Continual Learning Through Synaptic Intelligence

Friedemann Zenke, Ben Poole
Surya Ganguli

https://fzenke.net

Joint work with

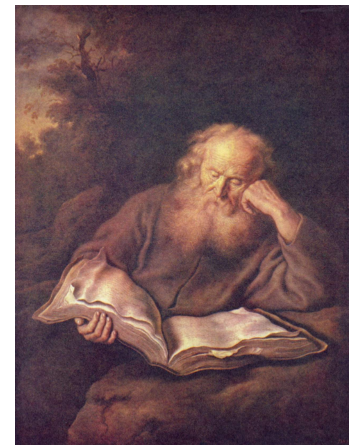
Ben Poole
poole@cs.stanford.edu

Poster: #46

Surya Ganguli
sganguli@stanford.edu

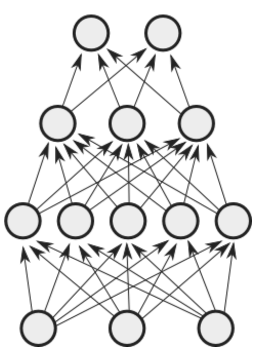
Humans learn continually

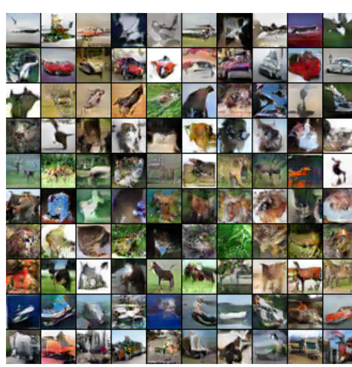
Humans: Continual learning, non stationary data

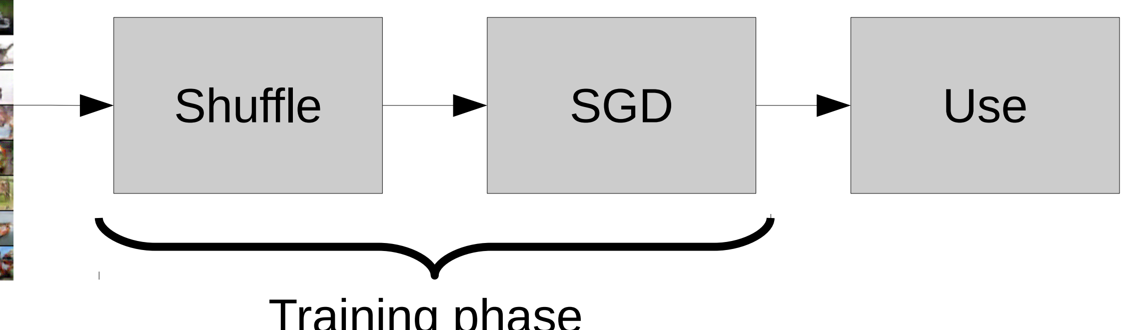


L Muntz, 1898
J-H Fragonard, 1770
S Koninck, 1643

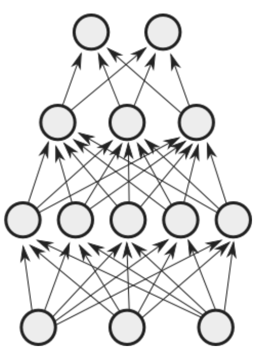
Machines: Training phase, stationary data







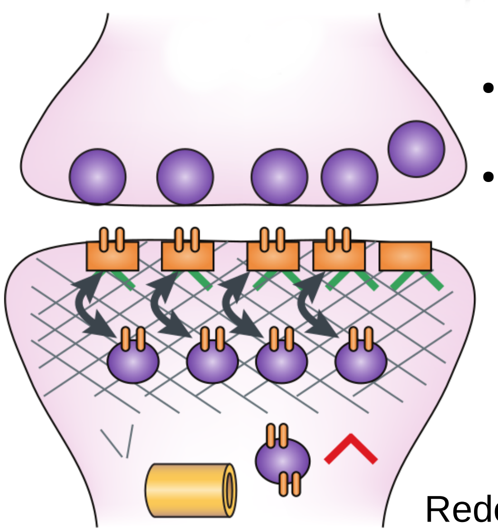
Problem: Catastrophic forgetting



- Non-stationary data
- New memories overwrite old ones
- Capacity not the issue

Synapses, biology's "model parameters", are complex

Biology: Synapse
Complex biochemical dynamical system



- High-dimensional state space
- Non-linear dynamics on different timescales

Redondo & Morris (2010)

Machines: Parameter
Single scalar value

$$w \quad \text{or} \quad \theta, J, \ldots$$

- Individual parameter one-dimensional state space

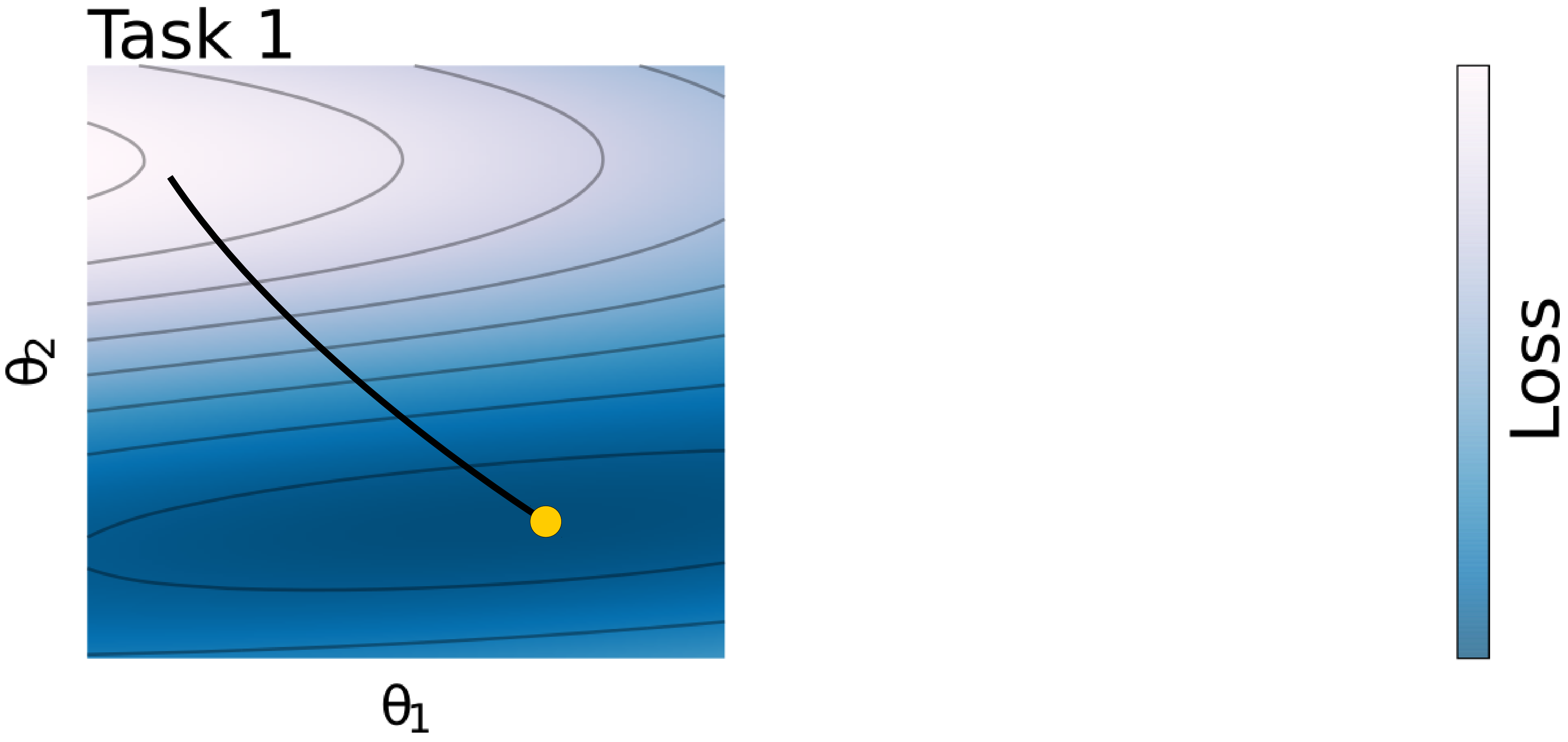
Computational neuroscience
Synaptic complexity is good for continual learning

- Fusi et al. (2005)
- Lahiri & Ganguli (2013)
- Benna & Fusi (2016)

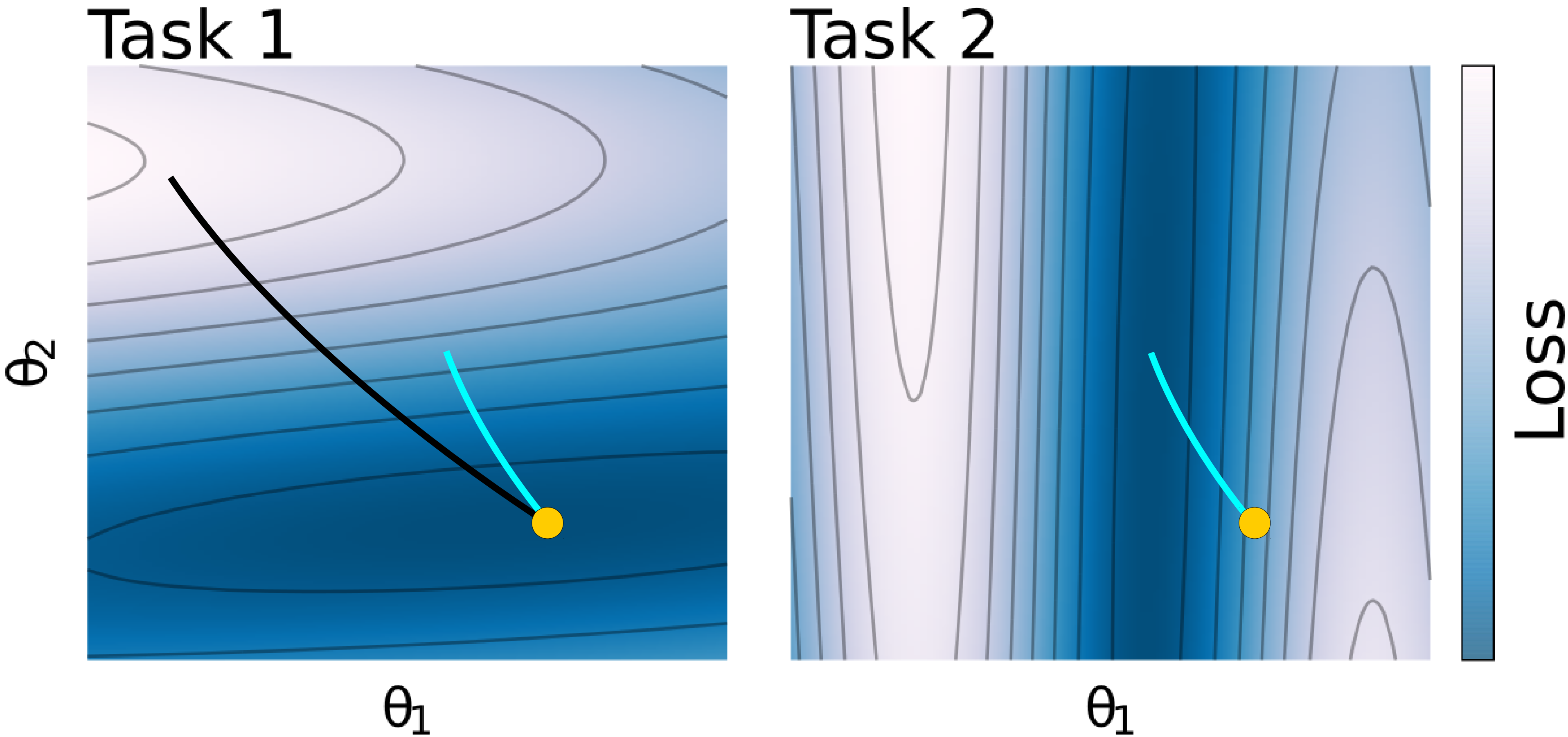
Existing approaches to alleviate catastrophic forgetting

- **Architectural:** Modify architecture
  Use specific nonlinearities (Goodfellow et al., 2013; Srivastava et al., 2013), Progressive Nets (Rusu et al., 2016), Fine tuning (Donahue et al., 2014)
- **Functional:** Regularize activations or outputs of network
  Learning without Forgetting (Li & Hoiem, 2016), Less-forgetting Learning (Jung et al., 2016)
- **Structural:** Regularize parameters of network
  Elastic weight consolidation (Kirkpatrick et al., 2017)

