# The Role of Representation Learning in Continual Learning

## 1 Preliminary

Let $d, k, r, t \in N, \epsilon \in (0, 1/2)$. The continual learning problem is defined over k environments $D_1, ..., D_k$. After learning data $D_t$ from task $t$, the learner $f(x)$ consists of two parts:

1. representation function $R_{w_t}(x) \in \mathbb{R}^r$
2. task-dependent linear classifiers $v_t \in \mathbb{R}^r$

Define the matrix $V_t = [v_1, v_2, ..., v_t]$ where each column is a classifier. The prediction of the t-th environment is made by $f(x) = \langle v_t, R_{w_t}(x) \rangle$.

**Assumption 1** (Global Representation)**.** *There exists a function $R^*$ and a sequence of linear classifiers $v_1^*, ..., v_k^* \in \mathbb{R}^r$ such that for any $(x, y) \sim D_i(i \in [k])$, the label $y$ satisfies:*

$$y = \langle v_i^*, R^*(x) \rangle + z, \; z \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

## 2 Proof for Upper bound of All Tasks

Suppose we can find $R_{w_k}$ such that for all $k$:

$$|R_{w_k}(x)v_k - y_k|^2 < \delta \tag{2}$$

$$R_{w_k}V_{k-1} = R_{w_{k-1}}V_{k-1} \tag{3}$$

Then first, for $V_{k-2}$:

$$|R_{w_k}(x)V_{k-2} - R_{w_{k-1}}(x)V_{k-2}|^2 \leq |R_{w_k}(x)V_{k-1} - R_{w_{k-1}}(x)V_{k-1}|^2 = 0 \tag{4}$$

Due to triangle inequality, we can see

$$|R_{w_k}(x)V_{k-2} - R_{w_{k-2}}(x)V_{k-2}|^2 \leq |R_{w_k}(x)V_{k-2} - R_{w_{k-1}}(x)V_{k-2}|^2 \tag{5}$$

$$+|R_{w_{k-1}}(x)V_{k-2} - R_{w_{k-2}}(x)V_{k-2}|^2 = 0 + 0 = 0 \tag{6}$$

And again:

$$|R_{w_k}(x)V_{k-3} - R_{w_{k-2}}(x)V_{k-3}|^2 \leq |R_{w_k}(x)V_{k-2} - R_{w_{k-2}}(x)V_{k-2}|^2 = 0 \tag{7}$$

Due to triangle inequality, we can see

$$|R_{w_k}(x)V_{k-3} - R_{w_{k-3}}(x)V_{k-3}|^2 \leq |R_{w_k}(x)V_{k-3} - R_{w_{k-2}}(x)V_{k-3}|^2 \tag{8}$$
$$+|R_{w_{k-2}}(x)V_{k-3} - R_{w_{k-3}}(x)V_{k-3}|^2 = 0 \tag{9}$$

Therefore, we can prove that

$$|R_{w_k}(x)V_{k-t} - R_{w_{k-t}}(x)V_{k-t}|^2 = 0, \quad \forall t \in [0, 1, 2, ..., k-1] \tag{10}$$

From Eq. 10 we can see that:

$$|R_{w_k}(x)v_{k-t} - R_{w_{k-t}}(x)v_{k-t}|^2 = 0 \tag{11}$$

Combine Eq. 11 and 2, we come up our goal:

$$|R_{w_k}(x)v_{k-t} - y_{k-t}|^2 \leq |R_{w_k}(x)v_{k-t} - R_{w_{k-t}}(x)v_{k-t}|^2 + |R_{w_{k-t}}(x)v_{k-t} - y_{k-t}|^2 < \delta \tag{12}$$

Hence, we prove that for all previous task $k-t$, the error is bounded.

