Overcoming catastrophic forgetting in neural networks

James Kirkpatrick[a,1], Razvan Pascanu[a], Neil Rabinowitz[a], Joel Veness[a], Guillaume Desjardins[a], Andrei A. Rusu[a], Kieran Milan[a], John Quan[a], Tiago Ramalho[a], Agnieszka Grabska-Barwinska[a], Demis Hassabis[a], Claudia Clopath[b], Dharshan Kumaran[a], and Raia Hadsell[a]

[a]DeepMind, London EC4 5TW, United Kingdom; and [b]Bioengineering Department, Imperial College London, London SW7 2AZ, United Kingdom

PNAS, Proceedings of the National Academy of Sciences
citation: 4

2017/04/28
Katy@Datalab

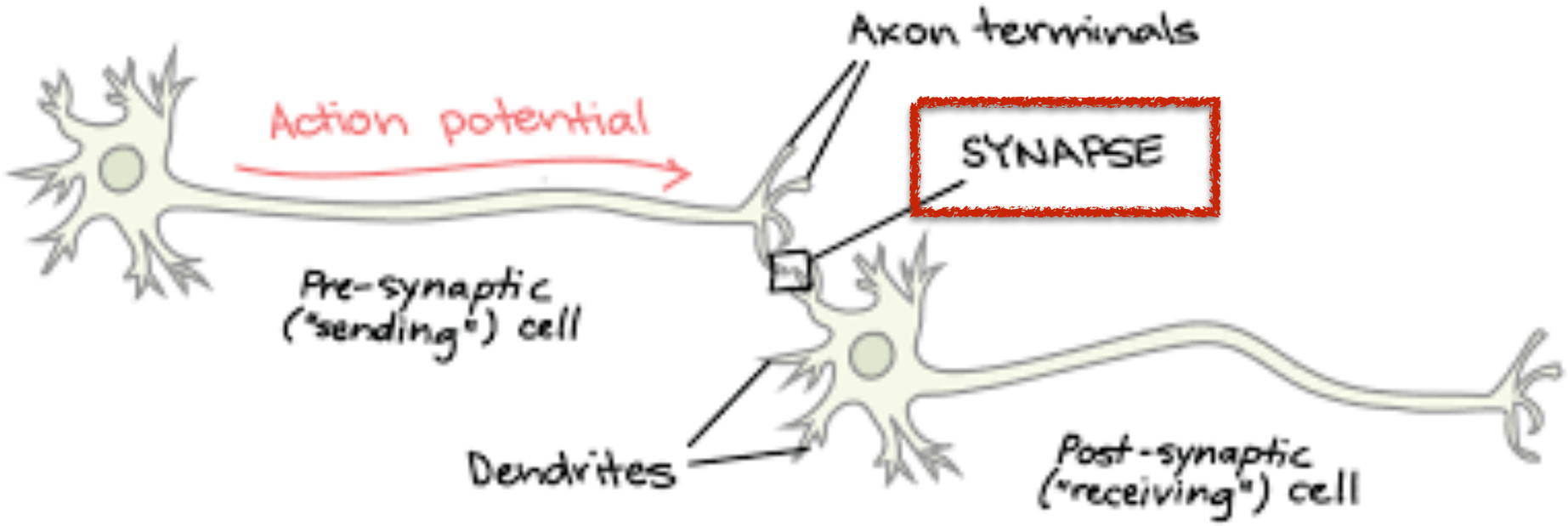
Background

- **Catastrophic forgetting** is forgetting key information needed to solve a previous task when training on a new task.

The kind of consolidation that occur in the brain

- Continual learning in the neocortex relies on **task-specific synaptic consolidation**