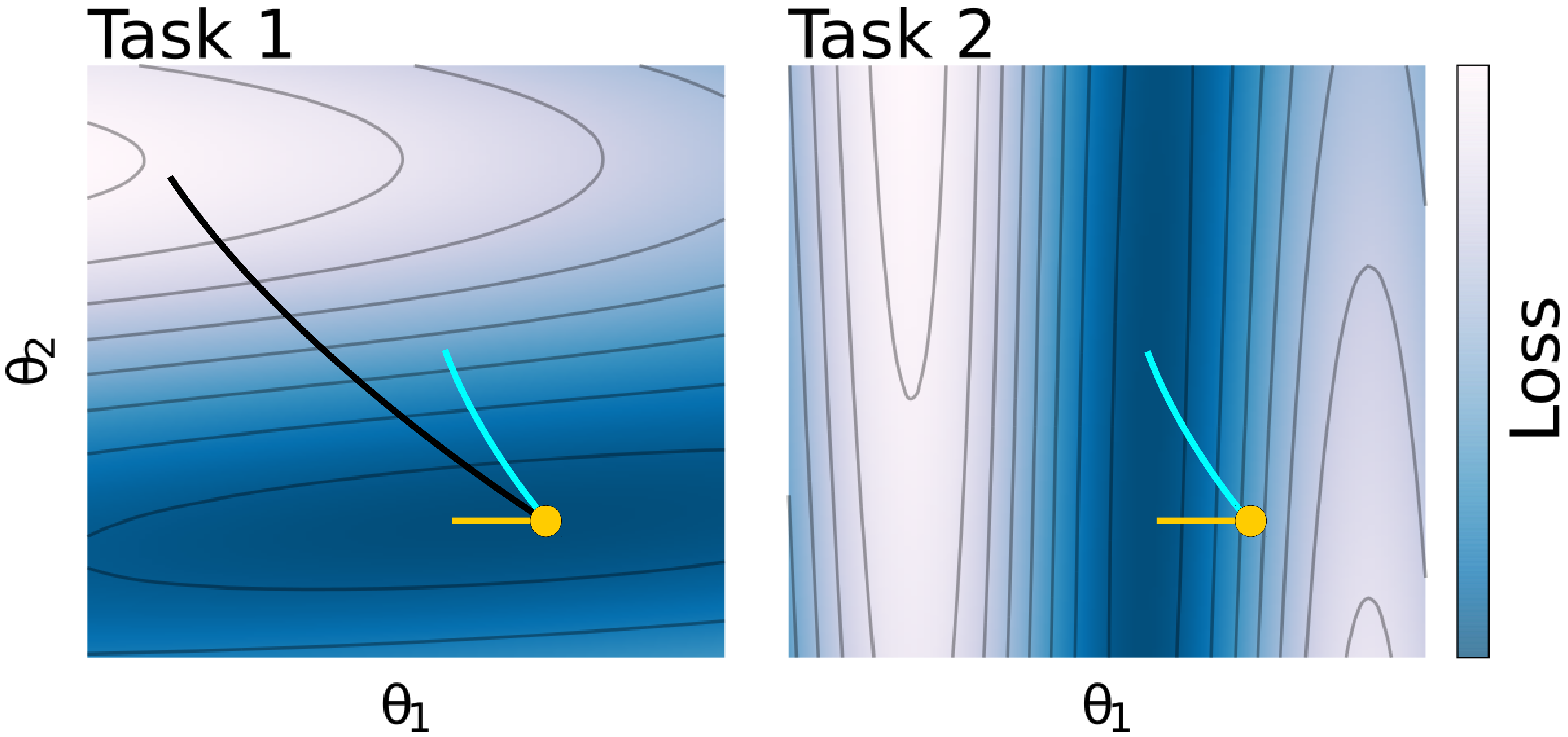
Problem: Catastrophic forgetting



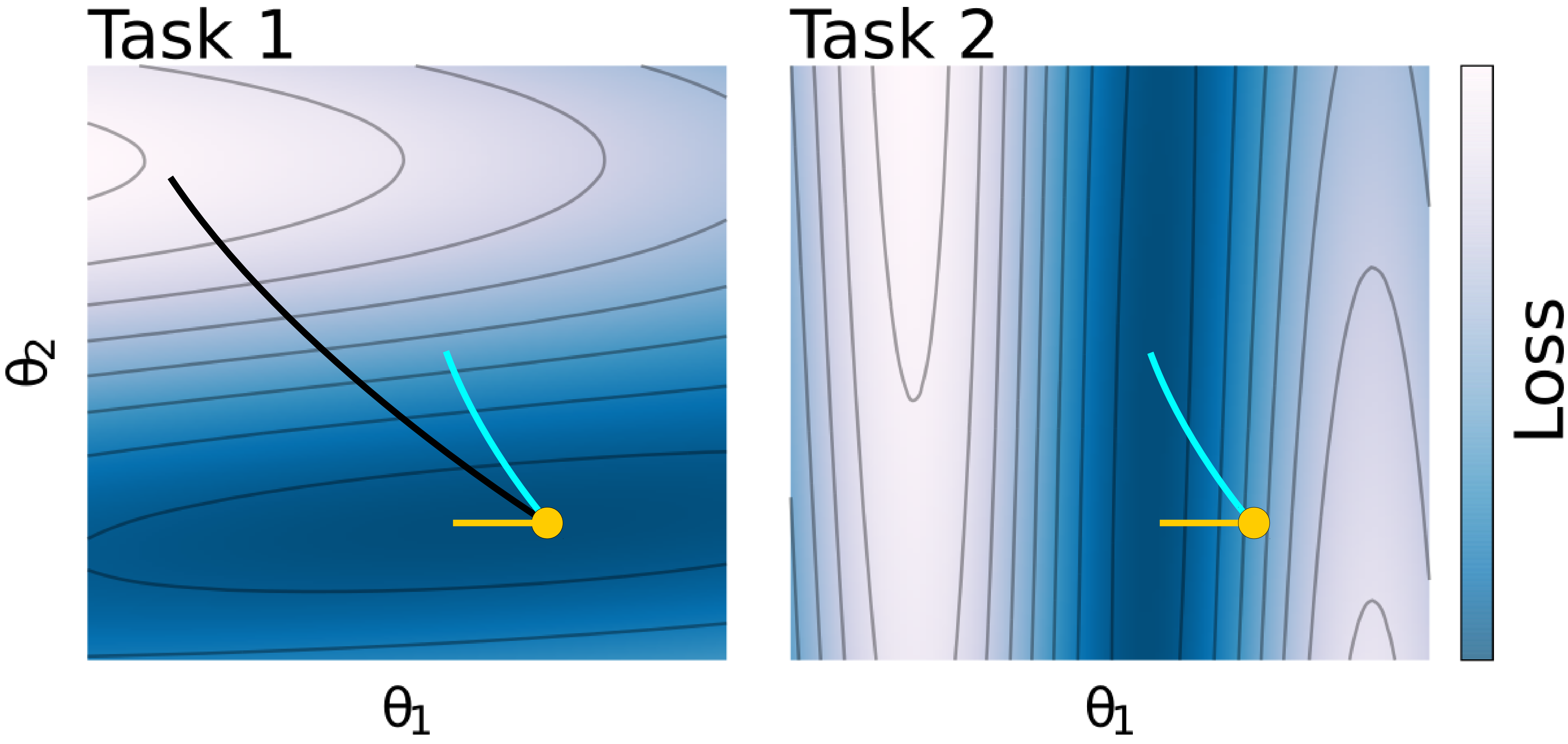
Problem: Catastrophic forgetting



Problem: Catastrophic forgetting



Problem: Catastrophic forgetting



Elastic Weight Consolidation (EWC)

$$L(\theta) = L_2(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta^*_{1,i})^2$$

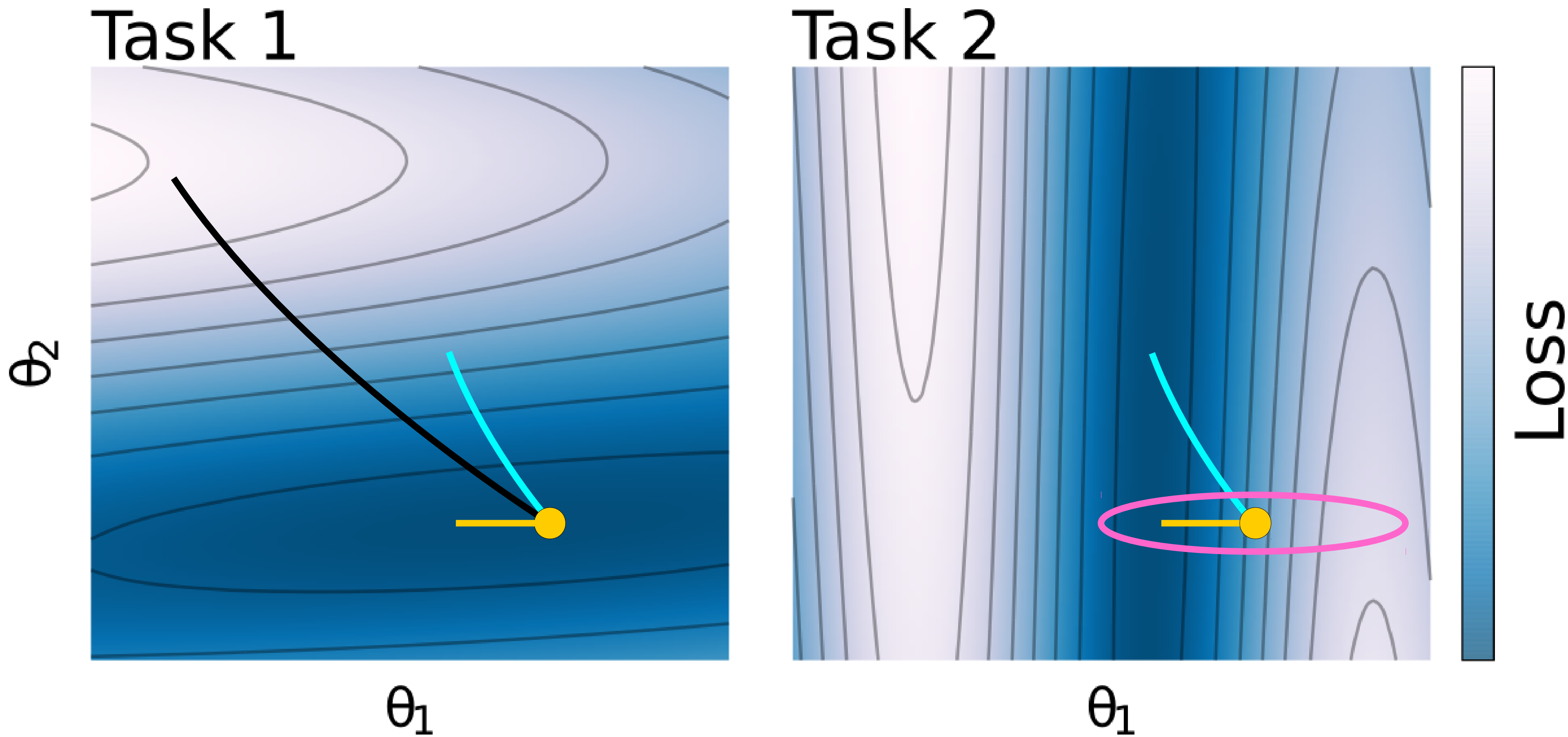
- Idea: Approximate $L_1(\theta)$ with quadratic penalty term
  - Each parameter "remembers" its previous value $\theta^*_{1,i}$
  - … and a local measure of curvature of $L_1(\theta)$

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017)

Elastic Weight Consolidation (EWC)

$$L(\theta) = L_2(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \boxed{\theta^*_{1,i}})^2$$

