Optimal strategies in navigation and learning: statistical physics meets control theory

Francesco Mori

UNIVERSITY OF OXFORD

NEW COLLEGE OXFORD

LPTMC Seminar

21st Nov 2024

Overview of my research

Forward problems

Multiscale dynamics

Inverse problems

Control theory

Fundamental principles of **nonequilibrium** systems

Overview of my research

Forward problems

Multiscale dynamics

Inverse problems

Control theory

Fundamental principles of nonequilibrium systems

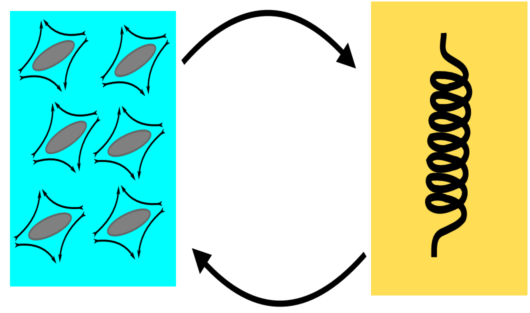
Active systems

Animal behavior

Learning systems

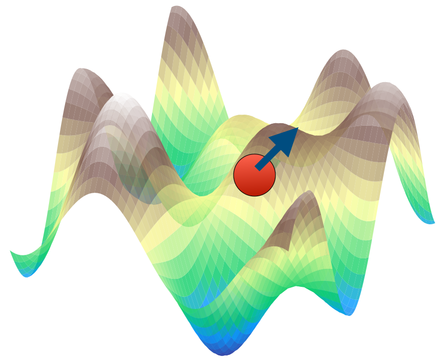
A

Mechanical control of active systems



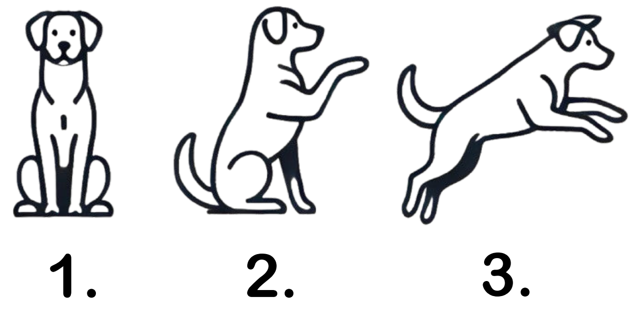
B

Navigation of active particles



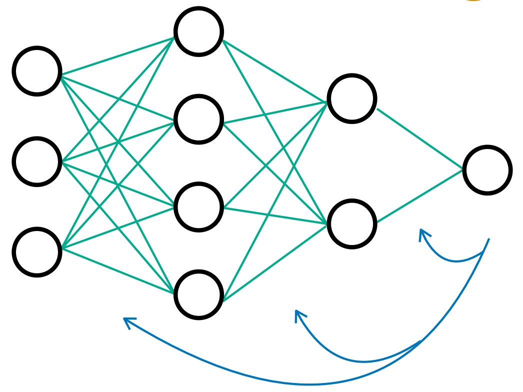
C

Optimal strategies in biological learning



D

Optimal protocols in machine learning



Active dynamics in learning processes

Part I:
Animal navigation

"Optimal switching strategies for navigation in stochastic settings."

FM, L. Mahadevan.

arXiv preprint *arXiv:2311.18813*

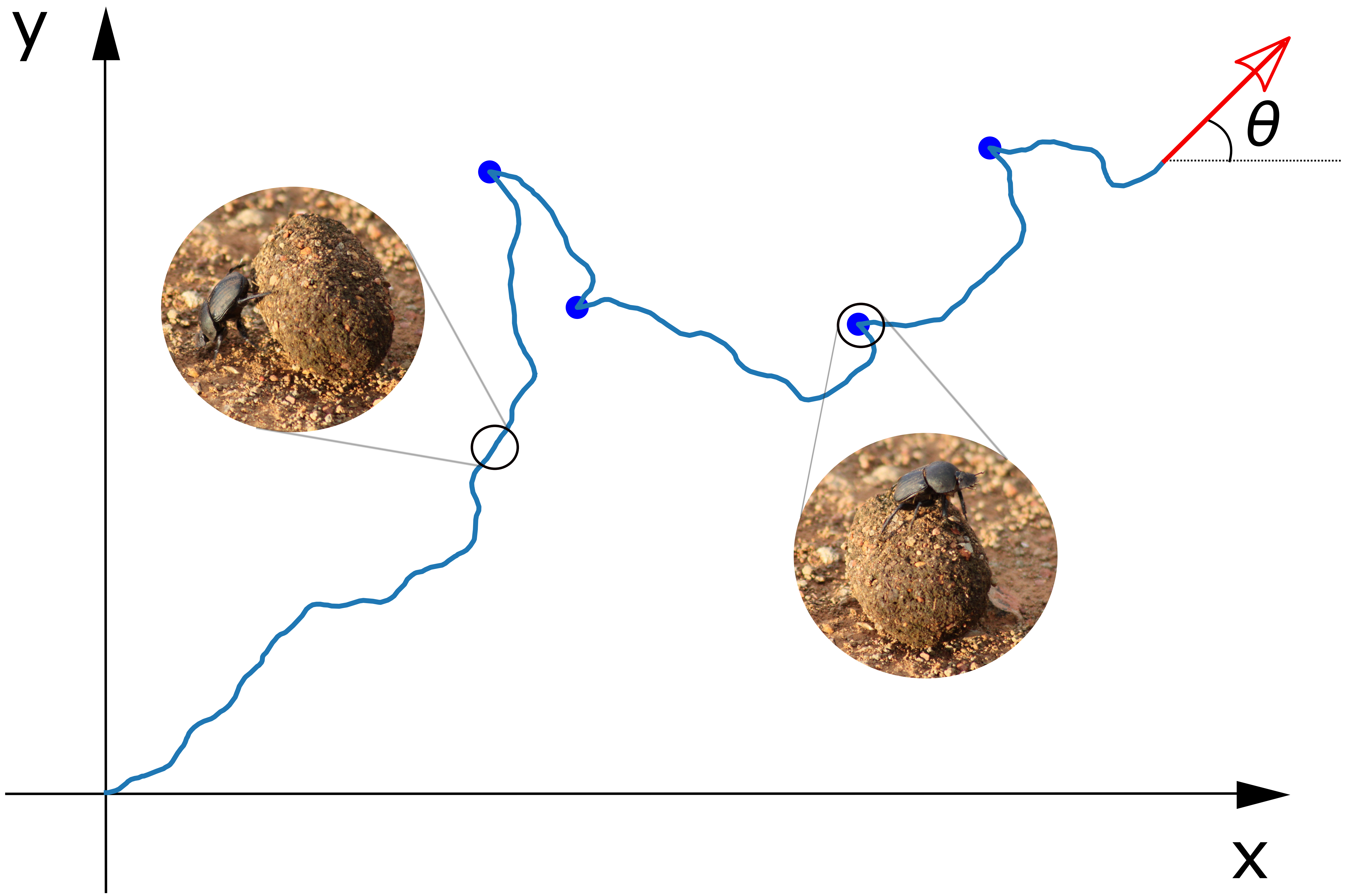
L. Mahadevan
(Harvard)

Navigation in noisy environments

Credit: MdeVicente CC0 1.0 Universal

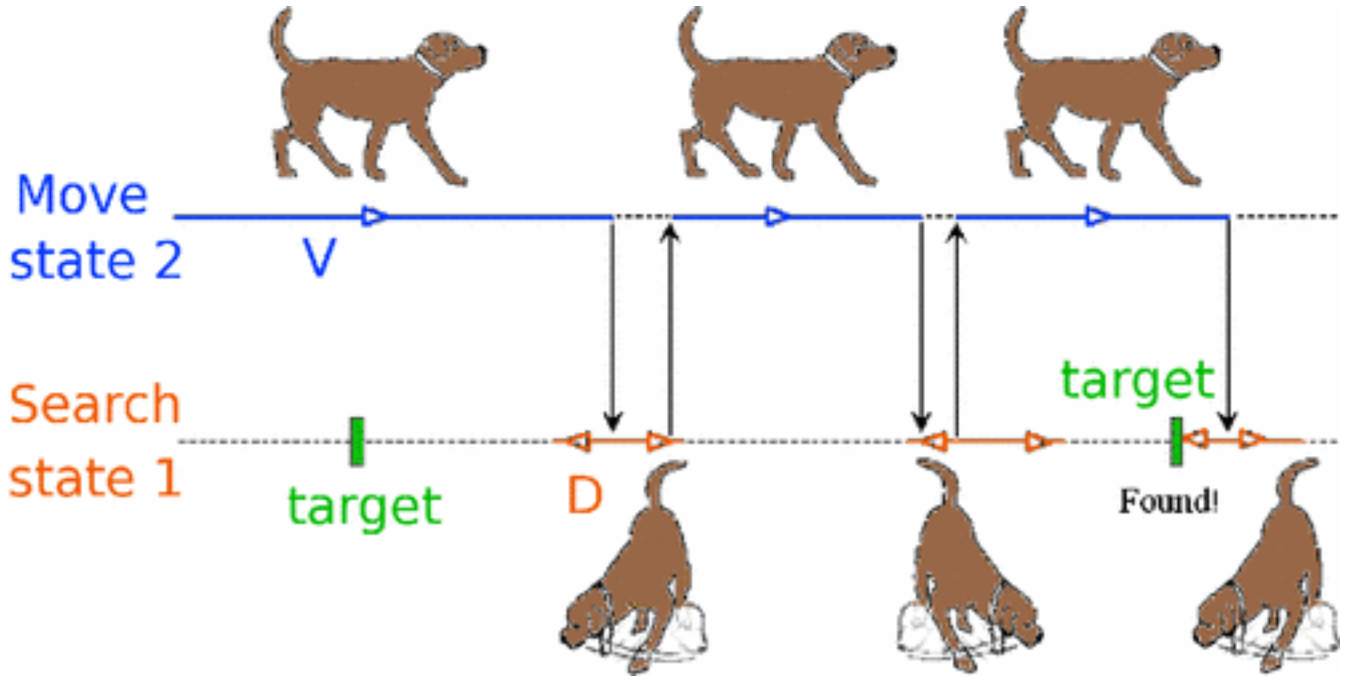
Navigation in noisy environments

Synthetic data



Switching behaviour in navigation

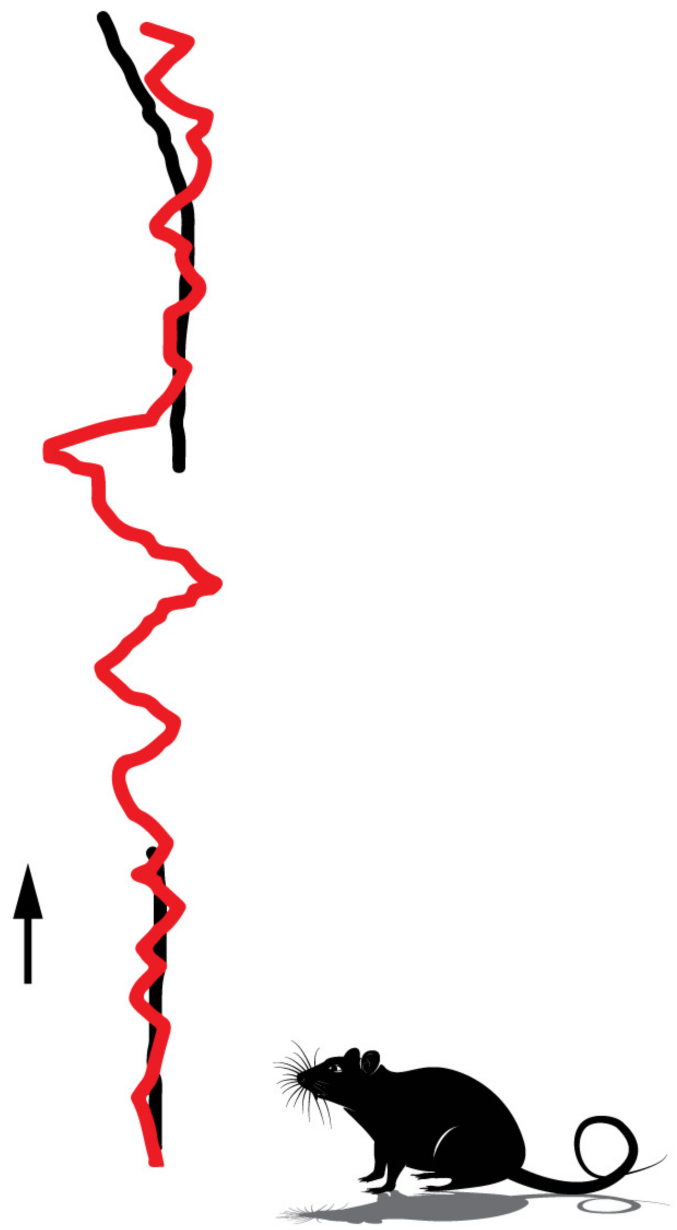
Intermittent search strategies



Thesen, Aud, Johan B. Steen, and Kjell B. Døving. *Journal of Experimental Biology* 180.1 (1993): 247-251.

O. Bénichou, C. Loverdo, M. Moreau, and R. Voituriez, Rev. Mod. Phys. **83**, 81

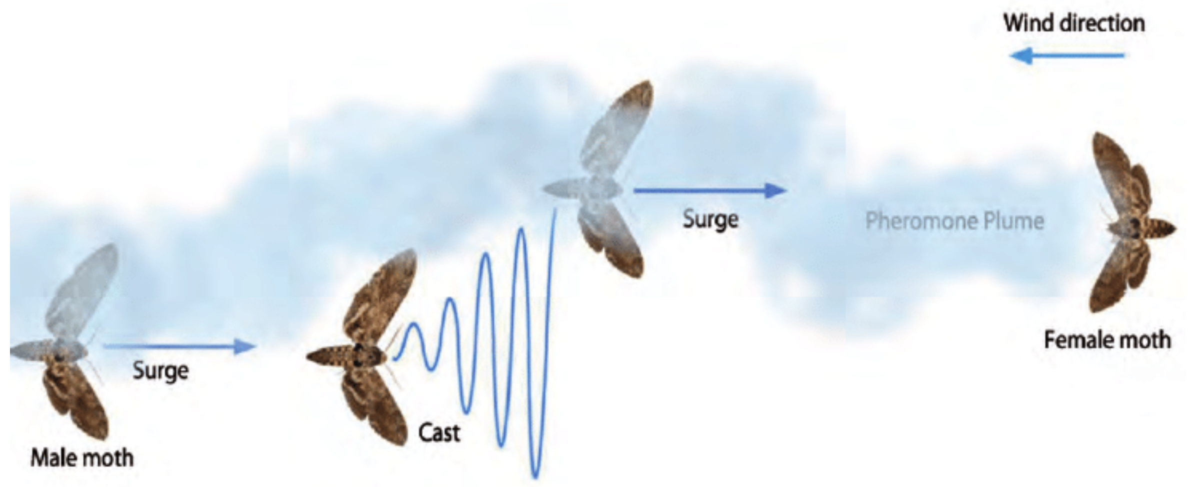
Trailing



A. G. Khan, M. Sarangi, and U. S. Bhalla, *Nature communications* ***3***, 703 (2012).

G. Reddy, B. I. Shraiman, and M. Vergassola, *PNAS 119*(1), e2107431118 (2022).

Zigzagging vs casting



E. Balkovsky, and B. I. Shraiman. *PNAS* 99, 12589 (2002).

L.L. López, et al. IntechOpen, 2011.

G. Reddy, V. N. Murthy, and M. Vergassola. *Ann. Rev. Cond. Matt. Phys.* 13 (2022): 191-213.

Egocentric vs geocentric strategies

Desert ant

Human with a map

Egocentric

How to optimally switch?

Geocentric

O. Peleg and L. Mahadevan. *Royal Society open science* 3.7, 160128 (2016).

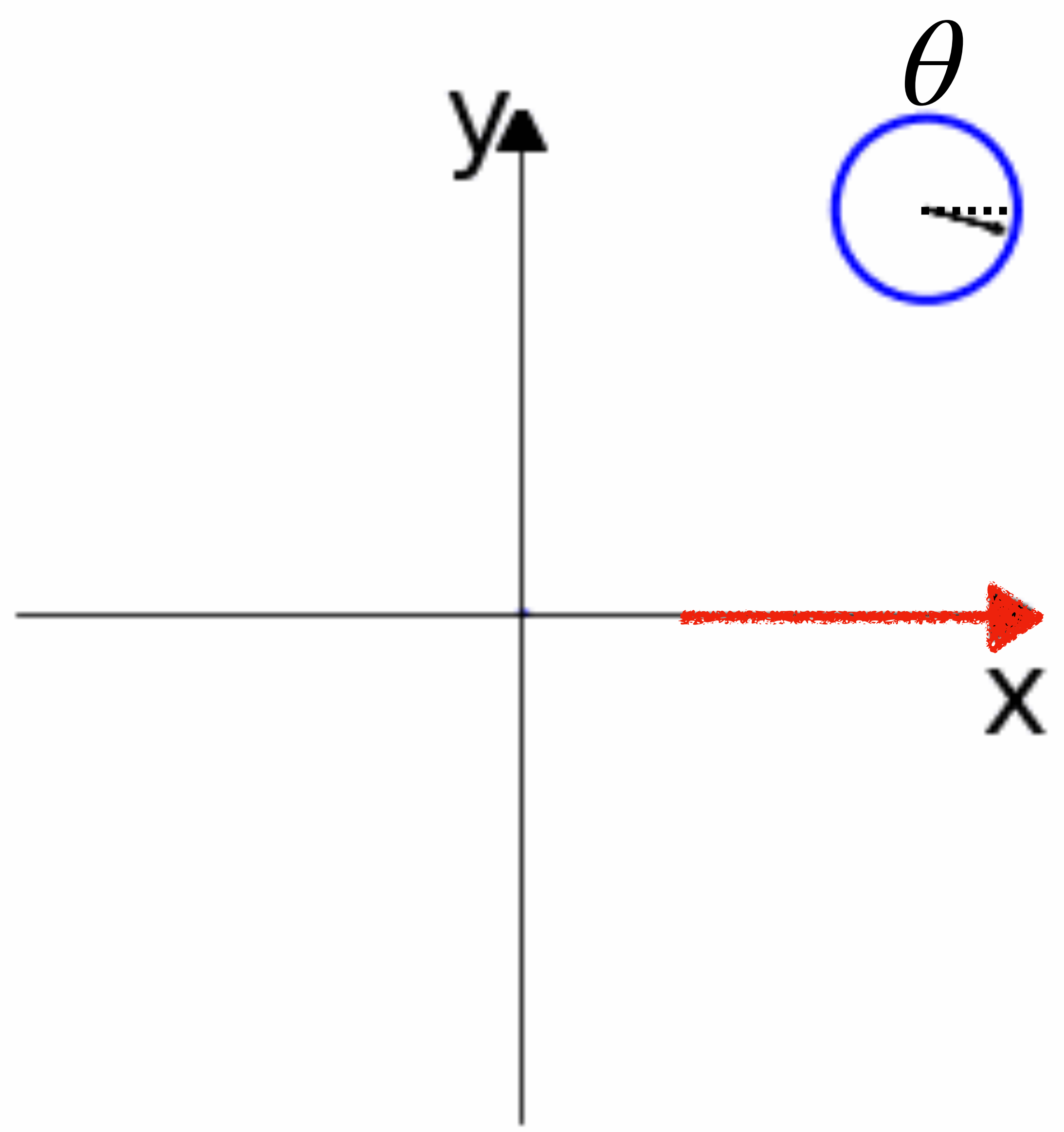
Active Brownian motion

$$\begin{cases} \dot{x}(t) = v_0 \cos[\theta(t)]\,, \\ \dot{y}(t) = v_0 \sin[\theta(t)]\,, \\ \dot{\theta}(t) = \eta(t) \end{cases}$$

$$\boxed{\begin{aligned} \langle \eta(t) \rangle &= 0 \\ \langle \eta(t)\eta(t') \rangle &= 2D\delta(t-t') \end{aligned}}$$

GOAL

Maximise the displacement in the x direction (keep θ close to zero)



Model

Active Brownian motion
with corrections

GOAL

Maximise the displacement in the x direction (keep θ close to zero)

$$\theta(t+dt) = \begin{cases} \theta(t) + \sqrt{2D}dt\eta(t) & \text{if } s(\theta, t) = 0\,, \\ 0 & \text{if } s(\theta, t) = 1\,. \end{cases}$$

Correction policy

Optimal control

Active Brownian motion
with corrections

Reward function

$$\mathcal{F}_{\theta_0,t_0}[s] = \left\langle \underbrace{v_0 \int_{t_0}^{t_f} \cos(\theta(\tau))d\tau}_{\text{Reward}} - \underbrace{x_c(N(t_f) - N(t_0))}_{\text{Cost}} \right\rangle_{\theta_0}$$

$t_f =$ Time horizon

$N(t) =$ number of reorientation events up to time t

$x_c =$ Cost per reorientation

Optimal control

Reward function

$$\mathscr{F}_{\theta_0,t_0}[s] = \langle v_0 \int_{t_0}^{t_f} d\tau \;\; \cos(\theta(\tau)) - x_c(N(t_f) - N(t_0)) \rangle$$

Optimal reward

$$\boxed{J(\theta, t) = \max_s \mathscr{F}_{\theta,t}[s]}$$

Maximisation over all possible strategies

Optimal strategy

$$\boxed{s^*(\theta, t) = \mathrm{argmax}_s \, \mathscr{F}_{\theta,t}[s]}$$

Previous works

UNIVERSITY OF OXFORD

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

CrossMark
click for updates

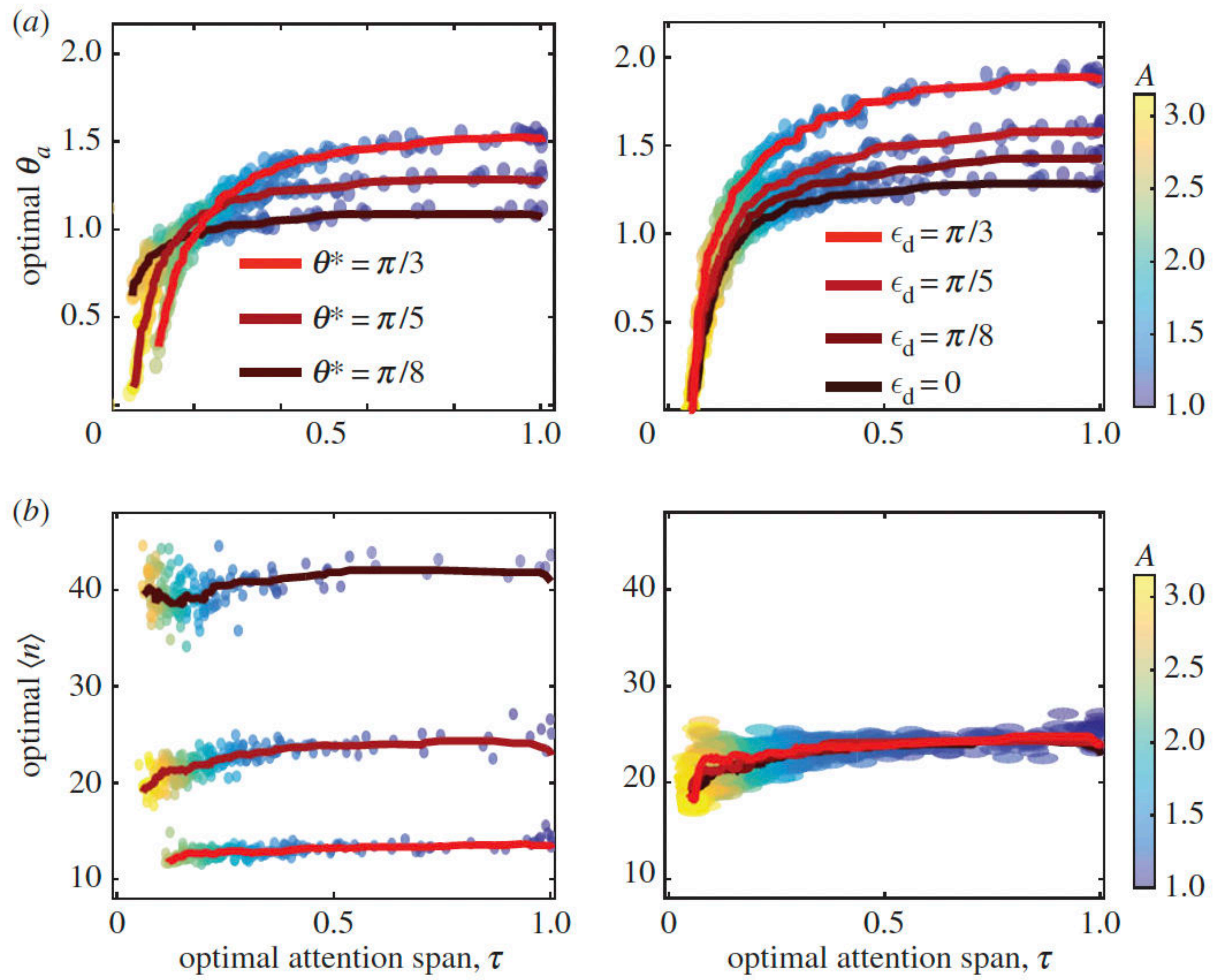
Cite this article: Peleg O, Mahadevan L. 2016 Optimal switching between geocentric and egocentric strategies in navigation. *R. Soc. open sci.* **3**: 160128.
http://dx.doi.org/10.1098/rsos.160128

Optimal switching between geocentric and egocentric strategies in navigation

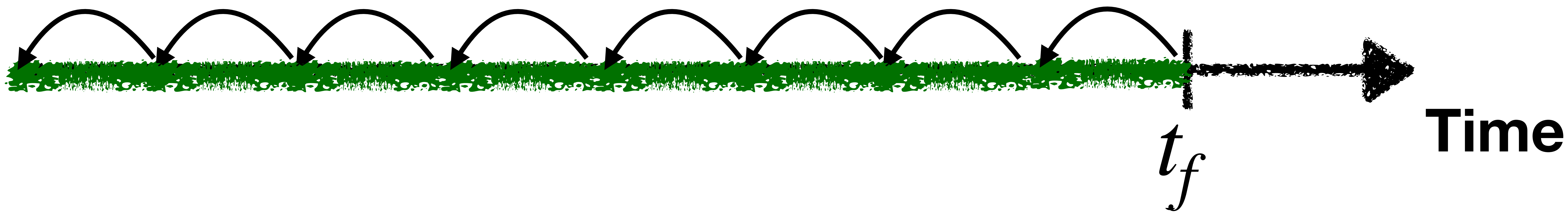
O. Peleg[1] and L. Mahadevan[1,2,3]

[1]Paulson School of Engineering and Applied Sciences, [2]Department of Physics, and [3]Department of Organismic and Evolutionary Biology, Kavli Institute for NanoBio Science and Technology, Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA 02138, USA

OP, 0000-0001-9481-7967



Dynamic programming



R. Bellman, R. W. Kalaba, et al., Dynamic programming and modern control theory (1965)

Bellman equation

B. De Bruyne (QRT)

Optimal reward

$$\boxed{J(\theta, t) = \max_s \mathscr{F}_{\theta,t}}$$

$$-\partial_t J(\theta, t) = D\partial_\theta^2 J(\theta, t) + v_0 \cos(\theta)\,, \quad \theta \in \Omega(t)\,,$$

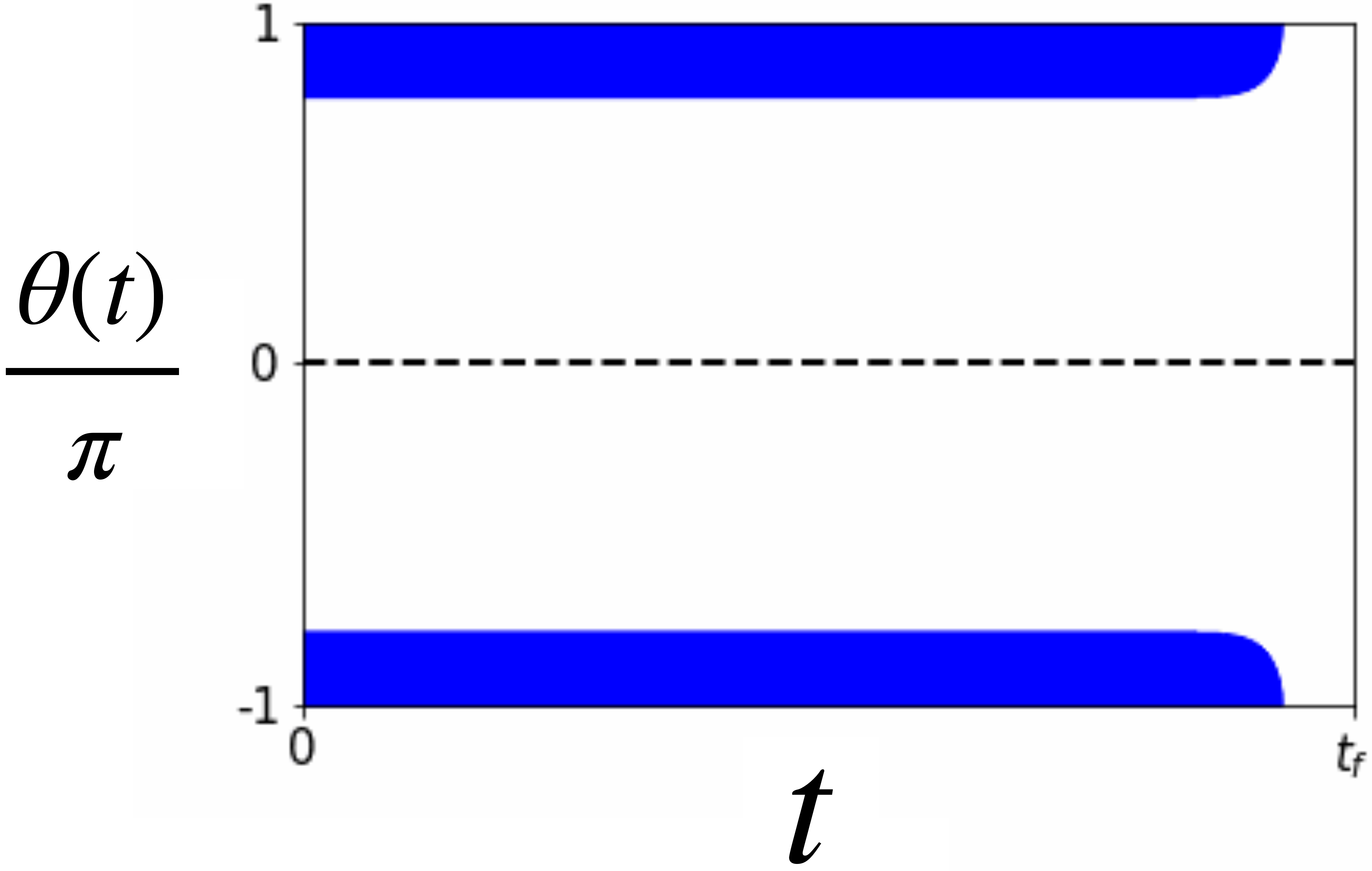
$$\Omega(t) = \{\theta : J(\theta, t) \geq J(0,t) - x_c\}$$