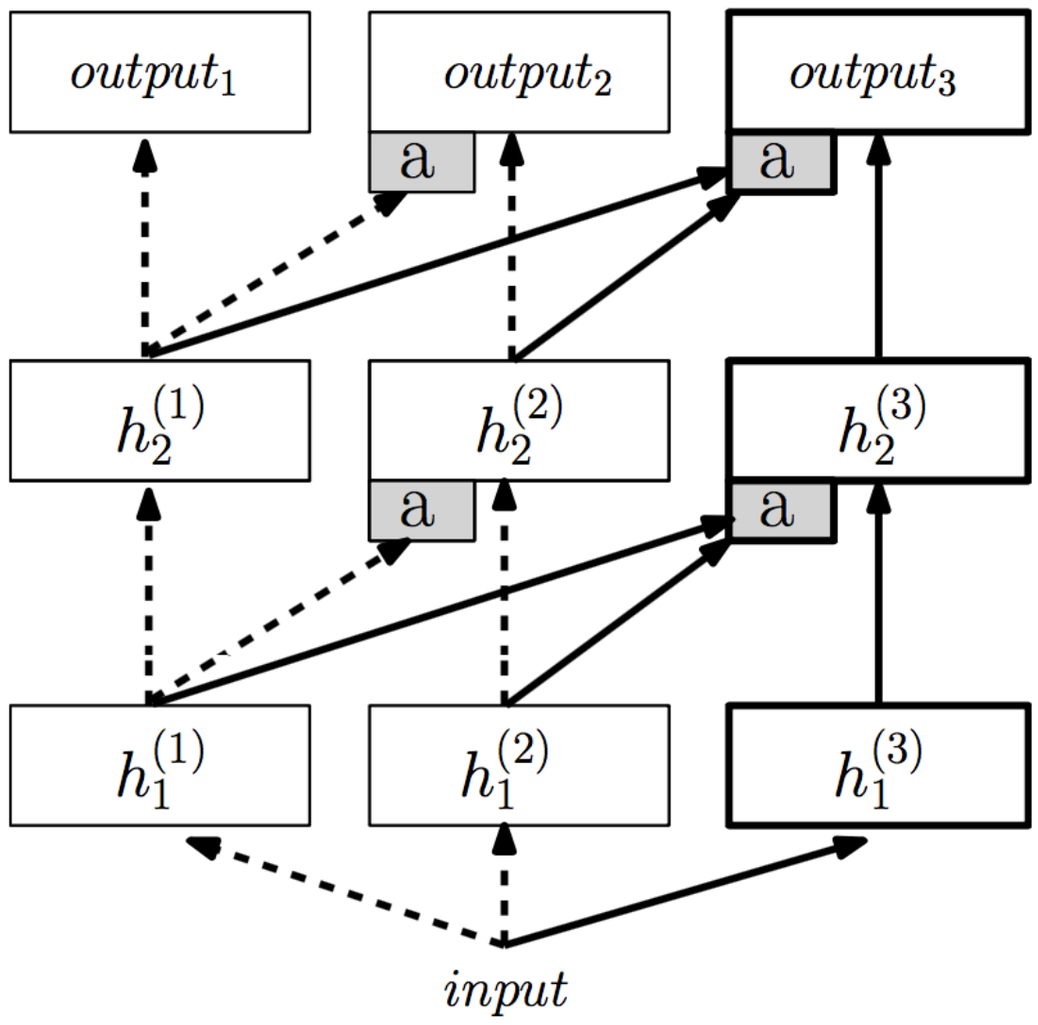
- **Synaptic consolidation:** connections between neurons are less likely to be overwritten if they have been important in previously learnt tasks.



- knowledge is durably encoded by rendering a proportion of synapses less plastic and therefore stable over long timescales.

Related Work

- Ensemble of DNNs: Rusu, Andrei A., et al. "Progressive neural networks." arXiv preprint arXiv: 1606.04671 (2016).



- Fernando, Chrisantha, et al. "**Pathnet**: Evolution channels gradient descent in super neural networks." arXiv preprint arXiv:1701.08734 (2017).

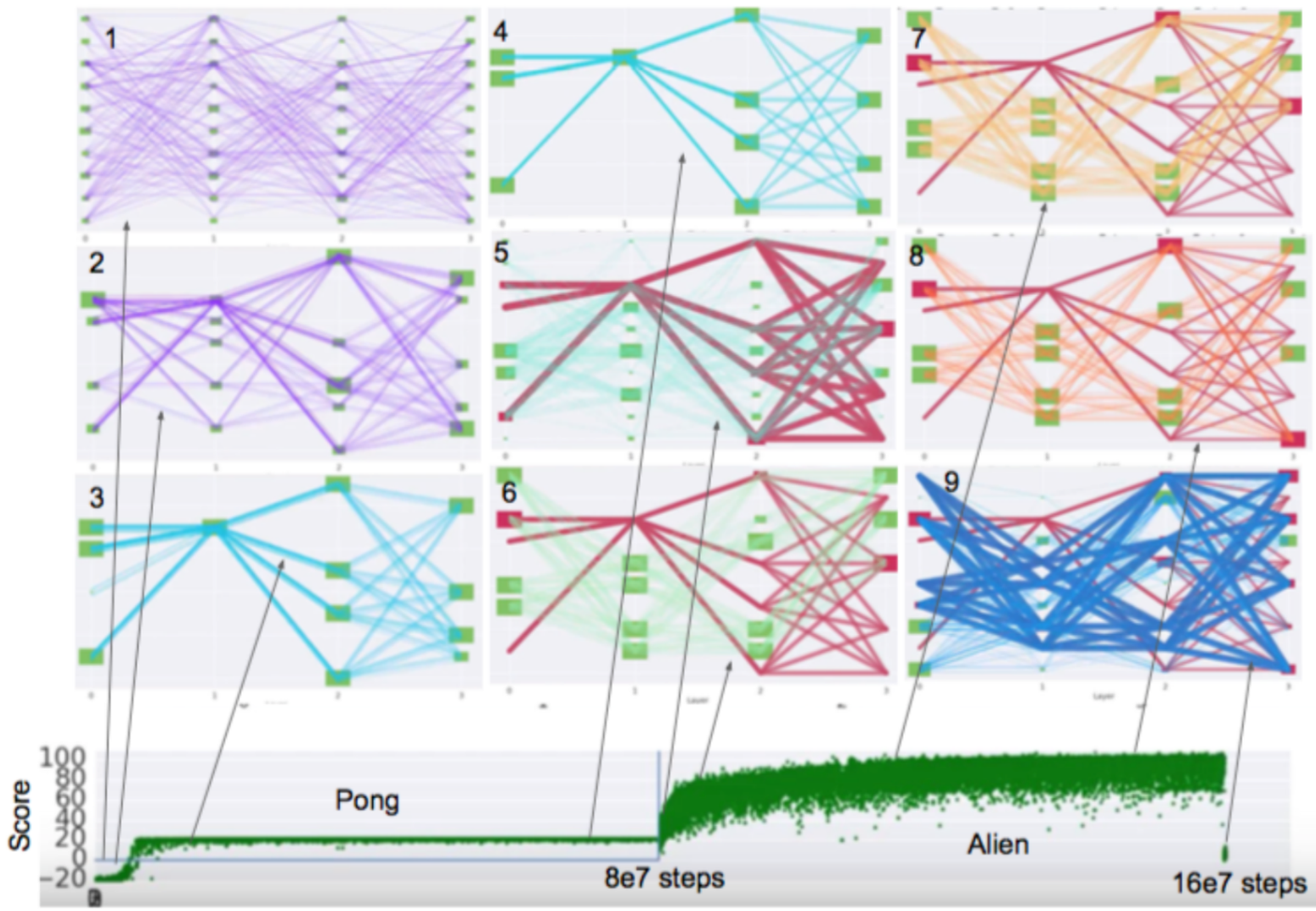


Figure 1: A population of randomly initialized pathways (purple lines in Box 1) are evolved whilst learning task A, Pong. At the end of training, the best pathway is fixed (dark red lines in Box 5) and a new population of paths are generated (light blue lines in Box 5) for task B. This population is then trained on Alien and the optimal pathway that is evolved on Alien is subsequently fixed at the end of training, shown as dark blue lines in Box 9.

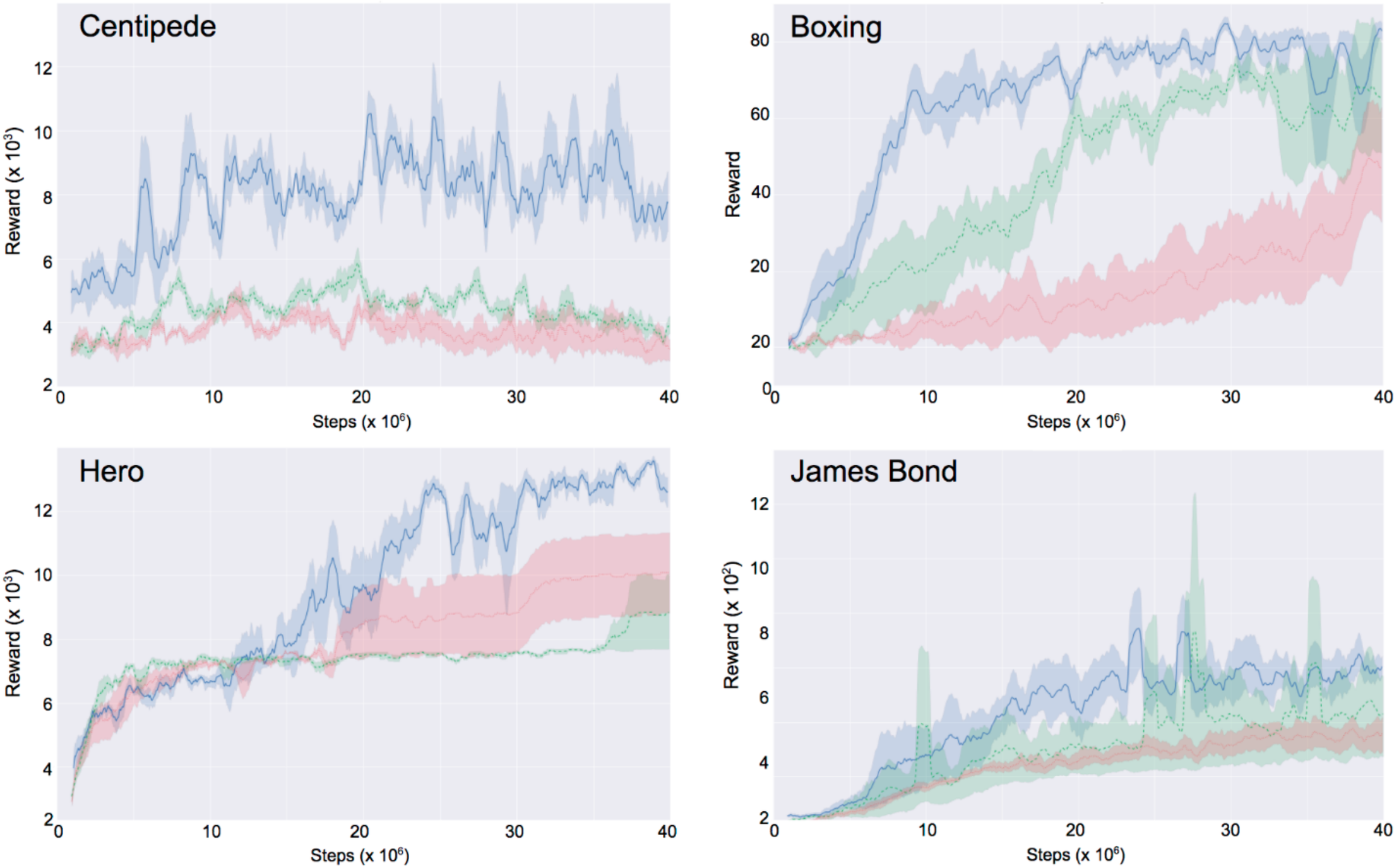
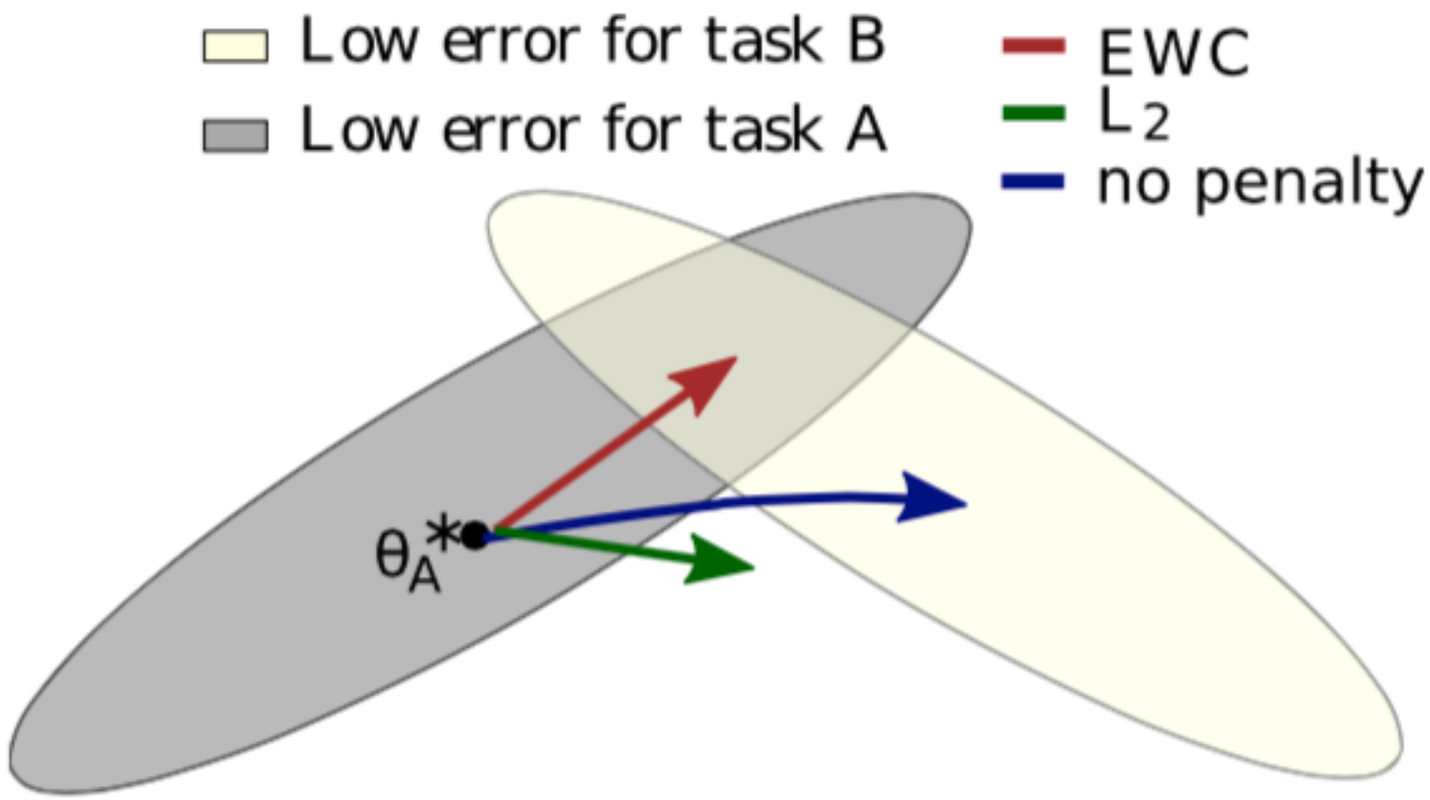


Figure 6: PathNet shows positive transfer from River Raid to Centipede, Boxing, Hero and James Bond. The graphs show the reward over the first 40 million steps of training. Blue shows the results from the best five hyperparameter settings of PathNet out of 243. Compare these to the best five hyperparameter setting runs for independent learning (red) and fine-tuning (green) controls out of 45.

Intuition

- They implement an algorithm that performs a similar operation in artificial neural networks by **constraining important parameters to stay close to their old values**.

