Learning in the Null Space: Small Singular Values for Continual Learning

Cuong Anh Pham[1], Praneeth Vepakomma[1,2], Samuel Horváth[1]

[1]MBZUAI [2]MIT

{cuong.pham, praneeth.vepakomma, samuel.horvath}@mbzuai.ac.ae

Summary

We introduce a new Continual Learning approach that utilizes small singular values, ensuring the LoRA-style weight updates are approximately lying in the null space of the input of previous tasks → better forgetting rates.

Challenges

- Catastrophic Forgetting
- Stability-Plasticity Balance

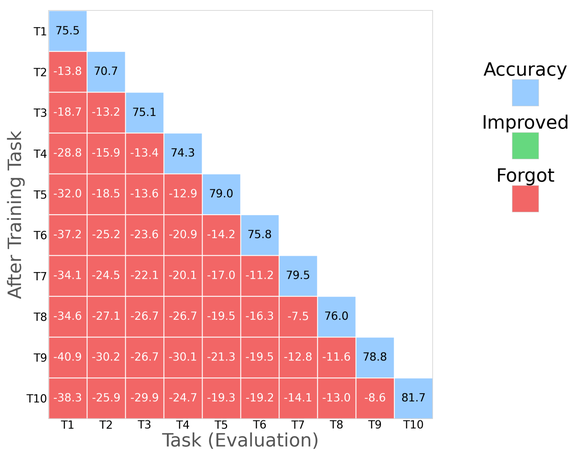


Figure: Continual Learning from scratch

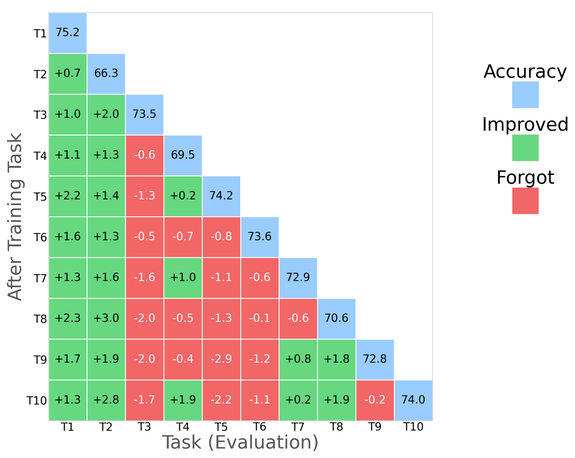


Figure: Continual Learning with reduced forgetting

Optimization Problem

- Let the weight update when learning task t is:

$$\Delta \mathbf{W}_t = \mathbf{W}_t - \mathbf{W}_{t-1}$$

- Let $\mathcal{I}_t$ denote the set of inputs collected from tasks $1, 2, \ldots, t-1$.

Stability Constraint (Output Preservation).

$$\left\|x^\top \Delta \mathbf{W}_t\right\|_2^2 \leq \varepsilon, \quad \forall x \in \mathcal{I}_t, \tag{1}$$

where ε controls the allowable output perturbation.

Plasticity Objective (New Task Learning).
Under the stability constraint, the learning problem for task t can be formulated conceptually as

$$\begin{aligned} \min_{\mathbf{W}_t} \quad & \mathcal{L}_{CE}(D_t, \mathbf{W}_t) \\ \text{s.t.} \quad & \left\|x^\top \Delta \mathbf{W}_t\right\|_2^2 \leq \varepsilon, \quad \forall x \in \mathcal{I}_t. \end{aligned} \tag{2}$$

