Rethinking Gradient Projection Continual Learning: Stability / Plasticity Feature Space Decoupling

Zhen Zhao, Zhizhong Zhang, Xin Tan, Jun Liu, Yanyun Qu, Yuan Xie, Lizhuang Ma

TUE-AM-354

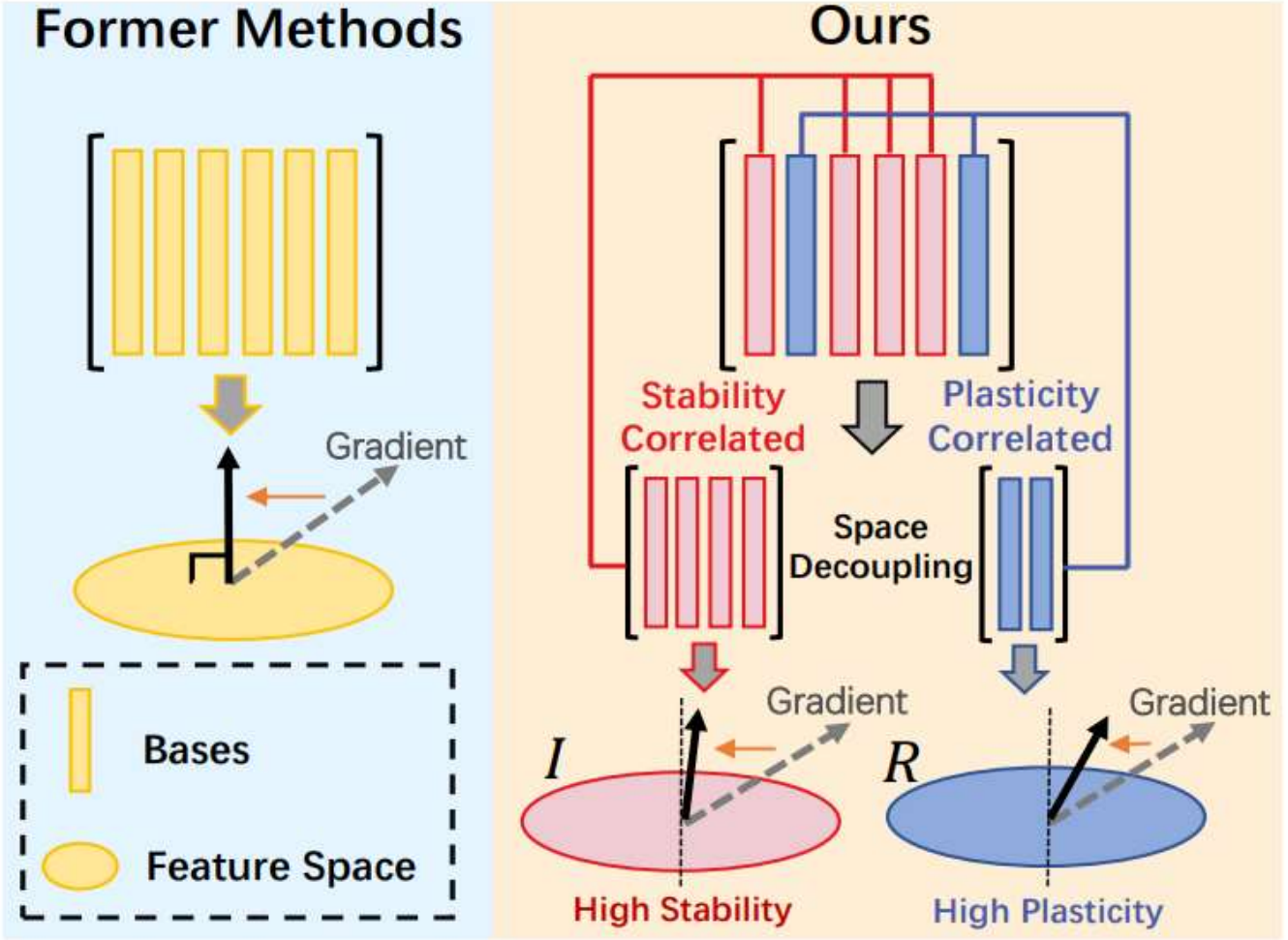
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Quick Preview