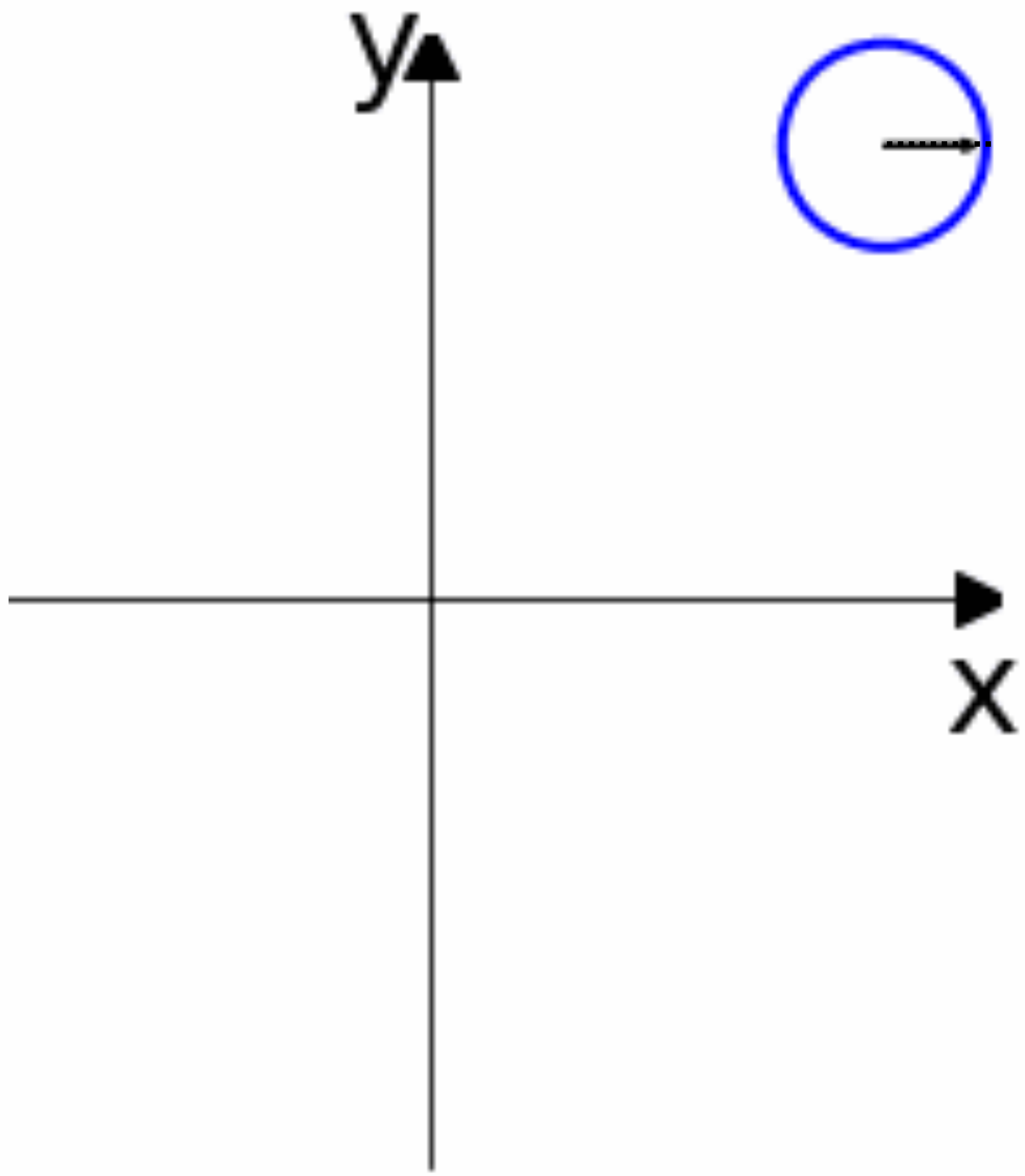
Optimal strategy

$$s^*(\theta, t) = \begin{cases} 1 & \text{if } \theta \notin \Omega(t)\,, \\ 0 & \text{if } \theta \in \Omega(t)\,, \end{cases}$$

B. De Bruyne and FM, *Physical Review Research* **5,** 013122 (2023).

Optimal reorientation policy





Infinite time horizon limit

$$-\partial_t J(\theta,t) = D\partial_\theta^2 J(\theta,t) + v_0\cos(\theta)\,, \quad \theta \in \Omega(t)\,,$$

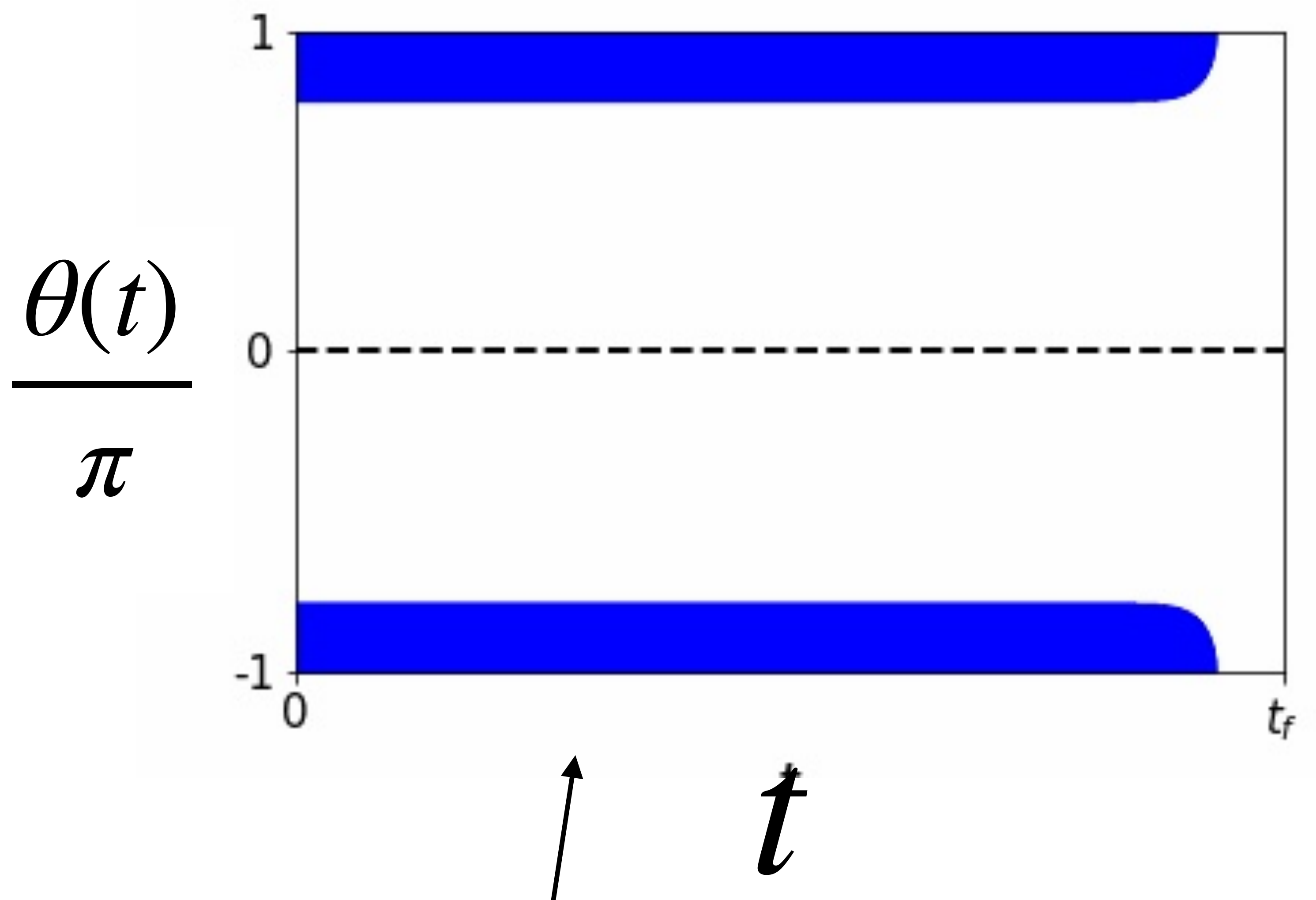
$$\Omega(t) = \{\theta : J(\theta,t) \geq J(0,t) - x_c\}$$

Ansatz

$$\boxed{\begin{array}{c} J(\theta,t) = j(\theta) + v^*(t_f - t) \\ \Omega(t) = [-\theta_a, \theta_a] \end{array}}$$

$$\frac{1}{2}\sin(\theta_a)\theta_a + \cos(\theta_a) = 1 - \frac{Dx_c}{v_0}\,.$$

$$v^* = v_0\frac{\sin(\theta_a)}{\theta_a}$$

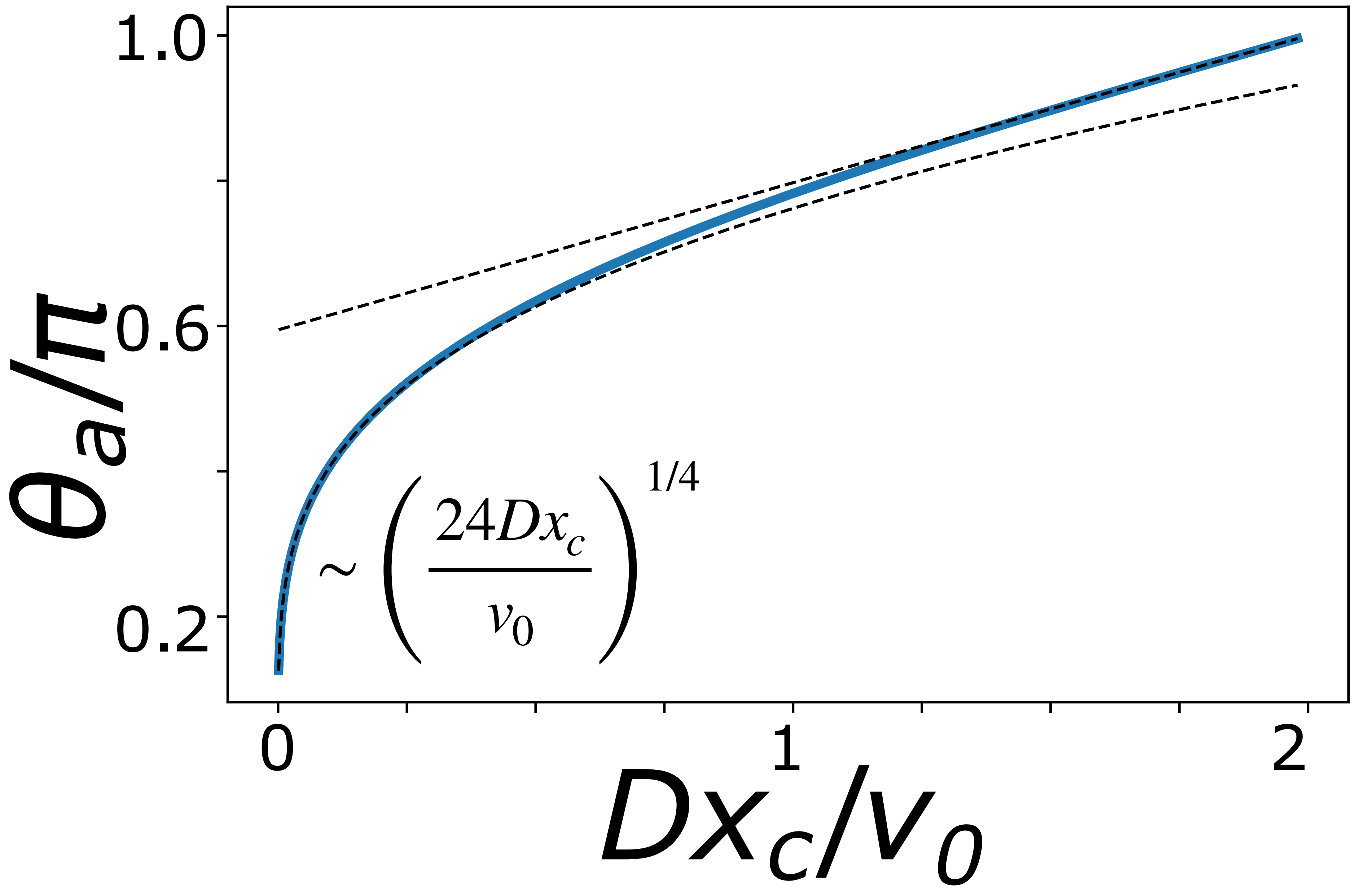


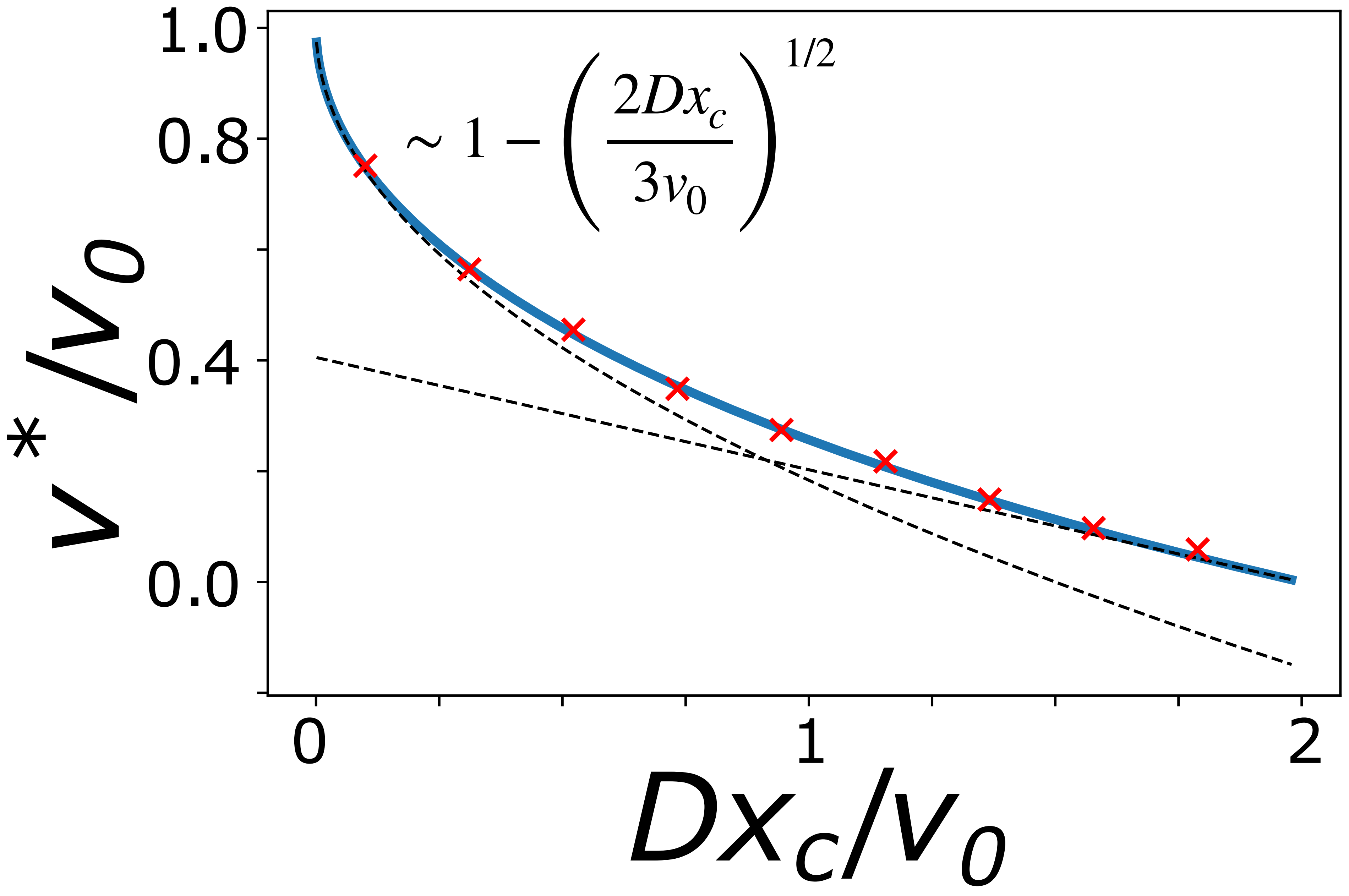
$$\boxed{(t_f - t) \gg D^{-1}}$$

Infinite time horizon limit

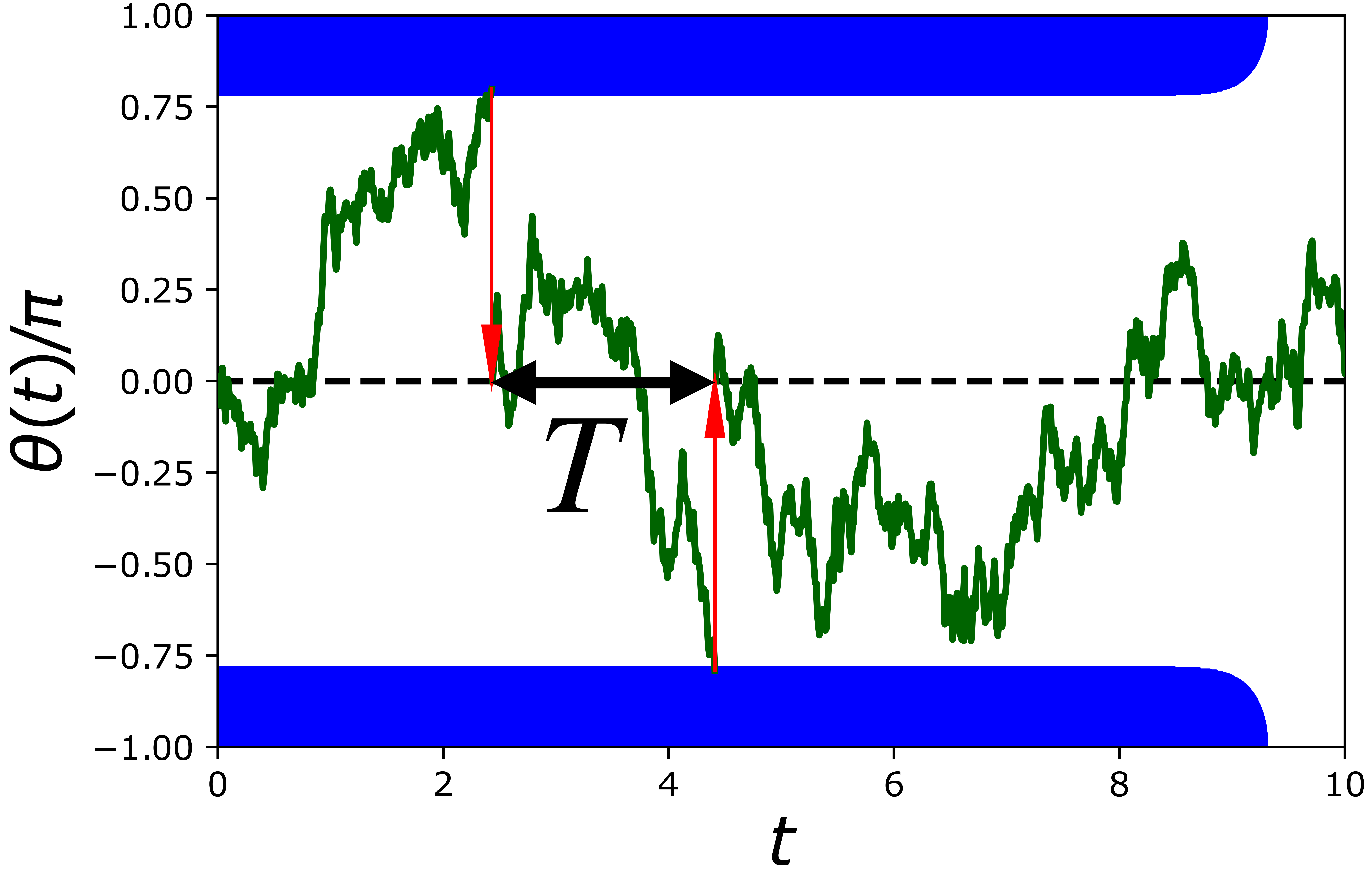
$$\frac{1}{2}\sin(\theta_a)\theta_a + \cos(\theta_a) = 1 - \frac{Dx_c}{v_0}.$$

$$v^* = v_0\frac{\sin(\theta_a)}{\theta_a}$$

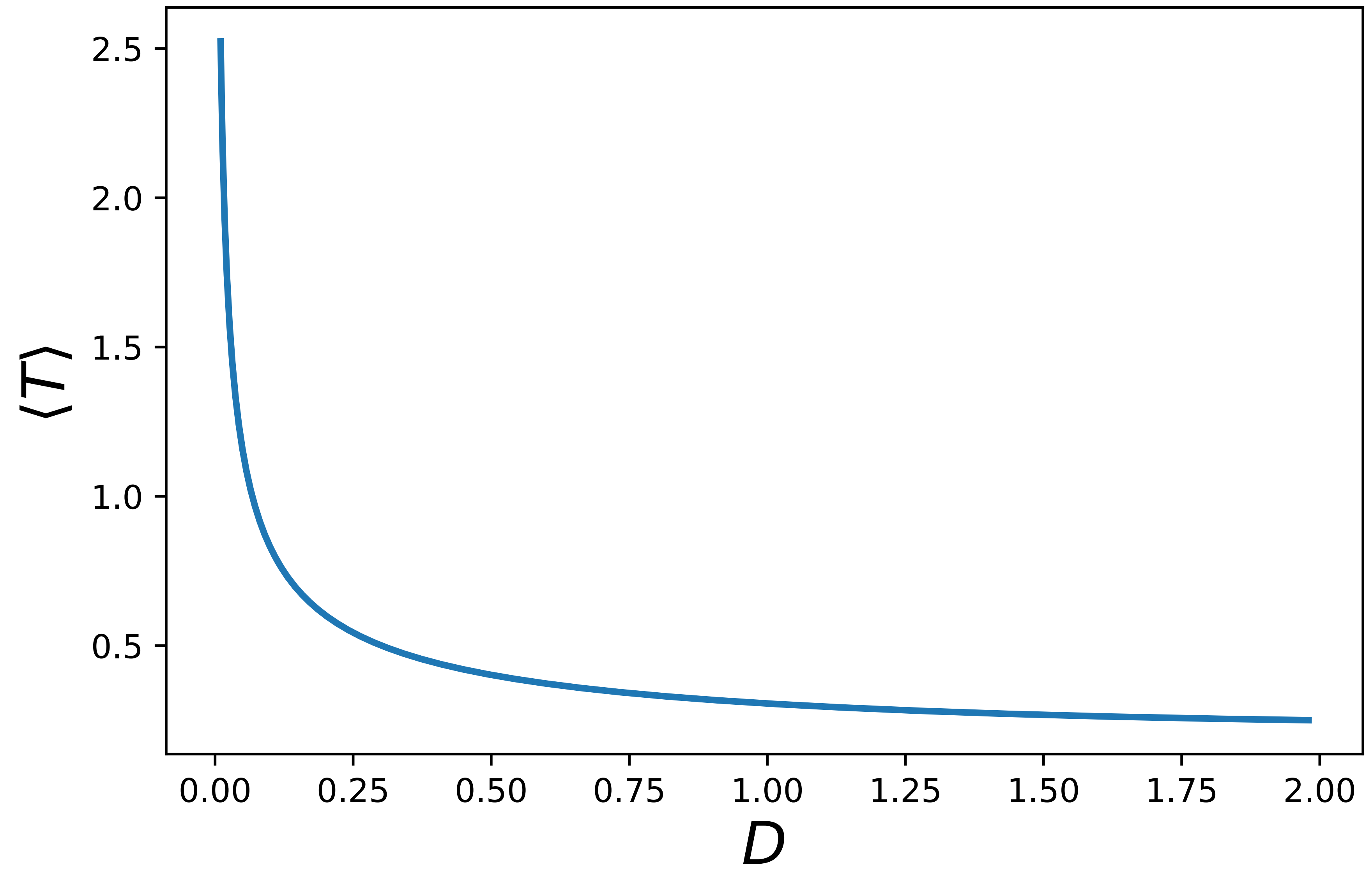




Time between reorientations



$$\langle T \rangle = \frac{\theta_a^2}{2D} \sim \frac{1}{\sqrt{D}}$$



Measurement noise

$$\begin{cases} \dot{x}(t) = v_0 \cos[\theta(t)]\,, \\ \dot{y}(t) = v_0 \sin[\theta(t)]\,, \\ \theta_1(t) = \theta(t) + \theta_n(t) \end{cases}$$

θ = real angle

$$\dot{\theta} = \eta(t)$$

$$\langle \eta(t)\eta(t') \rangle = 2D\delta(t - t')$$

θ_n = measurement noise

$$\dot{\theta}_n = \eta_n(t)$$

$$\langle \eta_n(t)\eta_n(t') \rangle = 2D_n\delta(t - t')$$

Strategy: reorient if $|\theta_1| > \theta_a$

Measurement noise

$$\begin{cases} \dot{x}(t) = v_0 \cos[\theta(t)] \,, \\ \dot{y}(t) = v_0 \sin[\theta(t)] \,, \\ \theta_1(t) = \theta(t) + \theta_n(t) \end{cases}$$

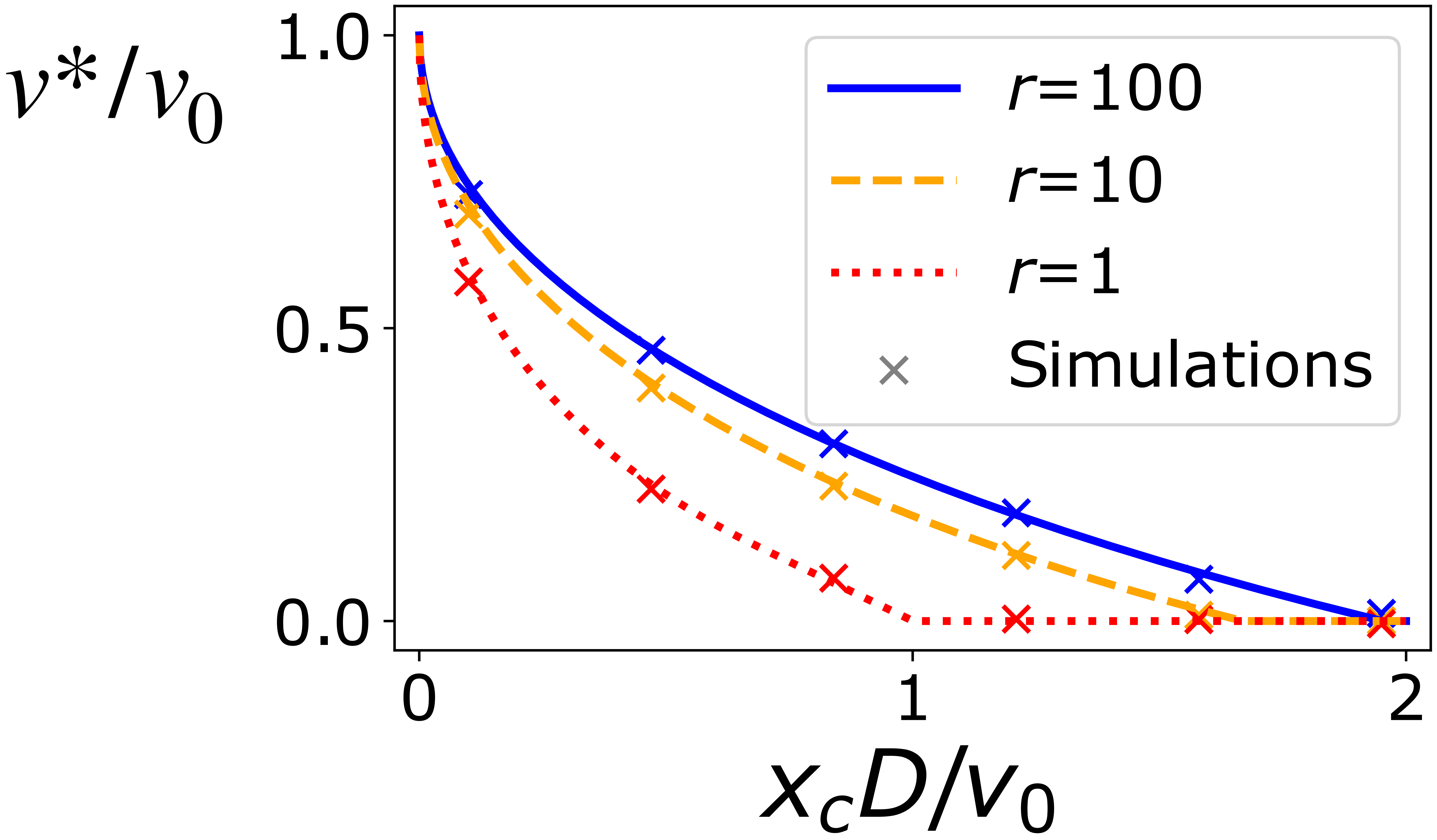
$$\boxed{\textbf{Strategy:}\ \text{reorient if}\ |\theta_1| > \theta_a}$$

$$v = \frac{2v_0}{\theta_a^2}(1 + 1/r)\left[1 - z - \frac{\cos\left(\frac{\theta_a}{1+1/r}\right)}{\cosh\left(\frac{\theta_a\sqrt{1/r}}{1+1/r}\right)}\right]$$

$$\boxed{r = D/D_n} \qquad z = \frac{Dx_c}{v_0}$$

Measurement noise

$$r = D/D_n$$



Conclusions and perspectives

- Minimal model of animal navigation inspired by the dung beetle

- Applications other systems: olfactory navigation, search processes,…

- Experimental verification (**Lund Vision Group**)

DALL.E

Part II:
Multi-task learning

"Optimal protocols for continual learning via statistical physics and control theory"