Diagonal of Fisher at on Task 1

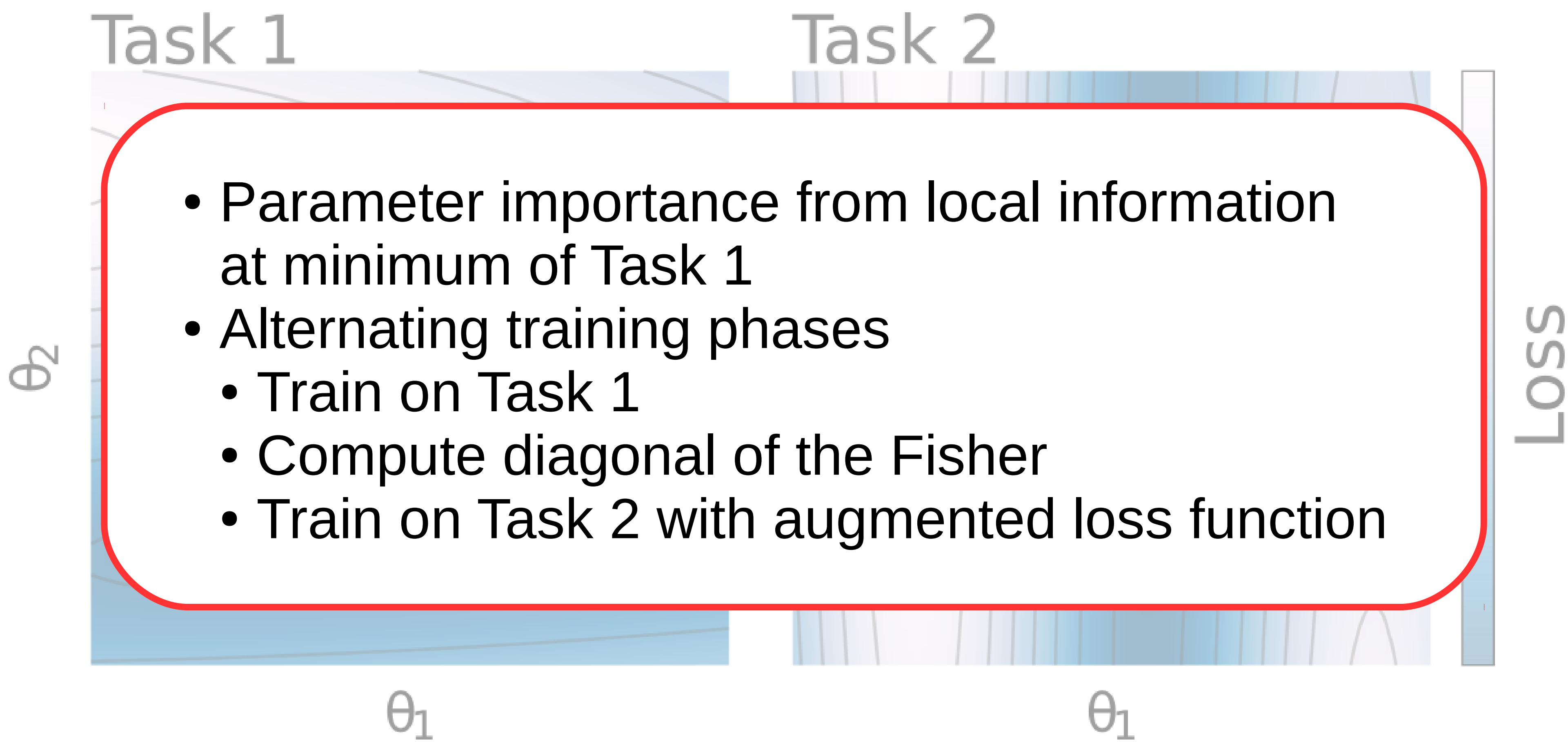


Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017)

Elastic Weight Consolidation (EWC)

$$L(\theta) = L_2(\theta) + \boxed{\sum_i \frac{\lambda}{2} F_i (\theta_i - \boxed{\theta^*_{1,i}})^2}$$

Diagonal of Fisher at



- Parameter importance from local information at minimum of Task 1
- Alternating training phases
  - Train on Task 1
  - Compute diagonal of the Fisher
  - Train on Task 2 with augmented loss function

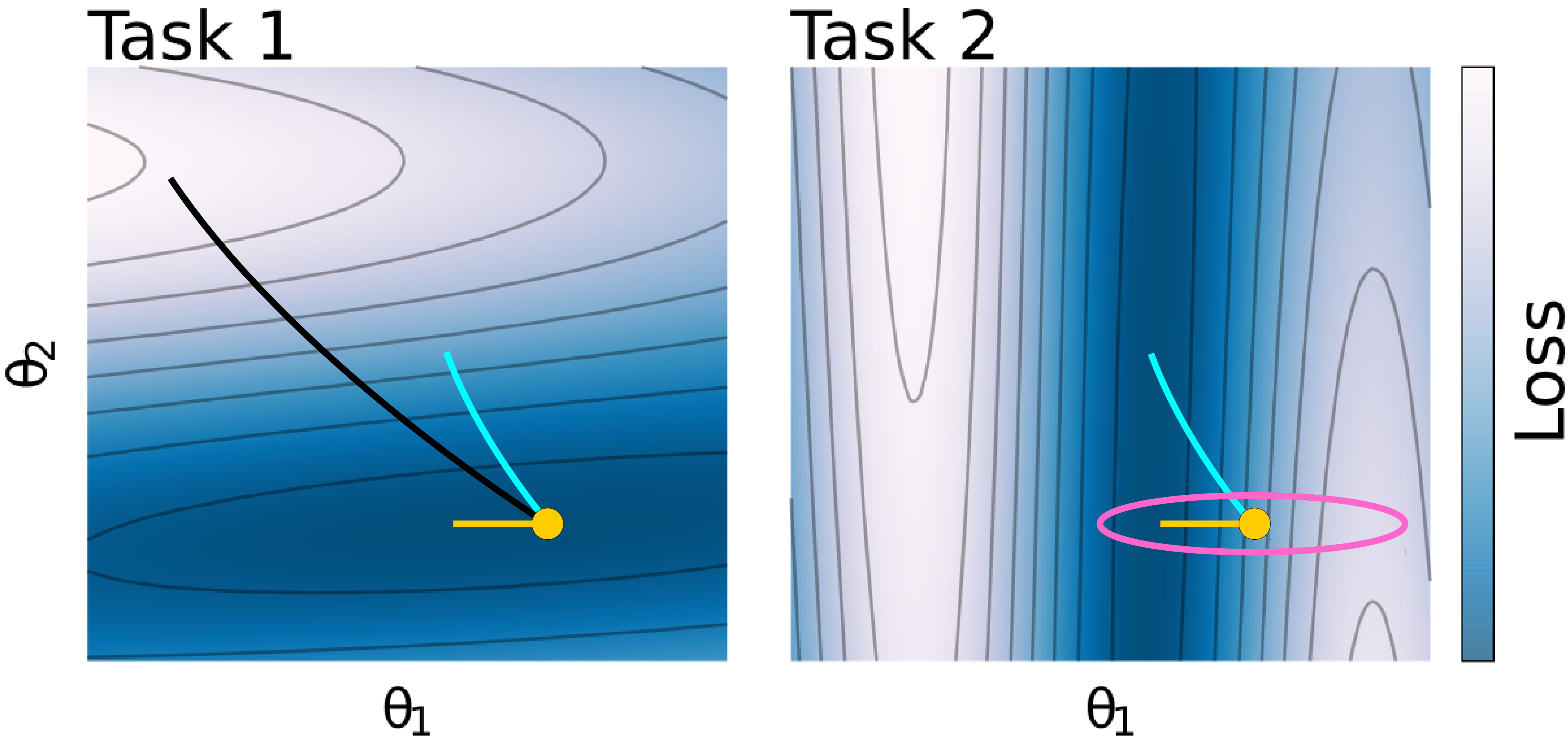
Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017)

Our Contribution

Our approach: Parameter importance on-line from learning trajectory

$$L(\theta) = L_2(\theta) + \boxed{c \sum_i \Omega_i \left(\theta_i - \boxed{\theta^*_{1,i}}\right)^2}$$

From learning trajectory



Total change in loss is given by the path integral over the gradient field

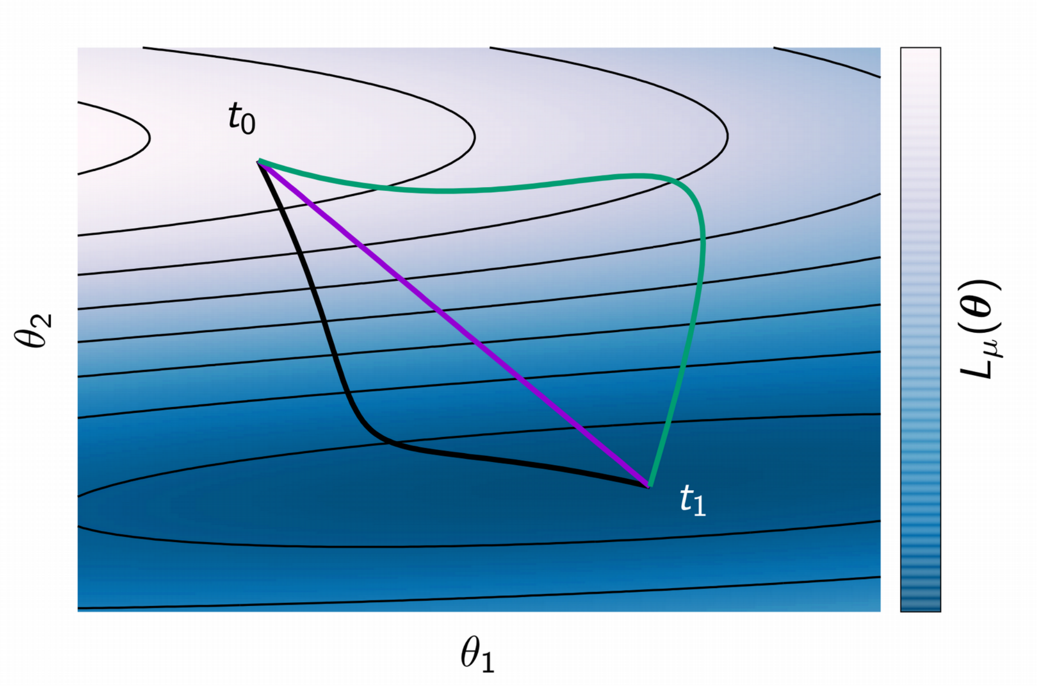
$$\int_C \boldsymbol{g}(\boldsymbol{\theta}(t))d\boldsymbol{\theta} = \int_{t_0}^{t_1} \boldsymbol{g}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{\theta}'(t)dt \ = L(t_1) - L(t_0)$$

$$= \sum_k \int_{t_0}^{t_1} g_k(t)\theta_k'(t)dt \equiv -\sum_k \omega_k$$

$\boldsymbol{g}$: Gradient

$\boldsymbol{\theta}$: Parameters

$\boldsymbol{\theta'}$: Updates



Total change in loss is given by the path integral over the gradient field

$$\int_C \boldsymbol{g}(\boldsymbol{\theta}(t))d\boldsymbol{\theta} = \int_{t_0}^{t_1} \boldsymbol{g}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{\theta}'(t)dt = L(t_1) - L(t_0)$$

$$= \sum_k \underbrace{\int_{t_0}^{t_1} g_k(t)\theta'_k(t)dt}_{} \equiv -\sum_k \omega_k$$

- Is a parameter-specific quantity
- Can be computed on-line during training (running sum)