decouple feature space to stability-correlated and plasticity-correlated

stability-correlated → strict constraint

plasticity-correlated → loose constraint



Outline

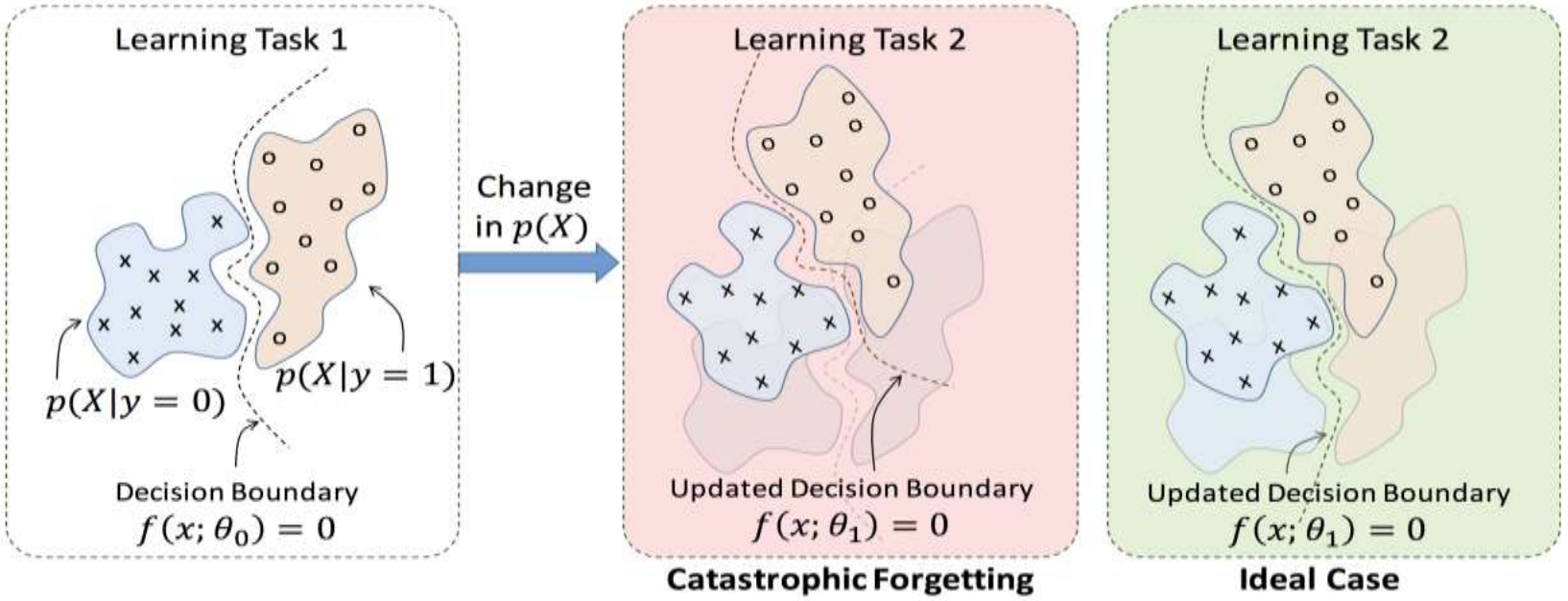
I. Continual Learning

II. Gradient Projection

III. Feature Space Decoupling

IV. Experiments

Continual Learning



- **Stability**:

  preserve old knowledge

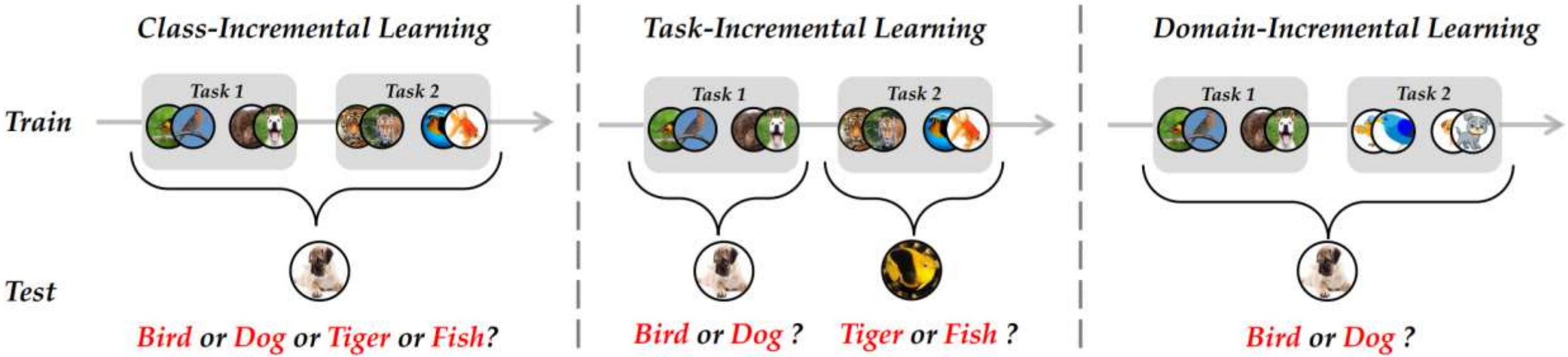
- **Plasticity**:

  learn novel concepts

- Continual Learning：continually learn novel tasks and avoid forgetting

- Key Problem：Catastrophic Forgetting、Stability-Plasticity Dilemma

Continual Learning



Class-Incremental **Task-Incremental** **Domain-Incremental**

Outline

I. Continual Learning

II. Gradient Projection

III. Feature Space Decoupling

IV. Experiments

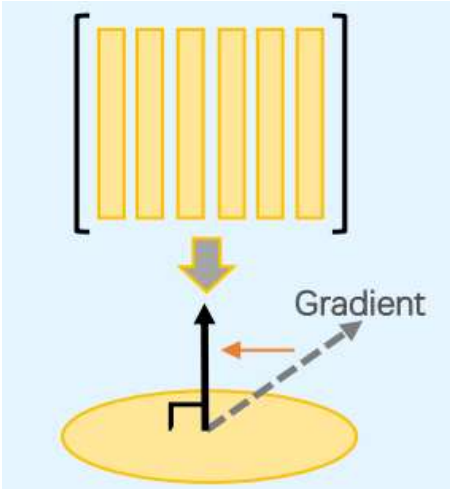
Gradient Projection

Project gradient update into the orthogonal direction of the input space

- **Orthogonal-based:**

  OWM / GPM / TRGP / etc

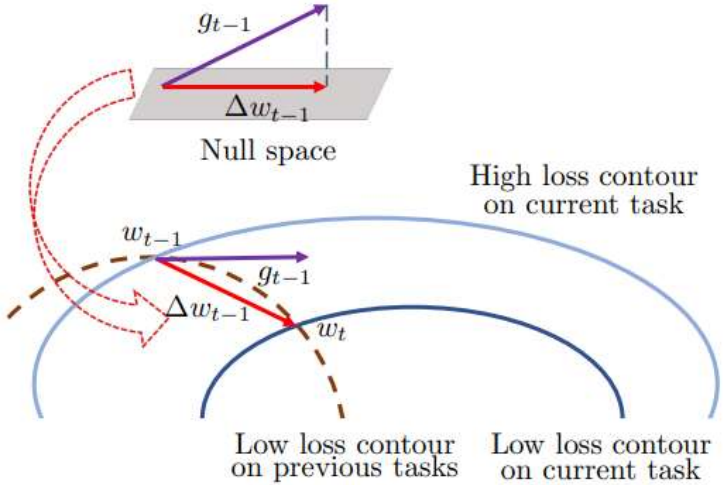
  (NMI'19) (ICLR'21) (ICLR'22)



- **Null space-based:**

  Adam-NSCL / AdNS / etc

  (CVPR'21) (ECCV'22)



