



Loss of Plasticity in Deep Continual Learning

In deep *continual* learning, a network may fail to learn new tasks, even though the same architecture can easily learn the same task from random initialization.

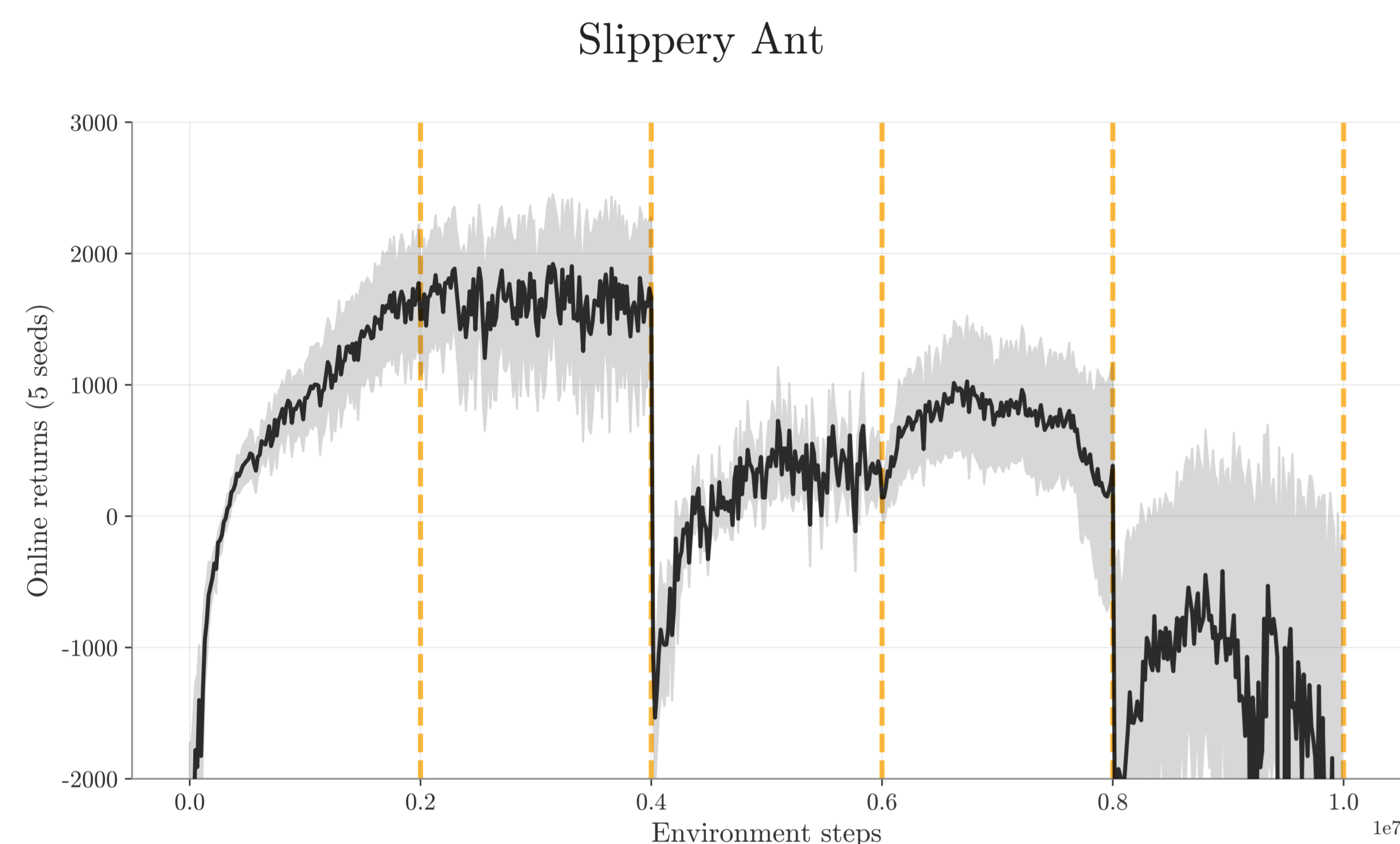


Figure 1. When the friction changes periodically, standard ppo fails to perform well.

Goal of our work

Explain why loss of plasticity happens in deep continual learning!

τ -Trainability through the lens of spectral collapse

We consider a sequence of tasks $\{\tau = 0, 1, \dots, T-1\}$ with shared parameters $\theta \in \mathbb{R}^m$. For task τ . Let $f_\tau(\theta)$: evaluation metric (accuracy, return), $L_\tau(\theta)$ the differentiable surrogate loss. The learner receives a budget of K gradient steps. Let initialization for task τ be the final weights at the end of the previous optimization: $\theta_\tau^{(0)} = \theta_{\tau-1}^{(K)}$.

Successful training on task τ

$$f_\tau(\theta_\tau^{(K)}) \geq \sup_{\theta} f_\tau(\theta) - v,$$

Definition (τ -trainability). A network is τ -trainable if

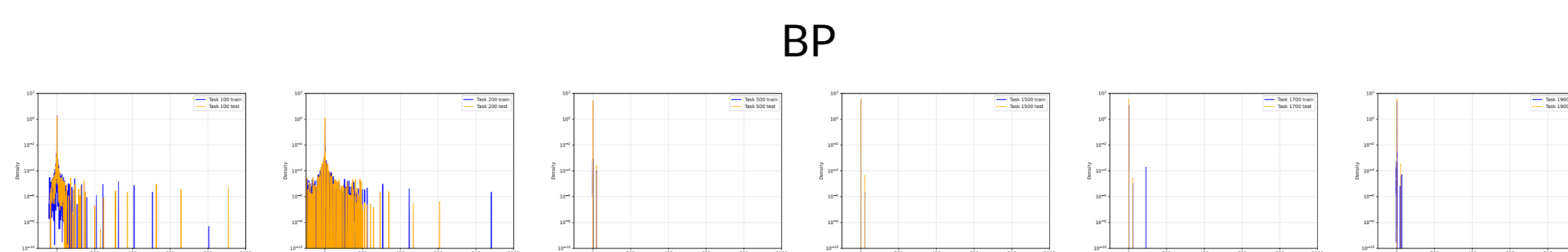
$$\text{rank}(H_\tau^{(0)}) \geq \rho_\tau,$$

for a task/optimizer-dependent threshold ρ_τ . The probability that this event holds is denoted as τ -trainability. Note: being τ -trainable is a necessary condition for successful training on task τ .

[Learn more about \$\tau\$ -trainability as a framework for neural network plasticity in our paper!](#)

Spectral Collapse

Algorithms that lose plasticity:



Algorithms that preserve plasticity:

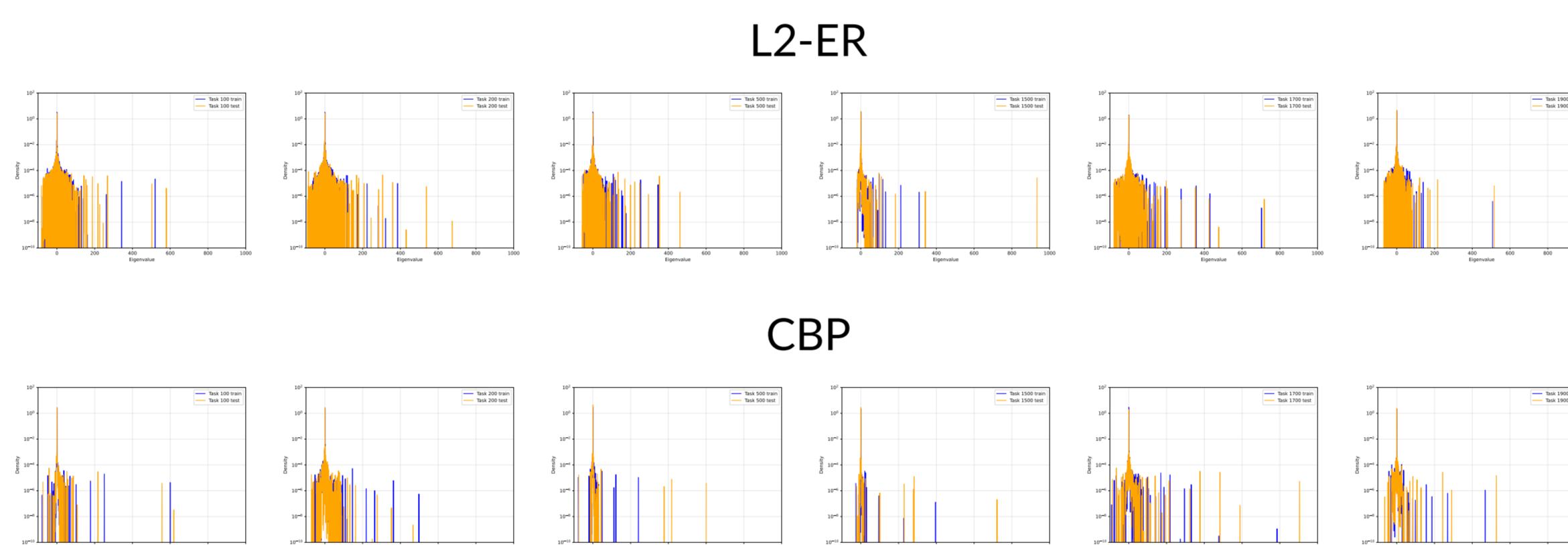


Figure 2. Columns show Hessian eigenspectra at the initialization of tasks 100, 200, 500, 1500, 1700, and 1900.

Empirical finding. Spectral Collapse (measured through the ϵ -rank of a Hessian) strongly correlates with **trainability**: a large ϵ -rank is a *necessary* condition for high task performance.

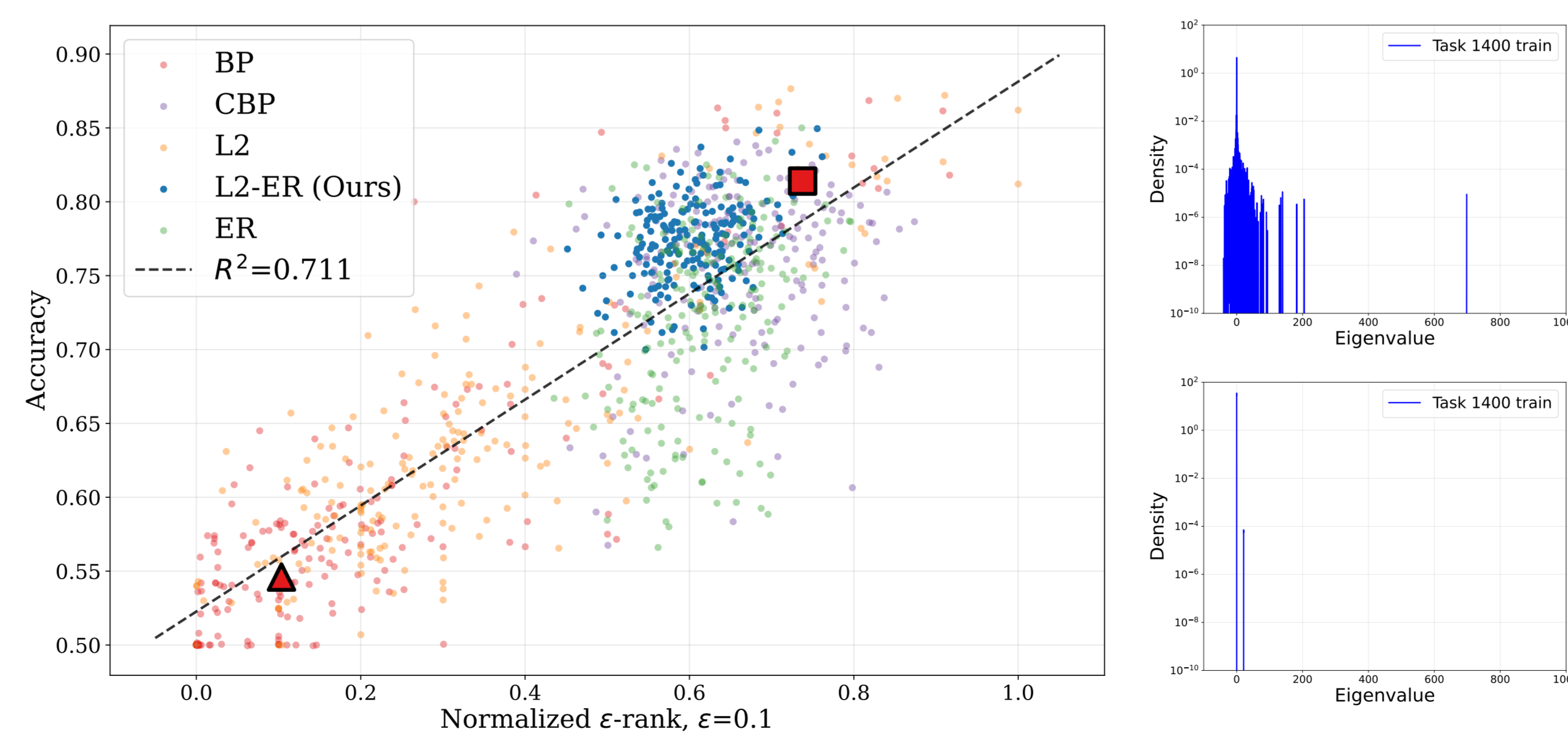


Figure 3. (Left) Accuracy (i.e., successful training) vs. curvature (normalized Hessian ϵ -rank) on Continual ImageNet. Linear fit (dotted) highlights the positive association. (Top right) L2-ER spectrum *before* collapse at task 1400. (Bottom right) BP spectrum *after* collapse at task 1400.

Regularizing Curvature to Preserve Plasticity (L2-ER)

Idea. If spectral collapse kills plasticity, then explicitly keeping the Hessian spectrum “spread out” should keep the network trainable. We operationalize this by regularizing a *Kronecker-factored* generalized Gauss-Newton (GGN) approximation of the Hessian.

$$H_{\text{GGN}} \approx \hat{H}_{\text{GGN}} = \left(\frac{1}{N} \sum_{i=1}^N a_i^{(\ell)} a_i^{(\ell)\top} \right) \otimes \left(\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C g_{i,c}^{(\ell)} g_{i,c}^{(\ell)\top} \right), \quad (1)$$

which separates *feature statistics* and *gradient statistics* into Kronecker factors.

L2-ER objective. We augment the task loss with penalties on both effective feature rank and parameter norm:

$$\min_{\theta} L_\tau(\theta) = \mathbb{E}_{(x,y) \sim d_\tau} [\ell_\tau(F_\theta(x), y)] - \text{erank} \left(\sum_{\ell} \mathbb{E}[a_\ell a_\ell^\top] \right) + \lambda \|\theta\|_2^2.$$

- The *input variance term* regularizes the GGN approximation rank.
- The *L2 term* controls parameter norm growth and helps ensure that increases in $\text{rank}(H_{\text{GGN}})$ translate into increases in the true Hessian rank and τ -trainability. [see theorem B.3].

Experiments

