k=1}^{n_{1a}} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{d}_i^{(k)}} \frac{\partial \boldsymbol{d}_i^{(k)}}{\partial \boldsymbol{a}_{1Ai}^{(j)}} \frac{\partial \boldsymbol{a}_{1Ai}^{(j)}}{\partial \boldsymbol{z}_{1Ai}^{(j)}} \\
&= \frac{1}{m} \sum_{k=1}^{n_{1a}} [(\hat{y}_{ai} - y_{ai})^T + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}] \theta_{2a}^{(:,k)} \boldsymbol{M}_j \mathrm{sgn}(\boldsymbol{a}_{1Ai}^{(j)}) \\
&= \frac{1}{m} [(\hat{y}_{ai} - y_{ai})^T + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}] \theta_{2a} \boldsymbol{U} \boldsymbol{M}_j \mathrm{sgn}(\boldsymbol{a}_{1Ai}^{(j)}),
\end{aligned}
$$

where

$$
\boldsymbol{U} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n_{1a} \times 1} .
$$

Using the chain rule and we can get the gradients of the loss with respect to $\theta_{1a}^{(kj)}$:

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \theta_{1a}^{(kj)}} &= \sum_{i=1}^{m} \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}_{1Ai}^{(k)}} \frac{\partial \boldsymbol{z}_{1Ai}^{(k)}}{\partial \theta_{1a}^{(kj)}} \\
&= \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}_{ai} - y_{ai})^T + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}] \theta_{2a} \boldsymbol{U} \boldsymbol{M}_k \mathrm{sgn}(\boldsymbol{a}_{1Ai}^{(k)}) \boldsymbol{x}_i^{(j)}.
\end{aligned}
$$

In vectorial form:

$$
\frac{\partial \boldsymbol{L}}{\partial \theta_{1a}} = \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}_{ai} - y_{ai})^T + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}] \theta_{2a} \boldsymbol{U} (\boldsymbol{M} \circ \mathrm{sgn}(\boldsymbol{a}_{1Ai})) \boldsymbol{x}_i^T.
$$

To sum up, the gradients of the overall loss function with respect to the parameters at each layer corresponding to a batch of samples are:

$$
\begin{aligned}
\frac{\partial \boldsymbol{L}}{\partial \theta_{2b}} &= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi}) \boldsymbol{s}^T \\
\frac{\partial \boldsymbol{L}}{\partial \theta_{2a}} &= \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}_{ai} - y_{ai}) + \theta_{2b}^T (\hat{y}_{bi} - y_{bi})] \boldsymbol{d}_i \\
\frac{\partial \boldsymbol{L}}{\partial \theta_{1b}} &= \frac{\gamma}{m\sqrt{\sigma_B^2 + \epsilon}} \sum_{i=1}^{m} ((\theta_{2b}^T (\hat{y}_{bi} - y_{bi})) \circ \mathrm{sgn}(\boldsymbol{z}_{1Bi})) \boldsymbol{x}_i^T \\
\frac{\partial \boldsymbol{L}}{\partial \theta_{1a}} &= \frac{1}{m} \sum_{i=1}^{m} [(\hat{y}_{ai} - y_{ai})^T + (\hat{y}_{bi} - y_{bi})^T \theta_{2b}] \theta_{2a} \boldsymbol{U} (\boldsymbol{M} \circ \mathrm{sgn}(\boldsymbol{a}_{1Ai})) \boldsymbol{x}_i^T \\
\frac{\partial \boldsymbol{L}}{\partial \beta} &= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b} \boldsymbol{R} \\
\frac{\partial \boldsymbol{L}}{\partial \gamma} &= \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_{bi} - y_{bi})^T \theta_{2b} \hat{\boldsymbol{x}}_i^T
\end{aligned}
$$

■