Gradient Projection

Project gradient update into the orthogonal direction of the input space

$$\boldsymbol{R}_1^l = [(\boldsymbol{X}_{1,1}^l)^T, (\boldsymbol{X}_{2,1}^l)^T, ..., (\bar{\boldsymbol{X}}_{n_s,1}^l)^T]$$

$$\boldsymbol{R}_1^l = \boldsymbol{U}_1^l \boldsymbol{\Sigma}_1^l (\boldsymbol{V}_1^l)^T$$

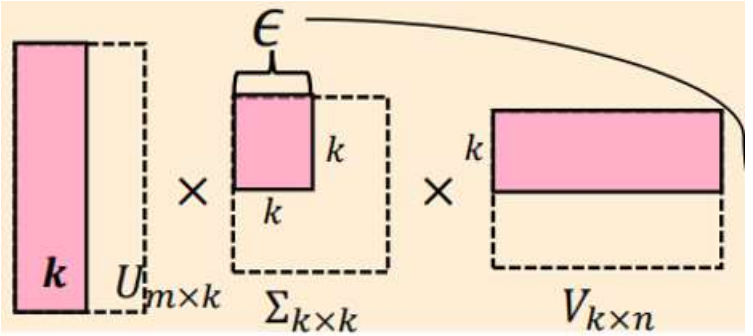
$$||(\boldsymbol{R}_1^l)_k||_F^2 \geq \epsilon_{th}^l ||\boldsymbol{R}_1^l||_F^2.$$

$$\boldsymbol{M}^l = [\boldsymbol{u}_{1,1}^l, \boldsymbol{u}_{2,1}^l, ..., \boldsymbol{u}_{k,1}^l]$$

$$\nabla_{\boldsymbol{W}_2^l} L_2 = \nabla_{\boldsymbol{W}_2^l} L_2 - \boldsymbol{M}^l (\boldsymbol{M}^l)^T (\nabla_{\boldsymbol{W}_2^l} L_2)$$

Avoid Forgetting

$$\begin{aligned} \boldsymbol{\theta}_t^l x_{j,i}^l &= (\boldsymbol{\theta}_{t-1}^l + \Delta\boldsymbol{\theta}_{t-1}^l) x_{j,i}^l \\ &= \boldsymbol{\theta}_{t-1}^l x_{j,i}^l + \Delta\boldsymbol{\theta}_{t-1}^l x_{j,i}^l \\ &= \boldsymbol{\theta}_{t-1}^l x_{j,i}^l. \end{aligned}$$

