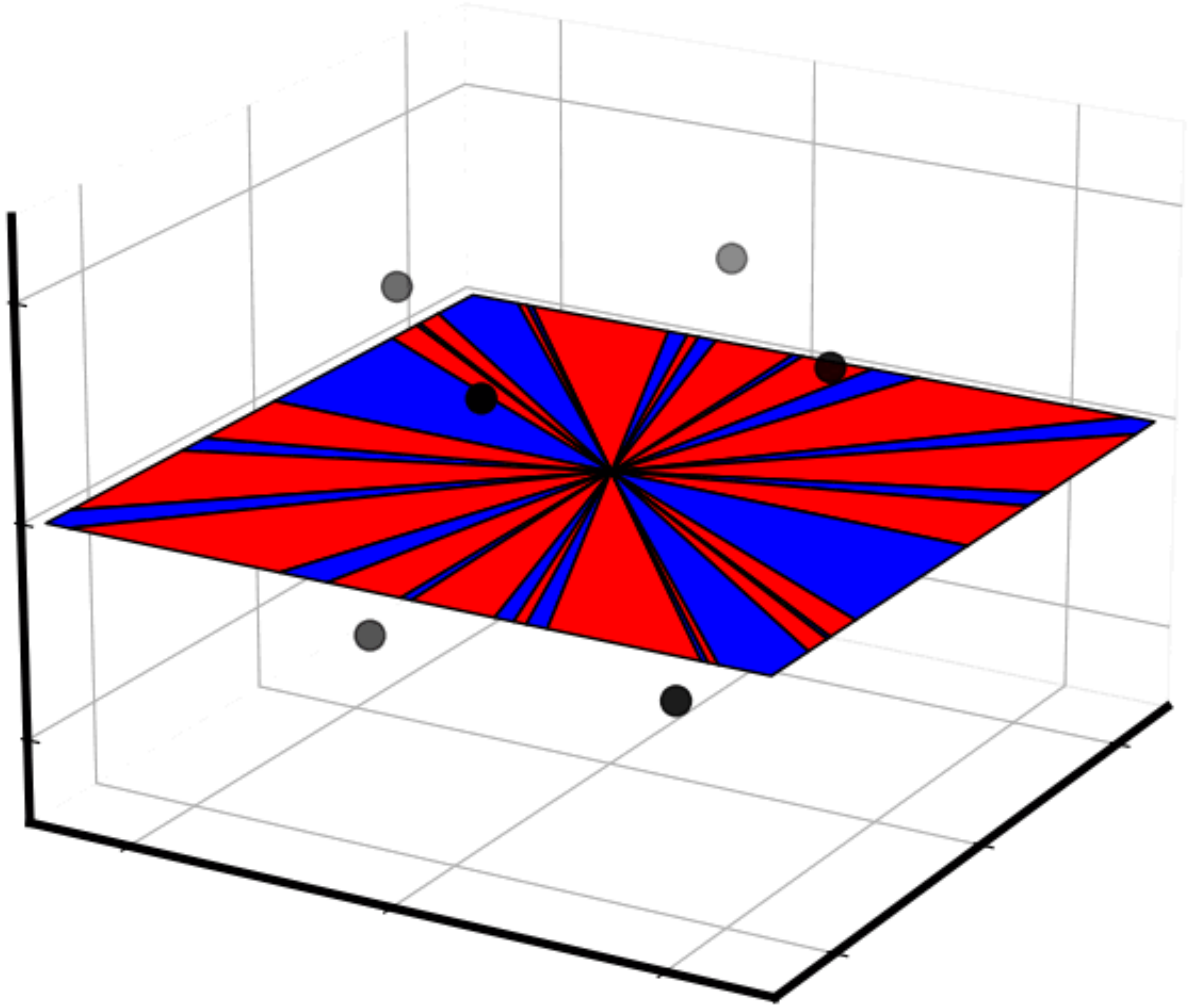
Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning

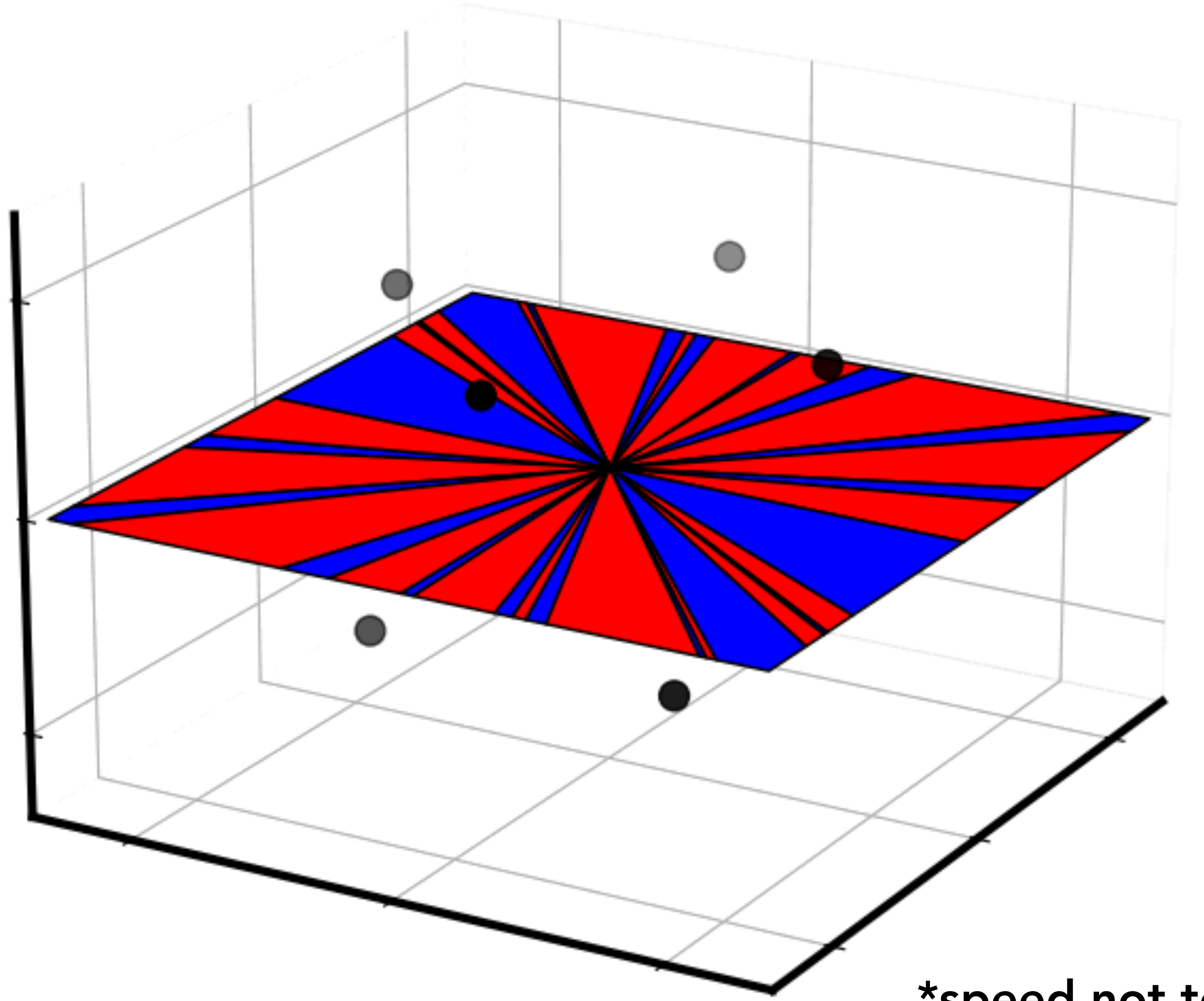
NeurIPS 2024 Spotlight

Daniel Kunin*, Allan Raventós*, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, Surya Ganguli

