# Elastic Weight Consolidation: Derivation and Extension

Peng YUN

20200907

## 1 Derivation

The prior-based methods, like EWC[1], consider minimizing the statistical risk of all seen tasks as finding the most probable parameters given data $\mathcal{D} = \cup_{t=1}^{\mathcal{T}} \mathcal{D}$. They approximate the statistical risk by the negative logarithm of posterior probability:

$$-\log p(\boldsymbol{\theta}|\mathcal{D}) = -[\log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D})] \tag{1}$$

They consider the incremental learning in a sequential manner and consider $\mathcal{D} = \cup_{t=1}^{\mathcal{T}-1}\mathcal{D}^{(t)} \cup \mathcal{D}^{\mathcal{T}} = \mathcal{D}_A \cup D_B$, where $\mathcal{D}_A = \cup_{t=1}^{\mathcal{T}-1}\mathcal{D}^{(t)}$ denotes old tasks $\mathcal{D}_B = \mathcal{D}^{\mathcal{T}}$ denotes new tasks. Therefore, we have

$$\begin{aligned}
\log p(\boldsymbol{\theta}|\mathcal{D}) &= \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D}) \\
&= \log p(\mathcal{D}_A, \mathcal{D}_B|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D}_A, \mathcal{D}_B) \\
&= \log p(\mathcal{D}_A|\boldsymbol{\theta}) + \log p(\mathcal{D}_B|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D}_A) - \log p(\mathcal{D}_B) \\
&= [\log p(\mathcal{D}_A|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D}_A)] + \log p(\mathcal{D}_B|\boldsymbol{\theta}) - \log p(\mathcal{D}_B) \\
&= \log p(\boldsymbol{\theta}|\mathcal{D}_A) + \log p(\mathcal{D}_B|\boldsymbol{\theta}) - \log p(\mathcal{D}_B) \\
&= \log p(\boldsymbol{\theta}|\mathcal{D}_A) + \log p(\mathcal{D}_B|\boldsymbol{\theta}) + \text{const.}
\end{aligned} \tag{2}$$

where $\log p(\boldsymbol{\theta}|\mathcal{D}_A)$ is the posterior probability of parameters on old-task data, $\log p(\mathcal{D}_B|\boldsymbol{\theta})$ is the likelihood of new-task data. To optimize $\log p(\boldsymbol{\theta}|\mathcal{D})$, our objective function:

$$\begin{aligned}
\boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) \\
\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) &= -\log p(\boldsymbol{\theta}|\mathcal{D}_A) - \log p(\mathcal{D}_B|\boldsymbol{\theta})
\end{aligned} \tag{3}$$

We have new-task dataset $\mathcal{D}_B$ and the parametric model $f(\cdot;\boldsymbol{\theta})$ at hand. After training the old task, we have the optimal parameter $\boldsymbol{\theta}^*$. When training the new task, the old-task dataset $\mathcal{D}_A$ is intractable. Let's first consider the

term $-\log p(\mathcal{D}_B|\boldsymbol{\theta})$.

$$\begin{aligned}
-\log p(\mathcal{D}_B|\boldsymbol{\theta}) &= -\sum_i \log p(x_B^{(i)}, y_B^{(i)}|\boldsymbol{\theta}) \\
&= -\sum_i [\log p(y_B^{(i)}|\boldsymbol{\theta}, x_B^{(i)}) + \log p(x_B^{(i)}|\boldsymbol{\theta})] \\
&= -\sum_i [\log p(y_B^{(i)}|\boldsymbol{\theta}, x_B^{(i)}) + \text{const.}] \\
&= -\sum_i \log p(y_B^{(i)}|\boldsymbol{\theta}, x_B^{(i)}) + \text{const.}
\end{aligned} \tag{4}$$

We consider the conditional probability of output given input: $p(y|x, \boldsymbol{\theta}) \sim \mathcal{M}(f(x, \boldsymbol{\theta}))$, where $\mathcal{M}$ denotes the Multinomial distribution:

$$\begin{aligned}
p(y|x, \boldsymbol{\theta}) &= \Pi_{c \in C} f_c(x, \boldsymbol{\theta})^{y_c} \\
\log p(y|x, \boldsymbol{\theta}) &= \sum_{c \in C} y_c \log f_c(x, \boldsymbol{\theta})
\end{aligned} \tag{5}$$

where the $y_c$ and $f_c(\cdot)$ represent the $c$-th component of $y$ and $f(\cdot)$, and $C$ represent the task-related class set. Therefore, the term $-\log p(\mathcal{D}_B|\boldsymbol{\theta})$ will be:

$$-\log p(\mathcal{D}_B|\boldsymbol{\theta}) = -\sum_i \sum_{c \in C_B} y_{B,c}^{(i)} \log f_c(x_B^i, \boldsymbol{\theta}) + \text{const.} \tag{6}$$

Minimizing the term $-\log p(\mathcal{D}_B|\boldsymbol{\theta})$ is the same as minimizing the new-task loss function for classification problems.