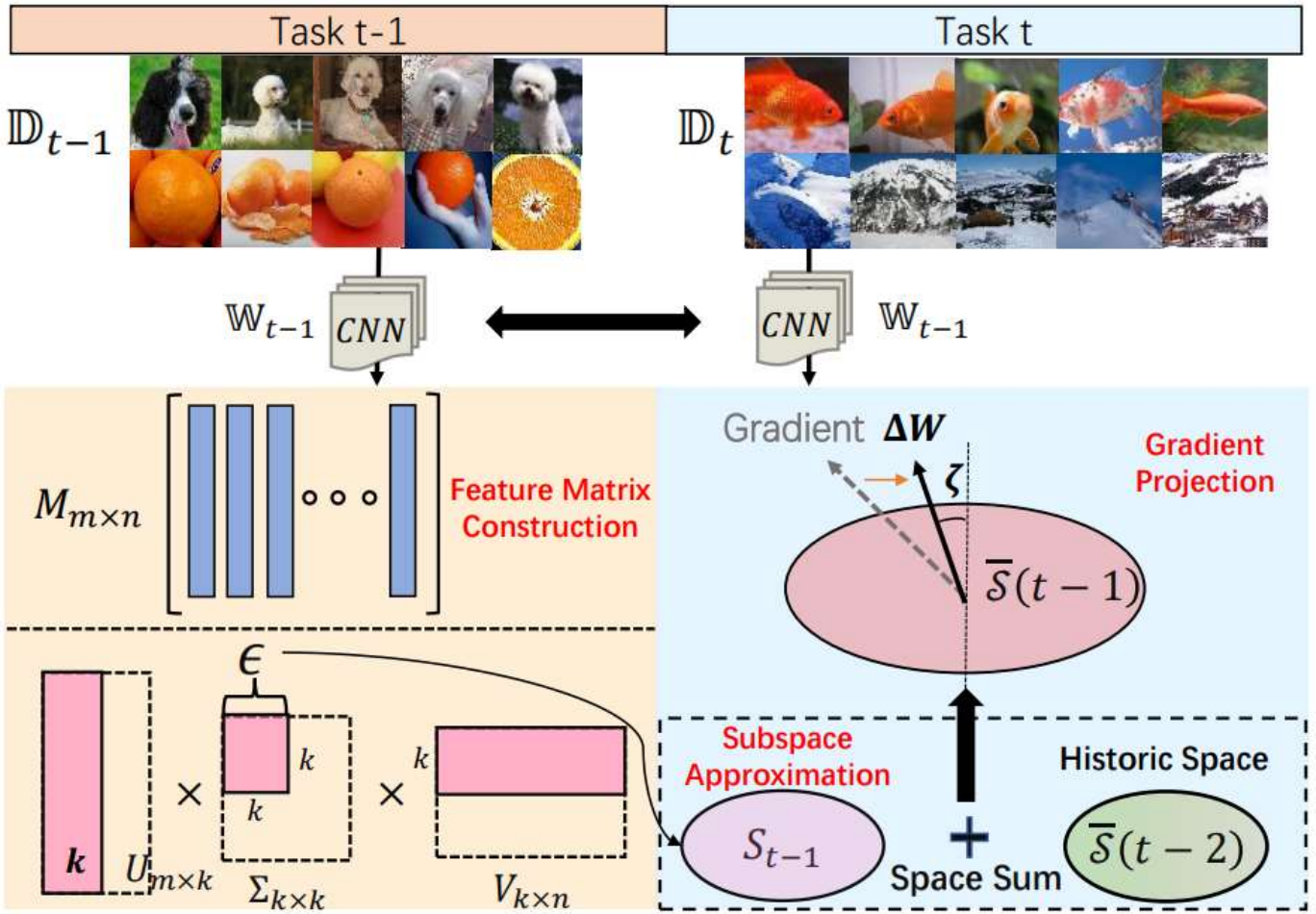


Gradient Projection

- OWM
- GPM
- TRGP
- Adam-NSCL
- AdNS

……

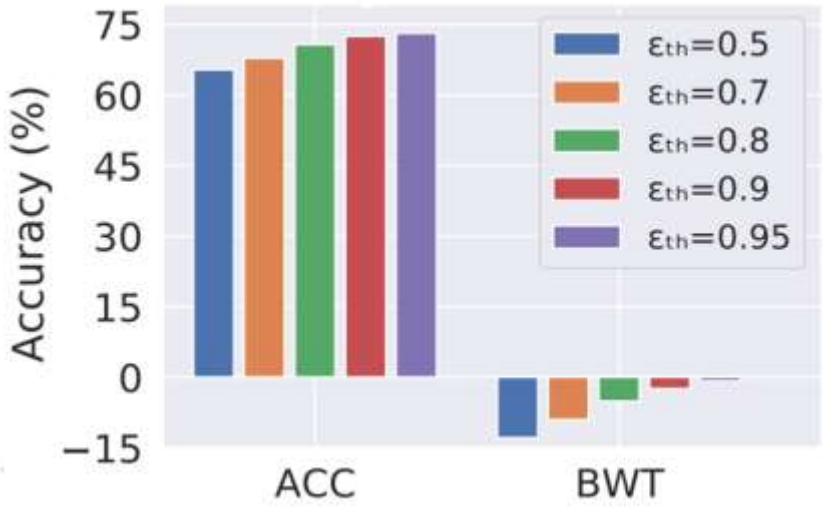
Feature Space Continual Learning Paradigm



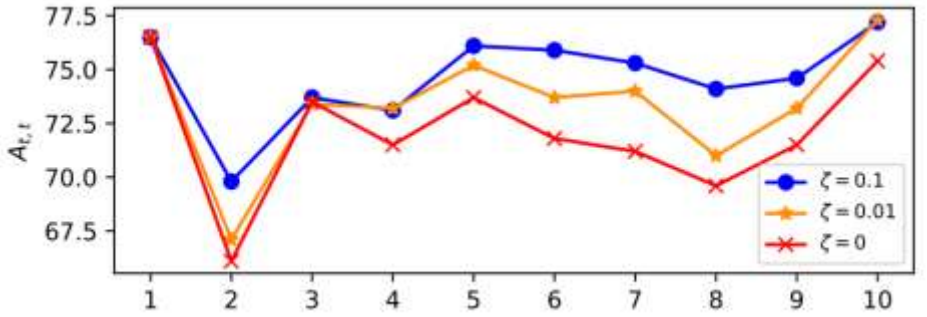
Gradient Projection

Stability–Plasticity Dilemma in Gradient Projection

$\varepsilon \uparrow$, **stability** $\uparrow$, **plasticity** $\downarrow$



$\zeta \uparrow$, **stability** $\downarrow$, **plasticity** $\uparrow$



Outline

I. Continual Learning

II. Gradient Projection

III. Feature Space Decoupling

IV. Experiments

Feature Space Decoupling

Motivation from GPM: GPM directly delete common bases when updating the feature space

Task-Shared bases is more important than Task-Specific bases

At the end of the task 2 training, we update the GPM with new task-specific bases (of CGS). To obtain such bases, we construct $\boldsymbol{R}_2^l = [\boldsymbol{x}_{1,2}^l, \boldsymbol{x}_{2,2}^l, ..., \boldsymbol{x}_{n_s,2}^l]$ using data from task 2 only. However, before performing SVD and subsequent k-rank approximation, from $\boldsymbol{R}_2^l$ we eliminate the common directions (bases) that are already present in the GPM so that newly added bases are unique and orthogonal to the existing bases in the memory. To do so, we perform the following step :

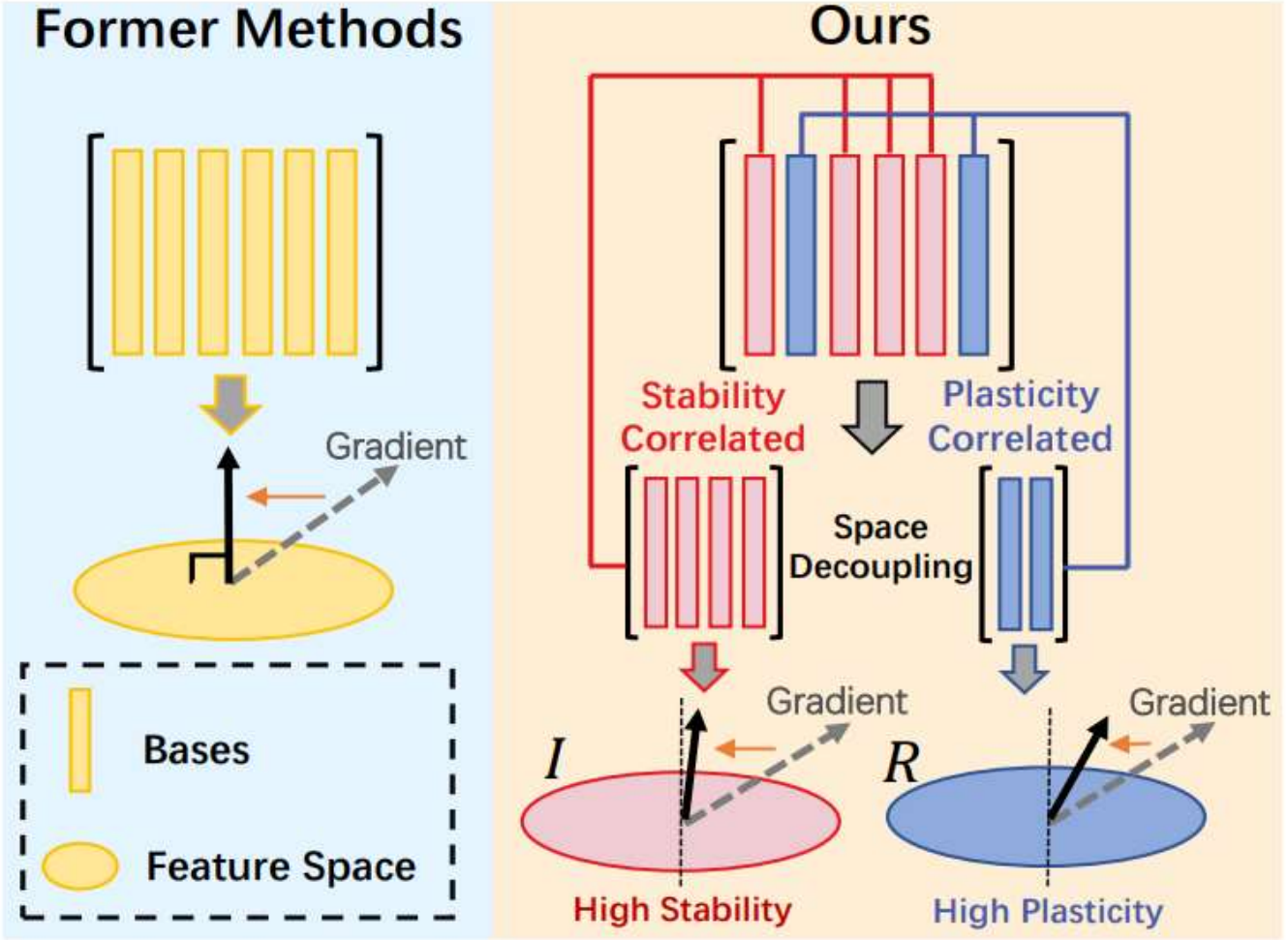
$$\hat{\boldsymbol{R}}_2^l = \boldsymbol{R}_2^l - \boldsymbol{M}^l(\boldsymbol{M}^l)^T(\boldsymbol{R}_2^l) = \boldsymbol{R}_2^l - \boldsymbol{R}_{2,Proj}^l. \tag{8}$$