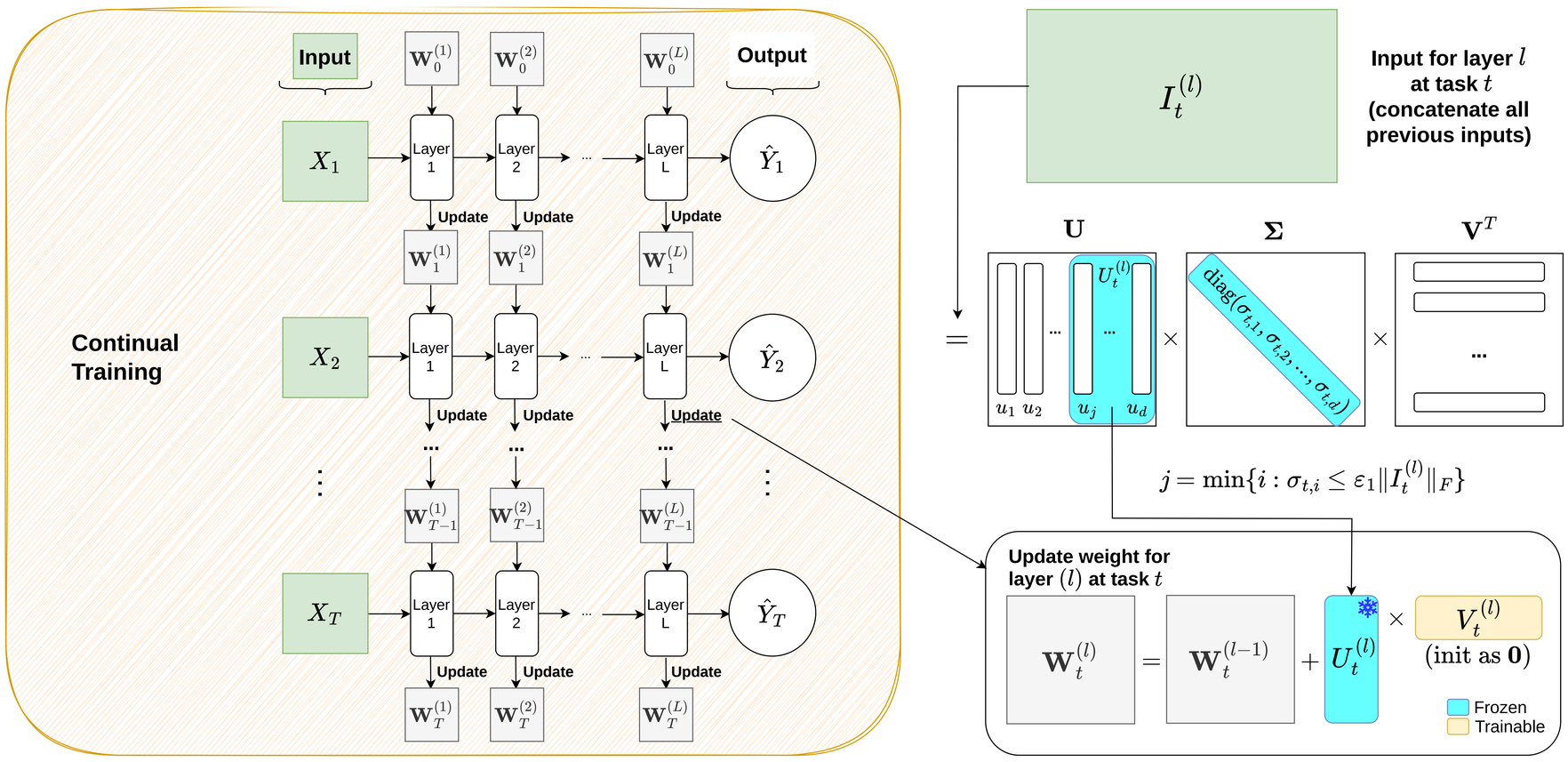


Figure: Overview of NESS (Null-space Estimated from Small Singular values)

Our Proposed Method

Construction of the Stability Subspace.
Using SVD:

$$I_t = [X_1 : X_2 : \cdots : X_{t-1}] = \tilde{U}_t \Sigma_t \tilde{V}_t^\top \tag{3}$$

where $\sigma_{t,1} \geq \sigma_{t,2} \geq \cdots \geq \sigma_{t,d} \geq 0$.

Selecting the Small-Singular-Value Subspace.
Let $\varepsilon_1 > 0$ and define

$$j = \min\left\{i : \sigma_{t,i} \leq \varepsilon_1 \|I_t\|_F\right\}. \tag{4}$$

$U_t = [u_{t,j} : \cdots : u_{t,d}]$ span an approximate null subspace. Defined the update as

$$\Delta \mathbf{W}_t = U_t V_t, \tag{5}$$

where U_t is fixed and only V_t is trainable.

Efficiency Trick.
Instead of combined input I_t, use the covariance matrix:

$$C_t = I_t I_t^\top = \sum_{i=1}^{d} x_{t,i} x_{t,i}^\top \tag{6}$$

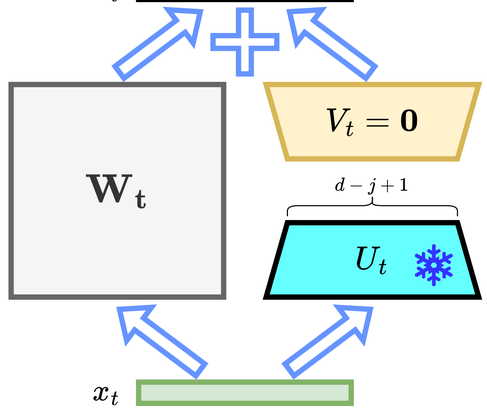


Figure: LoRA-style weights update

Theoretical Guarantees

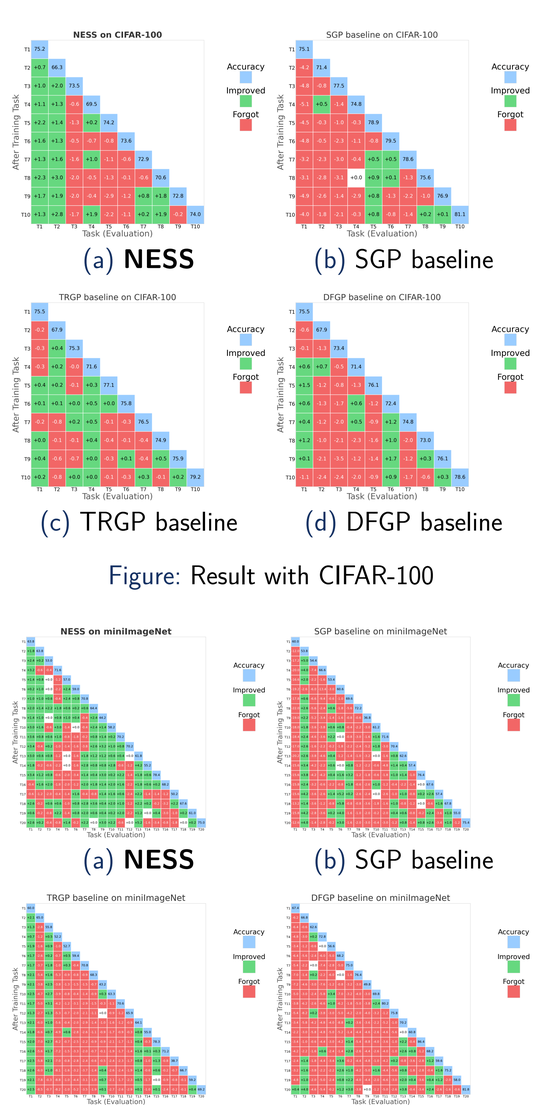
Explicit Stability Bound.
For every previous input $x \in \mathcal{I}_t$,

$$\|x^\top \Delta \mathbf{W}_t\|_2 \leq \varepsilon_1 \|I_t\|_F \|V_t\|_2. \tag{7}$$

If we enforce the bound of $\|V_t\|_2$ via **weight decay**, then $\|x^\top \Delta W_t\|_2^2 \leq \varepsilon, \quad \forall x \in \mathcal{I}_t$.

→ Updates lie in an approximate null subspace of previous inputs, ensuring bounded interference across tasks.

Sample Results



(a) **NESS** (b) SGP baseline

(c) TRGP baseline (d) DFGP baseline

Figure: Result with CIFAR-100

(a) **NESS** (b) SGP baseline

(c) TRGP baseline (d) DFGP baseline

Figure: Result with MiniImageNet

Experimental Results

Table: The averaged accuracy (ACC) and **backward transfer (BWT)** over all tasks on different datasets

Method	CIFAR-100 (10 Tasks)		5-datasets (5 Tasks)		MiniImageNet (20 Tasks)	
	ACC(%)	**BWT(%)↑**	ACC(%)	**BWT(%)↑**	ACC(%)	**BWT(%)↑**
OWM‡ [13]	50.94 ± 0.60	-30 ± 1	–	–	–	–
EWC‡ [11]	68.88 ± 0.80	-2 ± 1	88.64 ± 0.26	-4 ± 1	52.01 ± 2.53	-12 ± 3
HAT‡ [10]	72.06 ± 0.50	0 ± 0	91.32 ± 0.18	-1 ± 0	59.78 ± 0.57	-3 ± 0
A-GEM‡ [7]	63.98 ± 1.22	-15 ± 2	84.04 ± 0.33	-12 ± 1	57.24 ± 0.72	-12 ± 1
GPM [8]	71.63 ± 0.67	-0.28 ± 0.54	90.61 ± 0.57	-1.17 ± 0.26	63.56 ± 2.42	-1.39 ± 1.37
SGP [16]	75.99 ± 0.16	-1.18 ± 0.37	90.48 ± 0.48	-1.82 ± 0.13	64.45 ± 2.18	-0.53 ± 0.68
TRGP [15]	75.21 ± 0.32	**0.06 ± 0.17**	92.78 ± 0.65	**-0.09 ± 0.08**	62.74 ± 2.13	-1.23 ± 0.77
FS-DGPM† [14]	74.10 ± 0.09	-3.03 ± 0.31	–	–	–	–
DFGP [17, 18] (mixup=0.01)	73.24 ± 0.24	-0.95 ± 0.18	91.47 ± 0.22	-2.27 ± 0.31	68.50 ± 1.50	-0.11 ± 1.36
DFGP [17, 18] (mixup=0.05)	73.77 ± 0.33	-1.03 ± 0.41	90.22 ± 0.46	-3.87 ± 0.53	68.64 ± 2.25	-0.43 ± 1.60
DFGP [17, 18] (mixup=0.001)	73.32 ± 0.50	-1.09 ± 0.49	91.59 ± 0.50	-1.86 ± 0.25	67.75 ± 1.81	-1.11 ± 1.50
DFGP [17, 18] (mixup=0.0001)	73.16 ± 0.46	-1.11 ± 0.52	91.51 ± 0.15	-1.70 ± 0.42	68.39 ± 1.07	-0.16 ± 1.42
NESS (with SAM)	72.56 ± 0.07	-0.17 ± 0.51	90.98 ± 0.07	-0.86 ± 0.28	63.48 ± 1.38	-0.26 ± 0.67
NESS (with SGDm: m=0.9)	72.46 ± 0.26	0.03 ± 0.40	90.20 ± 0.47	-0.58 ± 0.15	63.72 ± 0.46	**0.41 ± 0.58**

Contact Information

- Web: https://pacman-ctm.github.io/
- Email: cuong.pham@mbzuai.ac.ae

Full Paper

Code