ELASTIC WEIGHT CONSOLIDATION (**EWC**)



- θ_A^* refers to the configuration of θ that performs well at A

- slow down the learning for the weights that were important to previous task(s)

Key Question

- How to determine which weights are most important to task A?

loss for B only

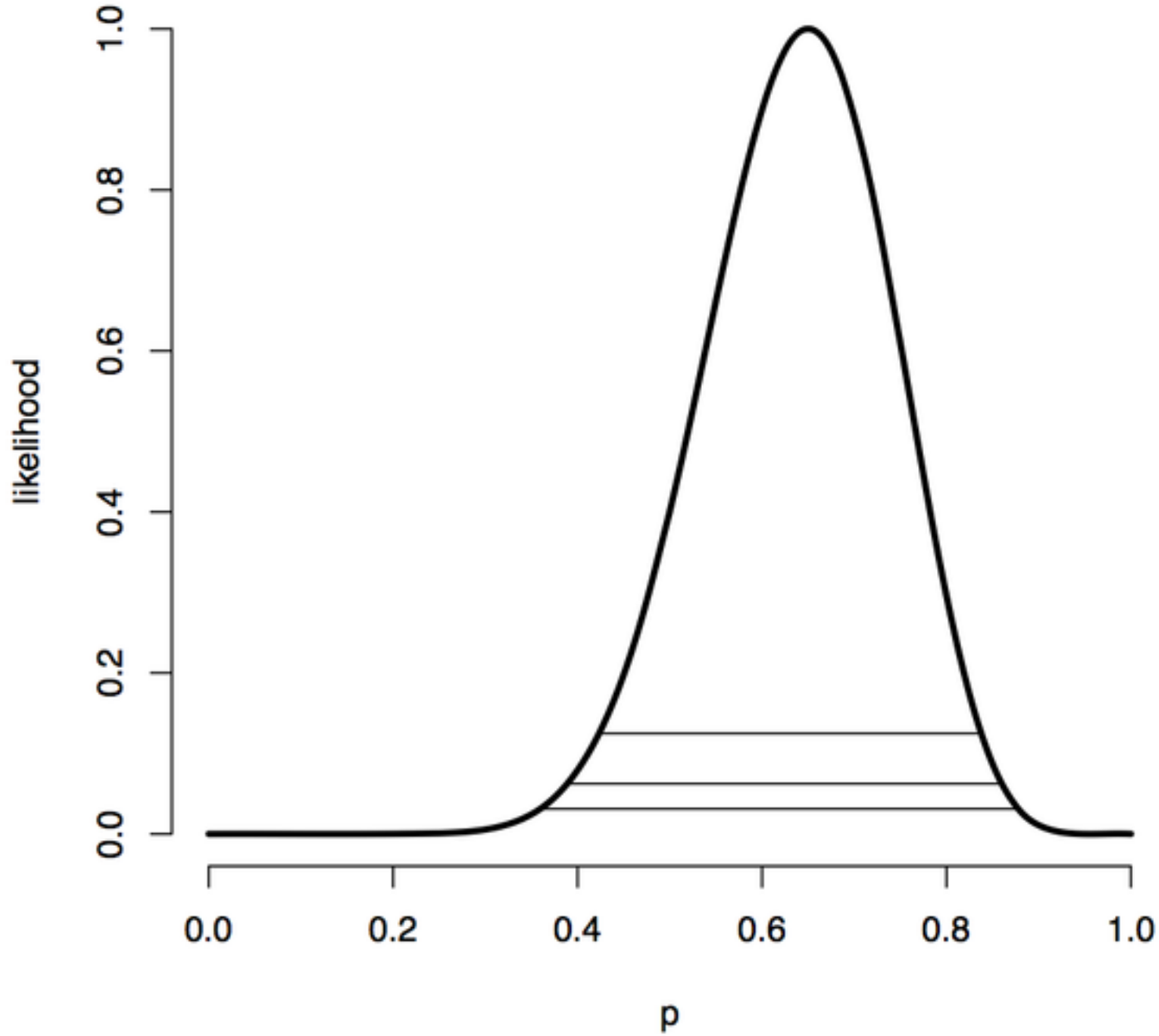
$$\theta^{A,B} = \text{argmin}_\theta \, \mathcal{L}_B(\theta) + \frac{1}{2} \sum_i F^A_{i,i} (\theta_i - \theta^A_i)^2$$

regulariser makes sure we don't catastrophically forget about A.

$$\mathbf{F}_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta] = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(X)].$$

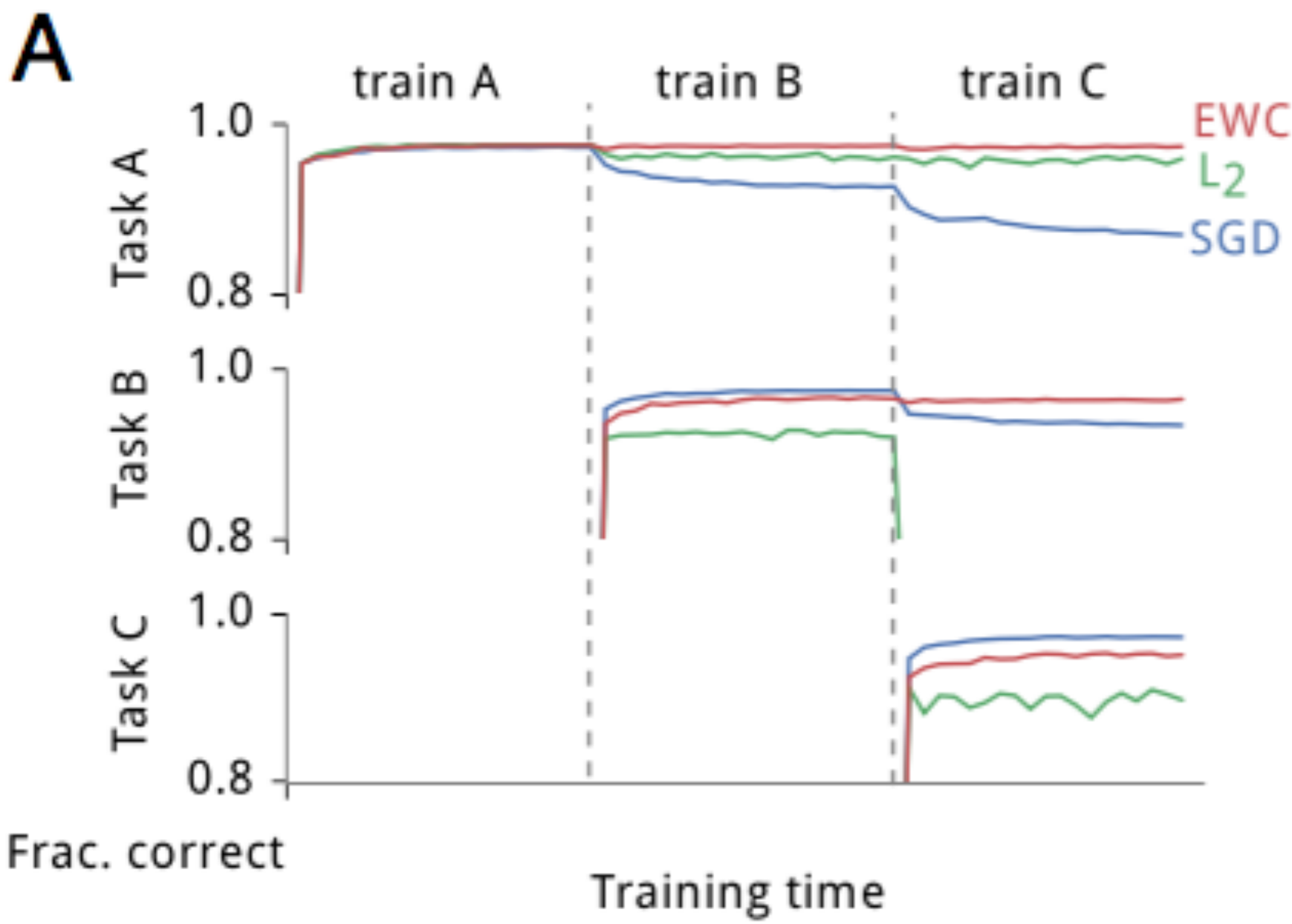
- The bigger the fish information(negative second-order derivatives), the more pointy of the distribution of the likelihood

