Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes

Maxim Kodryan*, Ekaterina Lobacheva*, Maksim Nakhodnov*, Dmitry Vetrov

Introduction

Normalization (Batch Norm, Layer Norm, …) ⇒ *scale invariance* (SI):

$$F(c\theta) = F(\theta),\ \forall c > 0$$

Introduction

Normalization (Batch Norm, Layer Norm, …) ⇒ *scale invariance* (SI):

$$F(c\theta) = F(\theta), \forall c > 0$$

SI model is effectively defined on the **sphere**!

Introduction

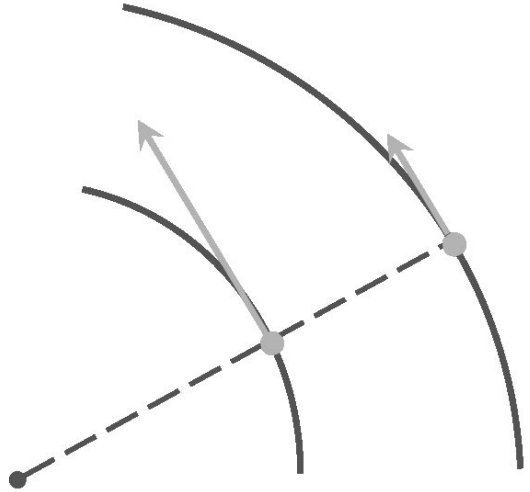
Normalization (Batch Norm, Layer Norm, …) ⇒ *scale invariance* (SI):

$$F(c\theta) = F(\theta),\ \forall c > 0$$

SI model is effectively defined on the **sphere**!

SGD optimization with a **fixed** LR ⇒ **varying** *effective learning rate* (ELR):

$\mathrm{ELR} = \mathrm{LR}/||\theta||^2$ — learning rate on the *unit sphere*.



Introduction

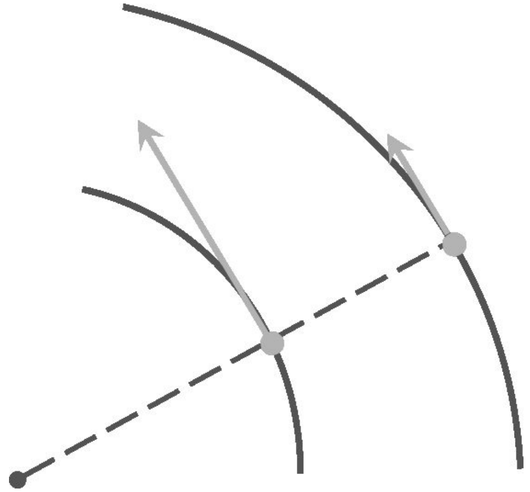
Normalization (Batch Norm, Layer Norm, …) ⇒ *scale invariance* (SI):

$$F(c\theta) = F(\theta),\ \forall c > 0$$

SI model is effectively defined on the **sphere**!

SGD optimization with a **fixed** LR ⇒ **varying** *effective learning rate* (ELR):

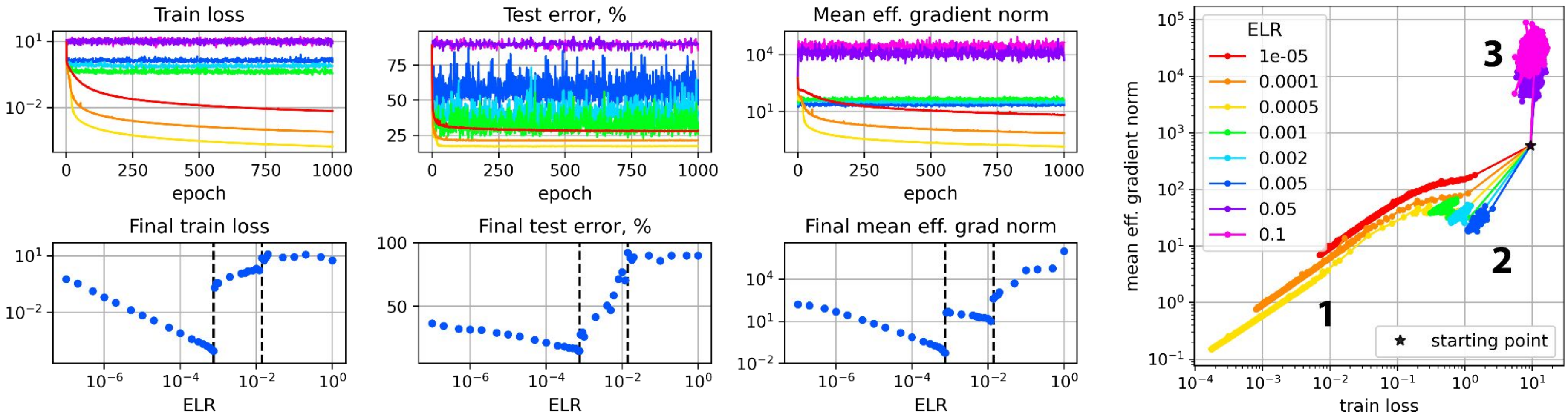
$\mathrm{ELR} = \mathrm{LR}/||\theta||^2$ — learning rate on the *unit sphere*.



Let's optimize SI models directly on the sphere!
(with a <u>fixed</u> effective learning rate)

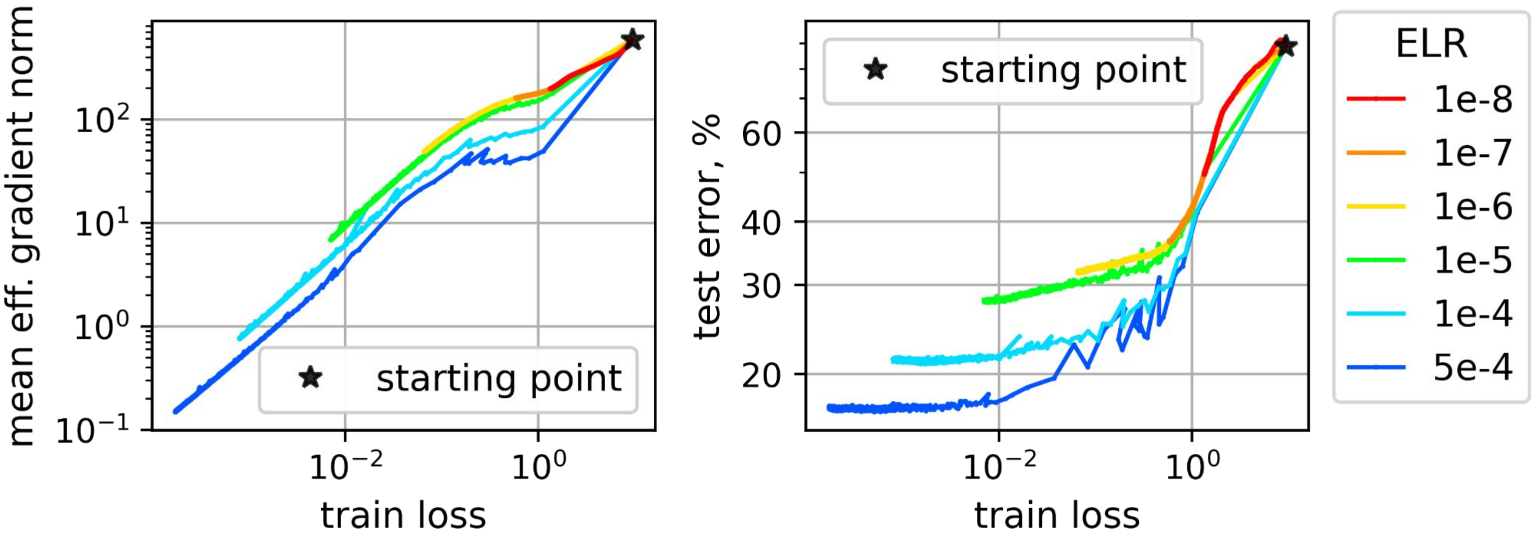
Three regimes of optimization on the sphere

1 Convergence **2** Chaotic equilibrium **3** Divergence



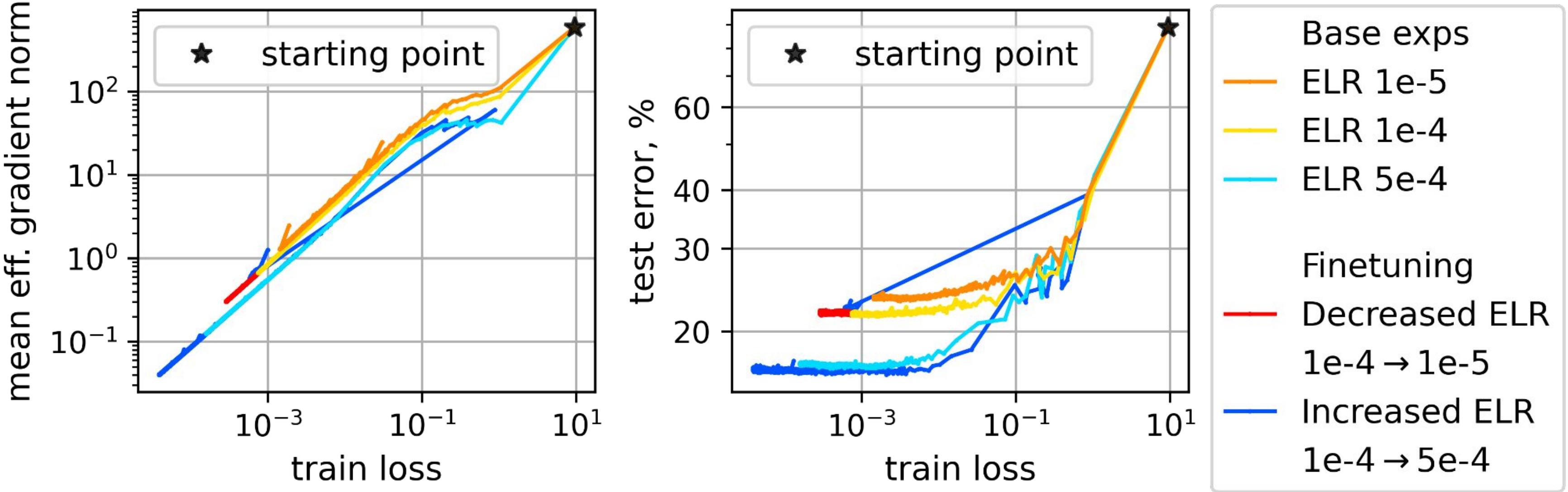
Regime 1: convergence

- Small ELR ⇒ ***convergence***
- Different optima depending on the ELR: **higher** ELR = **better** final solution (in terms of sharpness/generalization)



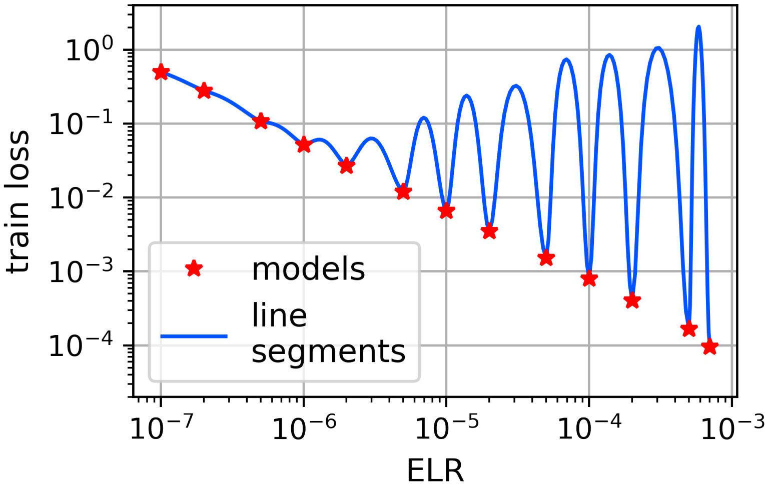
Regime 1: convergence

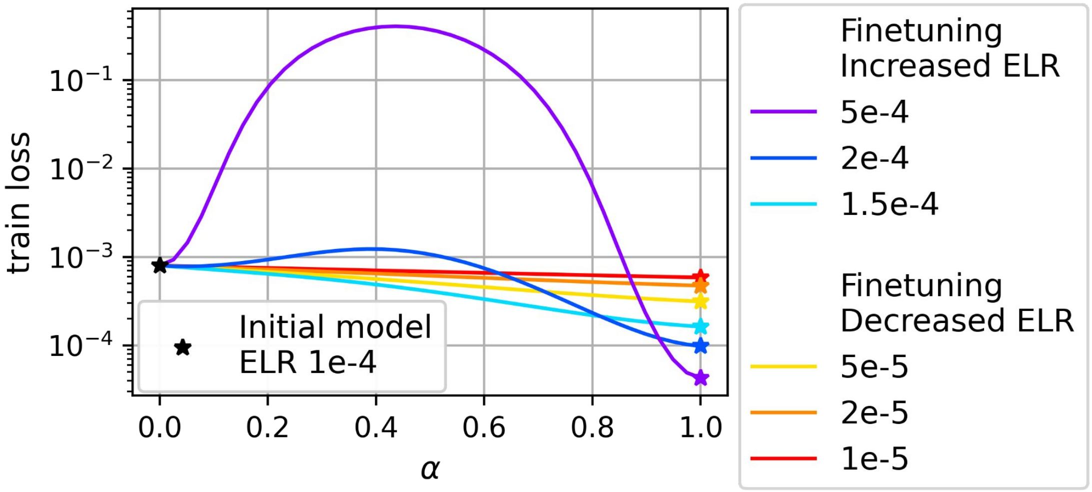
- Fine-tuning with **lower** ELR ⇒ the **same** trajectory
- Fine-tuning with **higher** ELR ⇒ jumping out into a **new wider** basin



Regime 1: convergence

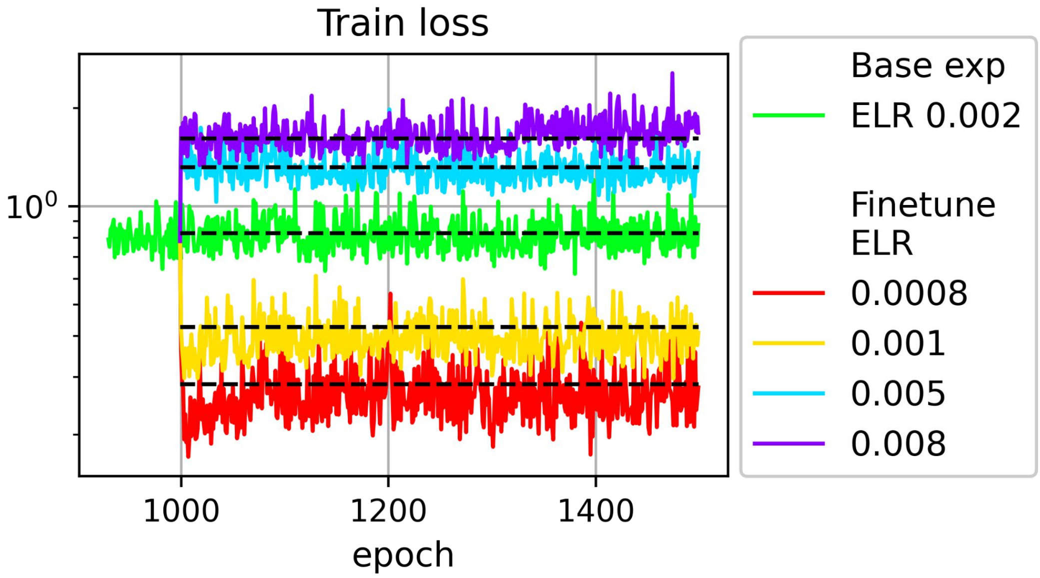
- Optima from **different** ELRs are **not** linearly connected (LC)
- Pre-trained and fine-tuned (FT) weights **are** LC for **low** FT ELR and **not** for **high** FT ELR





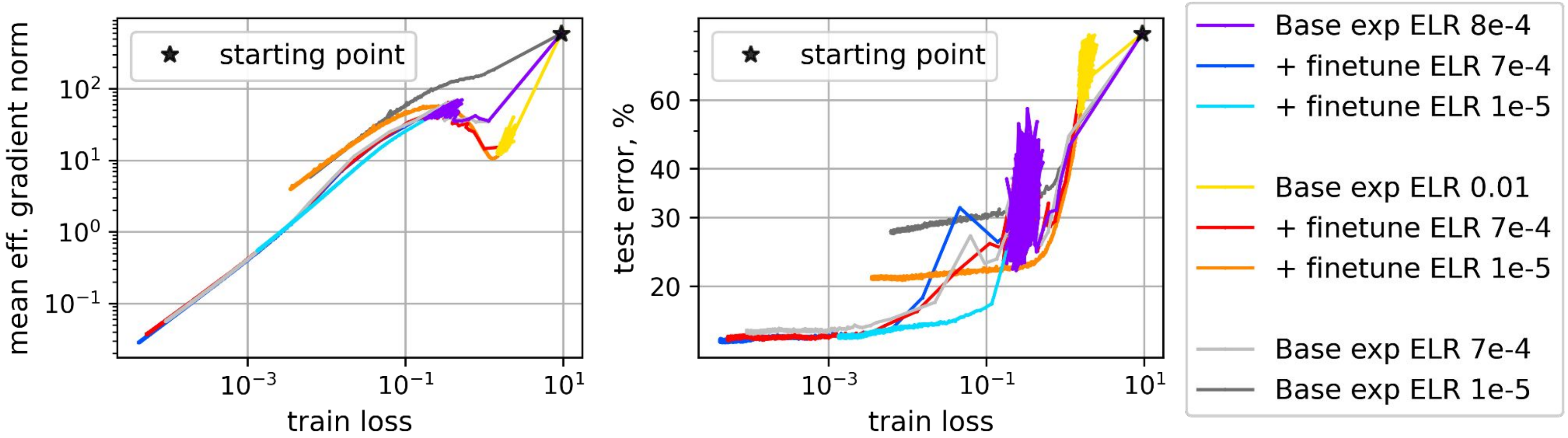
Regime 2: chaotic equilibrium

- Medium ELR ⇒ ***chaotic equilibrium***: loss noisily stabilizes at some *level*
- Changing the **ELR** ⇒ changing the **level**
- Optimization is “hopping along the walls” of the loss landscape



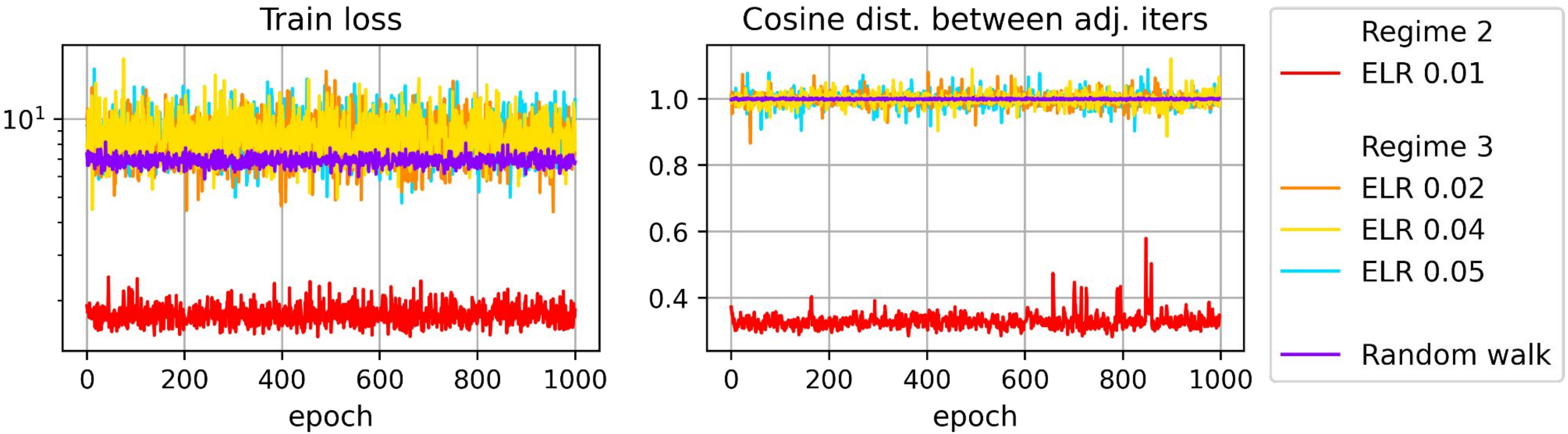
Regime 2: chaotic equilibrium

- Fine tuning with regime-1 ELRs from **low** regime-2 ELRs ⇒ convergence to the **widest** optimum
- Fine tuning with regime-1 ELRs from **high** regime-2 ELRs ⇒ convergence to **various** optima



Regime 3: divergence

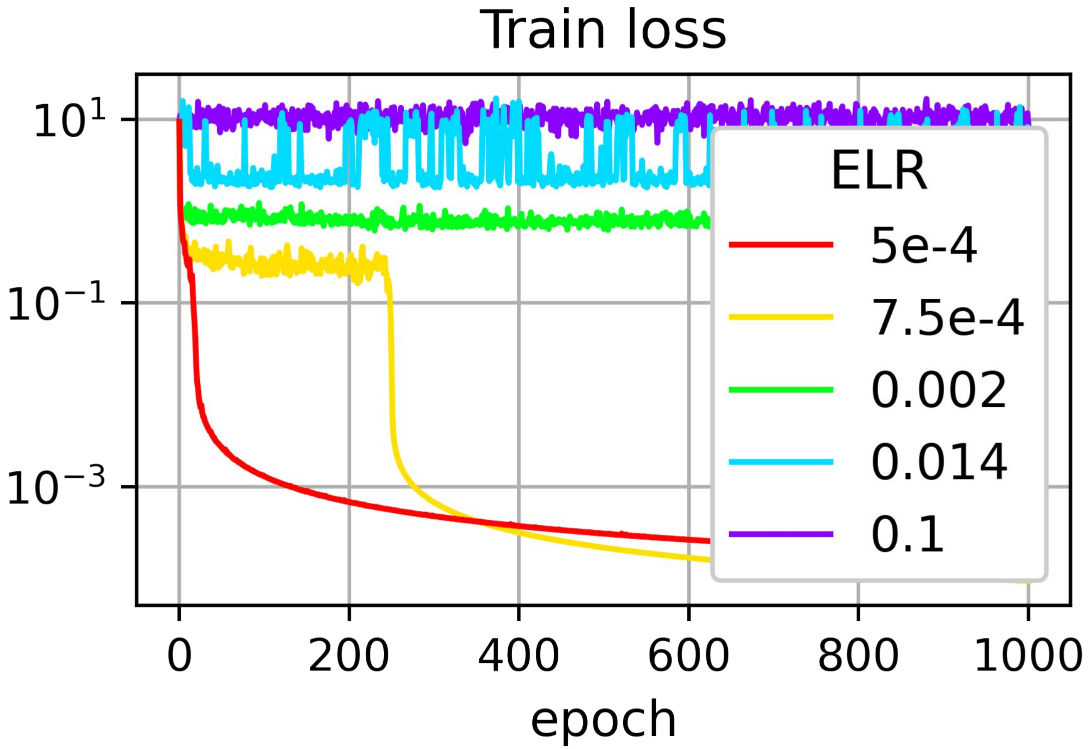
- High ELR ⇒ ***divergence***: near random guess behavior (close to random walk)



Transitions between regimes

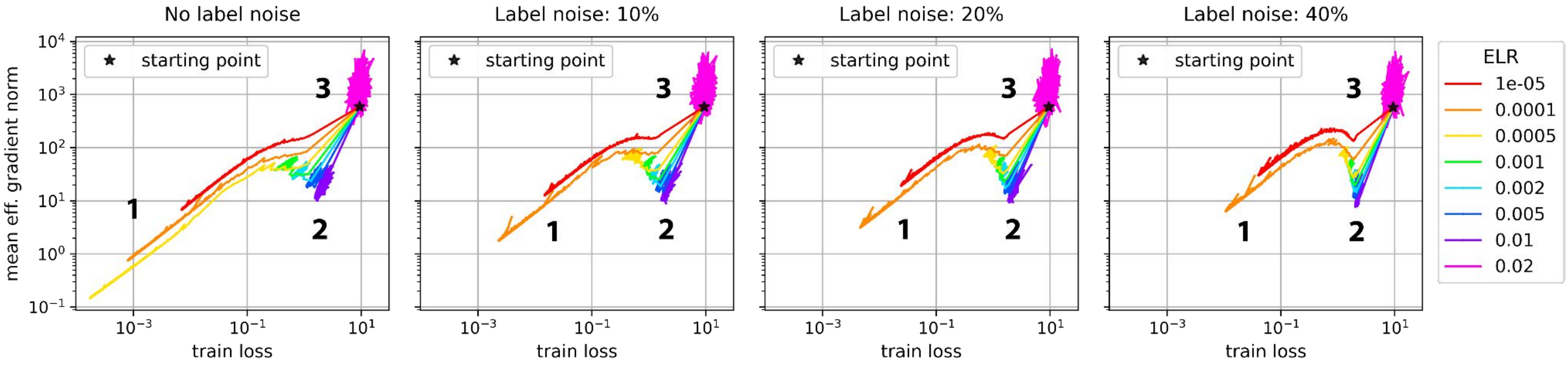
Typical regime ELRs:
- Regime 1
- Regime 2
- Regime 3

Transition ELRs:
- Regime 1/2
- Regime 2/3

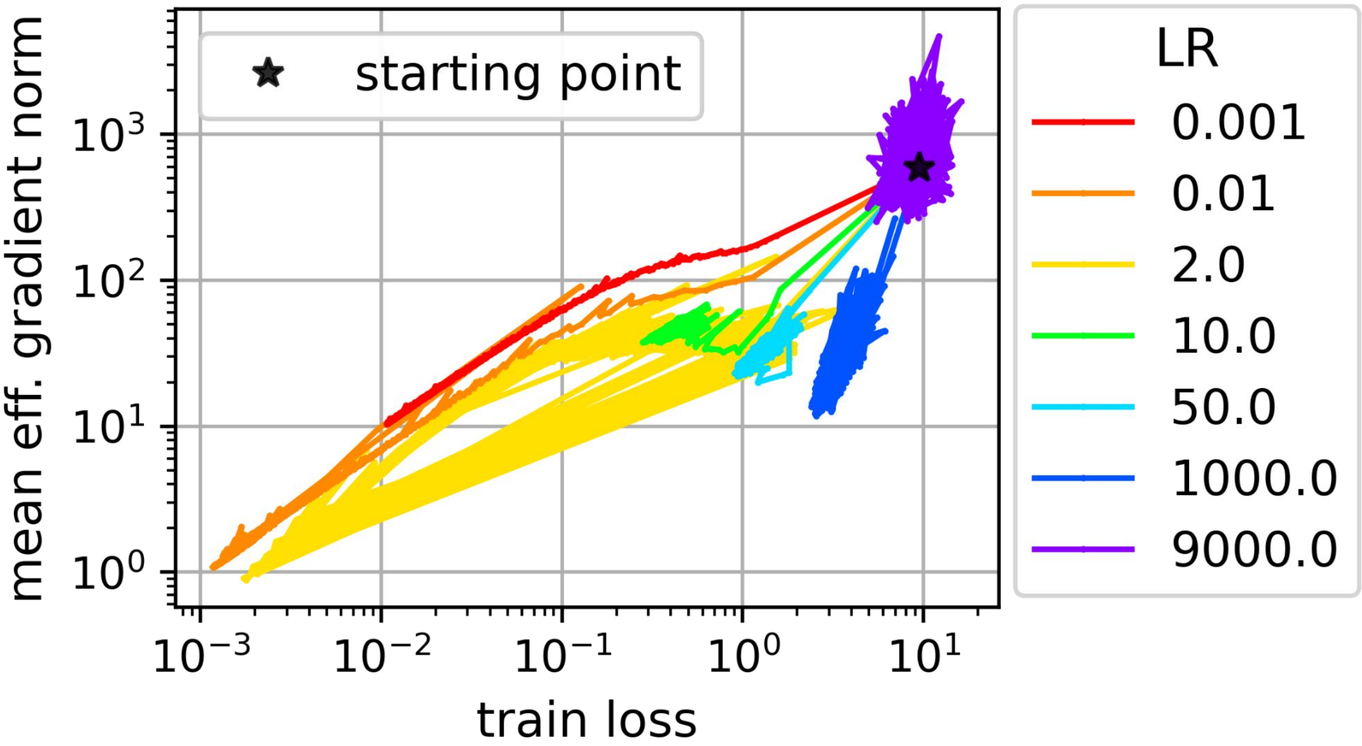


Transition between regimes 1/2 and DD

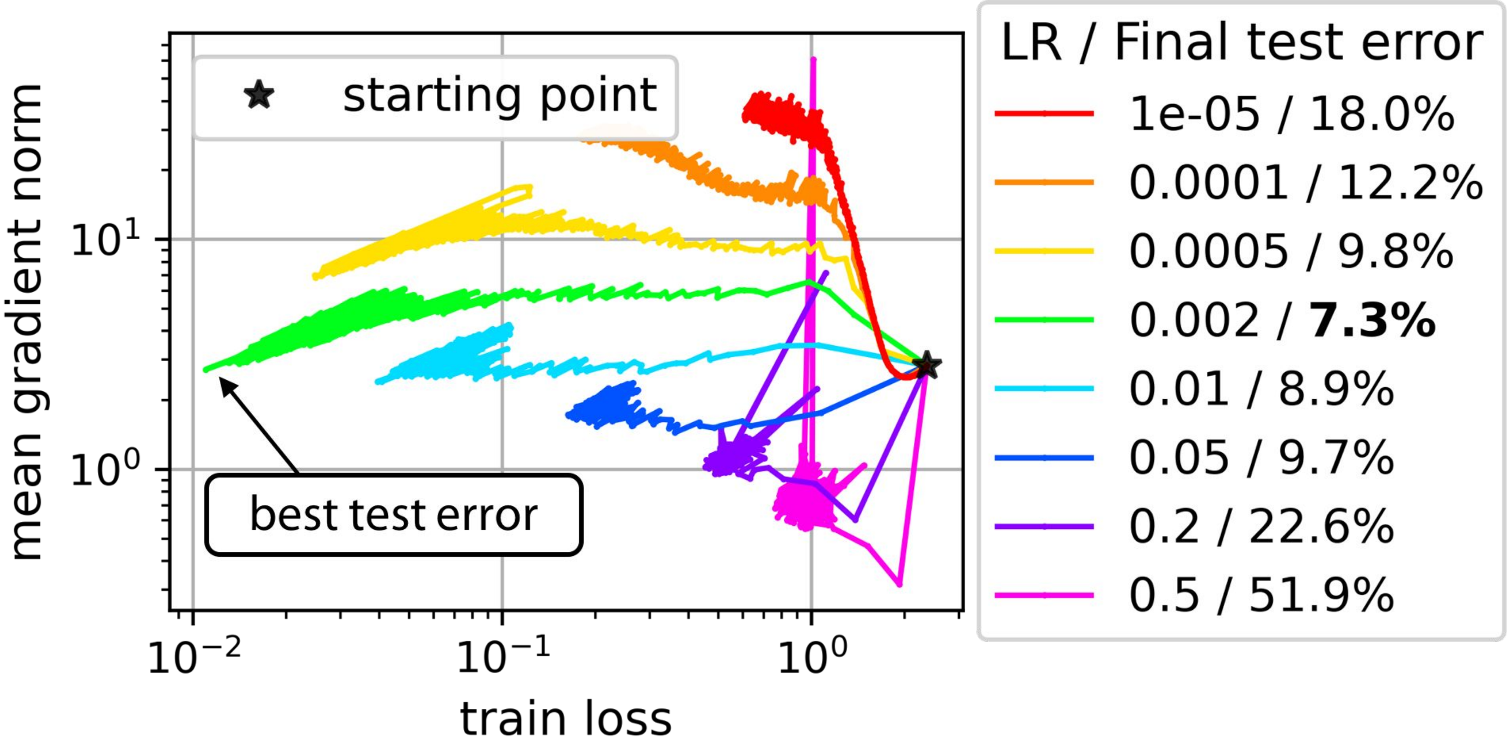
Double Descent peak ⇔ high-sharpness zone ⇔ transition between R1 and R2



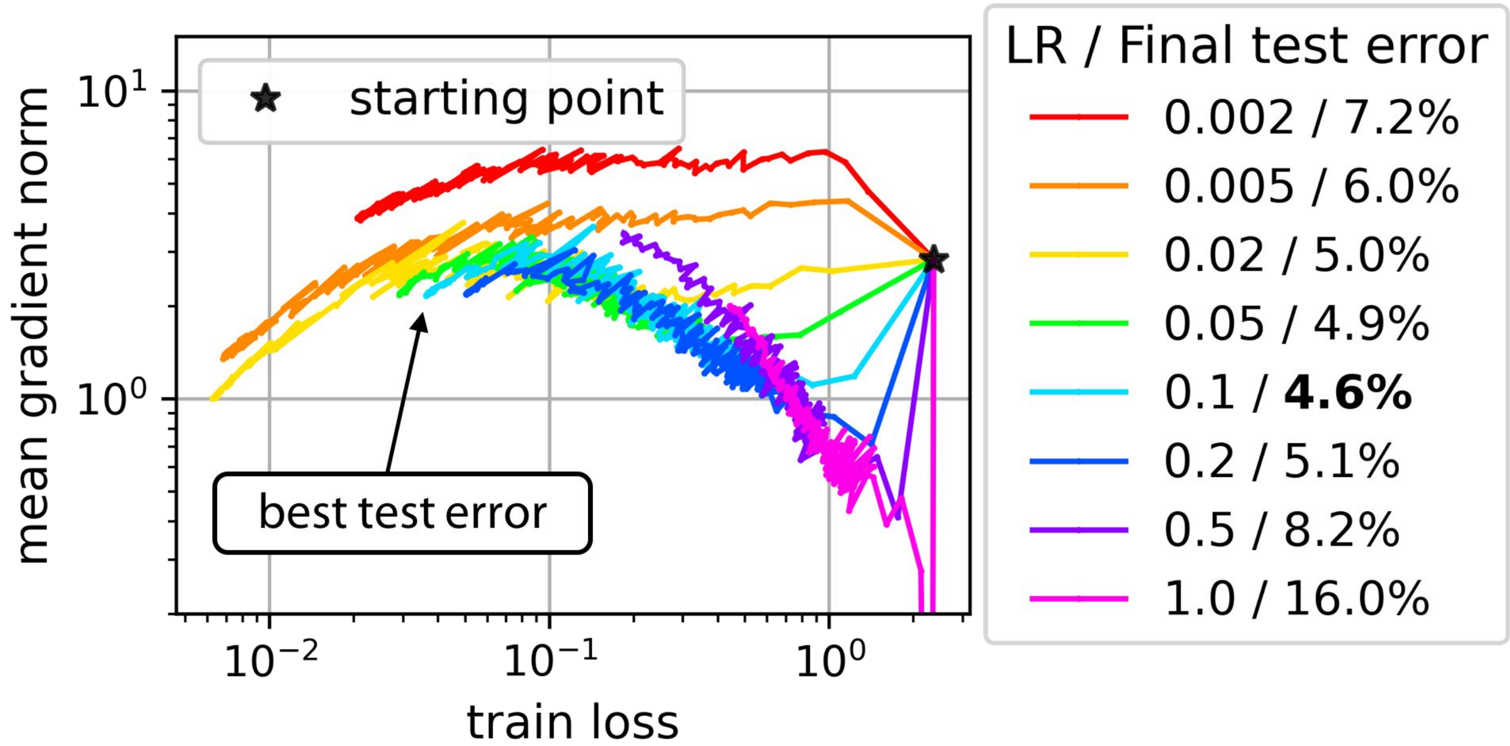
Regimes when training SI model in the whole space



Regimes in standard training: fixed LR



Regimes in standard training: cosine LR schedule



Theoretical analysis

$$F(\theta_1, \ldots, c\theta_i, \ldots, \theta_n) = F(\theta_1, \ldots, \theta_n), \, \forall c > 0, \, i = \overline{1, n}$$

Theoretical analysis

$$F(\theta_1, \ldots, c\theta_i, \ldots, \theta_n) = F(\theta_1, \ldots, \theta_n),\ \forall c > 0,\ i = \overline{1, n}$$

$$F(\theta_1, \ldots, \theta_n) \to \min_{(\theta_1, \ldots, \theta_n) \in \mathcal{S}(\rho)}$$

Theoretical analysis

$$F(\theta_1, \ldots, c\theta_i, \ldots, \theta_n) = F(\theta_1, \ldots, \theta_n),\ \forall c > 0,\ i = \overline{1, n}$$

$$F(\theta_1, \ldots, \theta_n) \to \min_{(\theta_1, \ldots, \theta_n) \in \mathcal{S}(\rho)}$$

$$\text{Fix: radius}\ \rho, \text{total LR}\ \eta \ \Rightarrow\ \text{total ELR}\ \tilde{\eta} \equiv \eta/\rho^2\ \text{is fixed}$$

Theoretical analysis

$$F(\theta_1, \ldots, c\theta_i, \ldots, \theta_n) = F(\theta_1, \ldots, \theta_n),\ \forall c > 0,\ i = \overline{1, n}$$

$$F(\theta_1, \ldots, \theta_n) \rightarrow \min_{(\theta_1, \ldots, \theta_n) \in \mathcal{S}(\rho)}$$

Fix: radius ρ, total LR η $\Rightarrow$ total ELR $\tilde{\eta} \equiv \eta/\rho^2$ is fixed

Define:

ELR for θ_i as $\tilde{\eta}_i \equiv \eta/||\theta_i||^2$

Eff. grad. norm for θ_i as $\tilde{g}_i \equiv ||\nabla_{\theta_i} F|| \cdot ||\theta_i||$

Theoretical analysis

$$\text{ELRs relation: } \sum_{i=1}^{n} 1/\tilde{\eta}_i = 1/\tilde{\eta}$$

Theoretical analysis

ELRs relation: $\sum_{i=1}^{n} 1/\tilde{\eta}_i = 1/\tilde{\eta}$

Eff. step sizes relation: $(\tilde{g}\tilde{\eta})^2 = \sum_{i=1}^{n} \omega_i (\tilde{g}_i \tilde{\eta}_i)^2$, $(\omega_1, \ldots, \omega_n) \in \Delta^{n-1}$

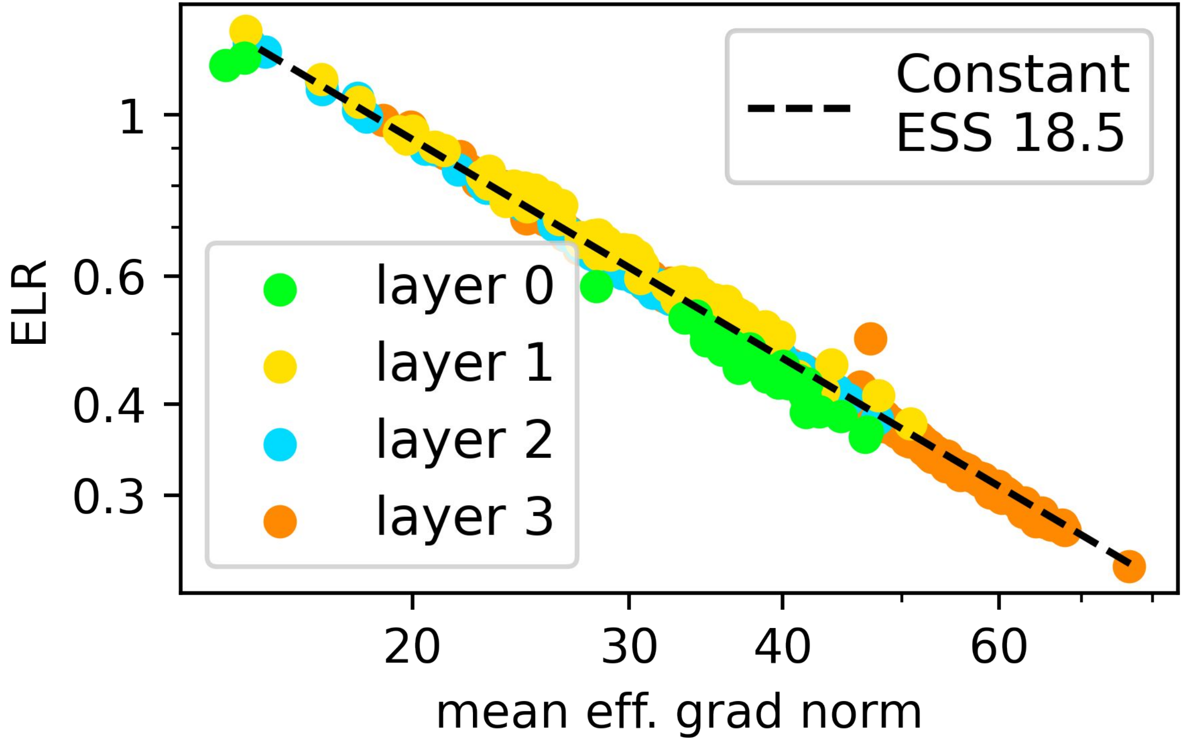
Theoretical analysis

ELRs relation: $\sum_{i=1}^{n} 1/\tilde{\eta}_i = 1/\tilde{\eta}$

Eff. step sizes relation: $(\tilde{g}\tilde{\eta})^2 = \sum_{i=1}^{n} \omega_i (\tilde{g}_i \tilde{\eta}_i)^2$, $(\omega_1, \ldots, \omega_n) \in \Delta^{n-1}$

ELRs dynamics: $\tilde{\eta}_i^{(t+1)} \leftarrow \tilde{\eta}_i^{(t)} \dfrac{1 + \left(\tilde{g}^{(t)} \tilde{\eta}\right)^2}{1 + \left(\tilde{g}_i^{(t)} \tilde{\eta}_i^{(t)}\right)^2}$

Regime 2: ESS alignment

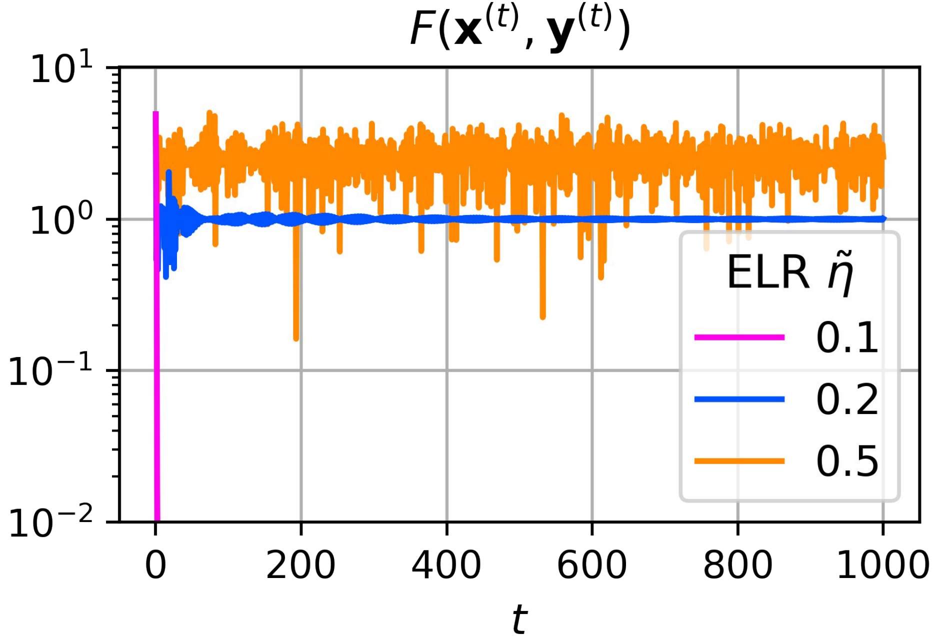


Toy example

$$F(x_1, y_1, \ldots, x_n, y_n) = \sum_{i=1}^{n} \alpha_i \frac{x_i^2}{x_i^2 + y_i^2}, \; \alpha_i > 0$$

$$\tilde{\eta} < 1/\sum_{i=1}^{n} \alpha_i \;\Rightarrow\; \text{Regime 1}$$

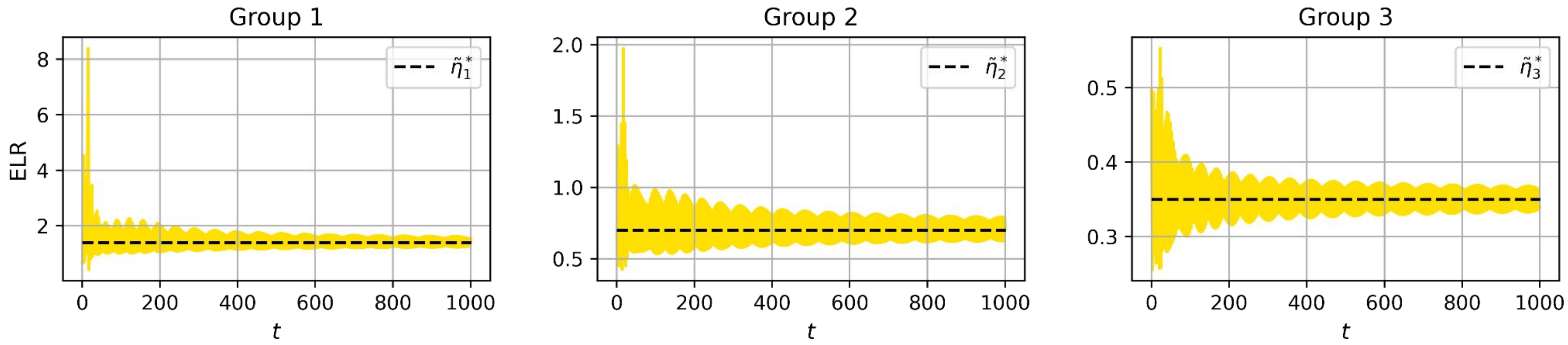
$$\tilde{\eta} > 1/\sum_{i=1}^{n} \alpha_i \;\Rightarrow\; \text{Regime 2/3}$$



Toy example

$$F(x_1, y_1, \ldots, x_n, y_n) = \sum_{i=1}^{n} \alpha_i \frac{x_i^2}{x_i^2 + y_i^2}, \ \alpha_i > 0$$

$$\text{Regime 2/3 with equilibrium ELRs } \tilde{\eta}_i^* \equiv \frac{\tilde{\eta} \sum_{j=1}^{n} \alpha_j}{\alpha_i}$$



Summing-up

- Scale-invariant models can be trained on the sphere in ***three regimes*** (depending on the ELR): *convergence, chaotic equilibrium, divergence*

Summing-up

- Scale-invariant models can be trained on the sphere in ***three regimes*** (depending on the ELR): *convergence, chaotic equilibrium, divergence*
- Each regime is specific:
  - **Regime 1:** convergence to different optima depending on *the ELR*
  - **Regime 2:** stabilization at some level; convergence to different optima (with a decreased ELR) depending on *that level*
  - **Regime 3:** near random guess behavior, divergence

Summing-up

- Scale-invariant models can be trained on the sphere in ***three regimes*** (depending on the ELR): *convergence, chaotic equilibrium, divergence*
- Each regime is specific:
  - **Regime 1:** convergence to different optima depending on *the ELR*
  - **Regime 2:** stabilization at some level; convergence to different optima (with a decreased ELR) depending on *that level*
  - **Regime 3:** near random guess behavior, divergence
- The regimes can be observed in standard NN training as well

Summing-up

- Scale-invariant models can be trained on the sphere in ***three regimes*** (depending on the ELR): *convergence, chaotic equilibrium, divergence*
- Each regime is specific:
    - **Regime 1:** convergence to different optima depending on *the ELR*
    - **Regime 2:** stabilization at some level; convergence to different optima (with a decreased ELR) depending on *that level*
    - **Regime 3:** near random guess behavior, divergence
- The regimes can be observed in standard NN training as well
- The regimes can be **analytically** derived for some functions

Summing-up

- Scale-invariant models can be trained on the sphere in ***three regimes*** (depending on the ELR): *convergence, chaotic equilibrium, divergence*
- Each regime is specific:
    - **Regime 1:** convergence to different optima depending on *the ELR*
    - **Regime 2:** stabilization at some level; convergence to different optima (with a decreased ELR) depending on *that level*
    - **Regime 3:** near random guess behavior, divergence
- The regimes can be observed in standard NN training as well
- The regimes can be **analytically** derived for some functions
- **Other results:** LC of regime-2 checkpoints, dependence on model width/data complexity, additional experiments, etc.