ETH zürich

Department of Computer Science

Unifying Regularisation Methods for Continual Learning

Frederik Benzing

AISTATS 2022, Oral

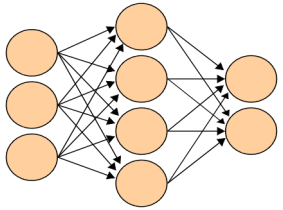
Science for ML

Gain **Empirical & Theoretical Understanding** of **Complex Algorithms**

- Understand
- Simplify & Consolidate
- Improve Algorithms

Continual Learning & Catastrophic Forgetting

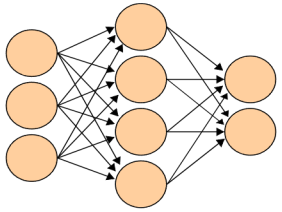
1. **Train to Distinguish Cats and Dogs**



“Cat“

✅

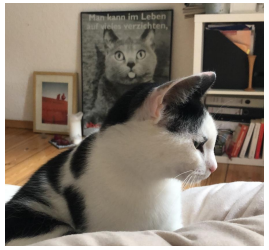
2. **Train to Distinguish Apples and Oranges**

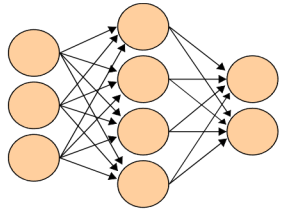


“Orange“

✅

3. **Problem: What is this?**





“Apple“

❌

McCloskey and Cohen, 1989;
Goodfellow et al., 2013

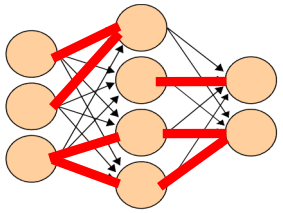
Approaches

- Many different approaches
- i.i.d. Training & Approximations / Relaxations of it
- **Hard Setting: No old data**
  - Regularisation Methods

e.g.
Kirkpatrick et al., 2017;
Chaudhry et al. 2019;
...

Regularisation Methods – EWC *(Kirkpatrick et al., 2017)*

- Parameter Importance



"Cat"

- Auxiliary Loss

$$\sum_{i=1}^{N} \omega_i \underbrace{\left(\theta_i - \theta_i^{(old)}\right)}_{\text{Deviation from old parameter}}^2$$

Importance

See also Li & Hoiem, 2017 for distillation-like regularisation

What does it mean to be important?

- **EWC: Fisher Information** *(Kirkpatrick et al., 2017; Nguyen et al., 2017)*

  (Approximate Laplace Posterior / Second-order Approximation of Loss)

- **SI:** Different Heurisitc Importance *(Zenke et al., 2017)*

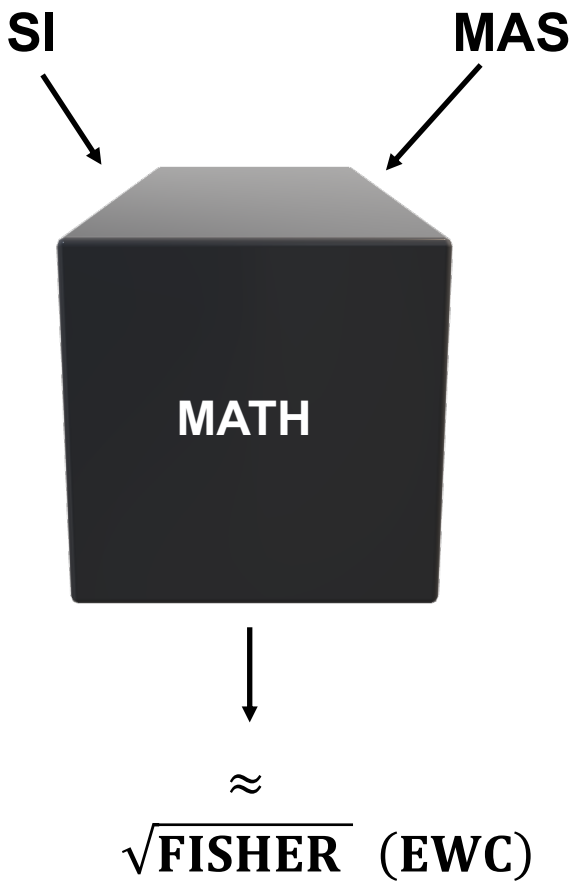
- **MAS:** Different Heuristic Importance *(Aljundi et al., 2018)*

