CPR: Classifier-Projection Regularization for Continual Learning

Sungmin Cha[1], Hsiang Hsu [2], Taebaek Hwang [1], Flavio P. Calmon [2] and Taesup Moon [3]

[1]Sungkyunkwan University, [2]Harvard University, [3]Seoul National University

csm9493@skku.edu, hsianghsu@g.harvard.edu, qxq9160@gmail.com, flavio@seas.Harvard.edu, tsmoon@snu.ac.kr

Introduction

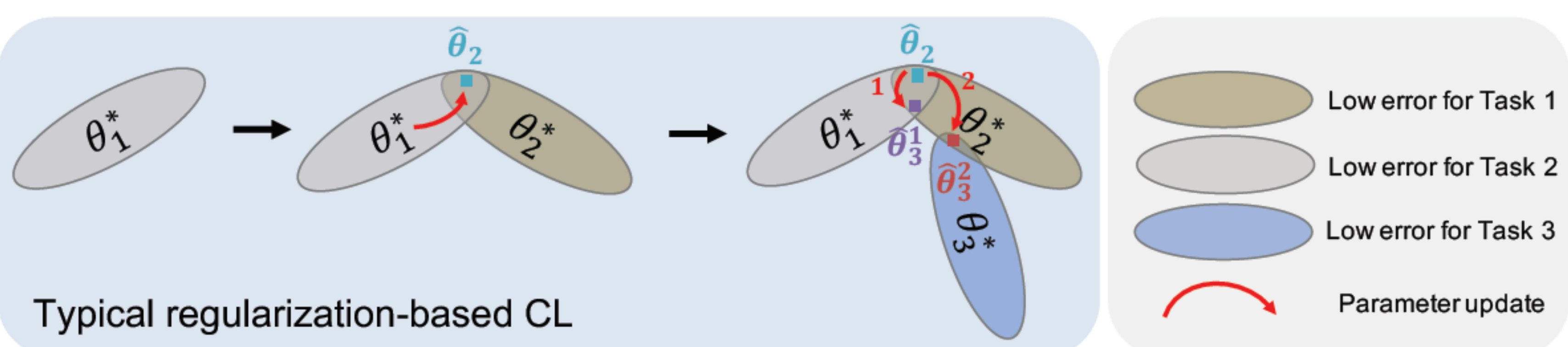
- **Algorithms for continual learning**
  - Dynamic network architecture-based (ex. PNN, DEN)
  - Dual memory system-based (Ex. GEM, iCaRL)
  - **Regularization-based (Reg-based)** (Ex. EWC, MAS, AGS-CL)

- **Stability-Plasticity dilemma of CL method**
  - **Stability**: overcoming catastrophic forgetting of previous tasks
  - **Plasticity**: learning news task well

- **Methods for converging to wide local minima**
  - Small mini-batch size, using a new optimizer (Ex. EntropySGD)
  - Regularizing the softmax output (Ex. **Entropy Maximization**)

Motivation

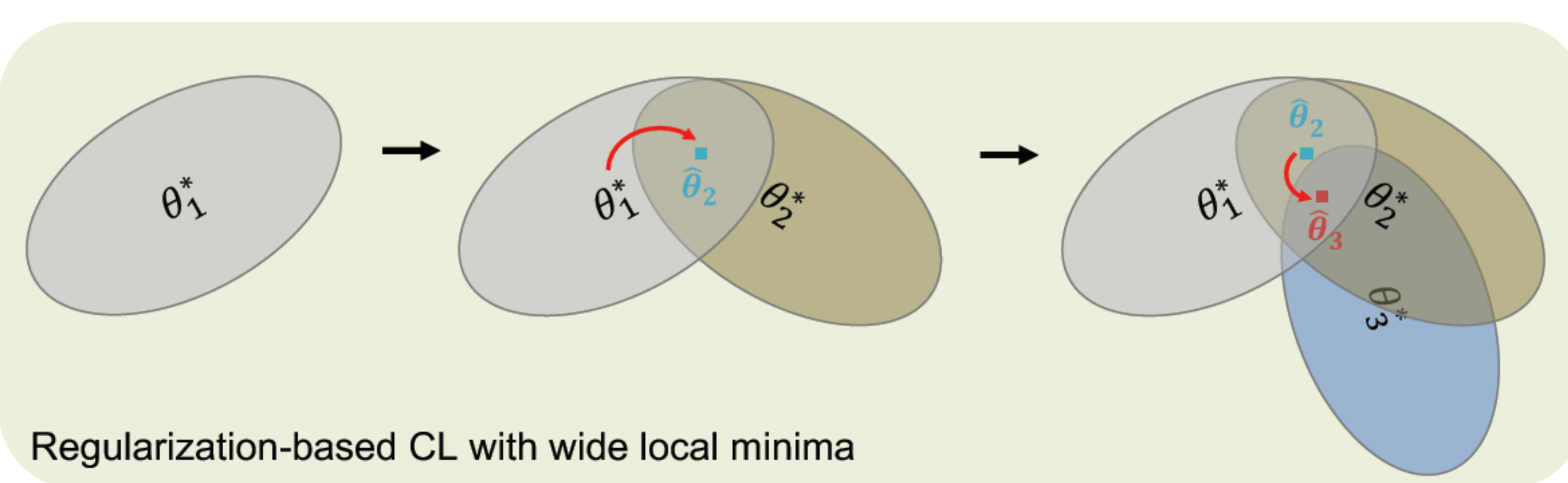
- **A geometric intuition of CPR**
  - Reg-based CL with **sharp (narrow) local minima**



ellipsoid: the uncertainty ellipsoid of parameter around the local minima

- It can **hurt** both stability and plasticity of CL

  - Reg-based CL with **wide (flat) local minima**



  - Promoting **wide local minima** during CL can be particularly beneficial for **regularization-based CL**
    - **Increase both the stability and plasticity at the same time!**

Method

- **Regularization-based continual learning**

$$L^t_{\text{CL}}(\boldsymbol{\theta}) = L^t_{\text{CE}}(\boldsymbol{\theta}) + \lambda \sum_i \Omega_i^{t-1}(\theta_i - \theta_i^{t-1})^2$$

  - λ : the regularization strength (hyperparameter)
  - $\{\theta_i^{t-1}\}$: the parameter learned until task $t-1$
  - $L^t_{\text{CE}}(\theta)$: the cross-entropy loss function for task t
  - $\boldsymbol{\Omega}^{t-1} = \{\Omega_i^{t-1}\}$: the set of estimates of the weight importance

- **Single-task wide local minima**

$$L_{\text{WLM}}(\boldsymbol{\theta}) = L_{\text{CE}}(\boldsymbol{\theta}) + \frac{\beta}{N}\sum_{n=1}^{N} D_{\text{KL}}(f_{\boldsymbol{\theta}}(\mathbf{x}_n) \| g)$$

  - β : the trade-off parameter (hyperparameter)
  - g : some probability distribution in Δ_M

- **CPR: Achieving wide local minima during CL**

$$L^t_{\text{CPR}}(\boldsymbol{\theta}) = L^t_{\text{CE}}(\boldsymbol{\theta}) + \frac{\beta}{N}\sum_{n=1}^{N} D_{\text{KL}}(f_{\boldsymbol{\theta}}(\mathbf{x}_n^t) \| P_U) + \lambda \sum_i \Omega_i^{t-1}(\theta_i^t - \theta_i^{t-1})^2,$$

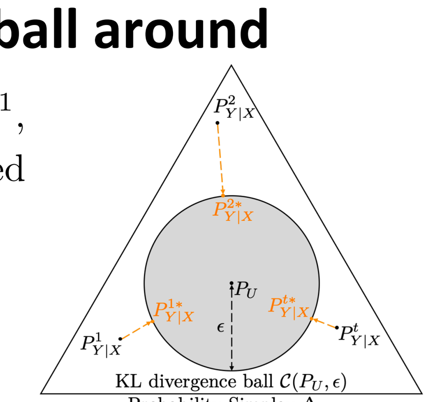
  - λ and β : the regularization parameters (hyperparameter)
  - P_U : the uniform distribution

CPR can be applied to **any** state-of-the-art reg-based CL!

- **Interpretation by information projection**
  - Information projection: $P^* = \underset{Q \in \mathcal{Q}}{\arg\min}\, D_{\text{KL}}(Q \| P)$
  - Classifier projection: $P^*_{Y|X} = \underset{Q_{Y|X} \in \mathcal{C}}{\arg\min}\, \mathbb{E}_{P_X}\left[D_{\text{KL}}(Q_{Y|X}(\cdot|X) \| P_{Y|X}(\cdot|X))\right]$
  - **CPR: Classifier projection onto a finite radius ball around**

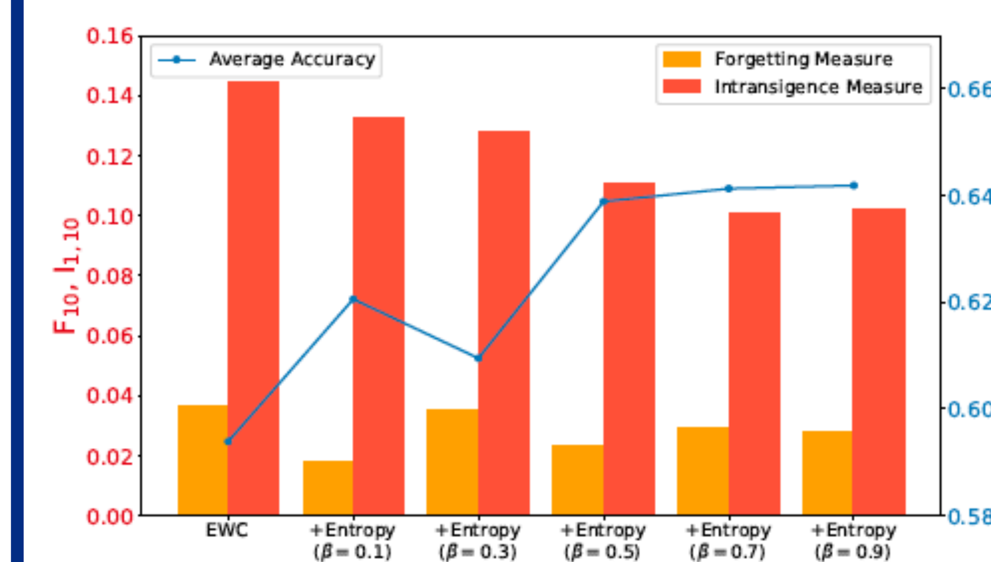
For any classifier $P^{t-1*}_{Y|X} \in \mathcal{C}(P_U, \epsilon)$ for task $t-1$ with data distribution P^{t-1}_X, and let any classifier for task t be $P^t_{Y|X} \notin \mathcal{C}(P_U, \epsilon)$ and $P^{t*}_{Y|X}$ be the projected classifier, then

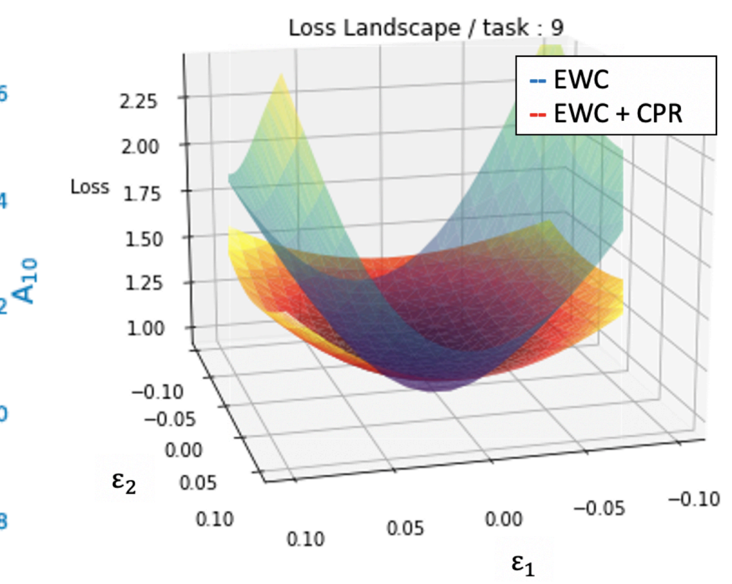
$$\mathbb{E}_{P^{t-1*}_{Y|X} P^{t-1}_X}\left[-\log P^t_{Y|X} P^{t-1}_X\right] \geq \mathbb{E}_{P^{t-1*}_{Y|X} P^{t-1}_X}\left[-\log P^{t*}_{Y|X} P^{t-1}_X\right].$$



Experimental Results

- **Quantifying the role of CPR during CL**
  - Selecting β for CPR / plotting the loss landscape





  - F_{10}: the measure for **stability**
  - $I_{1,10}$: the measure for **plasticity**

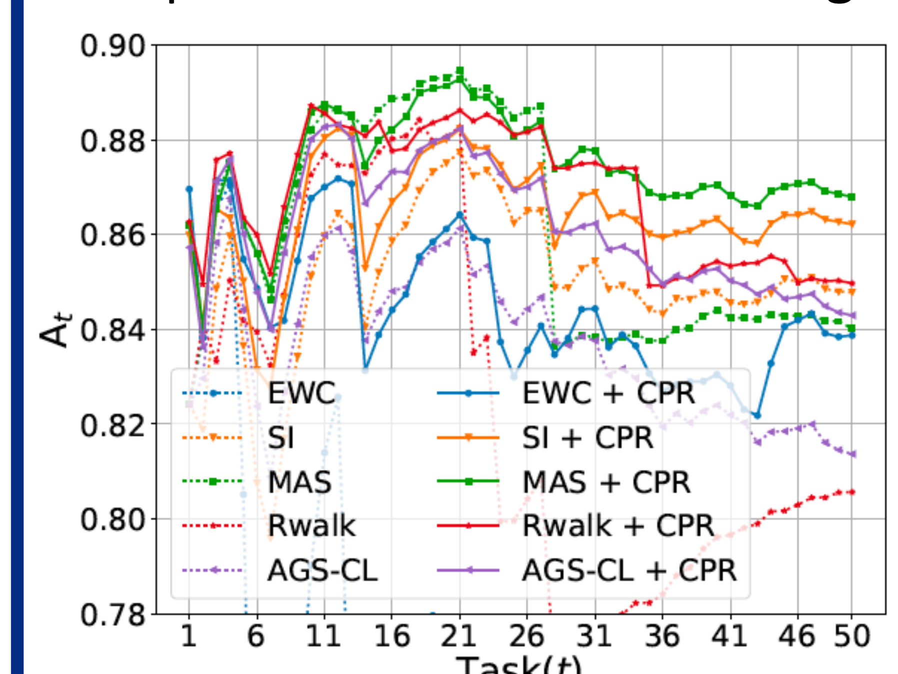
Make a model converge to **wide local minima**

- **Applying CPR to state-of-the-art reg-based CL methods**
  - Experimental results on **supervised learning** with various datasets