Natural way of assigning credit for a global change to local parameters

$$L(t_1) - L(t_0) = -\sum_k \omega_k^\mu$$

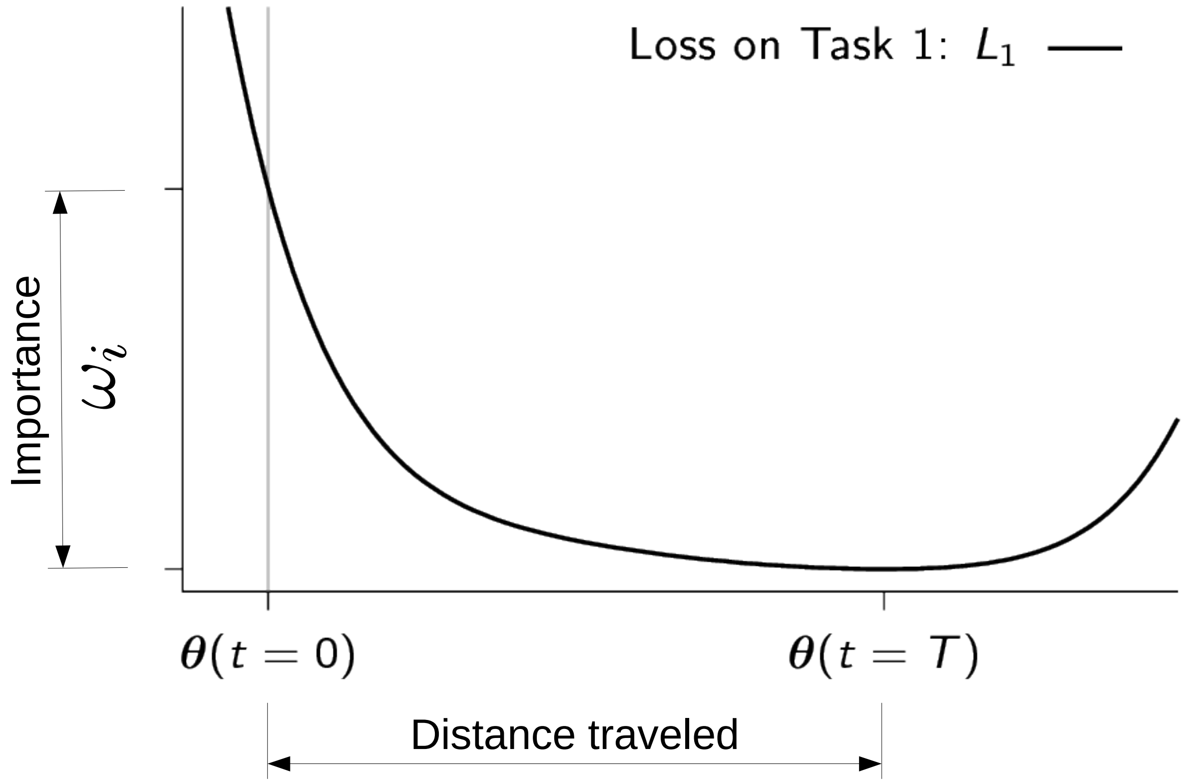
$\boldsymbol{g}$: Gradient

$\boldsymbol{\theta}$: Parameters

$\boldsymbol{\theta}'$: Updates

Leveraging per-parameter importance for continual learning

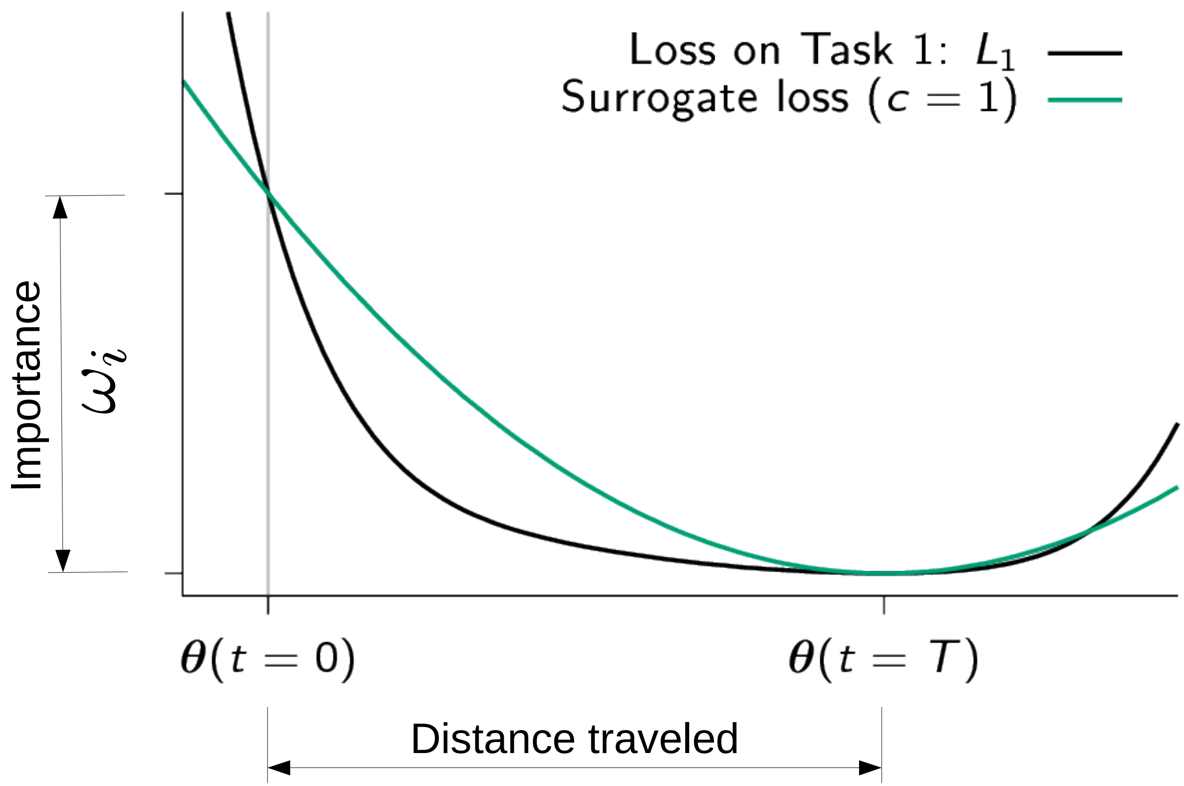
$$L(\theta) = L_2(\theta) + c\sum_i \Omega_i \left(\theta_i - \theta^*_{1,i}\right)^2$$



Leveraging per-parameter importance for continual learning

$$L(\theta) = L_2(\theta) + c \sum_i \Omega_i \left(\theta_i - \theta^*_{1,i}\right)^2$$

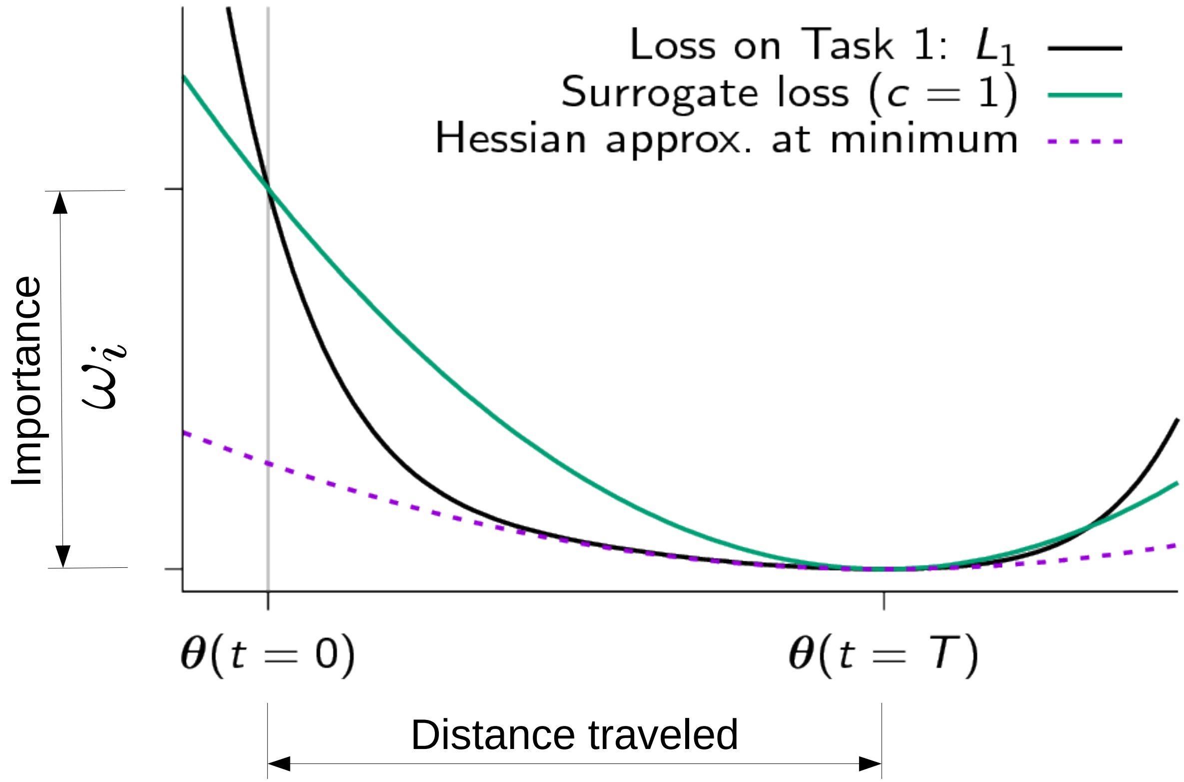
$$\boxed{\Omega_i \equiv \frac{\omega_i}{(\Delta_i)^2 + \epsilon}}$$



Leveraging per-parameter importance for continual learning

$$L(\theta) = L_2(\theta) + c \sum_i \Omega_i \left(\theta_i - \theta^*_{1,i}\right)^2$$

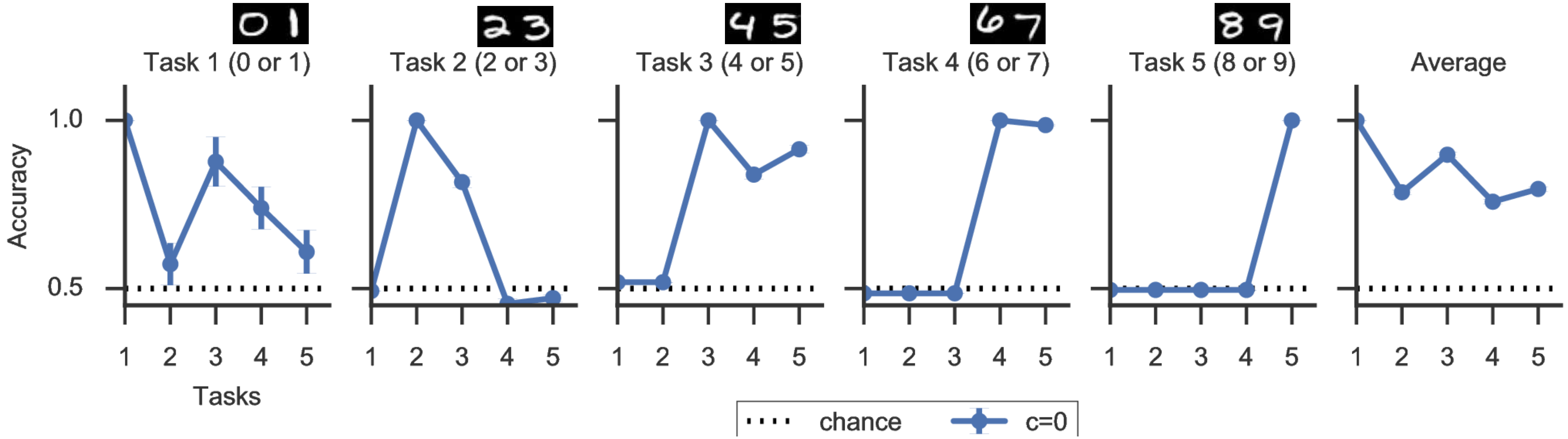
$$\boxed{\Omega_i \equiv \frac{\omega_i}{(\Delta_i)^2 + \epsilon}}$$



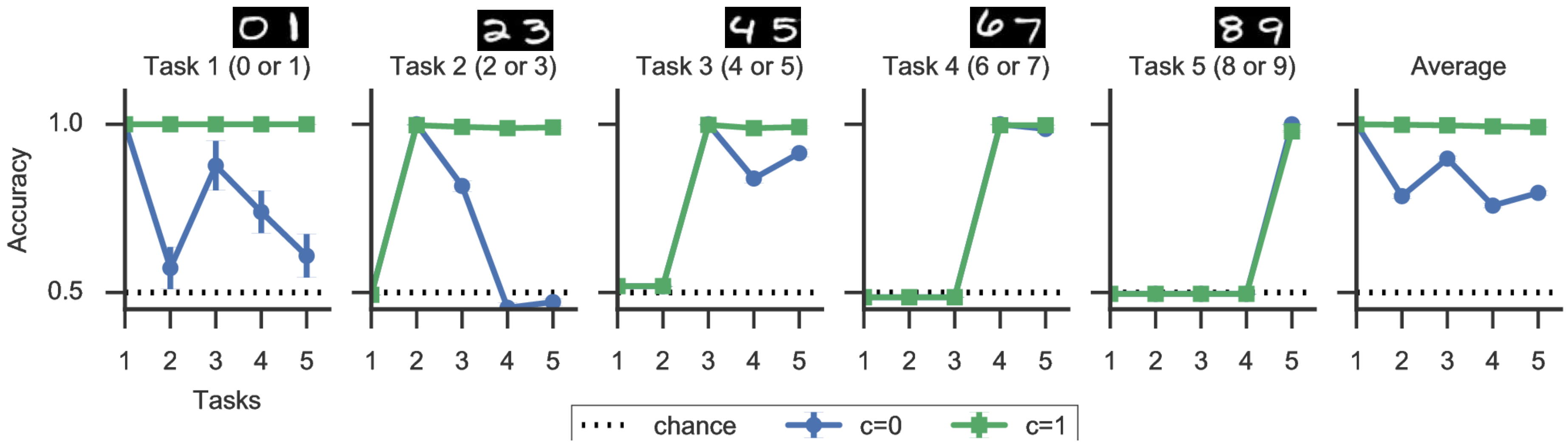
- Surrogate loss
  - Per-parameter importance
  - Distance traveled
- Different from local approximation
- Recovers Hessian for simple quadratic problems