- Follow ups

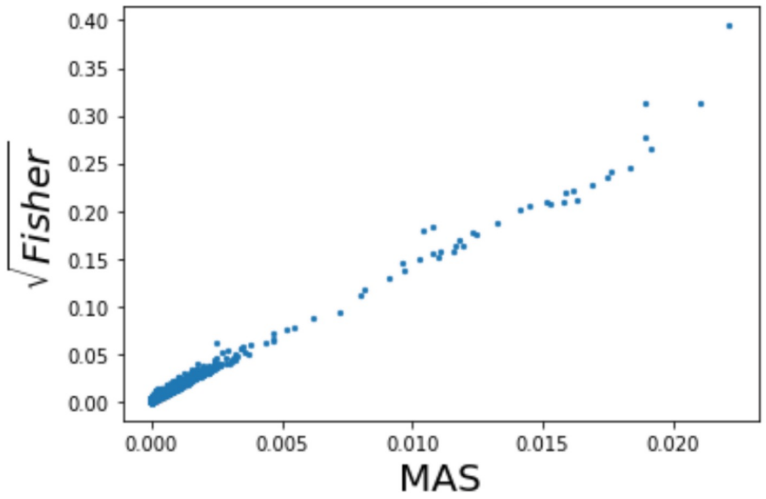
  Nguyen et al., 2017;
  Ritter et al., 2018;
  Chaudhry et al., 2018;
  Schwarz et al., 2018;
  Liu et al., 2018;
  Park et al., 2019;
  Yin et al., 2020