- likelihood is more sensitive to this parameter

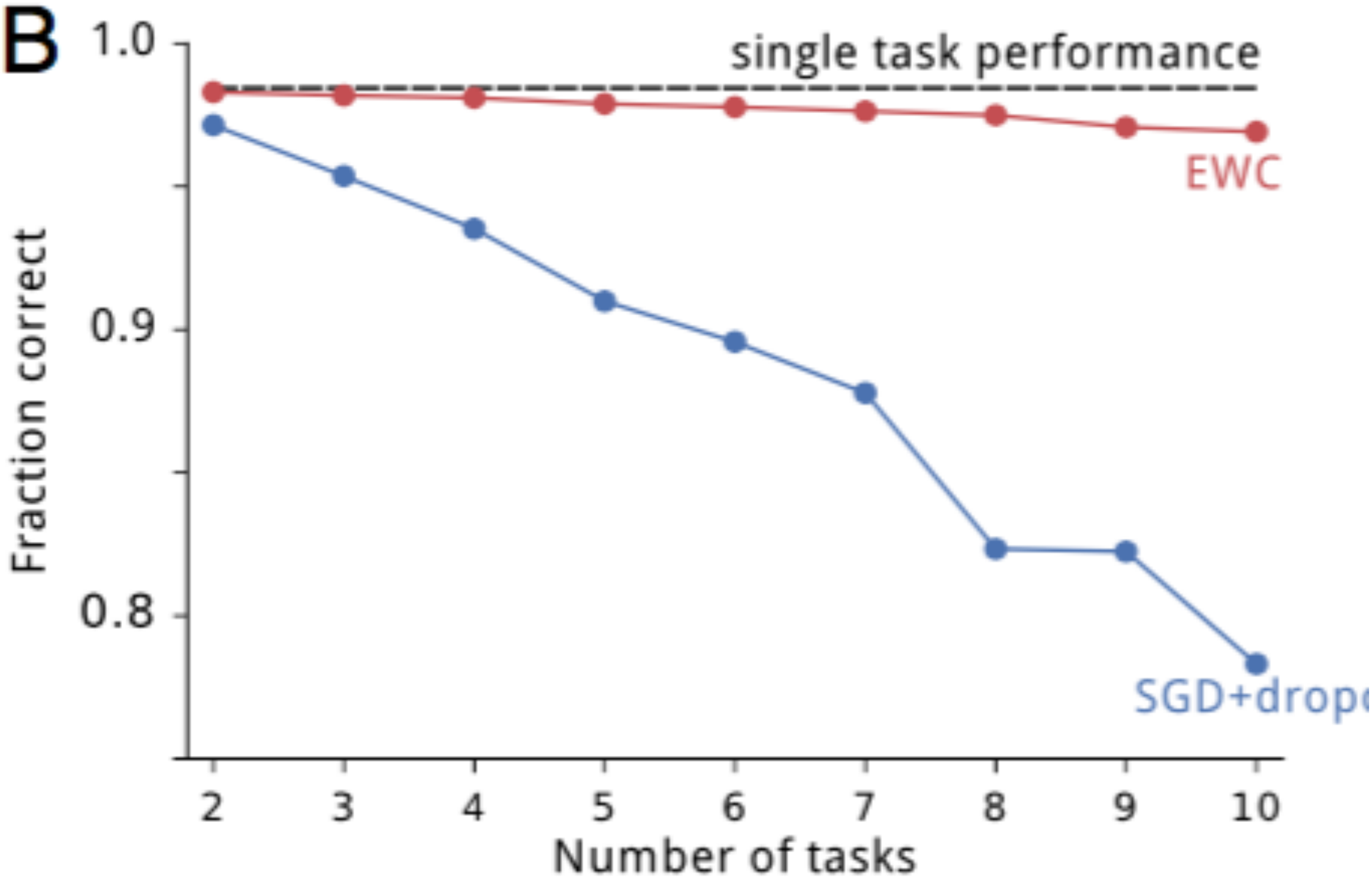


Experiment 1 over MNIST

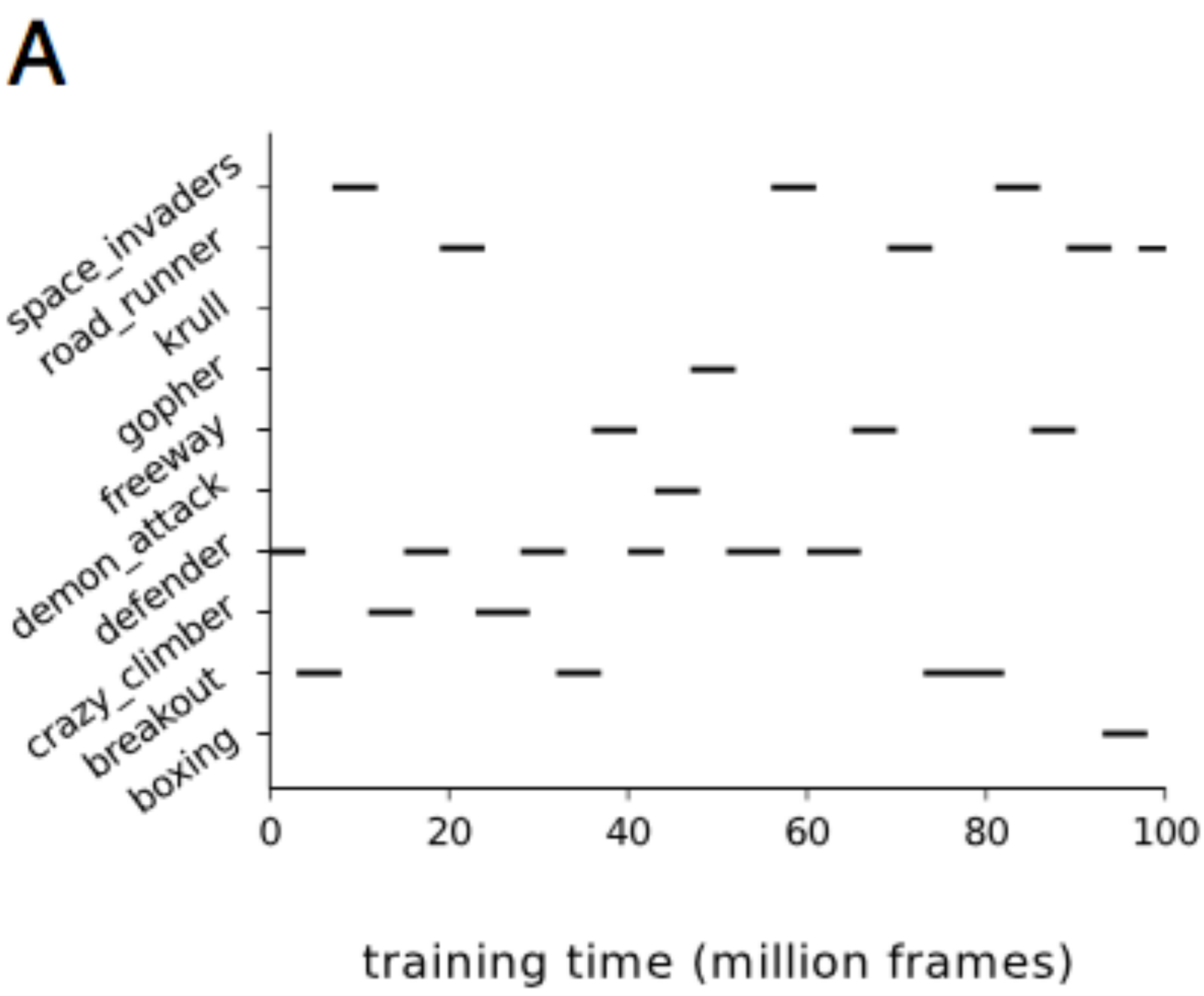
evaluation



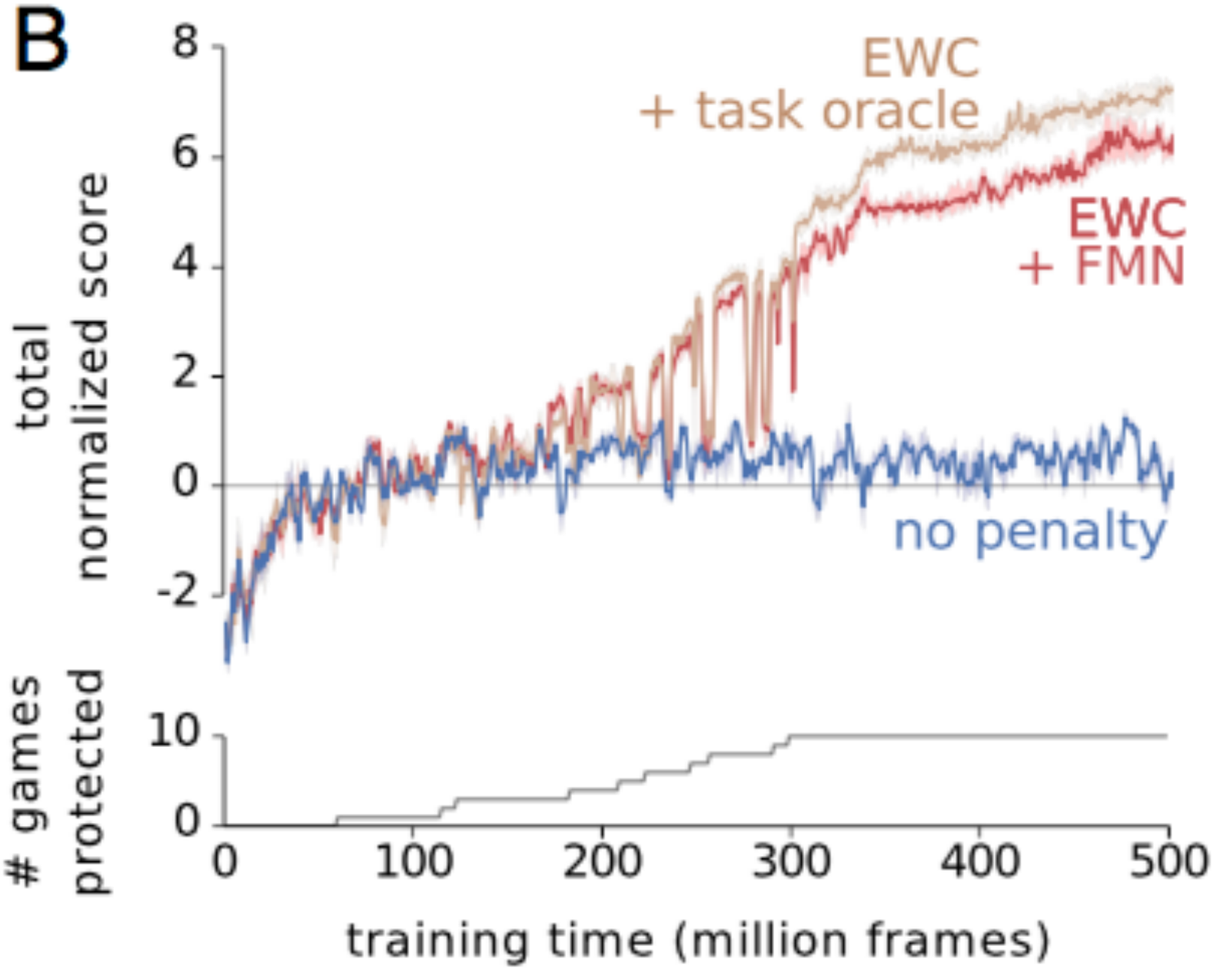
B


1.0
0.9
0.8
Fraction correct
single task performance
EWC
SGD+drop
2
3
4
5
6
7
8
9
10
Number of tasks

Experiment 2 on Atari Training Schedule



10: human level,
1: random agent



Conclusion

- The ability to learn tasks in succession without forgetting is a core component of biological and artificial intelligence.
- Weight uncertainty should inform learning rates.
- EWC allows the network to effectively squeeze in more functionality into a network with fixed capacity,