Optimal learning strategies via statistical physics and control theory

Francesco Mori

UNIVERSITY OF OXFORD

NEW COLLEGE OXFORD

ML Nosh Lunch

10th Feb 2025

“Optimal protocols for continual learning via statistical physics and control theory”

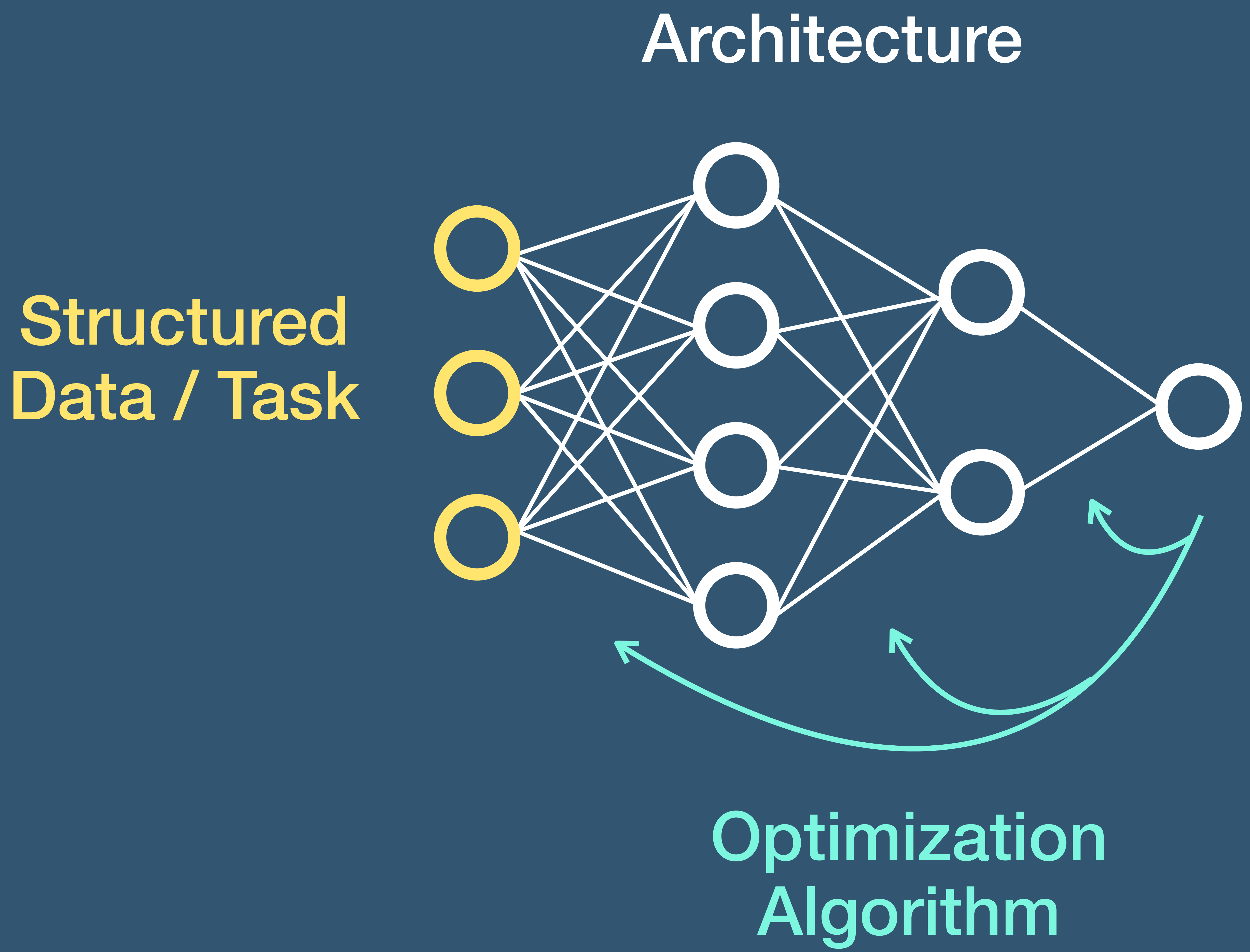
FM, Stefano Sarao Mannelli, Francesca Mignacco

Accepted at ICLR 2025

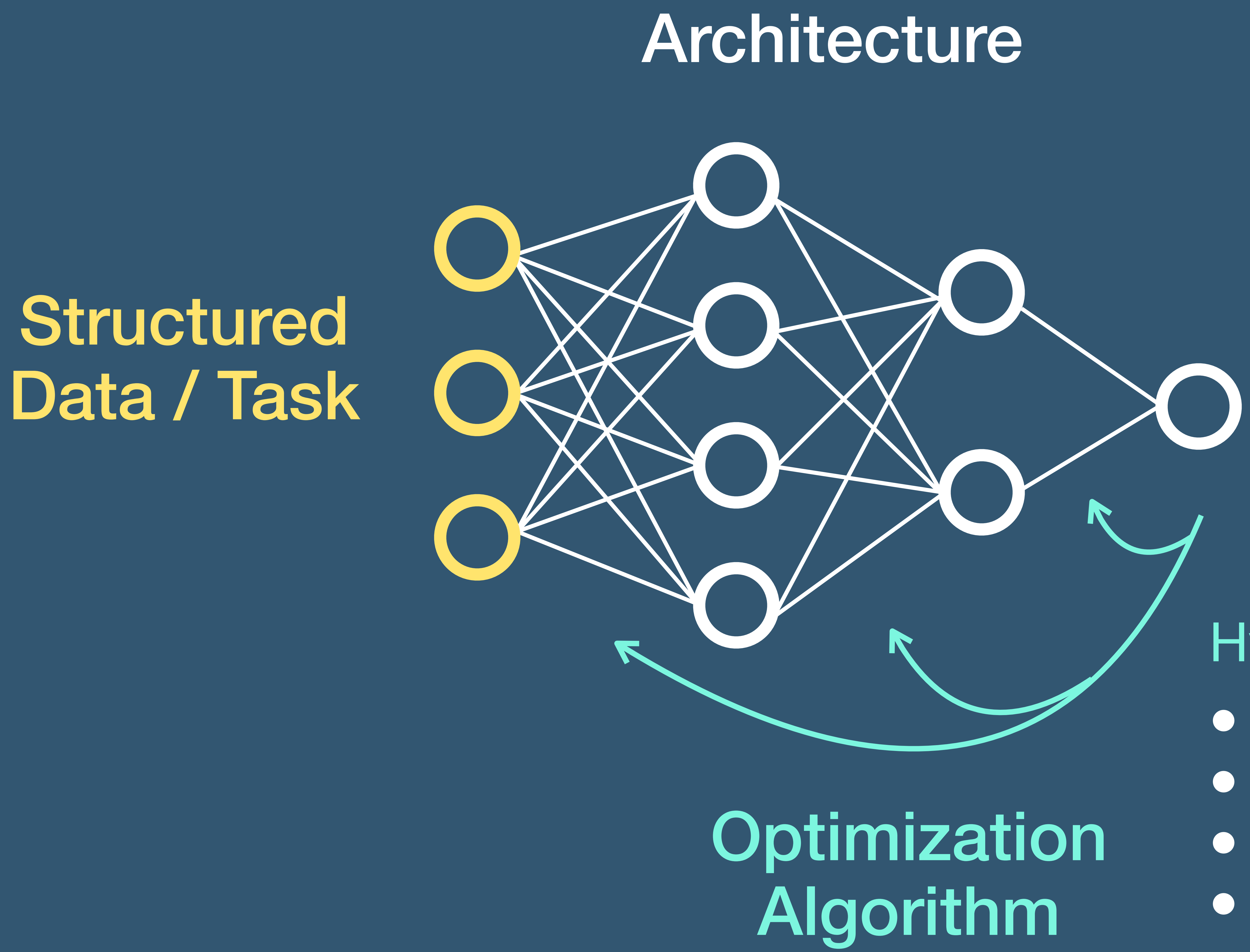
Francesca Mignacco
Princeton and CUNY

Stefano Sarao Mannelli
Chalmers University

Learning protocols



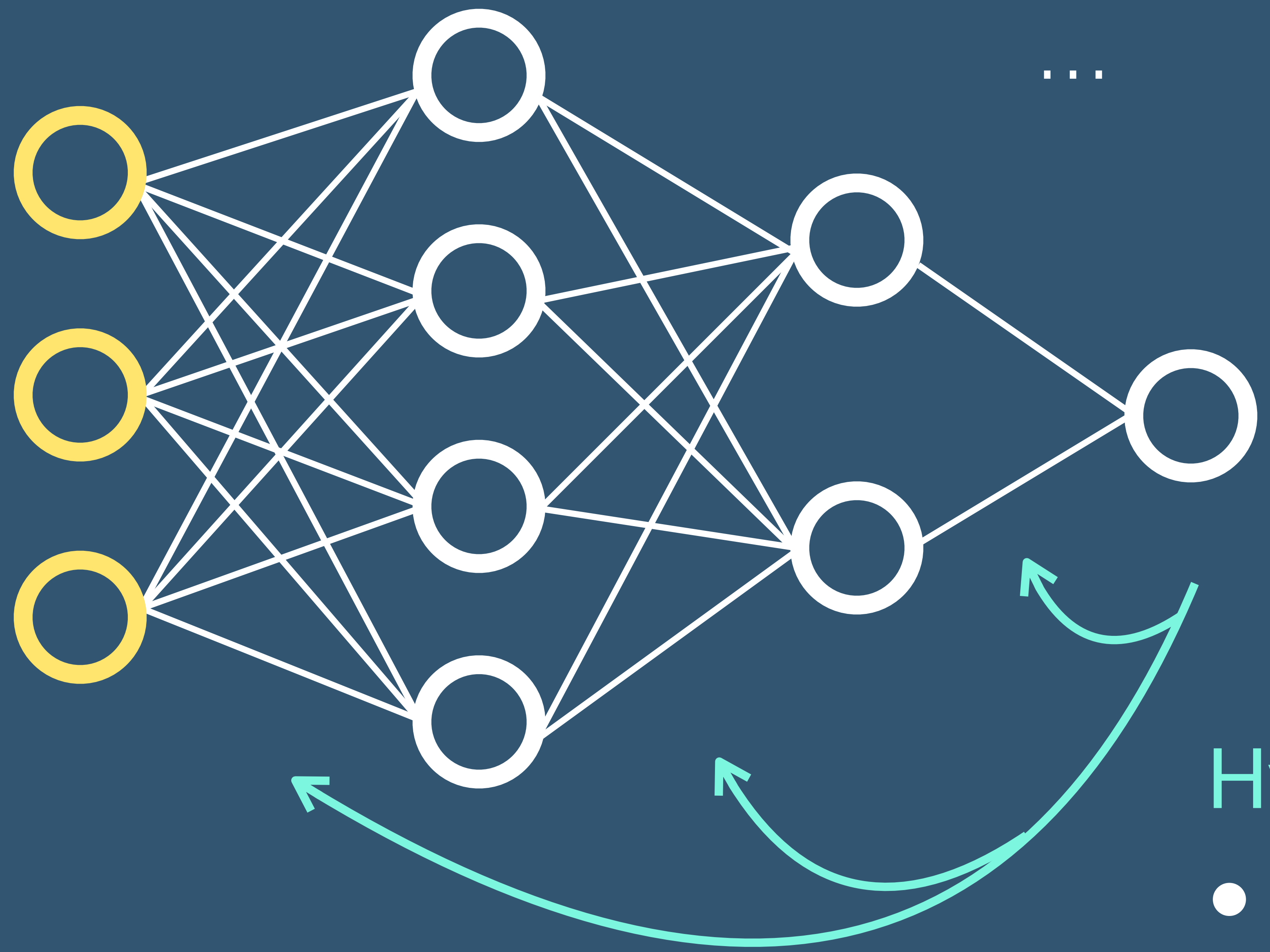
Learning protocols



Hyper-parameter schedules:

- Learning rate
- Momentum
- Batch size
- Regularization
- …

Learning protocols



Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

Hyper-parameter schedules:

- Learning rate
- Momentum
- Batch size
- Regularization
- …

Learning protocols

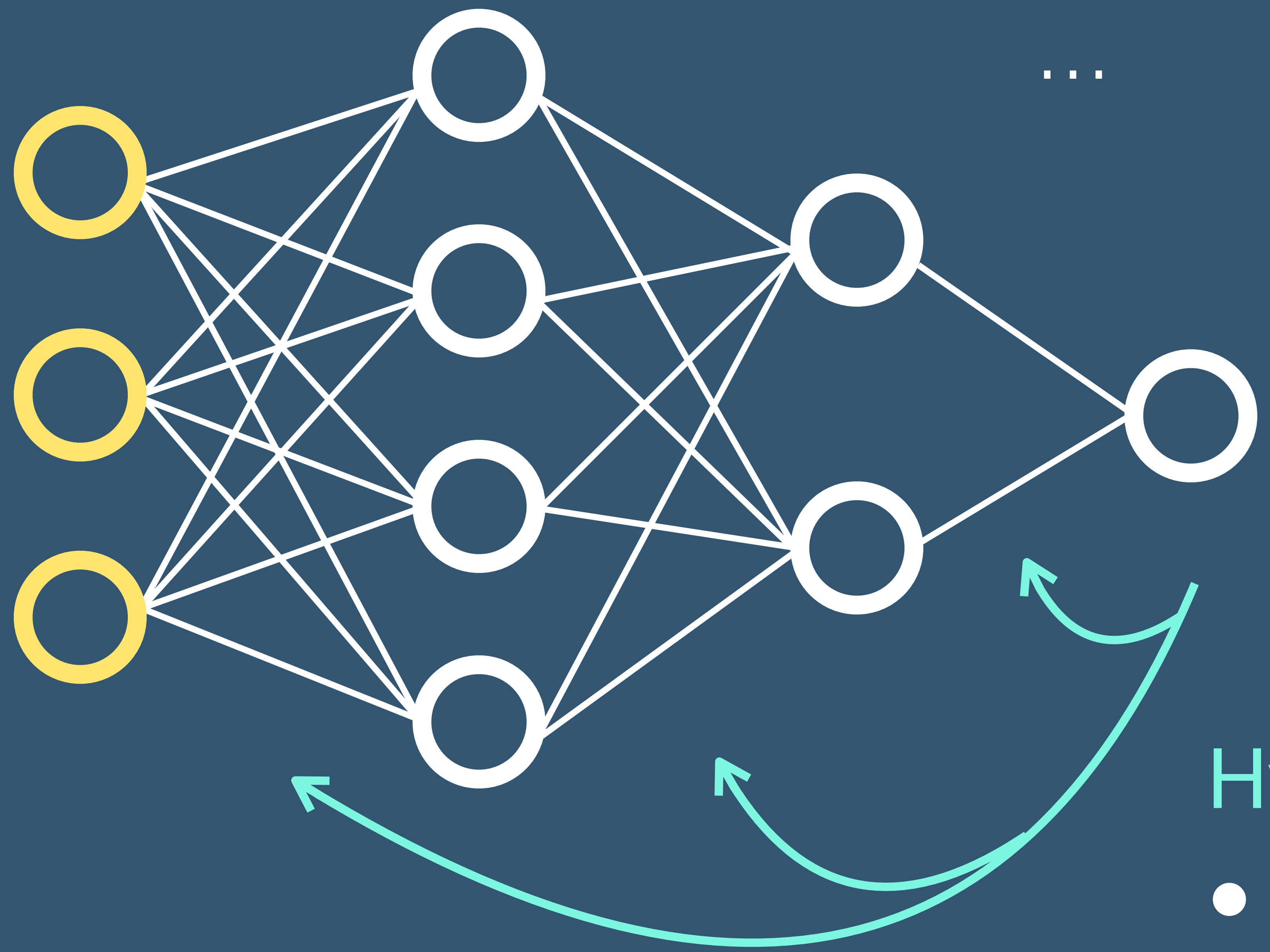
Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

Architecture

Structured
Data / Task



Dynamic data / task selection:

- Active learning
- Curriculum learning
- Transfer learning
- Multi-task learning
- …

Hyper-parameter schedules:

- Learning rate
- Momentum
- Batch size
- Regularization
- …

Optimization
Algorithm

Learning protocols

UNIVERSITY OF OXFORD

Goals:

- **Speed-up** convergence
- **Guide** the training towards better regions of parameters space

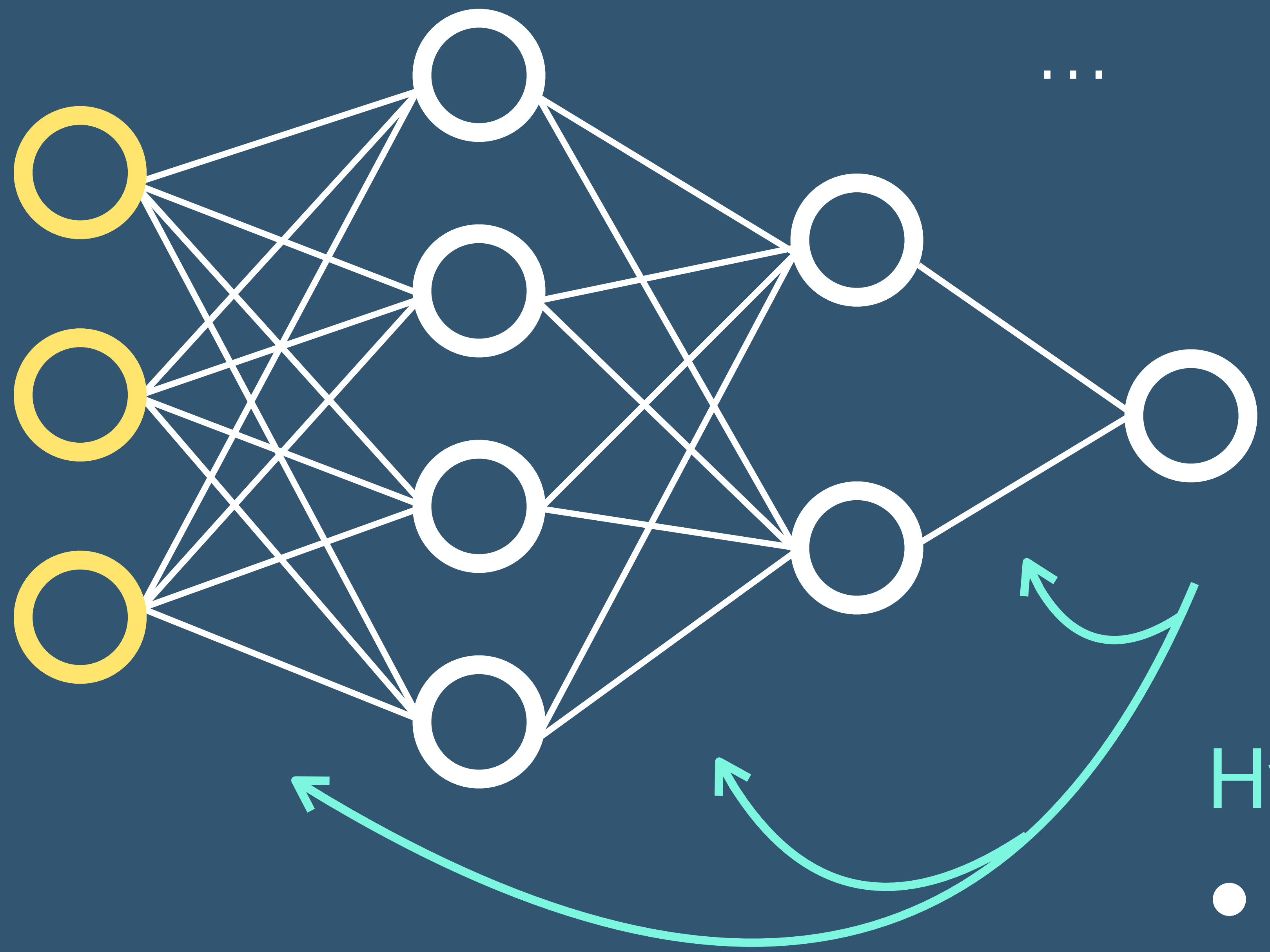
Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

Architecture

Structured Data / Task



Optimization Algorithm

Dynamic data / task selection:

- Active learning
- Curriculum learning
- Transfer learning
- Multi-task learning
- …

Hyper-parameter schedules:

- Learning rate
- Momentum
- Batch size
- Regularization
- …

Learning protocols

Goals:

- **Speed-up** convergence
- **Guide** the training towards better regions of parameters space

From smoother landscape to the target

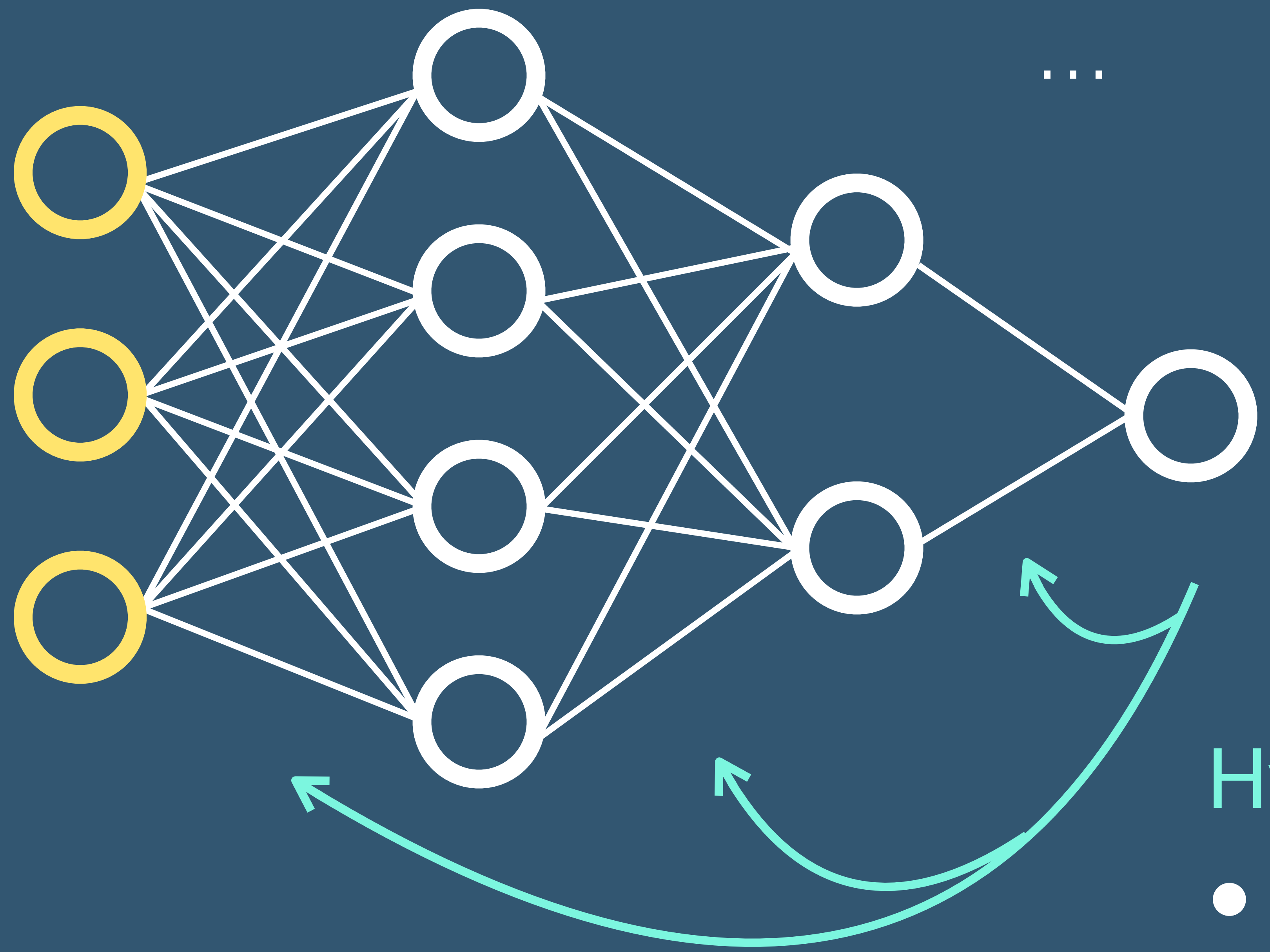
Architecture

Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

Structured Data / Task



In this talk:

Dynamic data / task selection:

- Active learning
- Curriculum learning
- Transfer learning
- Multi-task learning
- …

Optimization Algorithm

Hyper-parameter schedules:

- Learning rate
- Momentum
- Batch size
- Regularization
- …

Supervised learning - general setup

Dataset with labels

$$\mathscr{D} = \{x_i, y_i\}_{i=1}^{P}$$

Neural Network

$$\hat{y} = f_{\mathbf{w}}(x)$$

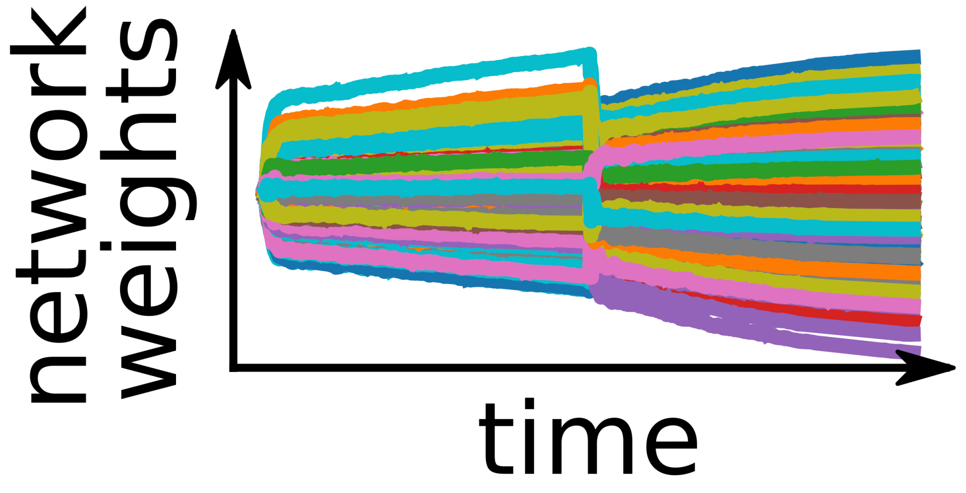
Simplest example: $\hat{y} = \mathbf{erf}(\mathbf{w}^\intercal x)$