# 3 Idea for Nonlinear Representation

**Assumption 2** (Distribution assumption)**.** *For any $i \in [k]$, the distribution of $x$, $p(x)$, is the same. However, $p(y|x)$ is different.*

**Goal** : Find $R_{w_t}$, $V'_{t-1}$ and $v_t$ such that for all $t > 1$,
$|R_{w_t}(X_t)v_t - \mathbf{y}_t|^2 < \epsilon$ and $R_{w_t}(x)V'_{t-1} = R_{w_{t-1}}(x)V_{t-1}$, where $\epsilon \in (0, \frac{1}{2})$, $V'_{t-1}$ is learned in task $t$, $X_t \in \mathbb{R}^{n_1 \times d}$ and $\mathbf{y}_t \in \mathbb{R}^{n_1}$ means the $n_1$ collected samples of input matrix and output vector.

**Algorithm** : The optimization goal is as follows:

$$\begin{aligned} &\min_{w_t, v_t} |R_{w_t}(X_t)v_t - \mathbf{y}_t|^2 \\ &\text{subject to } R_{w_t}(X_t)V'_{t-1} = R_{w_{t-1}}(X_t)V_{t-1} \end{aligned} \tag{13}$$

Follow section 5 in [1], we use Guassian width to characterize the complexity of representation function class $\Phi$ where $\forall R_w \in \Phi$.

**Definition 3.1** (Gaussian width)**.** *Given a set $\mathcal{K} \subset \mathbb{R}^M$, the Gaussian width of $\mathcal{K}$ is defined as*

$$\mathcal{G}(\mathcal{K}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})} \sup_{\mathbf{v} \in \mathcal{K}} \langle \mathbf{v}, \mathbf{z} \rangle \tag{14}$$

*and we use Gaussian width to measure the complexity of $\Phi$ that depends on the input data $\mathcal{X}$:*

$$\mathcal{F}_{\mathcal{X}}(\Phi) = \{\mathbf{a} \in \mathbb{R}^{n_1} : |\mathbf{a}|_F = 1, \exists R_w, R'_w \in \Phi s.t. \mathbf{a} \in span([R_w, R'_w])\} \tag{15}$$

**Theorem 3.1.** *Similar to Claim 5.3 in [1], Let $\hat{R}_{w_t}$ and $\hat{v}_1, ..., \hat{v}_t$ be the optimal solution to 13, Then with probability at least $1-\delta$ we have*

$$|R^*_{w_t}(X_t)v^*_t - \hat{R}_{w_t}(X_t)\hat{v}_t|^2 \le \sigma^2(\mathcal{G}(\mathcal{F}_{\mathcal{X}}(\Phi)) + \sqrt{log\frac{1}{\delta}})^2 \tag{16}$$

*Proof.* By the optimality of $\hat{R}_{w_t}$ and $\hat{v}_1, ..., \hat{v}_t$ for 13, we know

$$\begin{aligned} &|\hat{R}_{w_t}(X_t)\hat{v}_t - \mathbf{y}_t|^2 \le |R^*_{w_t}(X_t)v^*_t - \mathbf{y}_t|^2 \text{ and} \\ &\hat{R}_{w_t}(X_t)\hat{V}_{t-1} = R_{w_{t-1}}(X_t)V_{t-1} \end{aligned} \tag{17}$$

Recall Assumption 1, we get

$$|R^*_{w_t}(X_t)v^*_t + \mathbf{z}_t - \hat{R}_{w_t}(X_t)\hat{v}_t|^2 \le |\mathbf{z}_t|^2 \tag{18}$$

which indicates

$$|R^*_{w_t}(X_t)v^*_t - \hat{R}_{w_t}(X_t)\hat{v}_t|^2 \le 2\langle \mathbf{z}_t, R^*_{w_t}(X_t)v^*_t - \hat{R}_{w_t}(X_t)\hat{v}_t \rangle \tag{19}$$

Denote $\mathbf{a} = R^*_{w_t}(X_t)v^*_t - \hat{R}_{w_t}(X_t)\hat{v}_t \in \mathbb{R}^{n_1}$, the above equality reads $|\mathbf{a}|^2_F \le 2\langle \mathbf{z}_t, \mathbf{a}\rangle$. We can get

$$|\mathbf{a}|_F \le 2\langle \mathbf{z}_t, \frac{\mathbf{a}}{|\mathbf{a}|_F}\rangle \le 2 \underset{\bar{\mathbf{a}}\in\mathcal{F}_{\mathcal{X}}(\Phi)}{sup} \langle \bar{\mathbf{a}}, \mathbf{z}_t\rangle \tag{20}$$