Lazy Regime



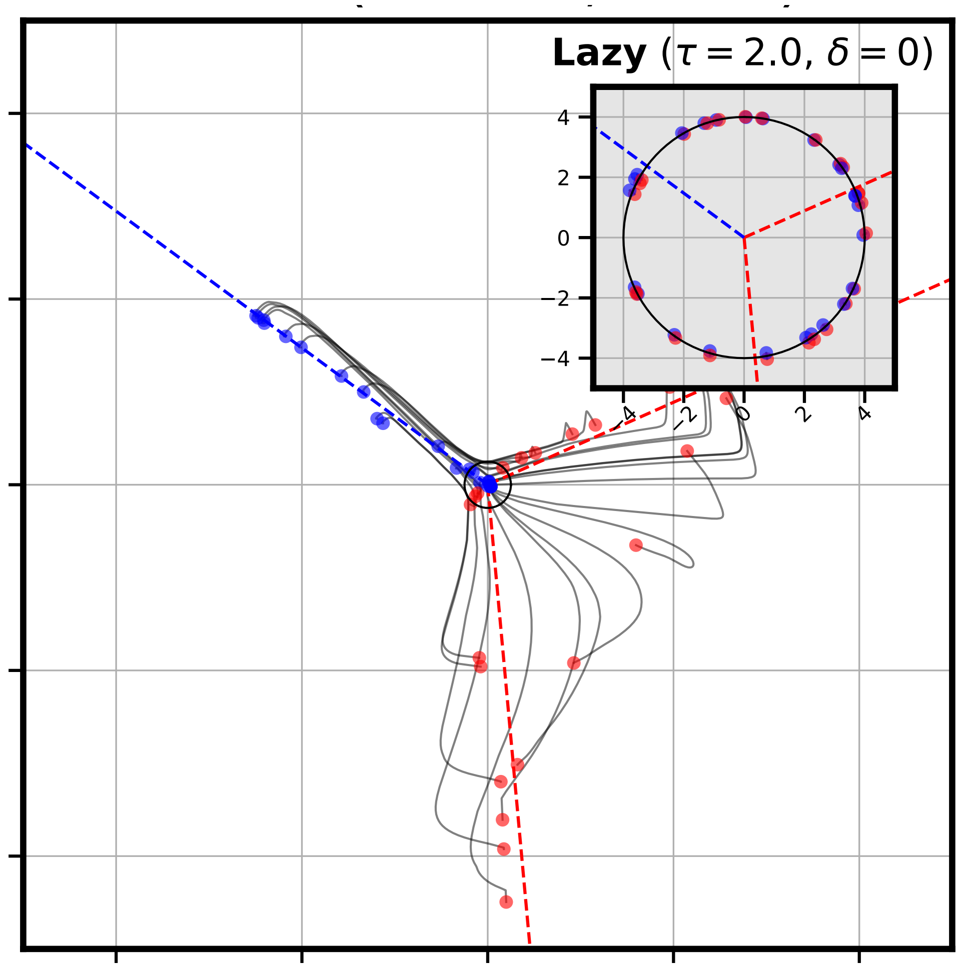
Rich Regime



*speed not to scale

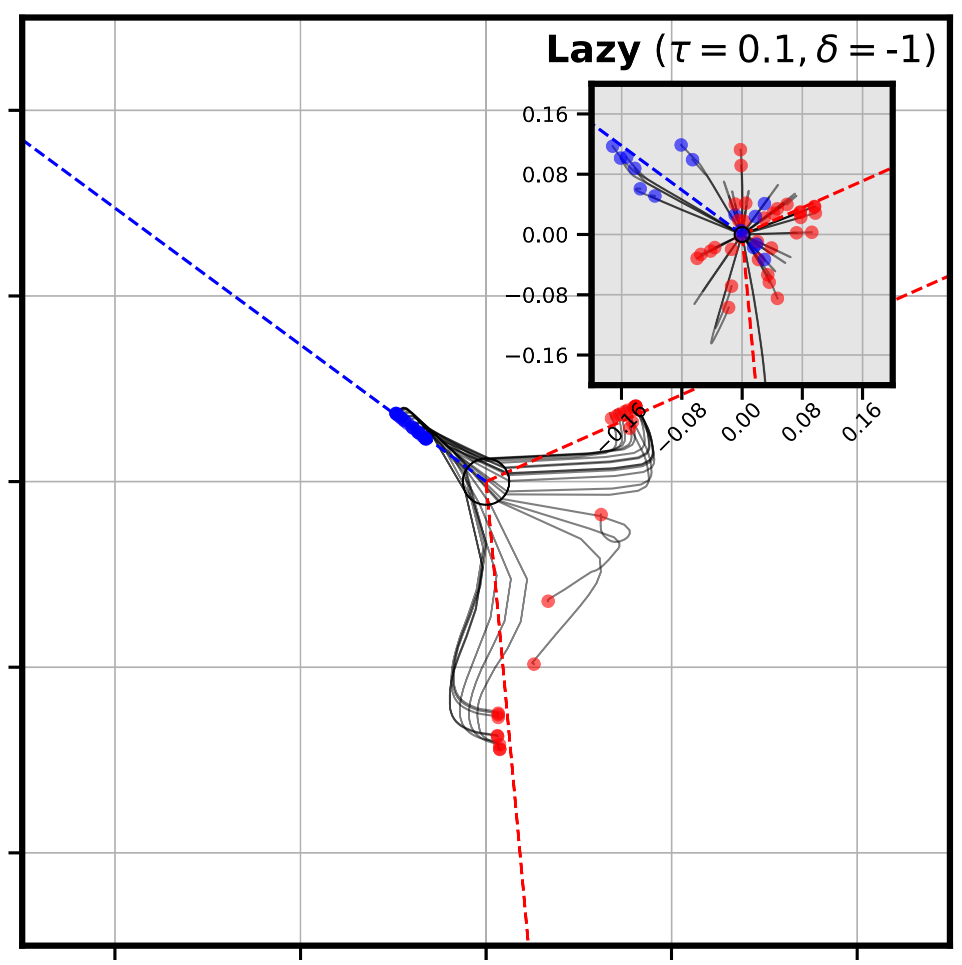
