On the Computation of the Fisher Information in Continual Learning

Blog post, ICLR 2025

Gido van de Ven

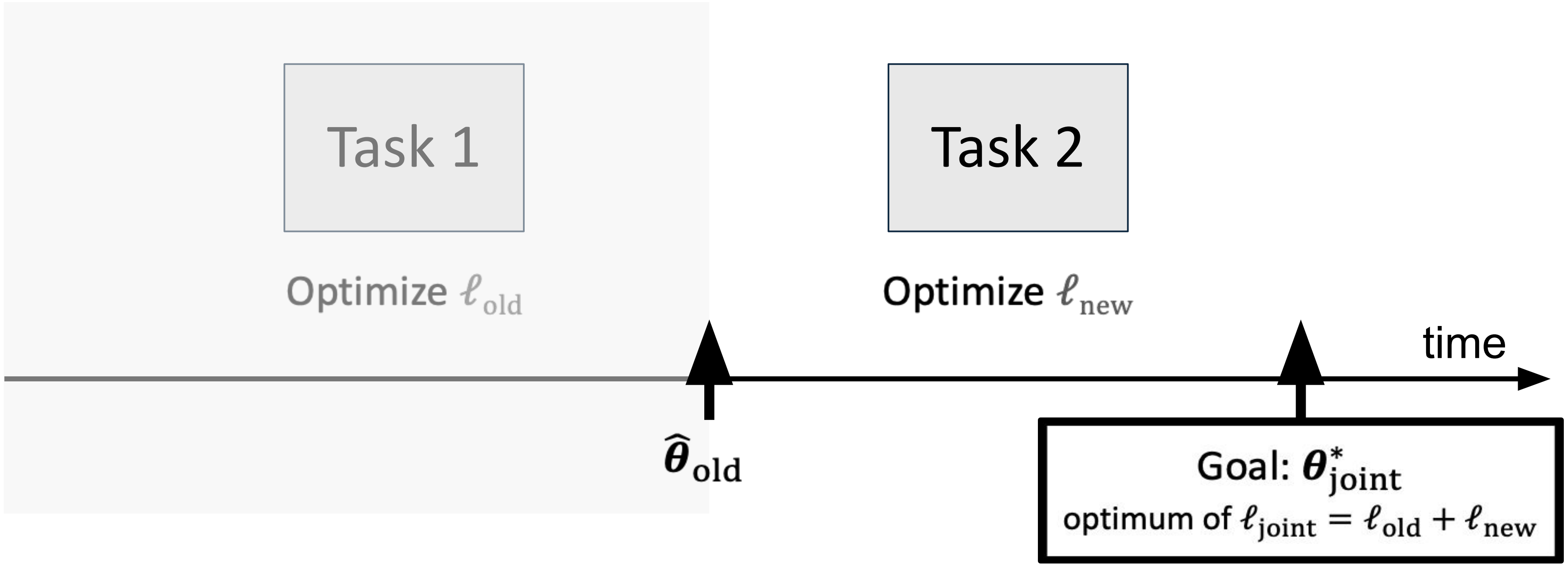
–

April 2025

Funding acknowledgement: This work has been supported by a senior post-doctoral fellowship from the Research Foundation – Flanders (FWO) under grant number 1266823N.

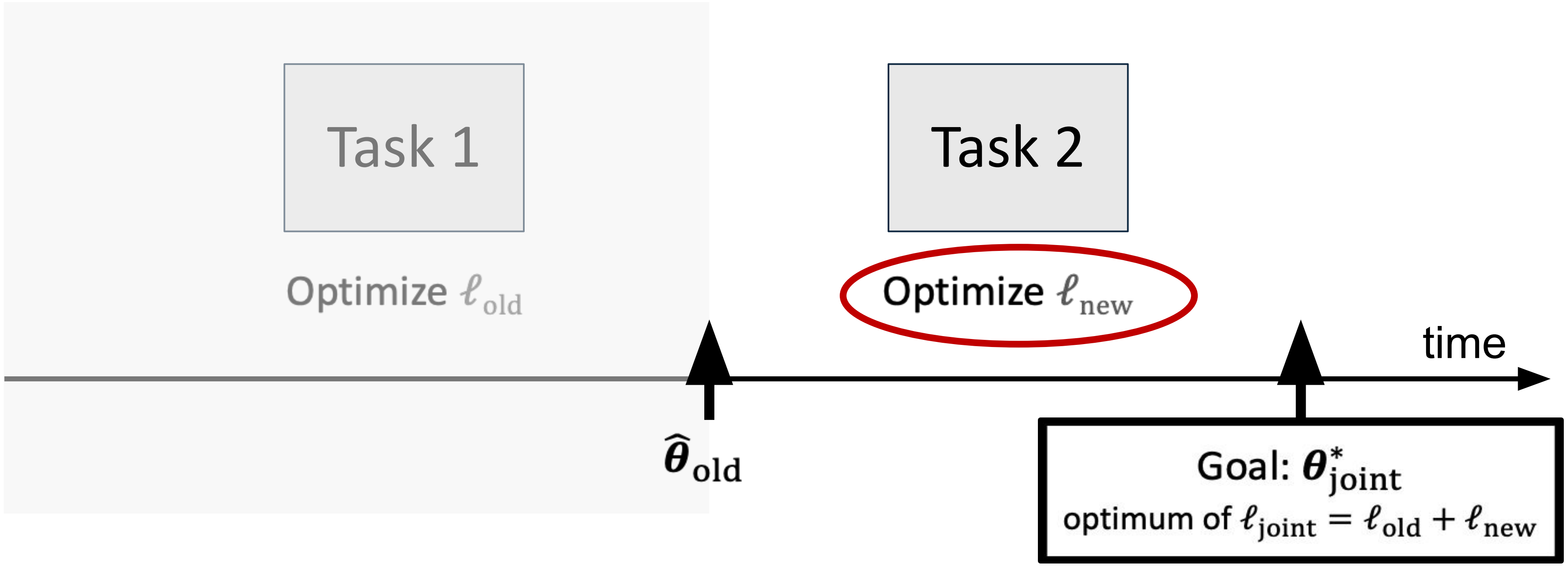
The Continual Learning Problem

→ Optimize the parameters $\boldsymbol{\theta}$ of a neural network $f_{\boldsymbol{\theta}}$ for two tasks that are observed one after the other



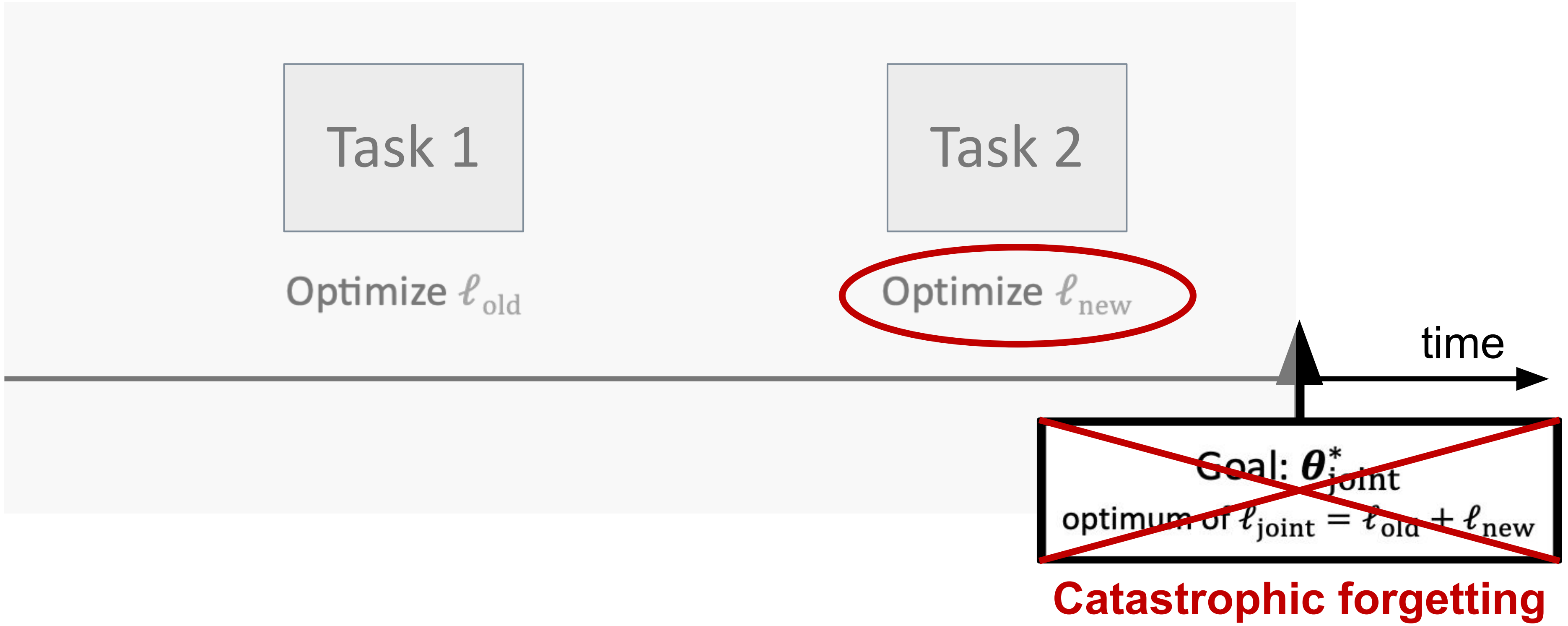
The Continual Learning Problem

→ Optimize the parameters $\boldsymbol{\theta}$ of a neural network $f_{\boldsymbol{\theta}}$ for two tasks that are observed one after the other



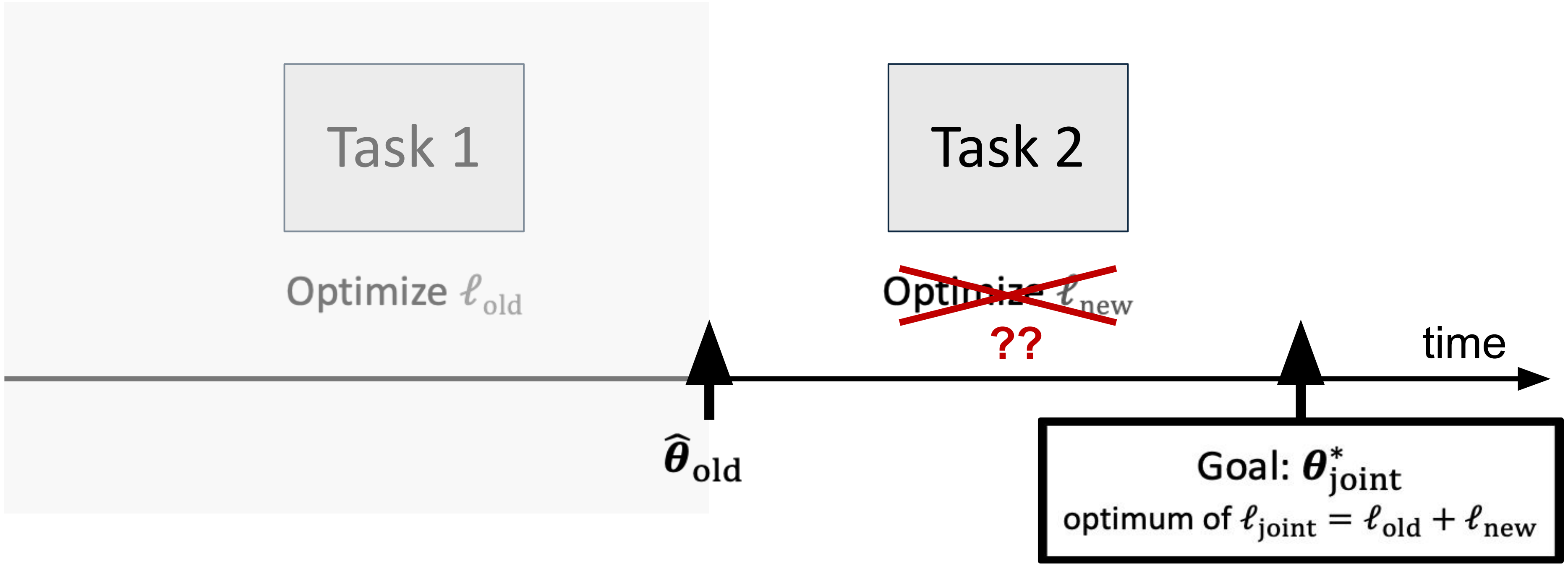
The Continual Learning Problem

→ Optimize the parameters $\boldsymbol{\theta}$ of a neural network $f_{\boldsymbol{\theta}}$ for two tasks that are observed one after the other



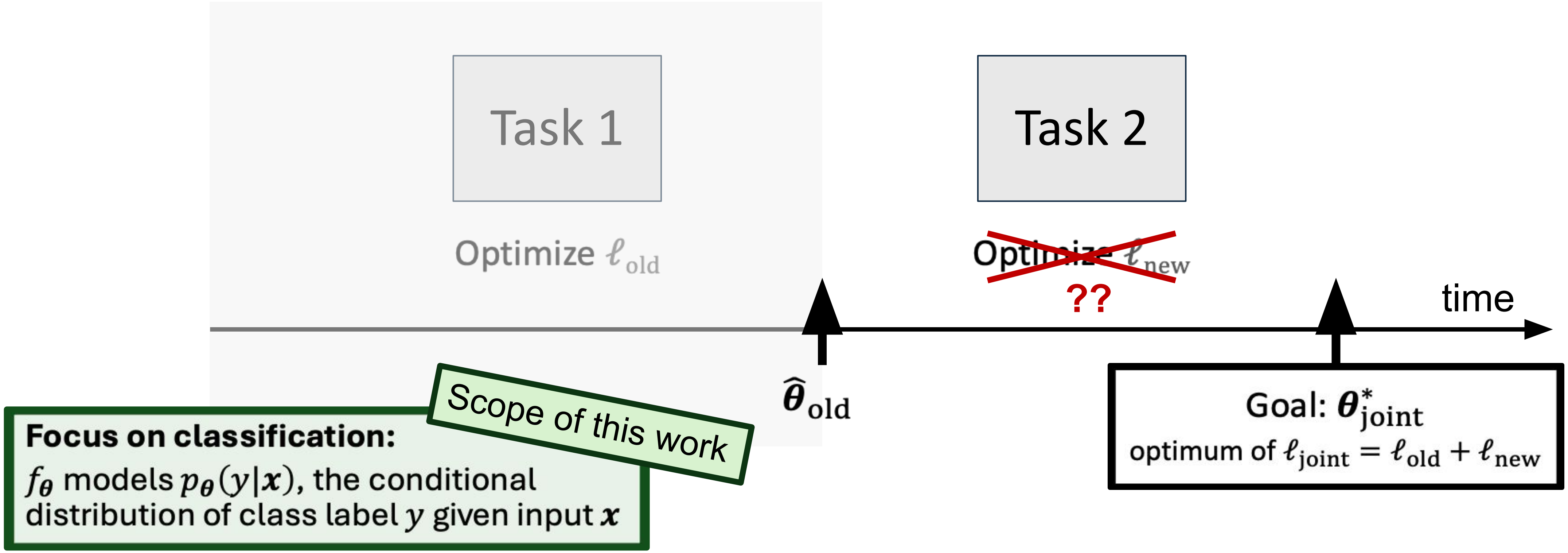
The Continual Learning Problem

→ Optimize the parameters $\boldsymbol{\theta}$ of a neural network $f_{\boldsymbol{\theta}}$ for two tasks that are observed one after the other



The Continual Learning Problem

→ Optimize the parameters $\boldsymbol{\theta}$ of a neural network $f_{\boldsymbol{\theta}}$ for two tasks that are observed one after the other



Elastic Weight Consolidation (EWC)

- One of the most popular continual learning methods, >8000 citations (Google Scholar)
- Used as baseline in large proportion of continual learning studies

Overcoming catastrophic forgetting in neural networks

PNAS

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]

March, 2017

Elastic Weight Consolidation (EWC)

- One of the most popular continual learning methods, >8000 citations (Google Scholar)
- Used as baseline in large proportion of continual learning studies

- When training on a new task, EWC adds an extra term to the loss:

$$\ell_{\mathrm{EWC}}(\boldsymbol{\theta}) = \ell_{\mathrm{new}}(\boldsymbol{\theta}) + \frac{\lambda}{2} \sum_{i=1}^{N_{\mathrm{params}}} F_{\mathrm{old}}^{i,i} \left(\theta^{i} - \hat{\theta}_{\mathrm{old}}^{i}\right)^{2}$$

i^{th} diagonal element of the old network's
Fisher Information matrix on the old task

Elastic Weight Consolidation (EWC)

- One of the most popular continual learning methods, >8000 citations (Google Scholar)
- Used as baseline in large proportion of continual learning studies

- When training on a new task, EWC adds an extra term to the loss:

$$\ell_{\mathrm{EWC}}(\boldsymbol{\theta}) = \ell_{\mathrm{new}}(\boldsymbol{\theta}) + \frac{\lambda}{2} \sum_{i=1}^{N_{\mathrm{params}}} F_{\mathrm{old}}^{i,i} \left(\theta^i - \hat{\theta}_{\mathrm{old}}^i\right)^2$$

i^{th} diagonal element of the old network's Fisher Information matrix on the old task

My claim: EWC is usually implemented sub-optimally, and most of its currently reported results can likely be improved

A Closer Look at the Fisher Information

Following Martens, 2020 – *JMLR*, the i^{th} diagonal element of the network's Fisher Information matrix on the data of the old task, is defined as:

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{old}}}\left[\mathbb{E}_{y\sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}}\left[\left(\left.\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^{i}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}}\right)^{2}\right]\right]$$

A Closer Look at the Fisher Information

Following Martens, 2020 – *JMLR*, the i^{th} diagonal element of the network's Fisher Information matrix on the data of the old task, is defined as:

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{old}}} \left[\mathbb{E}_{y \sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}} \left[\left(\left. \frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

- There are two expectations:

(1) An **outer expectation** over $\mathcal{D}_{\text{old}}$, the input distribution of the old task

(2) An **inner expectation** over $p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x})$, the conditional distribution of y given $\boldsymbol{x}$, defined by the network after training on the old task

A Closer Look at the Fisher Information

Following Martens, 2020 – *JMLR*, the i^{th} diagonal element of the network's Fisher Information matrix on the data of the old task, is defined as:

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{old}}}\left[\mathbb{E}_{y\sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}}\left[\left(\left.\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}}\right)^2\right]\right]$$

- There are two expectations:

(1) An **outer expectation** over $\mathcal{D}_{\text{old}}$, the input distribution of the old task

(2) An **inner expectation** over $p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x})$, the conditional distribution of y given $\boldsymbol{x}$, defined by the network after training on the old task

A Closer Look at the Fisher Information

Following Martens, 2020 – *JMLR*, the i^{th} diagonal element of the network's Fisher Information matrix on the data of the old task, is defined as:

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{old}}}\left[\mathbb{E}_{y\sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}}\left[\left(\left.\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}}\right)^2\right]\right]$$

- There are two expectations:

(1) An **outer expectation** over $\mathcal{D}_{\text{old}}$, the input distribution of the old task

(2) An **inner expectation** over $p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x})$, the conditional distribution of y given $\boldsymbol{x}$, defined by the network after training on the old task

Different Ways to Compute the Fisher: *(1) Exact*

- **Outer expectation:** estimate using all training data D_{old}
- **Inner expectation:** compute exactly

$$F_{\text{old}}^{i,i} = \frac{1}{|D_{\text{old}}|} \sum_{\boldsymbol{x} \in D_{\text{old}}} \left(\sum_{y=1}^{N_{\text{classes}}} p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x}) \left(\left. \frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right)$$

I will refer to this option as **EXACT**

Definition

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x} \sim D_{\text{old}}} \left[\mathbb{E}_{y \sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}} \left[\left(\left. \frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

Different Ways to Compute the Fisher: *(2) Sampling data points*

- **Outer expectation:** estimate using n random samples from D_{old}
- **Inner expectation:** compute exactly

$$F_{\text{old}}^{i,i} = \frac{1}{n} \sum_{\boldsymbol{x} \in S_{D_{\text{old}}}^{(n)}} \left(\sum_{y=1}^{N_{\text{classes}}} p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}(y|\boldsymbol{x}) \left(\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right)$$

Definition

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\text{old}}} \left[\mathbb{E}_{y \sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}} \left[\left(\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

I will explore this option using $n = 500$, referring to it as **EXACT (*n*=500)**

Different Ways to Compute the Fisher: *(3) Sampling labels*

- **Outer expectation:** estimate using over all training data D_{old}
- **Inner expectation:** estimate using a single Monte Carlo sample

$$F_{\text{old}}^{i,i} = \frac{1}{|D_{\text{old}}|} \sum_{\boldsymbol{x} \in D_{\text{old}}} \left(\left. \frac{\partial \log p_{\boldsymbol{\theta}} (c_{\boldsymbol{x}} | \boldsymbol{x})}{\partial \theta^i} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2$$ with $c_{\boldsymbol{x}}$ randomly sampled from $p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}(.|\boldsymbol{x})$

I will refer to this option as **SAMPLE**

Definition

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x} \sim D_{\text{old}}} \left[\mathbb{E}_{y \sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}} \left[\left(\left. \frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

Different Ways to Compute the Fisher:
(4) Empirical Fisher

- **Outer expectation:** estimate using all training data D_{old}
- **Inner expectation:** approximate by computing the squared gradient only for the ground-truth label

$$F_{\text{old}}^{i,i} = \frac{1}{|D_{\text{old}}|} \sum_{(\boldsymbol{x},y)\in D_{\text{old}}} \left(\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2$$

I will refer to this option as **EMPIRICAL**

Definition

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\text{old}}} \left[\mathbb{E}_{y\sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}} \left[\left(\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i} \Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

Different Ways to Compute the Fisher:
(5) Batched approximation of Empirical Fisher

- **Outer expectation:** estimate by averaging over batched version of D_{old}
- **Inner expectation:** approximate using the square of mini-batch averaged gradients w.r.t. ground-truth labels

Most used one!

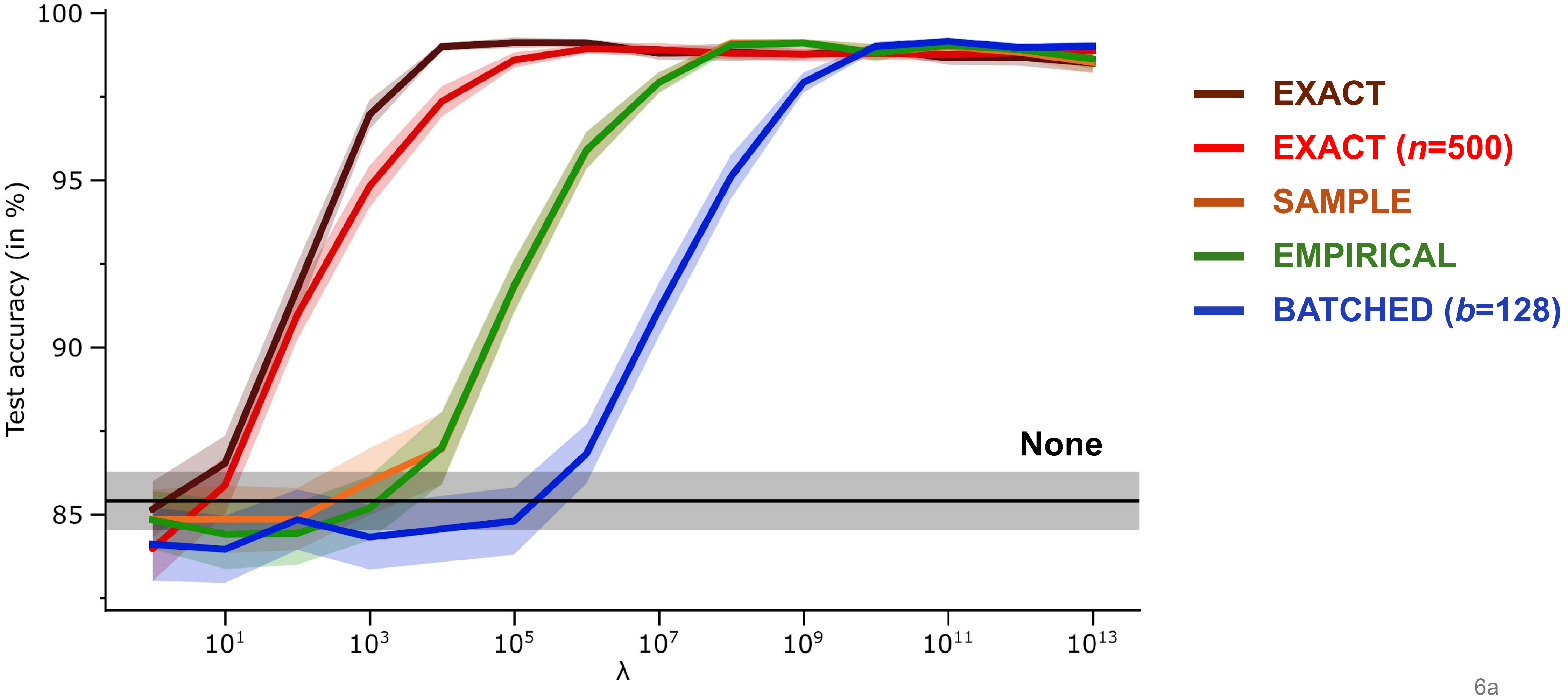
$$F_{\text{old}}^{i,i} = \frac{1}{\left|D_{\text{old}}^{(b)}\right|} \sum_{\mathcal{B} \in D_{\text{old}}^{(b)}} \left(\sum_{(\boldsymbol{x},y) \in \mathcal{B}} \frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2$$

I will explore this option using $b = 128$,
referring to it as **BATCHED (b=128)**

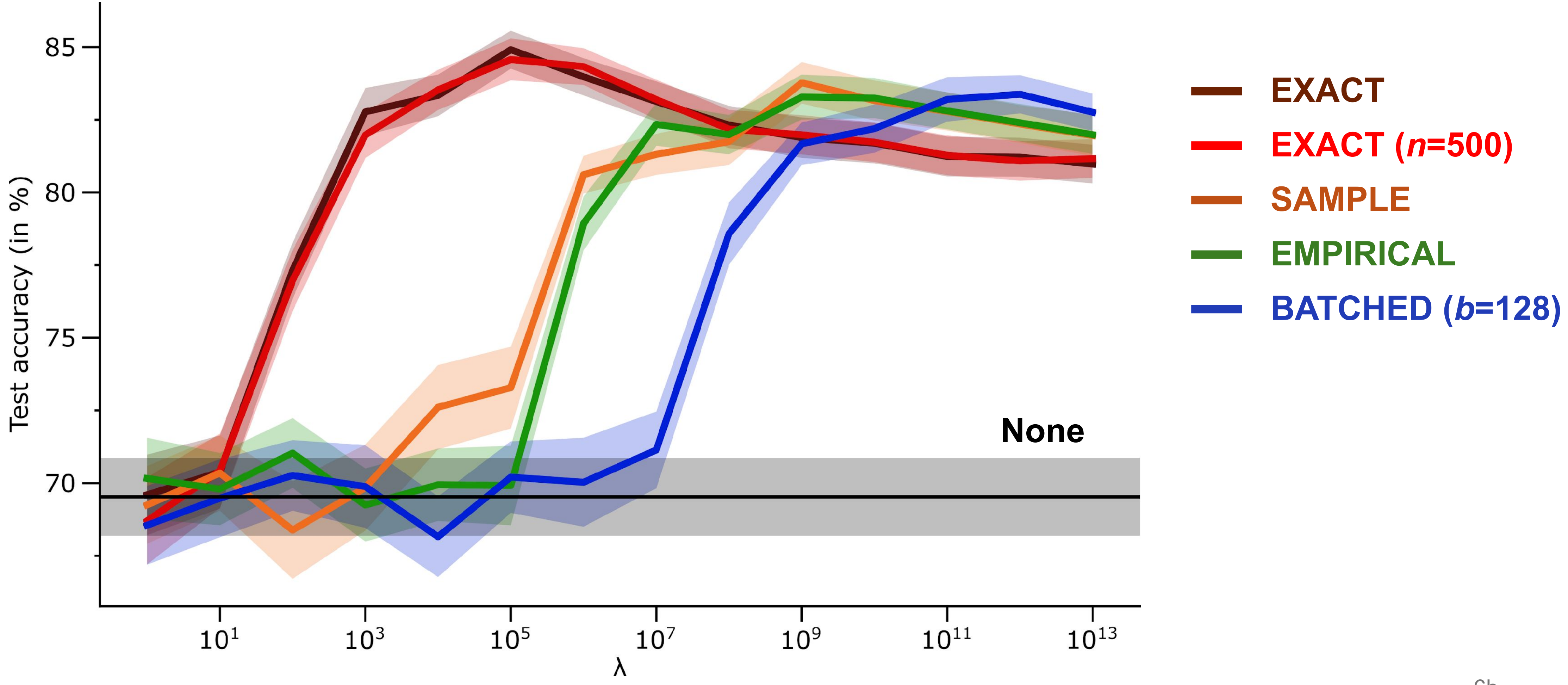
Definition

$$F_{\text{old}}^{i,i} = \mathbb{E}_{\boldsymbol{x} \sim D_{\text{old}}} \left[\mathbb{E}_{y \sim p_{\widehat{\boldsymbol{\theta}}_{\text{old}}}} \left[\left(\frac{\partial \log p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{\partial \theta^i}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{old}}} \right)^2 \right] \right]$$

Empirical Comparisons – Split MNIST



Empirical Comparisons – Split CIFAR-10



Conclusion

- The way in which the Fisher Information is computed can have substantial impact on the performance of EWC

Recommendations

(1) When using EWC, give details of how the Fisher is computed

(2) Do not simply *“use the best performing hyperparameters from another paper”,* if you cannot guarantee that the Fisher is computed in the same way

(3) It might be better to estimate the Fisher with fewer training samples, than to cut corners in some other way