Error (aka loss)

$$\mathscr{L} = \frac{1}{2}\left(\hat{y} - y\right)^2$$

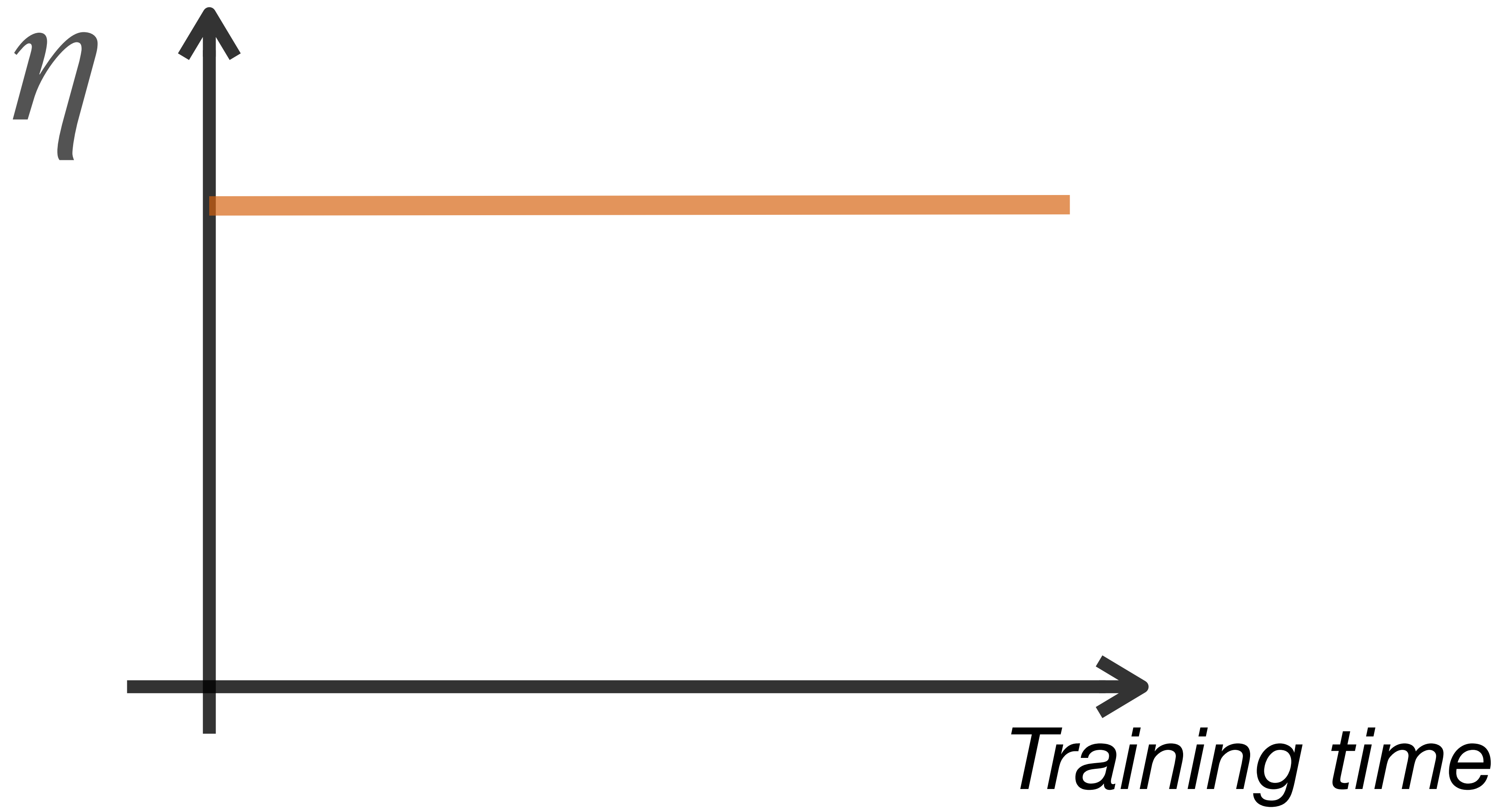
(Online) Stochastic gradient descent

$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

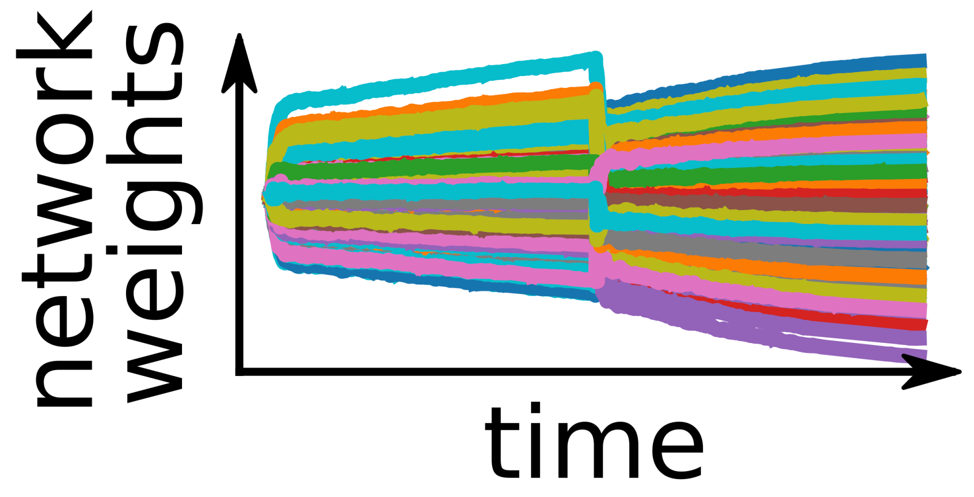


$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics

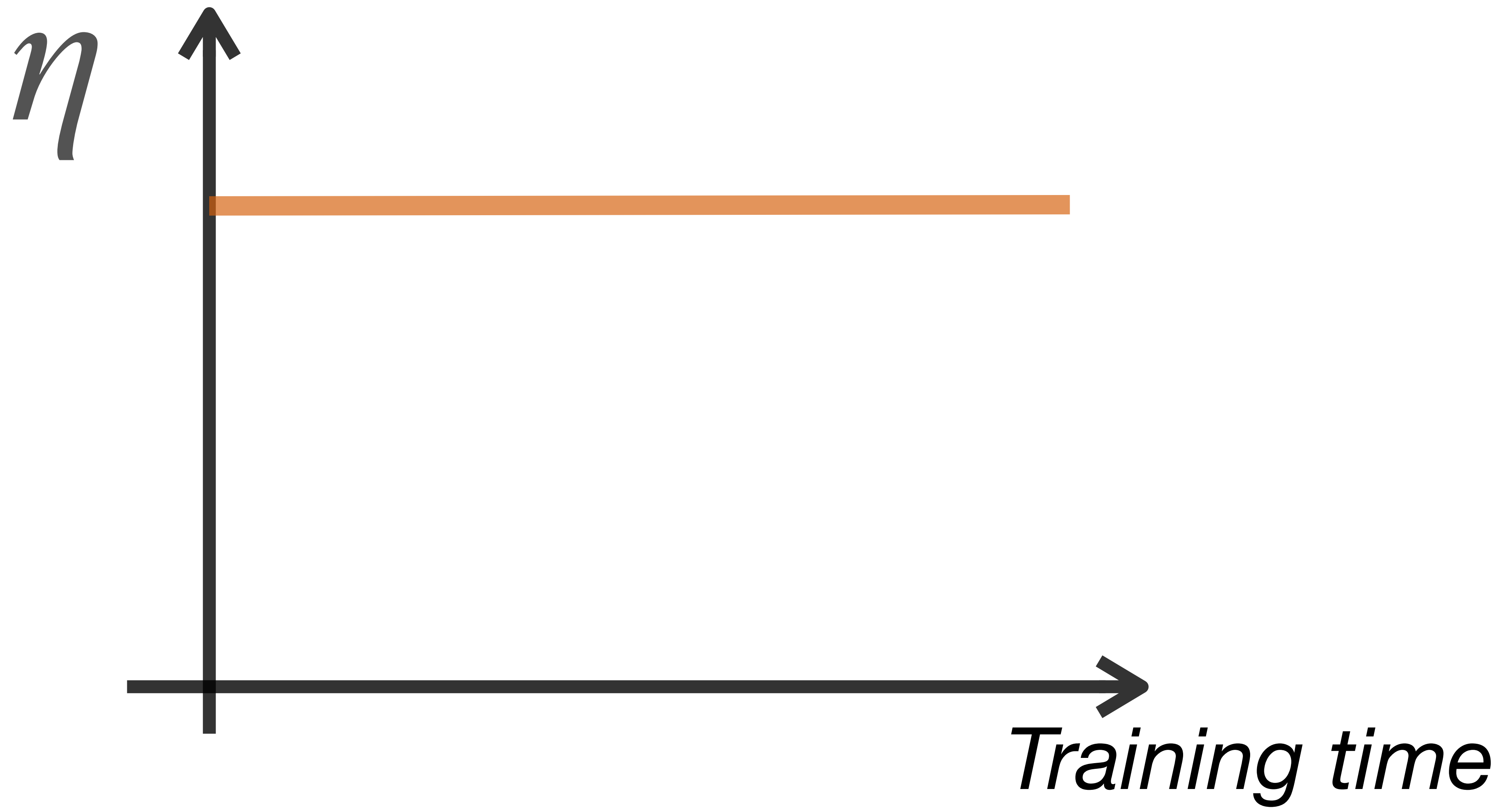


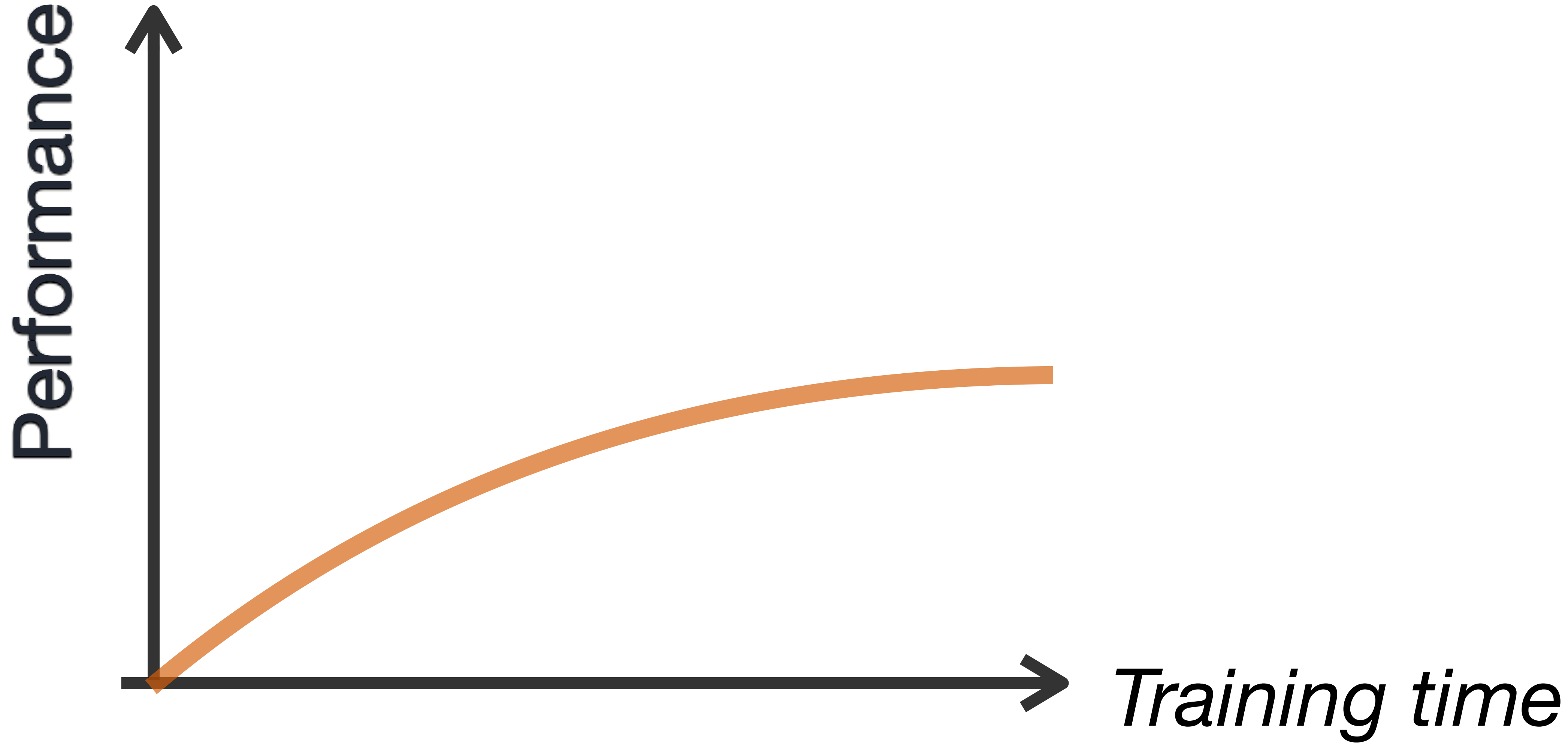
Training protocols example: learning rate



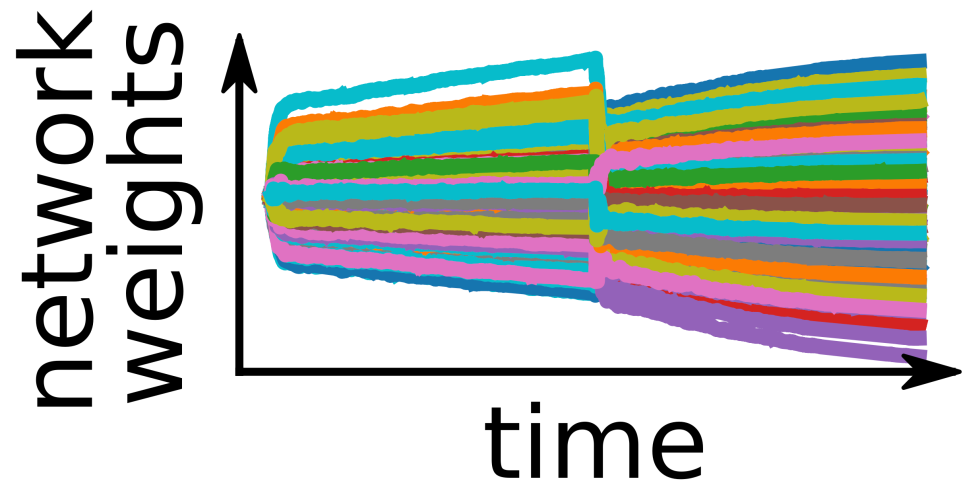
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics



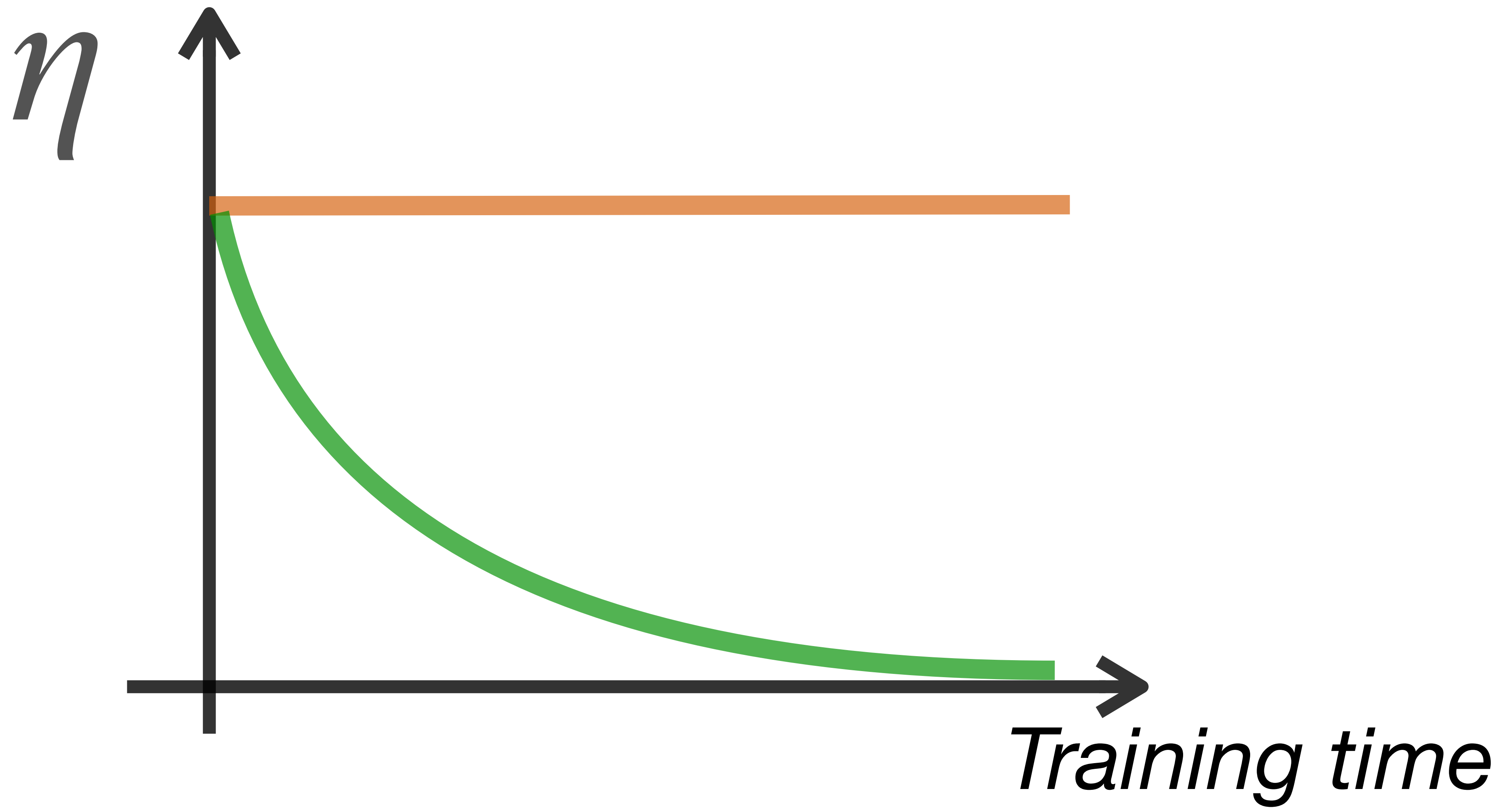


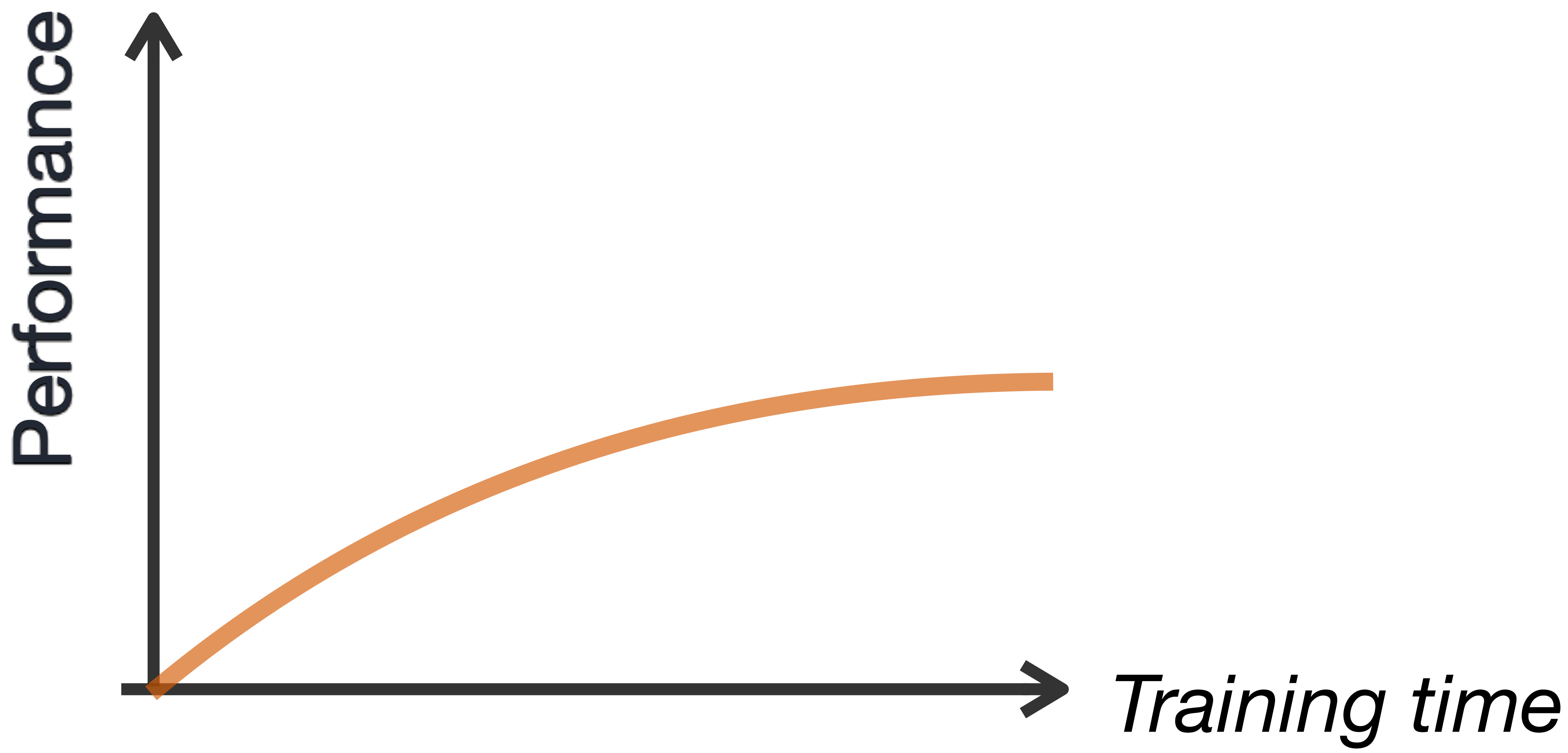
Training protocols example: learning rate



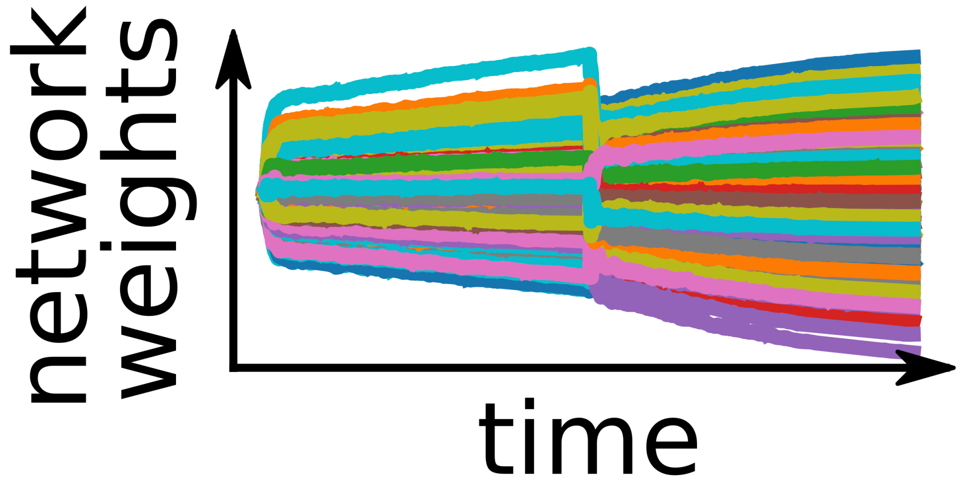
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics



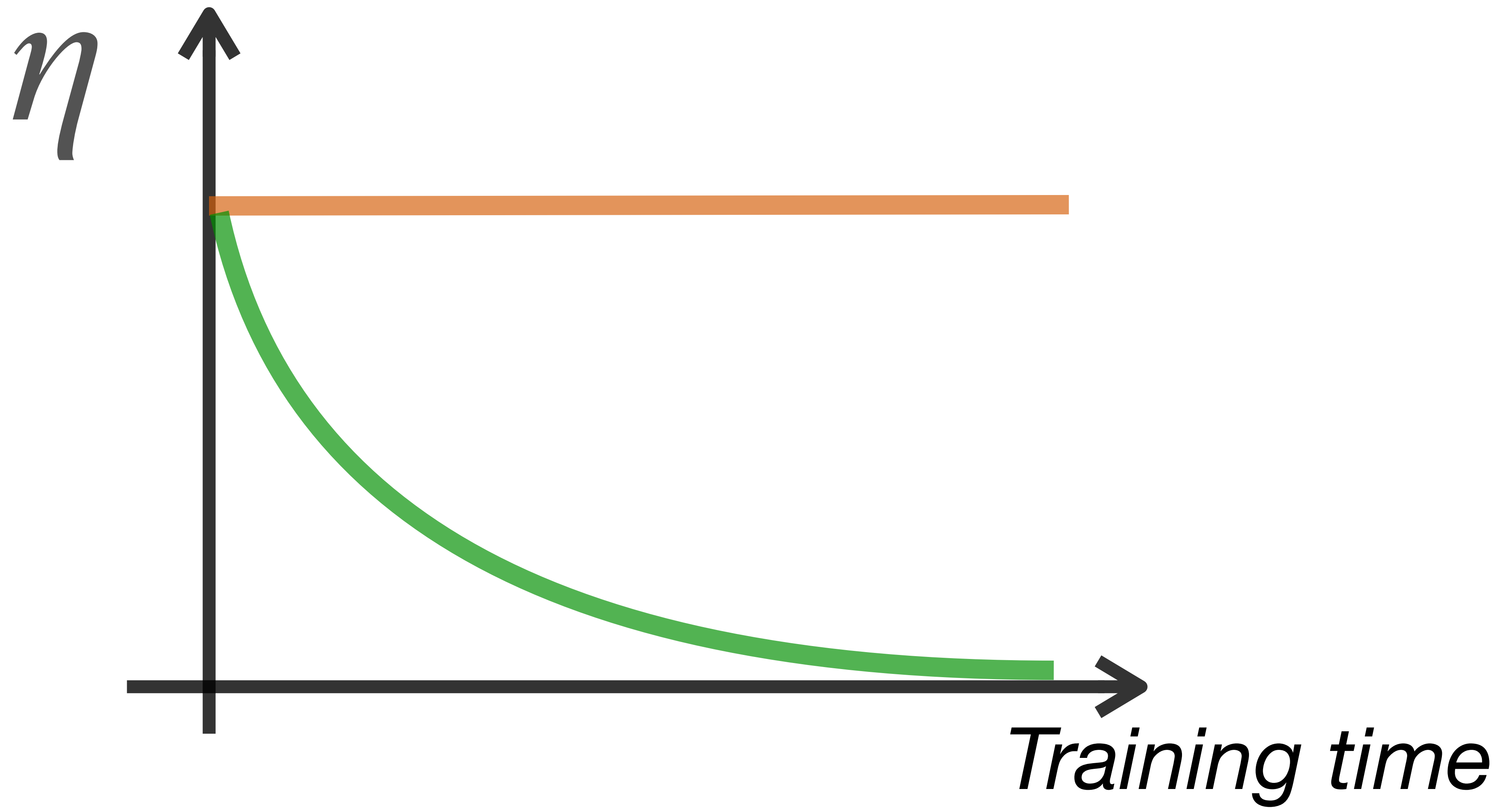


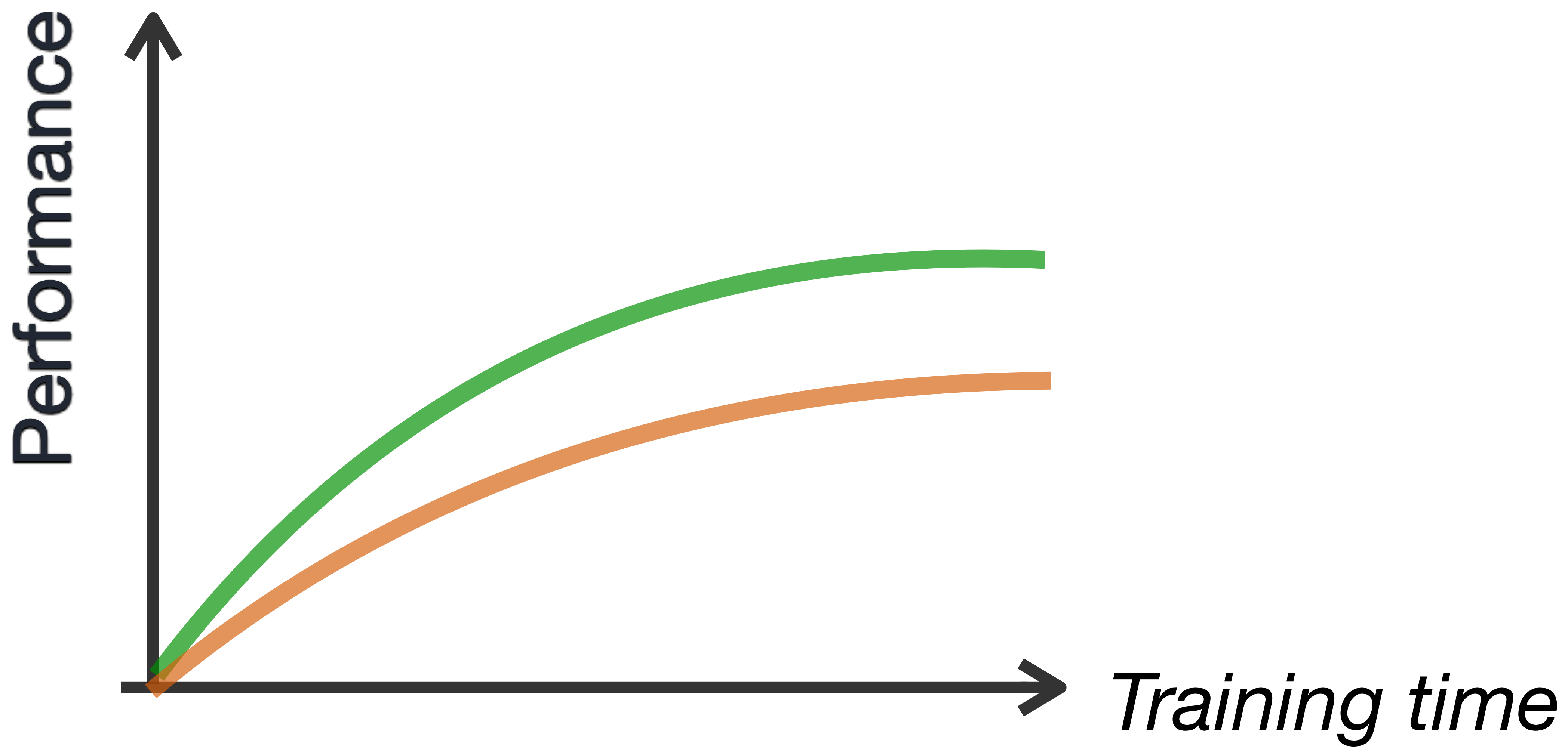
Training protocols example: learning rate



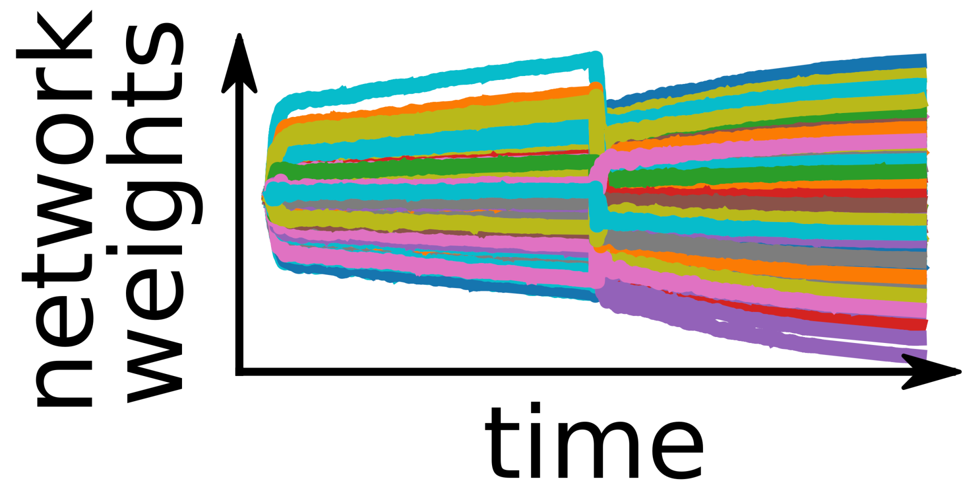
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics



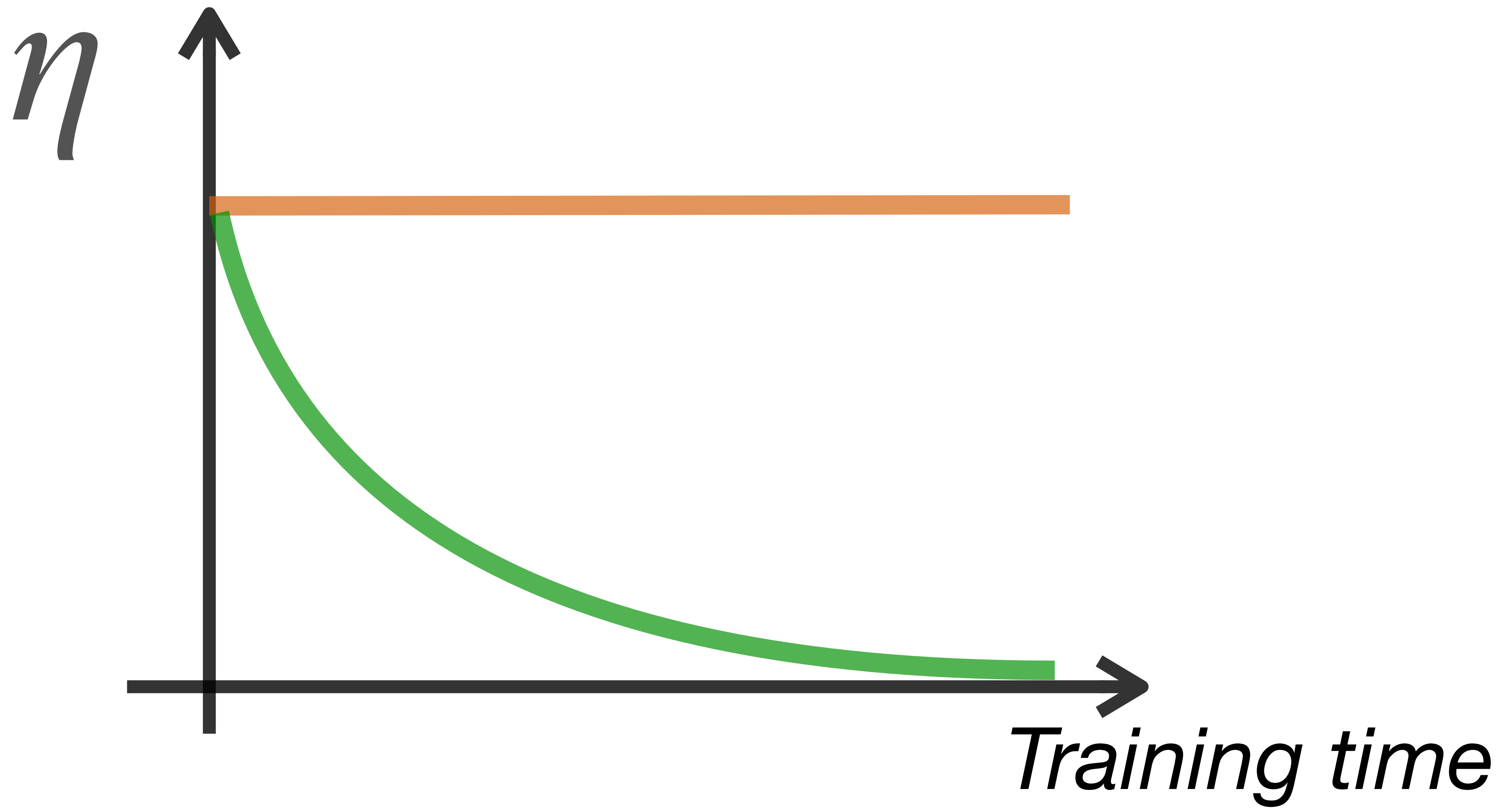


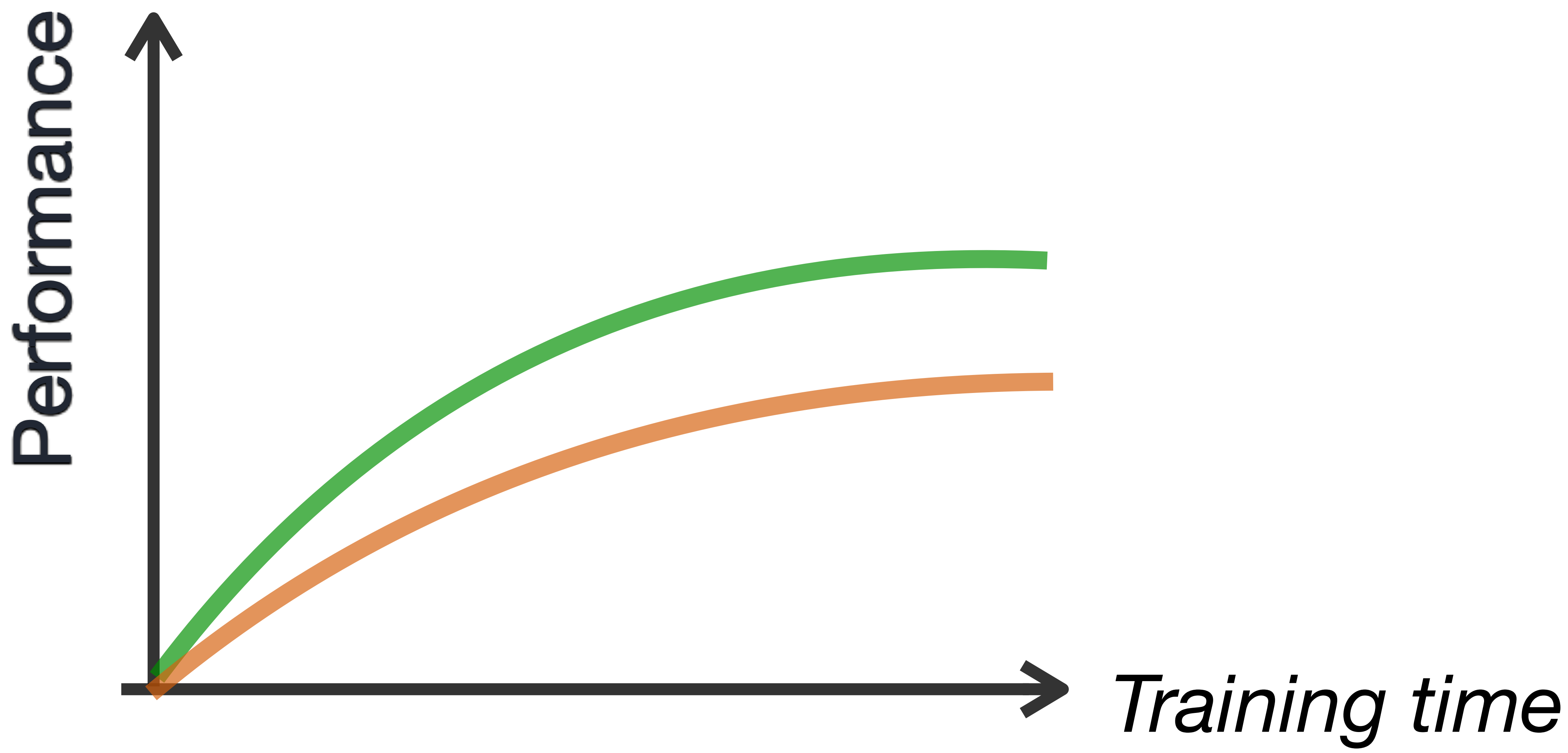
Training protocols example: learning rate



$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics





Can we compute the *optimal* * strategy ?

* In terms of the final performance

Training protocols

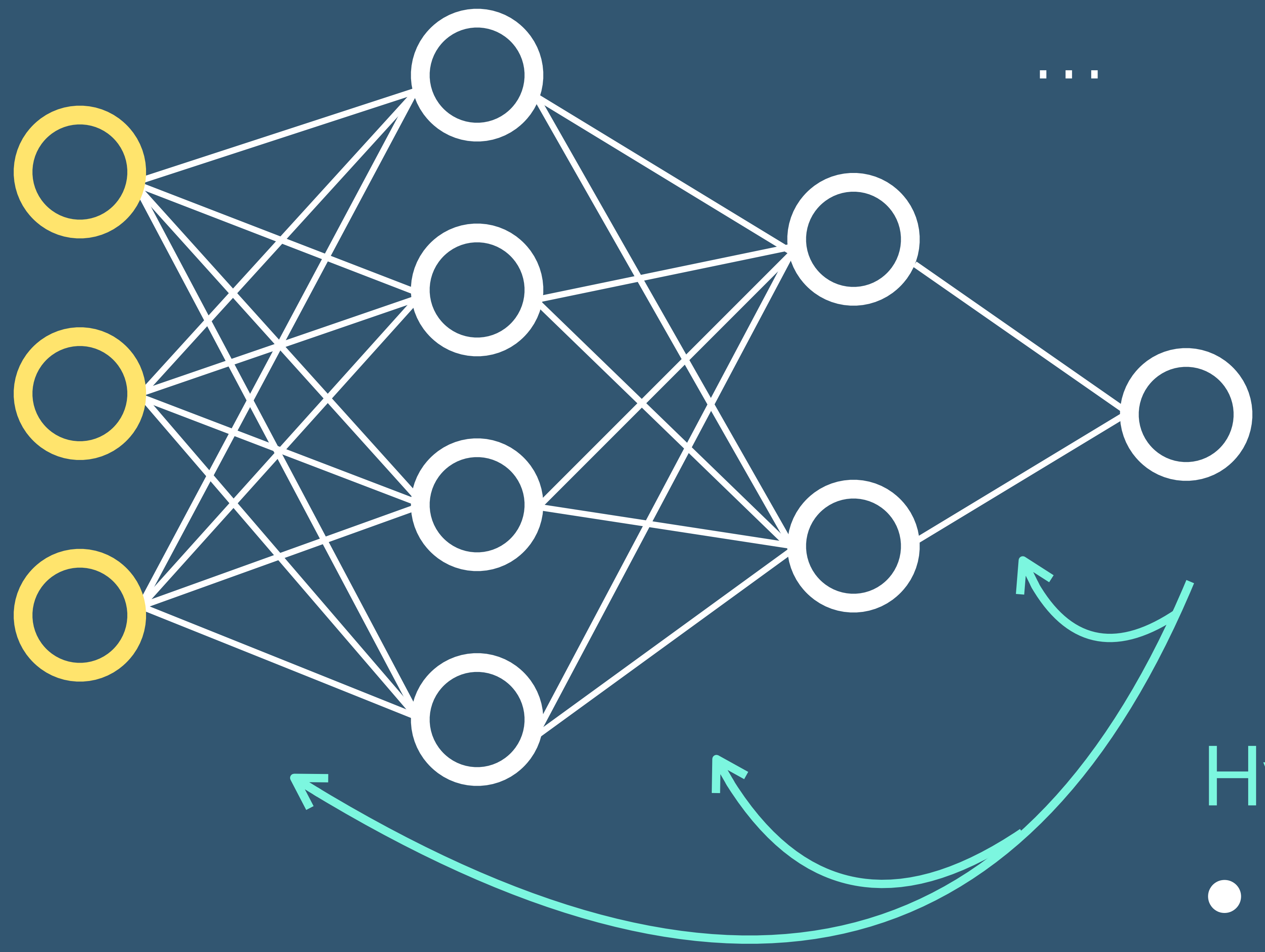
Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

Architecture

Structured Data / Task



Dynamic data / task selection:

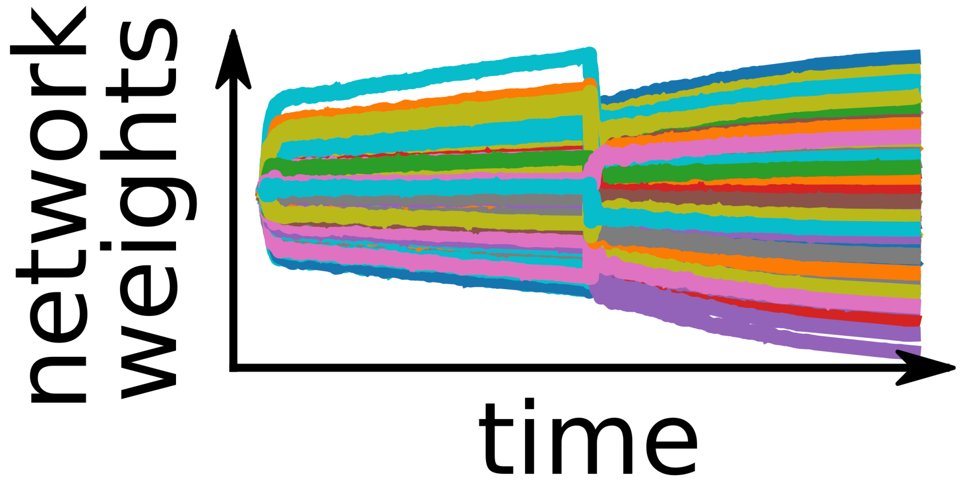
- Active learning
- Curriculum learning
- Transfer learning
- Multi-task learning
- …

Optimization Algorithm

Hyper-parameter schedules:

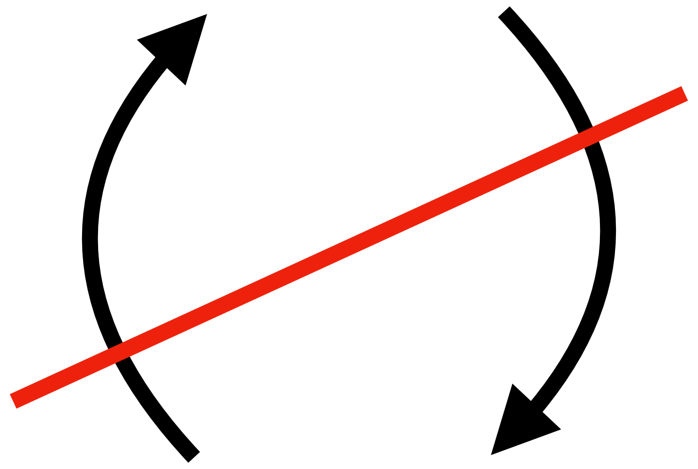
- Learning rate
- Momentum
- Batch size
- Regularization
- …

Dimensionality reduction + optimal control



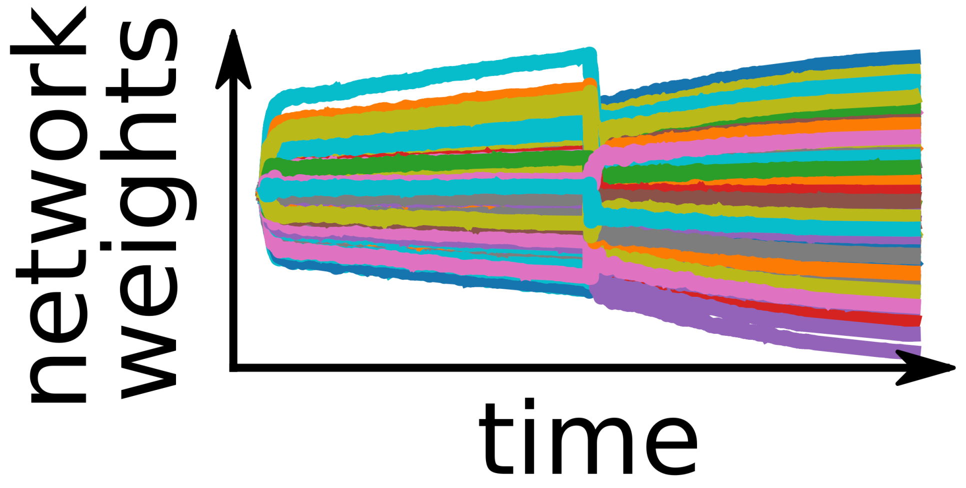
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional
complex dynamics



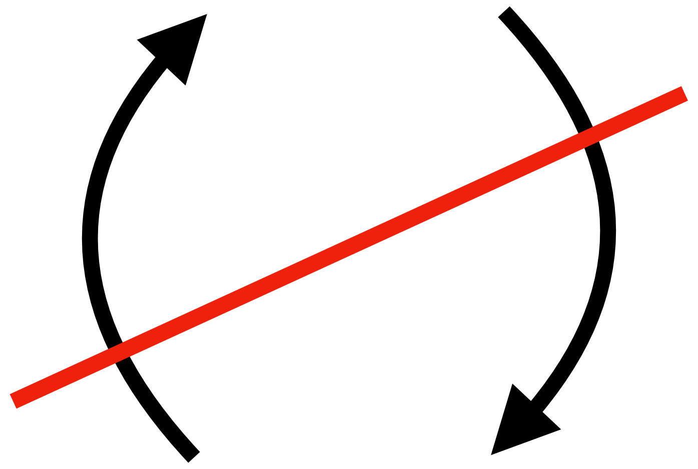
Control theory is
computationally
demanding

Dimensionality reduction + optimal control



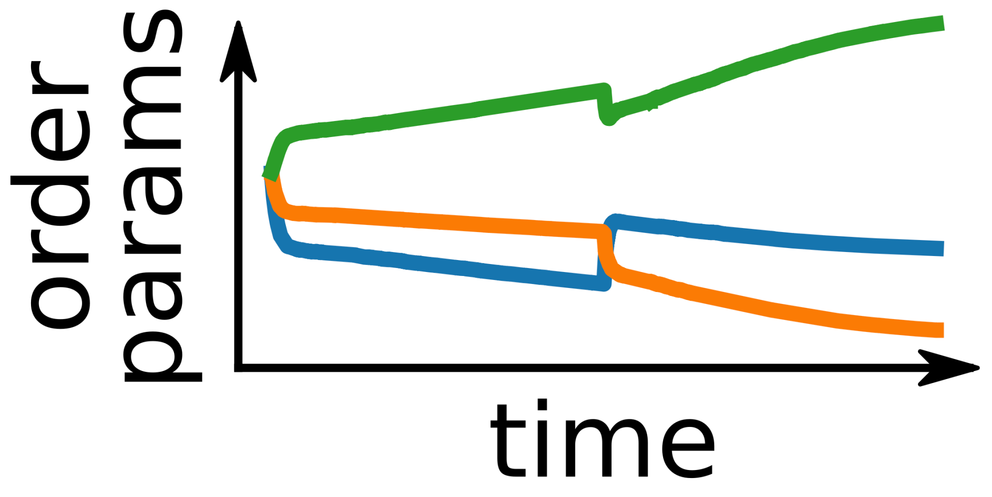
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics



Control theory is computationally demanding

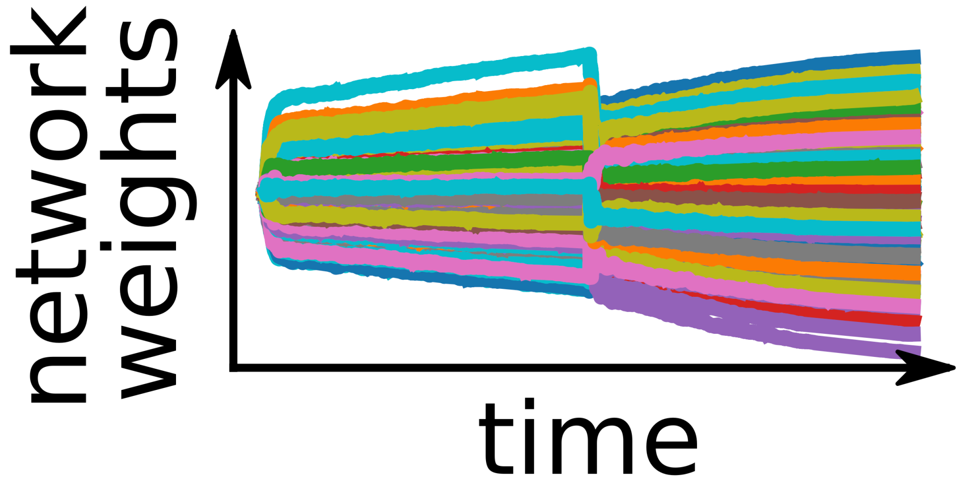
$$N \to \infty$$



$$\dot{\mathbf{Q}} = f(\mathbf{Q}, \mathbf{u})$$ control variables

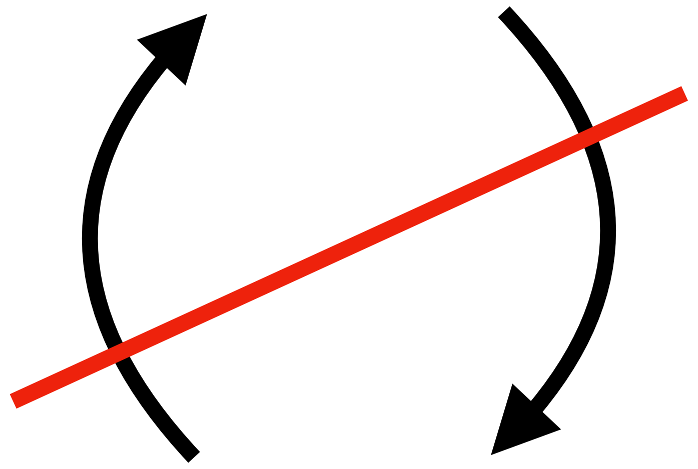
Low-dimensional effective description

Dimensionality reduction + optimal control



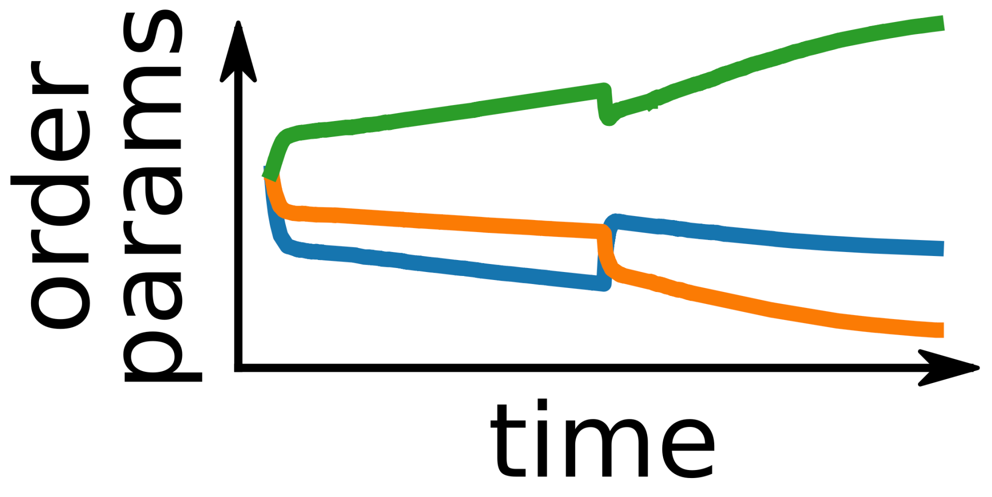
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu}$$

High-dimensional complex dynamics



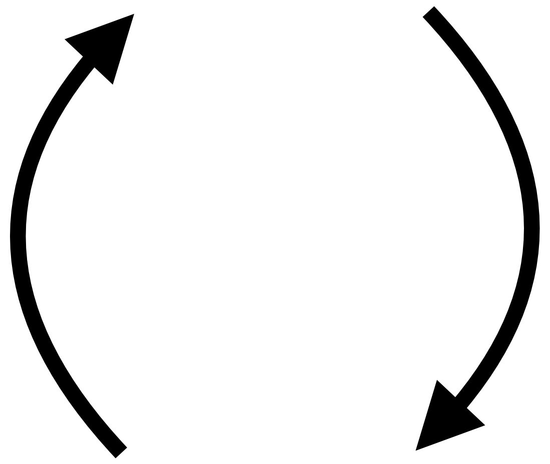
Control theory is computationally demanding

$$N \to \infty$$



$$\dot{\mathbf{Q}} = f(\mathbf{Q}, \mathbf{u})$$ control variables

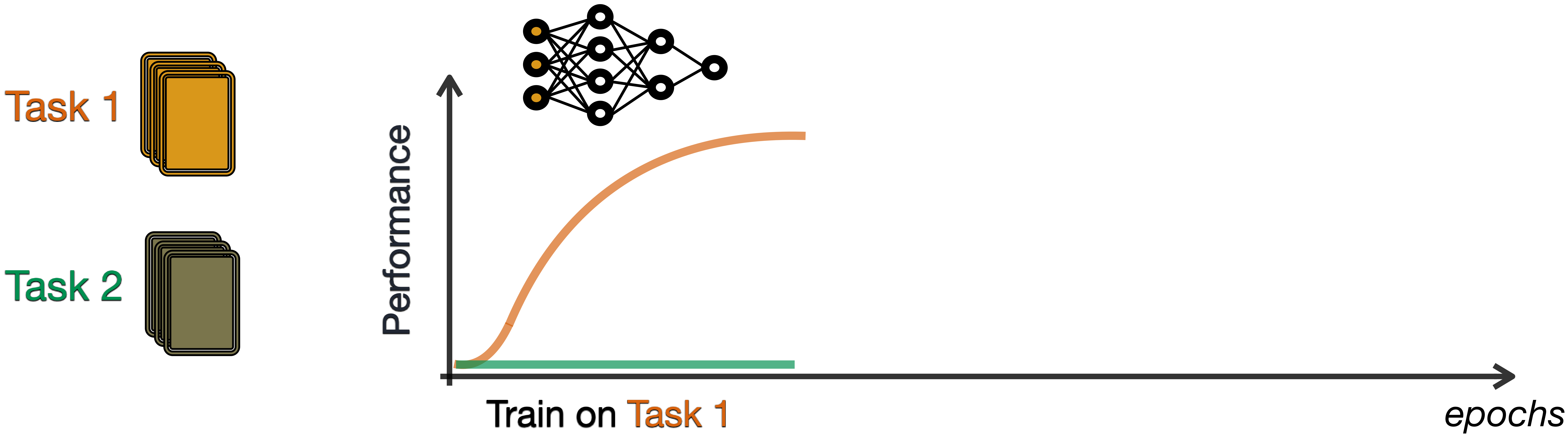
Low-dimensional effective description



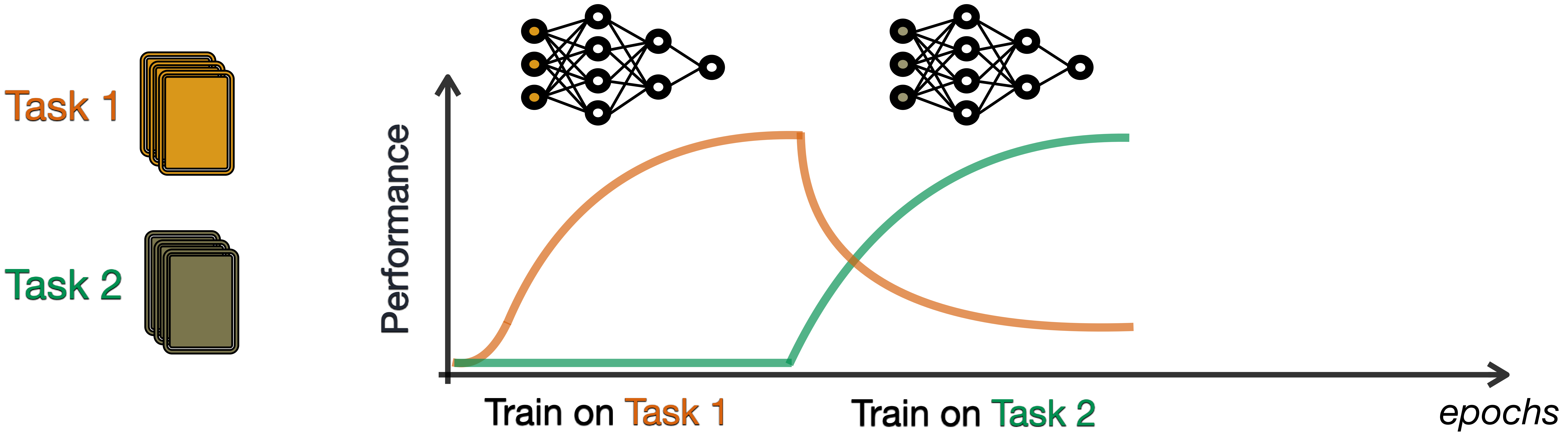
Control theory can be applied


Task 1
Task 2
Performance
epochs

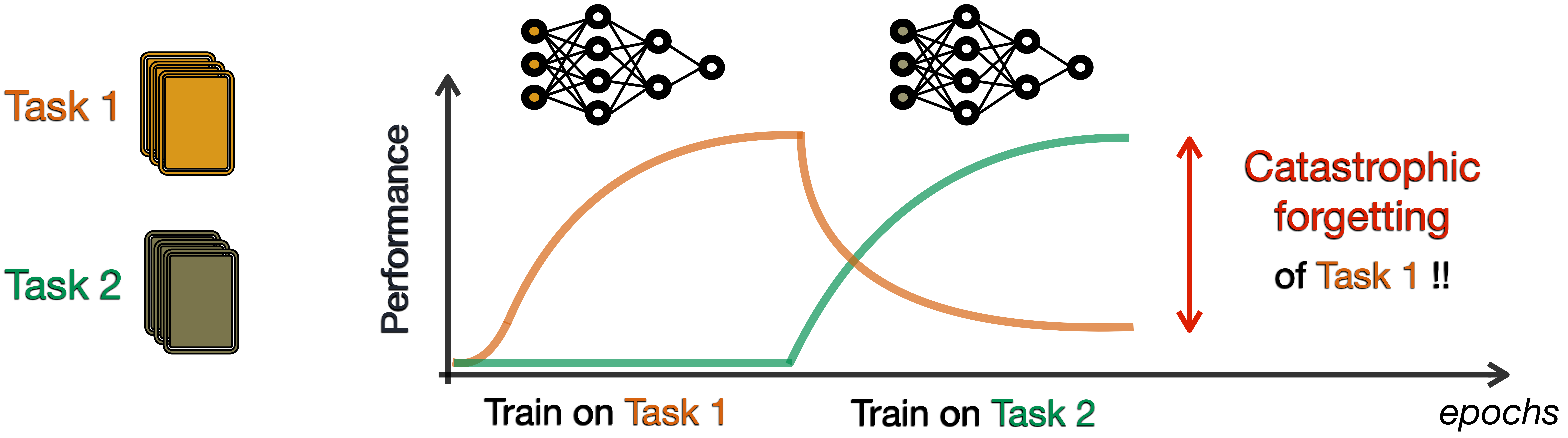
Continual learning & catastrophic forgetting