Now we will unfold the term $\log p(\boldsymbol{\theta}|\mathcal{D}_A)$. Please note that we have $\boldsymbol{\theta}^*$ at hand, and $\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}_*} = \mathbf{0}$.

$$\begin{aligned}
\log p(\boldsymbol{\theta}|\mathcal{D}_A) &= \log p(\mathcal{D}_A|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathcal{D}_A) \\
&= \log p(\mathcal{D}_A|\boldsymbol{\theta}) + \text{const.} \\
&= \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}^*} + \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}_*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
&\quad + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}_*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \text{const.} \\
&= \text{const.} + \mathbf{0} + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}_*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\
-\log p(\boldsymbol{\theta}|\mathcal{D}_A) &= -\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}_*} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \text{const.}
\end{aligned} \tag{7}$$

We denote $\boldsymbol{H} = \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(\mathcal{D}_A|\boldsymbol{\theta})|_{\boldsymbol{\theta}_*}$. Therefore, our objective function becomes:

$$\begin{aligned}
\boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) \\
\mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) &= \mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta}) - \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)
\end{aligned} \tag{8}$$

We approximate the Hessian matrix $\boldsymbol{H}$ with the fisher information matrix:

$$\begin{aligned} \boldsymbol{H} = -\mathbb{I}, \mathbb{I} = \mathbb{E}[(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathcal{D}_A|\theta)|_{\boldsymbol{\theta}^*})^T (\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathcal{D}_A|\theta)|_{\boldsymbol{\theta}^*})] \\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) = \mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbb{I} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \end{aligned} \tag{9}$$

It is hard to do integration and compute the real fisher information matrix, but we can approximate the fisher information matrix with sampling $\tilde{\mathcal{D}}_A$ from $\mathcal{D}_A$:

$$\begin{aligned} \mathbb{I} &= \mathbb{E}_{\tilde{\mathcal{D}}_A \sim \mathcal{D}_A}[(\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\tilde{\mathcal{D}}_A|\theta)|_{\boldsymbol{\theta}^*})^T (\frac{\partial}{\partial \boldsymbol{\theta}} \log p(\tilde{\mathcal{D}}_A|\theta)|_{\boldsymbol{\theta}^*})] \\ &= \sum_{\tilde{\mathcal{D}}_A} (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(x_A^{(j)}, y_A^{(j)}|\theta)|_{\boldsymbol{\theta}^*})^T (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(x_A^{(j)}, y_A^{(j)}|\theta)|_{\boldsymbol{\theta}^*}) \end{aligned} \tag{10}$$

For the logarithm of the joint distribution:

$$\begin{aligned} \log p(x, y|\boldsymbol{\theta}) &= \log p(y|x, \boldsymbol{\theta}) + \log p(x|\boldsymbol{\theta}) \\ &= \log p(y|x, \boldsymbol{\theta}) + \log p(x) \end{aligned} \tag{11}$$

We consider the conditional probability of output given input as Mutinomial distribution as before and plug the equation 5 into above equation:

$$\begin{aligned} \log p(x, y|\boldsymbol{\theta}) &= \sum_{c \in C} y_c \log f_c(x, \boldsymbol{\theta}) + \log p(x) \\ \frac{\partial}{\partial \boldsymbol{\theta}} \log p(x, y|\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{c \in C} y_c \log f_c(x, \boldsymbol{\theta}) \end{aligned} \tag{12}$$

Therefore, the fisher information matrix is approximated by:

$$\mathbb{I} = \sum_{\tilde{\mathcal{D}}_A} (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{c \in C} y_{A,c}^{(j)} \log f_c(x_{A,c}^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*})^T (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{c \in C} y_{A,c}^{(j)} \log f_c(x_{A,c}^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}) \tag{13}$$

As a result, the objective function becomes:

$$\begin{aligned} \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_B) \\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_B) &= \mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbb{I} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \end{aligned} \tag{14}$$

where the fisher information matrix $\mathbb{I}$ stores the prior information of old tasks. The incremental learning process of EWC is: train network with task-A data, compute the fisher information matrix $\mathbb{I}_A$ on task-A data; train network with task-B data and $\mathbb{I}_A$, compute the fisher information matrix $\mathbb{I}_B$ on task-B data... The process is carried forward inductively to learn all the tasks.

## 2 Extend to regression problems

In the last section, we revisited the incremental learning method EWC in classification problems. The original paper did not introduce how to extend their method to regression problems. Here we bridge this small gap with details. According to our previous derivation, we have :

$$\begin{aligned} \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) \\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) &= -\log p(\boldsymbol{\theta}|\mathcal{D}_A) - \log p(\mathcal{D}_B|\boldsymbol{\theta}) \end{aligned} \tag{15}$$

For the first term, we have:

$$-\log p(\mathcal{D}_B|\boldsymbol{\theta}) = -\sum_i \log p(y_B^{(i)}|\boldsymbol{\theta}, x_B^{(i)}) + \text{const.} \tag{16}$$

We consider the conditional probability of output given input for regression problem: $p(y|x, \boldsymbol{\theta}) \sim \mathcal{N}(f(x, \boldsymbol{\theta}), \sigma^2 \boldsymbol{I})$, where $\mathcal{N}$ denotes the Gaussian distribution:

$$\begin{aligned} p(y|x, \boldsymbol{\theta}) &= \frac{1}{\sigma\sqrt{(2\pi)^k}} exp(-\frac{1}{2\sigma^2}(y - f(x, \boldsymbol{\theta}))^T(y - f(x, \boldsymbol{\theta}))) \\ \log p(y|x, \boldsymbol{\theta}) &= -\frac{1}{2\sigma^2}||y - f(x, \boldsymbol{\theta})||_2^2 + \text{const.} \end{aligned} \tag{17}$$

Then the term $-\log p(\mathcal{D}_B|\boldsymbol{\theta})$ will be:

$$-\log p(\mathcal{D}_B|\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\sum_i ||y - f(x, \boldsymbol{\theta})||_2^2 + \text{const.} \tag{18}$$

Minimizing the term $-\log p(\mathcal{D}_B|\boldsymbol{\theta})$ is the same as minimizing the L2-norm new-task loss function for regression problems.

By similar method unfolding the second term $\log p(\boldsymbol{\theta}|\mathcal{D}_A)$ as classification solutions, we can get the objective function as the classification case:

$$\begin{aligned} \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) \\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) &= \mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \end{aligned} \tag{19}$$

We also approximate the Hessian matrix $\boldsymbol{H}$ with the fisher information matrix $\mathbb{I}$, which we can obtain by:

$$\mathbb{I} = \sum_{\tilde{\mathcal{D}}_A}(\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(x_A^{(j)}, y_A^{(j)}|\theta)|_{\boldsymbol{\theta}^*})^T(\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(x_A^{(j)}, y_A^{(j)}|\theta)|_{\boldsymbol{\theta}^*}) \tag{20}$$

We consider the conditional probability of output given input as Gaussian as before and plug the equation 17 into above equation:

$$\mathbb{I} = \sum_{\tilde{\mathcal{D}}_A}(\frac{1}{2\sigma^2})^2(\sum_j \frac{\partial}{\partial \boldsymbol{\theta}}||y_A^{(j)} - f(x_A^{(j)}, \boldsymbol{\theta})||_2^2|_{\boldsymbol{\theta}^*})^T(\sum_j \frac{\partial}{\partial \boldsymbol{\theta}}||y_A^{(j)} - f(x_A^{(j)}, \boldsymbol{\theta})||_2^2|_{\boldsymbol{\theta}^*}) \tag{21}$$

The final objective function have the same form as equation 14 but the term $\mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta})$ and $\mathbb{I}$ are computed with L2-norm instead of cross-entropy due to the Gaussian distribution assumption for regression problems. It is noted that the L2-norm can be replaced with L1-norm if we consider the $p(y|x, \boldsymbol{\theta})$ subjects to a Laplacian distribution with mean $f(x, \boldsymbol{\theta})$ and covariance matrix $\sigma^2 \boldsymbol{I}$.

# 3 Extend to object detection problems

Object detection consists of both classification and regression problem. According to our previous derivation, we have :

$$\begin{aligned} \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) \\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) &= -\log p(\boldsymbol{\theta}|\mathcal{D}_A) - \log p(\mathcal{D}_B|\boldsymbol{\theta}) \end{aligned} \tag{22}$$

For the first term, we have:

$$-\log p(\mathcal{D}_B|\boldsymbol{\theta}) = -\sum_i \log p(y_B^{(i)}|\boldsymbol{\theta}, x_B^{(i)}) + \text{const.} \tag{23}$$

We consider the output $y$ of detection problem consists of two independent part: classification $y_{cls}$ and regression output $y_{reg}$; and the classification output subjects to the Multinomial distribution $p(y_{cls}|x, \boldsymbol{\theta}) \sim \mathcal{M}(f_{cls}(x, \boldsymbol{\theta}))$; the regression output subjects to the Gaussian distribution $p(y_{reg}|x, \boldsymbol{\theta}) \sim \mathcal{N}(f_{reg}(x, \boldsymbol{\theta}), \sigma^2 \boldsymbol{I})$.

$$\begin{aligned} p(y|x, \boldsymbol{\theta}) &= p(y_{cls}, y_{reg}|x, \boldsymbol{\theta}) \\ &= p(y_{cls}|x, \boldsymbol{\theta}) p(y_{reg}|x, \boldsymbol{\theta}) \\ \log p(y|x, \boldsymbol{\theta}) &= \log p(y_{cls}|x, \boldsymbol{\theta}) + \log p(y_{reg}|x, \boldsymbol{\theta}) \end{aligned} \tag{24}$$

Therefore, we have

$$\begin{aligned} -\log p(\mathcal{D}_B|\boldsymbol{\theta}) = &-\sum_i \sum_{c \in C} y_{B,cls,c}^{(i)} \log f_{cls,c}(x_B^i, \boldsymbol{\theta}) \\ &+ \frac{1}{2\sigma^2} \sum_i ||y_{B,reg}^{(i)} - f_{reg}(x_B^i, \boldsymbol{\theta})||_2^2 + \text{const.} \end{aligned} \tag{25}$$

which is the linear combination of classification loss and regression loss for object detection problem with new-task data. The $\frac{1}{2}\sigma^2$ can be treated as the hyperparameter to balance the weights between classification and regression loss terms.

By similar method unfolding the second term $\log p(\boldsymbol{\theta}|\mathcal{D}_A)$, we can get the objective function as the classification case:

$$\begin{aligned} \boldsymbol{\theta} &= \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) \\ \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_A, \mathcal{D}_B) &= \mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \end{aligned} \tag{26}$$

We also approximate the Hessian matrix $\boldsymbol{H}$ with the fisher information matrix $\mathbb{I}$, which we can obtain by:

$$\mathbb{I} = \sum_i (\frac{\partial}{\partial \boldsymbol{\theta}} \log p(x_A^{(i)}, y_A^{(i)}|\boldsymbol{\theta})|_{\boldsymbol{\theta}^*})^T (\frac{\partial}{\partial \boldsymbol{\theta}} \log p(x_A^{(i)}, y_A^{(i)}|\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}) \tag{27}$$

We consider the conditional probability of output given input as before and plug the equation 24 into above equation:

$$\begin{aligned}
\mathbb{I} &= \sum_{\tilde{\mathcal{D}}_A} \{[\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,cls}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*} + \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,reg}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}]^T \\
&\times [\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,cls}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*} + \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,reg}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}]\} \\
&= \sum_{\tilde{\mathcal{D}}_A} \{\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,cls}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*} \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,cls}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}^T \\
&+ \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,reg}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*} \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,reg}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}^T \\
&+ \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,cls}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*} \sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{A,reg}^{(j)}|x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}^T\} \\
&= \sum_{\tilde{\mathcal{D}}_A} \{(\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{c \in C} y_{A,cls,c}^{(j)} \log f_c(x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*})^T (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{c \in C} y_{A,cls,c}^{(j)} \log f_c(x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*}) \\
&+ (\frac{1}{2\sigma^2})^2 (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} ||y_{A,reg}^{(j)} - f_{reg}(x_A^{(j)}, \boldsymbol{\theta})||_2^2|_{\boldsymbol{\theta}^*})^T (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} ||y_{A,reg}^{(j)} - f_{reg}(x_A^{(j)}, \boldsymbol{\theta})||_2^2|_{\boldsymbol{\theta}^*}) \\
&+ (\frac{1}{2\sigma^2})(\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{c \in C} y_{A,cls,c}^{(j)} \log f_c(x_A^{(j)}, \boldsymbol{\theta})|_{\boldsymbol{\theta}^*})^T (\sum_j \frac{\partial}{\partial \boldsymbol{\theta}} ||y_{A,reg}^{(j)} - f_{reg}(x_A^{(j)}, \boldsymbol{\theta})||_2^2|_{\boldsymbol{\theta}^*})\}
\end{aligned} \tag{28}$$

where the first term measures the amount of information for classification part, the second term is for regression part, parameters will cause great value in the last term if they are important for both classification and regression parts.

The final objective function have the same form as equation 14 but the term $\mathcal{L}(\mathcal{D}_B, \boldsymbol{\theta})$ represents the weighted linear combination of classification and regression losses. The term $\mathbb{I}$ are approximated with three terms considering both classification and regression problems. As before, the L2-norm can be replaced with L1-norm if we consider the $p(y|x, \boldsymbol{\theta})$ subjects to a Laplacian distribution with mean $f(x, \boldsymbol{\theta})$ and covariance matrix $\sigma^2 \boldsymbol{I}$. It also can be extended to other robust functions, like huber loss, for better performance.

# References


[1] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ra-

malho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017.