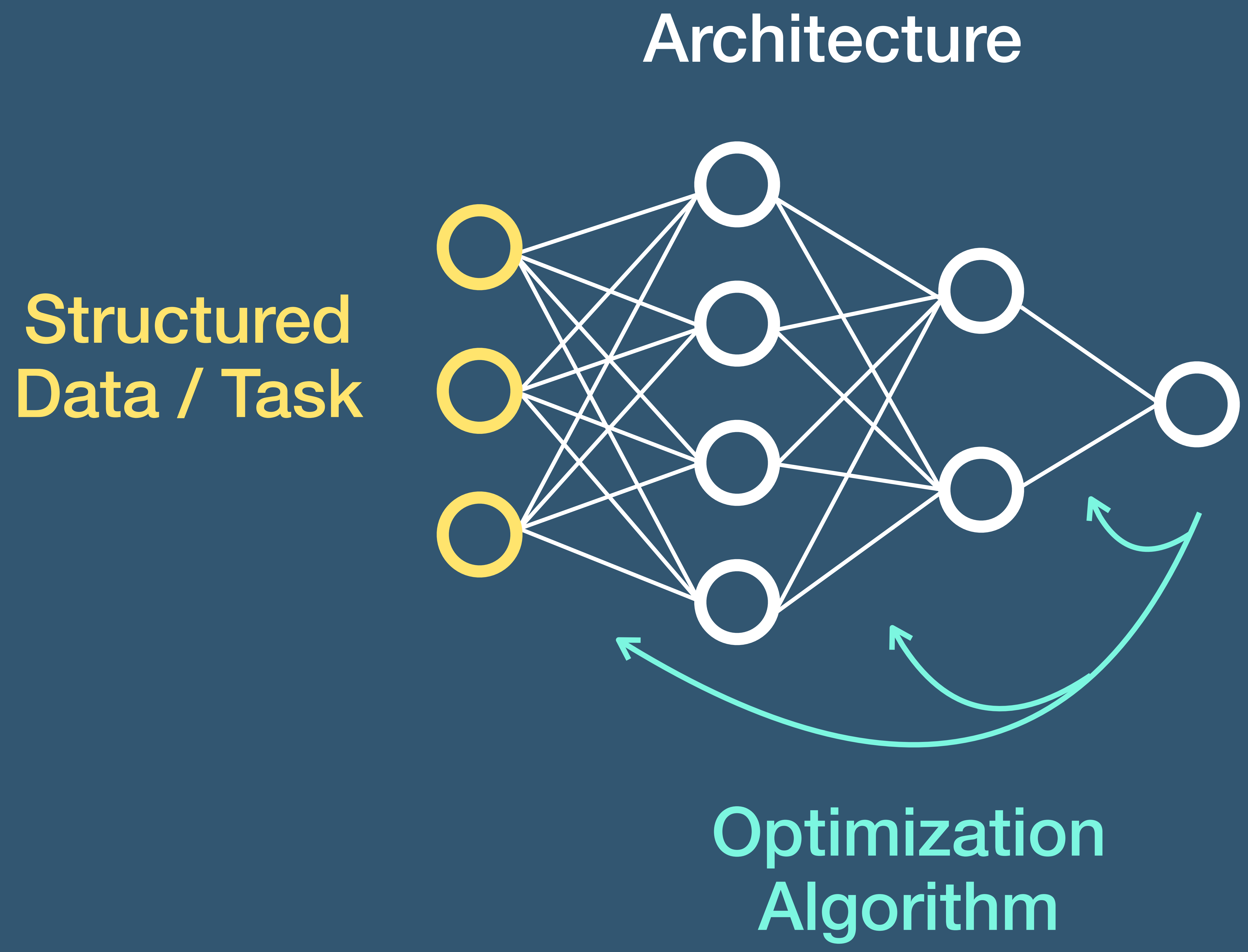
FM, Stefano Sarao Mannelli, Francesca Mignacco

arXiv preprint *arXiv:2409.18061*

Francesca Mignacco
Princeton and CUNY

Stefano Sarao Mannelli
Chalmers University

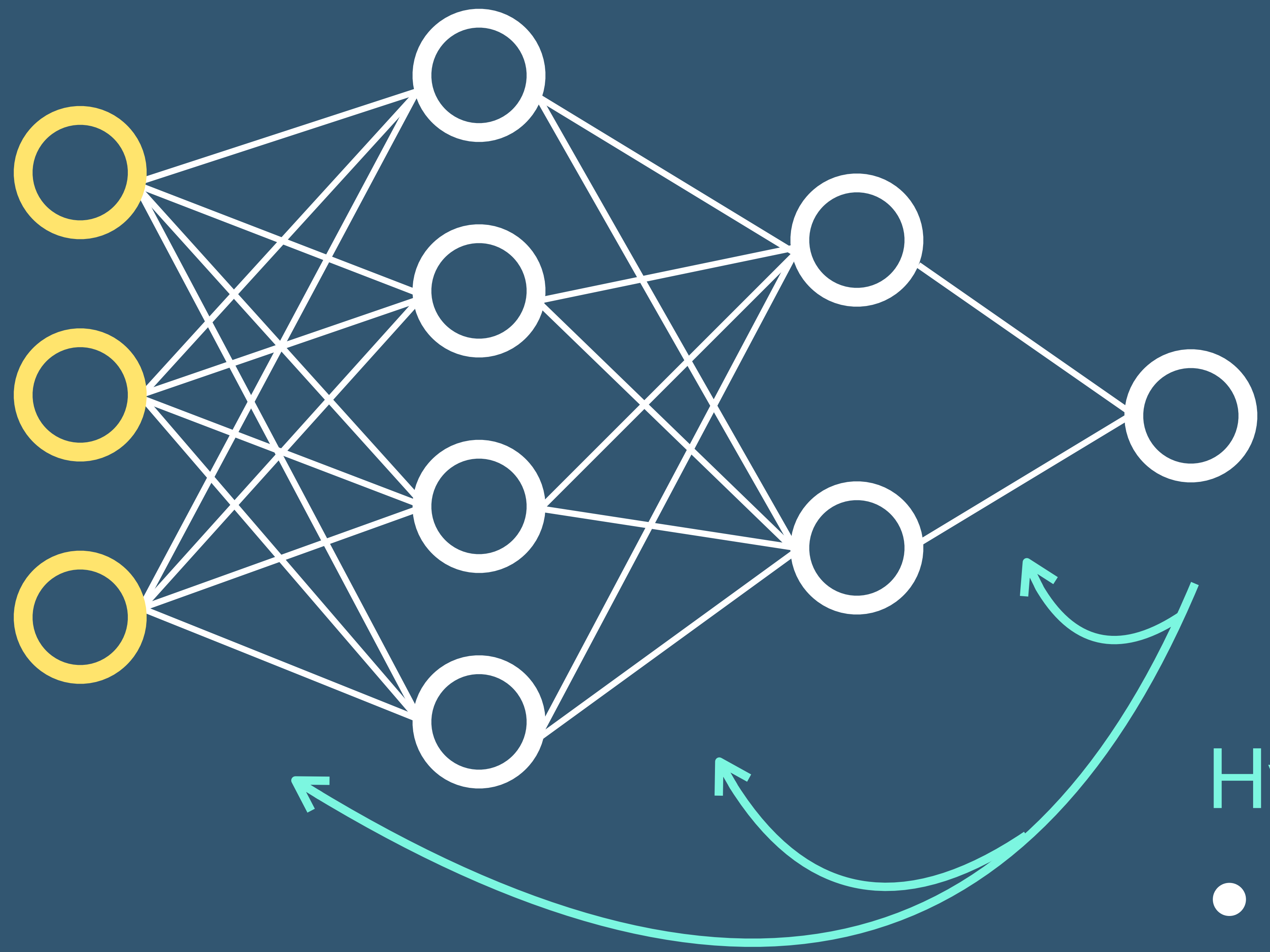
Learning protocols



Learning protocols

Architecture

Structured
Data / Task



Optimization
Algorithm

Hyper-parameter schedules:

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

Learning protocols

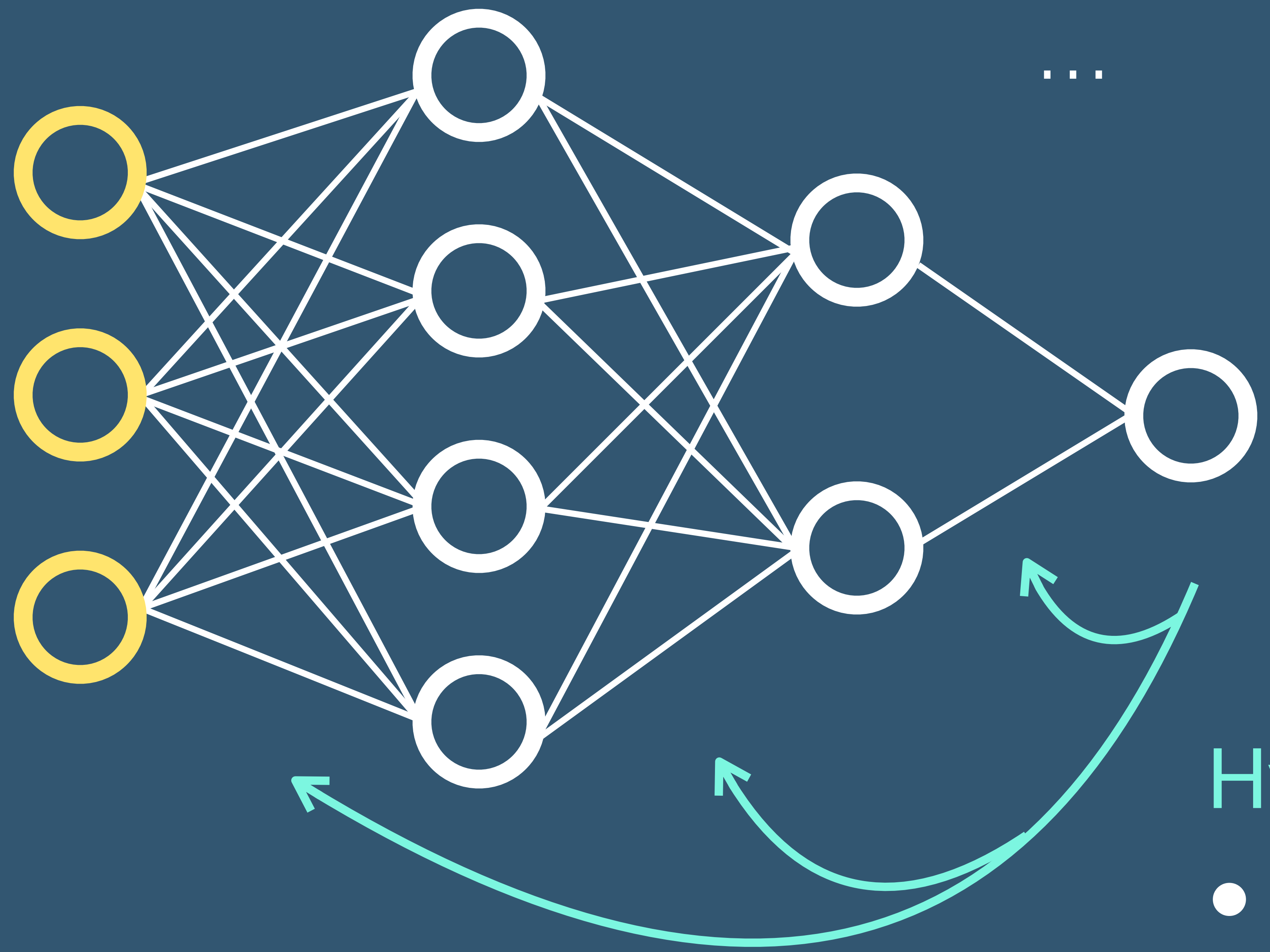
Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

Architecture

Structured Data / Task



Hyper-parameter schedules:

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

Optimization Algorithm

Learning protocols

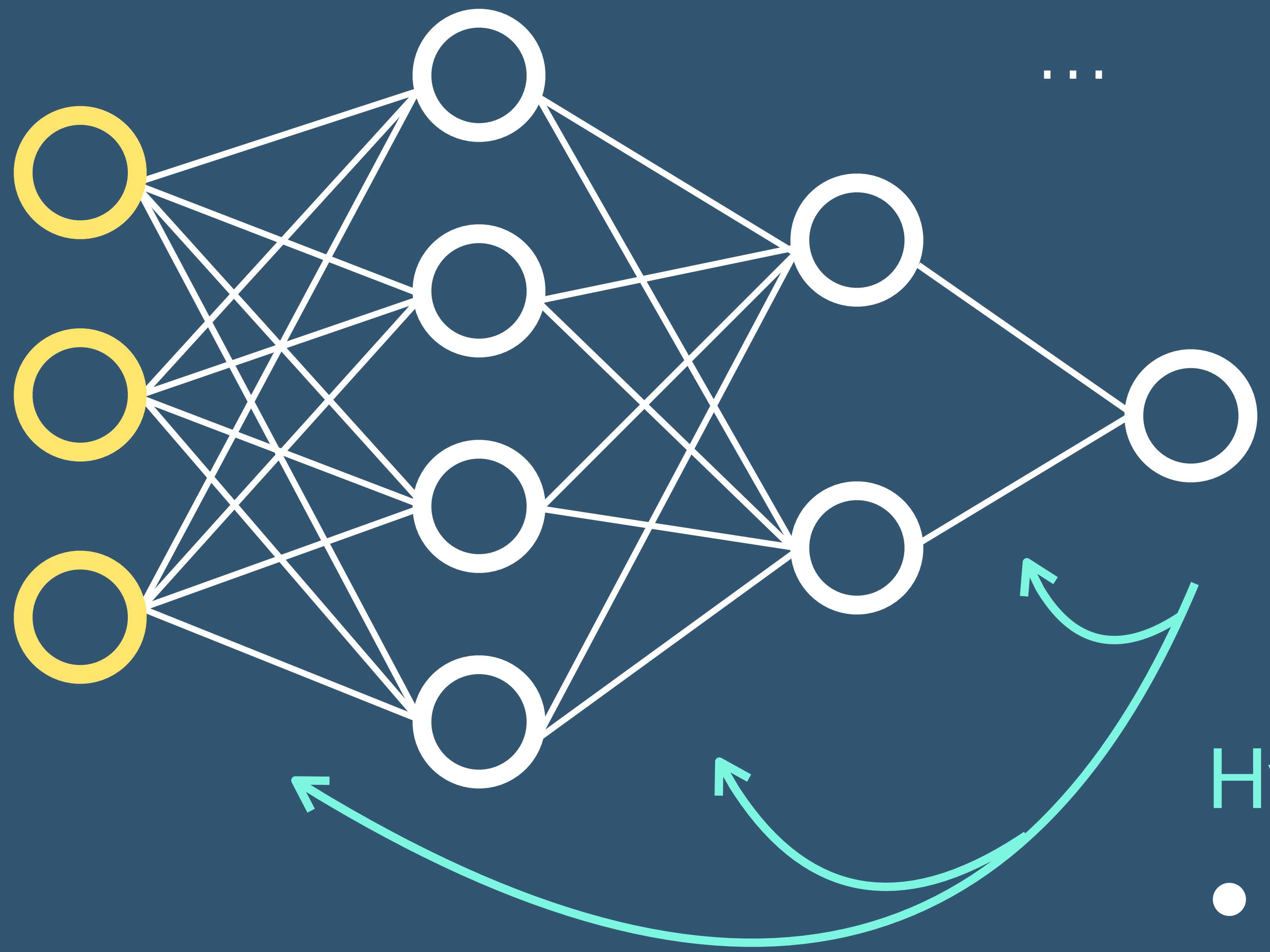
Model optimization:

- 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

From smoother landscape to the target

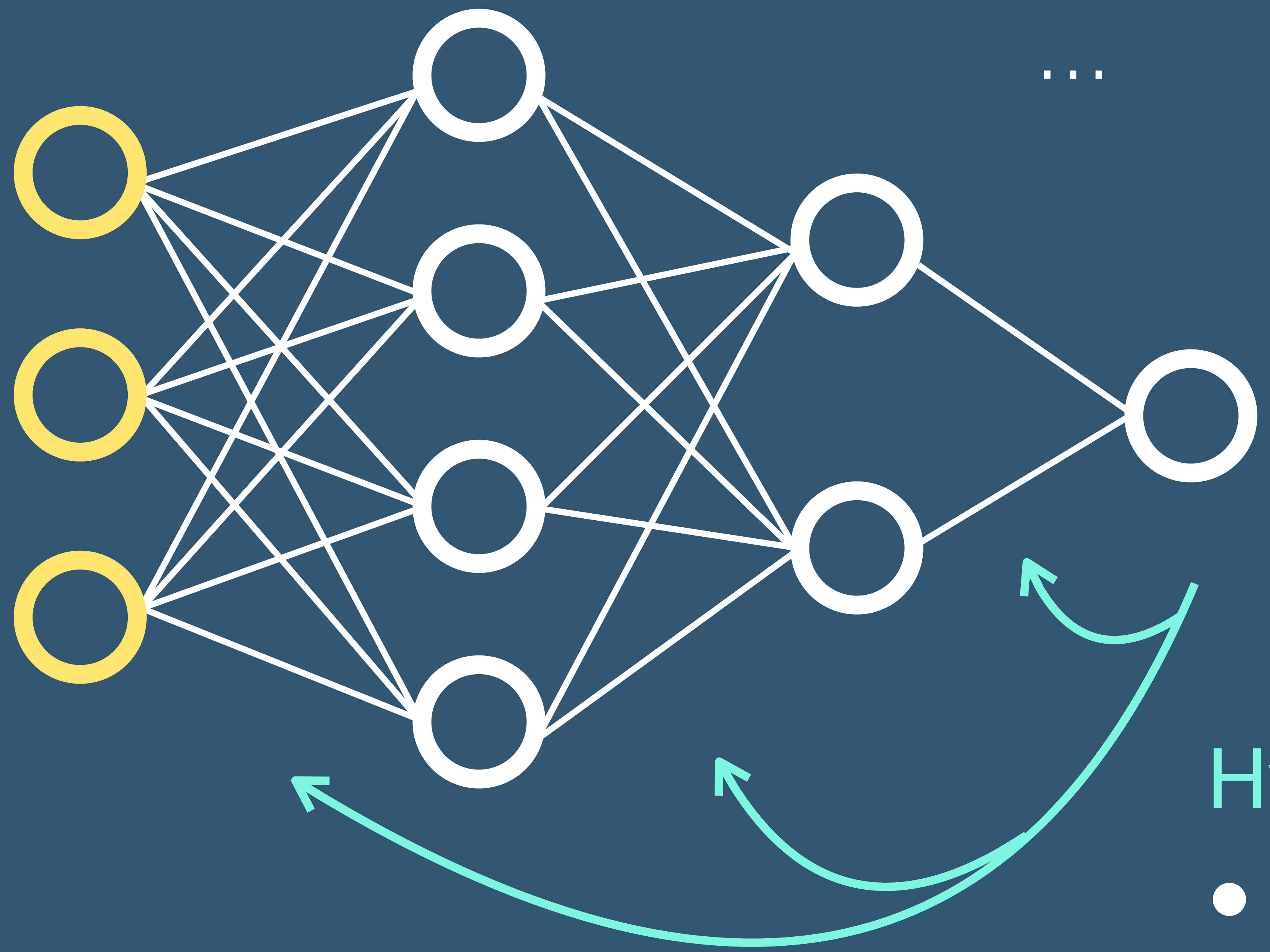
Architecture

Model optimization:

- Pruning
- Knowledge distillation
- Dropout

…

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
- …

Learning protocols

Goals:

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

From smoother landscape to the target

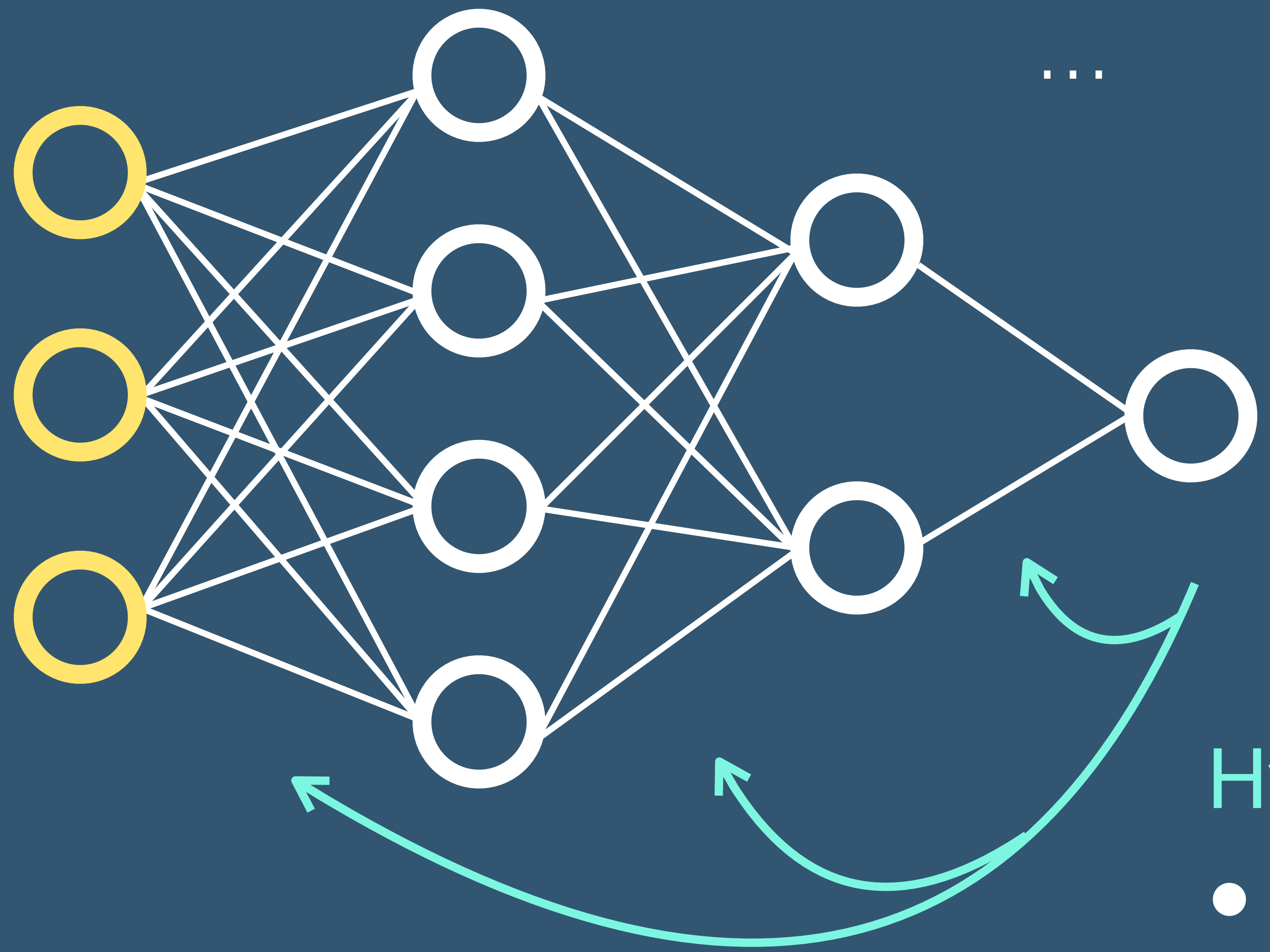
Architecture

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
- …

Model-based approaches to data / task selection

(Non-exhaustive list)

- Curriculum learning [Weinshall et al (2020); Saglietti et al. (2022); Abbe et al. (2023); Cornacchia et al (2023); Lee et al. (2024); Mannelli et al. (2024); …]
- Transfer learning [Dhifallah & Lu (2021); Gerace et al. (2022, 2023, 2024); …]
- Continual learning & catastrophic forgetting [Lee et al. (2021, 2022); Shan et al (2024); …]
- Active learning [Cui et al. (2020); …]
- …

Model-based approaches to data / task selection

(Non-exhaustive list)

- Curriculum learning [Weinshall et al (2020); Saglietti et al. (2022); Abbe et al. (2023); Cornacchia et al (2023); Lee et al. (2024); Mannelli et al. (2024); …]
- Transfer learning [Dhifallah & Lu (2021); Gerace et al. (2022, 2023, 2024); …]
- Continual learning & catastrophic forgetting [Lee et al. (2021, 2022); Shan et al (2024); …]
- Active learning [Cui et al. (2020); …]
- …

Can we compute the *optimal** strategy ?

* In terms of the final performance

Supervised learning - general setup

Dataset with labels

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

Neural Network

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

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

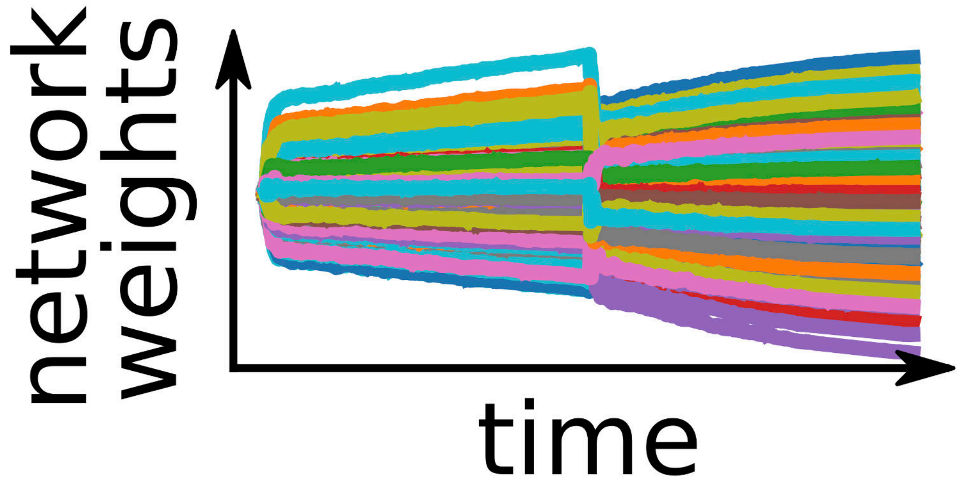
Error (aka loss)

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

Stochastic gradient descent

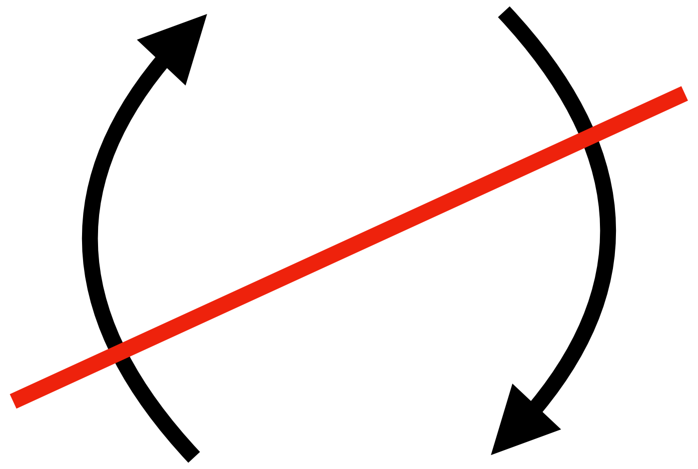
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla \mathscr{L}^{\mu}$$

Dimensionality reduction + optimal control



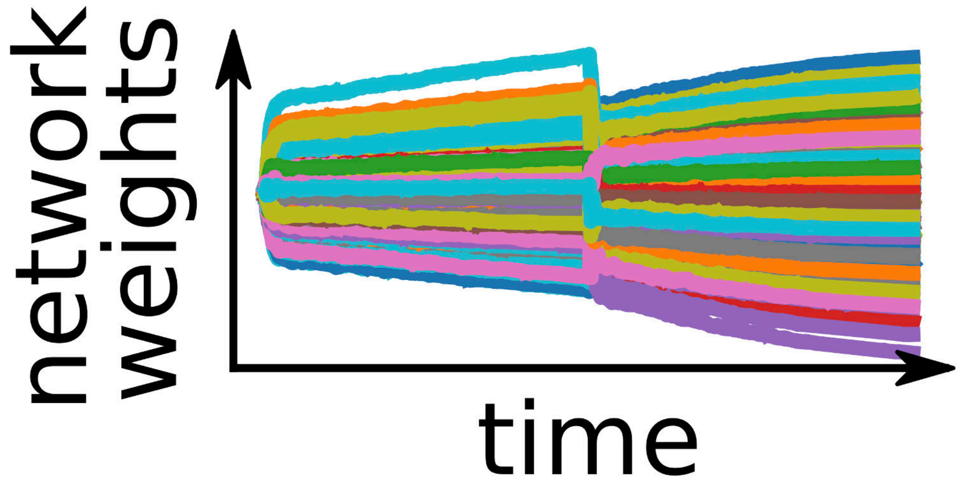
$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} - \eta \nabla \mathscr{L}^{\mu}$$

High-dimensional complex dynamics



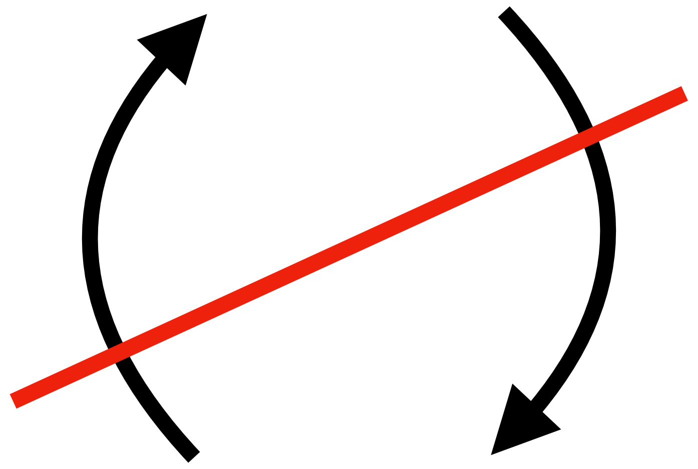
Control theory is computationally demanding

Dimensionality reduction + optimal control



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

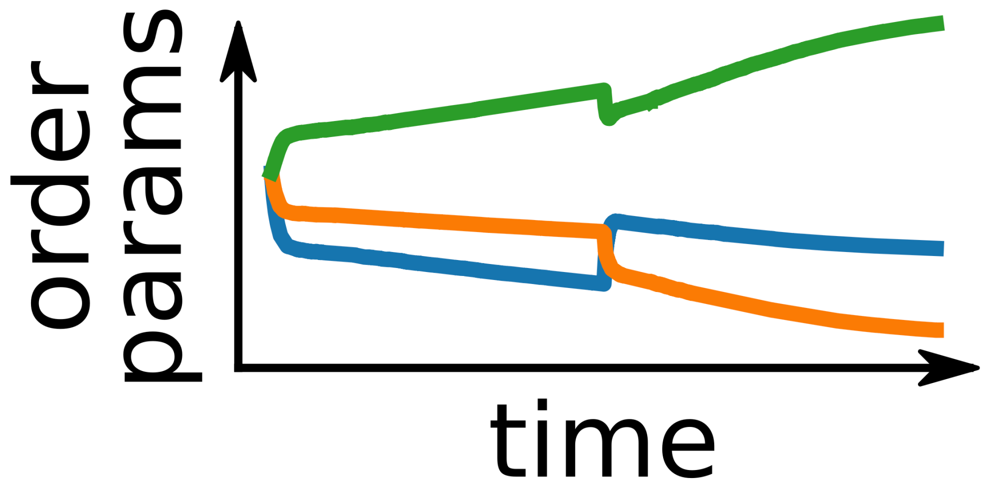
High-dimensional complex dynamics



Control theory is computationally demanding

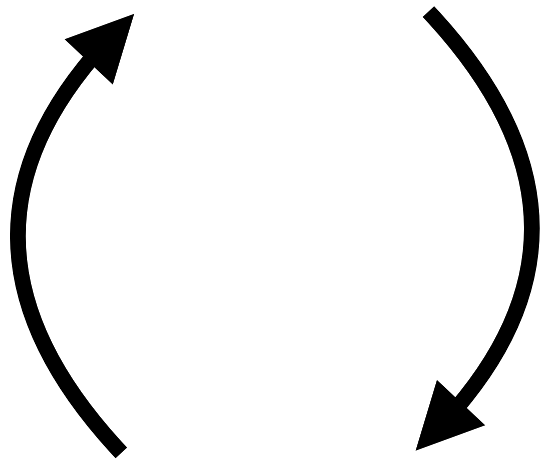
Statistical physics

[Saad & Solla (1995); Biehl & Schwarze (1995); Riegler & Biehl (1995); Goldt et al (2019); Refinetti (2020); Veiga et al (2022); Arnaboldi et al (2023); … Agoritsas et al (2018); Mignacco et al (2020); Mannelli et al (2021); Bonnaire et al (2023); …]