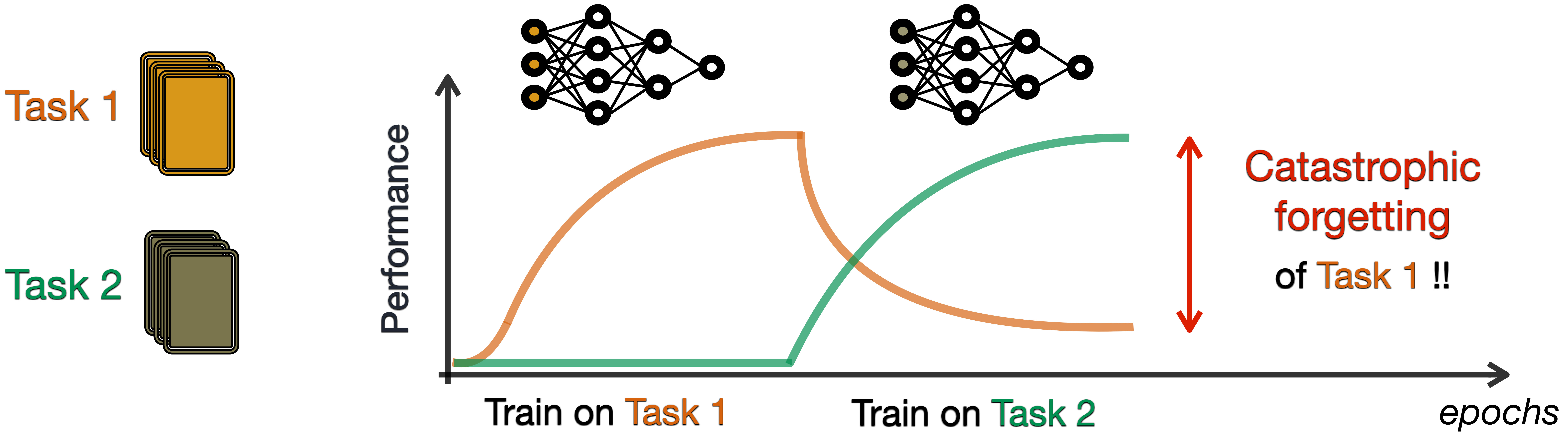
Continual learning & catastrophic forgetting



Continual learning & catastrophic forgetting



Continual learning & catastrophic forgetting



ML (empirical):

ML (theory):

A teacher-student model of continual learning

Introduced in: *Lee, Goldt, & Saxe (ICML 2021)*

Task 1

$\mathcal{D}_1 = \{x_i^{(1)}, y_i^{(1)}\}$

Task 2

$\mathcal{D}_2 = \{x_i^{(2)}, y_i^{(2)}\}$

$x \sim \mathcal{N}(0,1) \in \mathbf{R}^N \qquad N \gg 1$

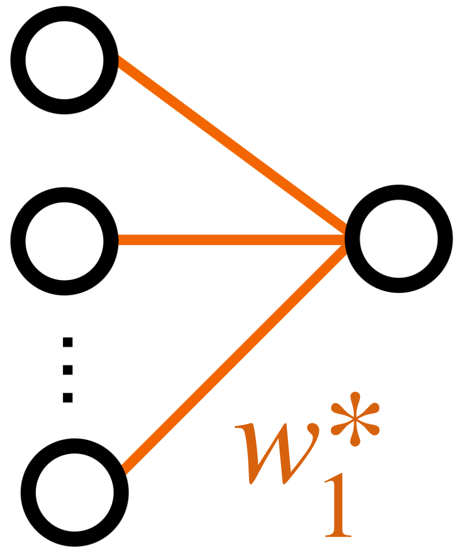
A teacher-student model of continual learning

Introduced in: *Lee, Goldt, & Saxe (ICML 2021)*

Task 1

Teacher 1

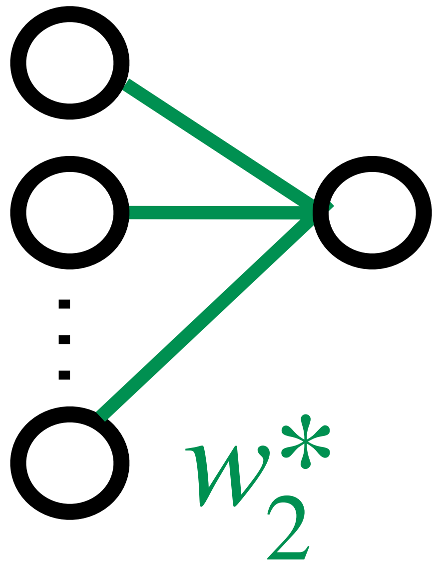
$$y^{(1)} = g_*\left(w_1^* \cdot \frac{x}{\sqrt{N}} \right)$$



Task 2

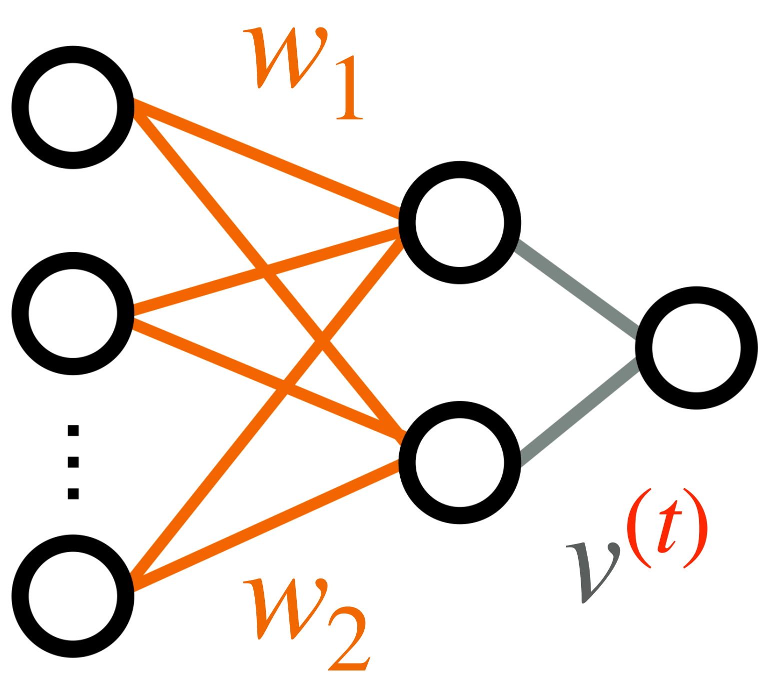
Teacher 2

$$y^{(2)} = g_*\left(w_2^* \cdot \frac{x}{\sqrt{N}} \right)$$



Student

(multi-head)



$$\hat{y}^{(t)} = \sum_{k=1}^{K=2} v_k^{(t)} g\left(w_k \cdot \frac{x}{\sqrt{N}} \right)$$

A teacher-student model of continual learning

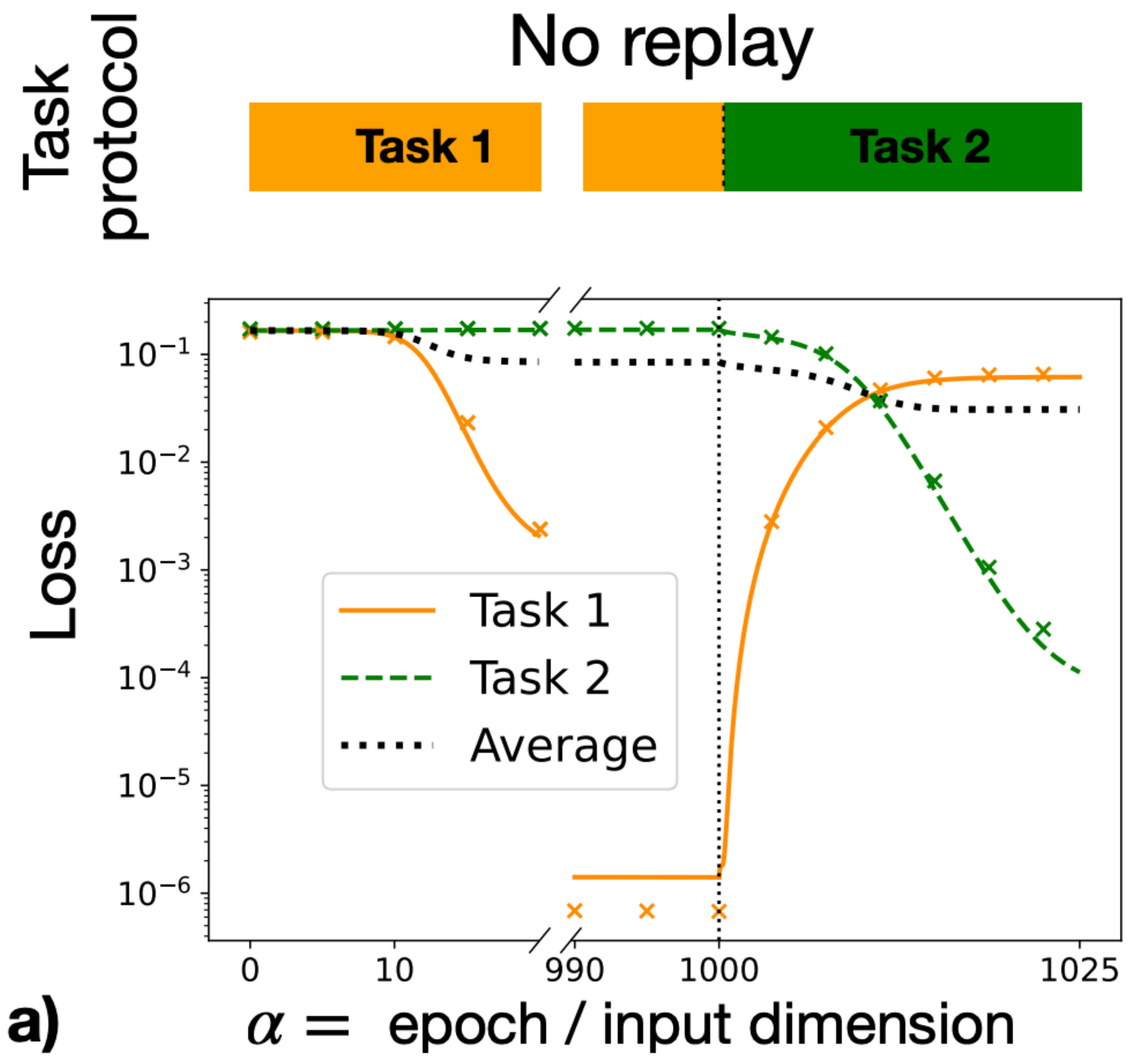
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla_{\mathbf{w}} \mathscr{L}^{\mu} \xrightarrow{N \to \infty} \frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} = f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

ODEs for the order parameters:

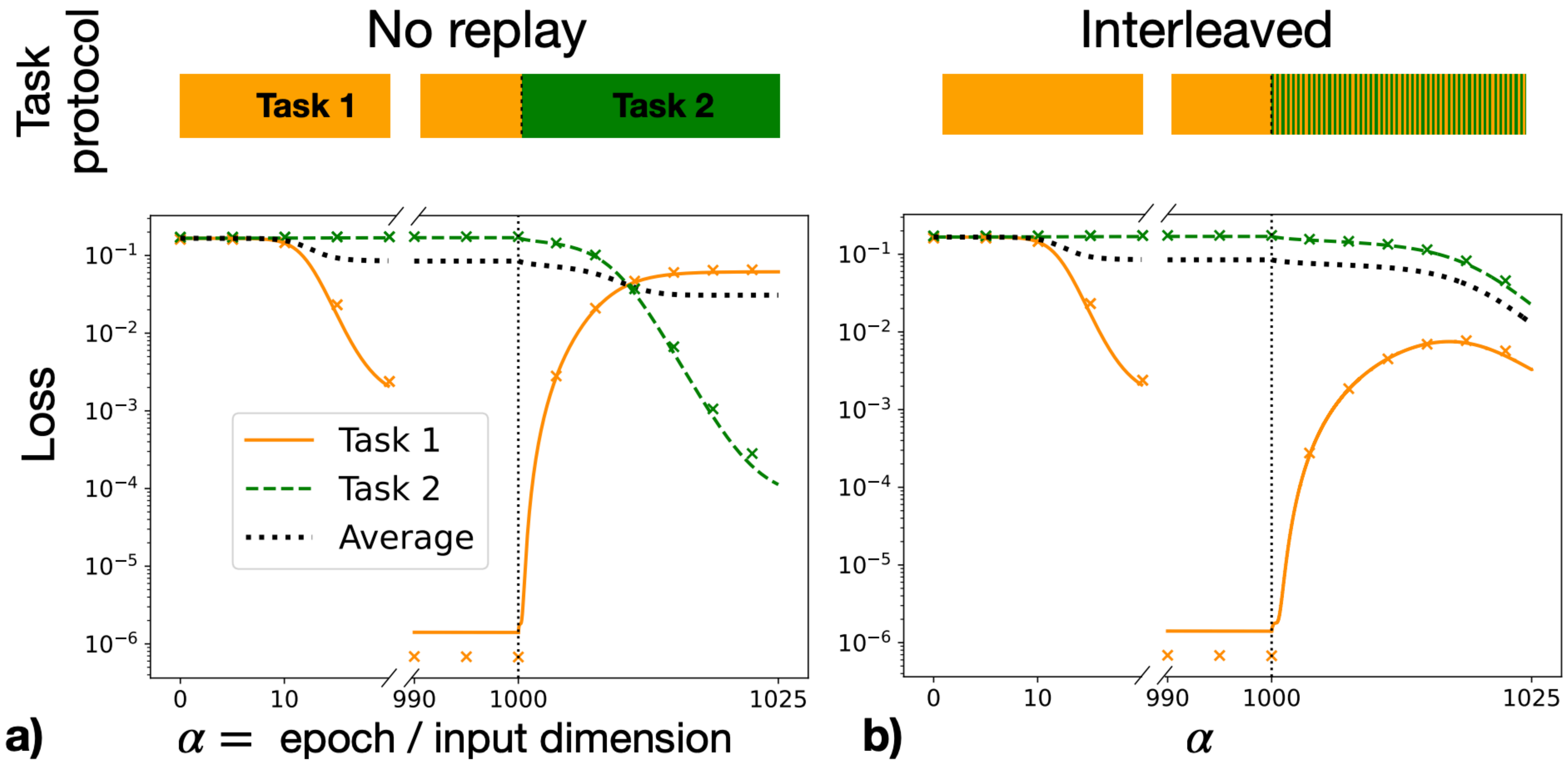
$$\alpha = \mu/N \qquad \alpha \in (0, \alpha_F]$$

Example: "Magnetisation" $M_{kt} = \dfrac{w_k \cdot w_*^t}{N}$

Results: optimal strategy vs benchmarks

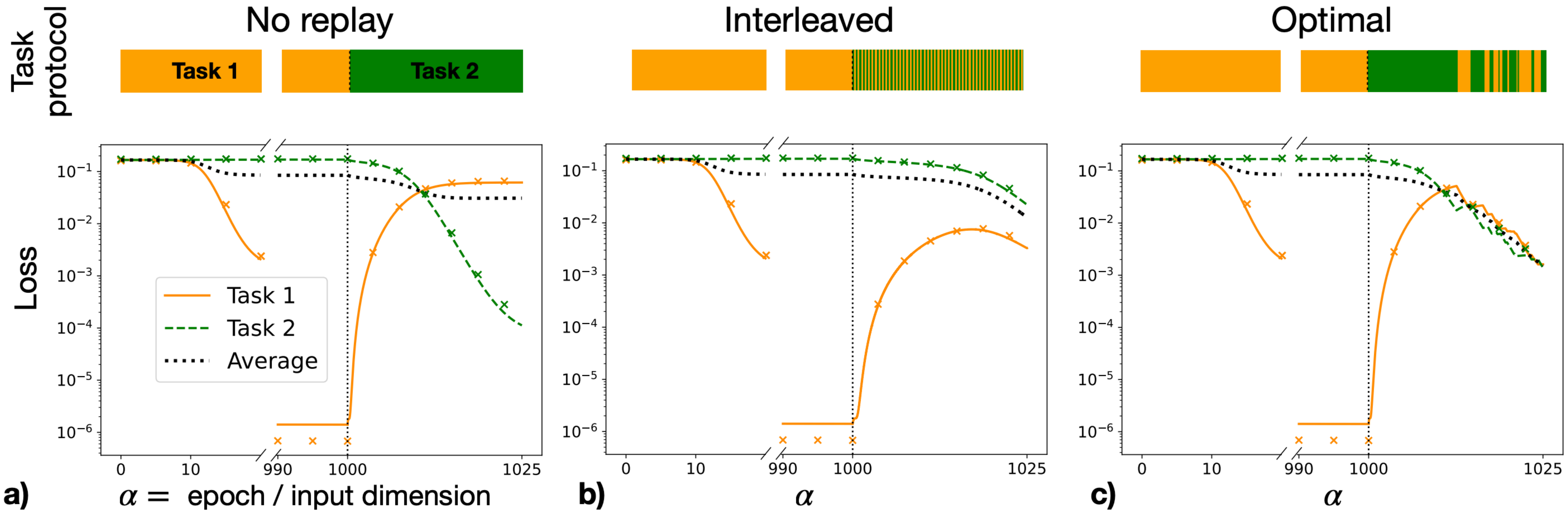


Results: optimal strategy vs benchmarks



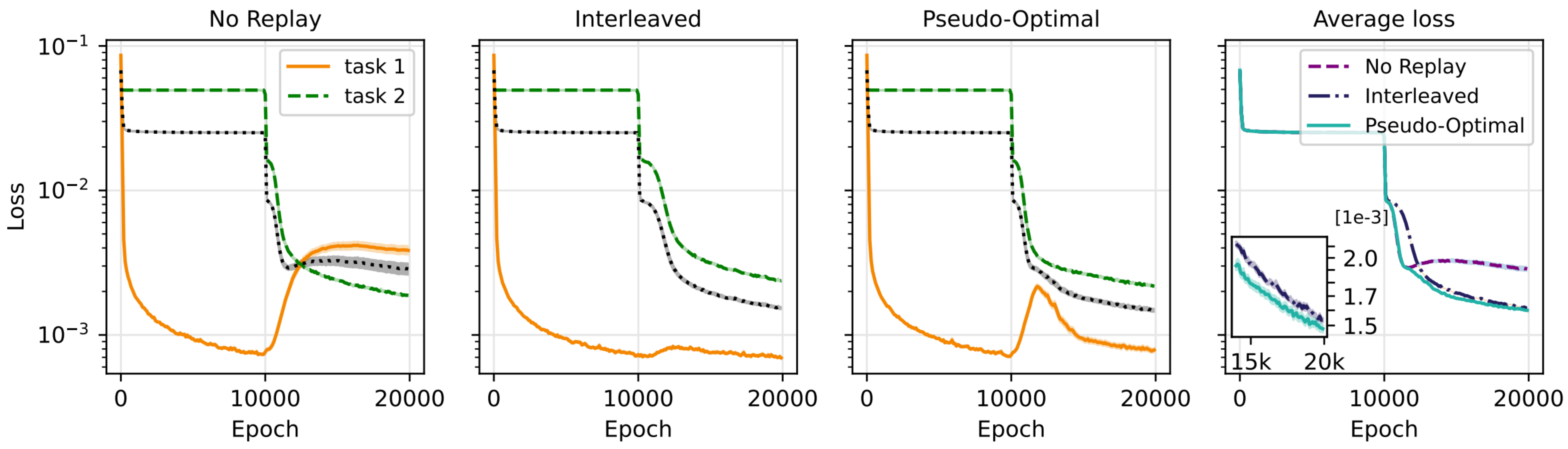
Results: optimal strategy vs benchmarks

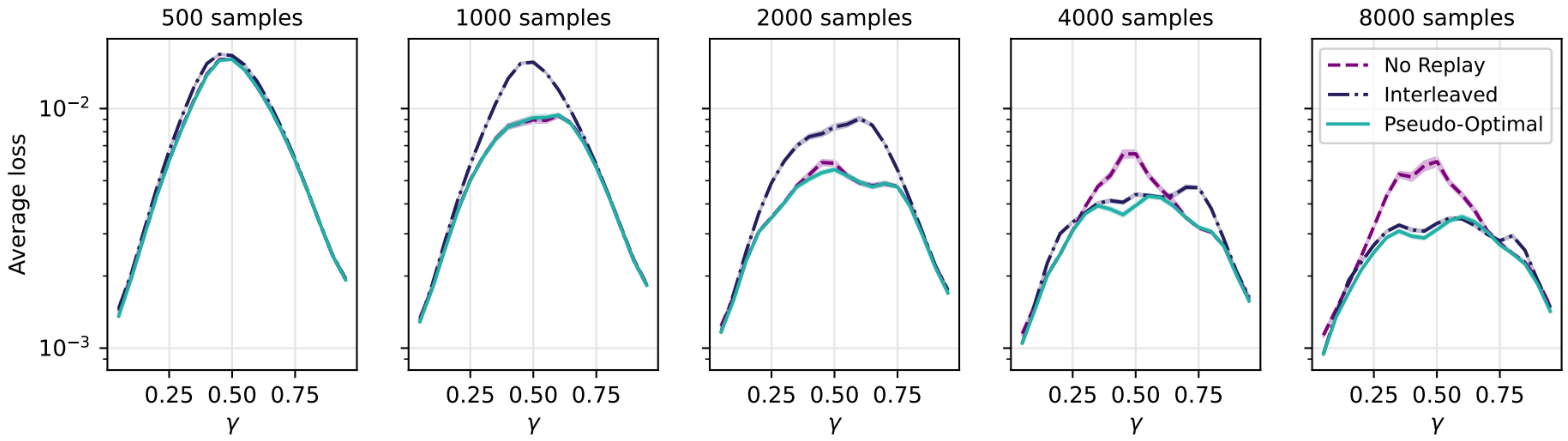
control = task



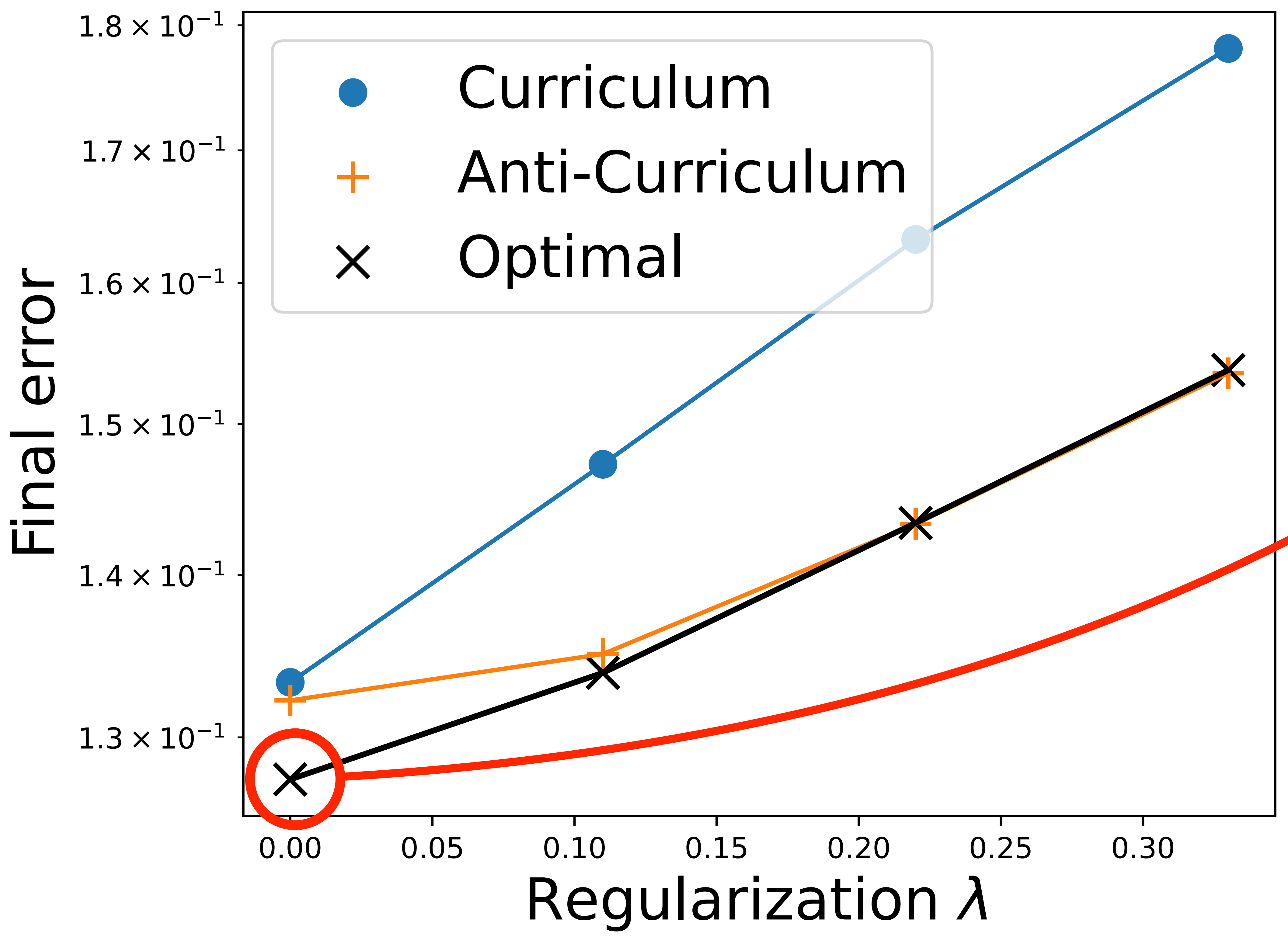
Experiments on Fashion MNIST

$$\mathcal{D}_1 = \{\boldsymbol{x}_i^{(1)}, y_i^{(1)}\}_i \qquad \mathcal{D}_2 = \{\boldsymbol{x}_i^{(2)}, y_i^{(2)}\}_i = \{\gamma \boldsymbol{x}_i^{(1)} + (1-\gamma)\tilde{\boldsymbol{x}}_i, \gamma y_i^{(1)} + (1-\gamma)\tilde{y}_i\}_i$$

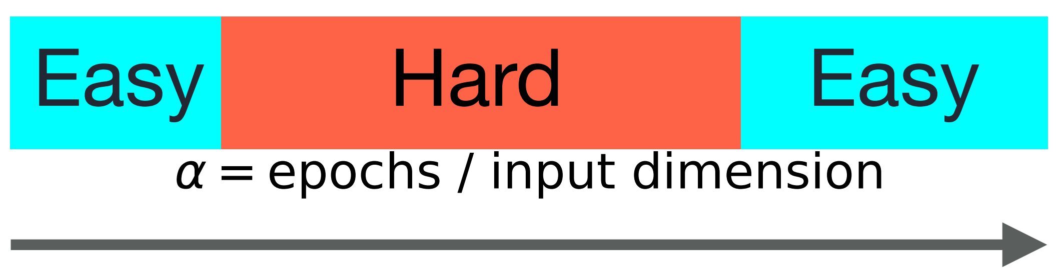




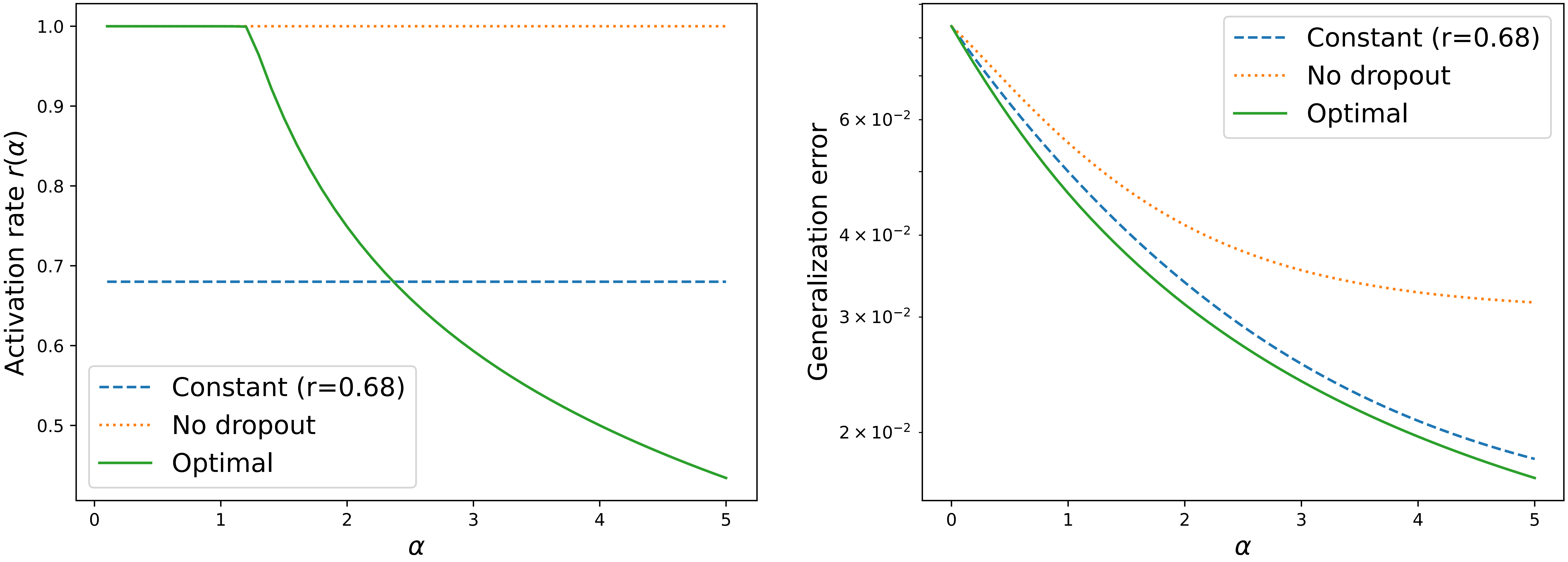
Optimal curriculum learning



Non-monotonic curriculum is optimal:



Optimal dropout



Conclusions & Perspectives

In summary:

Many open directions!