Unify

Theory

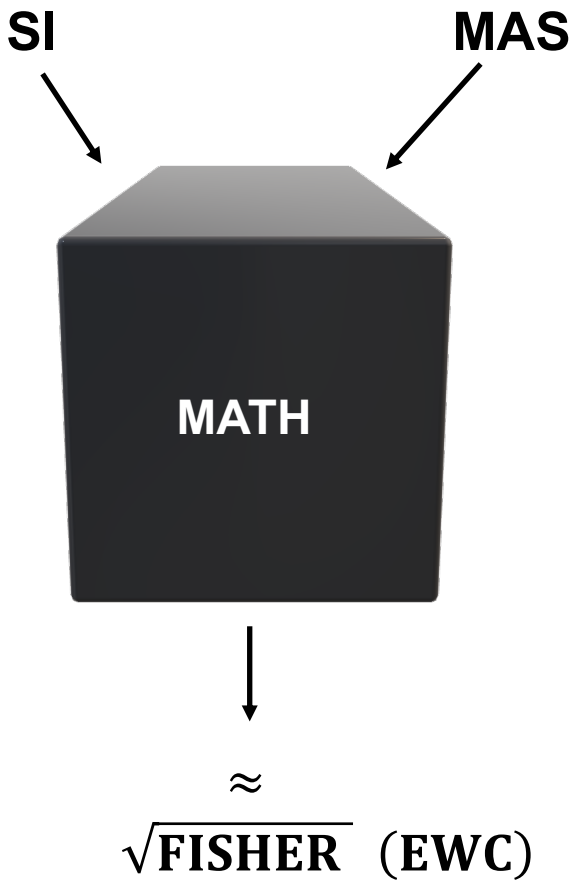


Practice

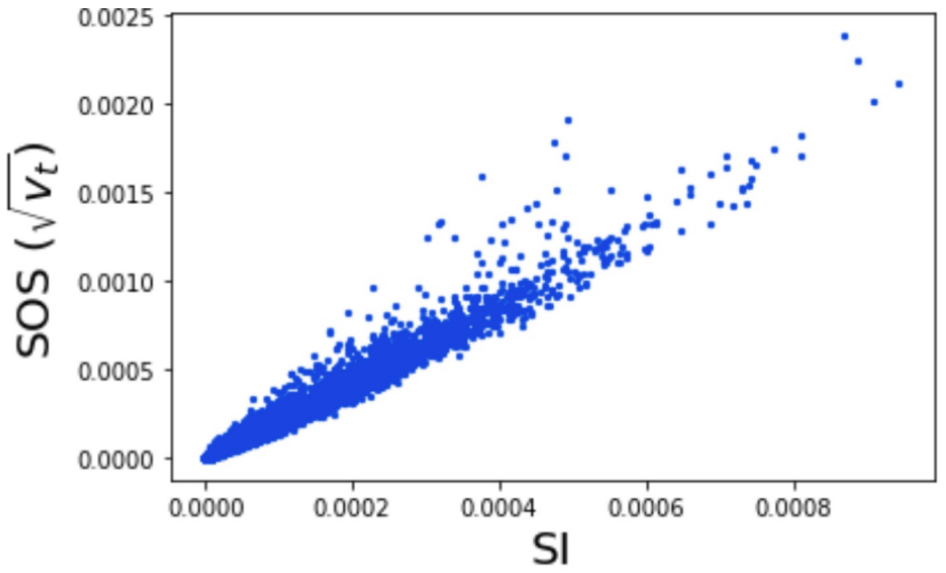


Unify

Theory



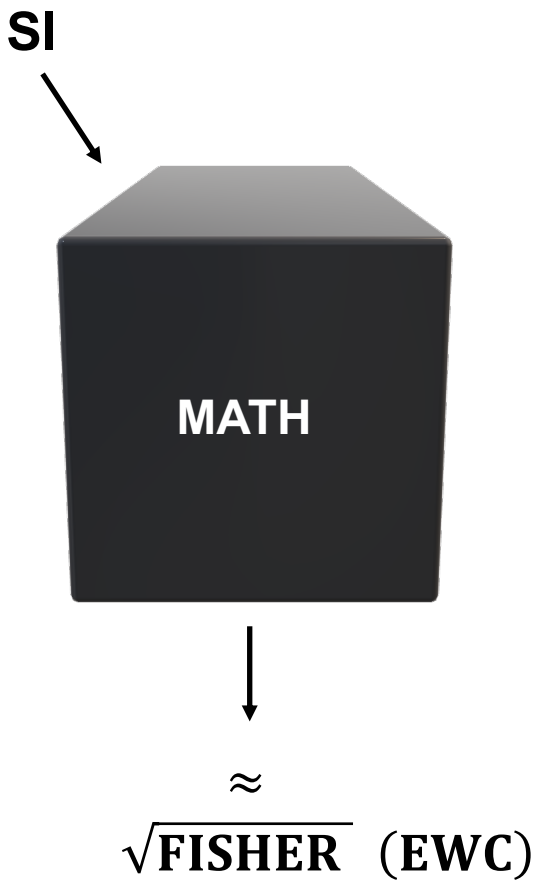
Practice



Unify

- Three key algorithms EWC, SI, MAS rely on same principle
- Theoretical Understanding of SI, MAS
- Unification / Simplification

What is it good for?

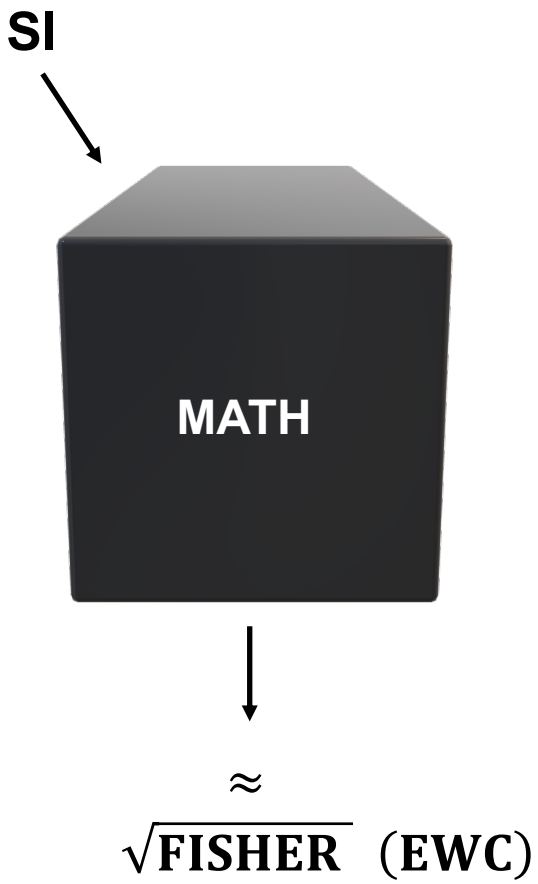


Opening The Black Box

- Approximation Depends on Several Assumptions
  - Batch Size
  - Learning rate schedule
  - Optimizer (Adam/SGD/…)
  - …

Whitfield et al., 1969; Rocel et al., 2015

What is it good for?



Opening The Black Box

- Approximation Depends on Several Assumptions
- Is this bad?
- **No, it allows predictions and improvements!**

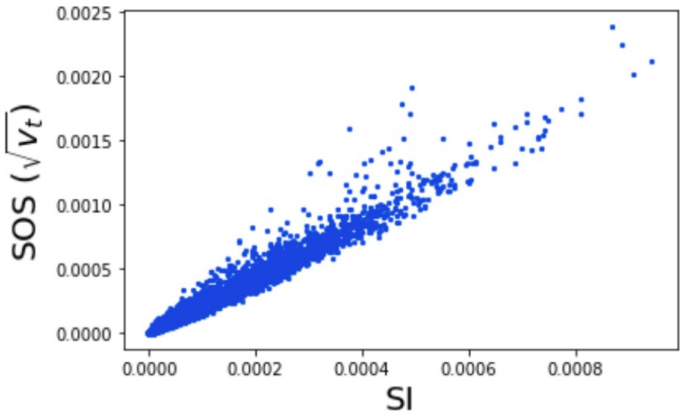
Whitfield et al (1969); Rocel et al (2015)