Motivating experiment

Chizat et al. 2019



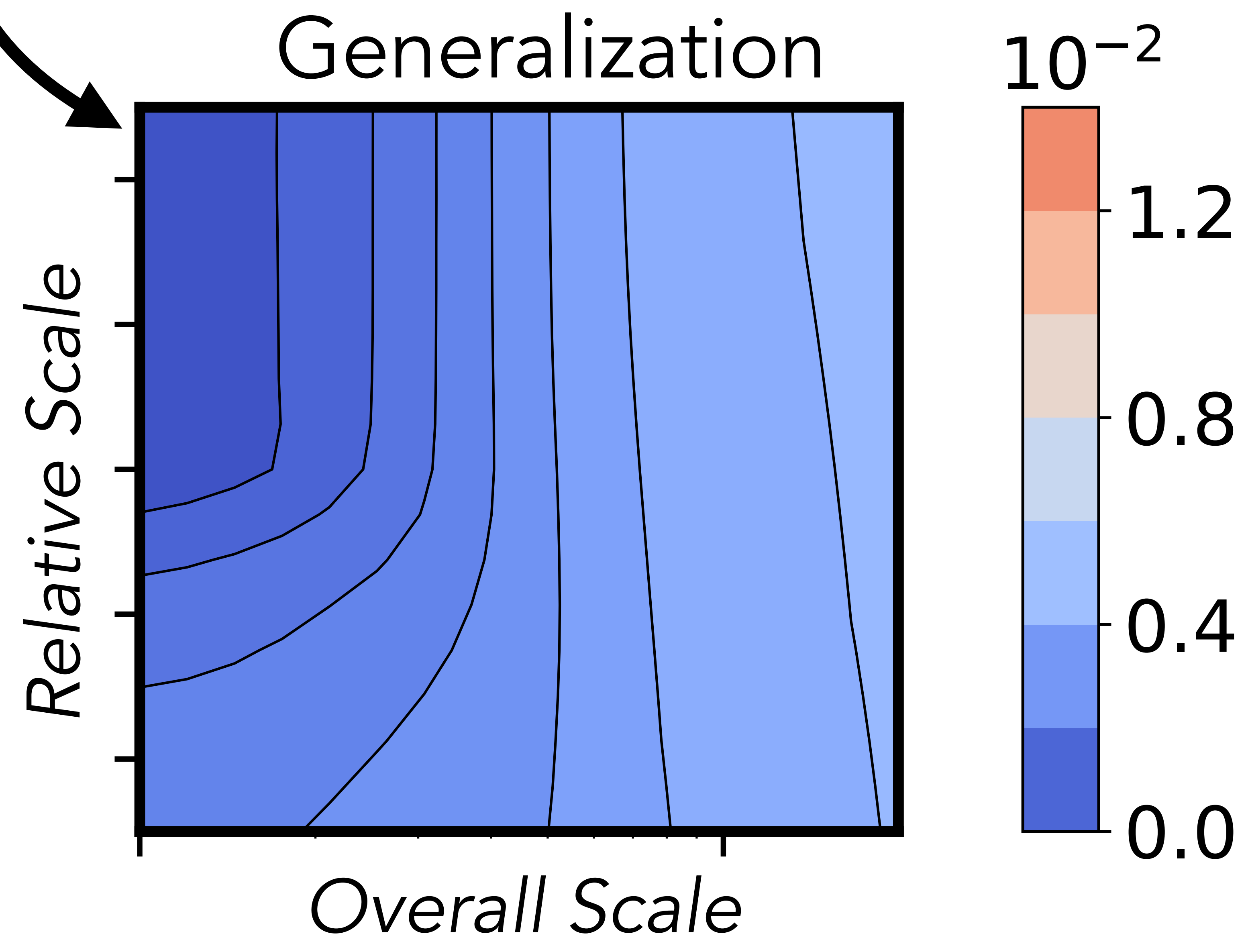
Overall scale determines feature learning

Kunin et al. 2024



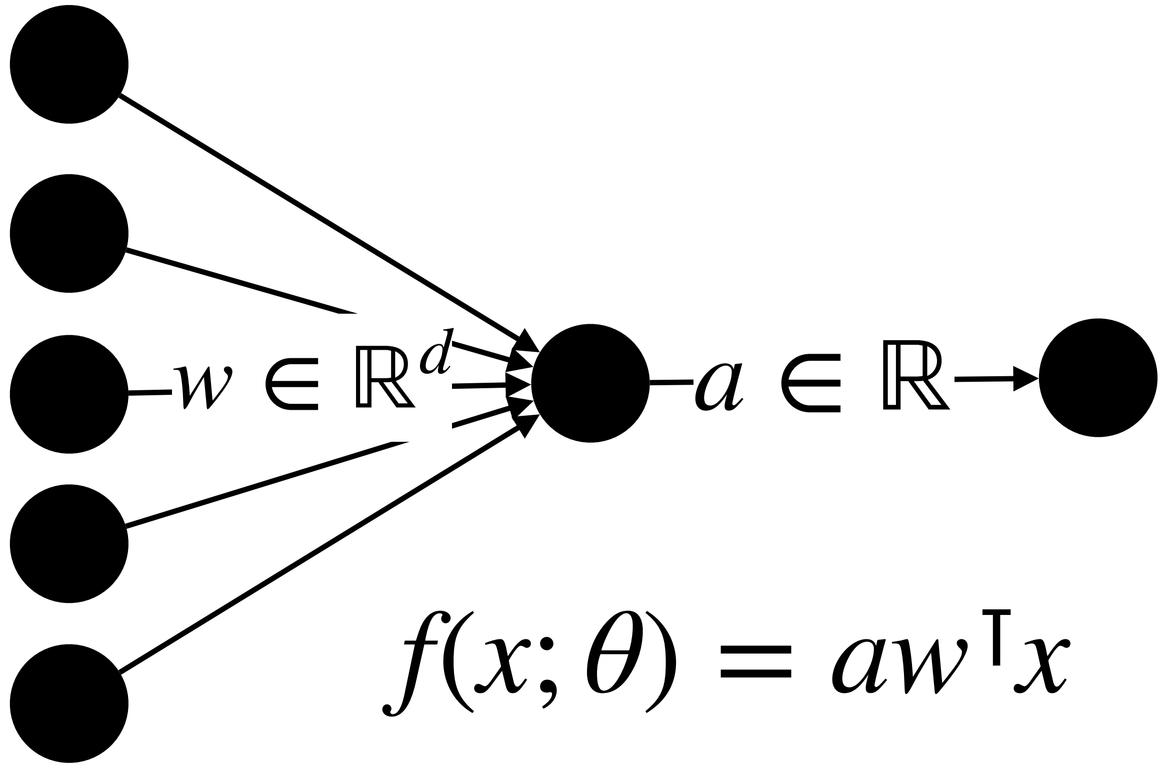
Overall scale <u>and relative scale</u> determine feature learning