Thank you!

Bonus Slides

Dropout

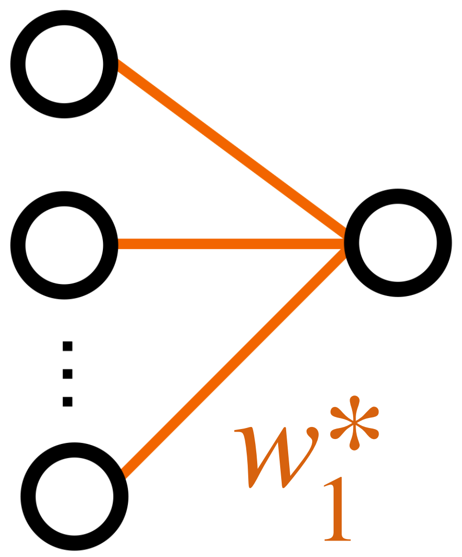
(in progress)

A teacher-student model of dropout

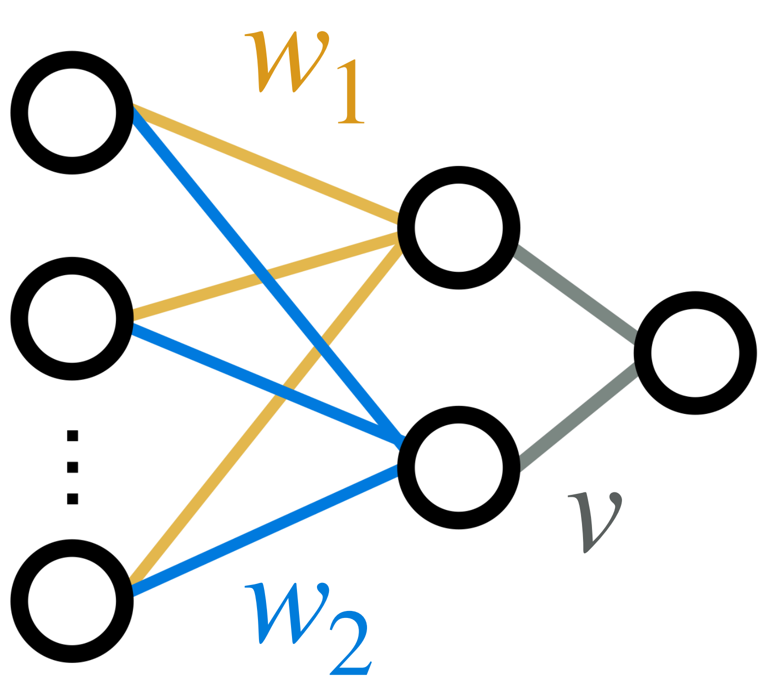
N. Srivastava, G. Hinton, et al., J. ML Res. (2014)

Teacher

$$y = g_*\left(w_1^* \cdot \frac{x}{\sqrt{N}}\right)$$



Student



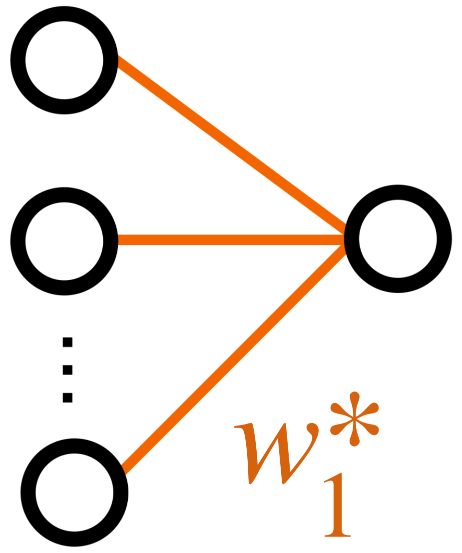
$$\hat{y} = \sum_{k=1}^{K=2} v_k\, g\left(w_k \cdot \frac{x}{\sqrt{N}}\right)$$

A teacher-student model of dropout

N. Srivastava, G. Hinton, et al., J. ML Res. (2014)

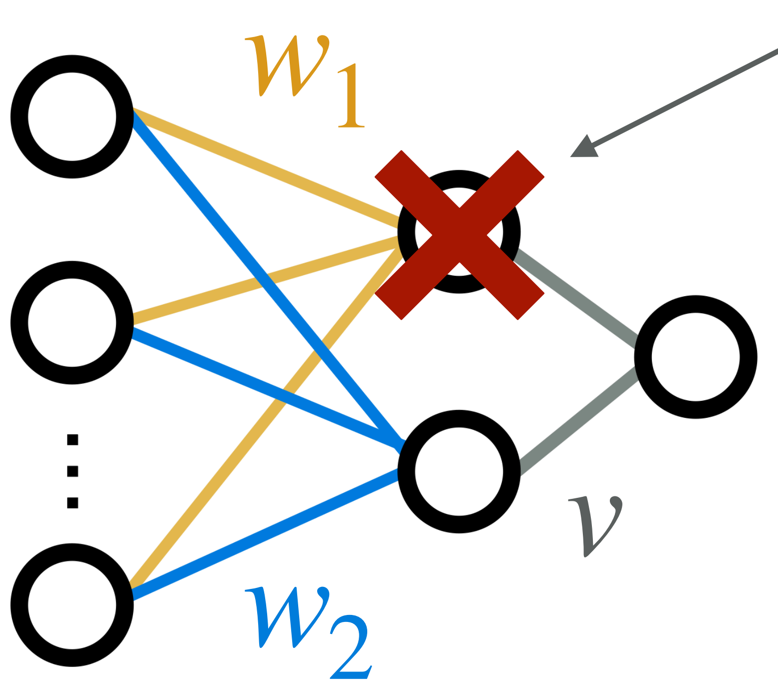
Teacher

$$y = g_*\left(w_1^* \cdot \frac{x}{\sqrt{N}}\right)$$



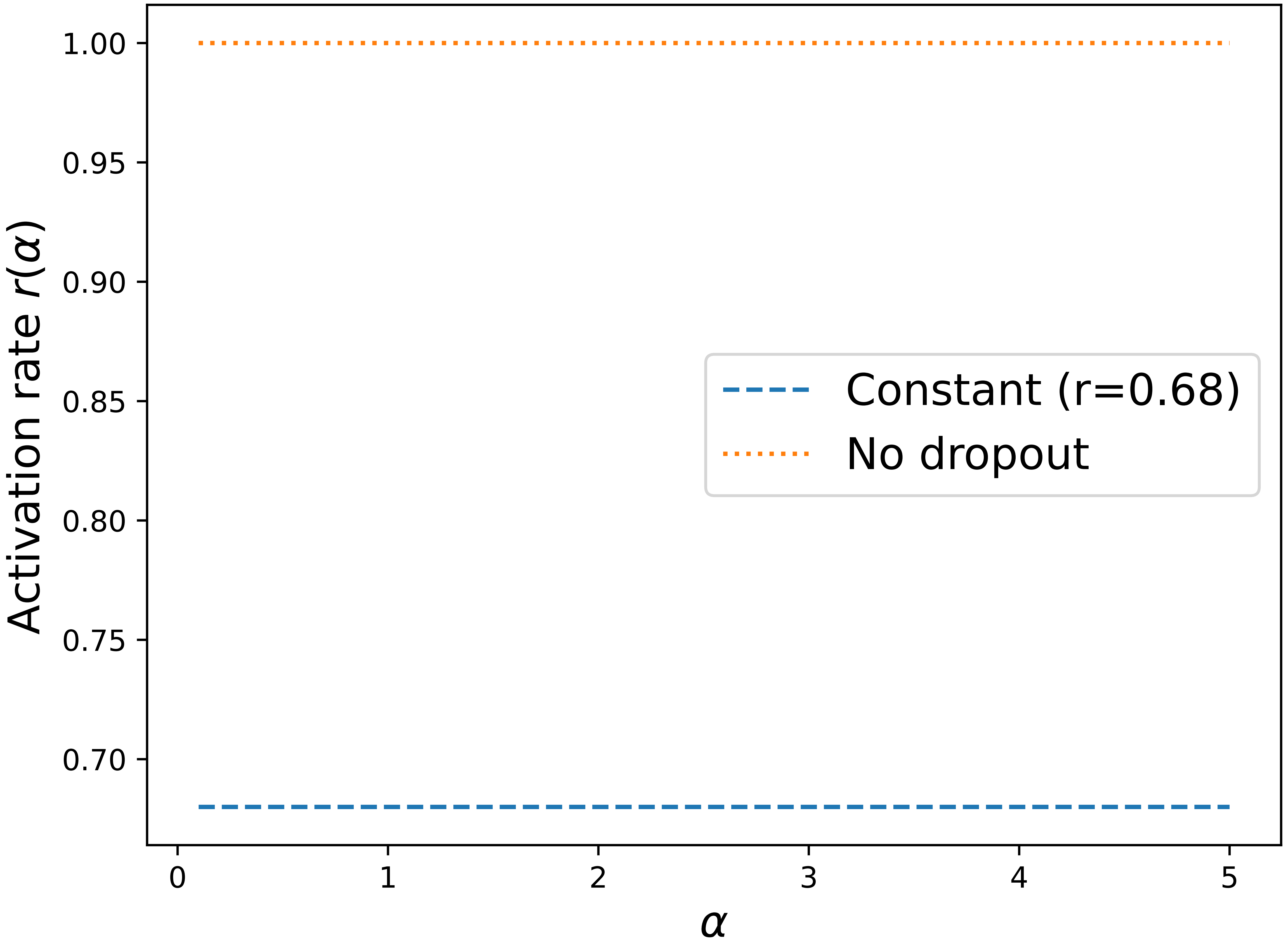
Student

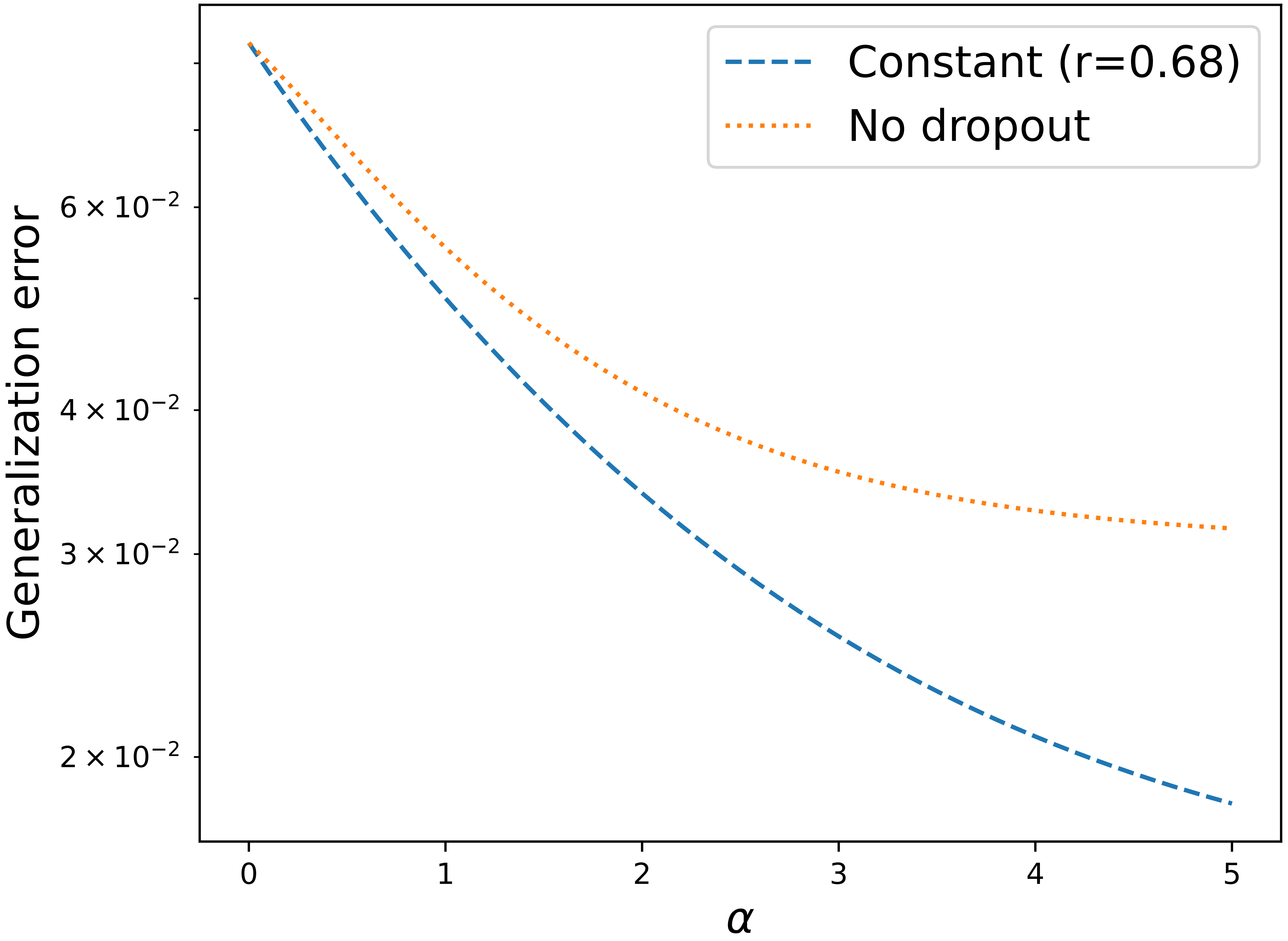
With probability $1 - r$



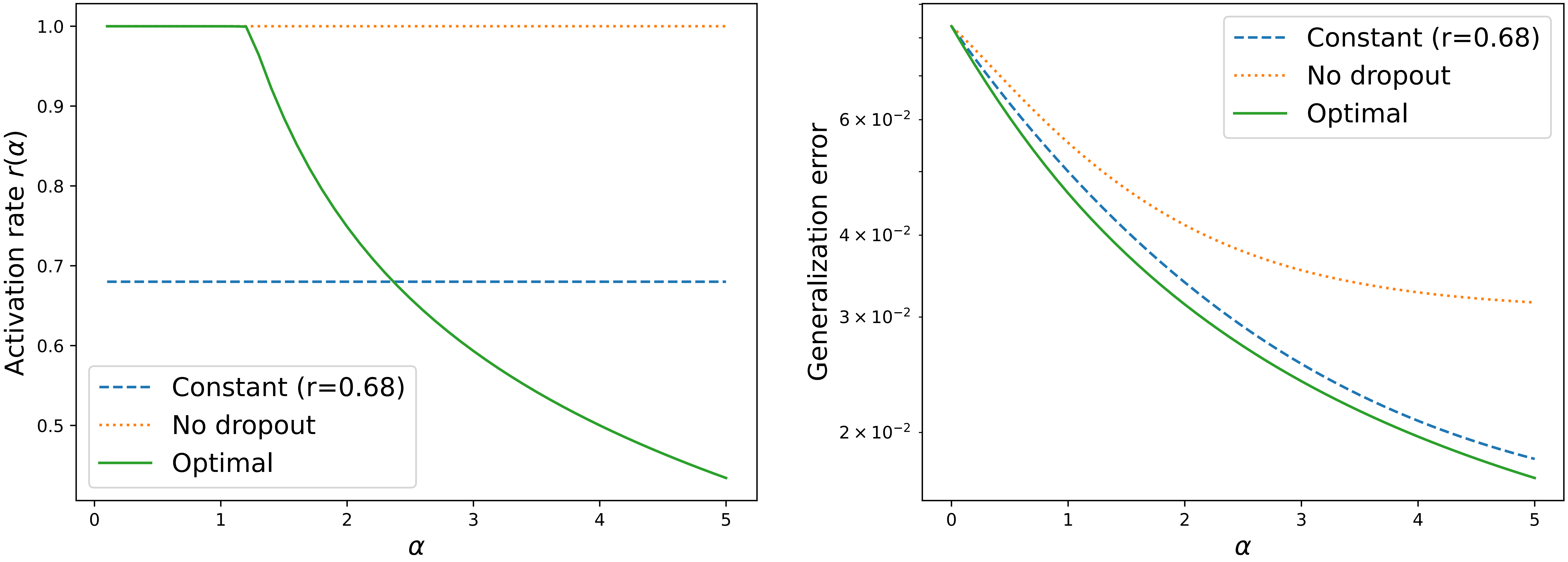
$$\hat{y} = v_2\, g\left(w_2 \cdot \frac{x}{\sqrt{N}}\right)$$

A teacher-student model of dropout

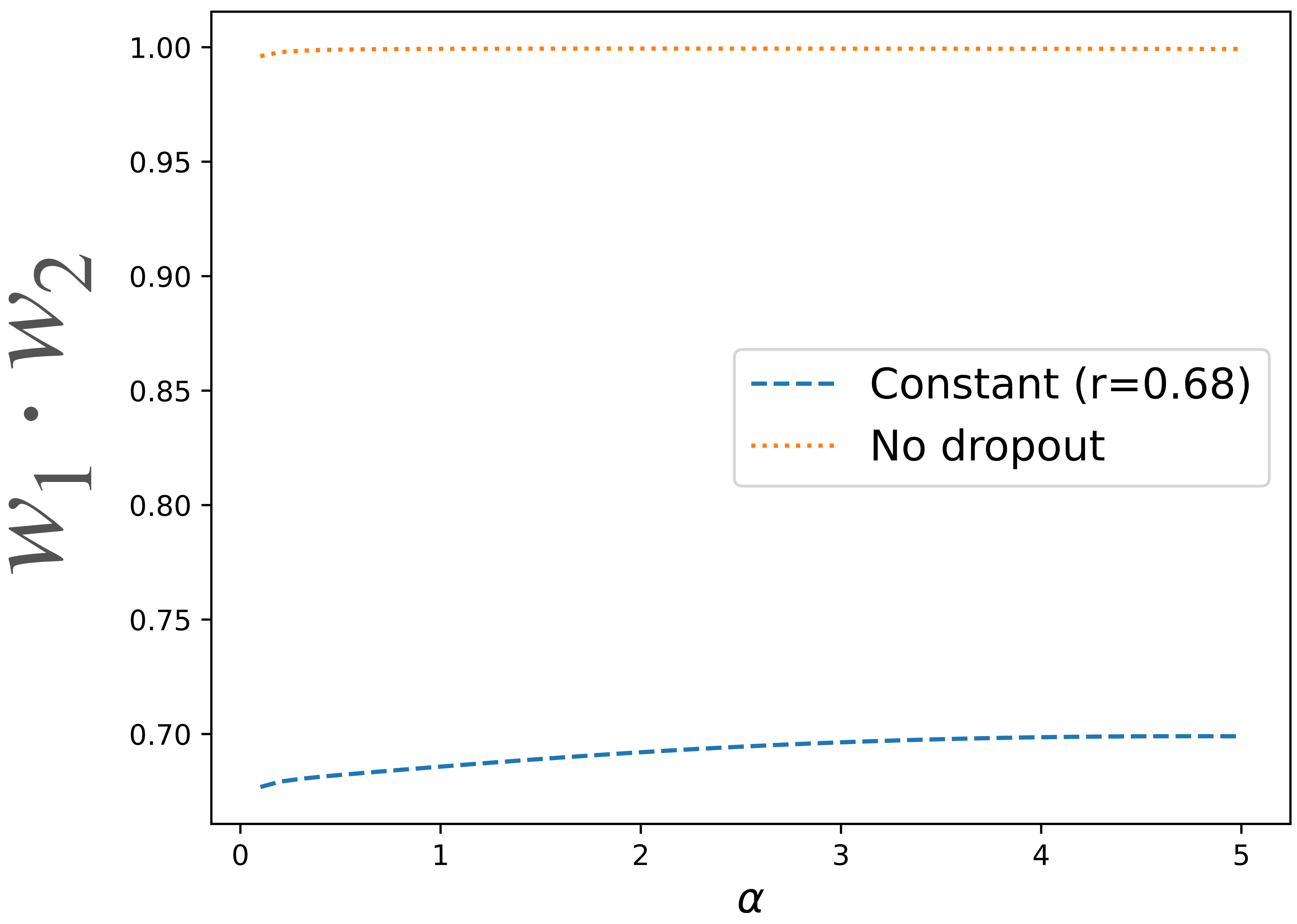


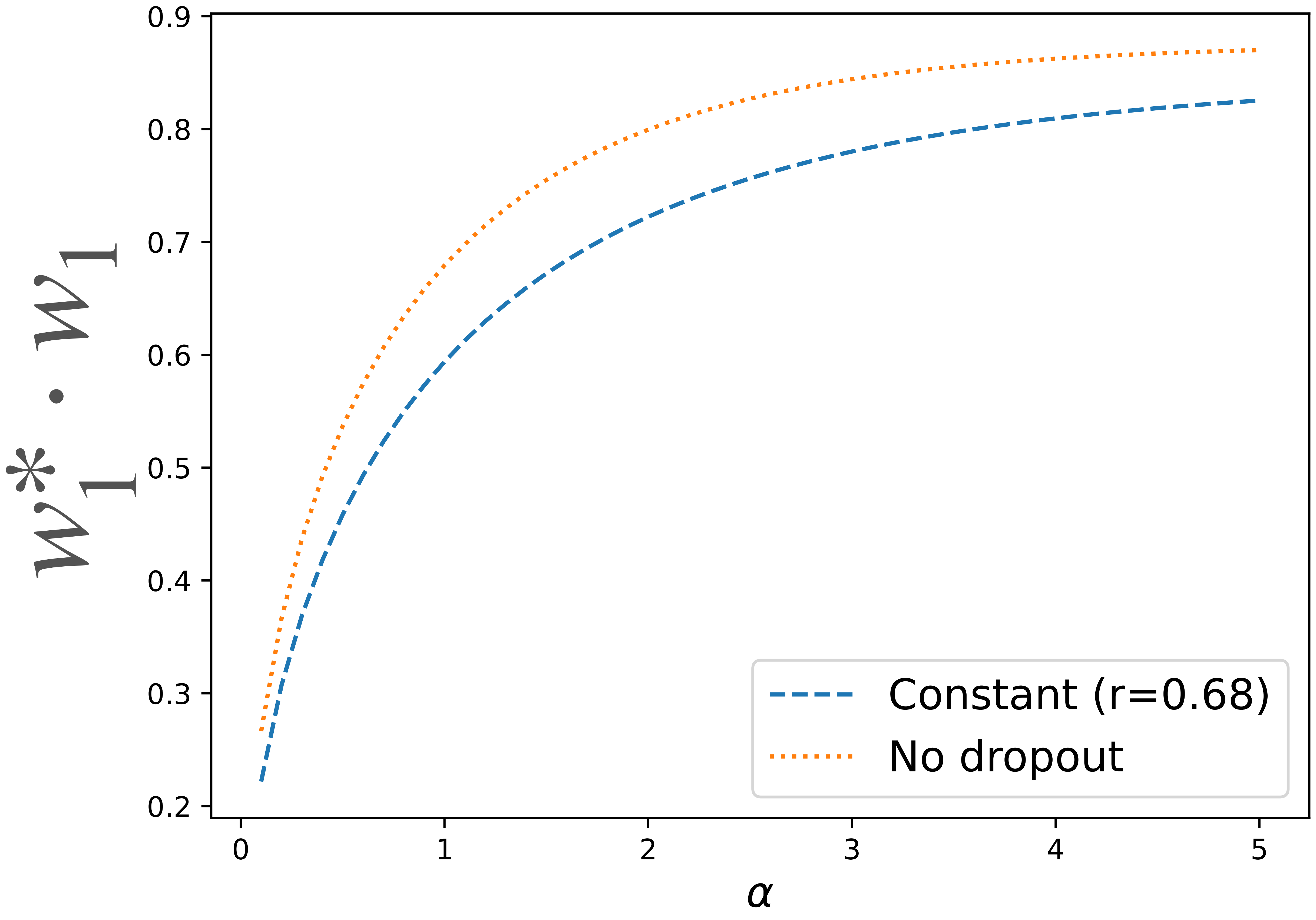


A teacher-student model of dropout

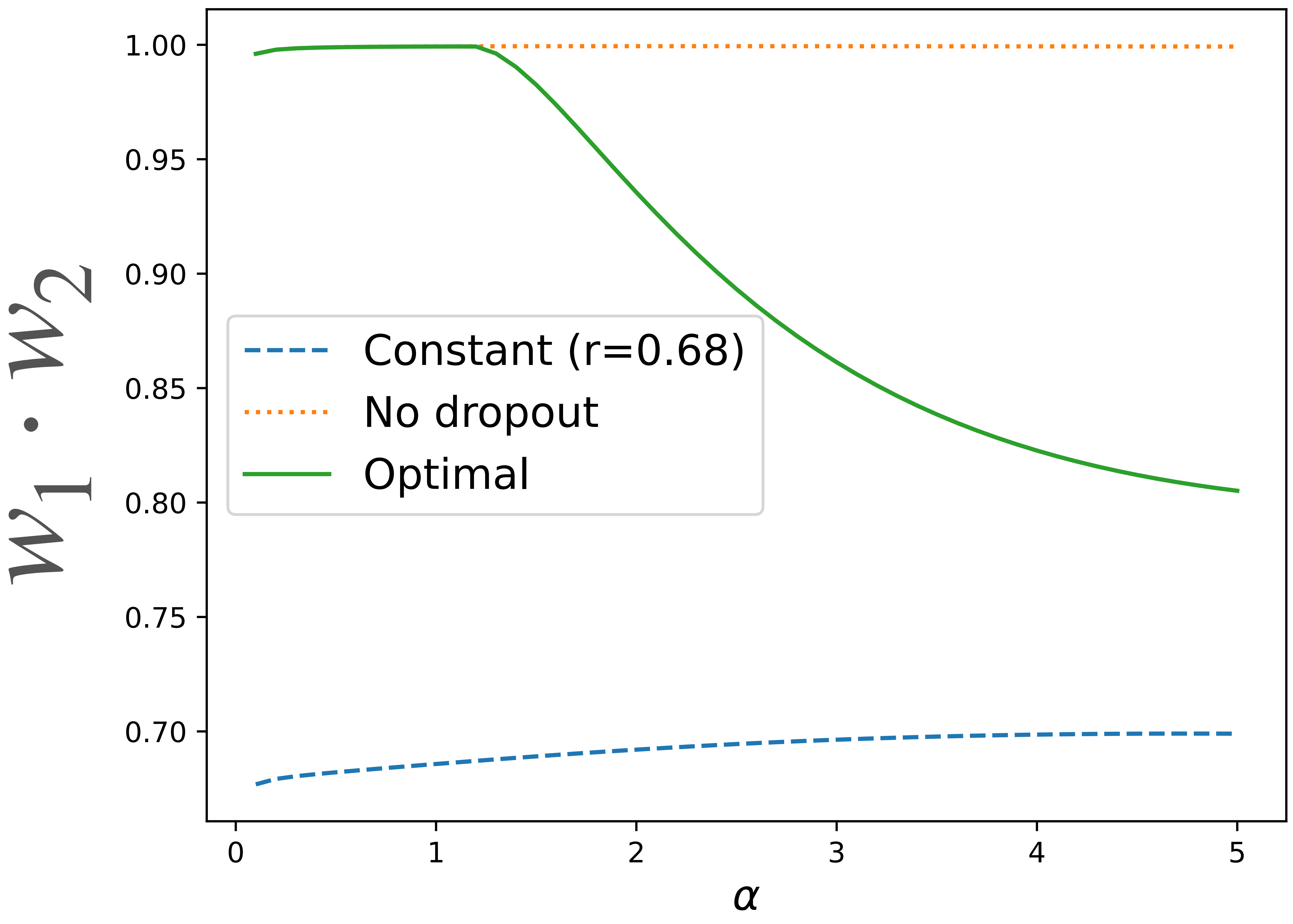


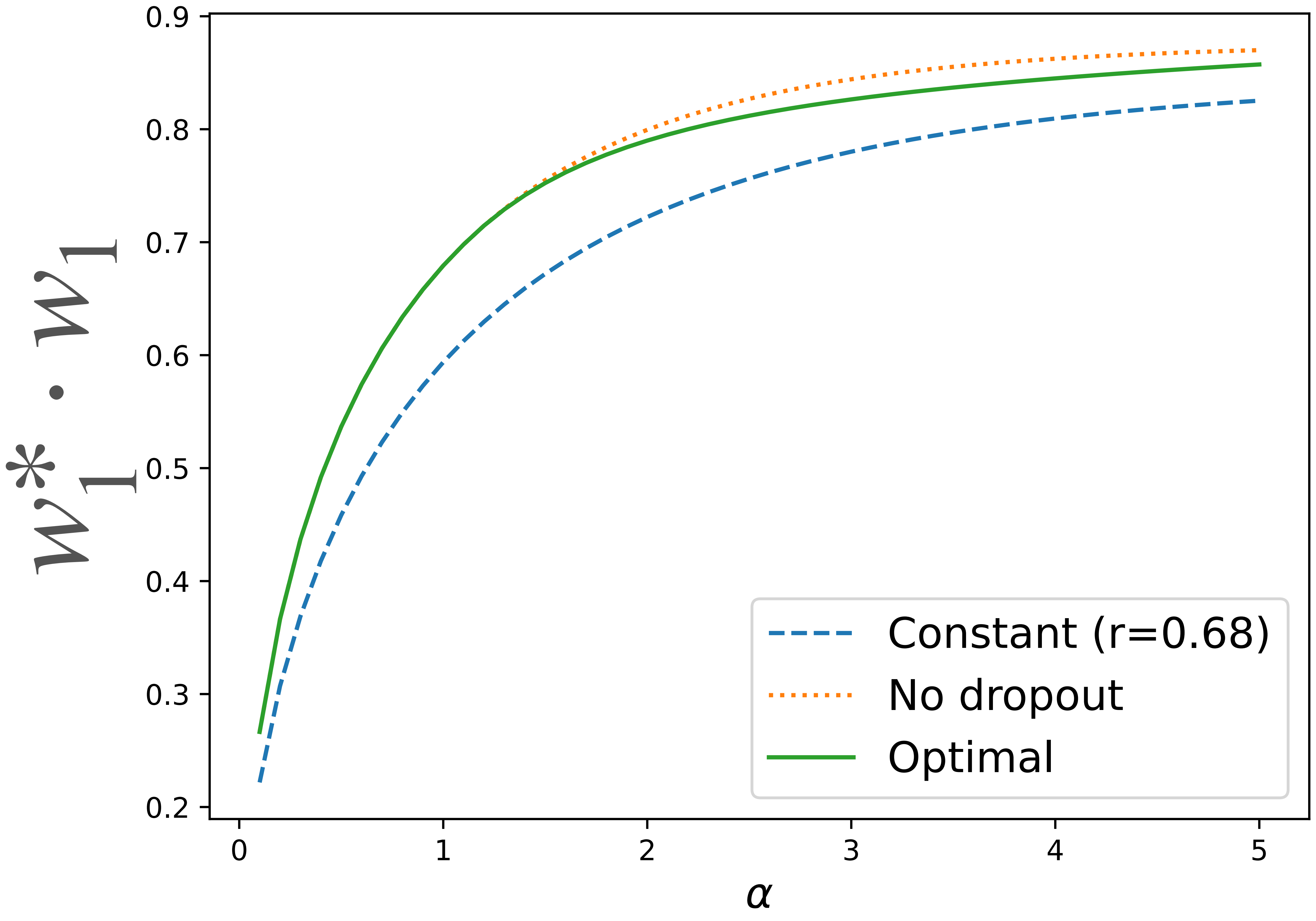
A teacher-student model of dropout





A teacher-student model of dropout





Curriculum learning

(in progress)

