Continual Learning through Control Minimization

Sander de Haan, Yassine Taoudi-Benchekroun, Pau Vilimelis Aceituno, Benjamin F. Grewe

Institute of Neuroinformatics, University of Zürich and ETH Zürich
ETH AI Center, ETH Zürich

Grewe Lab × ETH AI Center

TL;DR
We reframe continual learning as a control problem. Learning and preservation signals compete inside neural dynamics.
At equilibrium, weight updates implicitly encode the full prior-task Fisher Information Matrix without replay.

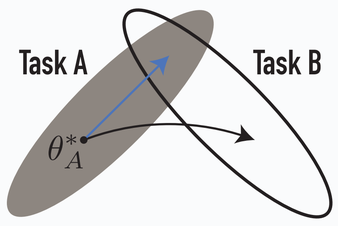
catastophic forgetting

Neural networks trained sequentially rapidly lose performance

Setup
- Assume Task A is optimally trained at parameters θ_A^*
- Assume Task B is now being trained, without replay

Parameter-based regularization methods protect important parameters for old tasks, usually measured by the Fisher evaluated over Task A data $\mathcal{D}_A$ at the optimal parameters θ_A^*

$$\mathcal{R}(\theta) \propto F_A \triangleq \mathbb{E}_{\mathcal{D}_A}\left[\nabla\mathcal{L}\,\nabla\mathcal{L}^\top\right]_{\theta_A^*}$$



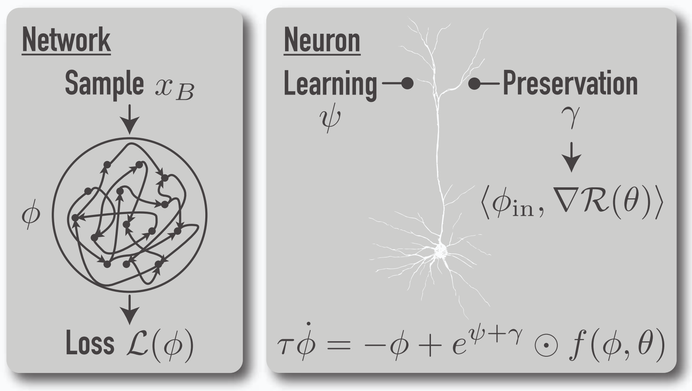
$$\nabla\mathcal{L}(\mathcal{X}) + \nabla\mathcal{R}(\theta)$$
(EWC, Kirkpatrick 2016)

Problems (Kim 2022; Masana 2022; Wu 2024)
- Stored curvature estimates misalign during training
- Parameter interactions are often discarded
- Regularization cannot perform class discrimination

Preservation is a correction after gradients computation
The learning signal is blind to preservation

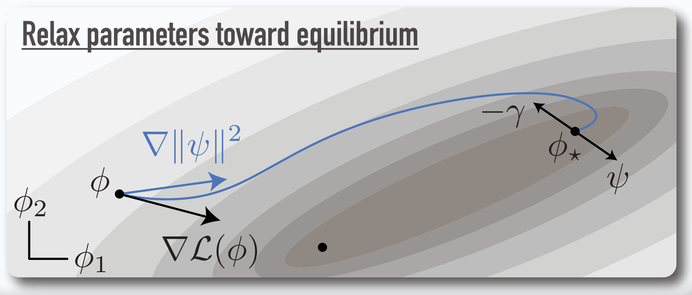
control minimization

Instead of correcting gradients, let learning and preservation compete before parameters update, by embeding preservation into the network's activity dynamics



$$\min_{\psi} \|\psi\|^2 \text{ s.t. } \underbrace{\phi = e^{\psi+\gamma} \odot f(\phi,\theta)}_{\text{equilibrium of } \phi}, \underbrace{\nabla_\phi\mathcal{L}(\phi) = \mathbf{0}}_{\text{loss minimum}}$$

Find the smallest learning signal that reaches equilibrium, taking the path of least resistance against preservation



continual-natural gradient

The control-minimization objective acts as a budget that the learning signal must allocate selectively, and directions that conflict with the previous tasks become expensive