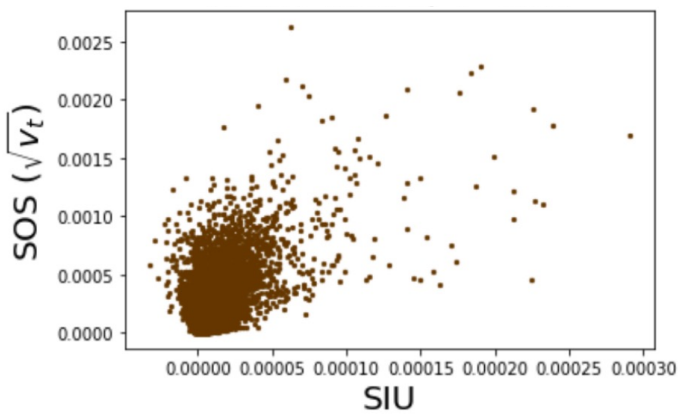
What is it good for? Experiments – SI

- SI has bias in approximation
  - *Removing bias should be good*

Algo.	P-MNIST	CIFAR
SI	97.2±0.1	74.4±0.2
SI Unbiased	96.3±0.1	72.5±0.3

- But: Bias needed for similarity to Fisher
  - *Removing bias is bad*

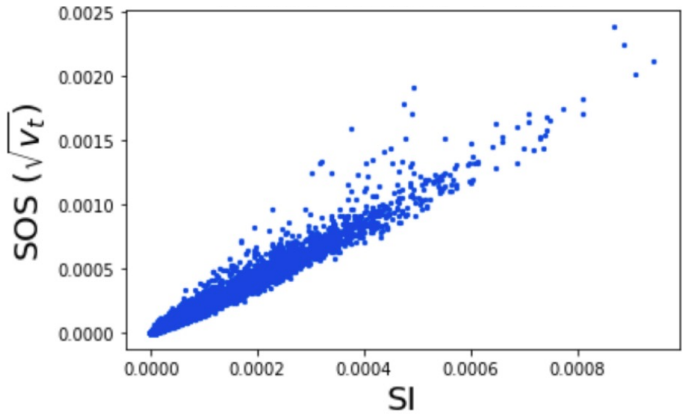


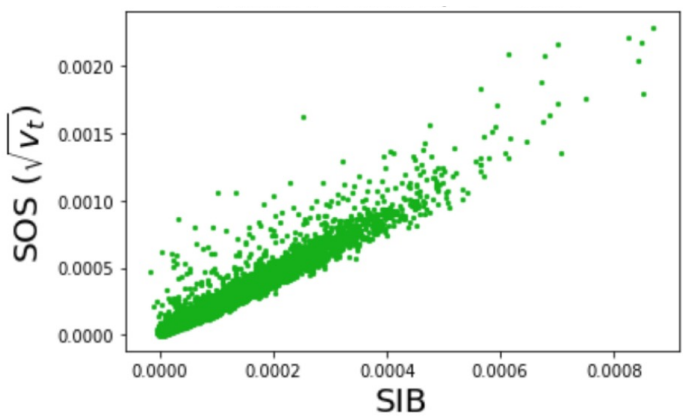


What is it good for? Experiments – SI

- SI has bias in approximation
  - *Removing bias should be good*
- But: Bias needed for similarity to Fisher
  - *Removing bias is bad*

Algo.	P-MNIST	CIFAR
SI	97.2±0.1	74.4±0.2
SI Unbiased	96.3±0.1	72.5±0.3
SI Bias-Only	97.2±0.1	75.1±0.1





What is it good for? Experiments – SI

- Approximation requires small batch size
- Otherwise link to Fisher is weak
- Prediction: Large Batch Size → Bad Performance

Algo.	P-MNIST	CIFAR
SI	97.2 ± 0.1	74.4 ± 0.2
SI(2048)	96.2 ± 0.1	70.0 ± 0.3
SOS(**2048**)	97.1 ± 0.1	74.4 ± 0.1

What is it good for? Experiments – SI

- Many Other Seemingly Small Choices affect SI, and break link to Fisher (e.g. learning rate decay)

Model	**SI**	**SOS**	**EWC***	**MAS***
Small	25.1 ± 4.6	44.3 ± 0.1	45.1	40.6
Base	46.0 ± 0.1	43.3 ± 0.3	42.4	46.9
Wide	40.0 ± 0.2	46.0 ± 0.1	31.1	45.1
Deep	21.6 ± 0.7	30.0 ± 0.1	29.1	33.6

Science for ML

Gain **Empirical & Theoretical Understanding** of **Complex Algorithms**

- Understand
- Simplify & Consolidate
- Improve Algorithms