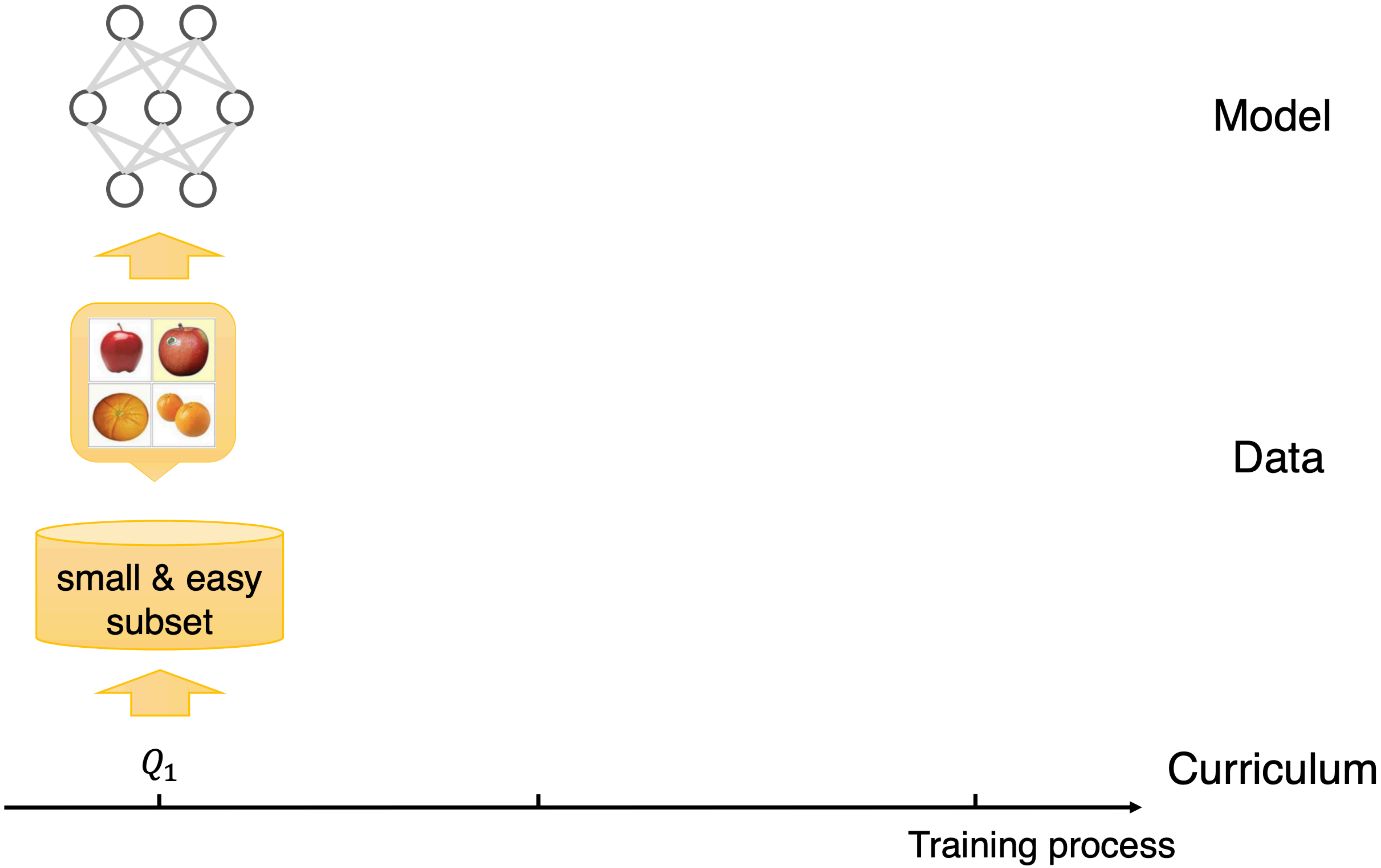


Image from: *Wang, Xin, Yudong Chen, and Wenwu Zhu. IEEE transactions on pattern analysis and machine intelligence 44.9 (2021):4555-4576.*

Curriculum learning

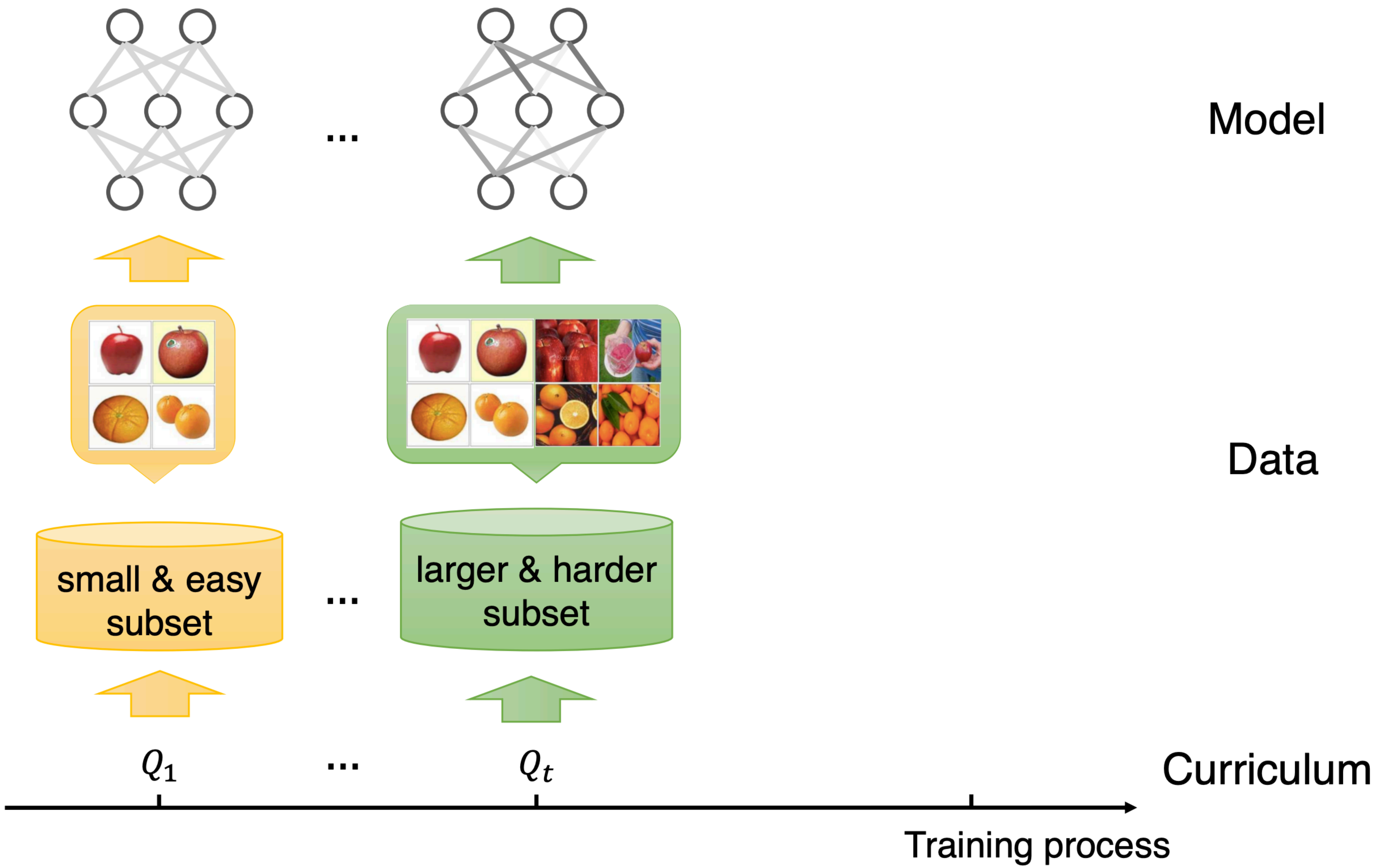


Image from: *Wang, Xin, Yudong Chen, and Wenwu Zhu. IEEE transactions on pattern analysis and machine intelligence 44.9 (2021):4555-4576.*

Curriculum learning

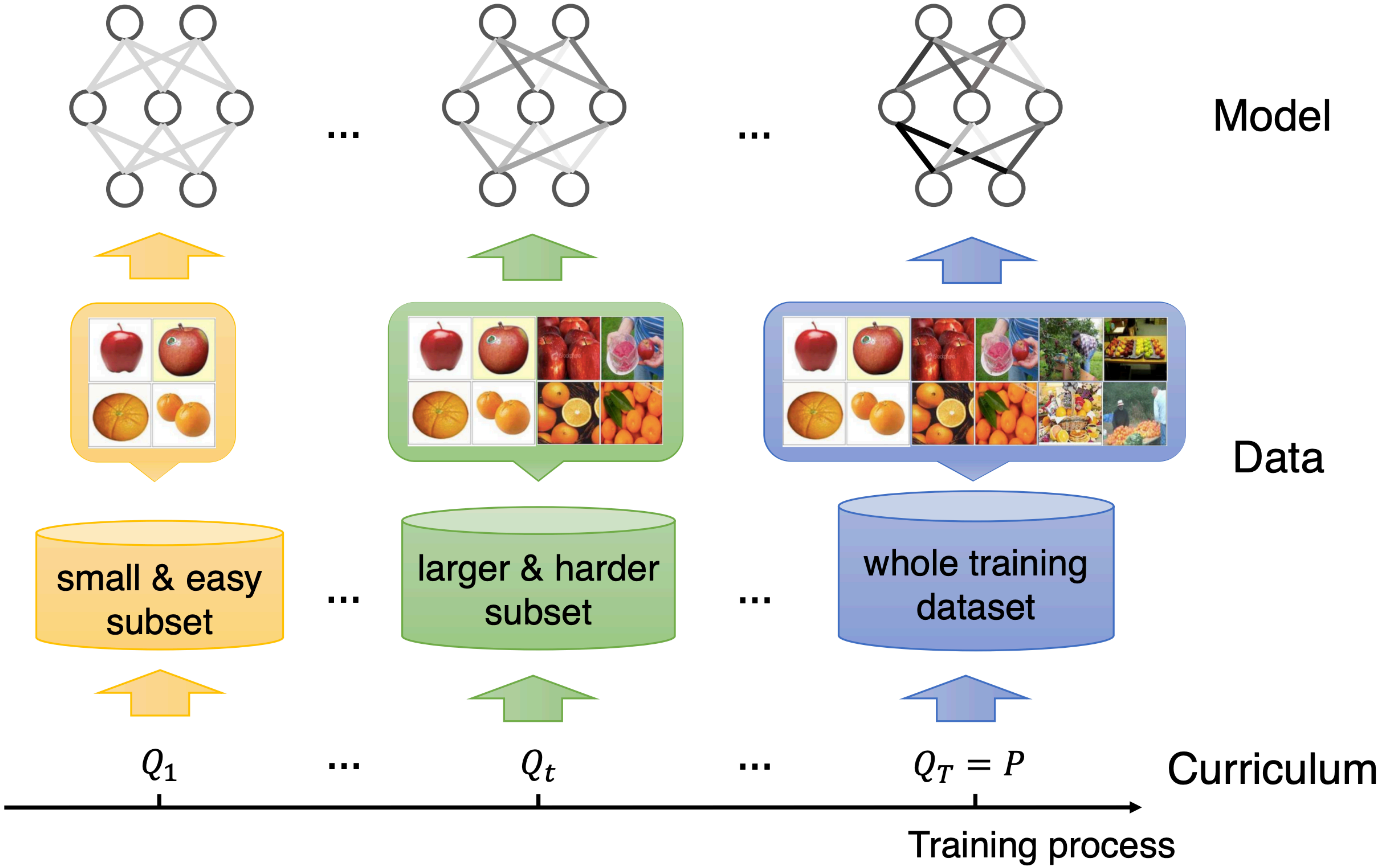


Image from: *Wang, Xin, Yudong Chen, and Wenwu Zhu. IEEE transactions on pattern analysis and machine intelligence 44.9 (2021):4555-4576.*

Curriculum learning

Animals:

Humans:

Curriculum learning

Animals:

Humans:

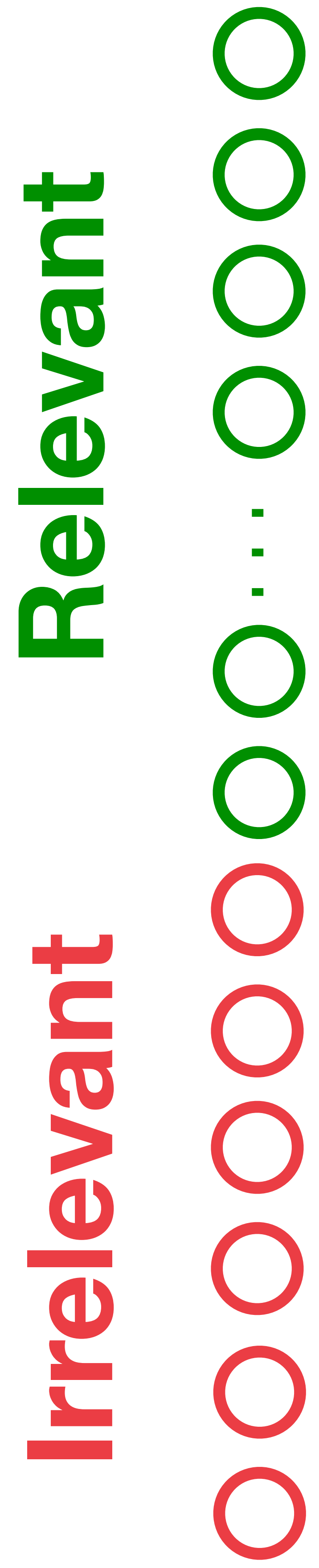
ML (empirical):

ML (theory):

A teacher-student model of curriculum learning

Introduced in: *Bengio, et al. (ICML 2009), Saglietti, et al. (NeurIPS 2022)*

Input: $\mathbf{x} = (\mathbf{x}_r, \mathbf{x}_i) \in \mathbb{R}^N$



$\mathbf{x}_r \in \mathbb{R}^{\rho N}$

Unit variance

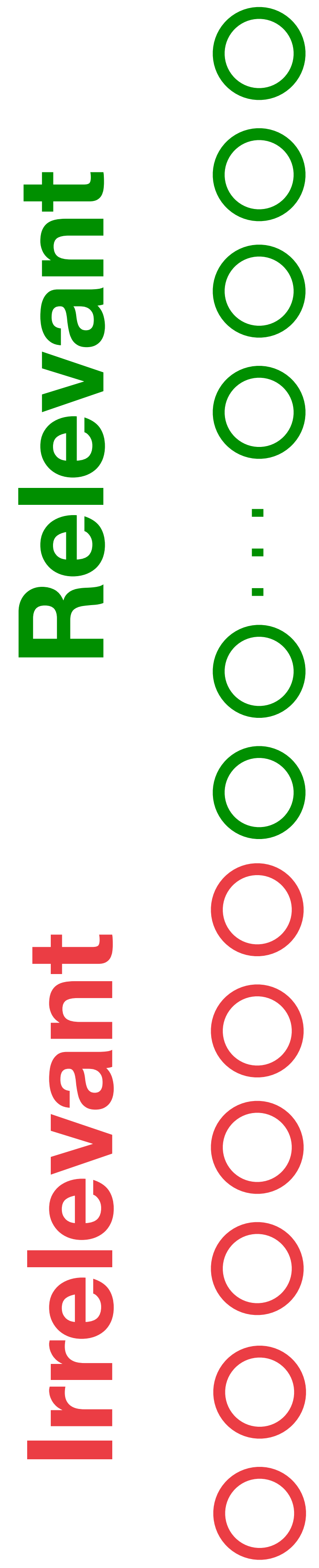
$\mathbf{x}_i \in \mathbb{R}^{(1-\rho)N}$

Variance Δ

A teacher-student model of curriculum learning

Introduced in: *Bengio, et al. (ICML 2009), Saglietti, et al. (NeurIPS 2022)*

Input: $\mathbf{x} = (\mathbf{x}_r, \mathbf{x}_i) \in \mathbb{R}^N$



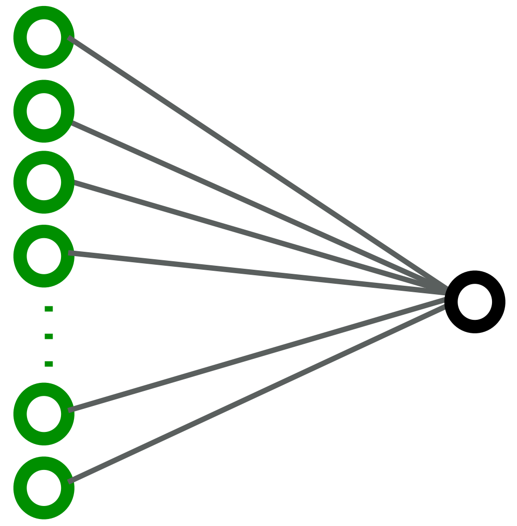
$\mathbf{x}_r \in \mathbb{R}^{\rho N}$

Unit variance

$\mathbf{x}_i \in \mathbb{R}^{(1-\rho)N}$

Variance Δ

Teacher



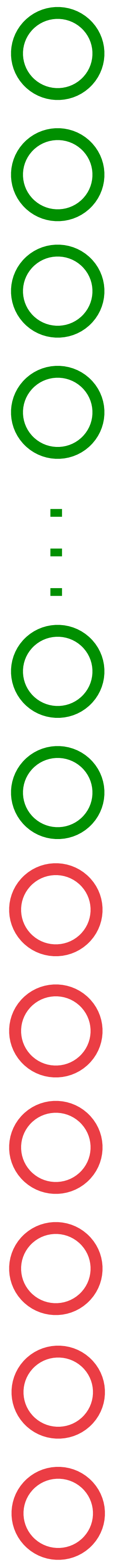
$y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}_r)$

A teacher-student model of curriculum learning

Introduced in: *Bengio, et al. (ICML 2009), Saglietti, et al. (NeurIPS 2022)*

Input: $\mathbf{x} = (\mathbf{x}_r, \mathbf{x}_i) \in \mathbb{R}^N$

Relevant



$\mathbf{x}_r \in \mathbb{R}^{\rho N}$

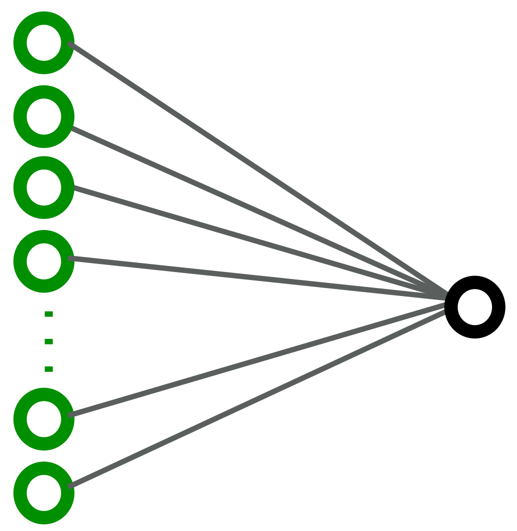
Unit variance

Irrelevant

$\mathbf{x}_i \in \mathbb{R}^{(1-\rho)N}$

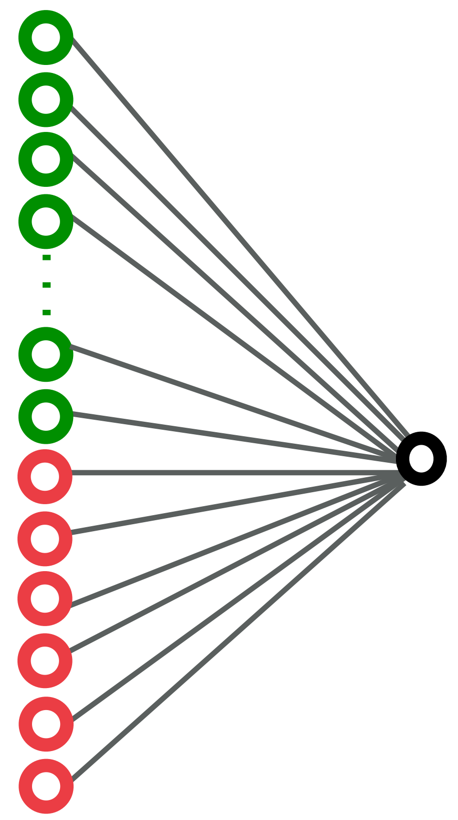
Variance Δ

Teacher



$$y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}_r)$$

Student



$$y = \mathrm{erf}\left(\frac{\mathbf{w} \cdot \mathbf{x}}{\sqrt{2}}\right)$$

Ridge-regularized MSE loss:

$$\mathscr{L} = \frac{1}{2}(y - \hat{y})^2 + \lambda \|\mathbf{w}\|_2^2$$

A teacher-student model of curriculum learning

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

Luca Saglietti[†,*], Stefano Sarao Mannelli[‡,*], and Andrew Saxe[‡,§]

The evolution of the dynamics can be tracked using four order parameters:

$$Q_r = \frac{1}{N}\boldsymbol{W}_r \cdot \boldsymbol{W}_r, \qquad R = \frac{1}{N}\boldsymbol{W}_r \cdot \boldsymbol{W}_T,$$

$$Q_i = \frac{1}{N}\boldsymbol{W}_i \cdot \boldsymbol{W}_i, \qquad T = \frac{1}{N}\boldsymbol{W}_T \cdot \boldsymbol{W}_T\,;$$

A teacher-student model of curriculum learning

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

Luca Saglietti[†,*], Stefano Sarao Mannelli[‡,*], and Andrew Saxe[‡,§]

The evolution of the dynamics can be tracked using four order parameters:

$$Q_r = \frac{1}{N} \boldsymbol{W}_r \cdot \boldsymbol{W}_r, \qquad R = \frac{1}{N} \boldsymbol{W}_r \cdot \boldsymbol{W}_T,$$

$$Q_i = \frac{1}{N} \boldsymbol{W}_i \cdot \boldsymbol{W}_i, \qquad T = \frac{1}{N} \boldsymbol{W}_T \cdot \boldsymbol{W}_T;$$

$$Q_r \leftarrow f_{Q_r}(Q_r, Q_i, R, T)$$

$$Q_i \leftarrow f_{Q_i}(Q_r, Q_i, R, T)$$

$$R \leftarrow f_R(Q_r, Q_i, R, T)$$

A teacher-student model of curriculum learning

An Analytical Theory of Curriculum Learning in Teacher-Student Networks

Luca Saglietti[†,*], Stefano Sarao Mannelli[‡,*], and Andrew Saxe[‡,§]

The evolution of the dynamics can be tracked using four order parameters:

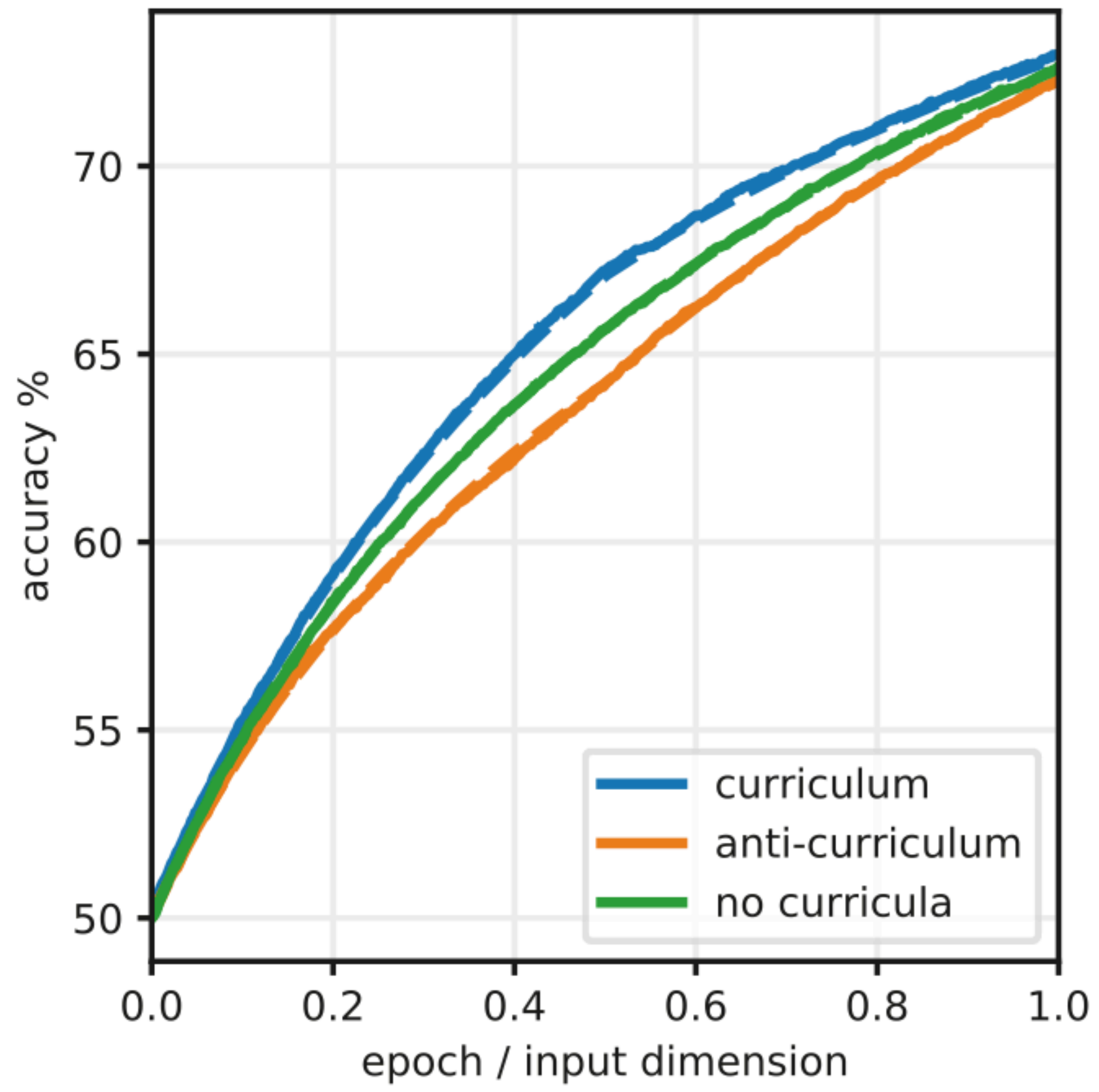
$$Q_r = \frac{1}{N}\boldsymbol{W}_r \cdot \boldsymbol{W}_r, \qquad R = \frac{1}{N}\boldsymbol{W}_r \cdot \boldsymbol{W}_T,$$