Experiments

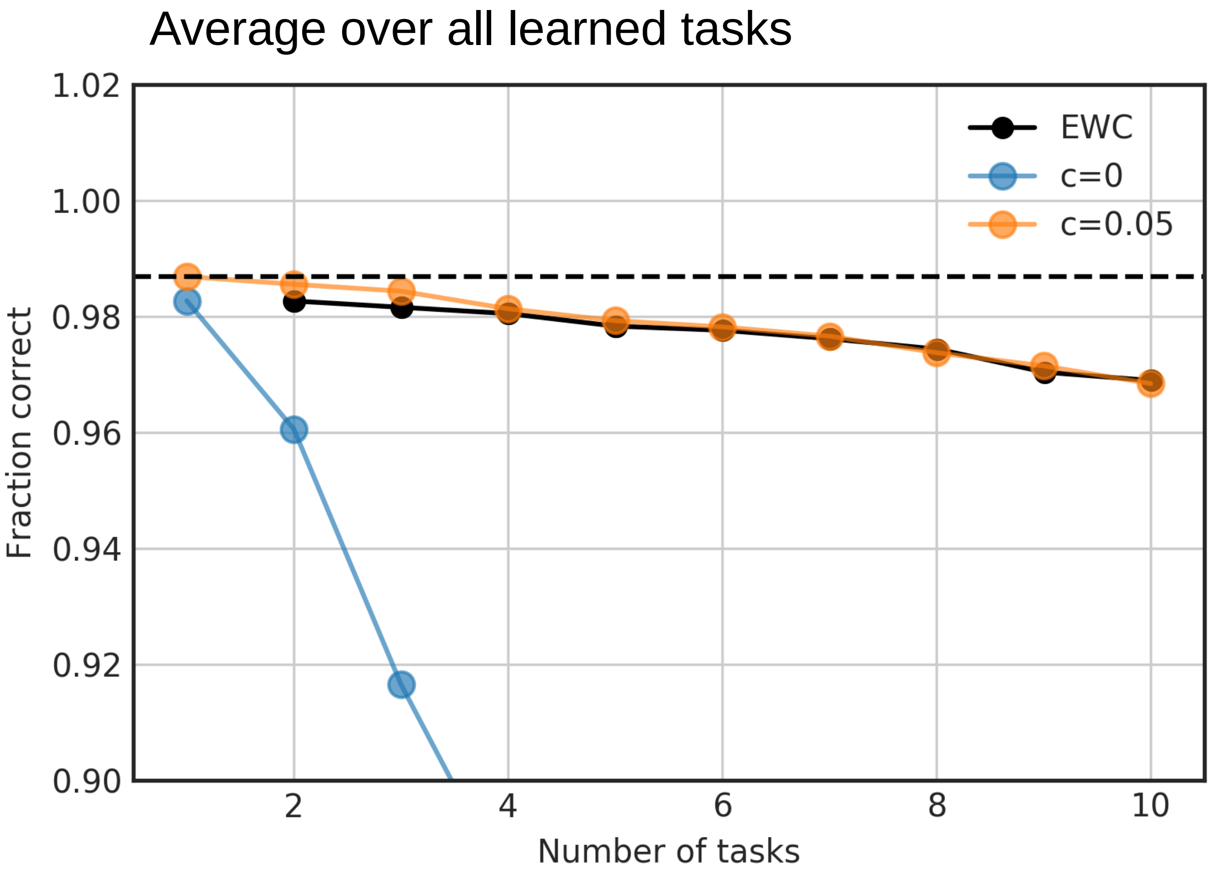
Catastrophic forgetting (split MNIST)



Catastrophic forgetting (split MNIST)



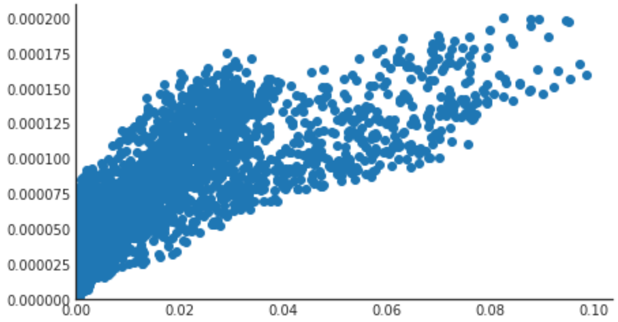
Permuted MNIST



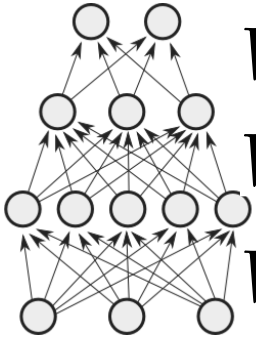
Fisher and our importance measure are correlated

W_1

ω_k



F_k

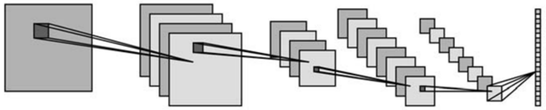


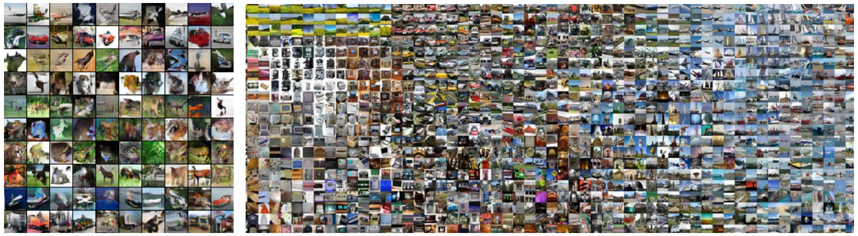
W_{out}

W_2

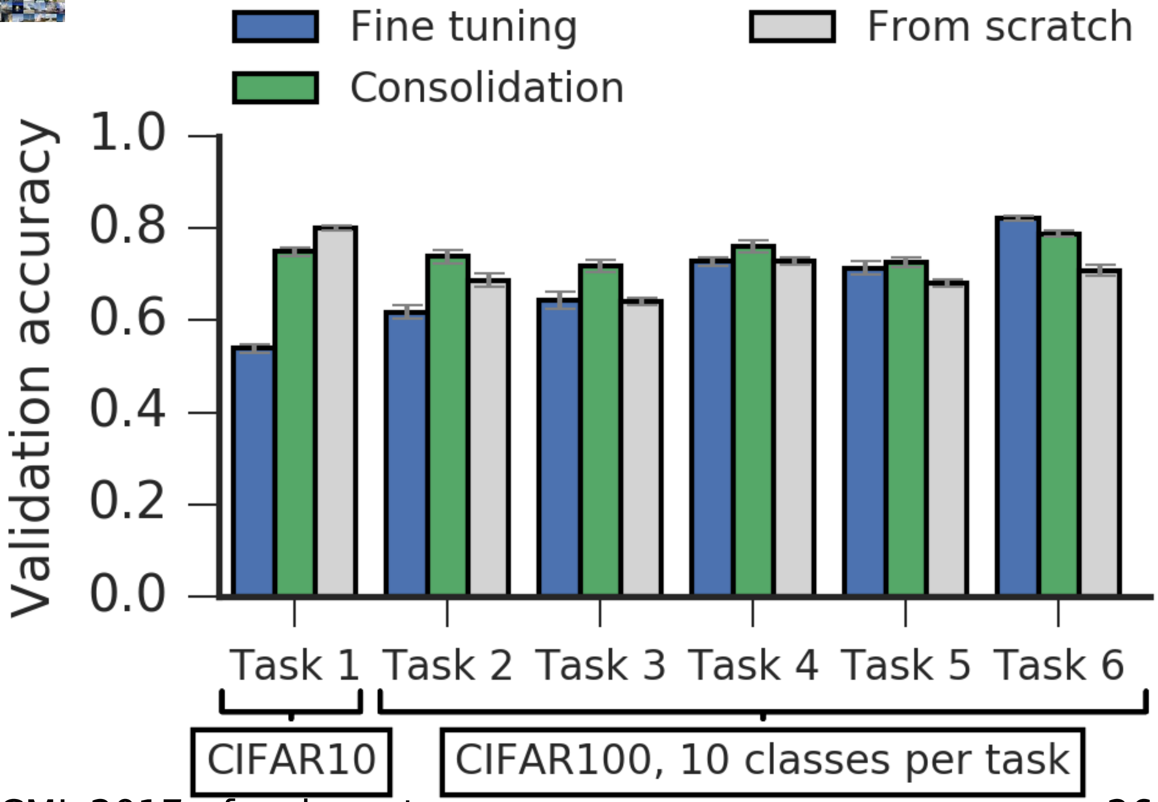
W_1

Works for CNNs: CIFAR10/100





After training on all tasks



Summary

- Individual synapses can estimate their importance as contribution to changes in loss
- They can do this on-line by efficiently computing the path integral over the entire parameter trajectory
- Exploiting this information intelligently
  - Alleviates catastrophic forgetting
  - Yields better generalization

Thanks

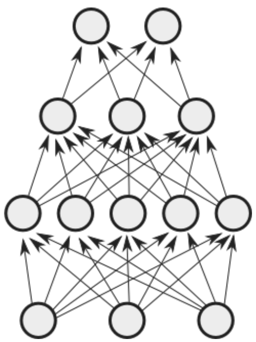
Poster: #46

Code: https://github.com/ganguli-lab/pathint

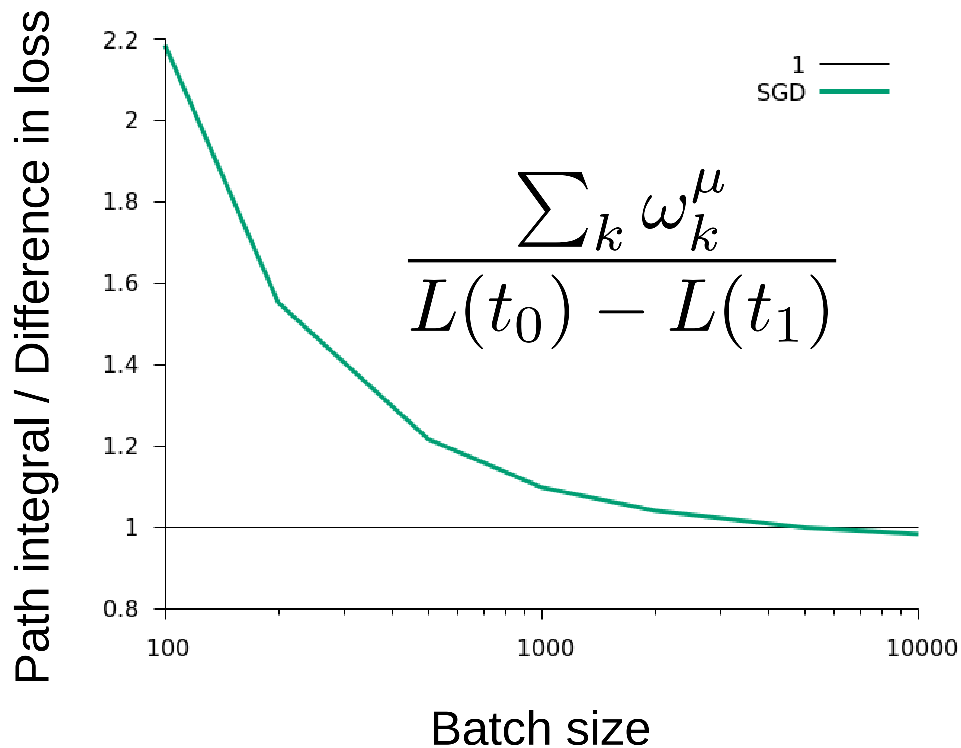
Funding: FNSNF Swiss National Science Foundation

wellcome

Sanity check: $-\sum_k \omega_k^\mu$ corresponds to difference in loss

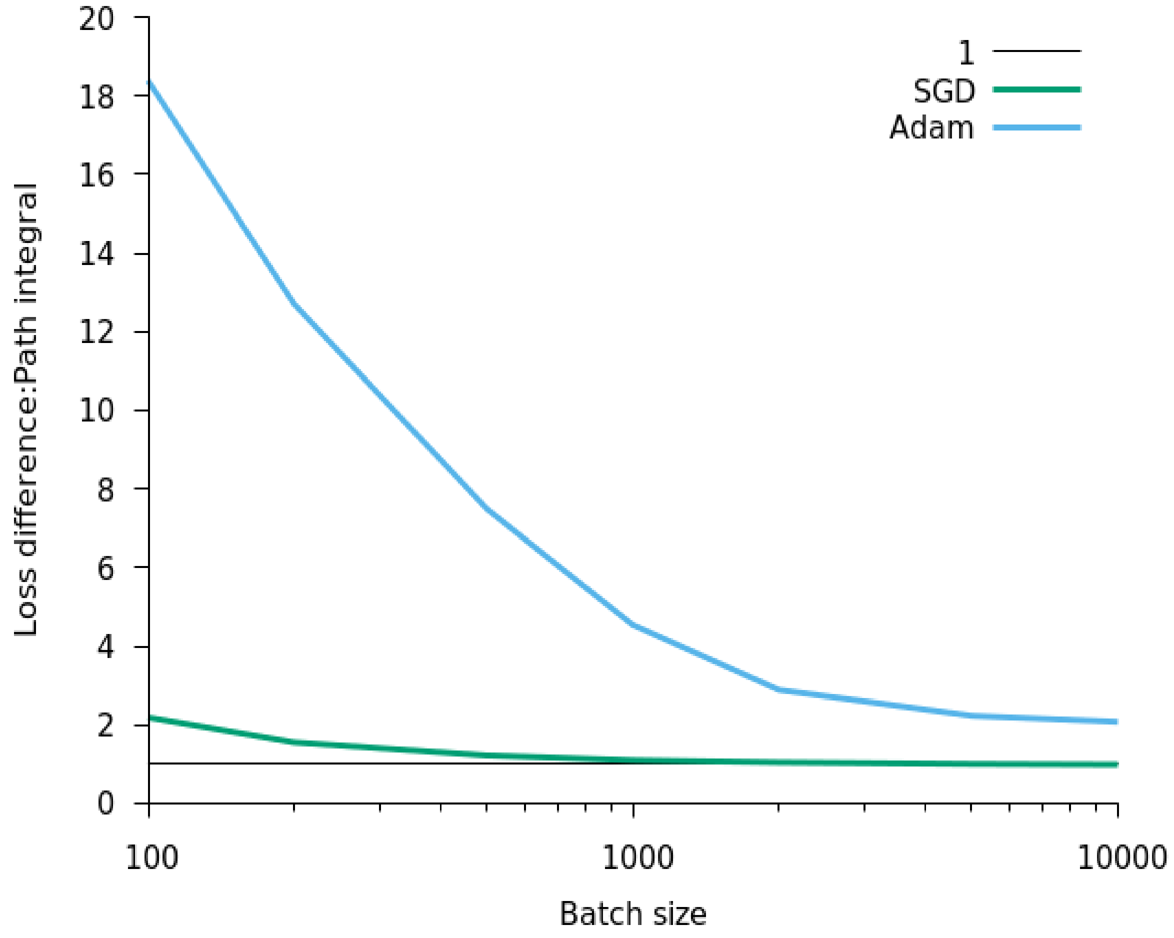


- MLP, MNIST
- 5.6M parameters

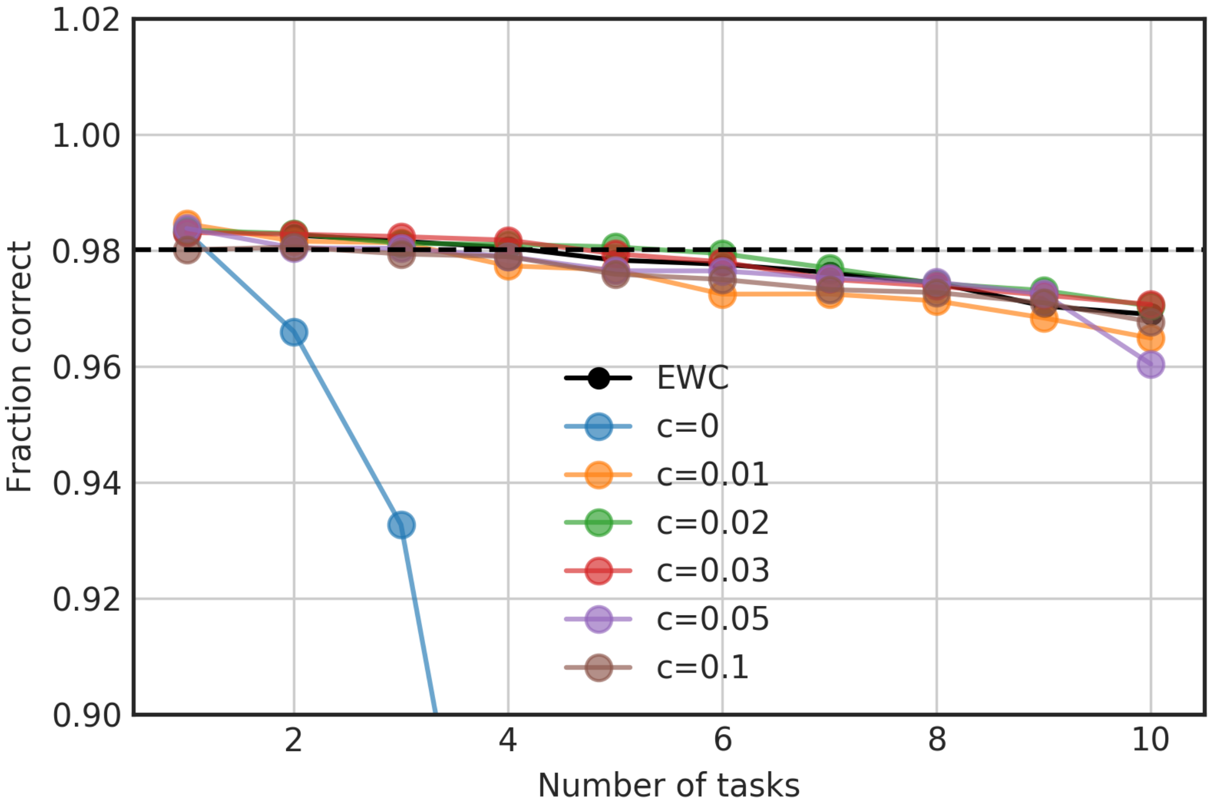


- Deviation due to noise from SGD
- Can be corrected with a multiplicative correction

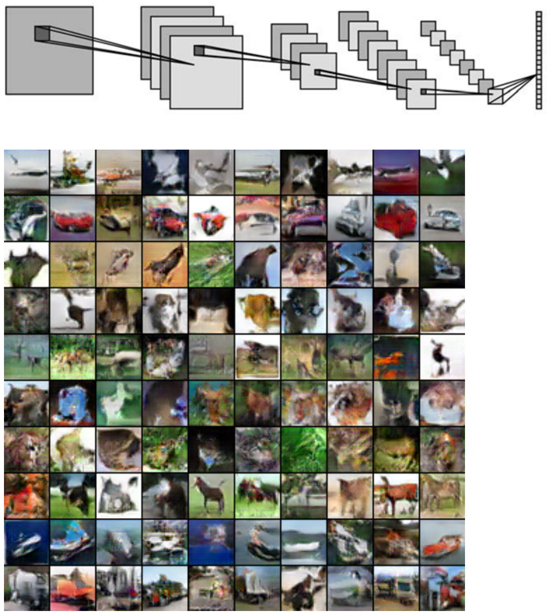
How well is the path integral approximated by SGD&Adam



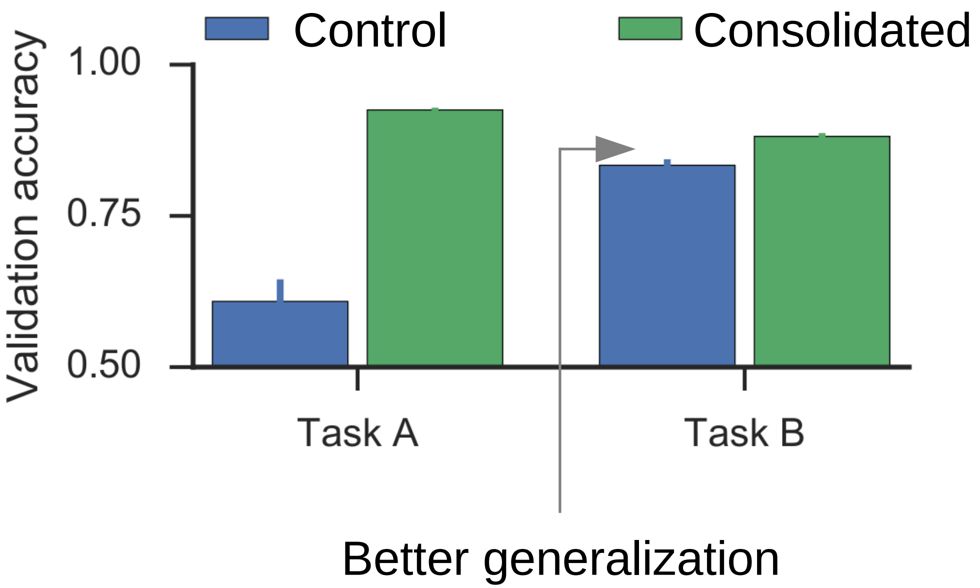
Permuted MNIST



Works for CNNs: Split CIFAR10

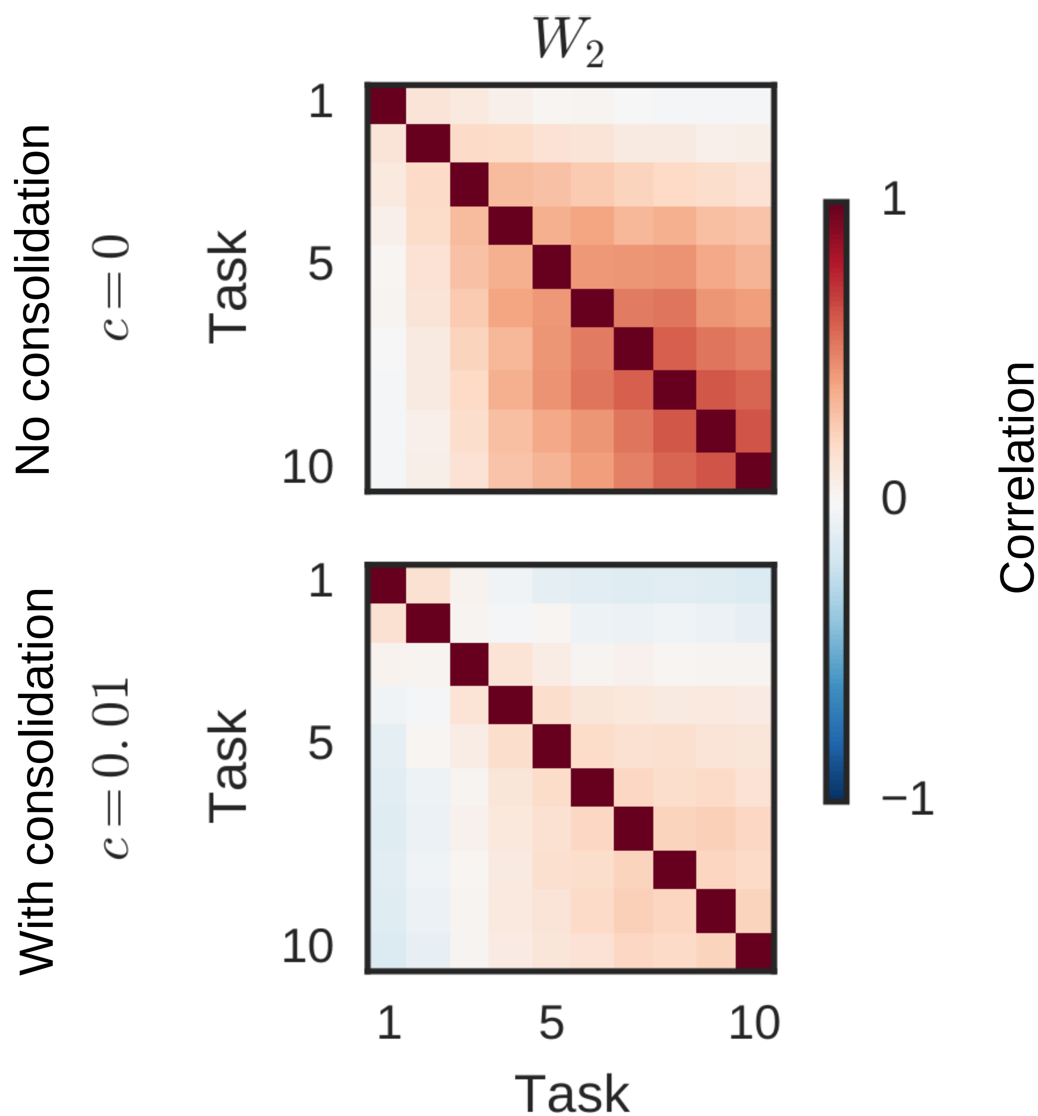


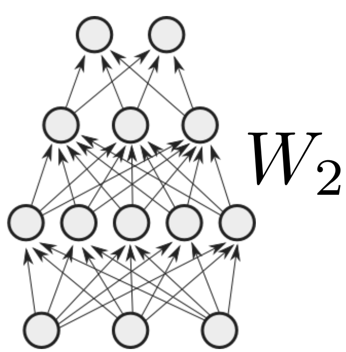
After training on Task A & B:



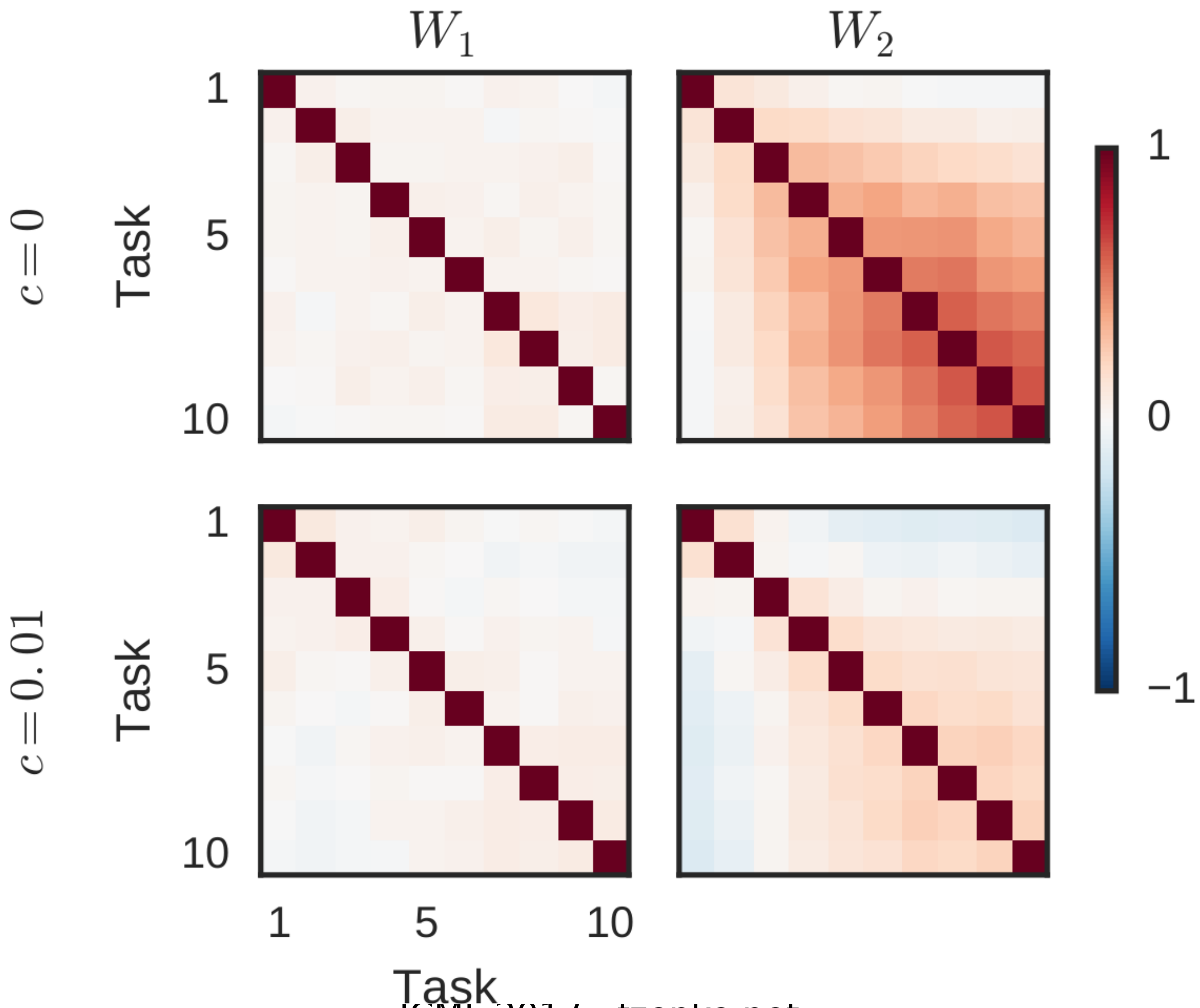
Task importance less correlated in hidden layers

Correlation matrix of the ω_k^μ

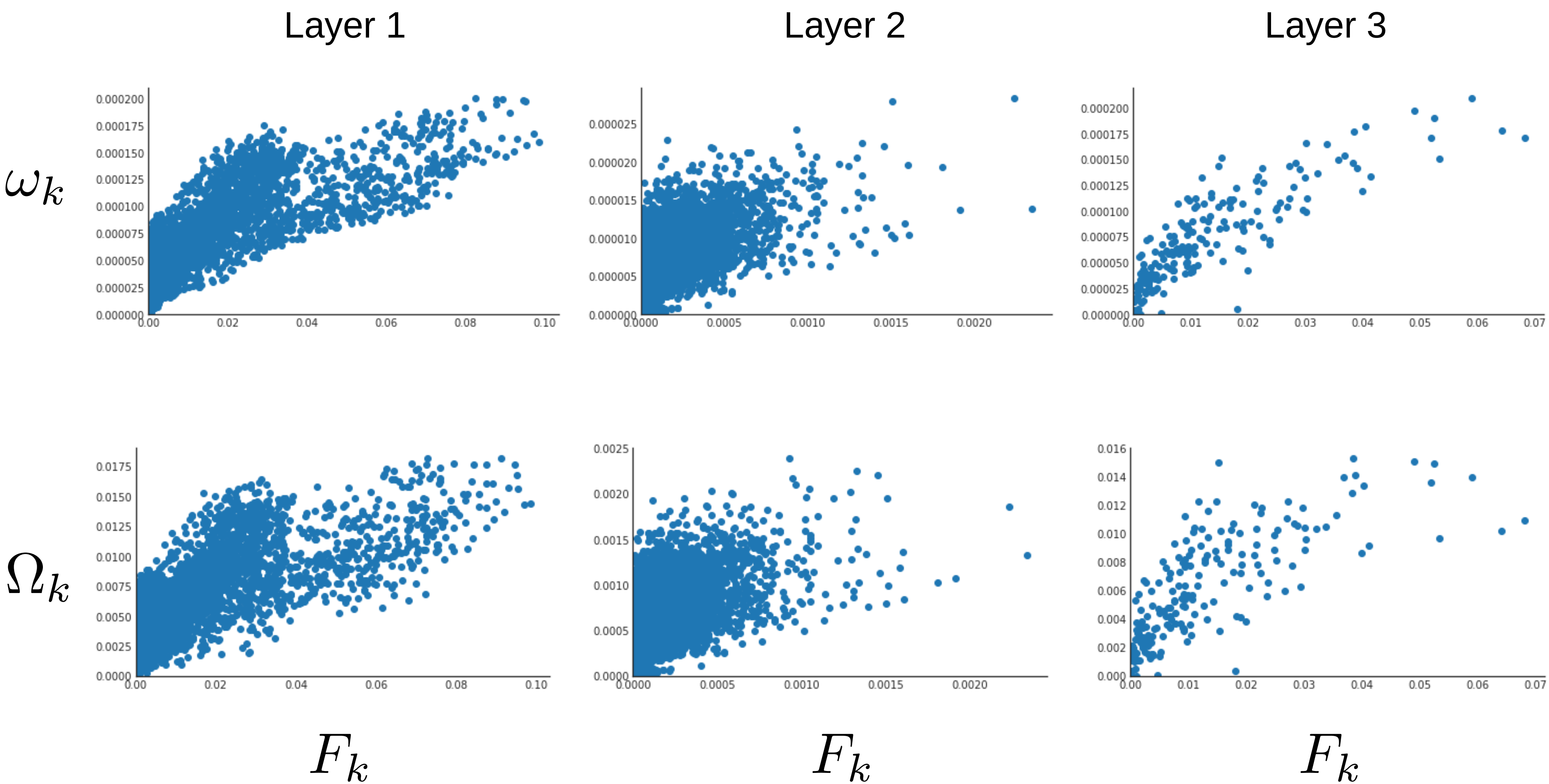




Task importance less correlated in hidden layers



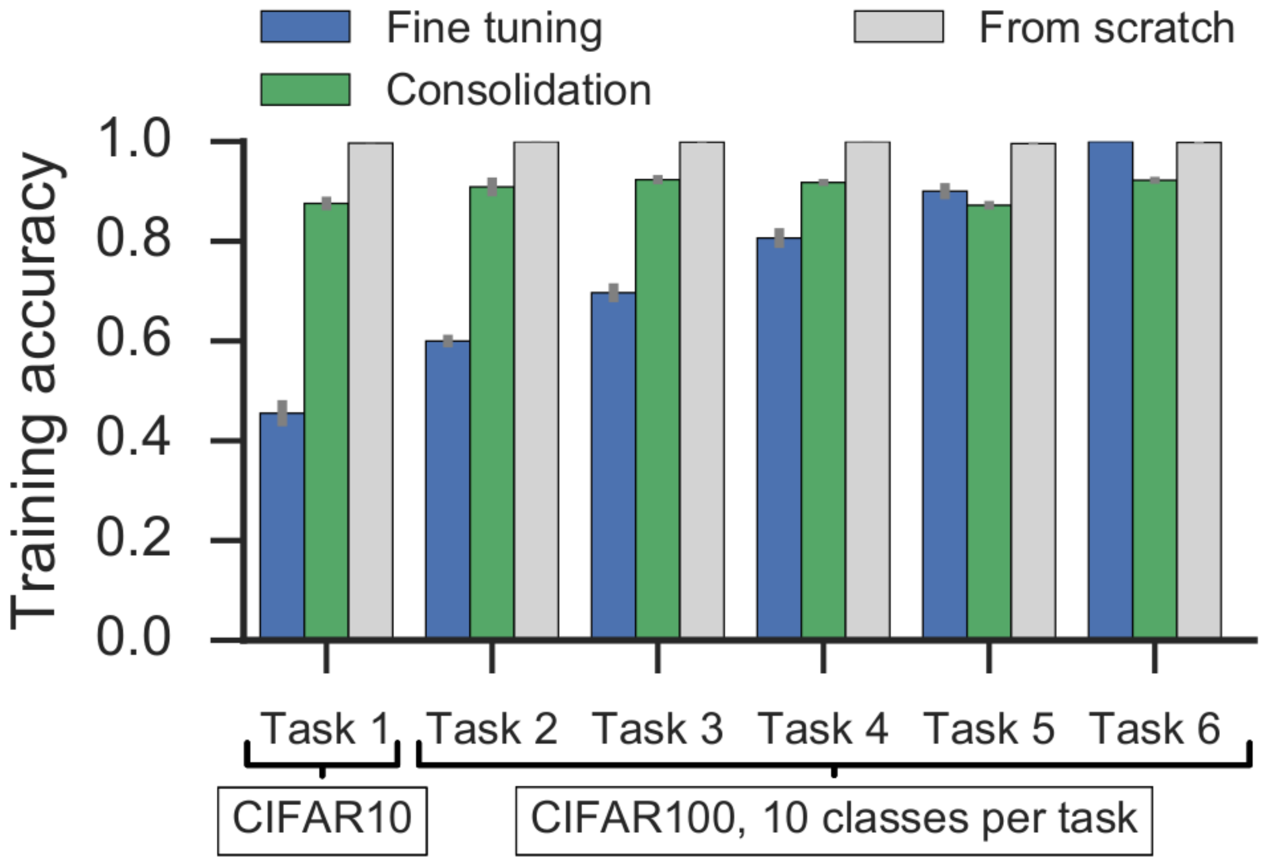
Fisher and our importance measure are correlated



Previous approaches

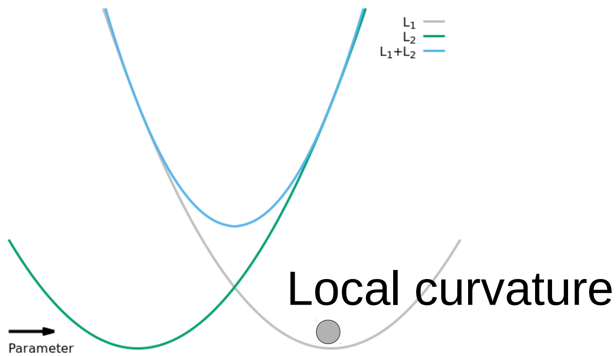
- **Architectural:** modify architecture to prevent forgetting
  Specific nonlinearities (Goodfellow et al., 2013; Srivasta et al., 2013), progressive nets (Rusu et al., 2016), fine tuning (Donahue et al., 2014)
  con: architectural complexity grows with tasks
- **Functional:** regularize activations or outputs of network
  LwF (Li & Hoiem, 2016), LFL (Jung et al., 2016)
  con: additional memory and computation to compare activations
- **Structural:** regularize parameters of network
  Elastic weight consolidation (Kirkpatrick et al., 2017)
  con: expensive to compute weights for regularization penalty

Training Error Split CIFAR10/100

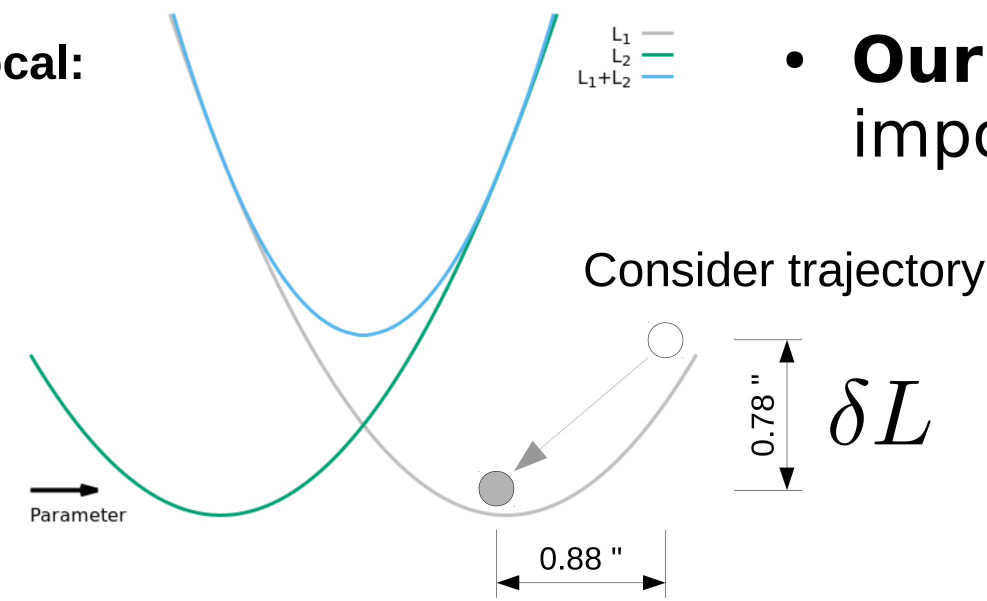


How to measure per-parameter importance: Local vs non-local

Local:



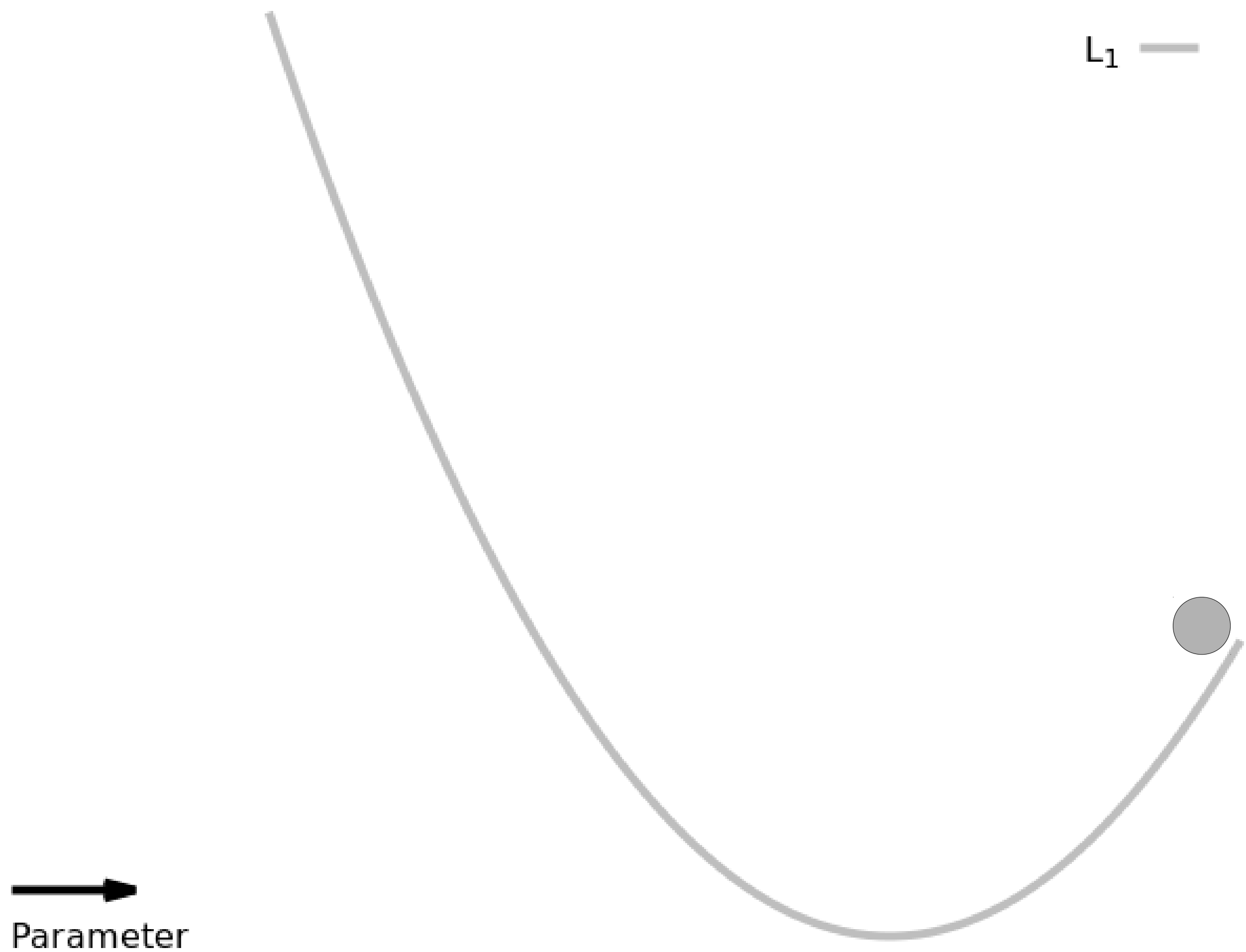
Non-local:



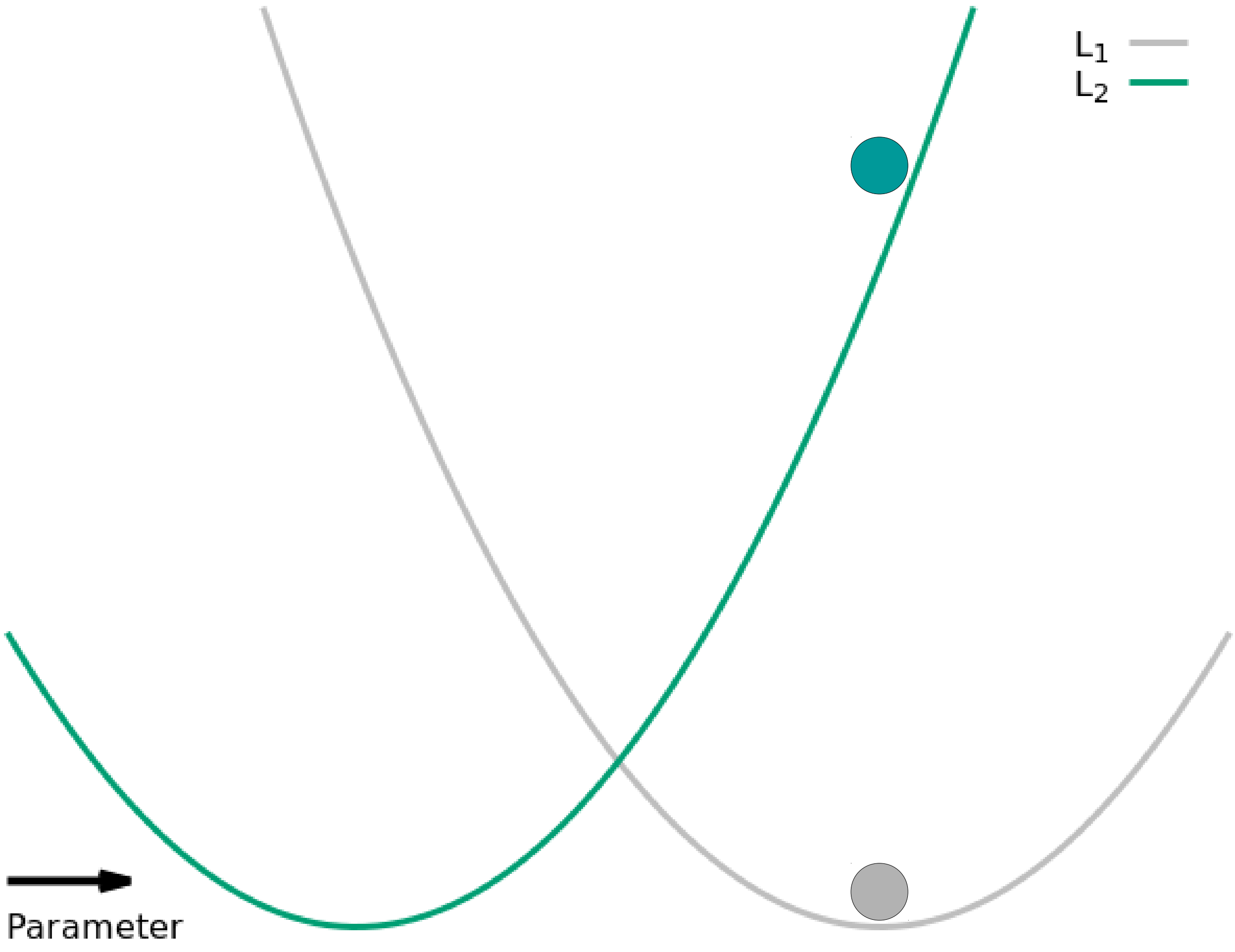
- **Our approach:** Estimate importance from trajectory

Need per-parameter contribution to changes in total loss L

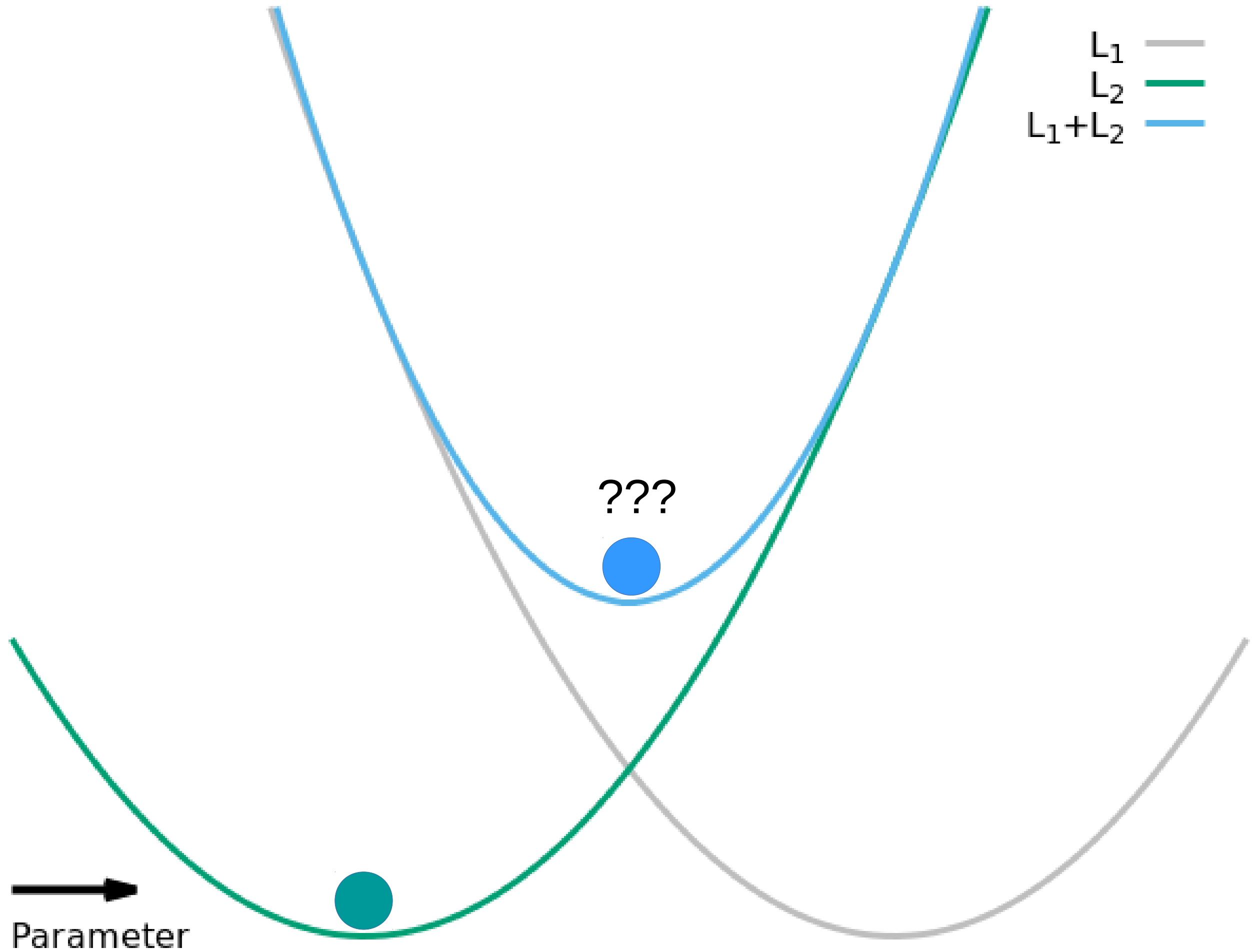
Problem: Catastrophic forgetting



Problem: Catastrophic forgetting



Problem: Catastrophic forgetting



$$L(\theta) = L_2 + L_1^{\text{approx}}$$