Dataset	Method	Average Accuracy (A_10) W/o CPR	Average Accuracy (A_10) W/ CPR	Average Accuracy (A_10) diff (W-W/o)	Forgetting Measure (F_10) W/o CPR	Forgetting Measure (F_10) W/ CPR	Forgetting Measure (F_10) diff (W/-W/o)	Intransigence Measure ($I_{1,10}$) W/o CPR	Intransigence Measure ($I_{1,10}$) W/ CPR	Intransigence Measure ($I_{1,10}$) diff (W-W/o)
CIFAR100 ($T=10$)	EWC	0.6002	0.6328	**+0.0326 (+5.2%)**	0.0312	0.0285	**-0.0027 (-8.7%)**	0.1419	0.1117	**-0.0302 (-21.3%)**
	SI	0.6141	0.6476	**+0.0336 (+5.5%)**	0.1106	0.0999	**-0.0107 (-9.7%)**	0.0566	0.0327	**-0.0239 (-42.2%)**
	MAS	0.6172	0.6442	**+0.0270 (+4.4%)**	0.0416	0.0460	**-0.0011 (-2.6%)**	0.1155	0.0778	**-0.0257 (-22.2%)**
	Rwalk	0.5784	0.6366	**+0.0581 (+10.0%)**	0.0937	0.0769	**-0.0169 (-18.0%)**	0.1074	0.0644	**-0.0430 (-40.0%)**
	AGS-CL	0.6369	0.6615	**+0.0246 (+3.9%)**	0.0259	0.0247	**-0.0012 (-4.63%)**	0.1100	0.0865	**-0.0235 (-24.4%)**
CIFAR10/100 ($T=11$)	EWC	0.6950	0.7055	**+0.0105 (+1.5%)**	0.0228	0.0181	**-0.0048 (-21.1%)**	0.1121	0.1058	**-0.0062 (-5.5%)**
	SI	0.7127	0.7186	**+0.0059 (+0.8%)**	0.0459	0.0408	**-0.0051 (-11.1%)**	0.0733	0.0721	**-0.0012 (-1.6%)**
	MAS	0.7239	0.7257	**+0.0017 (+0.2%)**	0.0479	0.0476	**-0.0003 (-0.6%)**	0.0603	0.0588	**-0.0015 (-2.5%)**
	Rwalk	0.6934	0.7046	**+0.0112 (+1.6%)**	0.0738	0.0707	**-0.0031 (-4.2%)**	0.0672	0.0589	**-0.0084 (-12.5%)**
	AGS-CL	0.7580	0.7613	**+0.0032 (+0.4%)**	0.0009	0.0009	**0**	0.0731	0.0697	**-0.0034 (-4.7%)**
Omniglot ($T=50$)	EWC	0.6632	0.8387	**+0.1755 (+26.5%)**	0.2096	0.0321	**-0.1776 (-84.7%)**	-0.0227	-0.0239	**-0.0012 (-5.3%)**
	SI	0.8478	0.8621	**+0.0143 (+1.7%)**	0.0247	0.0167	**-0.0079 (-32.0%)**	-0.0258	-0.0282	**-0.0065 (-25.3%)**
	MAS	0.8401	0.8679	**+0.0278 (+3.3%)**	0.0316	0.0101	**-0.0215 (-68.0%)**	-0.0247	-0.0314	**-0.0067 (-27.1%)**
	Rwalk	0.8056	0.8497	**+0.0440 (+5.5%)**	0.0644	0.0264	**-0.0380 (-59.0%)**	-0.0226	-0.0294	**-0.0068 (-30.1%)**
	AGS-CL	0.8553	0.8805	**+0.0253 (+3.0%)**	0	0	**0**	0.0323	0.0046	**-0.0277 (-85.8%)**
CUB200 ($T=10$)	EWC	0.5746	0.6098	**+0.0348 (+6.1%)**	0.0811	0.0807	**-0.0004 (-0.5%)**	0.1011	0.0667	**-0.0345 (-34.1%)**
	SI	0.6047	0.6232	**+0.0157 (+2.6%)**	0.0549	0.0474	**-0.0075 (-13.7%)**	0.0918	0.0827	**-0.0091 (-9.9%)**
	MAS	0.5842	0.6123	**+0.0281 (+4.8%)**	0.1188	0.1030	**-0.0158 (-13.3%)**	0.0575	0.0436	**-0.0139 (-24.2%)**
	Rwalk	0.6078	0.6324	**+0.0247 (+4.1%)**	0.0811	0.0601	**-0.0210 (-25.9%)**	0.0679	0.0621	**-0.0058 (-8.5%)**
	AGS-CL	0.5403	0.5623	**+0.0220 (+4.07%)**	0.0750	0.0692	**-0.0058 (-7.7%)**	0.1408	0.1241	**-0.0167 (-11.7%)**

CPR improves **accuracy, plasticity** and **stability** of reg-based CL methods for all datasets

- Experimental results on **Omniglot**



- Experimental results on **RL**