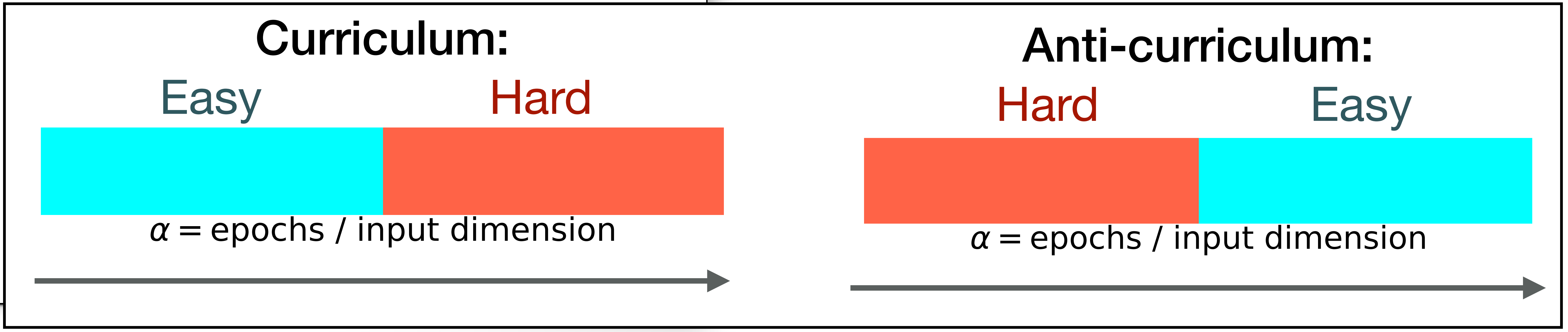
$$Q_i = \frac{1}{N}\boldsymbol{W}_i \cdot \boldsymbol{W}_i, \qquad T = \frac{1}{N}\boldsymbol{W}_T \cdot \boldsymbol{W}_T\,;$$

$$Q_r \leftarrow f_{Q_r}(Q_r, Q_i, R, T)$$

$$Q_i \leftarrow f_{Q_i}(Q_r, Q_i, R, T)$$

$$R \leftarrow f_R(Q_r, Q_i, R, T)$$

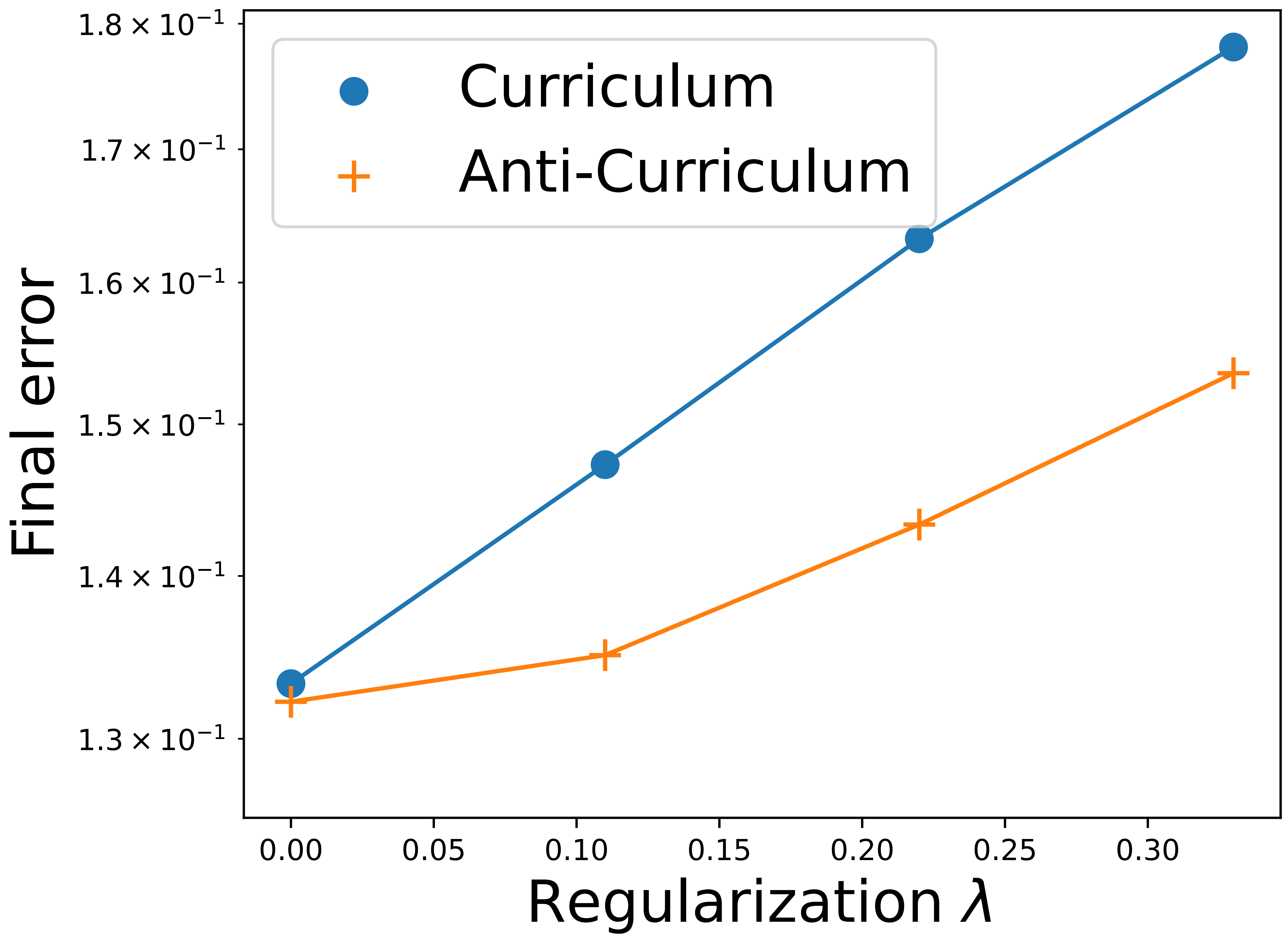




Optimal curriculum protocol

$\rho = 0.55,\ \eta = 2.58$

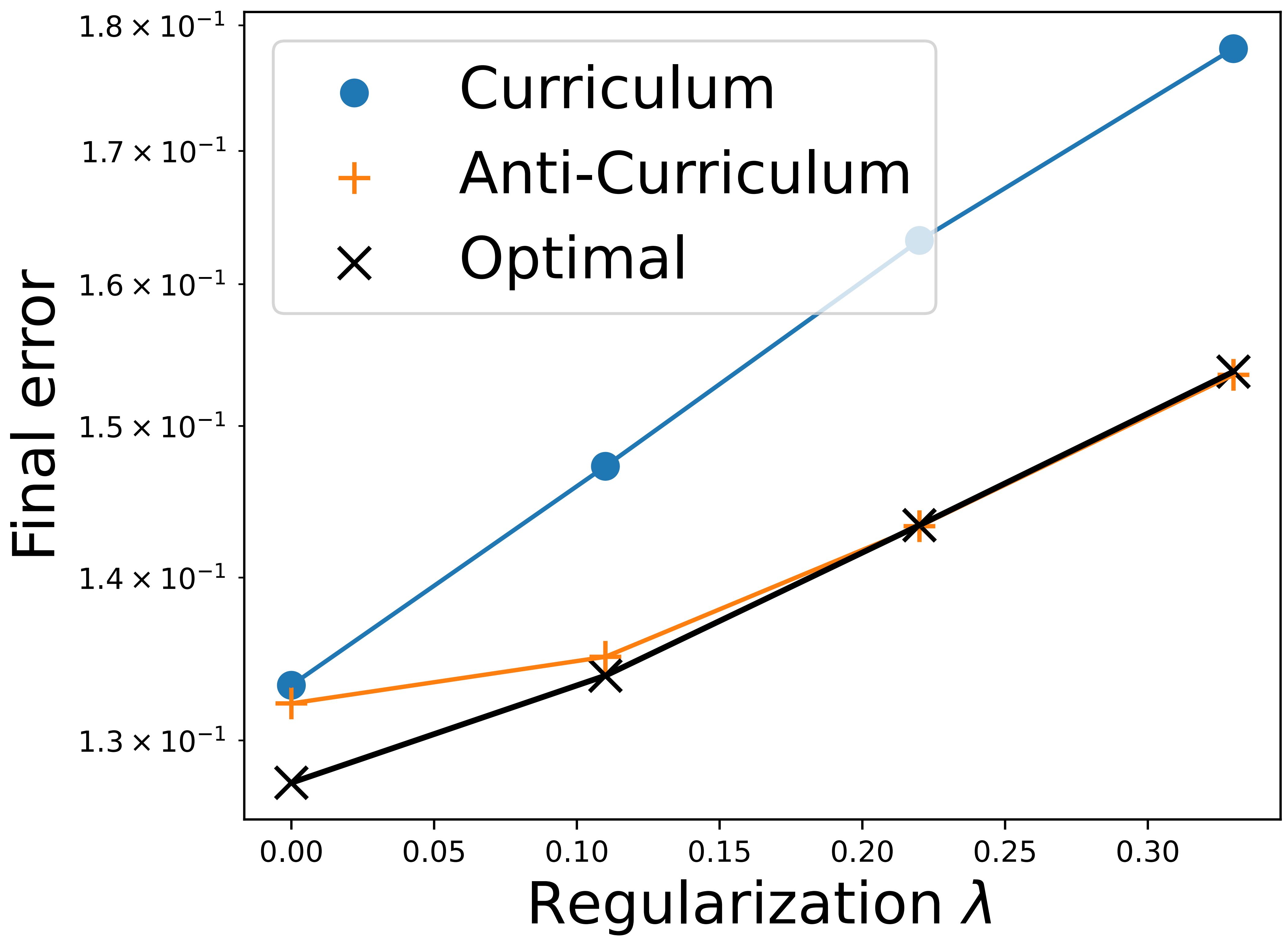
Control: $\mathbf{u} = \boldsymbol{\Delta}$



Optimal curriculum protocol

$\rho = 0.55, \eta = 2.58$

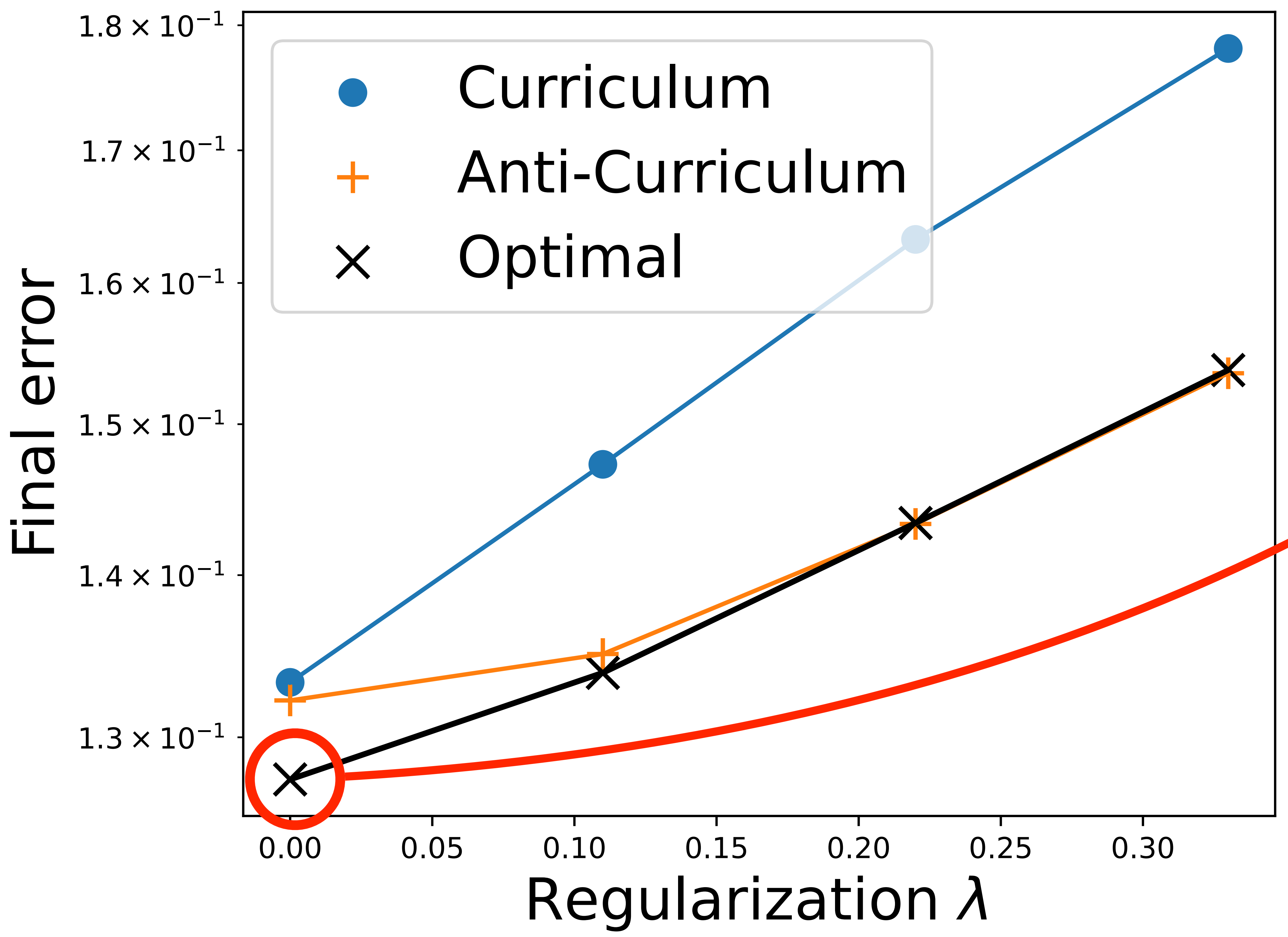
Control: $\mathbf{u} = \boldsymbol{\Delta}$



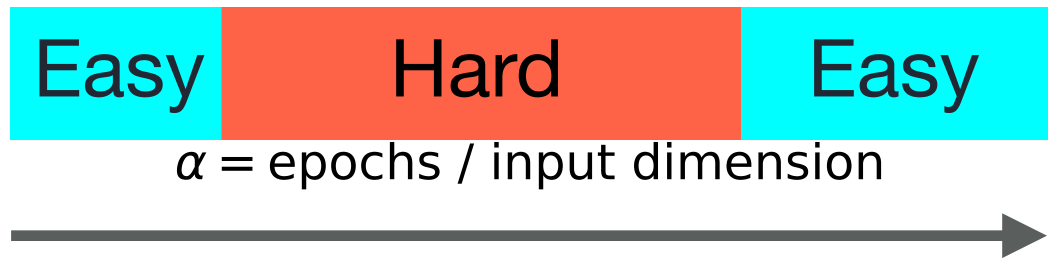
Optimal curriculum protocol

Control: $\mathbf{u} = \mathbf{\Delta}$

$\rho = 0.55, \eta = 2.58$



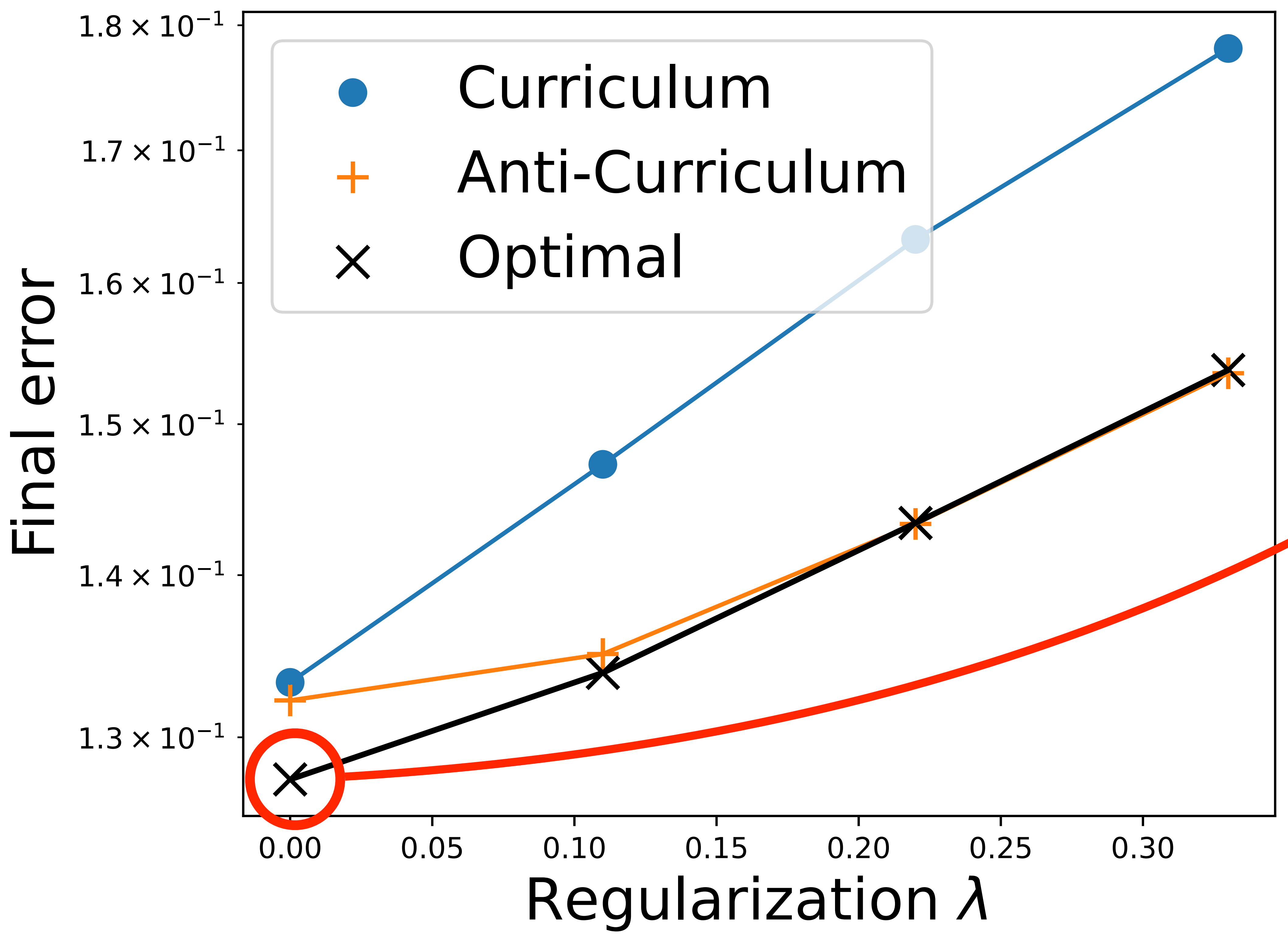
Non-monotonic curriculum is optimal:



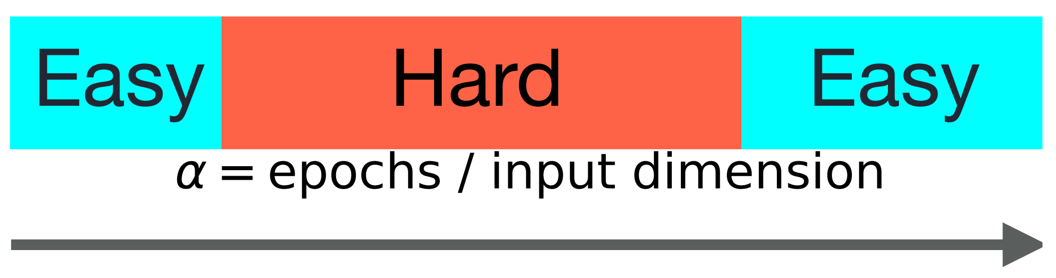
Optimal curriculum protocol

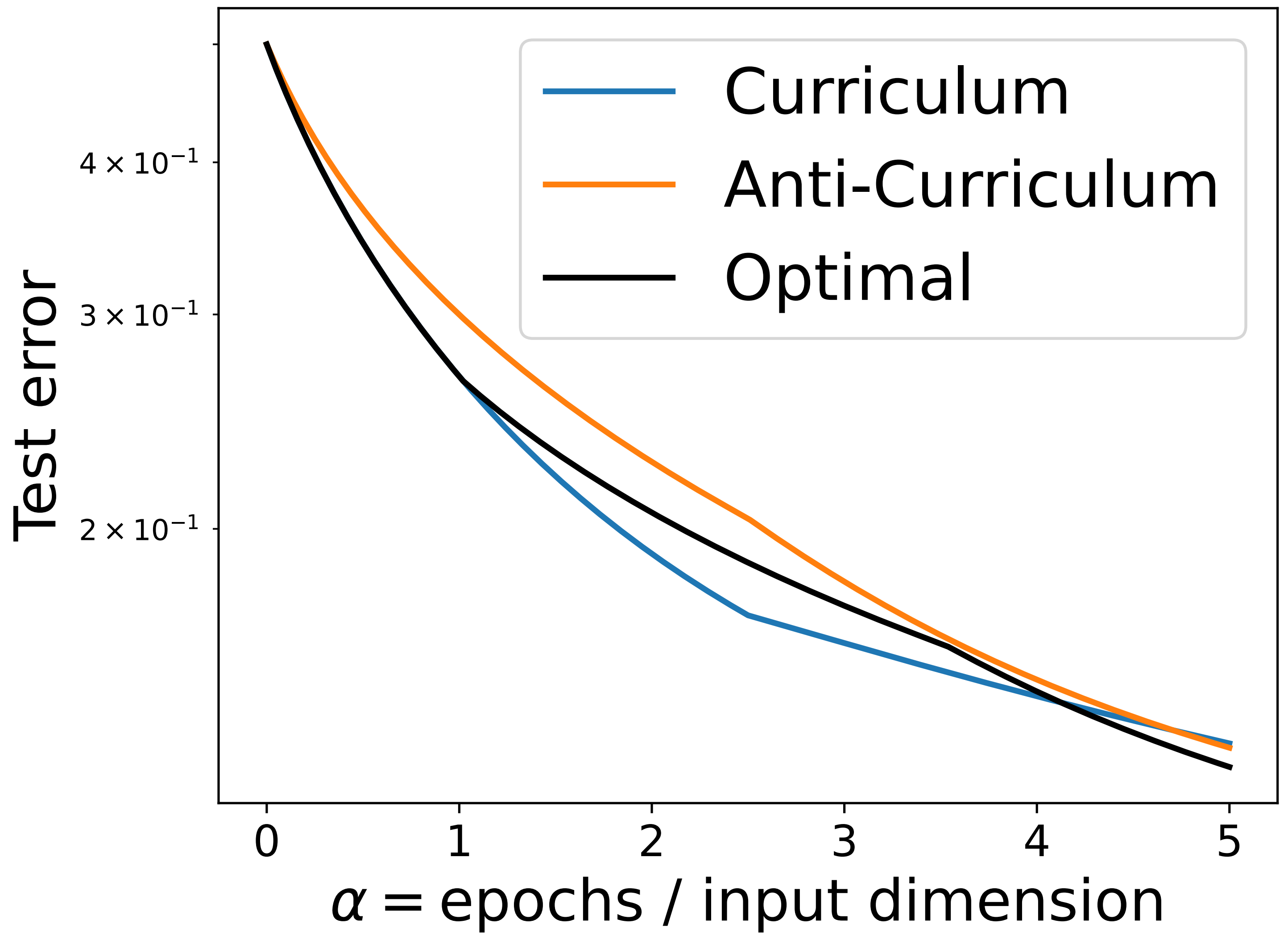
Control: $\mathbf{u} = \boldsymbol{\Delta}$

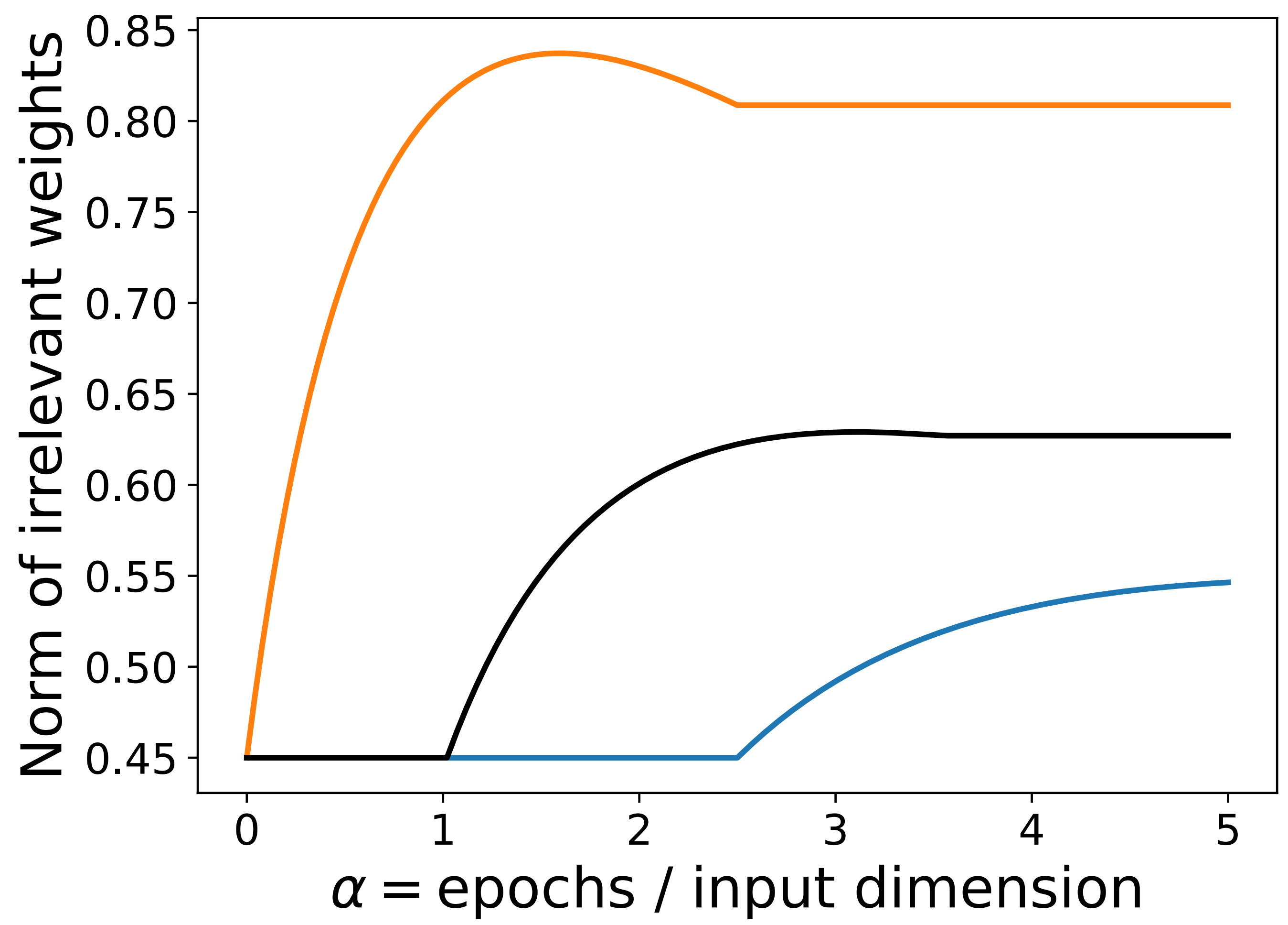
$\rho = 0.55,\ \eta = 2.58$

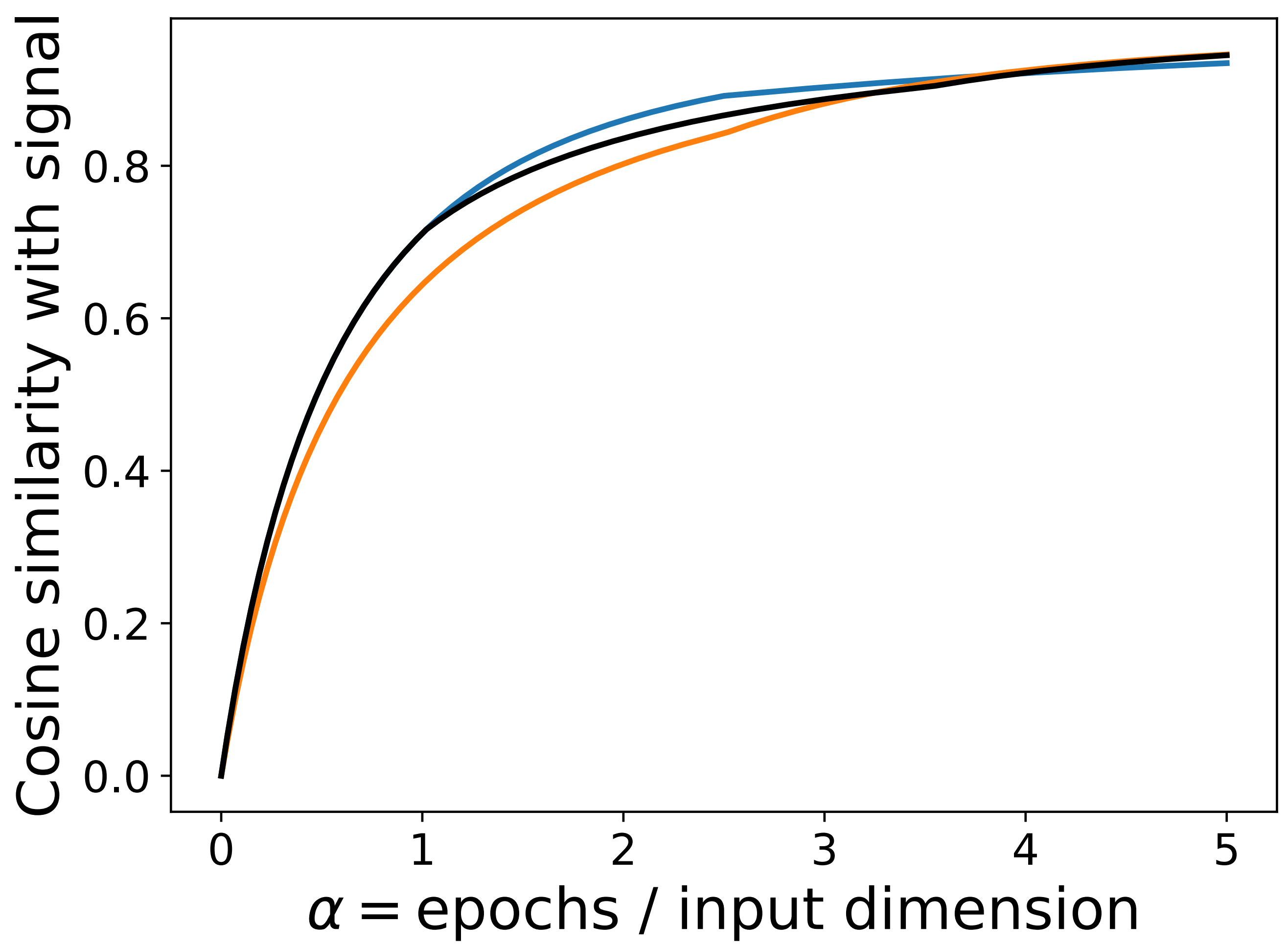


Non-monotonic curriculum is optimal:





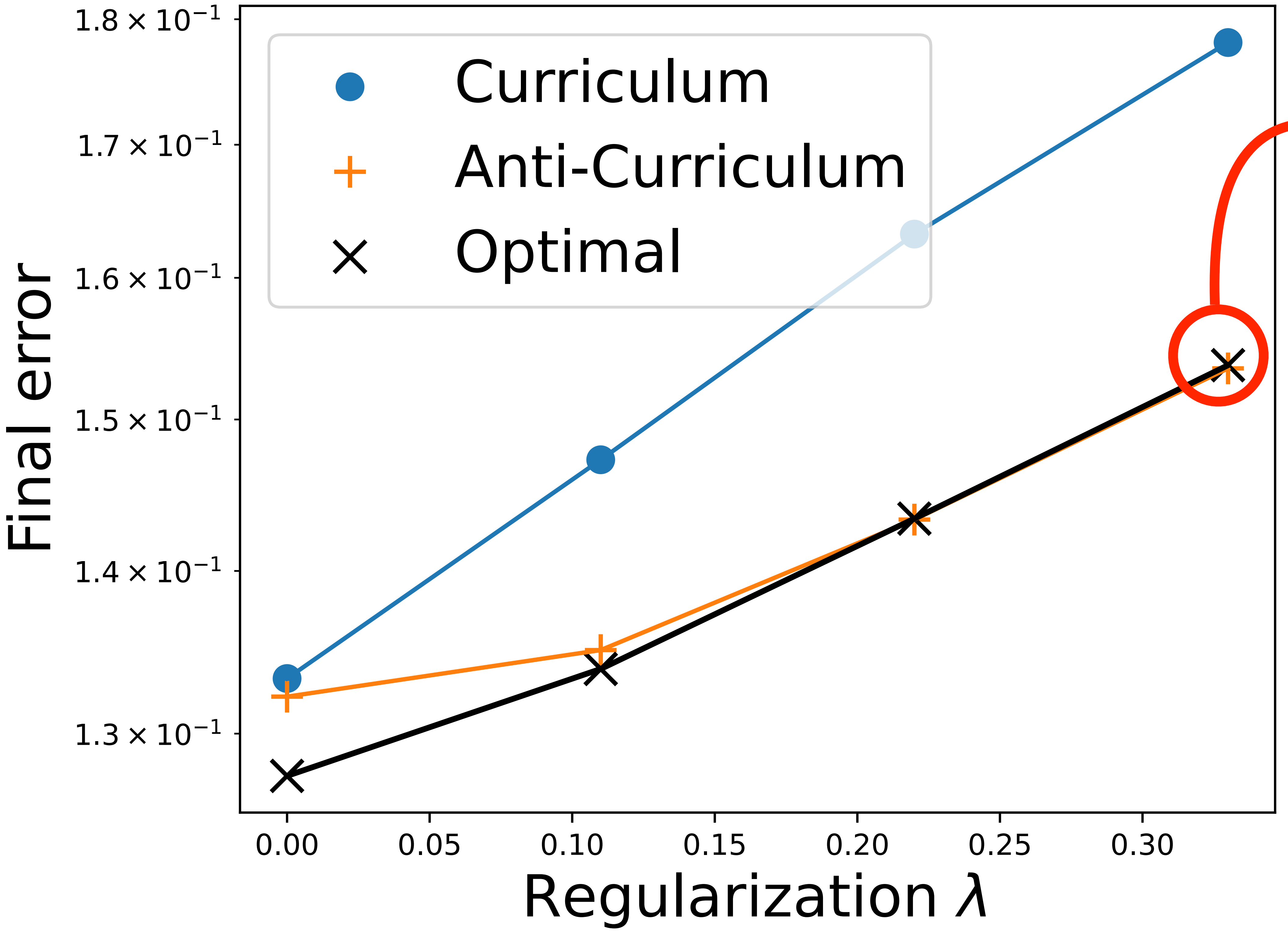




Optimal curriculum protocol

Control: $\mathbf{u} = \boldsymbol{\Delta}$

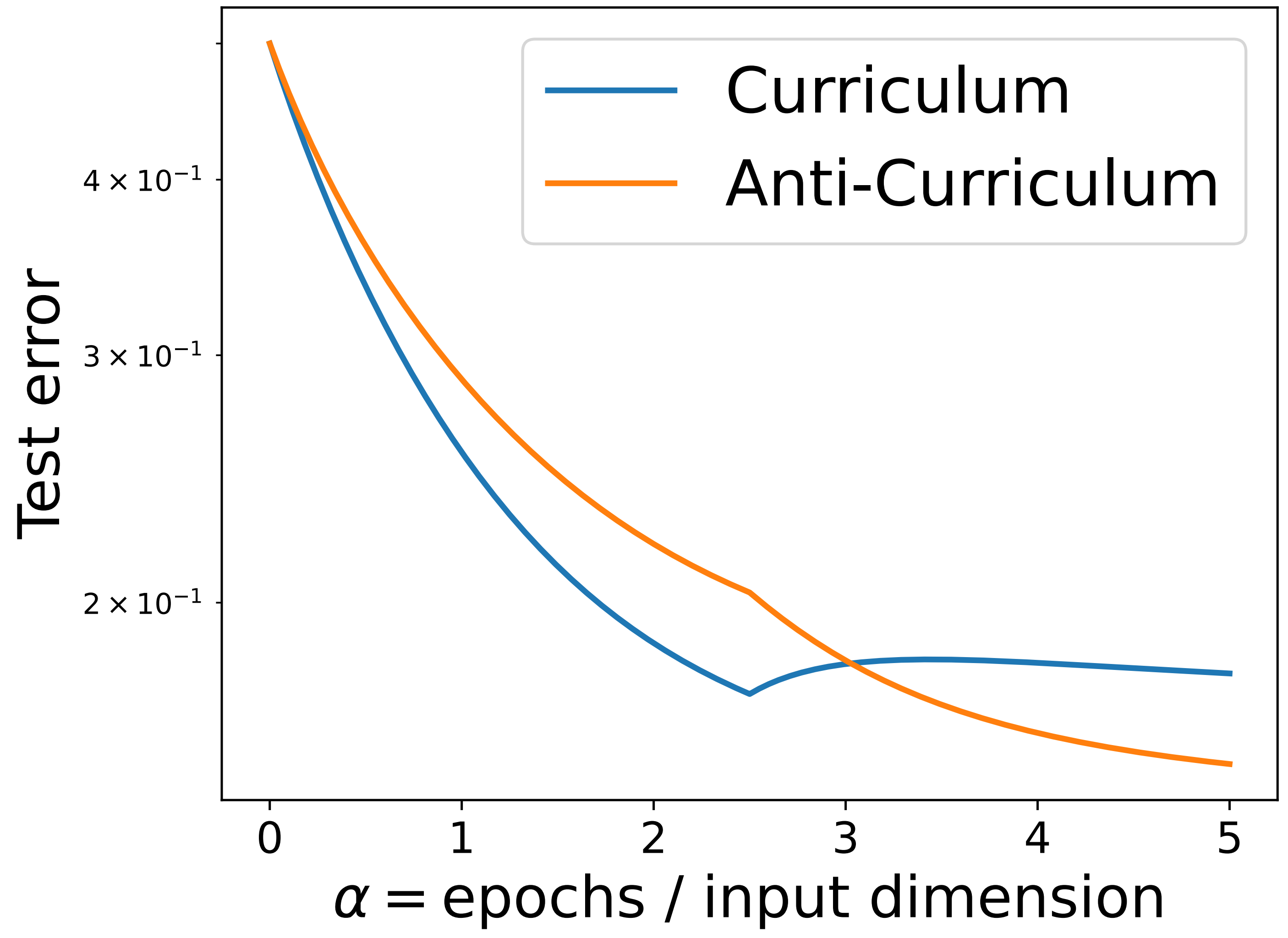
$\rho = 0.55, \eta = 2.58$

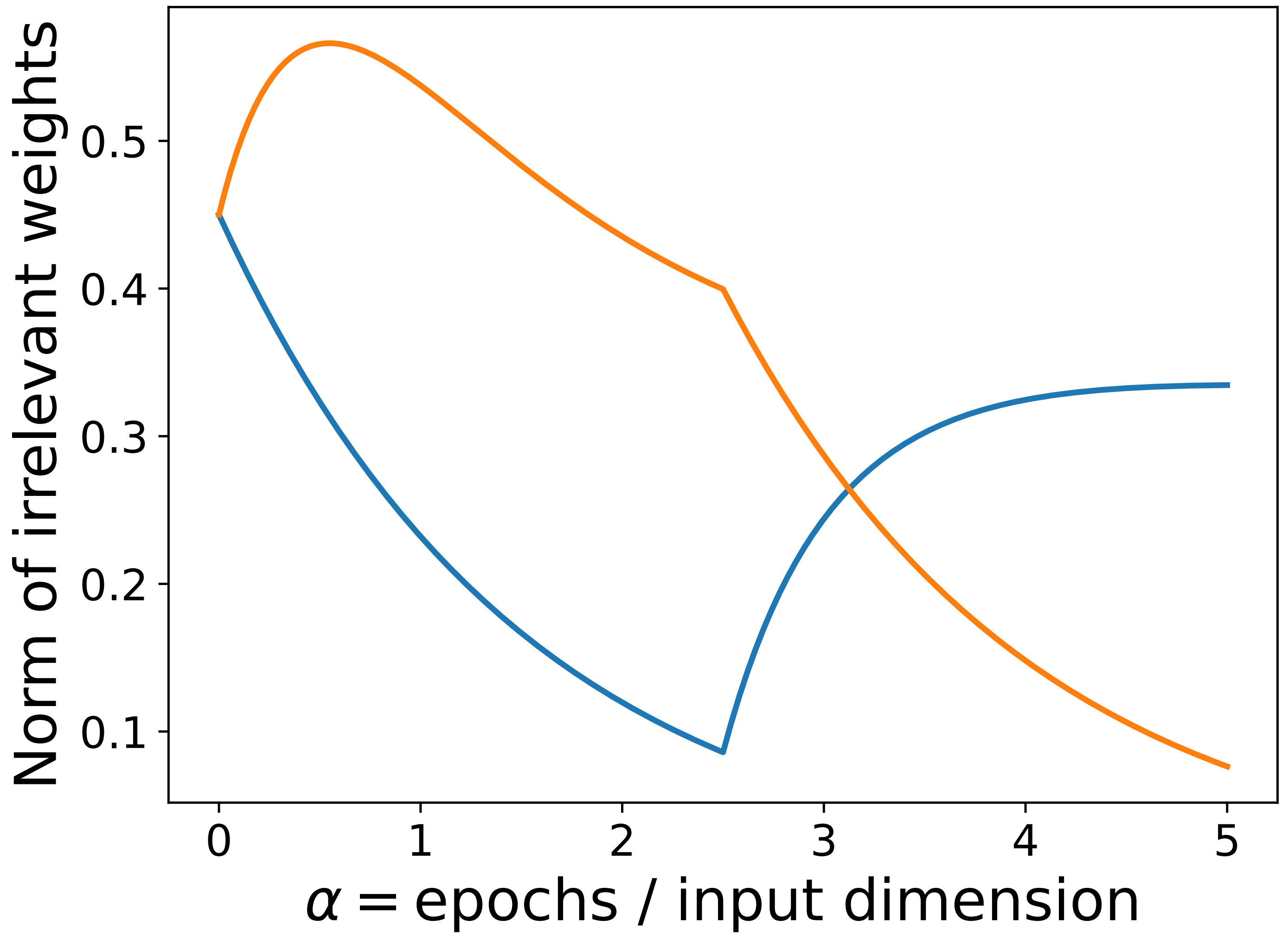


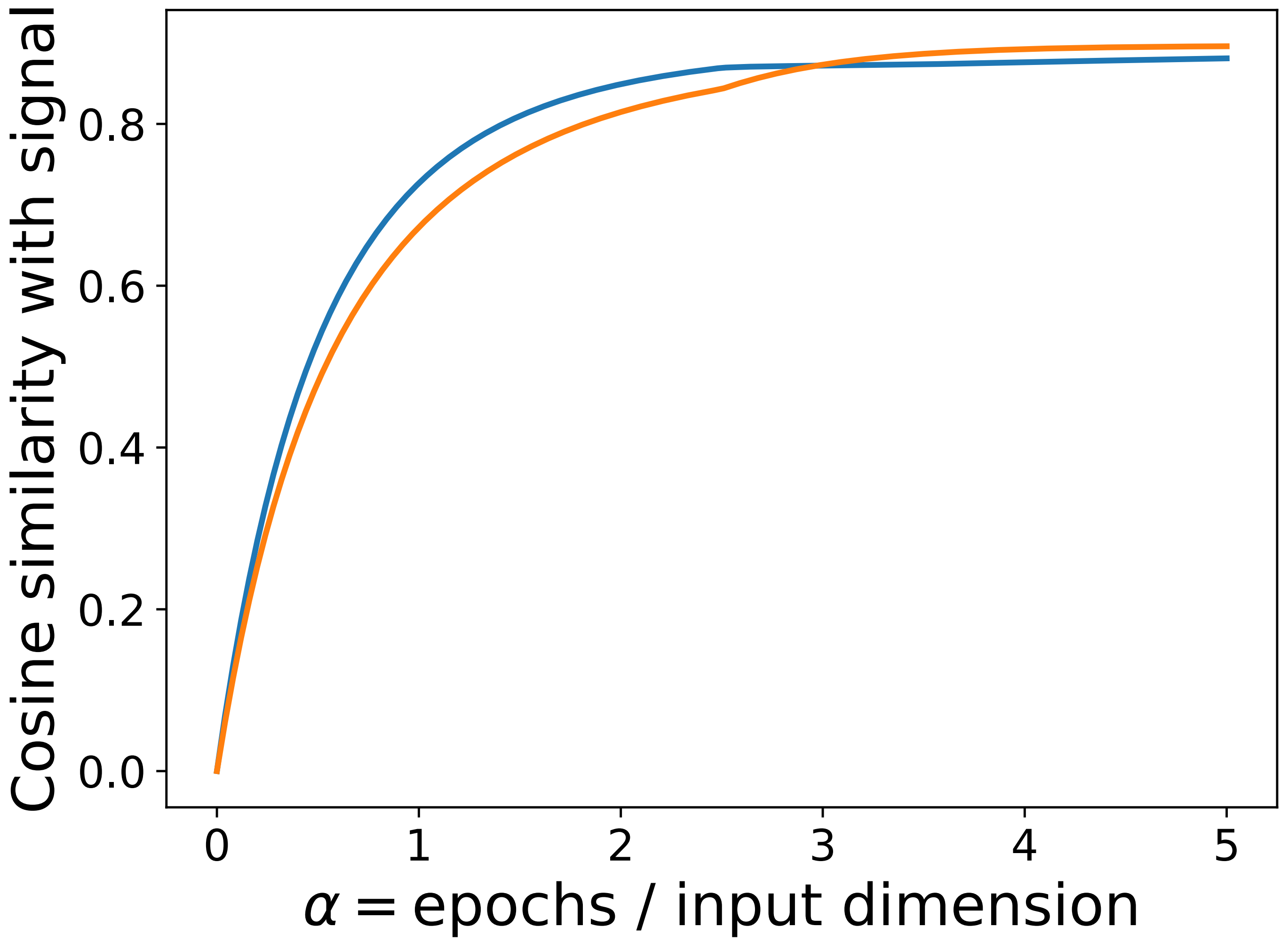
Anti-curriculum is optimal:

Hard	Easy

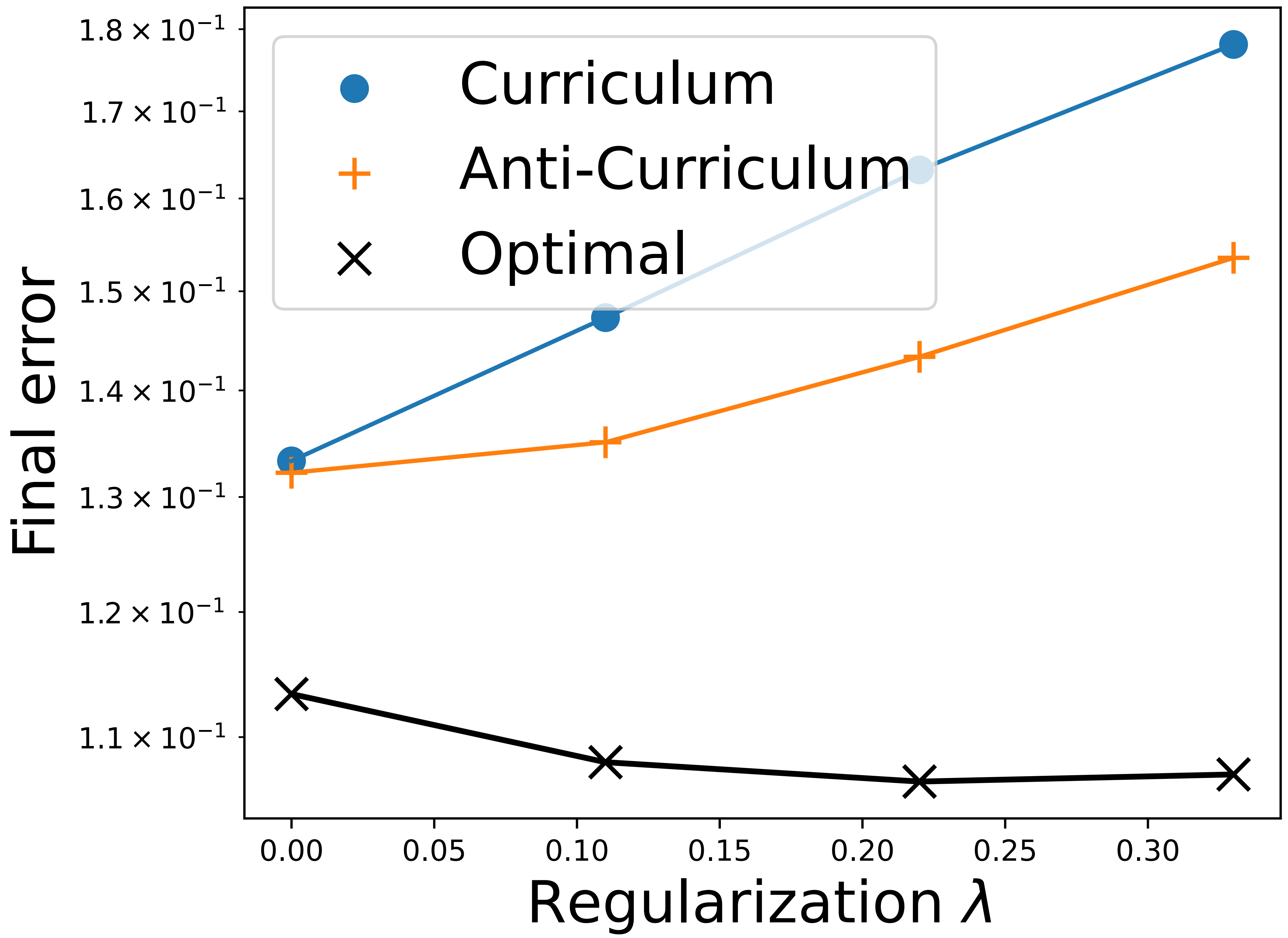
α = epochs / input dimension



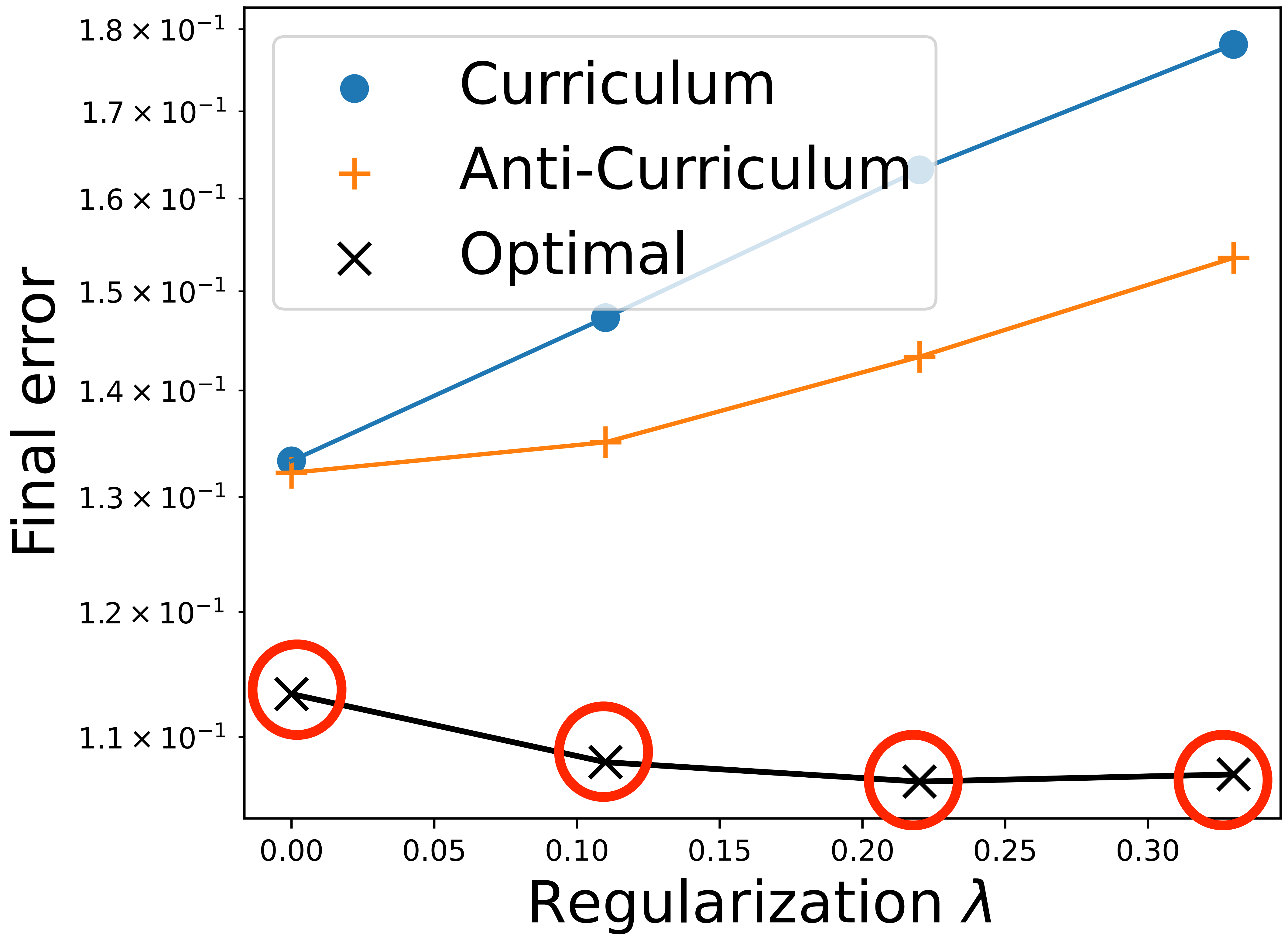




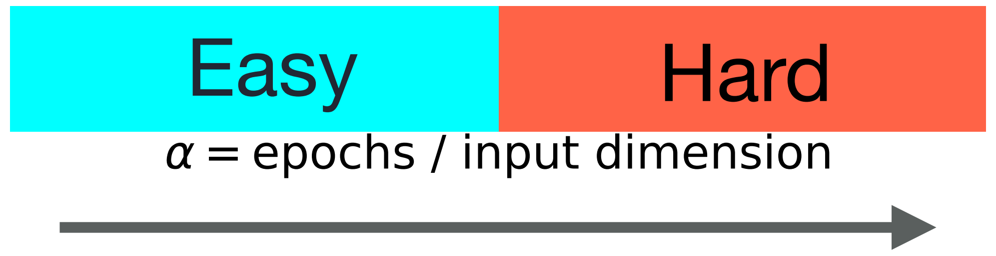
Optimal curriculum protocol



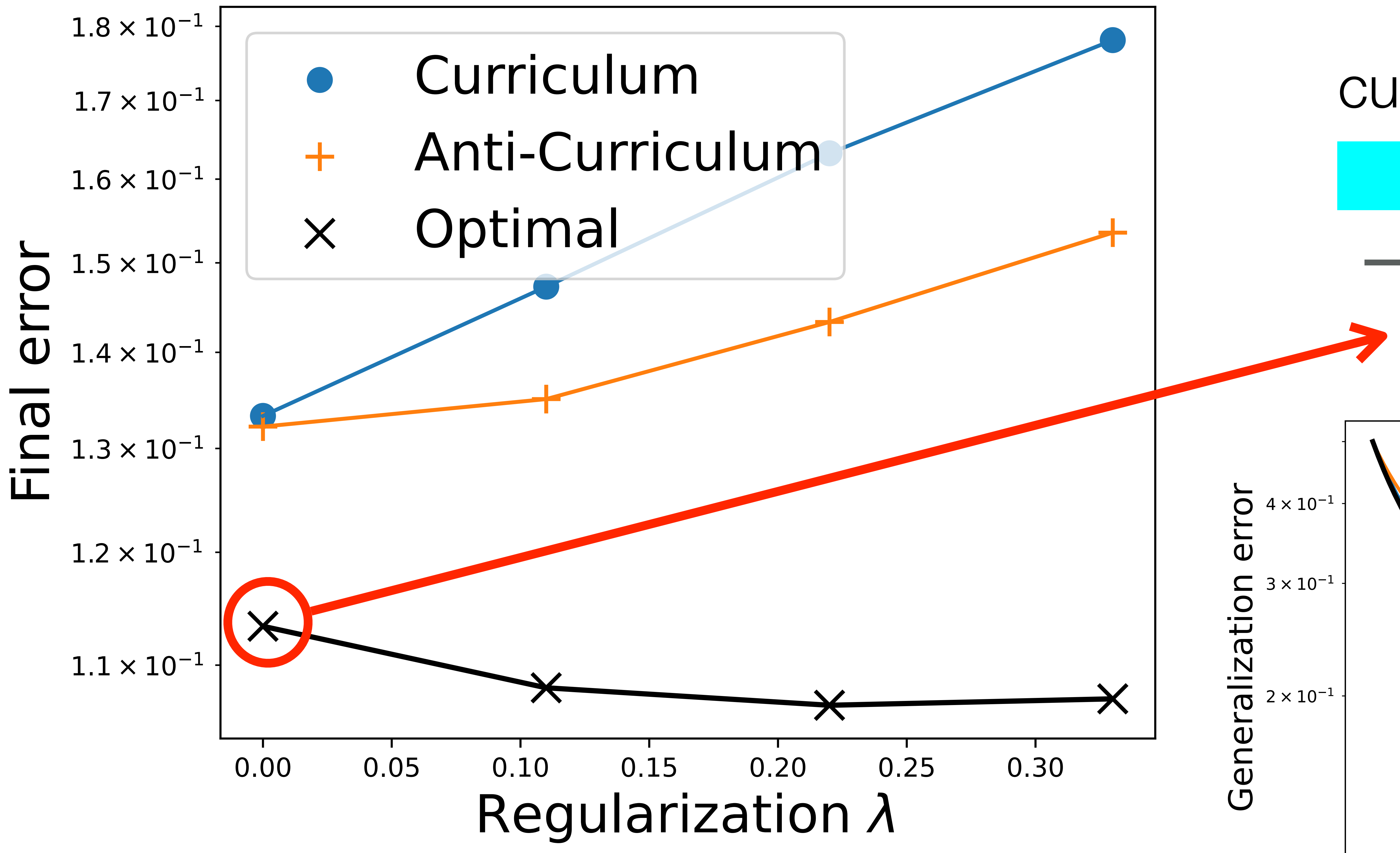
Optimal curriculum protocol



Easy-to-hard
curriculum is optimal:



Optimal curriculum protocol



Easy-to-hard
curriculum is optimal: