The Ideal Continual Learner: An Agent That Never Forgets

JOHNS HOPKINS MATHEMATICAL INSTITUTE *for* DATA SCIENCE

Liangzu Peng[1] Paris V. Giampouras[1] René Vidal[2,3]

[1]JHU [2]UPenn [3]NORCE Norwegian Research Centre

Penn UNIVERSITY *of* PENNSYLVANIA

Continual Learning: Problem Setup

- **Learning Tasks**: Given T tasks, $t = 1, \dots, T$

$$\mathcal{G}_t \coloneqq \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} L_t(w; D_t)$$

Loss and data for task t

Global minimizers for task t

- **Goal**: learn a model $\widehat{w}_T$ that solves all tasks when presented sequentially to the learner.
- **Catastrophic Forgetting Challenge**: model $\widehat{w}_T$ may perform poorly on previous tasks

Methods to Prevent Forgetting

- *Regularization-based*, e.g.,

$$\min_{w \in \mathbb{R}^n} L_t(w; D_t) + \delta \cdot \left\| w - \widehat{w}_{t-1} \right\|_2$$

- *Memory-based*, e.g., train with current data and part of previous data (rehearsal)
- *Expansion-based:* add new parameters for new tasks and learn only new parameters

✓ These methods greatly improve empirical performance in the deep learning context.

✗ However, there is limited theory explaining their empirical success.

The Ideal Continual Learner (ICL)

With $\mathcal{K}_0 \coloneqq \mathbb{R}^n$, **ICL** is a method that solves

$$\mathcal{K}_t \coloneqq \underset{w \in \mathcal{K}_{t-1}}{\operatorname{argmin}} L_t(w; D_t)$$

sequentially, for $t = 1, 2, \dots, T$.

Assumption 1 (Shared Multitask Model):

$$\cap_{t=1}^{T} \mathcal{G}_t \neq \emptyset$$

ICL Never Forgets by Design

Under Assumption 1, $\mathcal{K}_t = \cap_{i=1}^{t} \mathcal{G}_i$ for every $t = 1, \dots, T$, thus **ICL** never forgets.

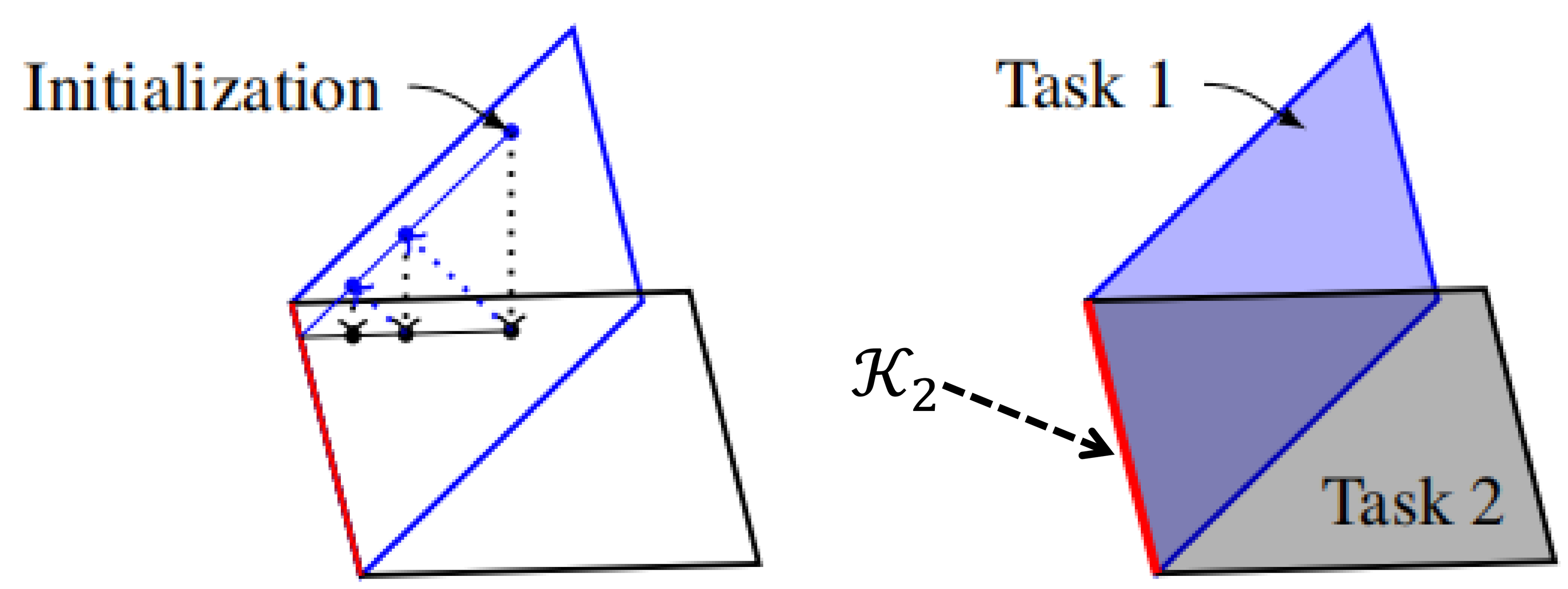
ICL Example: Continual Regression

$$L_t\big(w; (X_t, y_t)\big) \coloneqq \left\| X_t w - y_t \right\|_2^2$$

- $\mathcal{G}_t$ is an affine subspace:

$$\mathcal{G}_t = \{ w \in \mathbb{R}^n : X_t^\top X_t w = X_t^\top y_t \}$$

- Under Assumption 1, $\mathcal{K}_t$ is the intersection of the first t subspaces ($\mathcal{K}_t = \cap_{i=1}^{t} \mathcal{G}_i$):



Alternating Projection ICL

Representing $\mathcal{K}_t$. We need extra notations:

- $\widehat{w}_t \in \mathcal{K}_t$: shared solution to tasks $1, \dots, t$
- B_t: orthonormal basis of $\cap_{i=1}^{t} \operatorname{nullspace}(X_i)$

Then $w \in \mathcal{K}_t \Leftrightarrow w = \widehat{w}_t + B_t a$ for some a.

Implementing ICL. It suffices to maintain and update $(\widehat{w}_t, B_t)$ for every $t = 1, 2, \dots, T$:

- $(t = 1)$ Compute $(\widehat{w}_1, B_1)$ from (X_1, y_1) by using an SVD on X_1 and solving task 1
- $(t > 1)$ Compute $(\widehat{w}_t, B_t)$ from (X_t, y_t) and $(\widehat{w}_{t-1}, B_{t-1})$ by solving:

$$\min_{w \in \mathcal{K}_{t-1}} \left\| X_t w - y_t \right\|_2^2$$

$$\Leftrightarrow \min_{a} \left\| X_t(\widehat{w}_{t-1} + B_{t-1} a) - y_t \right\|_2^2$$

Our Paper Furthermore Discusses:

- generalization & optimization of **ICL**
- wider networks ⇒ less forgetting
- how **ICL** unifies many existing methods
- **ICL** connections to many other topics

Conclusion & Limitation:

- **ICL** is a general framework that never forgets
- Implementing **ICL** exactly is in general difficult, but approximating **ICL** is possible.

Acknowledgements: Work supported by the project ULEARN "Unsupervised Lifelong Learning" and co-funded under the grant number 316080 of the Research Council of Norway.