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

Low-dimensional effective description



Control theory can be applied

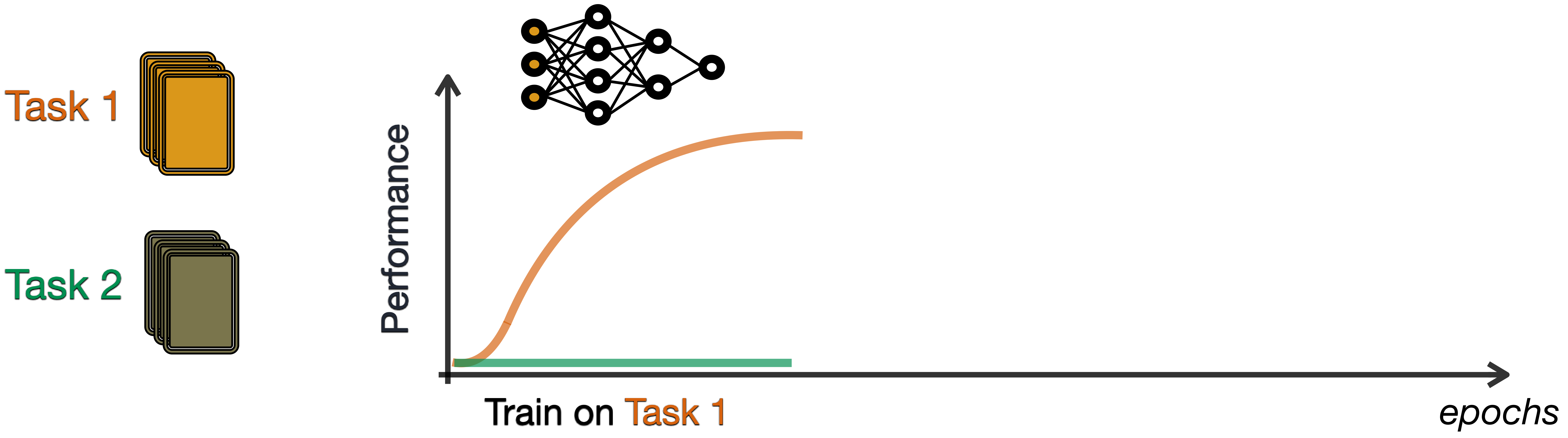
[Pontryagin (1962), Bellman (1965); Saad & Rattray(1997), Sivak & Crooks (2012); …]

Continual learning & catastrophic forgetting

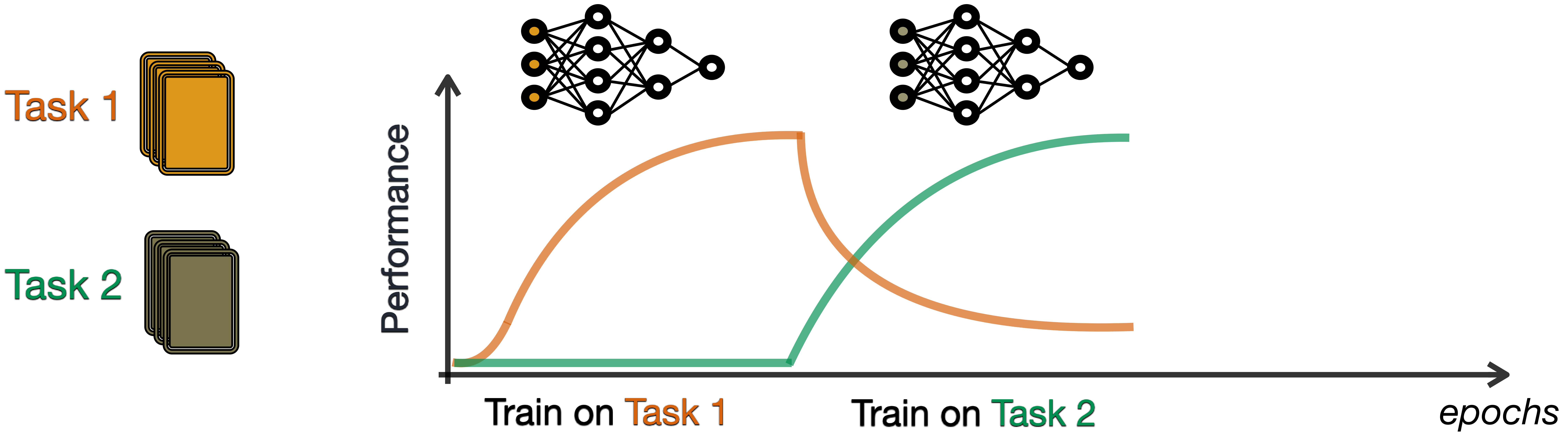
Task 1

Task 2

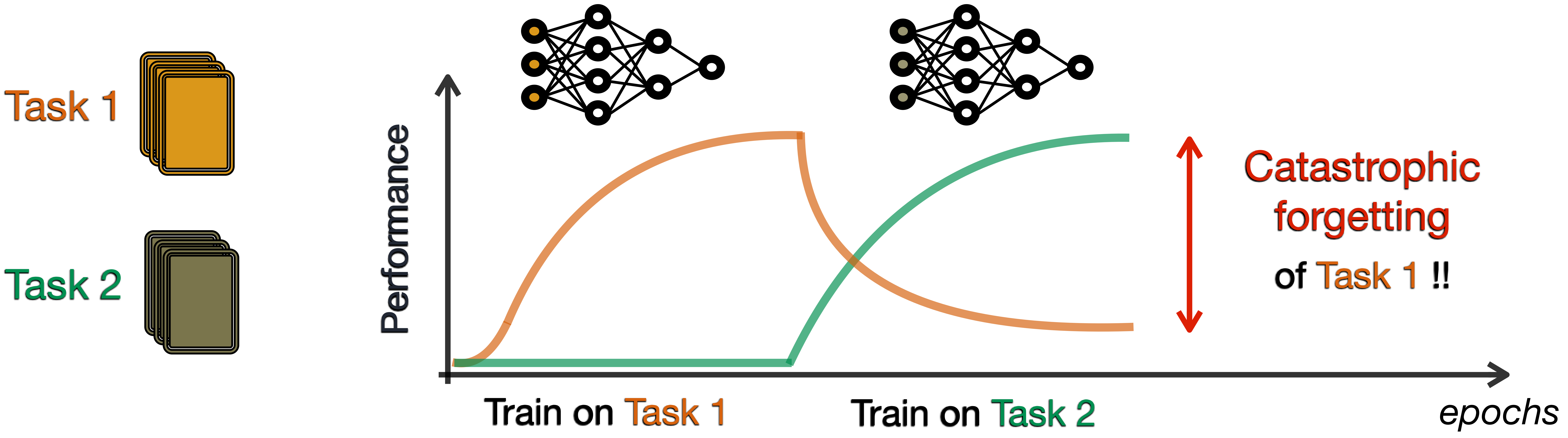
Continual learning & catastrophic forgetting



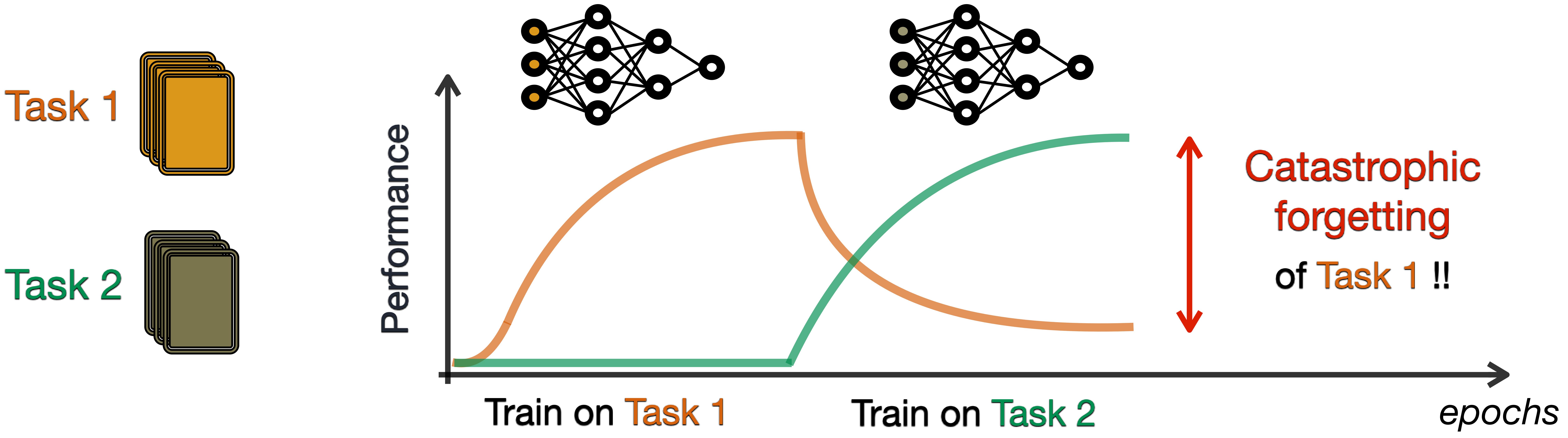
Continual learning & catastrophic forgetting



Continual learning & catastrophic forgetting



Continual learning & catastrophic forgetting



Humans & Animals

can learn sequentially without interference problems

[McClelland et al (1995); Barnett & Ceci (2002); Calvert et al., (2004); Mareschal et al (2007); Pallier et al (2003); … Flesch et al (2018); Cichon & Gan (2015); Yang et al., 2014) …]

ML (empirical):

Neural networks suffer from **catastrophic forgetting**

[Goodfellow et al (2014); Ruder & Planck (2014); Nguyen et al (2019); Parisi et al (2019); Mirzadeh et al (2020); Neyshabur et al (2020)]

ML (theory):

Key role of width, depth and **task similarity**

[Mirzadeh et al (2021); Lee et al. (2021, 2022); f Asanuma et al (2021); Doan et al (2021); Shan et al (2024)]

A teacher-student model of (**supervised**)continual learning

UNIVERSITY OF OXFORD

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

Task 1

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

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

$$N \gg 1$$

Task 2

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

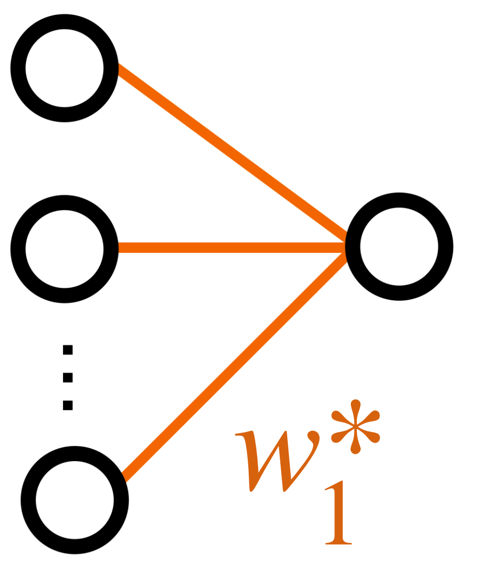
A teacher-student model of continual learning

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

Task 1

Teacher 1

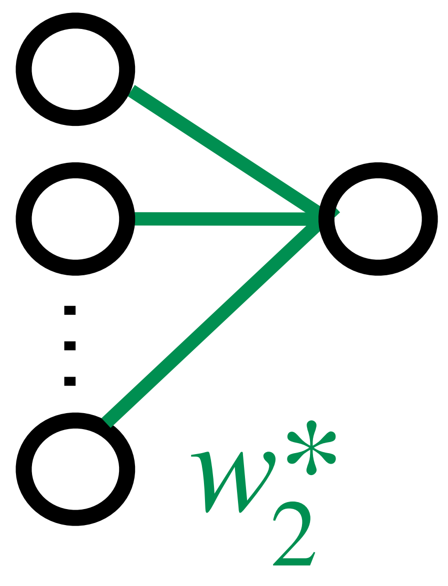
$$y^{(1)} = g_*\left(w_1^* \cdot \frac{x}{\sqrt{N}}\right)$$



Task 2

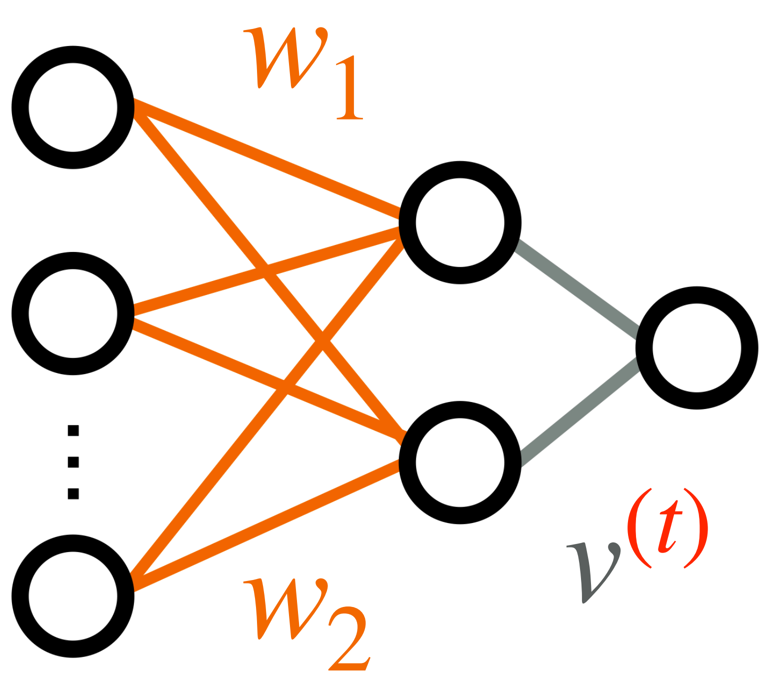
Teacher 2

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



Student

(multi-head)



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

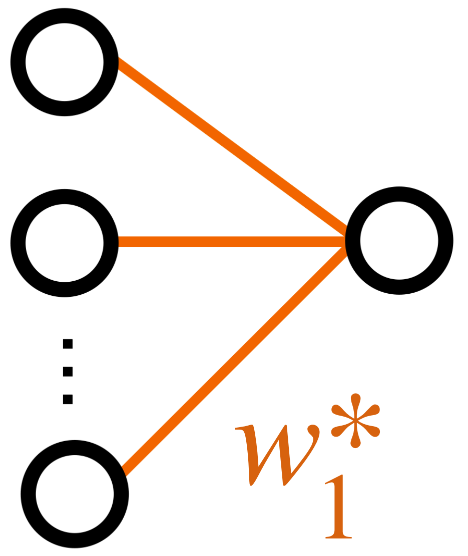
A teacher-student model of continual learning

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

Task 1

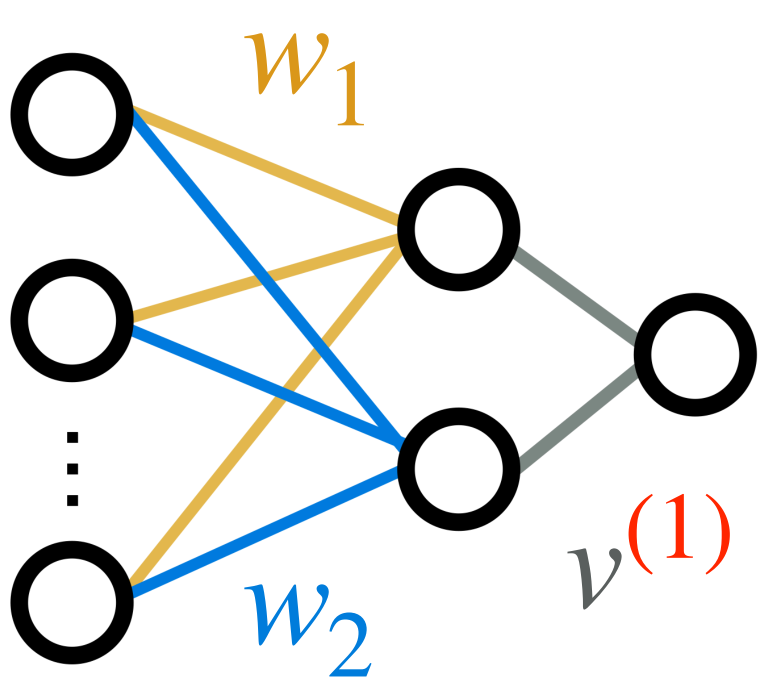
Teacher 1

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



Student

(multi-head)



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

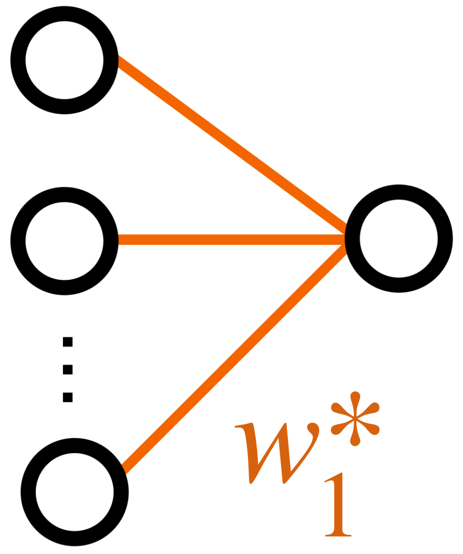
A teacher-student model of continual learning

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

Task 1

$$y^{(1)} = g_*\left(w^*_1 \cdot \frac{x}{\sqrt{N}}\right)$$

Teacher 1

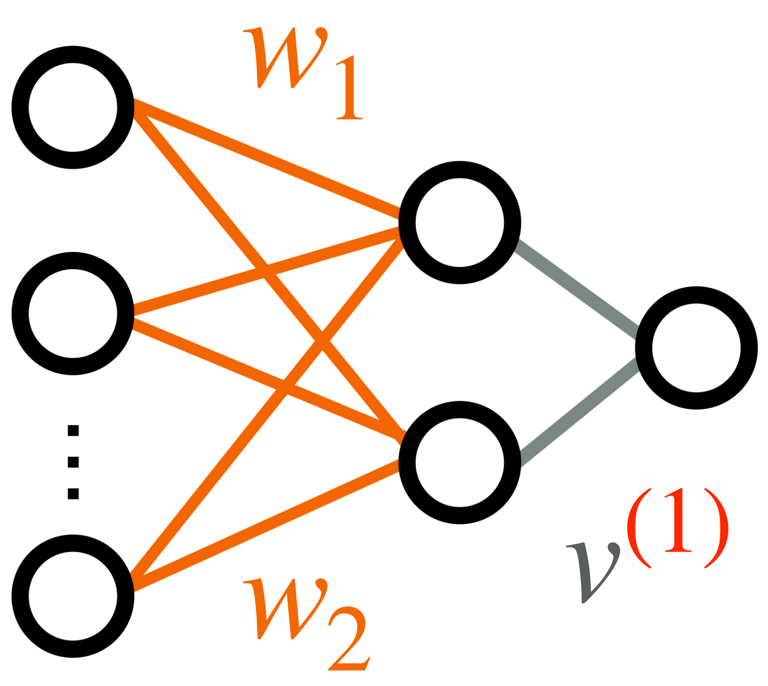


Training phase 1

Learned ✓

Student

(multi-head)



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

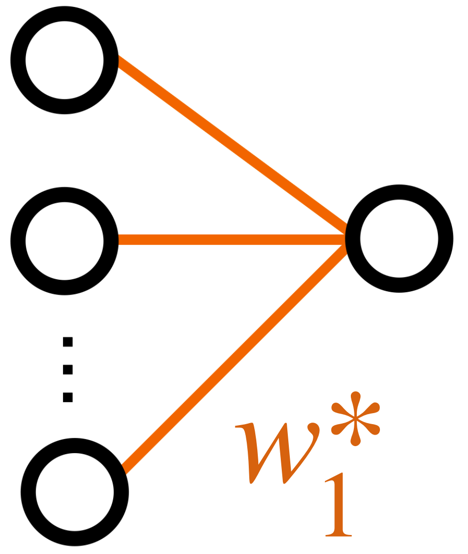
A teacher-student model of continual learning

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

Task 1

Teacher 1

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



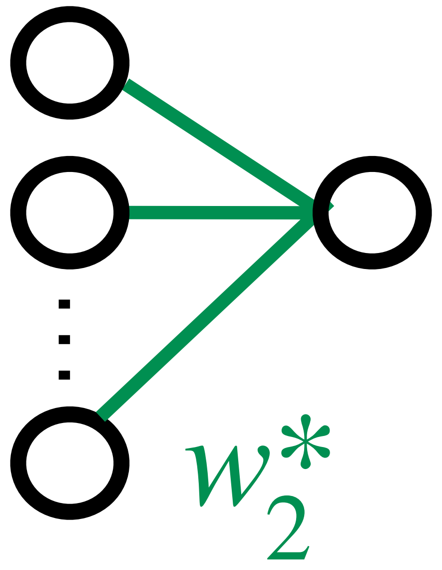
Training phase 1

Learned ✓

Task 2

Teacher 2

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

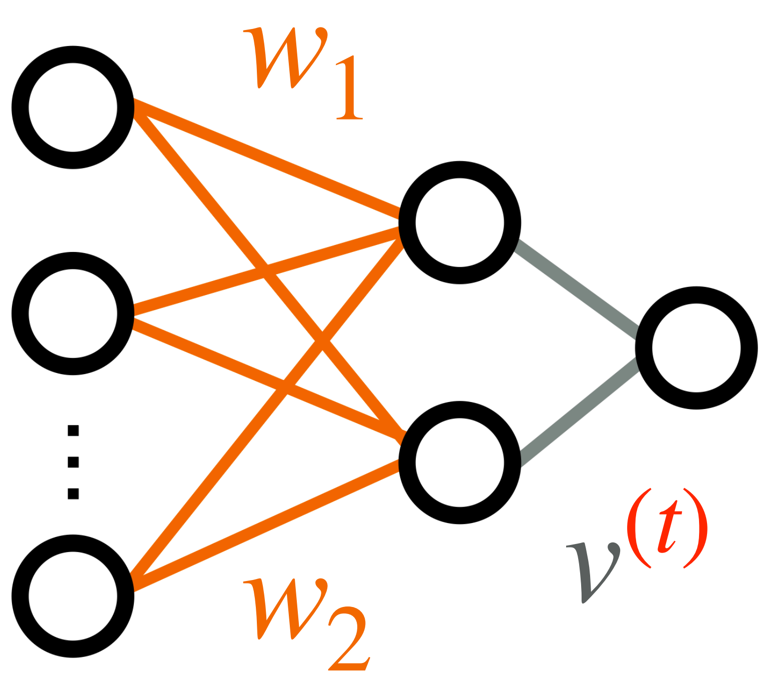


Training phase 2

To be learned without forgetting !!

Student

(multi-head)



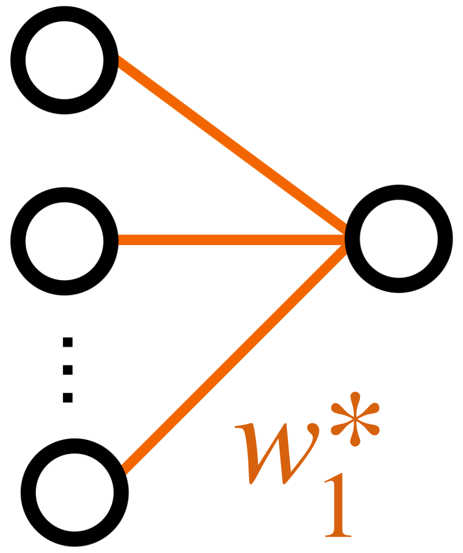
$$\hat{y}^{(t)} = \sum_{k=1}^{K=2} v_k^{(t)} g\left(w_k \cdot \frac{x}{\sqrt{N}}\right)$$

A teacher-student model of continual learning

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

Task 1

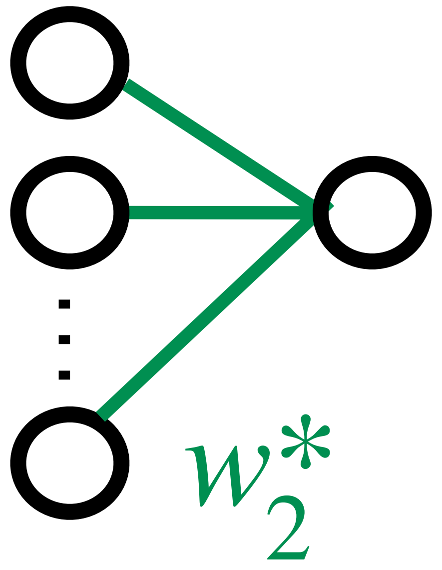
Teacher 1



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

Task 2

Teacher 2



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

Training phase 1

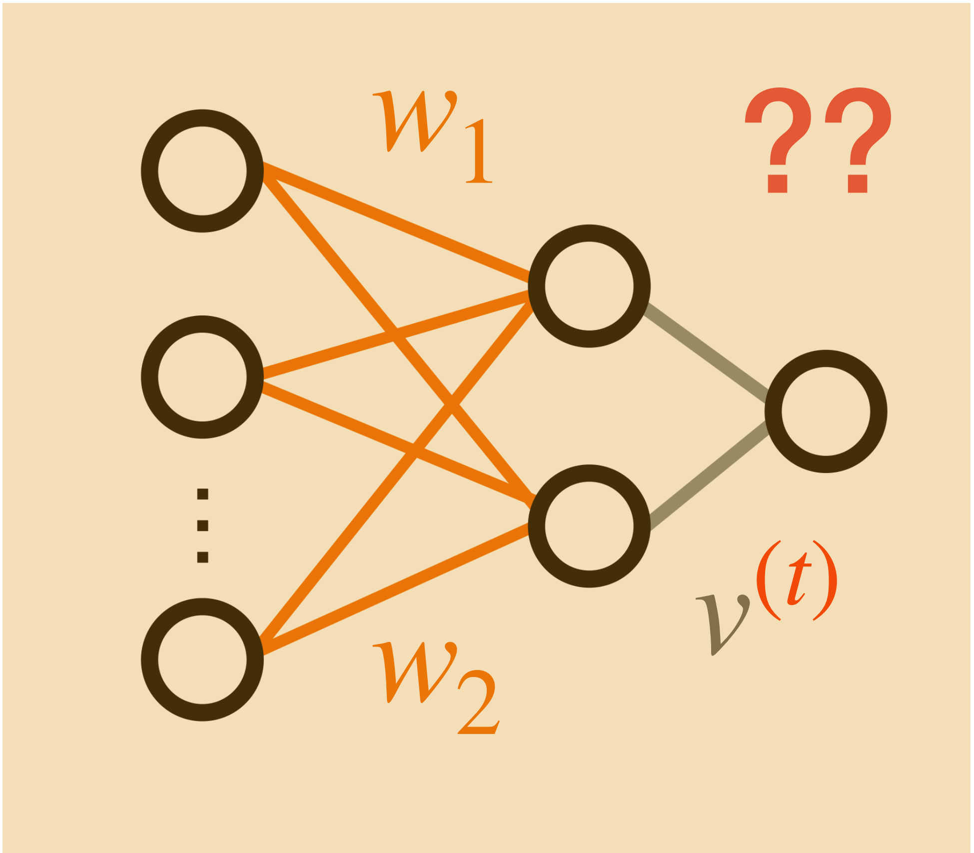
Learned ✓

Training phase 2

To be learned without forgetting !!

Student

(multi-head)



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

1. The error depends only on the order params

Generalization error on task t :

$$\varepsilon_t(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{W}_*) = \frac{1}{2}\mathbb{E}_x\left[\left(y^{(t)} - \hat{y}^{(t)}\right)^2\right]$$

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

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

Saad & Solla (1995); Biehl & Schwarze (1995);

1. The error depends only on the order params

Generalization error on task t :

$$\varepsilon_t\left(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{W}_*\right) = \frac{1}{2}\mathbb{E}_x\left[\left(y^{(t)} - \hat{y}^{(t)}\right)^2\right]$$

$$= \frac{1}{2}\sum_{k,h} v_k^{(t)} v_h^{(t)} \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}_*}\left[g(\lambda_k)g(\lambda_h)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}_*}\left[g_*(\lambda_*^{(t)})^2\right]$$

$$-\sum_k v_k^{(t)} \mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}_*}\left[g(\lambda_k)g_*(\lambda_*^{(t)})\right]$$

Pre-activations: $\lambda_k = \dfrac{w_k \cdot x}{\sqrt{N}}$ and $\lambda_*^{(t)} = \dfrac{w_*^{(t)} \cdot x}{\sqrt{N}}$

Saad & Solla (1995); Biehl & Schwarze (1995);

1. The error depends only on the order params

UNIVERSITY OF OXFORD

Generalization error on task t :

$$\varepsilon_t\left(\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{W}_*\right) = \mathbb{E}_x\left[f(\lambda, \lambda_*)\right]$$

Pre-activations: $\lambda_k = \dfrac{w_k \cdot x}{\sqrt{N}}$ and $\lambda_*^{(t)} = \dfrac{w_*^{(t)} \cdot x}{\sqrt{N}}$ are ***jointly Gaussian.***

Saad & Solla (1995); Biehl & Schwarze (1995);

1. The error depends only on the order params

Generalization error on task t :

$$\varepsilon_t\left(\boldsymbol{W},\boldsymbol{V},\boldsymbol{W}_*\right) = \mathbb{E}_x\left[f(\lambda,\lambda_*)\right]$$

$$P(\boldsymbol{\lambda},\boldsymbol{\lambda}_*) = \frac{1}{\sqrt{(2\pi)^{K+T}|\boldsymbol{C}|}}\exp\left(-\frac{1}{2}(\boldsymbol{\lambda},\boldsymbol{\lambda}_*)^\top\boldsymbol{C}^{-1}(\boldsymbol{\lambda},\boldsymbol{\lambda}_*)\right)$$

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{M} \\ \boldsymbol{M}^\top & \boldsymbol{S} \end{pmatrix}.$$

Pre-activations: $\lambda_k = \dfrac{w_k \cdot x}{\sqrt{N}}$ and $\lambda_*^{(t)} = \dfrac{w_*^{(t)} \cdot x}{\sqrt{N}}$ are ***jointly Gaussian.***

The second moments are given by the *order parameters* :

$$M_{kt} := \mathbb{E}_{\boldsymbol{x}}\left[\lambda_k\lambda_*^{(t)}\right] = \frac{\boldsymbol{w}_k\cdot\boldsymbol{w}_*^{(t)}}{N},$$

$$Q_{kh} := \mathbb{E}_{\boldsymbol{x}}\left[\lambda_k\lambda_h\right] = \frac{\boldsymbol{w}_k\cdot\boldsymbol{w}_h}{N},$$

$$S_{tt'} := \mathbb{E}_{\boldsymbol{x}}\left[\lambda_*^{(t)}\lambda_*^{(t')}\right] = \frac{\boldsymbol{w}_*^{(t')}\cdot\boldsymbol{w}_*^{(t)}}{N}.$$

Saad & Solla (1995); Biehl & Schwarze (1995);

2. Exact ODEs for the order parameters

$\mu = 1, \ldots, P =$ training epoch

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

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

Saad & Solla (1995); Biehl & Schwarze (1995);

2. Exact ODEs for the order parameters

$\mu = 1,\ldots,P =$ training epoch

Online SGD:

$$\boldsymbol{w}_k^{\mu+1} = \boldsymbol{w}_k^{\mu} - \eta^{\mu}\Delta^{(t_c)\mu}\, v_k^{(t_c)\mu}\, g'\left(\lambda_k^{\mu}\right)\frac{\boldsymbol{x}^{\mu}}{\sqrt{N}}\,,$$

$$\Delta^{(t)\mu} := \hat{y}^{(t)\mu} - y^{(t)\mu} = \sum_{k=1}^{K} v_k^{(t)} g(\lambda_k^{\mu}) - g_*(\lambda_*^{(t)\mu})$$

Pre-activations: $\lambda_k^{\mu} := \dfrac{\boldsymbol{x}^{\mu}\cdot\boldsymbol{w}_k^{\mu}}{\sqrt{N}}\,, \quad \lambda_*^{(t)\mu} := \dfrac{\boldsymbol{x}^{\mu}\cdot\boldsymbol{w}_*^{(t)}}{\sqrt{N}}\,.$

Saad & Solla (1995); Biehl & Schwarze (1995);

2. Exact ODEs for the order parameters

$\mu = 1,\ldots,P =$ training epoch

Online SGD: $$\boldsymbol{w}_k^{\mu+1} = \boldsymbol{w}_k^{\mu} - \eta^{\mu}\Delta^{(t_c)\mu}\, v_k^{(t_c)\mu}\, g'\left(\lambda_k^{\mu}\right)\frac{\boldsymbol{x}^{\mu}}{\sqrt{N}},$$

Example derivation: ODE for the "magnetization" $M_{kt} := \mathbb{E}_{\boldsymbol{x}}\left[\lambda_k \lambda_*^{(t)}\right] = \frac{\boldsymbol{w}_k \cdot \boldsymbol{w}_*^{(t)}}{N}$

$$\frac{\boldsymbol{w}_k^{\mu+1} \cdot \boldsymbol{w}_*^{(t)}}{N} - \frac{\boldsymbol{w}_k^{\mu} \cdot \boldsymbol{w}_*^{(t)}}{N} = -\frac{\eta^{\mu}}{N}\Delta^{(t_c)\mu} v_k^{(t_c)\mu} g'(\lambda_k^{\mu})\lambda_*^{(t)\mu}$$

Saad & Solla (1995); Biehl & Schwarze (1995);

2. Exact ODEs for the order parameters

$\mu = 1,\ldots,P =$ training epoch

Online SGD: $$\boldsymbol{w}_k^{\mu+1} = \boldsymbol{w}_k^{\mu} - \eta^{\mu}\Delta^{(t_c)\mu}\, v_k^{(t_c)\mu}\, g'\left(\lambda_k^{\mu}\right)\frac{\boldsymbol{x}^{\mu}}{\sqrt{N}},$$

Example derivation: ODE for the "magnetization" $M_{kt} := \mathbb{E}_{\boldsymbol{x}}\left[\lambda_k \lambda_*^{(t)}\right] = \dfrac{\boldsymbol{w}_k \cdot \boldsymbol{w}_*^{(t)}}{N}$

$$\frac{\boldsymbol{w}_k^{\mu+1}\cdot \boldsymbol{w}_*^{(t)}}{N} - \frac{\boldsymbol{w}_k^{\mu}\cdot \boldsymbol{w}_*^{(t)}}{N} = -\frac{\eta^{\mu}}{N}\Delta^{(t_c)\mu} v_k^{(t_c)\mu} g'(\lambda_k^{\mu})\lambda_*^{(t)\mu}$$

$\alpha = \mu/N =$ training "time", high-dimensional limit with $P/N = \mathcal{O}_N(1)$:

$$\frac{\mathrm{d}M_{kt}}{\mathrm{d}\alpha} = -\eta v_k^{(t_c)}\,\mathbb{E}_{\boldsymbol{\lambda},\boldsymbol{\lambda}_*}\left[\Delta^{(t_c)} g'(\lambda_k)\lambda_*^{(t)}\right] := f_{\boldsymbol{M},kt}$$

Saad & Solla (1995); Biehl & Schwarze (1995);

2. Exact ODEs for the order parameters

UNIVERSITY OF OXFORD

ODEs for the order parameters:

$$\mathbb{Q} = (\mathrm{vec}(\boldsymbol{Q}), \mathrm{vec}(\boldsymbol{M}), \mathrm{vec}(\boldsymbol{V}))^\top$$

$$\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} = f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

$$\alpha \in (0, \alpha_F]$$

Control

- Task
- Learning rate
- …

Saad & Solla (1995); Biehl & Schwarze (1995);

3. Meta-optimization

ODEs for the order parameters:

$$\mathbb{Q} = (\mathrm{vec}(\boldsymbol{Q}), \mathrm{vec}(\boldsymbol{M}), \mathrm{vec}(\boldsymbol{V}))^\top$$

$$\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} = f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

$$\alpha \in (0, \alpha_F]$$

Forward dynamics

Optimization objective:

$$h(\mathbb{Q}(\alpha_F)) = \sum_{t=1}^{T} c_t\, \varepsilon_t(\mathbb{Q}(\alpha_F))$$

$$\text{with} \quad c_t \geq 0 \text{ and } \sum_{t=1}^{T} c_t = 1$$

Cost functional:

[Pontryagin (1962)]

$$\mathcal{F}[\mathbb{Q}, \hat{\mathbb{Q}}, \boldsymbol{u}] = h\left(\mathbb{Q}(\alpha_F)\right) + \int_0^{\alpha_F} \mathrm{d}\alpha\, \hat{\mathbb{Q}}(\alpha)^\top \left[-\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} + f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)\right]$$

3. Meta-optimization

ODEs for the order parameters:

$$\mathbb{Q} = (\mathrm{vec}(\boldsymbol{Q}), \mathrm{vec}(\boldsymbol{M}), \mathrm{vec}(\boldsymbol{V}))^\top$$

$$\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} = f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

$$\alpha \in (0, \alpha_F]$$

Forward dynamics

Optimization objective:

$$h(\mathbb{Q}(\alpha_F)) = \sum_{t=1}^{T} c_t\, \varepsilon_t(\mathbb{Q}(\alpha_F))$$

$$\text{with} \quad c_t \geq 0 \text{ and } \sum_{t=1}^{T} c_t = 1$$

Cost functional: $\mathcal{F}[\mathbb{Q}, \hat{\mathbb{Q}}, \boldsymbol{u}] = h\left(\mathbb{Q}(\alpha_F)\right) + \int_0^{\alpha_F} \mathrm{d}\alpha\, \hat{\mathbb{Q}}(\alpha)^\top \left[-\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} + f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)\right]$

[Pontryagin (1962)]

$$-\frac{\mathrm{d}\hat{\mathbb{Q}}(\alpha)^\top}{\mathrm{d}\alpha} = \hat{\mathbb{Q}}(\alpha)^\top \nabla_{\mathbb{Q}} f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

$$\hat{\mathbb{Q}}(\alpha_F) = \nabla_{\mathbb{Q}} h(\mathbb{Q}(\alpha_F))$$

Backwards dynamics

3. Meta-optimization

ODEs for the order parameters:

$$\mathbb{Q} = (\mathrm{vec}(\boldsymbol{Q}), \mathrm{vec}(\boldsymbol{M}), \mathrm{vec}(\boldsymbol{V}))^\top$$

$$\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} = f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

$$\alpha \in (0, \alpha_F]$$

Forward dynamics

Optimization objective:

$$h(\mathbb{Q}(\alpha_F)) = \sum_{t=1}^{T} c_t\, \varepsilon_t(\mathbb{Q}(\alpha_F))$$

$$\text{with} \quad c_t \geq 0 \text{ and } \sum_{t=1}^{T} c_t = 1$$

Cost functional: $\mathcal{F}[\mathbb{Q}, \hat{\mathbb{Q}}, \boldsymbol{u}] = h\left(\mathbb{Q}(\alpha_F)\right) + \int_0^{\alpha_F} \mathrm{d}\alpha\, \hat{\mathbb{Q}}(\alpha)^\top \left[-\frac{\mathrm{d}\mathbb{Q}(\alpha)}{\mathrm{d}\alpha} + f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)\right]$

[Pontryagin (1962)]

$$-\frac{\mathrm{d}\hat{\mathbb{Q}}(\alpha)^\top}{\mathrm{d}\alpha} = \hat{\mathbb{Q}}(\alpha)^\top \nabla_{\mathbb{Q}} f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), \boldsymbol{u}(\alpha)\right)$$

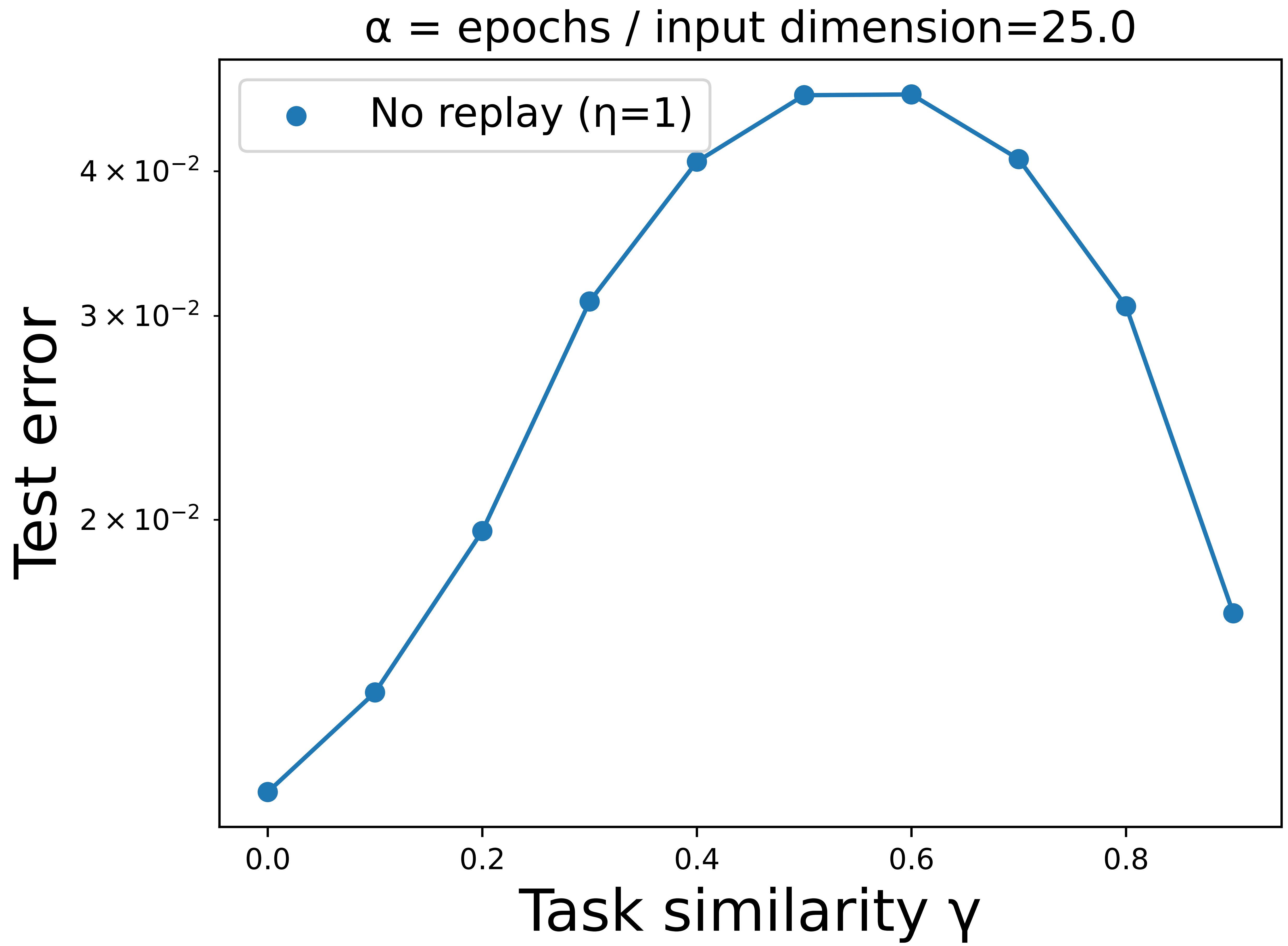
$$\hat{\mathbb{Q}}(\alpha_F) = \nabla_{\mathbb{Q}} h(\mathbb{Q}(\alpha_F))$$

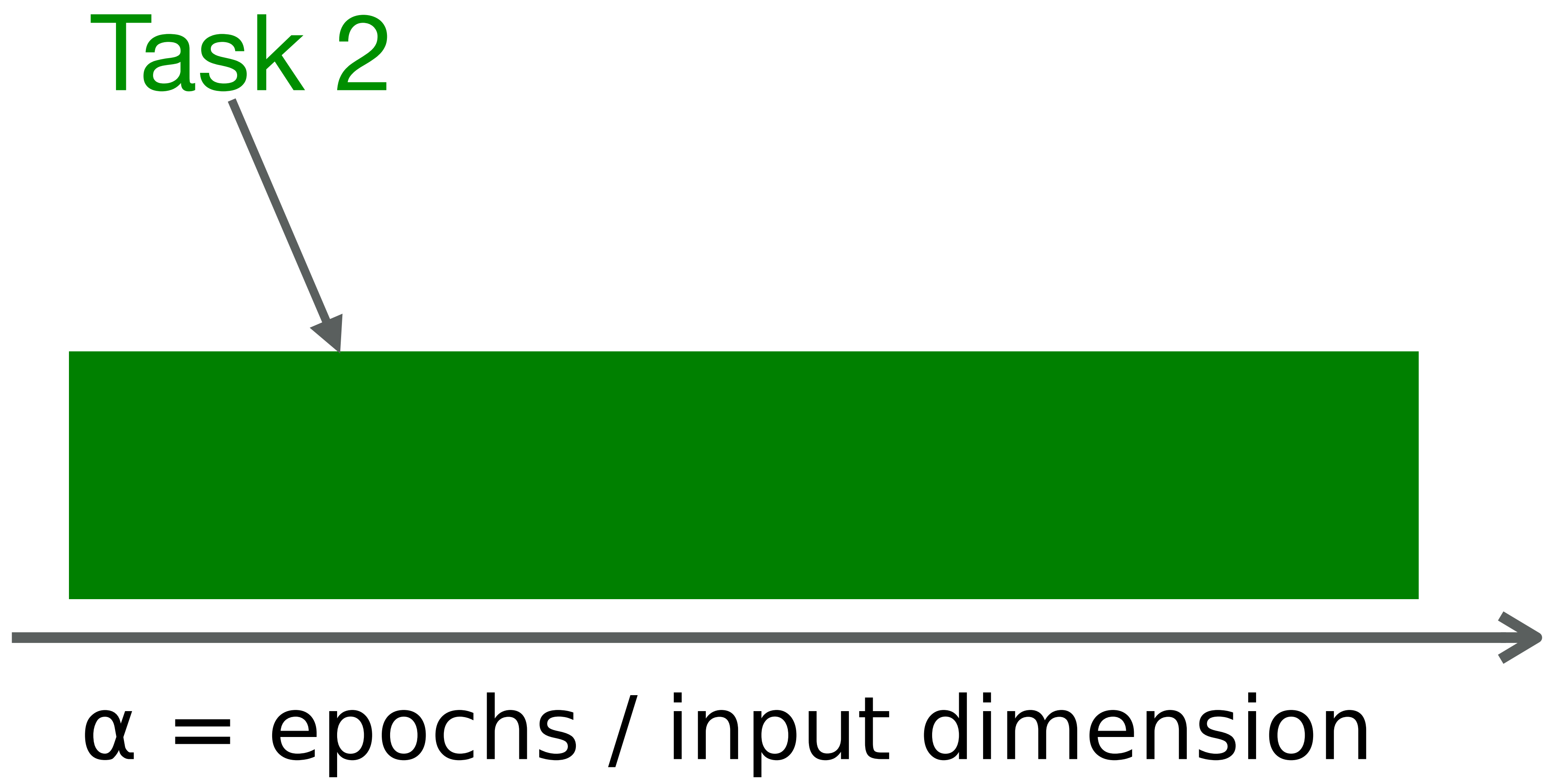
Backwards dynamics

Optimal protocol: $u^*(\alpha) = \mathrm{argmin}_u \left\{ \hat{\mathbb{Q}}(\alpha)^\top f_{\mathbb{Q}}\left(\mathbb{Q}(\alpha), u\right) \right\}$

Training only on the new task

Control: $\mathbf{u} =$ task





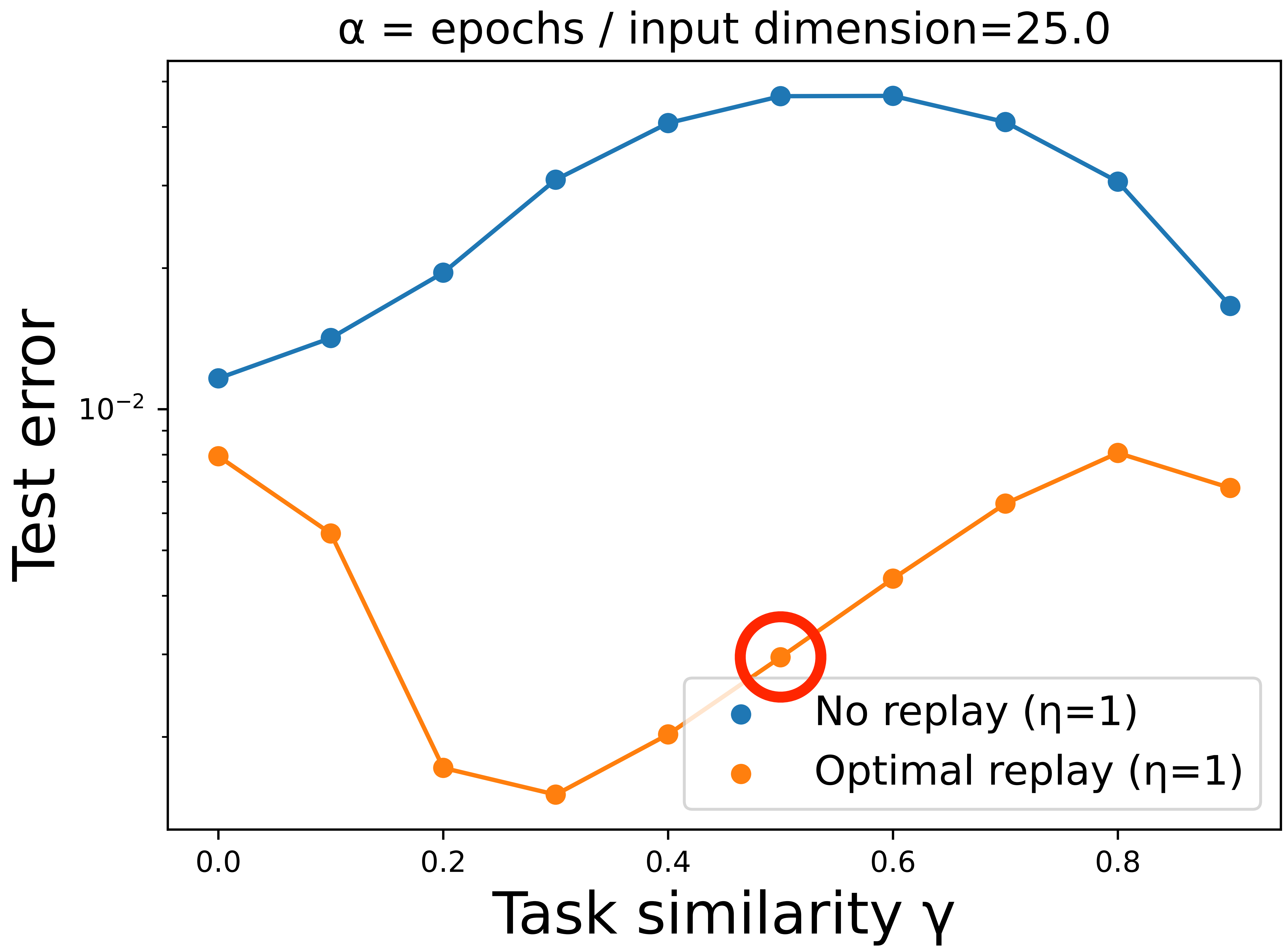
Intermediate task similarity leads to the worst forgetting.

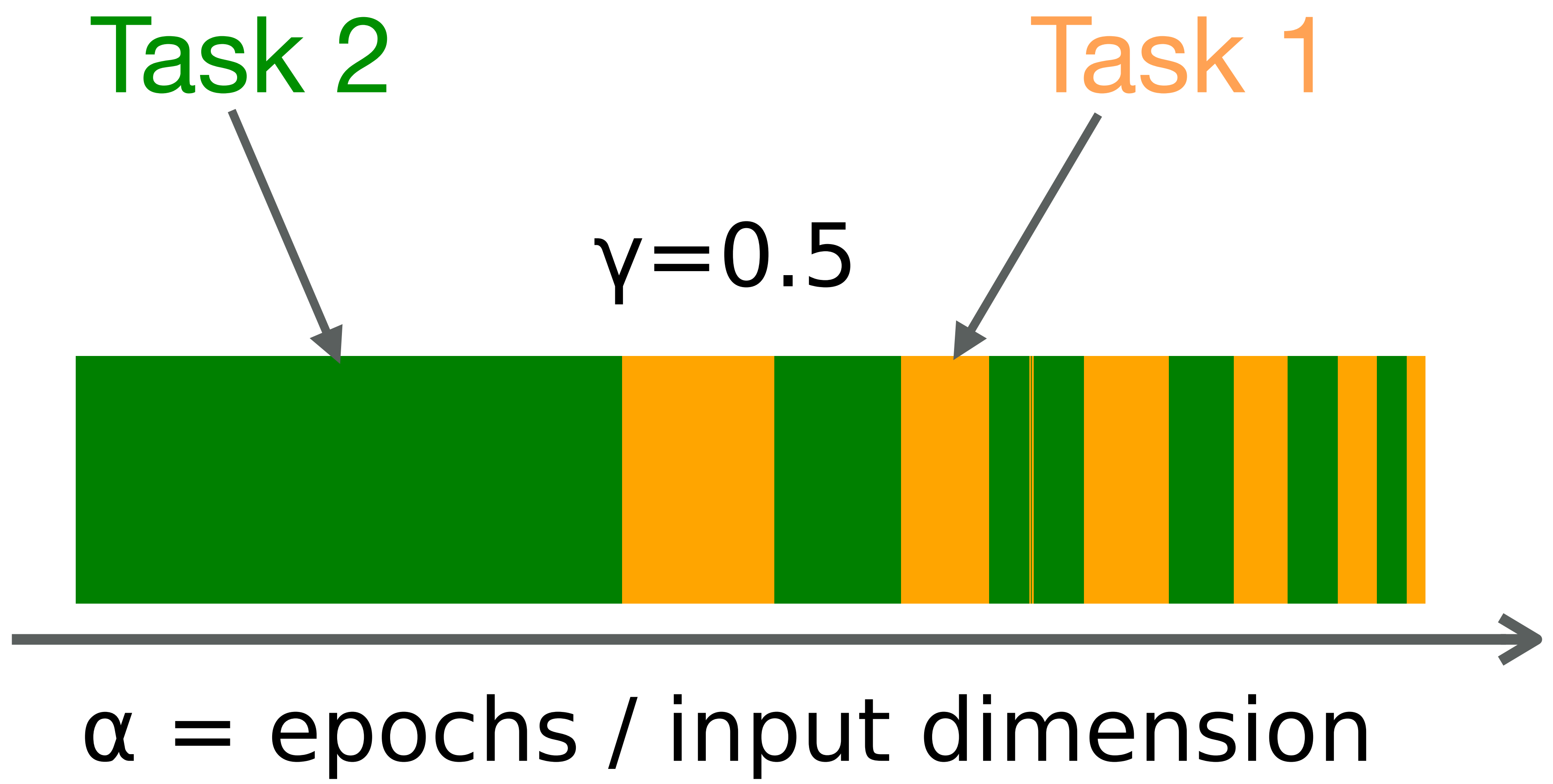
[Lee et al (2021); Ramesh et al (2020); Doan et al (2020); Nguyen et al (2019)]

[F. Mori, S. Sarao Mannelli, **FM**, *in preparation*]

The impact of replay on the performance

Control: $\mathbf{u} =$ task

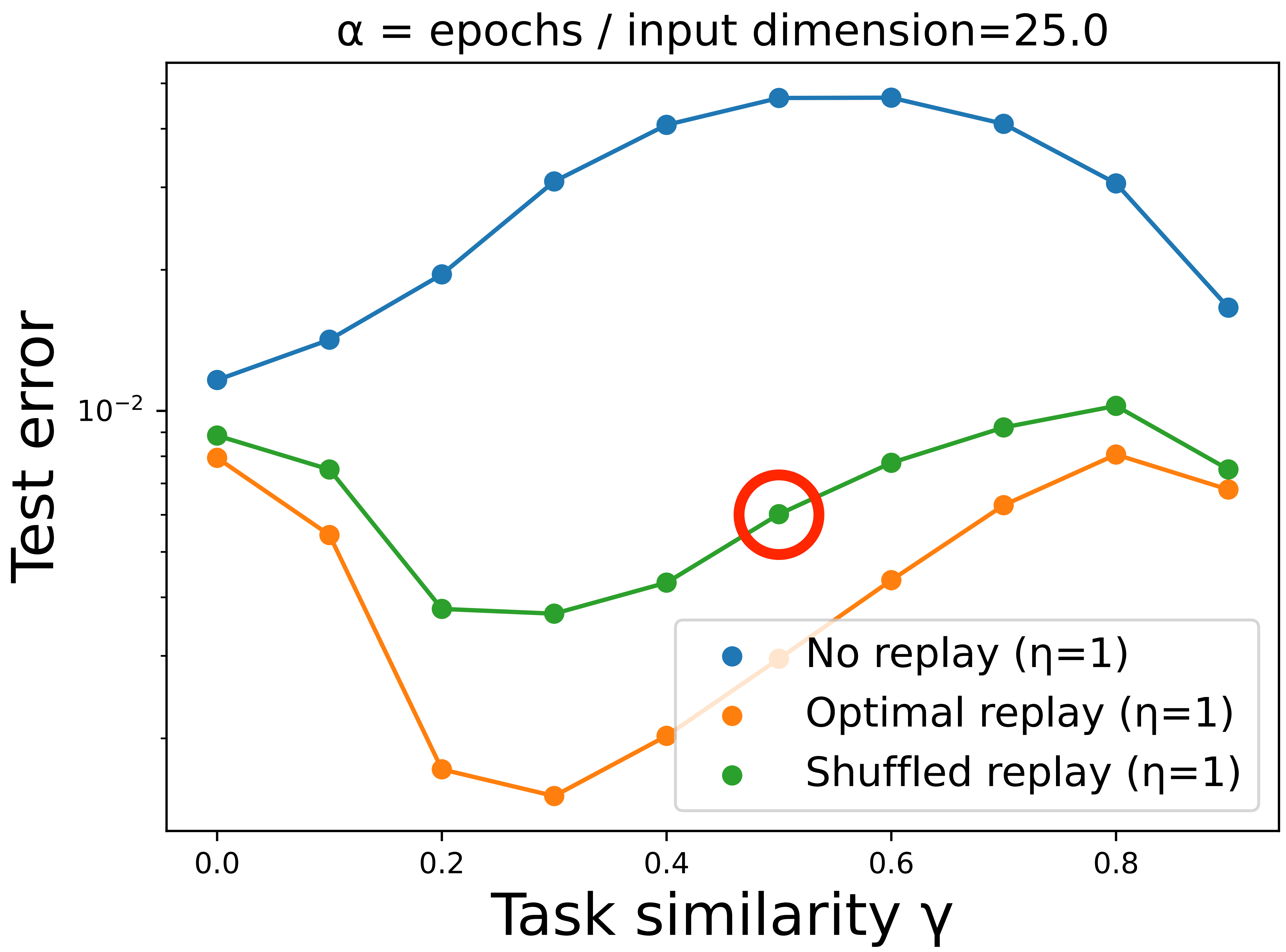




[F. Mori, S. Sarao Mannelli, **FM**, *in preparation*]

The impact of replay on the performance

Control: $\mathbf{u} =$ task



γ=0.5

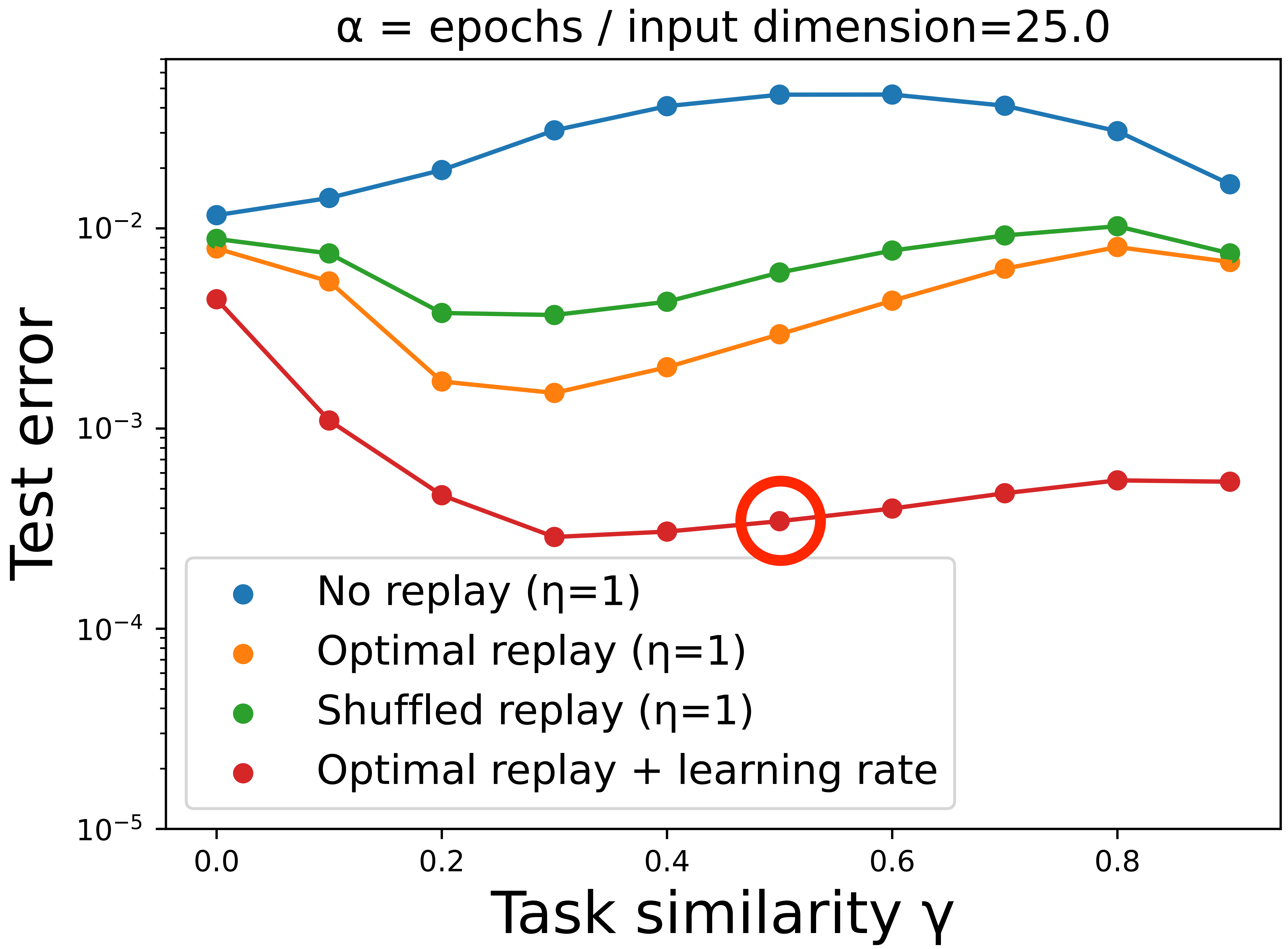
α = epochs / input dimension

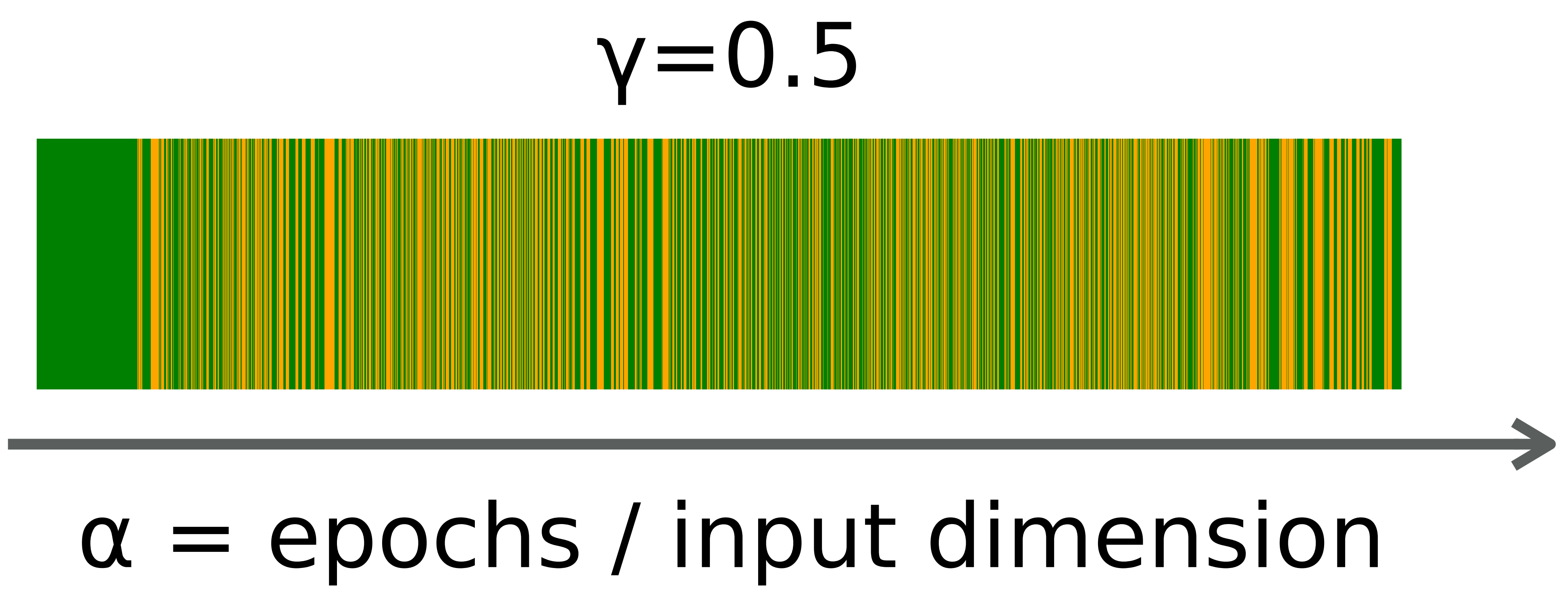
Order matters!

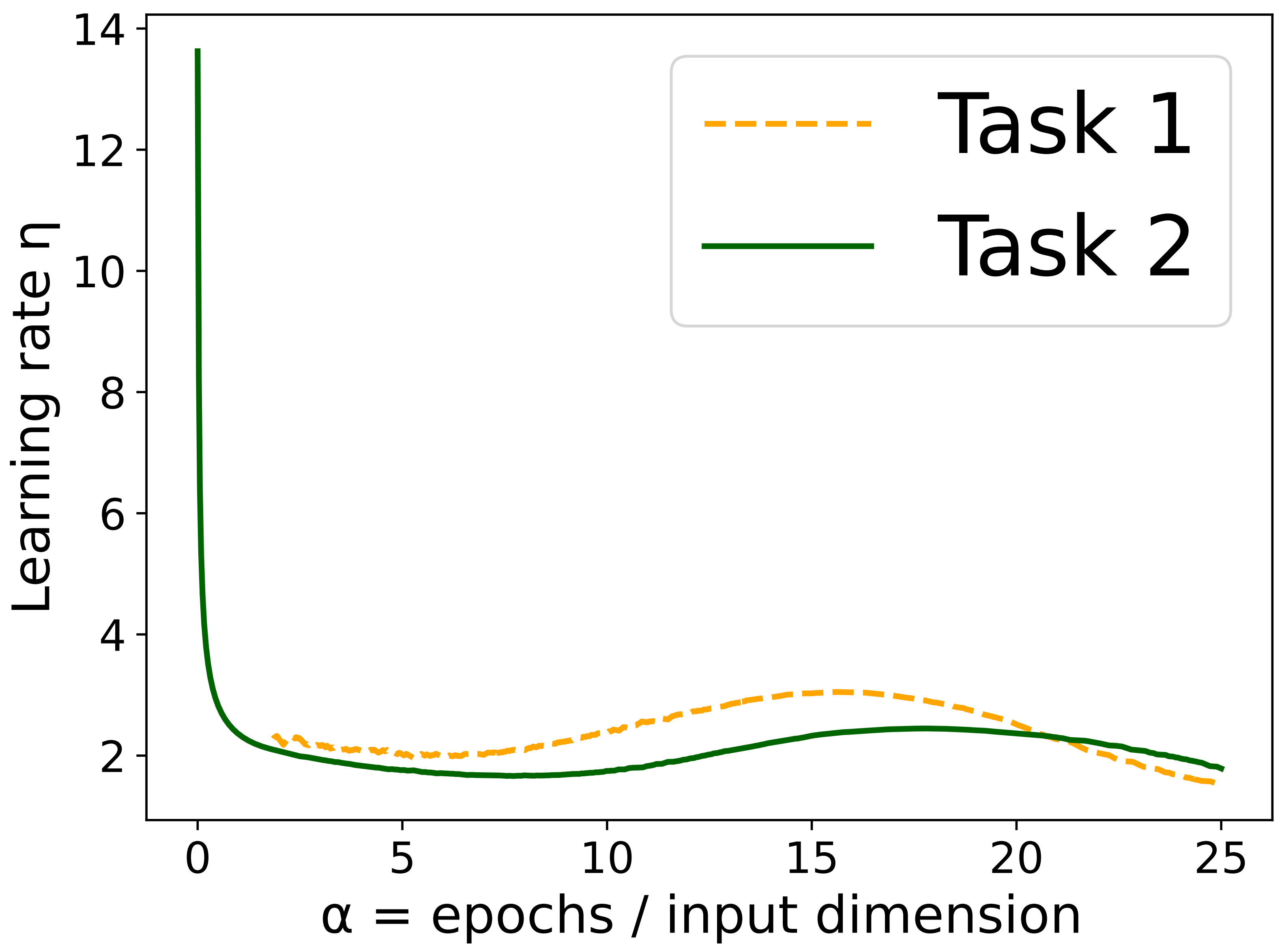
[F. Mori, S. Sarao Mannelli, **FM**, *in preparation*]

Optimizing replay + learning rate schedule

Control: $\mathbf{u} =$ task

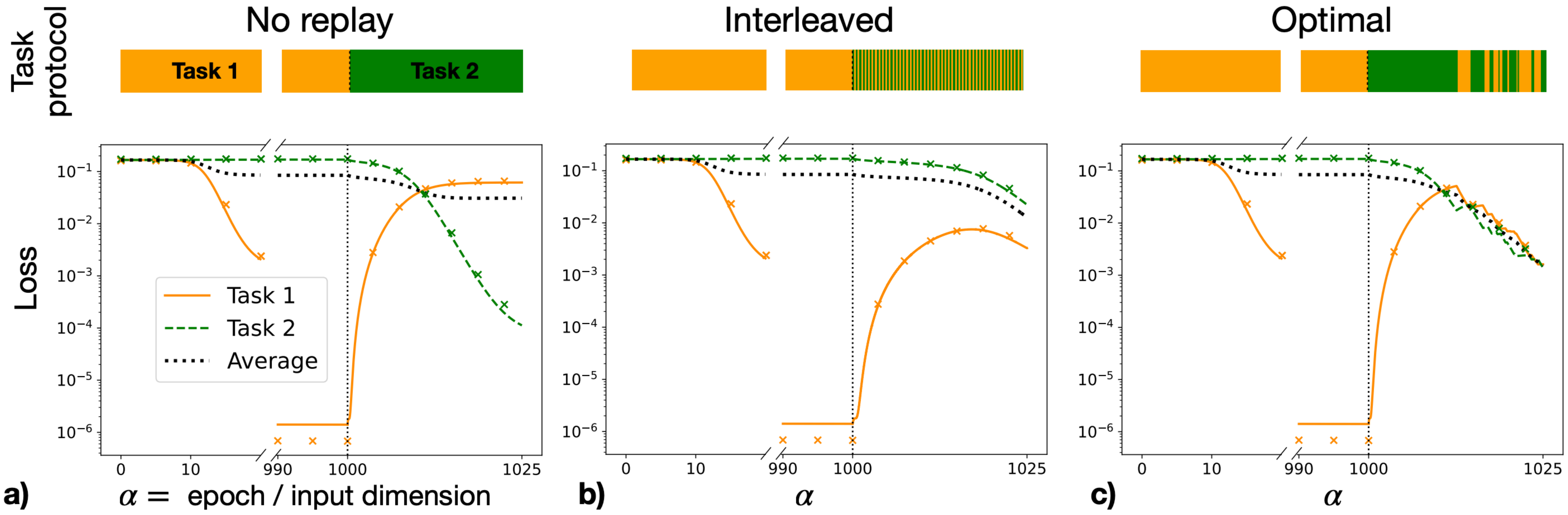




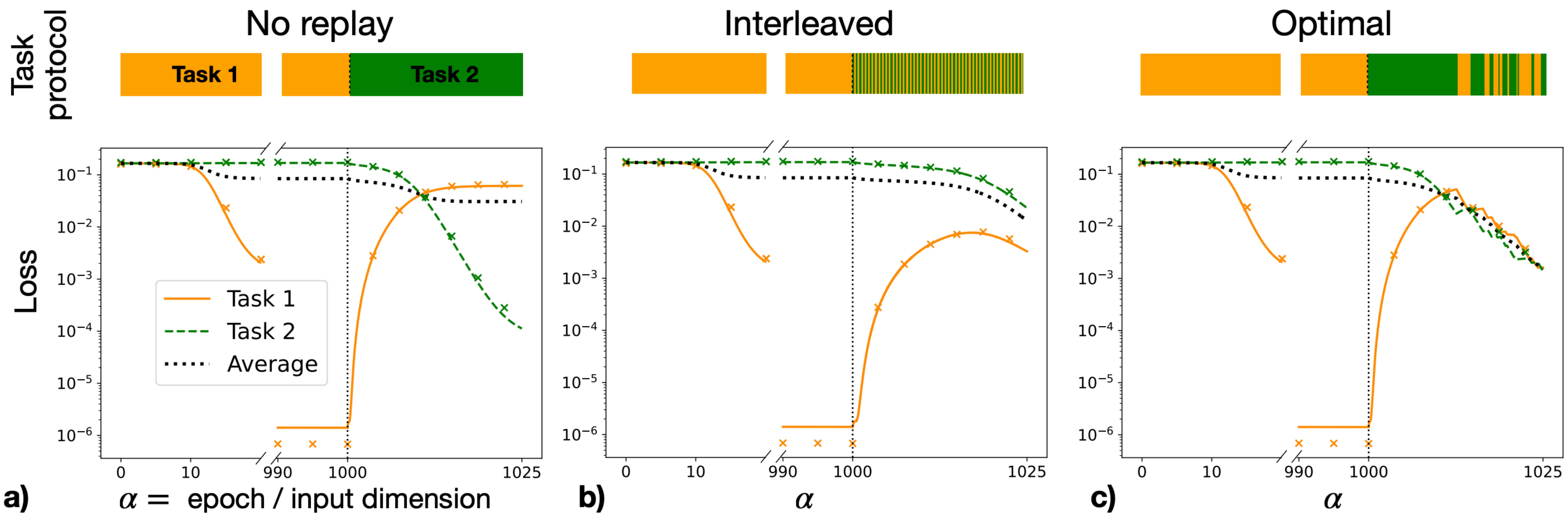


[F. Mori, S. Sarao Mannelli, **FM**, *in preparation*]

Results: optimal strategy vs benchmarks

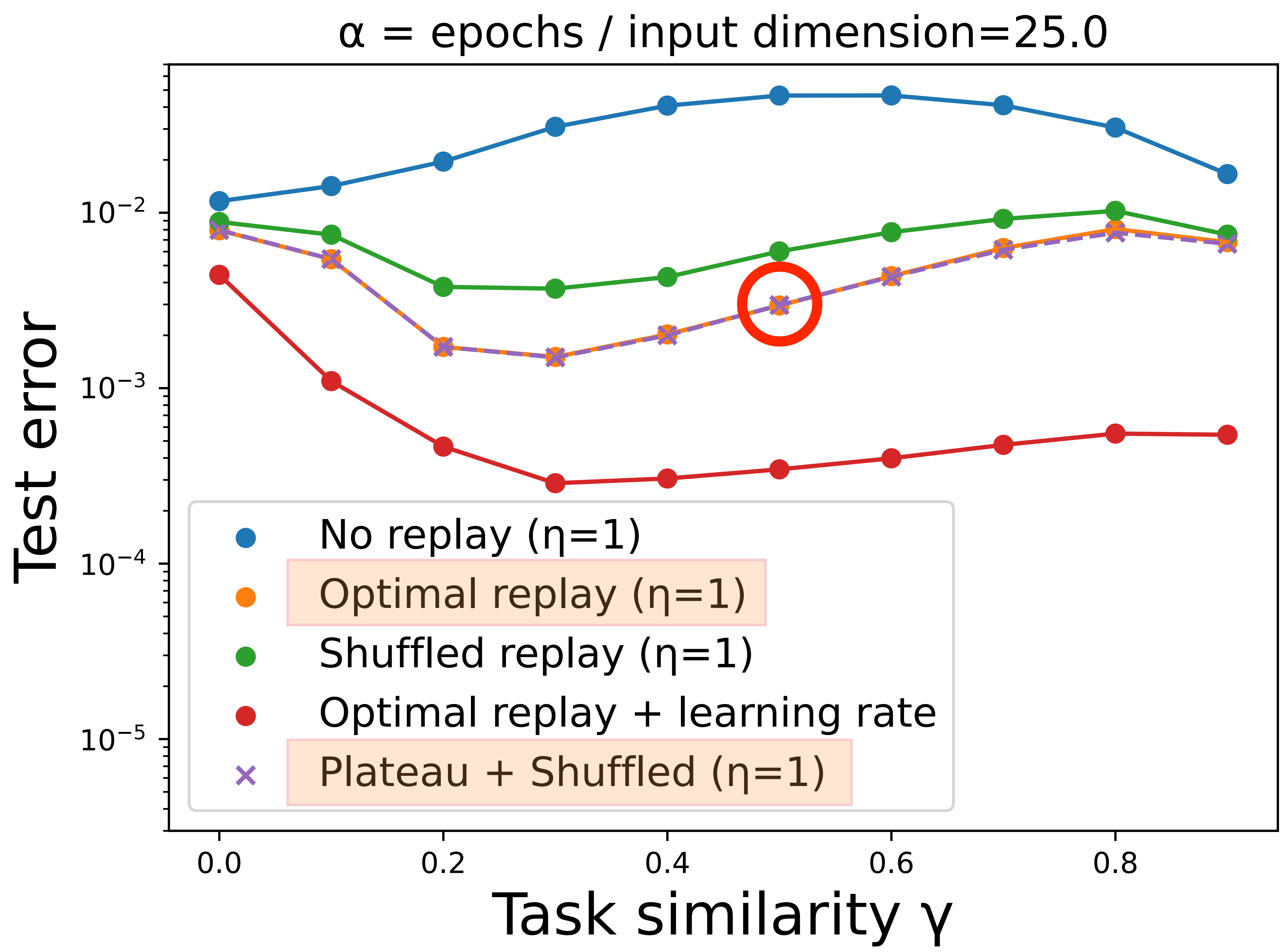


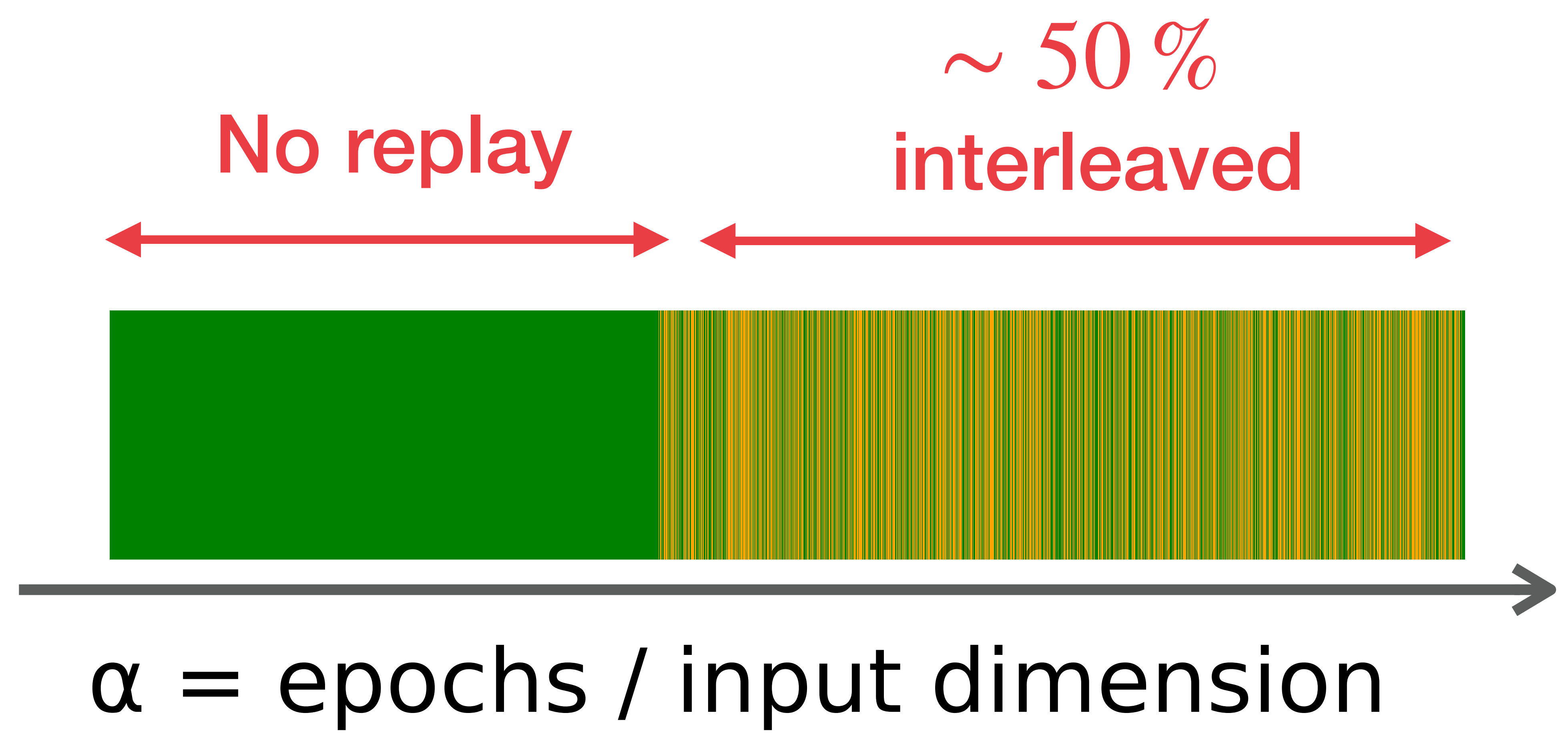
Results: optimal strategy vs benchmarks



Does the order of replayed episodes matter?

Learning the new **vs** *Replaying the old*





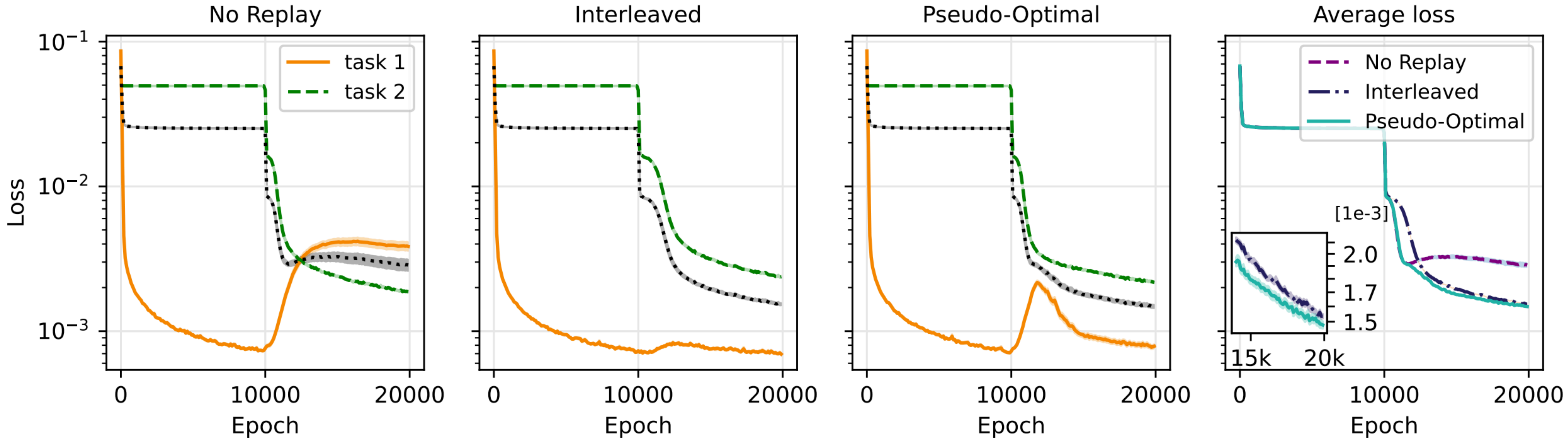
γ=0.5

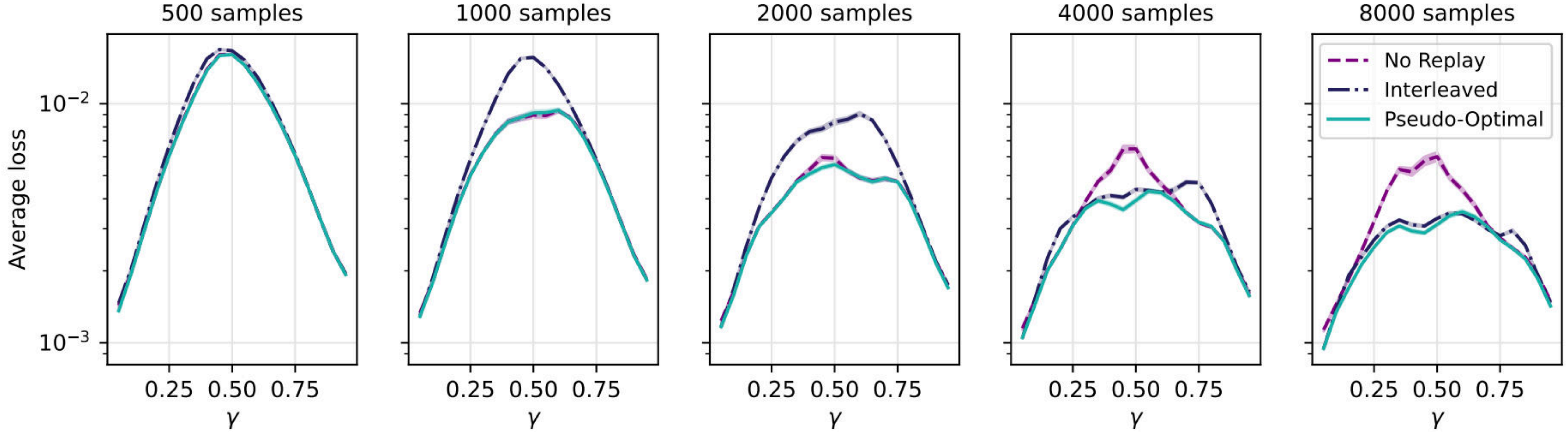
Order does not matter after the no-replay phase!

[F. Mori, S. Sarao Mannelli, **FM**, *in preparation*]

Experiments on Fashion MNIST

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





Conclusions & Perspectives

In summary:

- Optimal control of effective learning dynamics reveals **nontrivial training protocols**.
- Continual learning: **non-homogeneous replay** avoids forgetting.

Many open directions!

- Experiments beyond toy models: incorporate models of structured data;
- Batch learning;
- Other learning paradigms : shaping, transfer learning, active learning, …

Thank you!

Bonus Slides

Curriculum learning

(in progress)

Curriculum learning

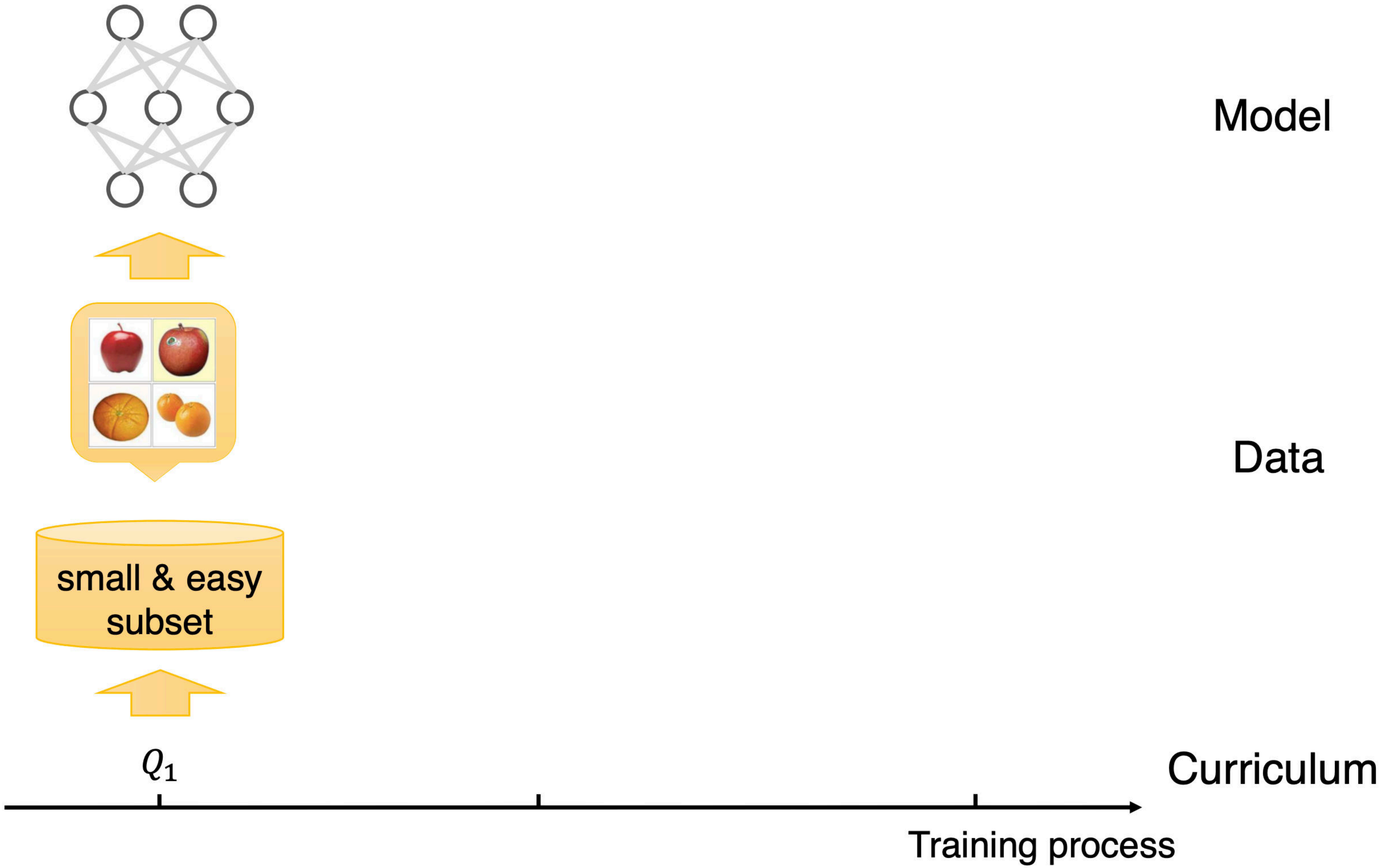


Image from: *Wang, Xin, Yudong Chen, and Wenwu Zhu. IEEE transactions on pattern analysis and machine intelligence 44.9 (2021):4555-4576.*

Curriculum learning

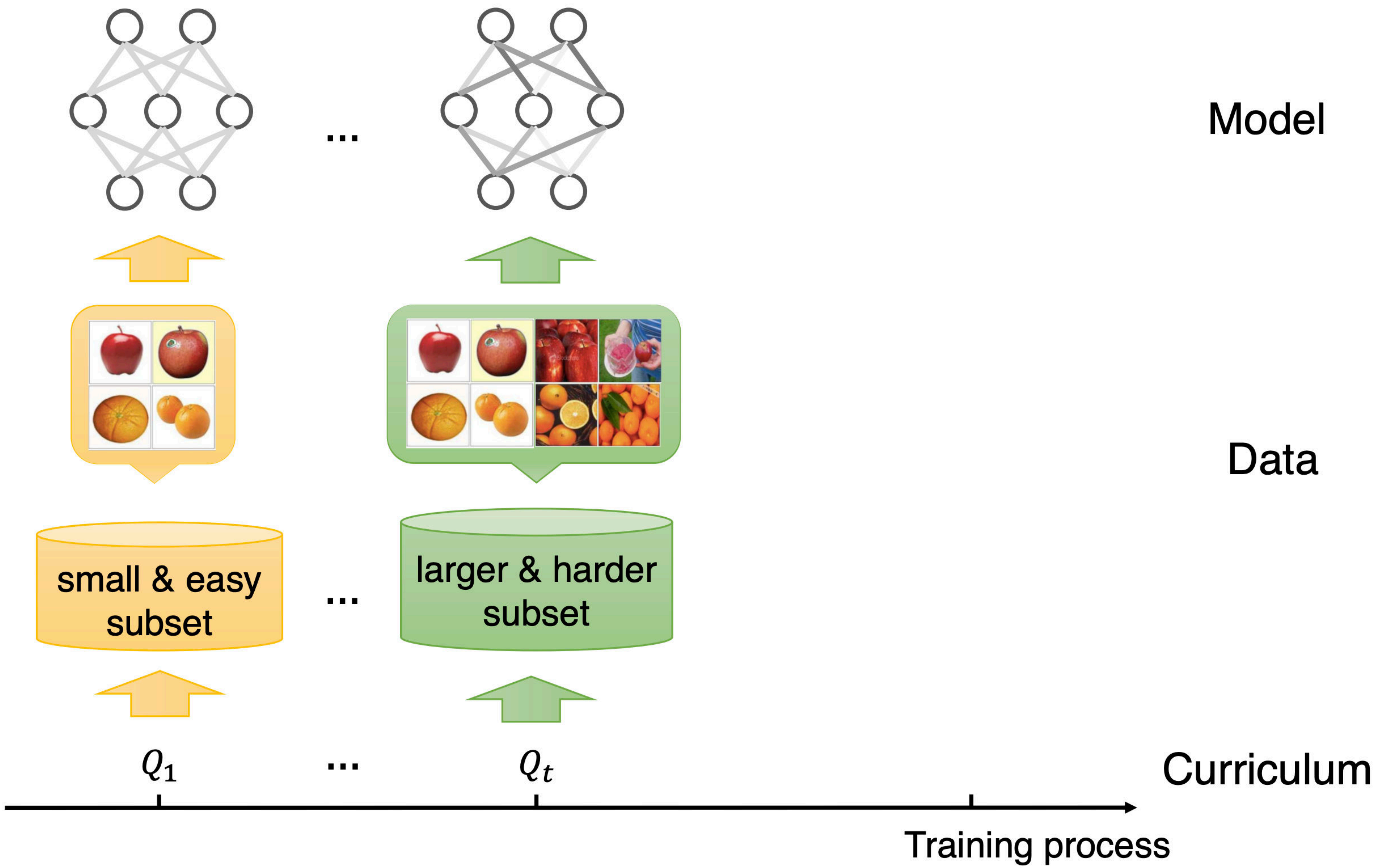


Image from: *Wang, Xin, Yudong Chen, and Wenwu Zhu. IEEE transactions on pattern analysis and machine intelligence 44.9 (2021):4555-4576.*

Curriculum learning

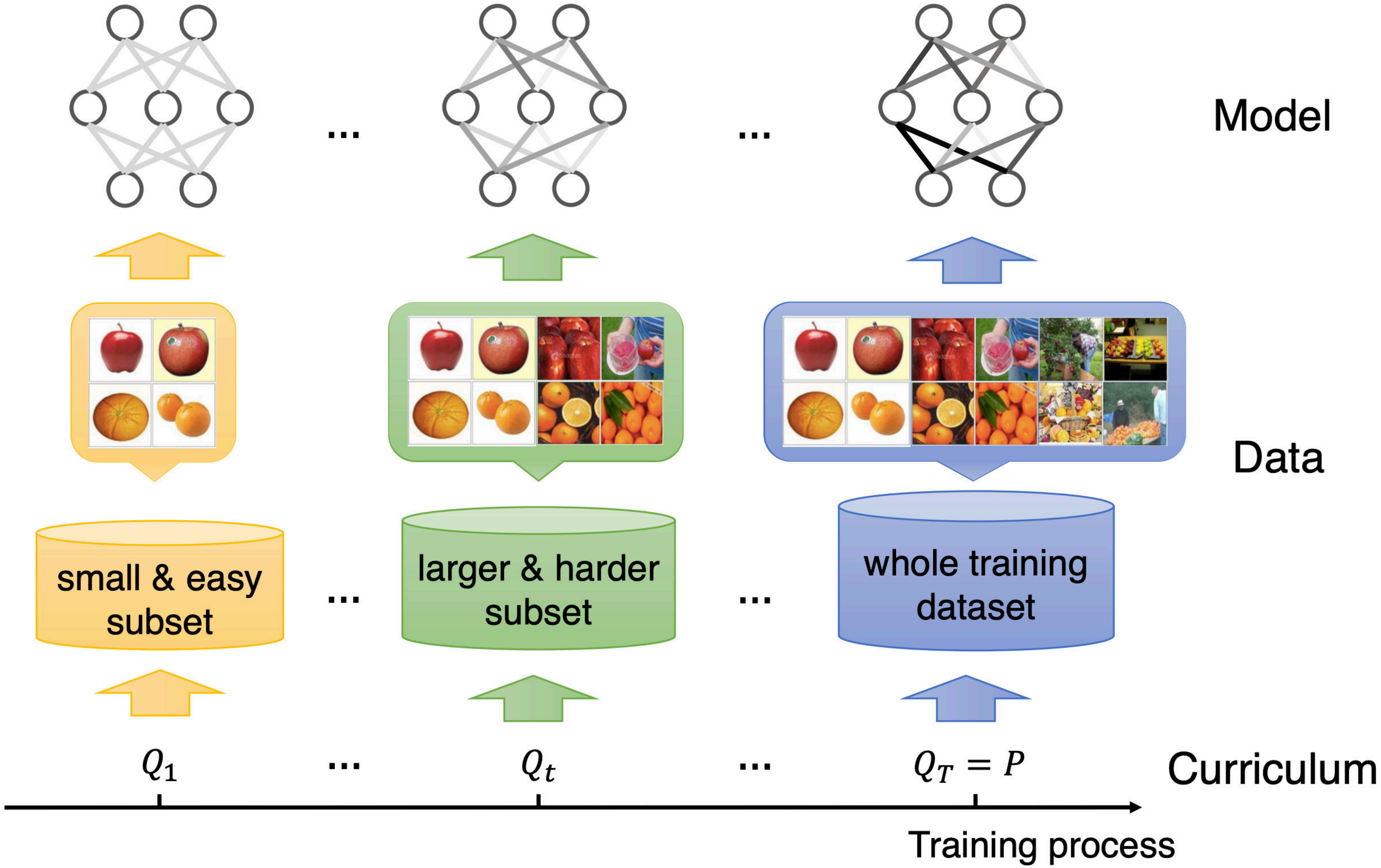


Image from: *Wang, Xin, Yudong Chen, and Wenwu Zhu. IEEE transactions on pattern analysis and machine intelligence 44.9 (2021):4555-4576.*

Curriculum learning

Animals:

- Conditional reflexes (dogs) [Pavlov (1927)]
- Shaping (rats, pigeons) [Skinner (1938)]
- Discrimination along a continuum (rats) [Lawrence (1952)]
- Cross-species auditory identification (rats, humans) [Liu et al. (2008)]

Humans:

- Discrimination along a continuum [Baker, Stanley (1954)]
- Past tense [Plunkett et al (1990; 1991)]
- Fading with auditory and visual stimuli [Pashler, Mozer (2013)]
- Eureka effect [Ahissar, Hochstein (1997)]

Curriculum learning

UNIVERSITY OF OXFORD

Animals:

- Conditional reflexes (dogs) [Pavlov (1927)]
- Shaping (rats, pigeons) [Skinner (1938)]
- Discrimination along a continuum (rats) [Lawrence (1952)]
- Cross-species auditory identification (rats, humans) [Liu et al. (2008)]

Humans:

- Discrimination along a continuum [Baker, Stanley (1954)]
- Past tense [Plunkett et al (1990; 1991)]
- Fading with auditory and visual stimuli [Pashler, Mozer (2013)]
- Eureka effect [Ahissar, Hochstein (1997)]

ML (empirical):

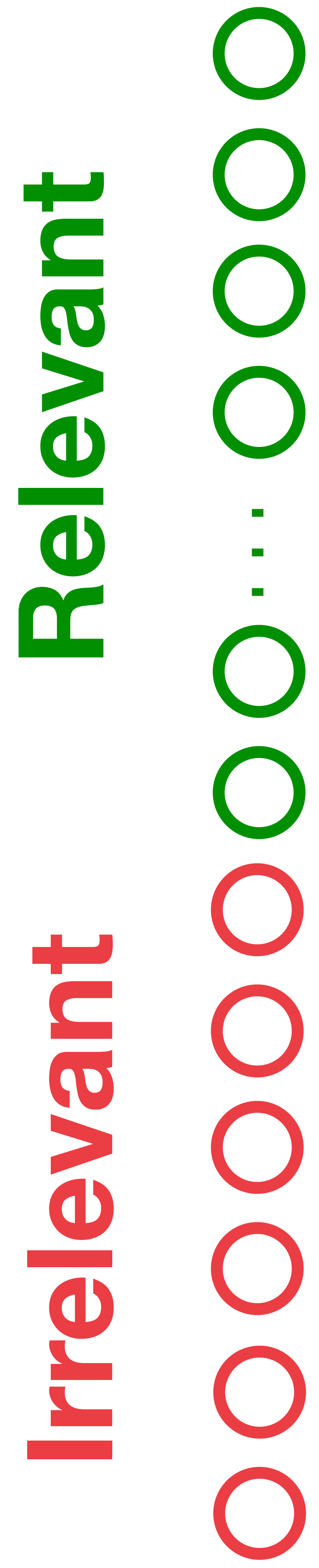
- Easy-to-hard training [Bengio et al (2009]
- Anti-curriculum [Zhang et al (2019); Hacohen & Weinshall (2019)]
- No effect of CL in vision benchmarks [Wu et al (2020)]
- Convincing results for LLMs and RL [Brown et al (2020); Narvekar et al (2020)]

ML (theory):

- Single-update advantage of easy samples [Weinshall et al (2020)]
- Speed benefit but limited performance upgrade in convex problems [Saglietti et al (2021); Lee et al. (2024)]
- Computational benefit in parity machines [Abbe et al. (2023); Cornacchia et al (2023)]
- Asymptotic benefit in non-convex models [Mannelli et al (2024)]

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

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

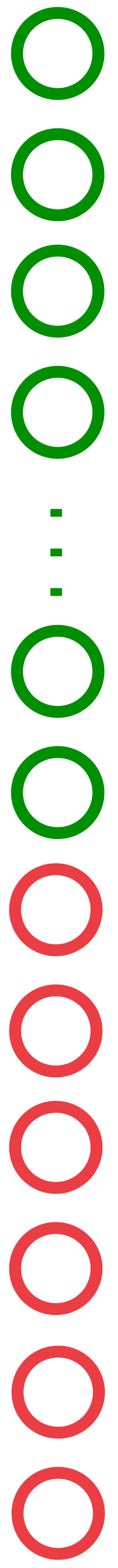
Variance Δ

A teacher-student model of curriculum learning

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

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

Relevant



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

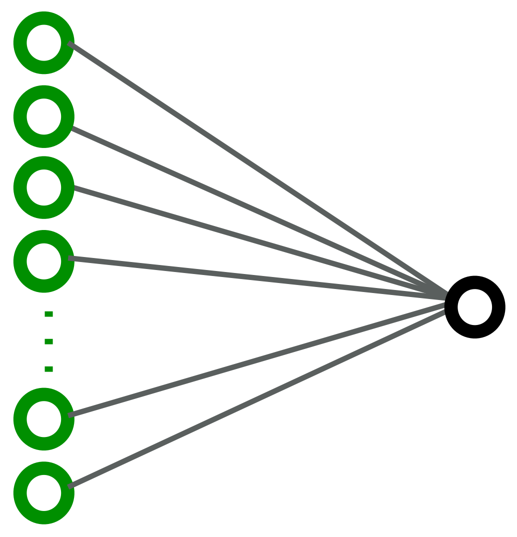
Unit variance

Irrelevant

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

Variance Δ

Teacher

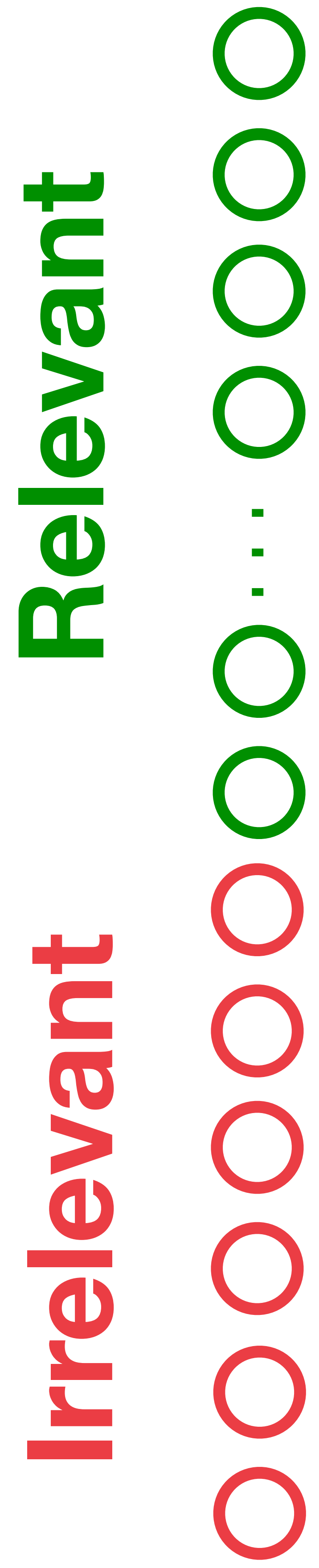


$y = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x}_r)$

A teacher-student model of curriculum learning

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

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



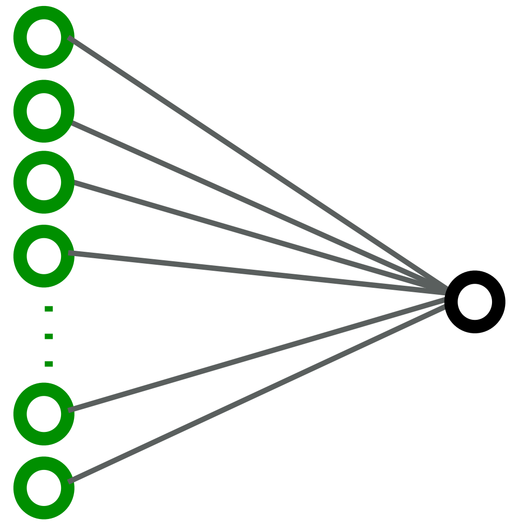
$\mathbf{x}_r \in \mathbb{R}^{\rho N}$

Unit variance

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

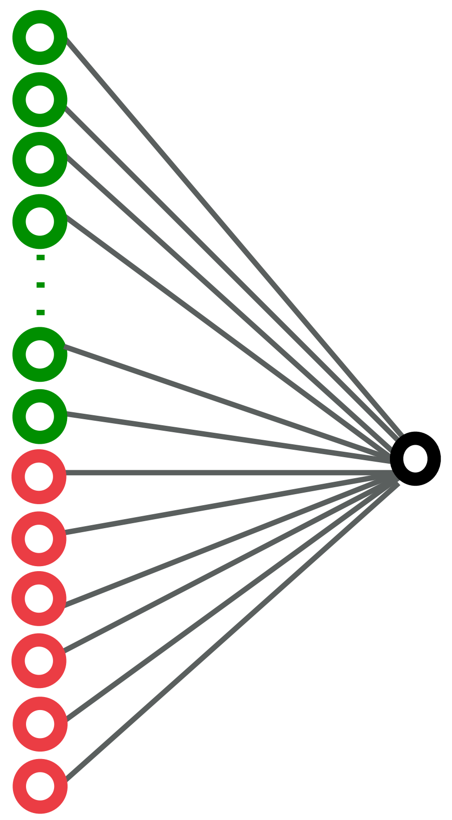
Variance Δ

Teacher



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

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

[Saglietti, et al (*NeurIPS* 2022)]

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[‡,§]

[Saglietti, et al (*NeurIPS* 2022)]

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

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

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

Online learning: [Biehl & Schwarze (1995); Saad & Solla (1995); …]

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

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

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

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:

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

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

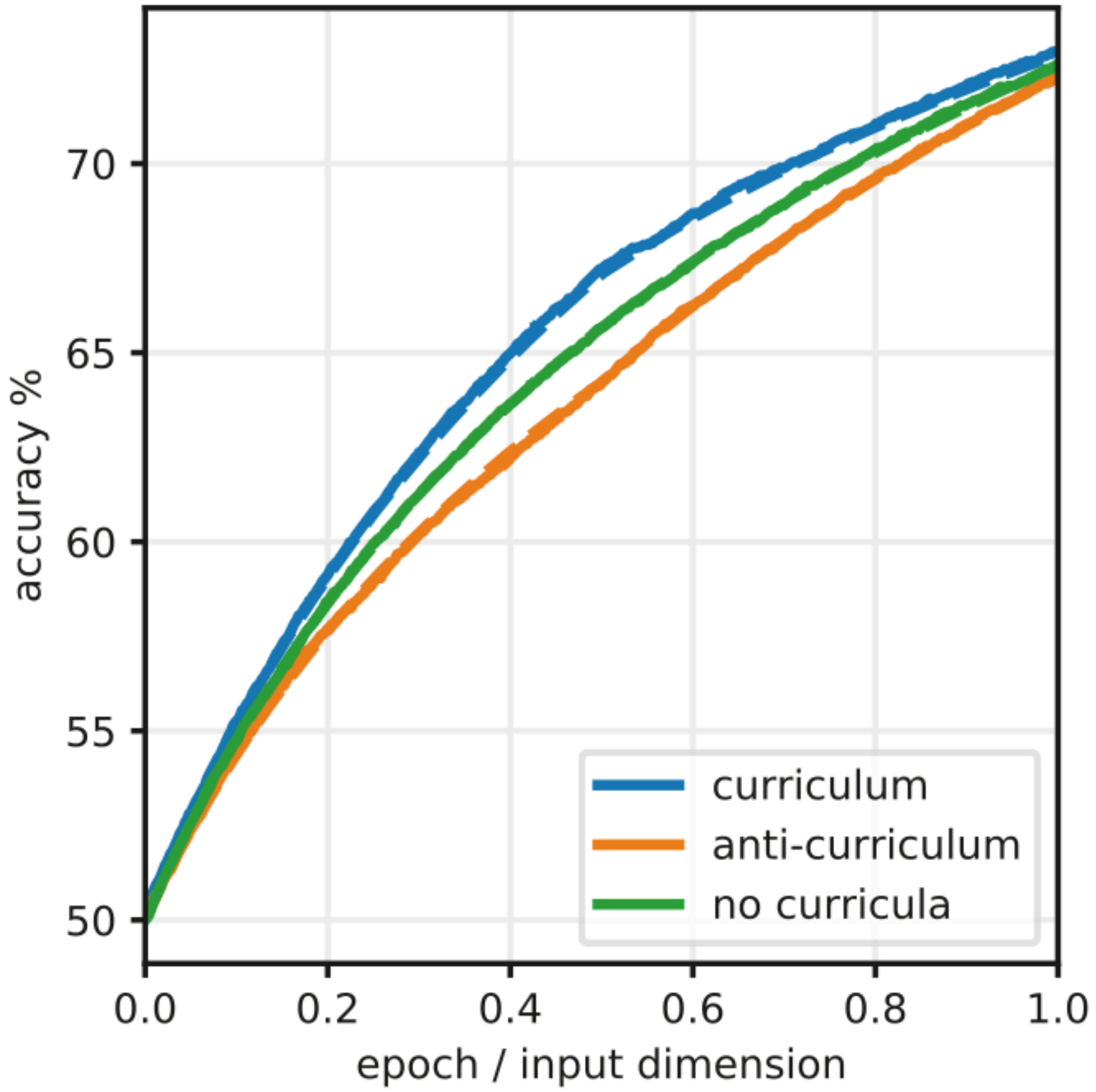
Online learning: [Biehl & Schwarze (1995); Saad & Solla (1995); …]

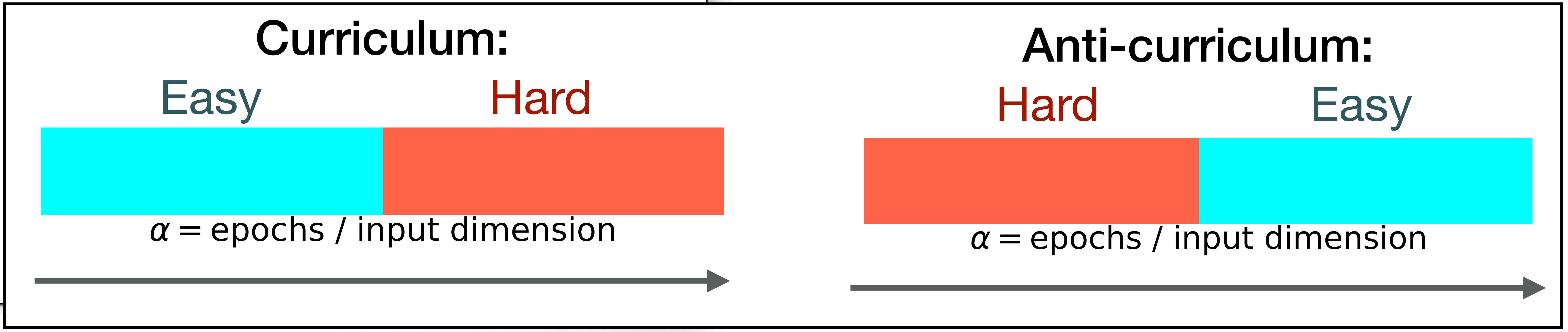
$$Q_r \leftarrow f_{Q_r}(Q_r, Q_i, R, T)$$

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

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

[Saglietti, et al (*NeurIPS* 2022)]



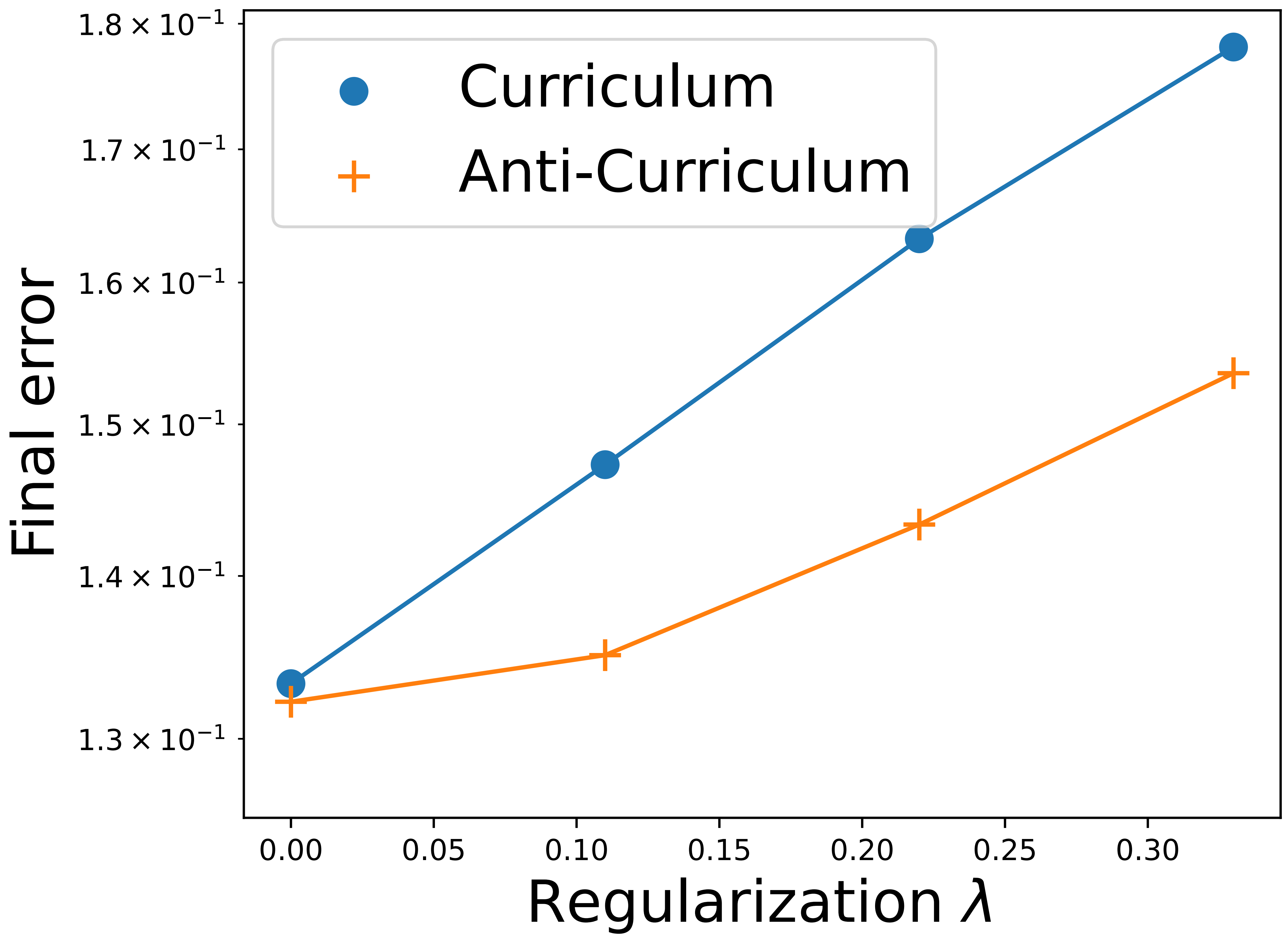


Optimal curriculum protocol

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

- Easy-to-hard curriculum is often *suboptimal*.

$\rho = 0.55, \eta = 2.58$



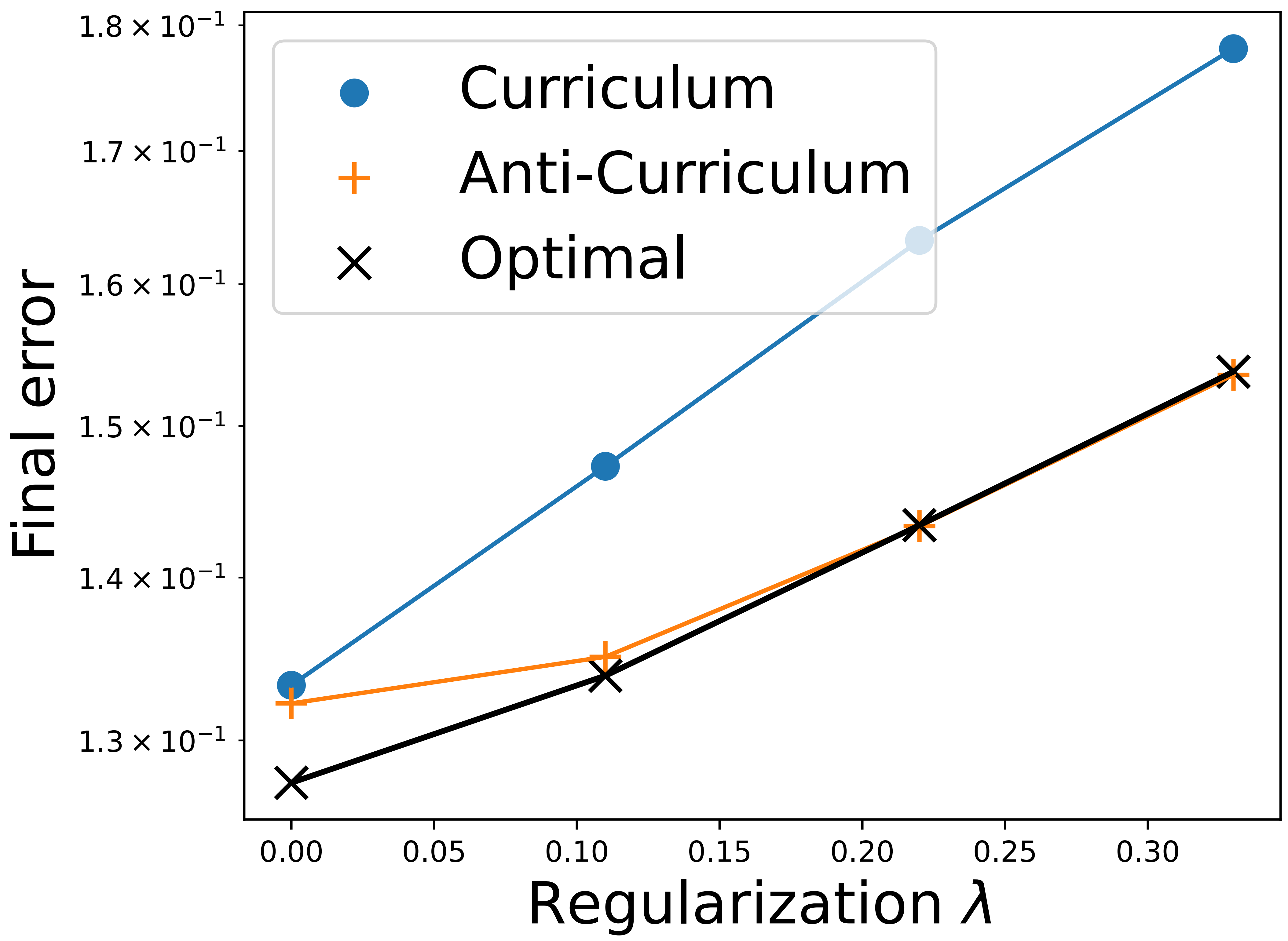
[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

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

- Easy-to-hard curriculum is often *suboptimal*.

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



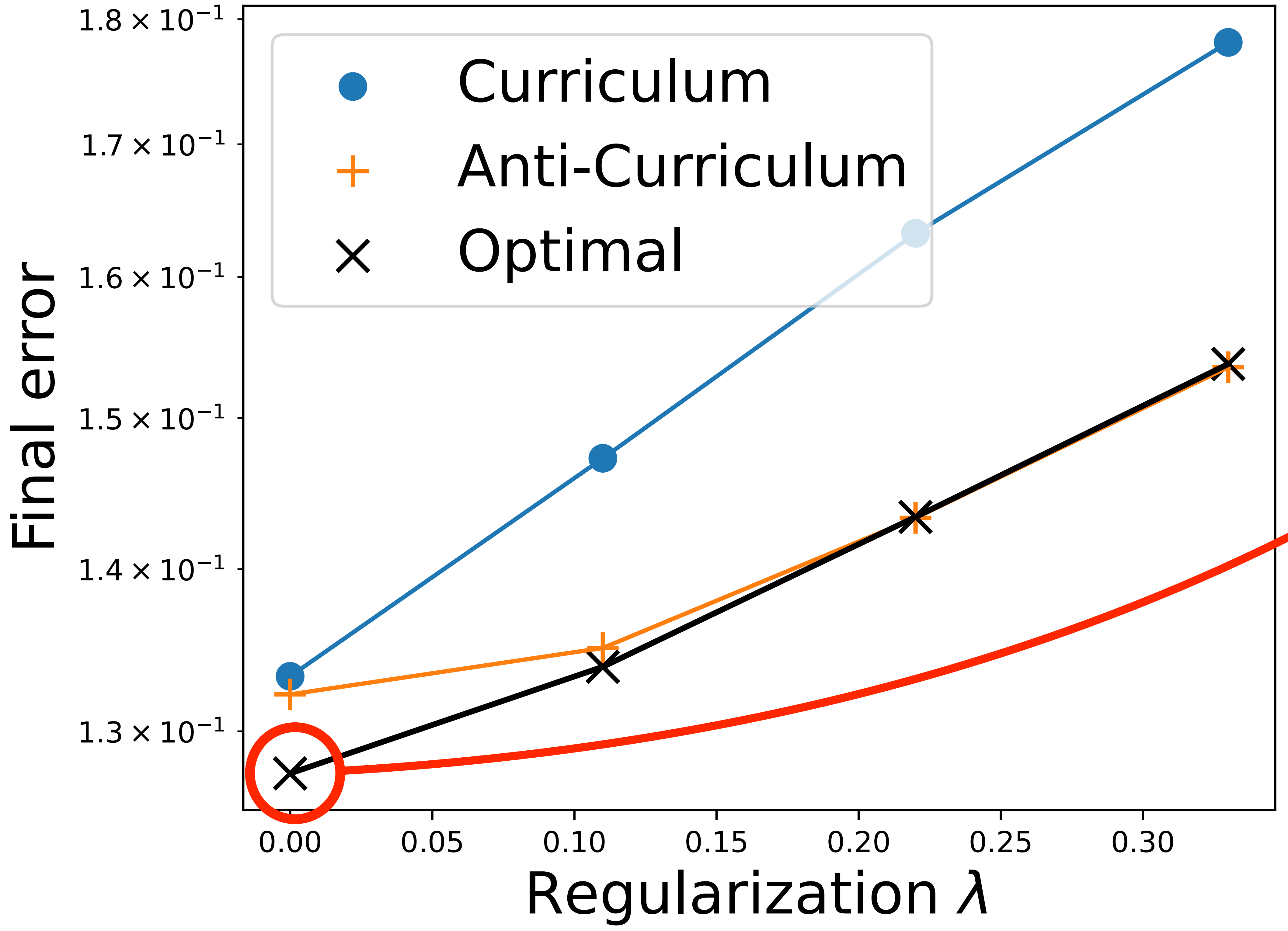
[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

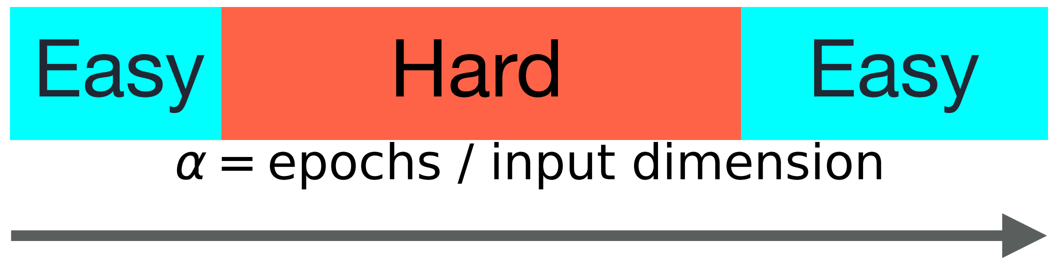
Control: $\mathbf{u} = \mathbf{\Delta}$

- Easy-to-hard curriculum is often *suboptimal*.

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



Non-monotonic curriculum is optimal:



[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

- Easy-to-hard curriculum is often *suboptimal*.

$\rho = 0.55, \eta = 2.58$

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

Non-monotonic curriculum is optimal:

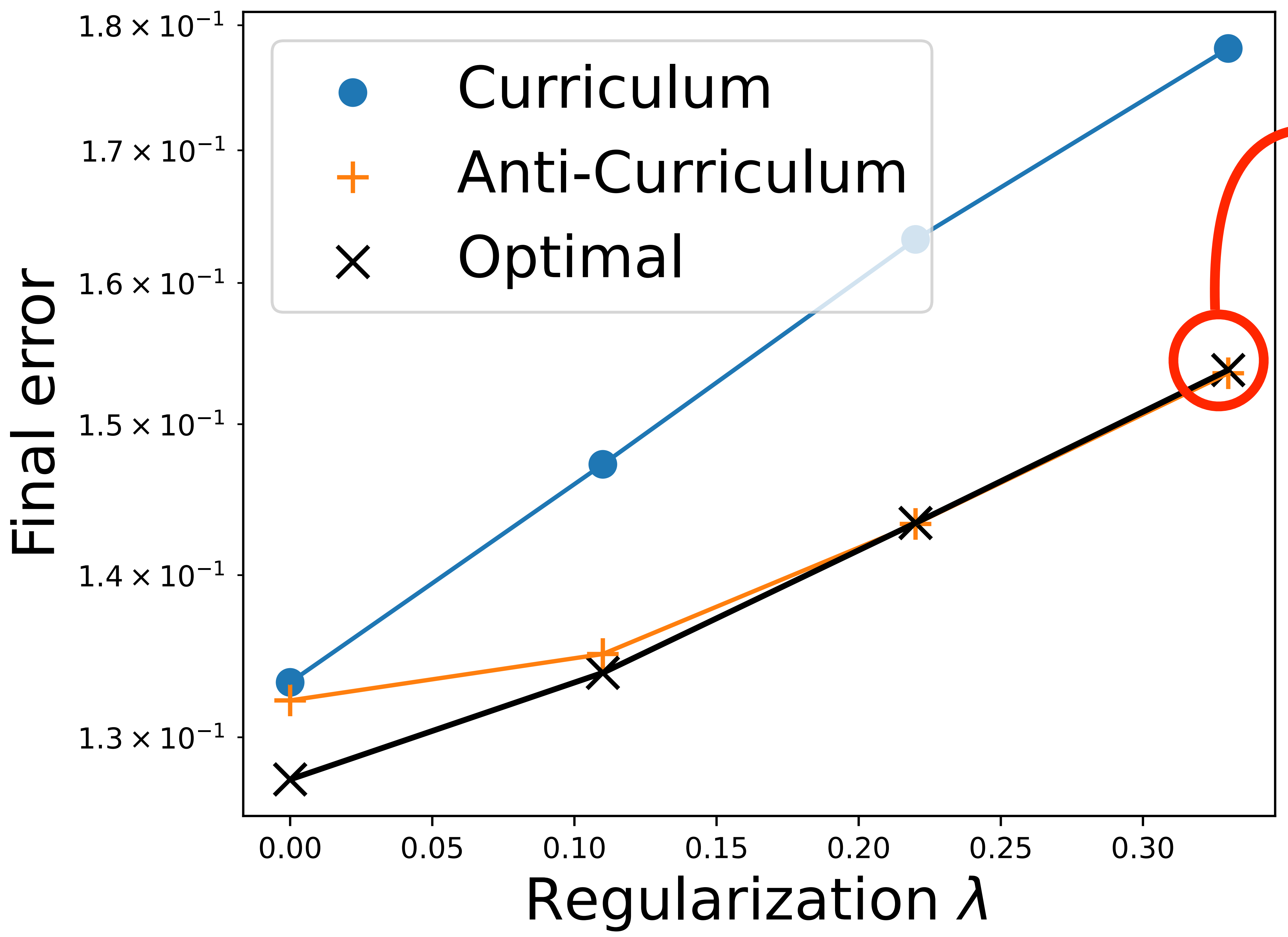
[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

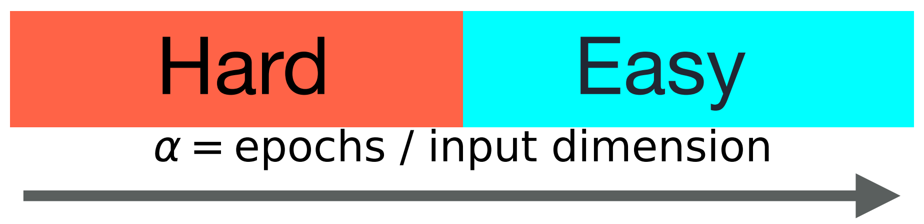
- Easy-to-hard curriculum is often *suboptimal*.

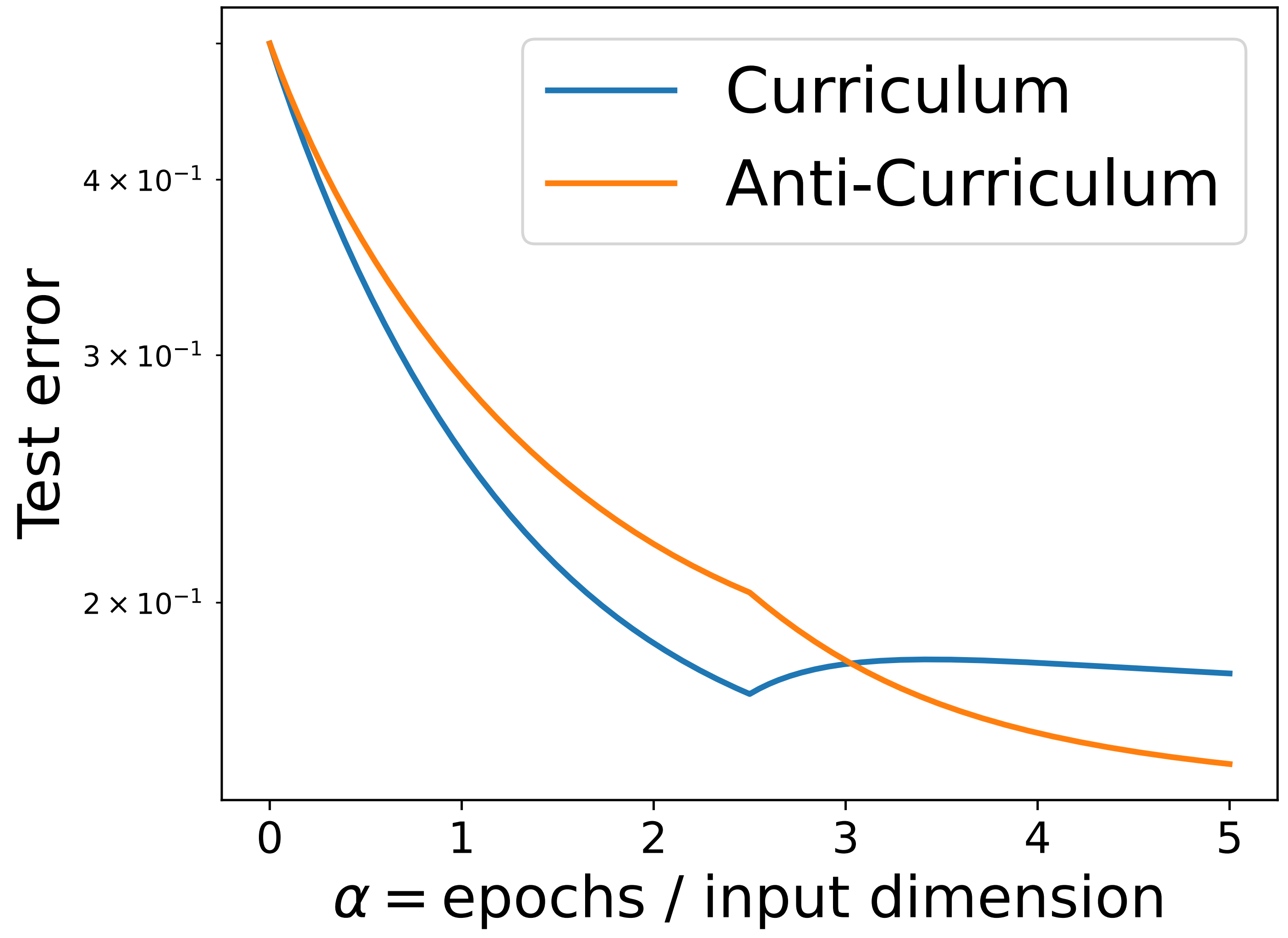
$\rho = 0.55,\ \eta = 2.58$

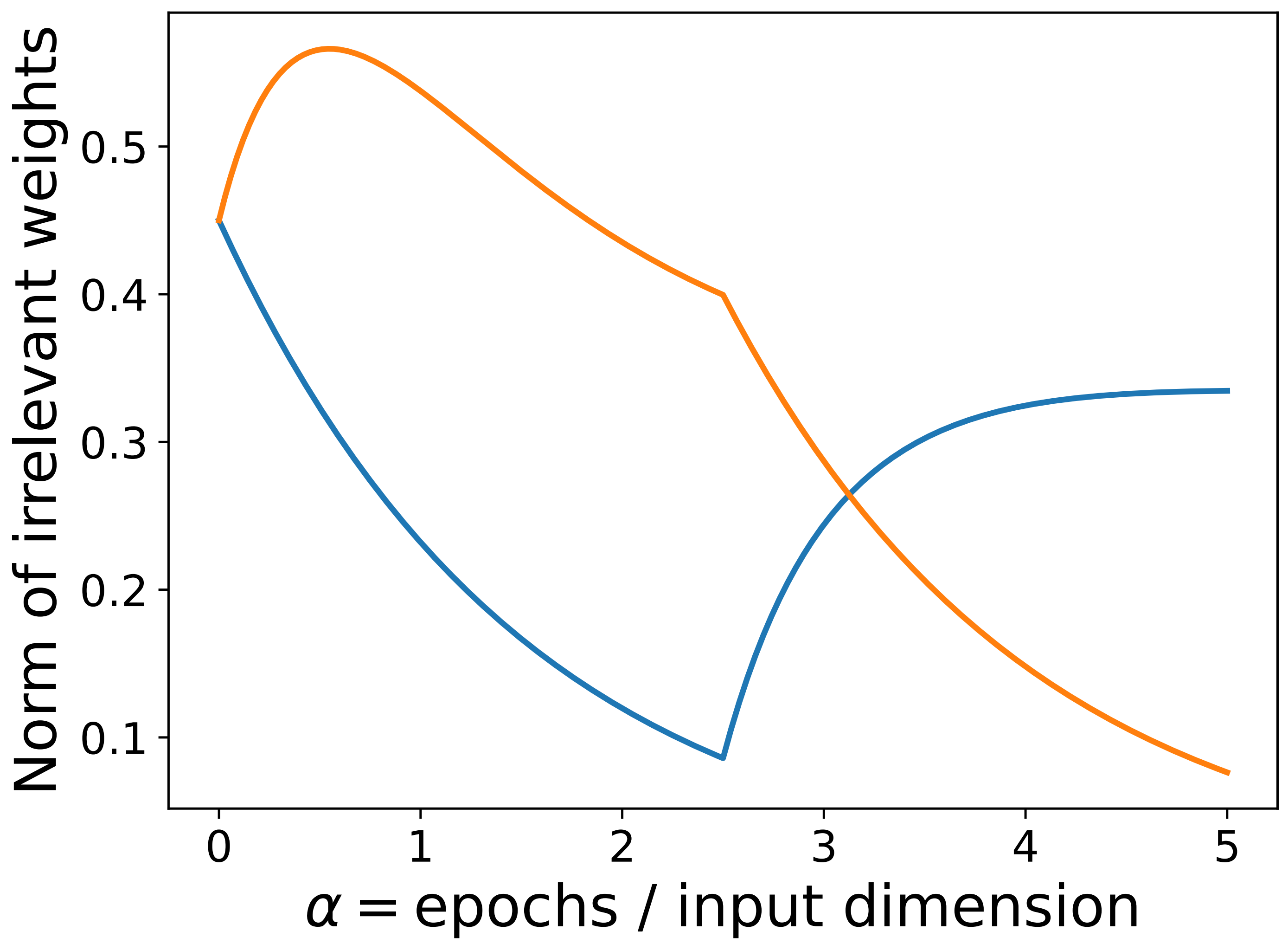
Control: $\mathbf{u} = \boldsymbol{\Delta}$

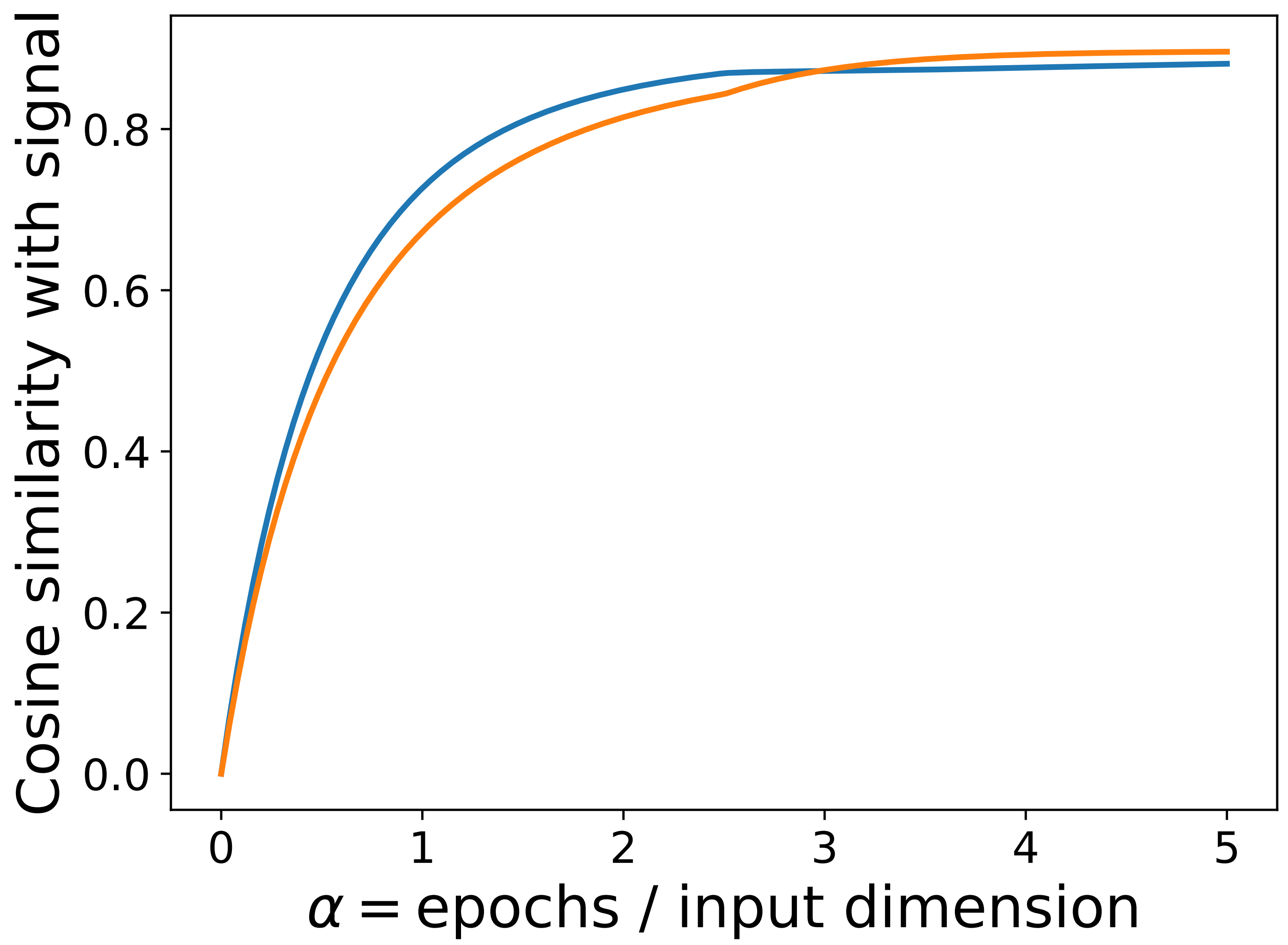


Anti-curriculum
is optimal:





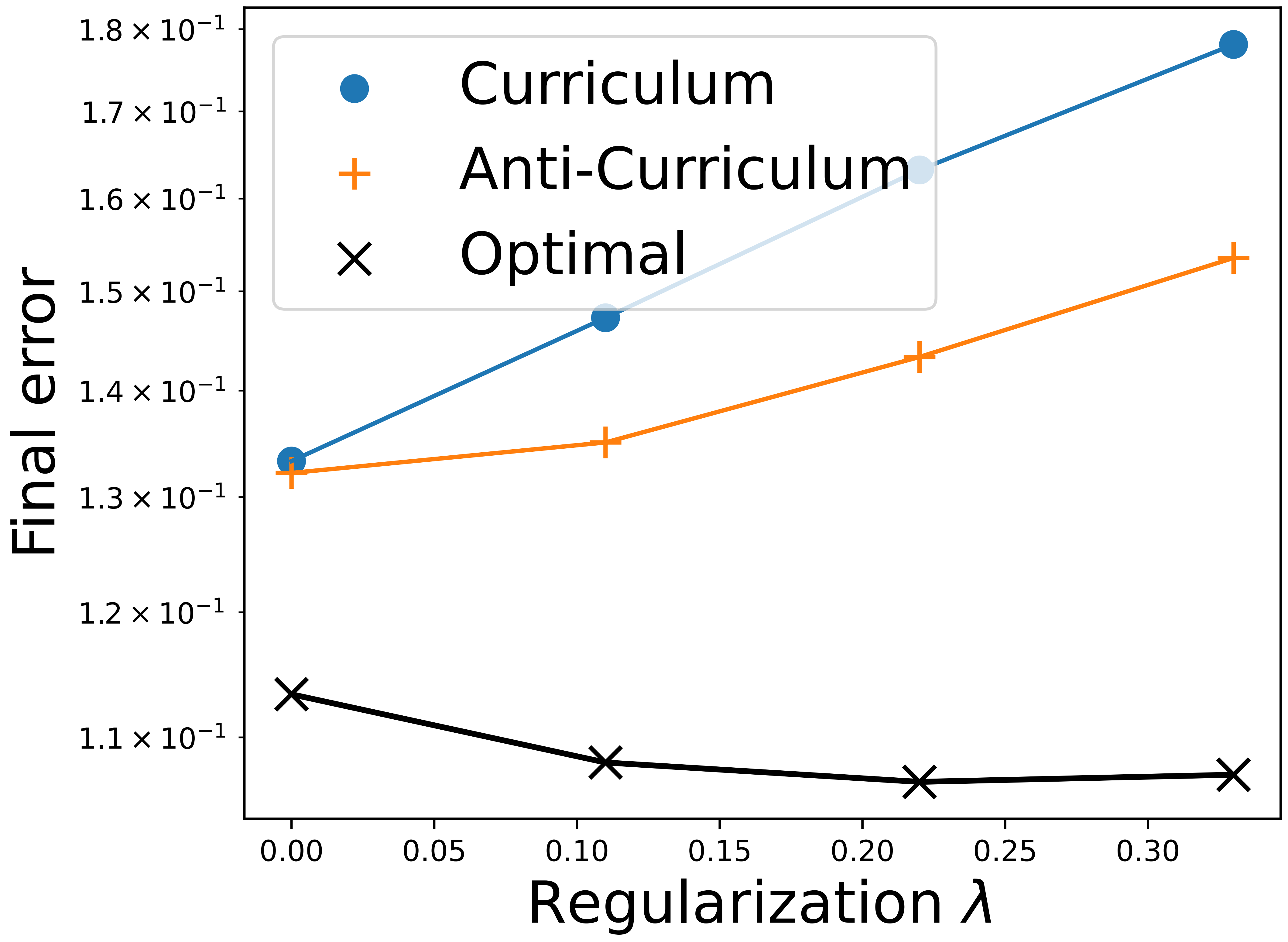




[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

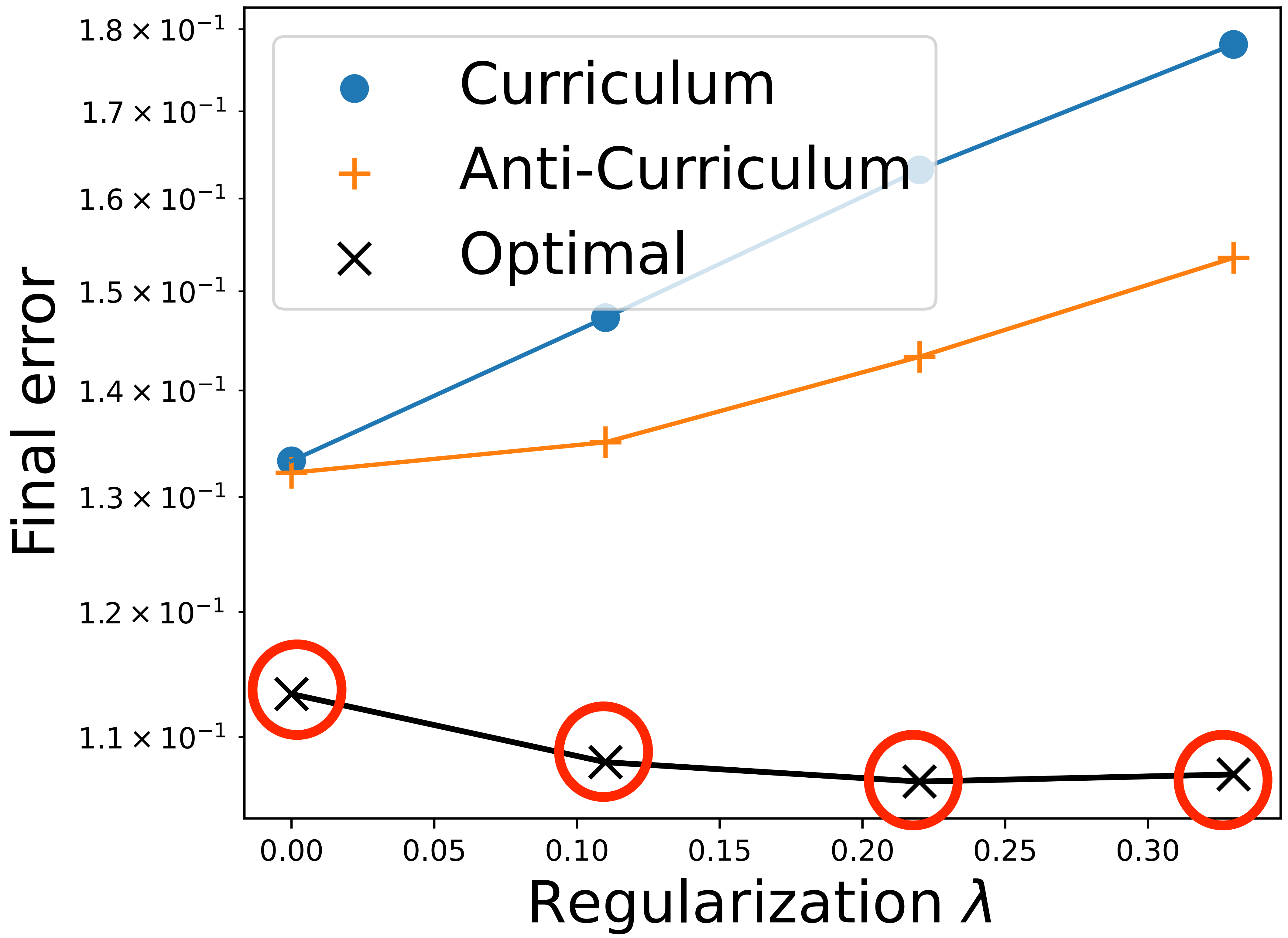
- Easy-to-hard curriculum becomes *optimal* **if we also optimize over the learning rate schedule.**



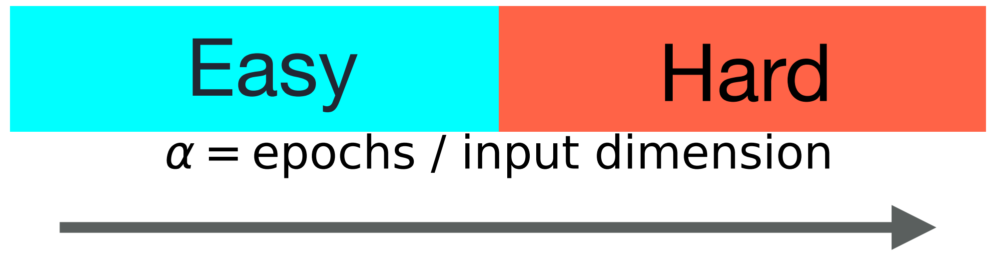
[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

- Easy-to-hard curriculum becomes *optimal* **if we also optimize over the learning rate schedule.**



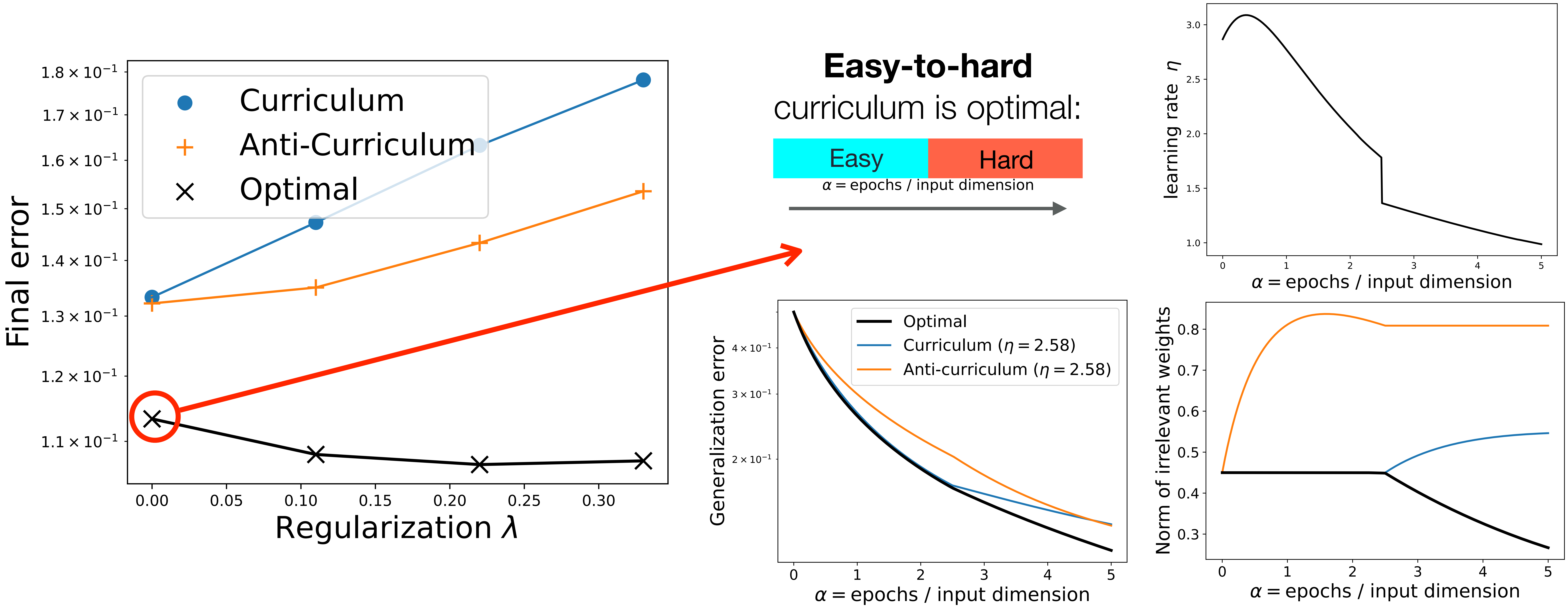
Easy-to-hard
curriculum is optimal:



[F. Mori, **FM**, *in preparation*]

Optimal curriculum protocol

- Easy-to-hard curriculum becomes *optimal* **if we also optimize over the learning rate schedule.**



[F. Mori, **FM**, *in preparation*]