Feature Space Decoupling

decouple feature space to stability-correlated and plasticity-correlated

stability-correlated → strict constraint

plasticity-correlated → loose constraint



Feature Space Decoupling

Subspace intersection and sum

$$\mathcal{P} \cap \mathcal{Q} = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \in \mathcal{P}, \boldsymbol{\alpha} \in \mathcal{Q}\}$$

$$\mathcal{P} + \mathcal{Q} = \{\boldsymbol{\alpha} + \boldsymbol{\beta} | \boldsymbol{\alpha} \in \mathcal{P}, \boldsymbol{\beta} \in \mathcal{Q}\}.$$

Stability subspace:

$$\mathcal{I}(t) = \sum_{1<i\leq t} \bar{\mathcal{S}}(i-1) \cap \mathcal{S}_i$$

(sum of intersection) → Task-Shared

Plasticity subspace:

$$\boldsymbol{R}(t) = \bar{\boldsymbol{S}}(t) - \mathrm{Proj}_{\mathcal{I}(t)}(\bar{\boldsymbol{S}}(t))$$

(orthogonal complementary of I) → Task-Specific

Feature Space Decoupling

Experimental evidence

Influence Score of gradient update:

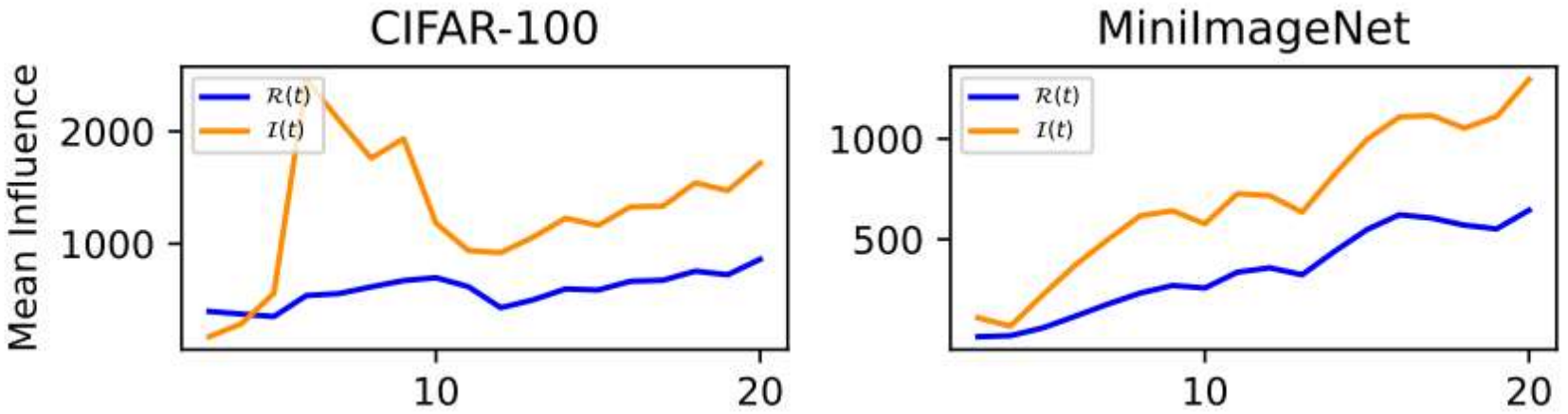
its interference with feature space

$$\omega(\boldsymbol{g}) = \sum_{j=1}^{t} \left\| \mathrm{Proj}_{\mathcal{S}_j}(\boldsymbol{g}) \right\|_F^2$$