By definition 3.1, we know $\mathbb{E}_{\mathbf{z}_t}[\underset{\bar{\mathbf{a}}\in\mathcal{F}_{\mathcal{X}}(\Phi)}{sup} \langle \mathbf{a}, \sigma^{-1}\mathbf{z}_t\rangle] = \mathcal{G}(\mathcal{F}_{\mathcal{X}}(\Phi))$. By Chebyshev's inequality, we probability at least $1-\delta$

$$\underset{\bar{\mathbf{a}}\in\mathcal{F}_{\mathcal{X}}(\Phi)}{sup} \langle \bar{\mathbf{a}}, \sigma^{-1}\mathbf{z}_t\rangle \le \mathbb{E}_{\mathbf{z}_t}[\underset{\bar{\mathbf{a}}\in\mathcal{F}_{\mathcal{X}}(\Phi)}{sup} \langle \mathbf{a}, \sigma^{-1}\mathbf{z_t}\rangle] + \sqrt{log\frac{1}{\delta}} = \mathcal{G}(\mathcal{F}_{\mathcal{X}}(\Phi)) + \sqrt{log\frac{1}{\delta}} \tag{21}$$

Therefore, original objective is bounded by

$$|R^*_{w_t}(X_t)v^*_t - \hat{R}_{w_t}(X_t)\hat{v}_t|^2 \le \sigma^2(\mathcal{G}(\mathcal{F}_{\mathcal{X}}(\Phi)) + \sqrt{log\frac{1}{\delta}})^2 \tag{22}$$

□

# 4 Few Shot Class Incremental Learning

## 4.1 Setting

Learner $f(x)$ is composed of two parts: the representation function $R(x)$ and linear classifier $v$. Few shot class incremental learning problem is defined over $T$ tasks $\{D_0, D_1, ..., D_T\}$, and the corresponding label of $D_t$ is $C_t$. Different tasks have no overlapped classes, i.e. $\forall i, j$ and $i \ne j$, $C_i \cap C_j = \varnothing$. First dataset $D_0$ is bigger than the rest, i.e. $|D_0| >> |D_t|$ for $t > 0$.

**Assumption 3** (Fixed Representation)**.** *Assume learner can learn good representation only from the first task. When learning task $t > 0$, we only train the classifier $v$ and fix $R(x)$.*

The goal is to learn new classes $C_t$ from only a few new training examples without forgetting knowledge about old classes $C_j, j < t$. At the evaluation session after learning task $t$, test dataset includes samples from previous tasks and current task, i.e. the label space of $C_0 \cup C_1 ... \cup C_t$. Let $\hat{C}_t = C_0 \cup C_1 ... \cup C_t = \{\hat{c}_t^1, \hat{c}_t^2, ..., \hat{c}_t^{n_t}\}$, where $n_t$ is the number of classes from 0-th task to $t$-th task.

## 4.2 Linear Representation

Task identity is unknown in the evalution session of class incremental learning, so learner needs to learn $v$ continually. Define $v_t$ as the classifier after learning task $t$. There are two kinds of classifier:

- (1) $v_t \in \mathbb{R}^m$. It outputs the predicted label $\hat{y}_t$ directly, which is the same as [1].
- (2) $v_t \in \mathbb{R}^{m \times n_T}$. It produces a output vector $o_t \in \mathbb{R}^{n_T}$, which need an activation function (e.g. softmax function) to predict the probability over all classes.

### 4.2.1 Ideas of (2)

By fixing $v_{t,i}$, $\forall i \in \hat{C}_{t-1}$, we can fix the probability of all classes in $\hat{C}_{t-1}$ when learning task $t$. However, this idea cannot maintain good performace of previous tasks since the probability of new tasks might be higher.

# References


[1] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.