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

Mohamed bin Zayed
University of
Artificial Intelligence

Problem Formulation

Consider a Continual Learning (CL) setting with:

- T **sequential tasks**: $\{\text{Task}_1, \text{Task}_2, \ldots, \text{Task}_T\}$
- Dataset $D_t = (X_t, Y_t) = \{(x_{t,i}, y_{t,i})\}_{i=1}^{N_t}$ at task t.
- Input $x_{t,i} \in X_t \in \mathbb{R}^d$ and output $y_{t,i} \in Y_t \in \mathbb{R}^{d_{\text{out}}}$

Tasks are learned sequentially; only data from D_t is available for training at task t.

The goal of CL: Learn parameters $\mathbf{W}_T \sim \mathbf{W}$ that perform well on all tasks $\{1, \ldots, T\}$. A common formulation of the objective is

$$\mathbf{W} = \arg\min_{\mathbf{W}} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t(\mathbf{W}), \tag{1}$$

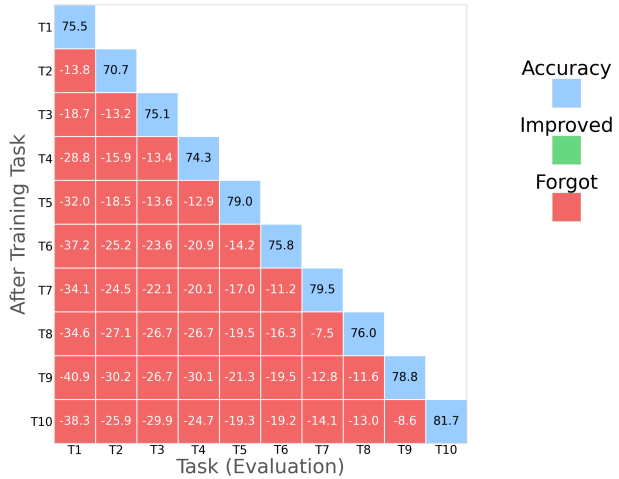
where $\mathbf{W} \in \mathbb{R}^{d \times d_{\text{out}}}$ and $\mathcal{L}_t(\mathbf{W})$ denotes the loss associated with task t.

- Eq (1) cannot be optimized directly.

Continual Learning

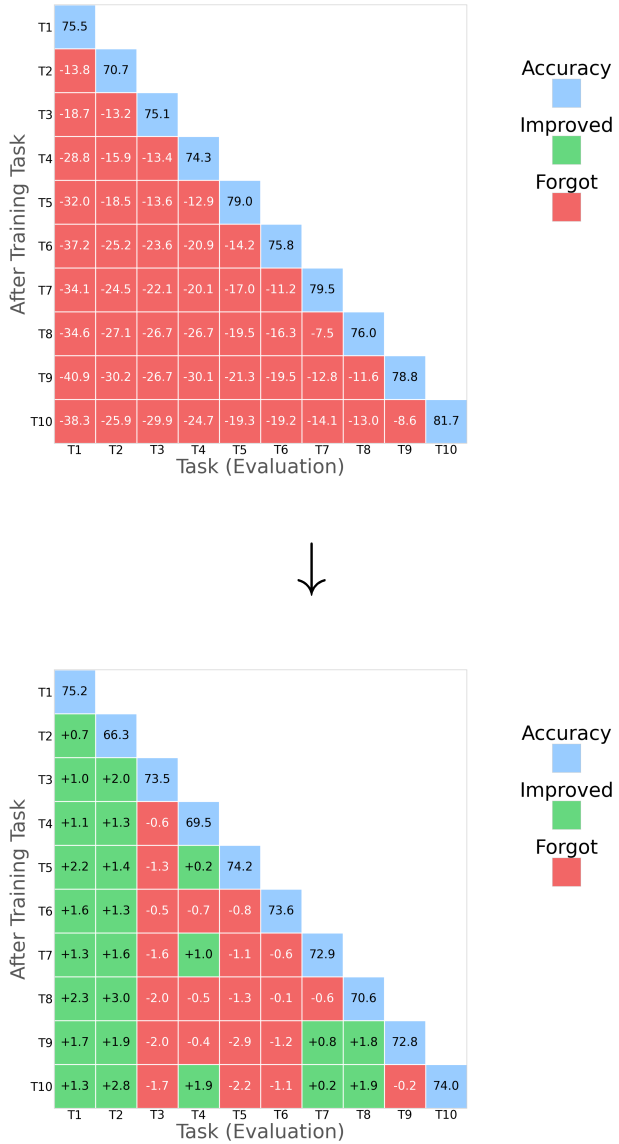
Challenges:

- Catastrophic Forgetting
- Stability-Plasticity balance



Challenges:

- Catastrophic Forgetting
- Stability-Plasticity balance



Continual Learning

Challenges:

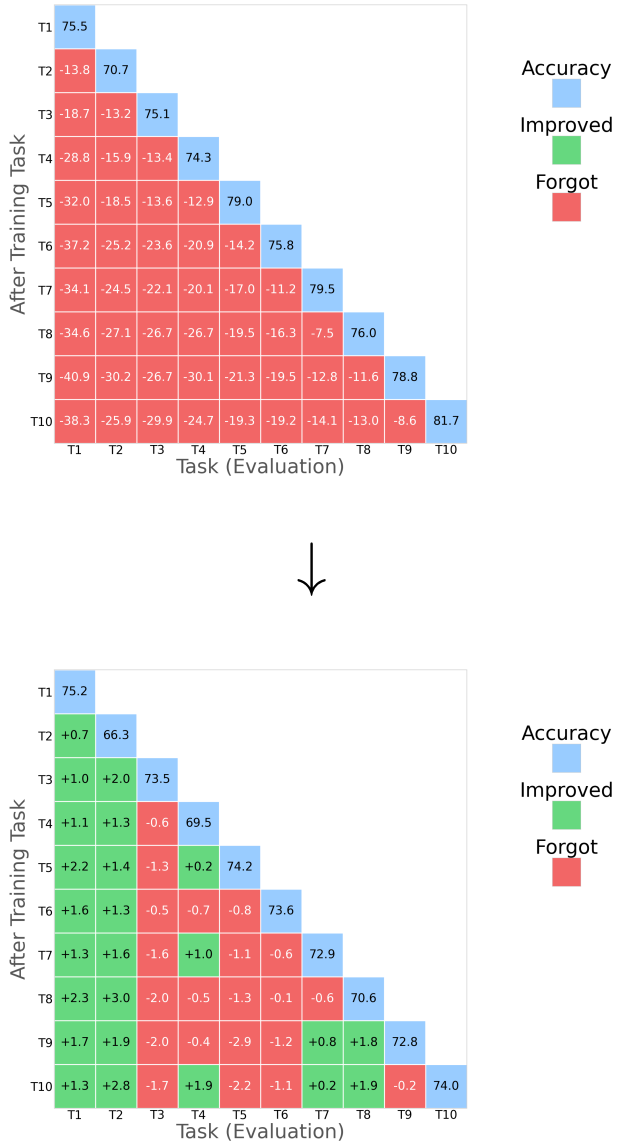
- Catastrophic Forgetting
- Stability-Plasticity balance

Related Works:

- Memory-based, Architecture-based, **Optimization-based**, Representation-based, Bayesian methods,...