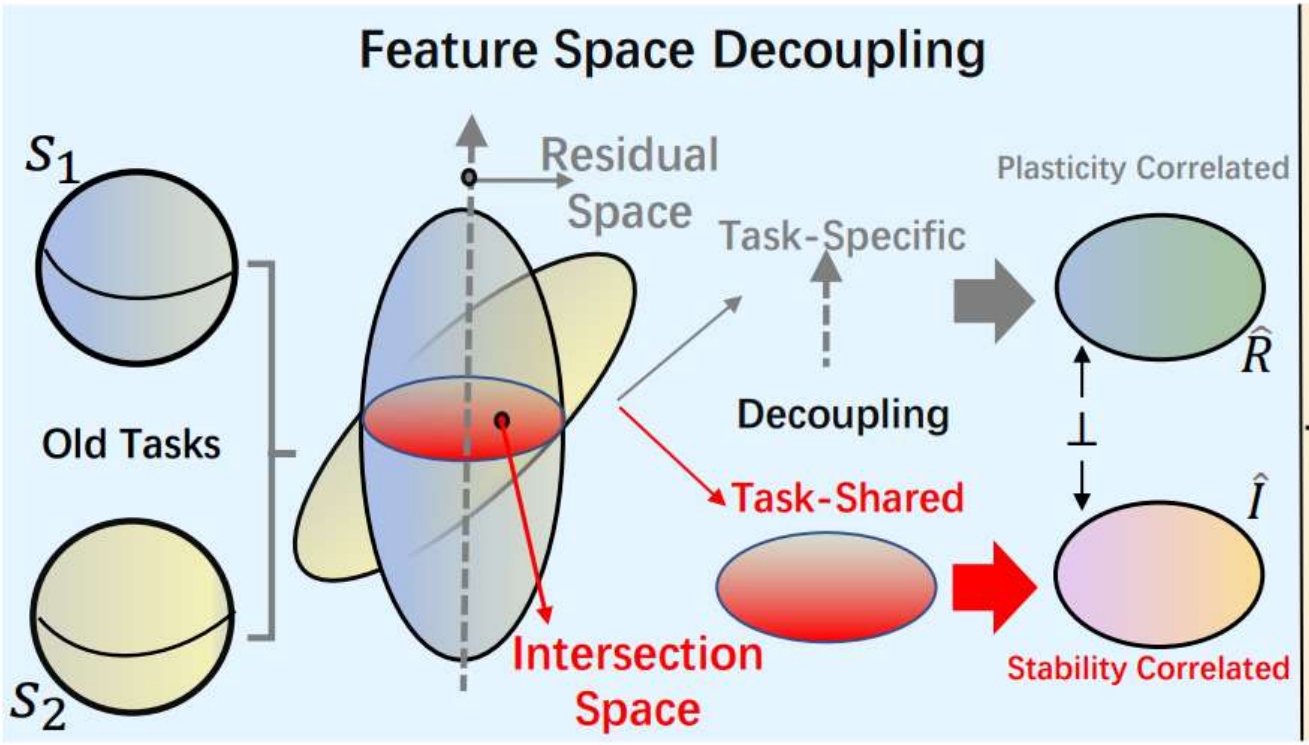
Influence Score of feature subspace:

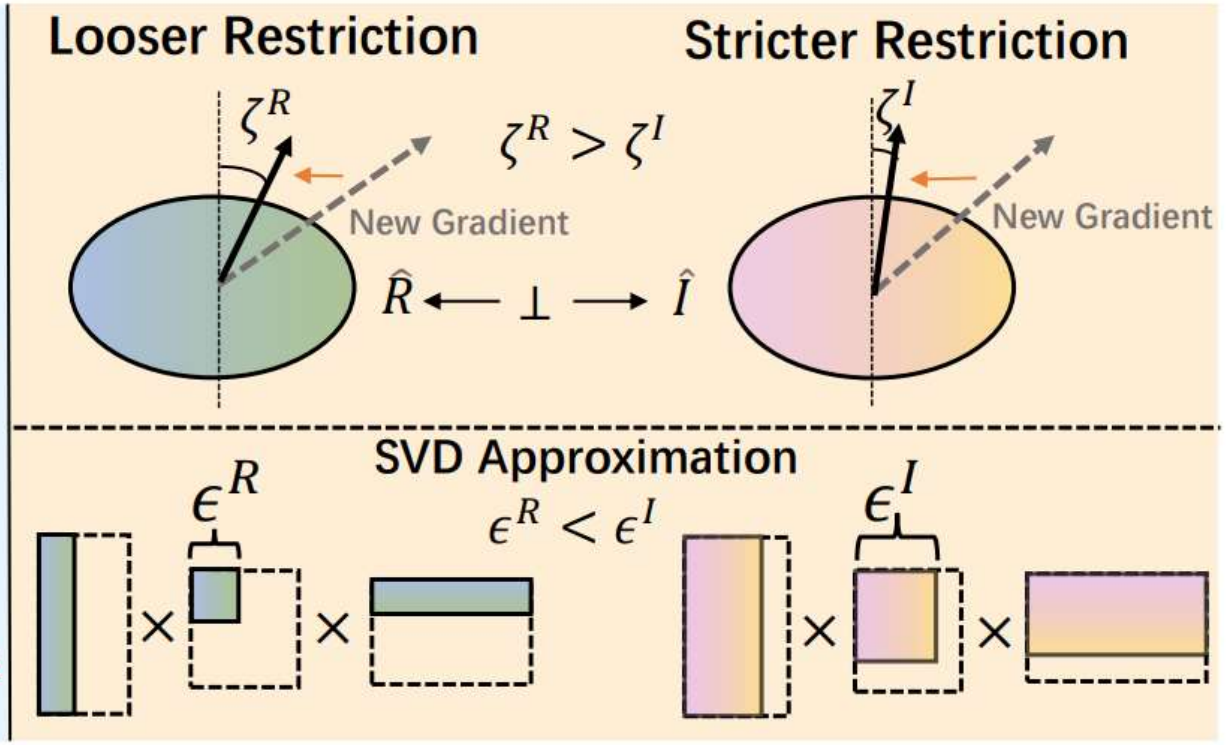
sum of Influence Score of gradient component within this subspace

$$\Omega(\mathcal{I}(t)) = \frac{\sum_i \omega(\boldsymbol{g}_i^{\mathcal{I}})}{dim(\mathcal{I}(t))} \qquad \Omega(\mathcal{R}(t)) = \frac{\sum_i \omega(\boldsymbol{g}_i^{\mathcal{R}})}{dim(\mathcal{R}(t))}$$