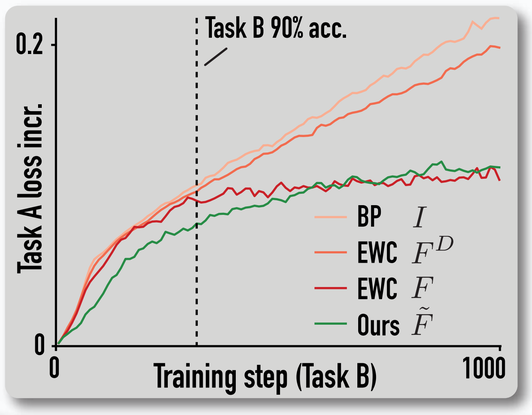
Through the network dynamics, the learning signal implicitly encodes all relevant second-order parameter interactions of the preservation signal at equilibrium

Learning therefore proceeds inside the geometry from the already-acquired knowledge

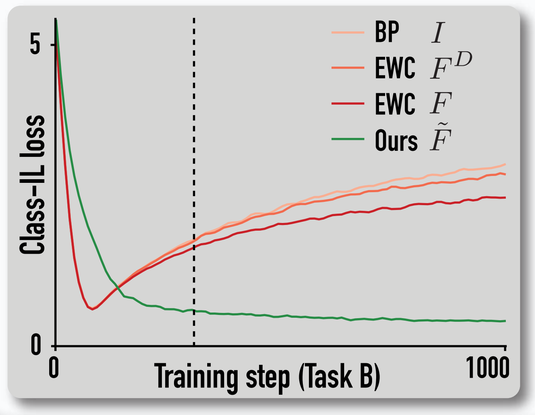
$$\Delta\theta \approx -\eta\tilde{F}_A^{-1}\nabla\mathcal{L}_B(x_B)$$

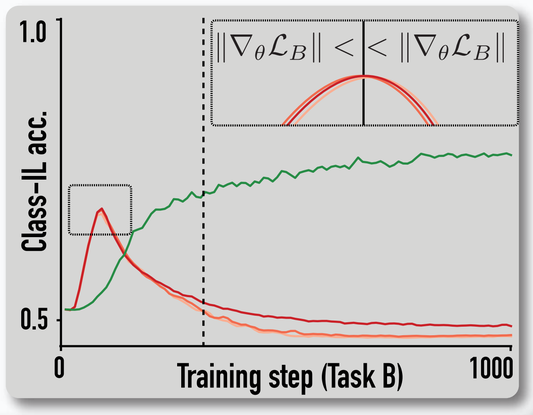
Properties
- Class discrimination becomes possible
  interference between tasks is filtered by the dynamics, which no parameter-space correction can remove
- Forgetting bounds tighten
  loss no longer increases with network size, and the new task's curvature no longer interferes with preservation
- Curvature stays current
  preservation is computed from the network dynamics at current parameters, not from a frozen prior optimum

implicit dynamical approximation tracks the existing curvature

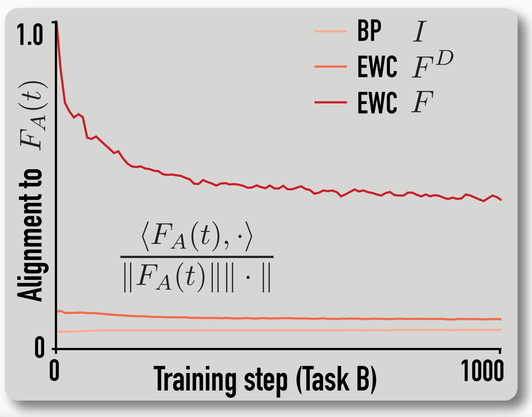


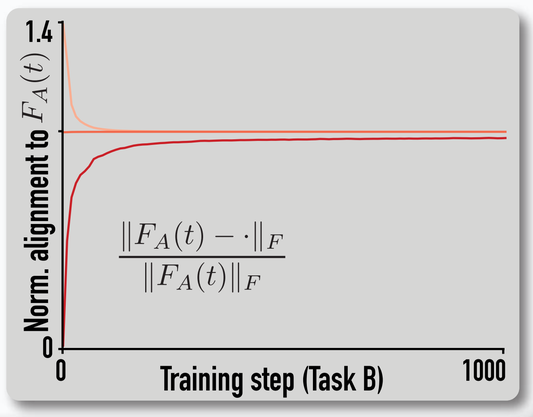
backpropagation fails regardless of curvature quality



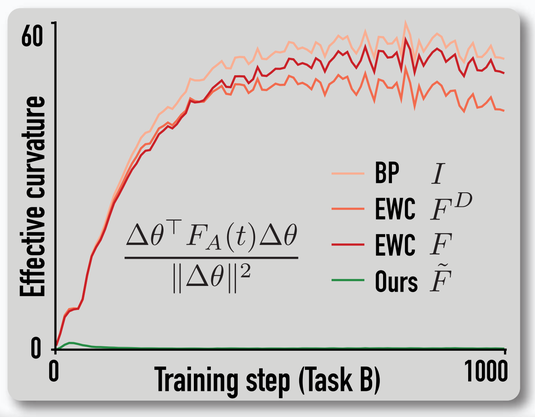


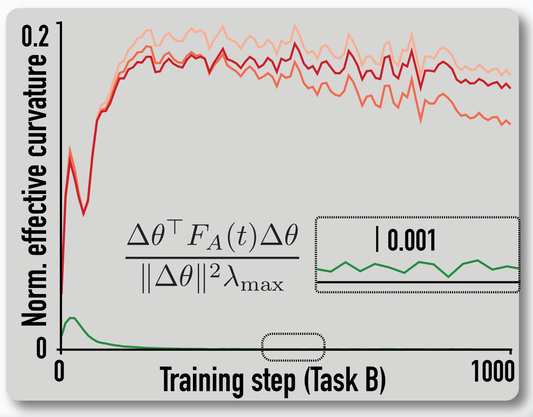
progressive mislignment between stored and true curvature





weights updates inside the previously learned curvature





	Method		Split-MNIST		Split-CIFAR10		Split-Tiny-ImageNet	
			Task-IL	*Class-IL*	*Task-IL*	*Class-IL*	*Task-IL*	*Class-IL*
Baselines	JOINT	(upper bound)	99.7±0.0	98.2±0.1	98.3±0.1	92.2±0.2	82.0±0.1	60.0±0.2
	SGD	(lower bound)	98.1±0.1	19.3±0.0	95.5±0.5	19.6±0.1	28.0±0.5	3.8±0.1
Parameter-based regularization	EWC	(Kirkpatrick et al., 2017)	97.2±0.6	19.8±0.0	95.5±1.2	20.9±1.1	34.2±0.4	4.6±0.1
	oEWC	(Schwarz et al., 2018)	**99.2±0.4**	20.2±0.8	95.9±0.7	21.2±2.0	35.5±0.3	4.8±0.0
	SI	(Zenke et al., 2017)	97.5±0.9	19.7±0.0	92.6±1.3	19.8±0.1	35.5±0.5	4.7±0.0
	CSQN	(Eeckt & Hamme, 2025)	98.2±0.3	19.2±0.0	95.4±0.7	20.2±1.2	35.5±0.5	4.8±0.1
	EFC	**(ours)**	98.3±0.7	**51.4±4.8**	**96.2±0.3**	**50.2±7.0**	**37.2±0.4**	**8.8±0.1**
Replay (200)	DER++	(Buzzega et al., 2020)	98.6±0.2	71.1±2.2	95.4±0.8	62.3±1.2	39.0±1.6	10.5±1.2

Table 1. Classification accuracies on continual learning benchmarks. We report mean ± standard deviation over five random seeds. JOINT trains on all tasks simultaneously (upper bound). SGD trains sequentially without any continual learning mechanism (lower bound).

Key Takeaways

- Preservation and learning compete inside the neural activity
- Parameter interactions are obtained for free by solving network dynamics
- Backprop always fails regardless of curvature quality