Optimization-based methods: orthogonality intuition.

- gradient-based: GPM [Saha et al., 2021], SGP [Saha and Roy, 2023], DFGP [Yang et al., 2023, 2025],... $\rightarrow$ might overly constrain gradient directions
- feature-based $\rightarrow$ more freedom on updates $\rightarrow$ **Ours**



Our Proposed Method (1/5): Notations

For task t in the CL setting, we aim to adapt the model to new data while limiting interference with previously learned tasks.

- 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 $1, 2, \ldots, t-1$.

Our Proposed Method (2/5): Optimization Problem

Stability Constraint (Output Preservation)

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

where ε controls the allowable output perturbation.

Our Proposed Method (2/5): Optimization Problem

Stability Constraint (Output Preservation)

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

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

$$\begin{aligned} \min_{\mathbf{W}_t} \quad & \mathcal{L}_{CE}(D_t, \mathbf{W}_t) \\ \text{s.t.} \quad & \left\| x^\top \Delta \mathbf{W}_t \right\|_2^2 \leq \varepsilon, \quad \forall x \in \mathcal{I}_t. \end{aligned} \tag{3}$$

Eq. (3) serves as a conceptual formulation of the stability–plasticity trade-off.

Our Proposed Method (2/5): Optimization Problem

Stability Constraint (Output Preservation)

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

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

$$\begin{aligned} \min_{\mathbf{W}_t} \quad & \mathcal{L}_{CE}(D_t, \mathbf{W}_t) \\ \text{s.t.} \quad & \left\| x^\top \Delta \mathbf{W}_t \right\|_2^2 \leq \varepsilon, \quad \forall x \in \mathcal{I}_t. \end{aligned} \tag{3}$$

Eq. (3) serves as a conceptual formulation of the stability–plasticity trade-off.

Rather than solving this constrained problem directly, we introduce a structured parameterization of $\Delta \mathbf{W}_t$ that satisfies the stability constraint.

Our Proposed Method (3/5): Stability Subspace Construction

Construction of the Stability Subspace

Using SVD and let

$$I_t = [X_1 : X_2 : \cdots : X_{t-1}] = \tilde{U}_t \Sigma_t \tilde{V}_t^\top \in \mathbb{R}^{d \times \sum_{i=1}^{t-1} N_i}, \tag{4}$$

where the singular values satisfy

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

Note: Restricting updates to directions associated with small singular values reduces forgetting with earlier tasks.

Our Proposed Method (4/5): Weight Updates Construction

Selecting the Small-Singular-Value Subspace.

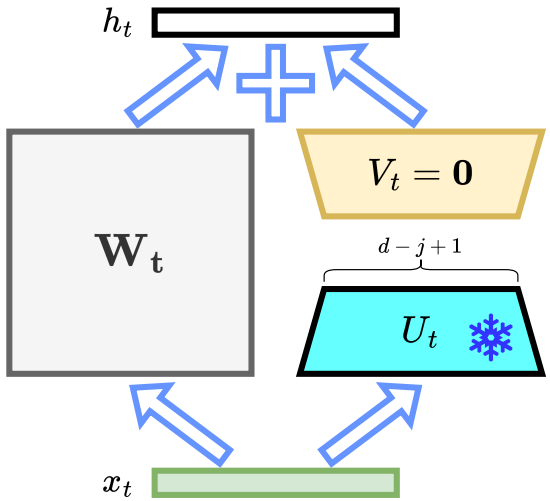
Let $\varepsilon_1 > 0$ be a threshold and define

$$j = \min\{i : \sigma_{t,i} \leq \varepsilon_1 \|I_t\|_F\}. \tag{5}$$

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

$$\Delta W_t = U_t V_t, \tag{6}$$

where U_t is fixed and only V_t is trainable.



LoRA-style weights update

Our Proposed Method (4/5): Weight Updates Construction

Selecting the Small-Singular-Value Subspace.

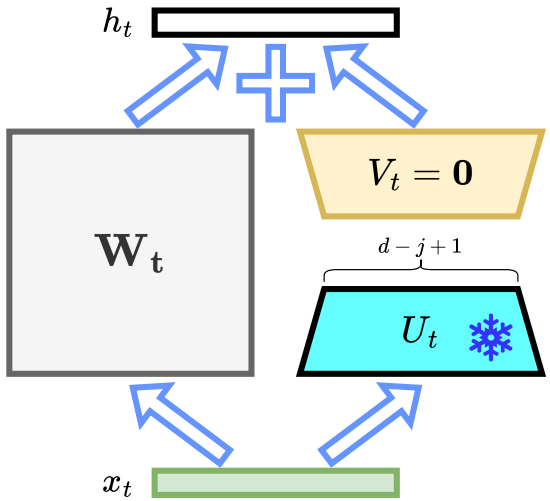
Let $\varepsilon_1 > 0$ be a threshold and define

$$j = \min\{i : \sigma_{t,i} \leq \varepsilon_1 \|I_t\|_F\}. \tag{5}$$

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

$$\Delta W_t = U_t V_t, \tag{6}$$

where U_t is fixed and only V_t is trainable.



LoRA-style weights update

Efficiency Trick

Instead of combined input I_t, use the covariance matrix:

$$C_t = I_t I_t^\top = \sum_{i=1}^{d} x_{t,i} x_{t,i}^\top \tag{7}$$

$\rightarrow$ Only need to save $d \times d$ matrix C_t
$\rightarrow$ Only need to save matrix $\tilde{U}_t$ and Σ_t

Our Proposed Method (5/5): Theoretical Guarantees

Stability Bound

Consider any previous input $x \in \mathcal{I}_t$. For every previous input x,

$$\|x^\top \Delta W_t\|_2 \leq \varepsilon_1 \|I_t\|_F \|V_t\|_2. \tag{8}$$

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

Hence, stability constraint in Eq. (2) is satisfied for every previous input.

Our Proposed Method (5/5): Theoretical Guarantees

Stability Bound

Consider any previous input $x \in \mathcal{I}_t$. For every previous input x,

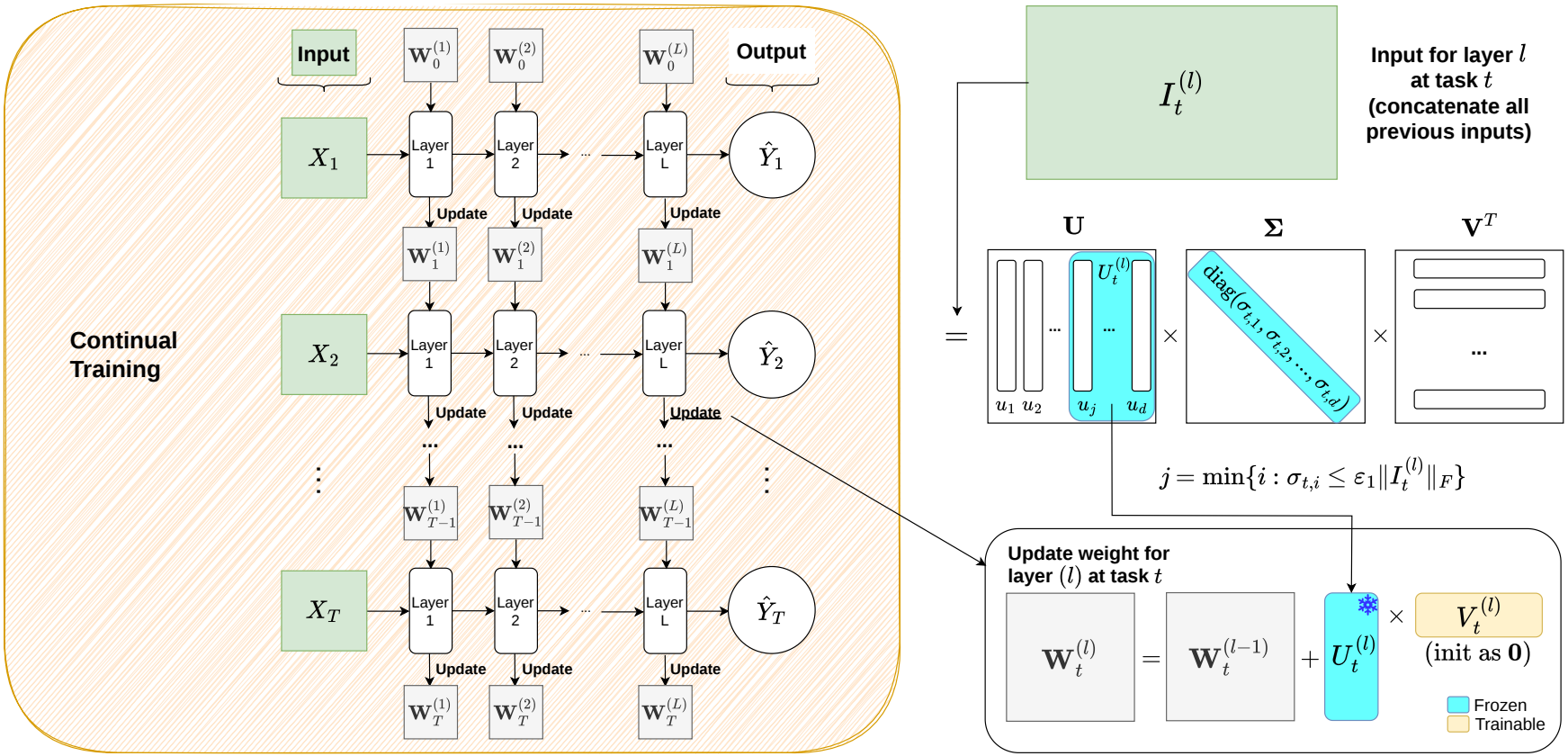
$$\|x^\top \Delta W_t\|_2 \leq \varepsilon_1 \|I_t\|_F \|V_t\|_2. \tag{8}$$

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

Hence, stability constraint in Eq. (2) is satisfied for every previous input.

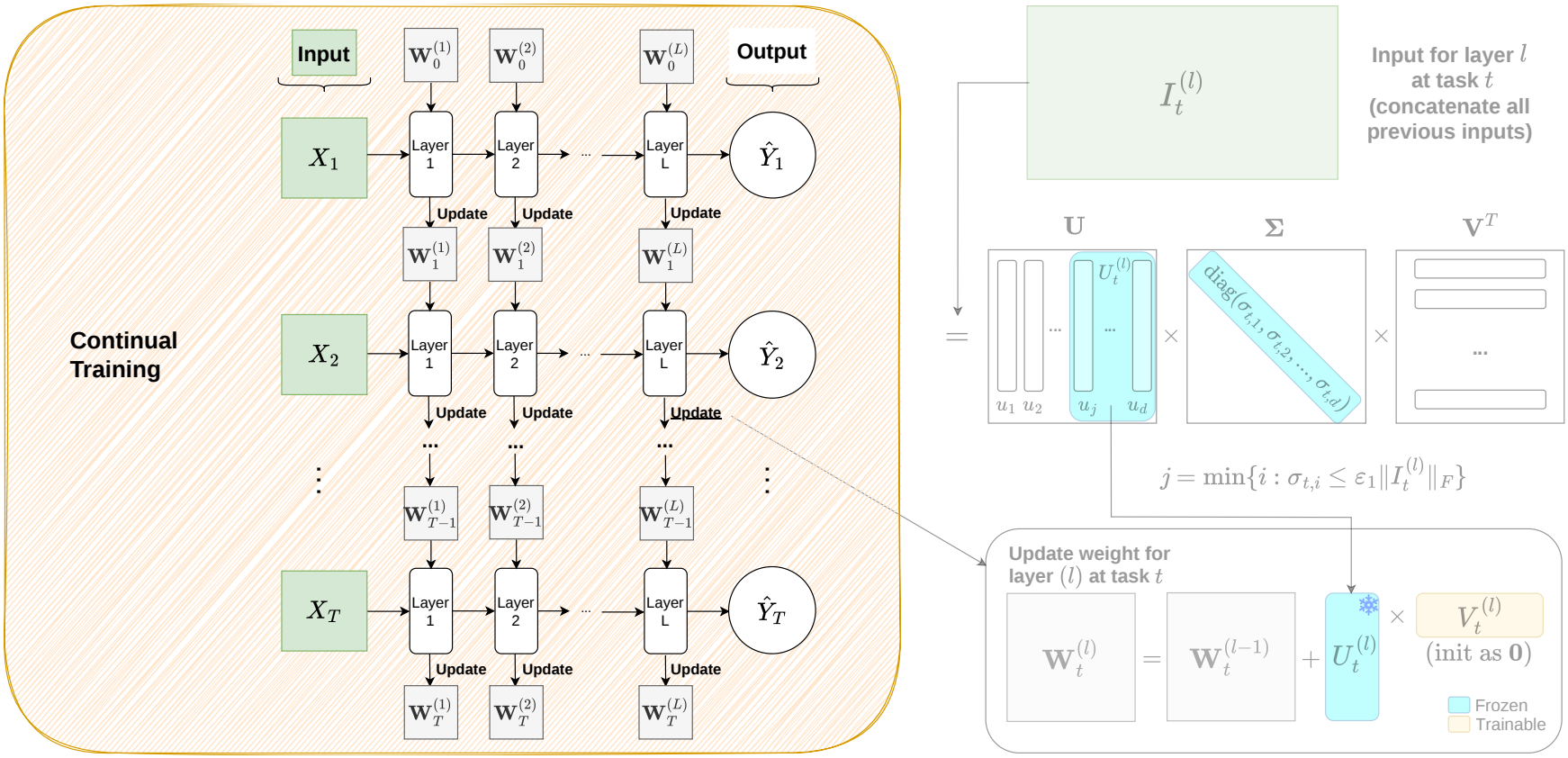
$\rightarrow$ Updates lie in an approximate null subspace of previous inputs, while their magnitudes remain controlled, ensuring bounded interference across tasks.