Feature Space Decoupling





$$\hat{\boldsymbol{I}}(t) = \mathcal{A}(\boldsymbol{I}(t); \epsilon^{\mathcal{I}})$$

$$\hat{\boldsymbol{R}}(t) = \mathcal{A}(\boldsymbol{R}(t); \epsilon^{\mathcal{R}})$$

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{t+1} =& \nabla_{\boldsymbol{\theta}}\mathcal{L}_{t+1} \\ &- \nabla_{\boldsymbol{\theta}}\mathcal{L}_{t+1}(1-\zeta^{\mathcal{I}})\hat{\boldsymbol{I}}(t)(\hat{\boldsymbol{I}}(t))^T \\ &- \nabla_{\boldsymbol{\theta}}\mathcal{L}_{t+1}(1-\zeta^{\mathcal{R}})\hat{\boldsymbol{R}}(t)(\hat{\boldsymbol{R}}(t))^T\end{aligned}$$

Outline

I. Continual Learning

II. Gradient Projection

III. Feature Space Decoupling

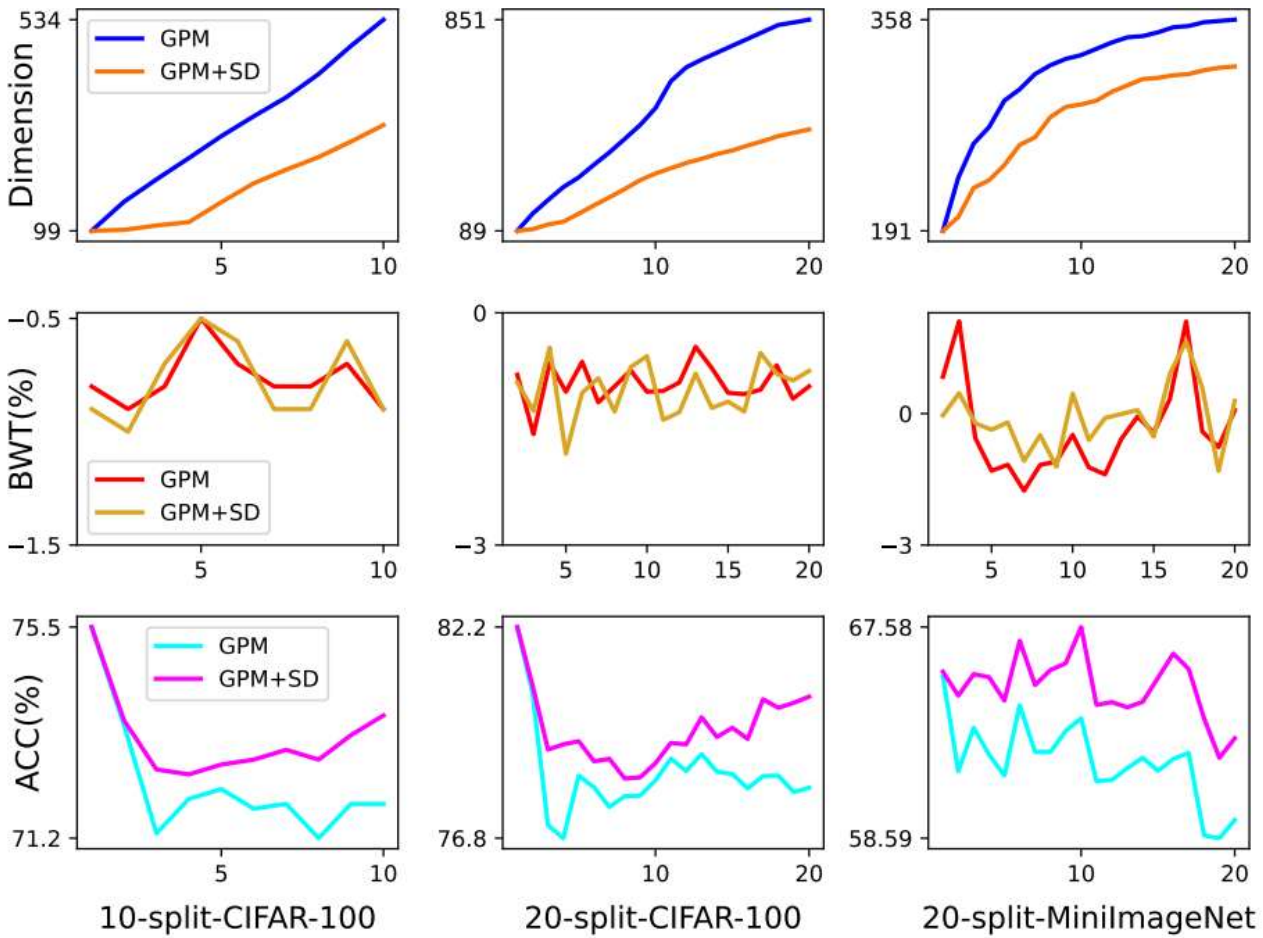
IV. Experiments

Experiments

Model	Venue	20-split-MiniImageNet		20-split-CIFAR-100		10-split-CIFAR-100	
		ACC(%)	BWT(%)	ACC(%)	BWT(%)	ACC(%)	BWT(%)
LWF [22]	PAMI'17	57.63	-8.72	74.38	-9.11	70.7	-6.27
EWC [17]	PANS'17	52.01	-12	71.66	-3.72	70.77	-2.83
MAS [2]	ECCV'18	50.12	-5.82	63.84	-6.29	66.93	-4.03
MUC-MAS [24]	ECCV'20	46.24	-3.79	67.22	-5.72	63.73	-3.38
GEM [25]	NIPS'17	-	-	68.89	-1.2	49.48	2.77
A-GEM [5]	ICLR'18	57.24	-12	61.91	-6.88	49.57	-1.13
*AdNS [18]	ECCV'22	60.82	-4.24	77.33	-3.25	**77.21**	-2.32
OWM [41]	NMI'19	47.48	-8.57	68.47	-3.37	68.89	-1.88
GPM [31]	ICLR'21	60.41± 0.61	-0.7±0.4	77.53±0.83	-0.97±0.59	72.48±0.4	-0.9±0
Adam-NSCL [38]	CVPR'21	59.07±1.1	-4.9±1.32	75.81±0.93	-3.98±0.85	74.97±1.15	-2.64±0.91
TRGP [23]	ICLR'22	**63.51±0.74**	-0.76±0.25	**80.68±0.7**	-0.87±0.46	74.46±0.32	-0.9±0.01
GPM+SD		62.39±0.56	-0.61±0.11	80.71±0.82	-0.73±0.27	73.53±0.44	-0.83±0.31
Adam-NSCL+SD		60.38±0.75	-4.81±1	76.5±1.02	-3.99±0.96	**75.97±0.66**	-2.88±0.89
TRGP+SD		**65.8±0.16**	-0.49±0.08	**83.84±0.12**	-0.72±0.2	75.5±0.35	-0.96±0.09