NESS (1/4): Overview



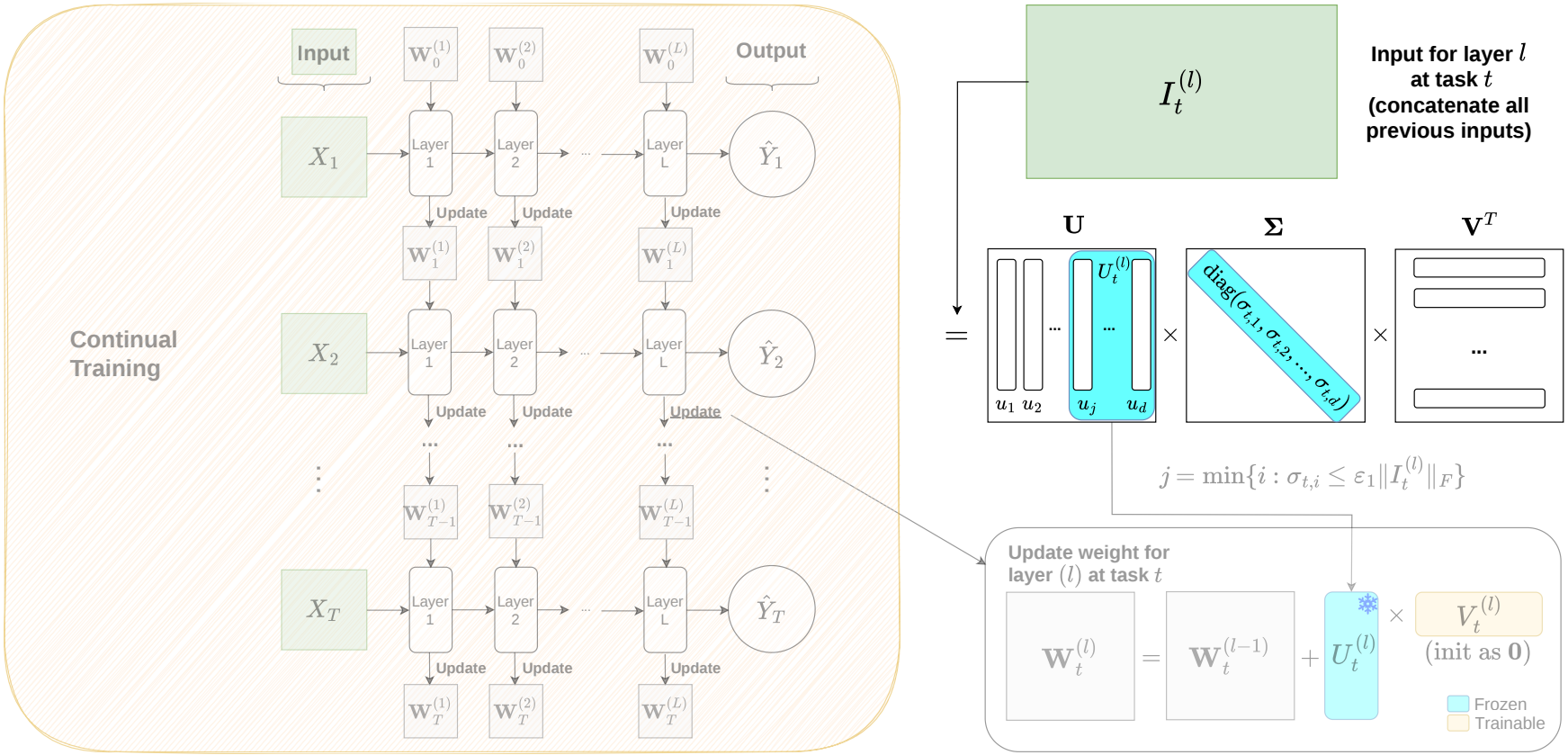
Overview of NESS (Null-space Estimated from Small Singular values)

NESS (2/4): Continual Learning



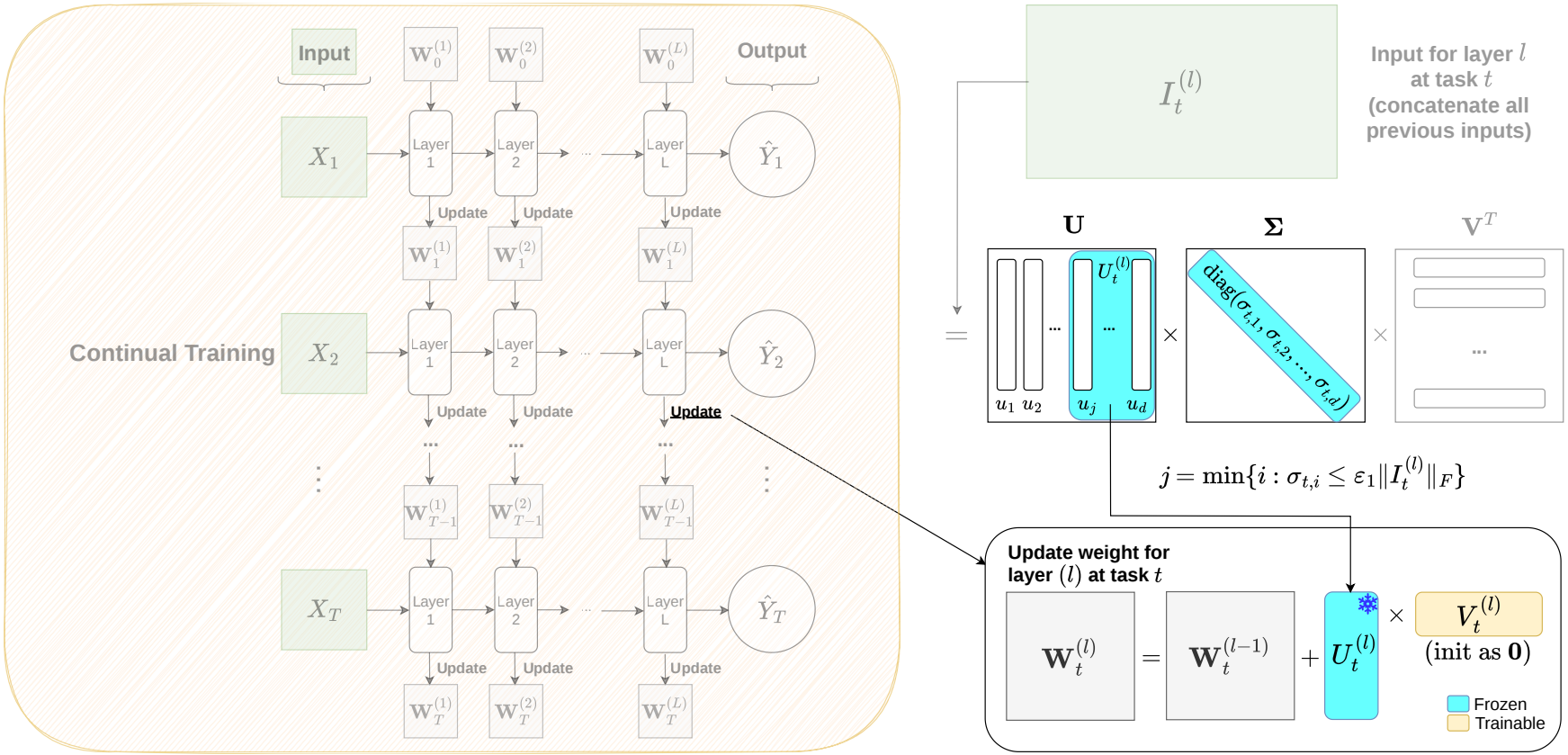
Overview of NESS: Continual Learning

NESS (3/4): SVD



Overview of NESS: SVD

NESS (4/4): LoRA-style Weight Updates



Overview of NESS: LoRA-style weight updates

Experimental Setup

Datasets (with models). CIFAR-100 (10 classes, using AlexNet), 5-datasets[1] (5 classes, using ResNet18), Split-miniImageNet (20 classes, using ResNet18).

Metrics.

- **Average Accuracy (ACC)**: represents the average test accuracy of the model trained on all tasks (defined as $\text{ACC} = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}$).
- **Backward Transfer (BWT)**: measures the forgetting of old tasks (defined as $\text{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}(A_{T,i} - A_{i,i})$).

with T denotes number of tasks and $A_{t,i}$ denotes the accuracy tested on task i after training with task t.

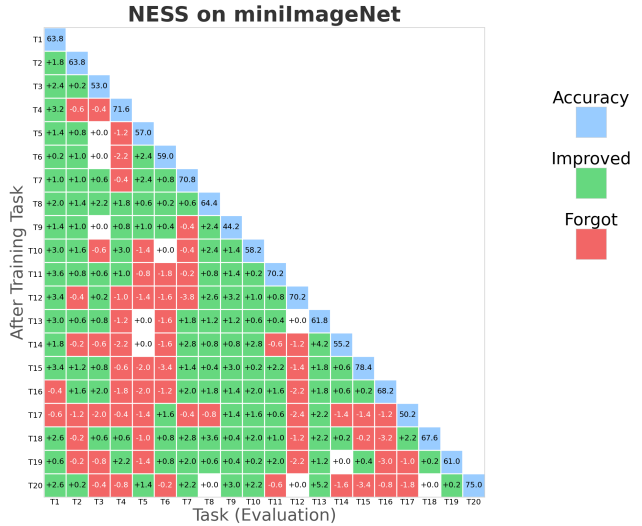
[1] including CIFAR-10, MNIST, SVHN, not-MNIST and FashionMNIST

Results

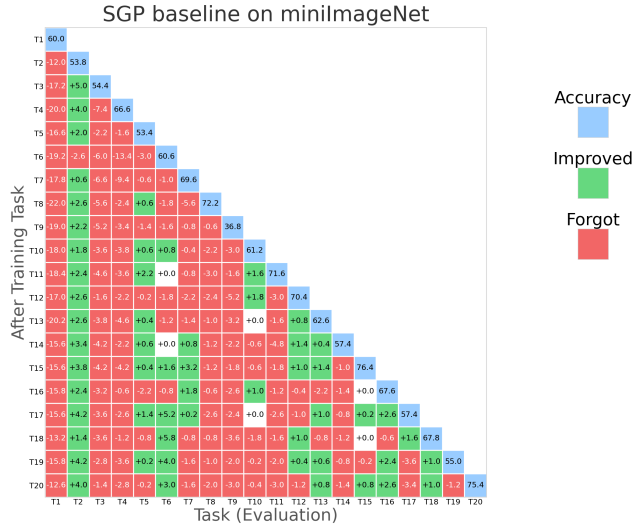
The averaged accuracy (ACC) and **backward transfer (BWT)** over all tasks on different datasets

Method	CIFAR-100 (10 Tasks)		5-datasets (5 Tasks)		MiniImageNet (20 Tasks)	
	ACC(%)	**BWT(%)↑**	ACC(%)	**BWT(%)↑**	ACC(%)	**BWT(%)↑**
OWM‡ [10]	50.94 ± 0.60	-30 ± 1	–	–	–	–
EWC‡ [3]	68.88 ± 0.80	-2 ± 1	88.64 ± 0.26	-4 ± 1	52.01 ± 2.53	-12 ± 3
HAT‡ [7]	72.06 ± 0.50	0 ± 0	91.32 ± 0.18	-1 ± 0	59.78 ± 0.57	-3 ± 0
A-GEM‡ [1]	63.98 ± 1.22	-15 ± 2	84.04 ± 0.33	-12 ± 1	57.24 ± 0.72	-12 ± 1
GPM [6]	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 [5]	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 [4]	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† [2]	74.10 ± 0.09	-3.03 ± 0.31	–	–	–	–
DFGP [8, 9] (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 [8, 9] (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 [8, 9] (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 [8, 9] (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**

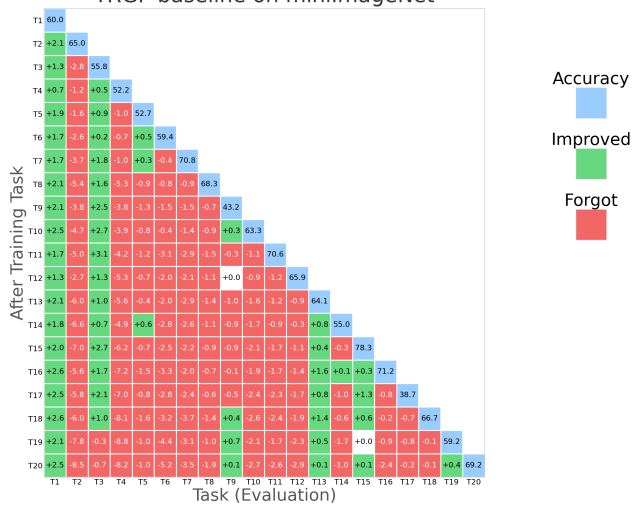
Results



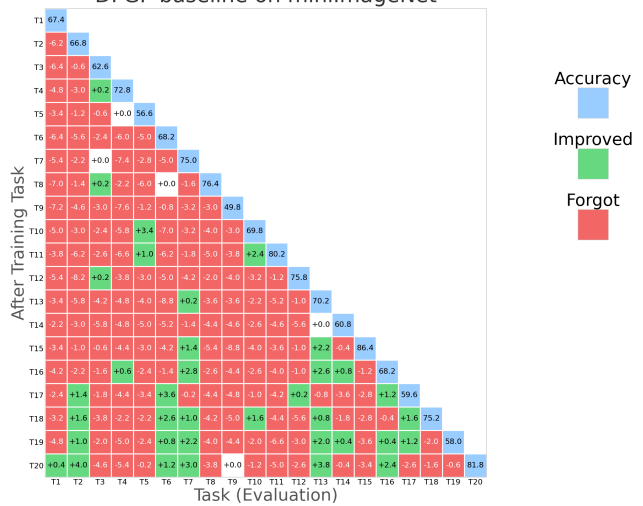
(a) **NESS**



(b) SGP [5]



(c) TRGP [4]



(d) DFGP [8, 9]

Limitations

- Not fully memory-free.
- The threshold ε_1 affects the trade-off between plasticity and stability.
- Omit the biases/batchnorm layers and focus only on linear/convolutional layers.

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 $\rightarrow$ better forgetting rates.

References I

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. *Advances in Neural Information Processing Systems*, 34:18710–18721, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. In *The Tenth International Conference on Learning Representations*, 2022.

Gobinda Saha and Kaushik Roy. Continual learning with scaled gradient projection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9677–9685, 2023.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR, 2018.

Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, and Xingwei Wang. Data augmented flatness-aware gradient projection for continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5630–5639, October 2023.

Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Revisiting flatness-aware optimization in continual learning with orthogonal gradient projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3895–3907, 2025. doi: 10.1109/TPAMI.2025.3539019.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Thank You!

Questions & Discussion

Paper (OpenReview)

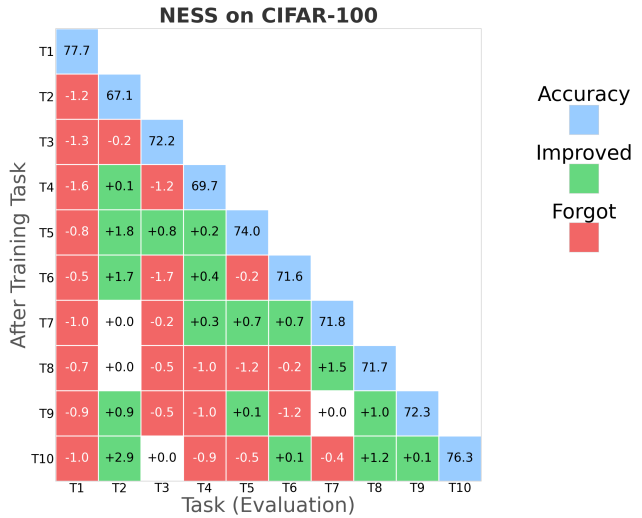
Code

My Website

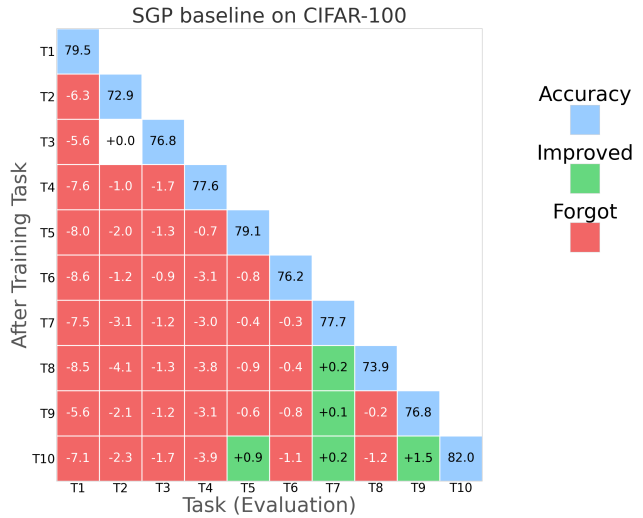
Contact Information:

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

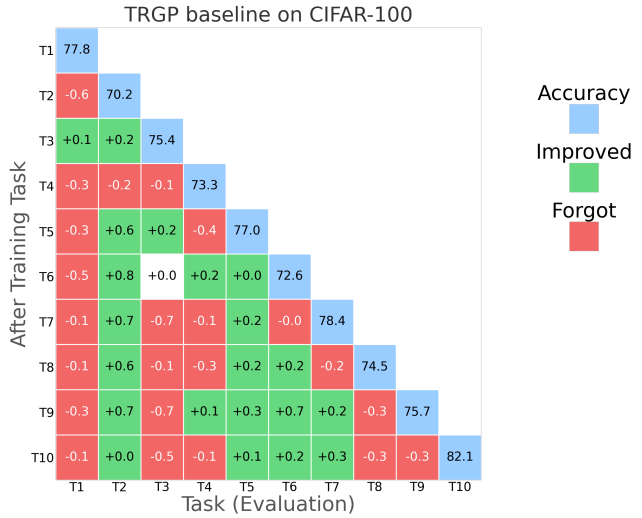
Appendix 1: Additional results (1/3)



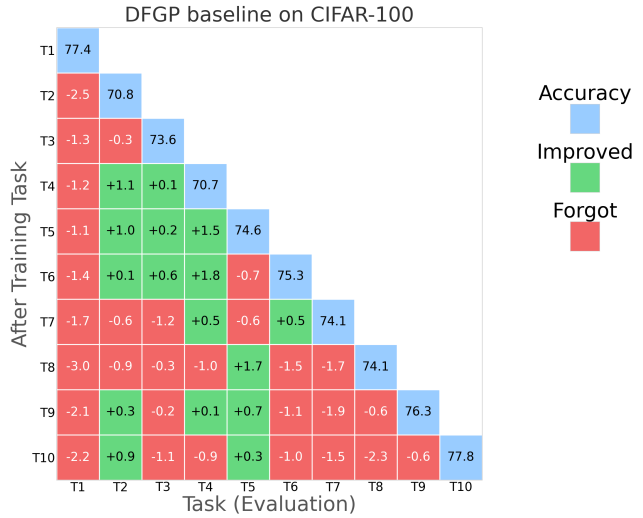
(a) **NESS**



(b) SGP baseline

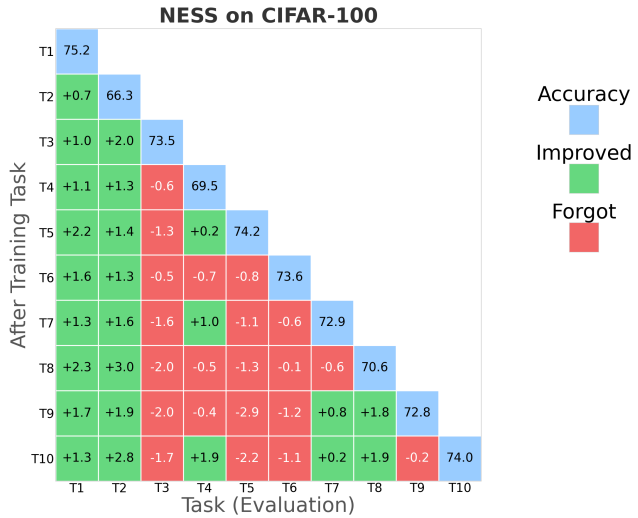


(c) TRGP baseline

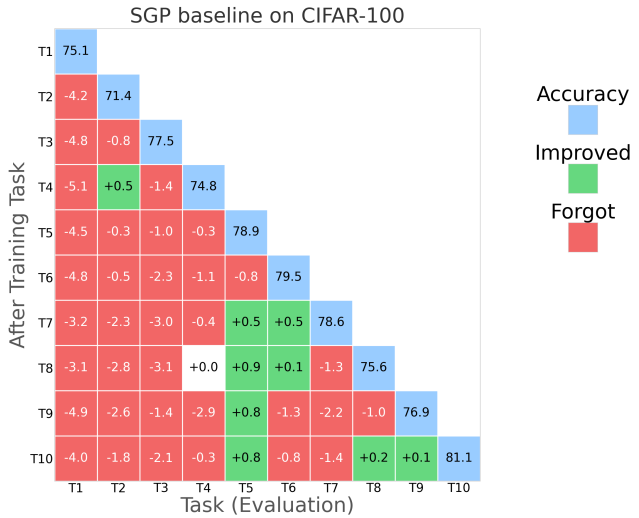


(d) DFGP baseline

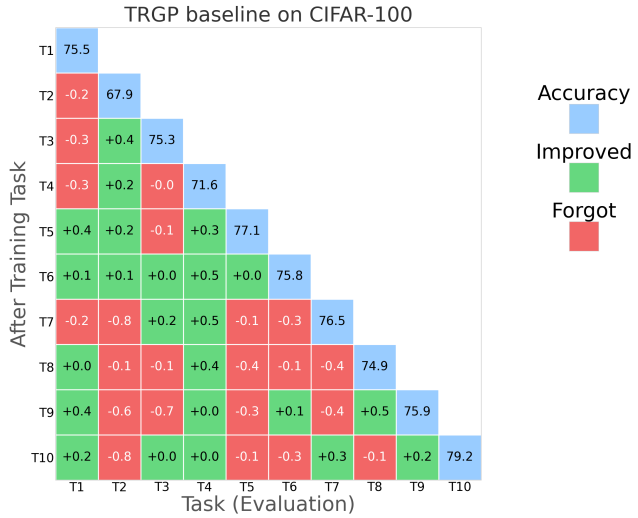
Appendix 1: Additional results (2/3)



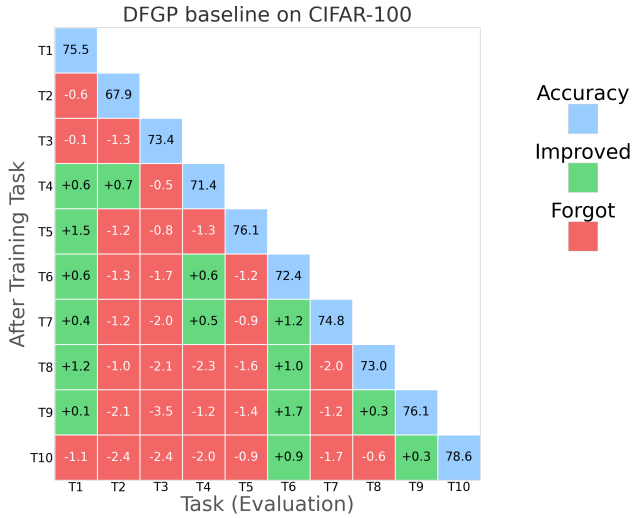
(a) **NESS**



(b) SGP baseline

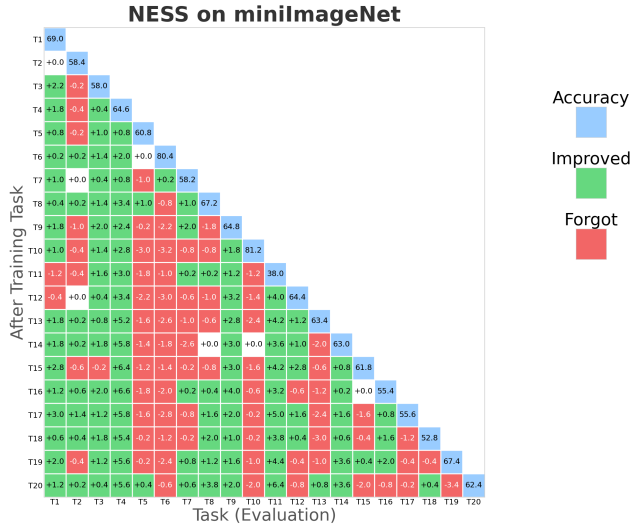


(c) TRGP baseline

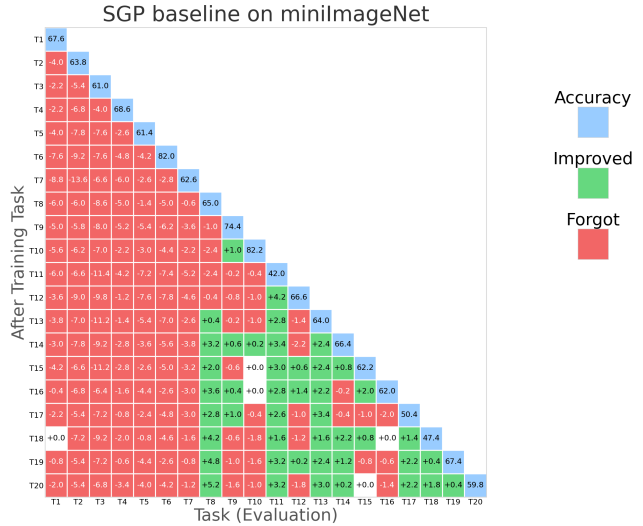


(d) DFGP baseline

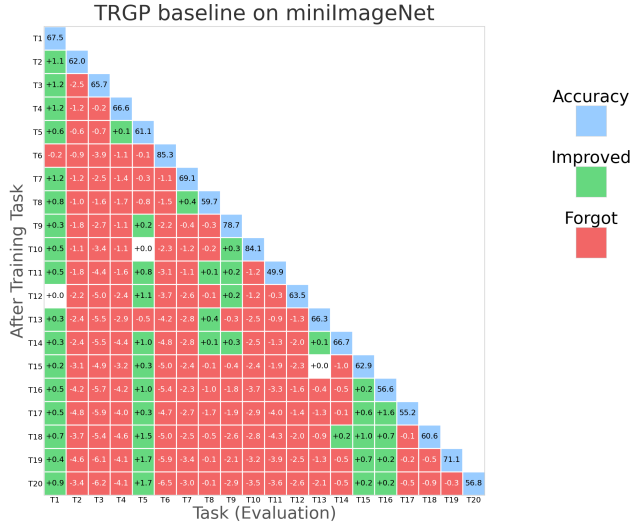
Appendix 1: Additional results (3/3)



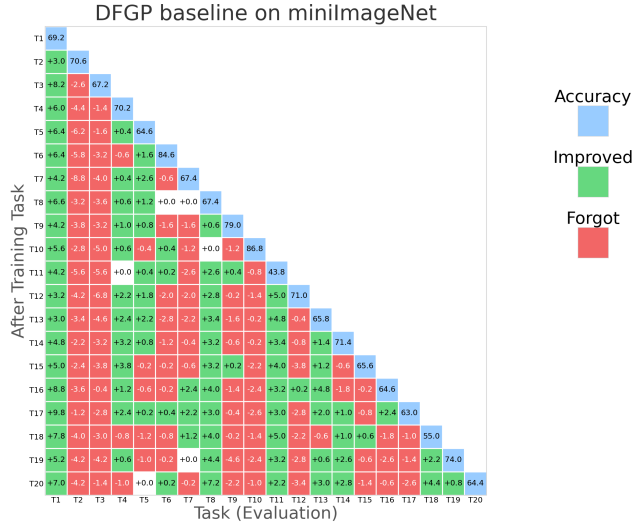
(a) **NESS**



(b) SGP baseline



(c) TRGP baseline



(d) DFGP baseline

Appendix 2: Proof for Explicit Stability Bound (Eq. (2))

Consider any previous input $x \in \mathcal{I}_t$. Since each such x corresponds to a column of I_t, there exists a standard basis vector e_i such that

$$x^\top = e_i^\top I_t^\top.$$

Using Eq. (4), $I_t^\top = \tilde{V}_t \Sigma_t \tilde{U}_t^\top$. Thus, $x^\top \Delta W_t = e_i^\top \tilde{V}_t \Sigma_t \tilde{U}_t^\top U_t V_t$. Because U_t contains only singular vectors corresponding to indices $j, \ldots, d$, $\tilde{U}_t^\top U_t = \begin{bmatrix} 0 \\ I \end{bmatrix}$, so only the small singular values remain active. Denoting the diagonal matrix of these values by Σ_{small}, we obtain $x^\top \Delta W_t = e_i^\top \tilde{V}_t \Sigma_{\text{small}} V_t$. Taking norms, $\|x^\top \Delta W_t\|_2 \leq \|\Sigma_{\text{small}}\|_2 \|V_t\|_2$ (notice that $\tilde{V}_t$ is orthogonal), with $\|\Sigma_{\text{small}}\|_2 \leq \varepsilon_1 \|I_t\|_F$ (by construction). Therefore, for every previous input x,

$$\|x^\top \Delta W_t\|_2 \leq \varepsilon_1 \|I_t\|_F \|V_t\|_2. \tag{9}$$

If we enforce

$$\|V_t\|_2 \leq \frac{\sqrt{\varepsilon}}{\varepsilon_1 \|I_t\|_F}, \tag{10}$$

then $\|x^\top \Delta W_t\|_2^2 \leq \varepsilon, \quad \forall x \in \mathcal{I}_t$. Hence, the stability constraint in Eq. (2) is satisfied for every previous input.