Main Results

Experiments

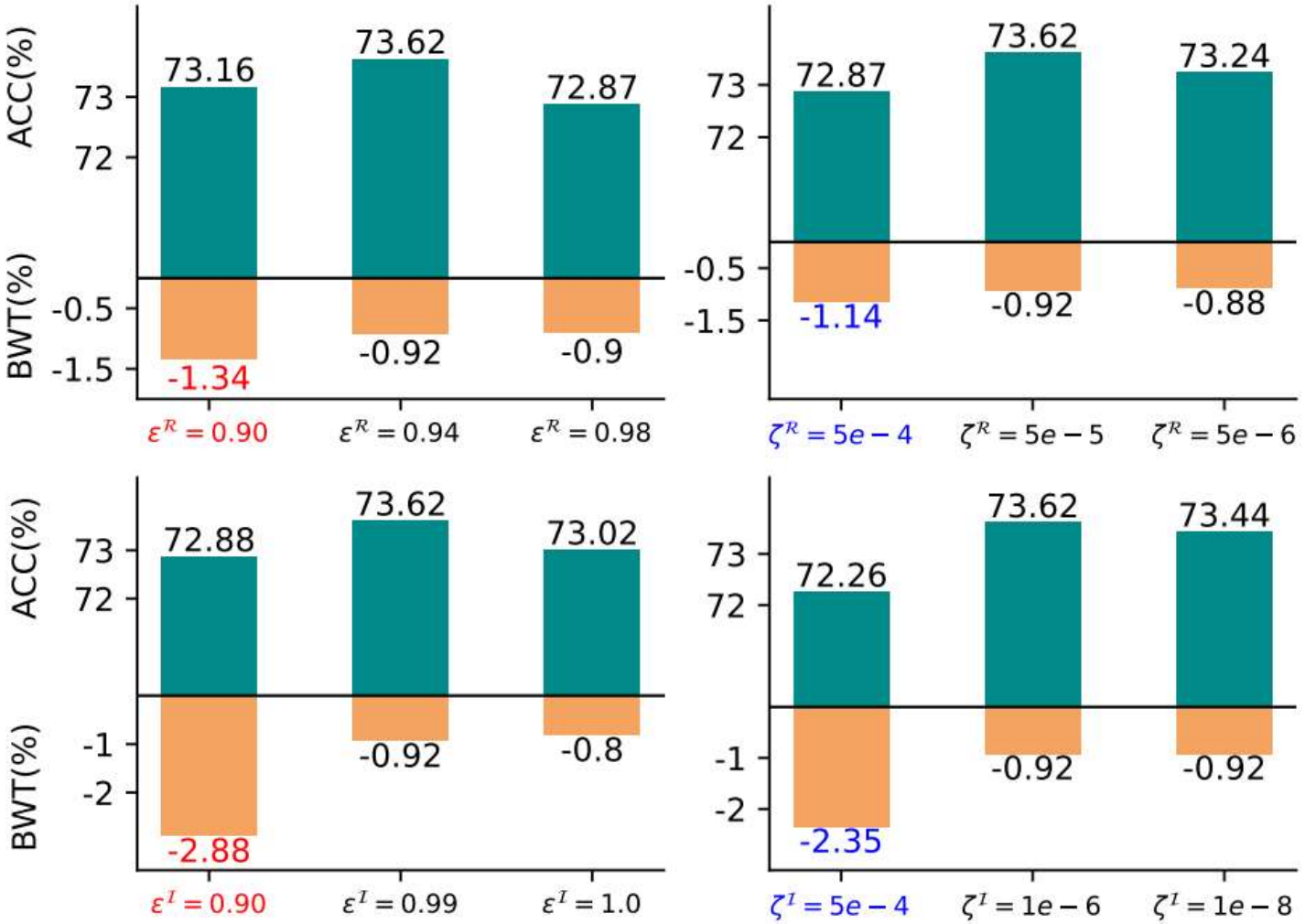


Dimension, BWT and ACC

Experiments

Stability-Plasticity

Comparison of BWT when same constraints are put on I and R



Experiments

Datasets	Methods					
	GPM	GPM+SD	TRGP	TRGP+SD	Adam-NSCL	Adam-NSCL+SD
10-split-CIFAR-100	0.25	0.27	0.41	0.45	2.71	2.93
20-split-CIFAR-100	0.31	0.36	0.48	0.53	4.14	4.53
20-split-MiniImageNet	0.58	0.66	0.81	0.92	11.28	12.95

Computational complexity

Citation

Image Citation

1. Kolouri S, Ketz N, Zou X, et al. Attention-based selective plasticity[J]. arXiv preprint arXiv:1903.06070, 2019.
2. Zhou D W, Wang Q W, Qi Z H, et al. Deep class-incremental learning: A survey[J]. arXiv preprint arXiv:2302.03648, 2023.
3. Wang S, Li X, Sun J, et al. Training networks in null space of feature covariance for continual learning[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2021: 184-193.
4. Saha G, Garg I, Roy K. Gradient projection memory for continual learning[J]. arXiv preprint arXiv:2103.09762, 2021.

Thanks !