The best generalization occurs at small overall scale and large relative scale



TL;DR: We derive exact solutions to a minimal model that transitions between lazy and rich learning, precisely elucidating how unbalanced initialization variances and learning rates determine the degree of feature learning in a finite-width network.

Part I: A Minimal model that transitions between lazy & rich



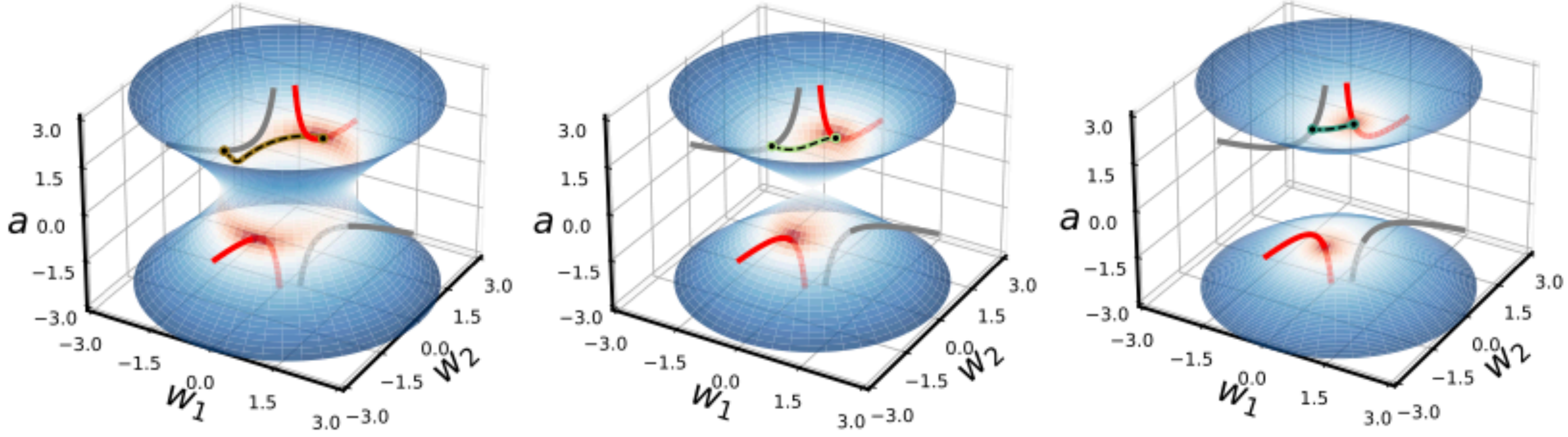
$$f(x;\theta) = aw^{\intercal}x$$

We study a two-layer linear network with a single hidden neuron defined by the map $f(x;\theta) = aw^{\intercal}x$ first proposed by Azulay et al. 2021. We show how to solve the gradient flow dynamics exactly by solving a Riccati equation and Bernoulli ODE.

The relative scale of the initialization determines the sign and magnitude of the conserved quantity

$$\delta = a^2 - \|w\|^2,$$

which constrains geometry of learning trajectories.

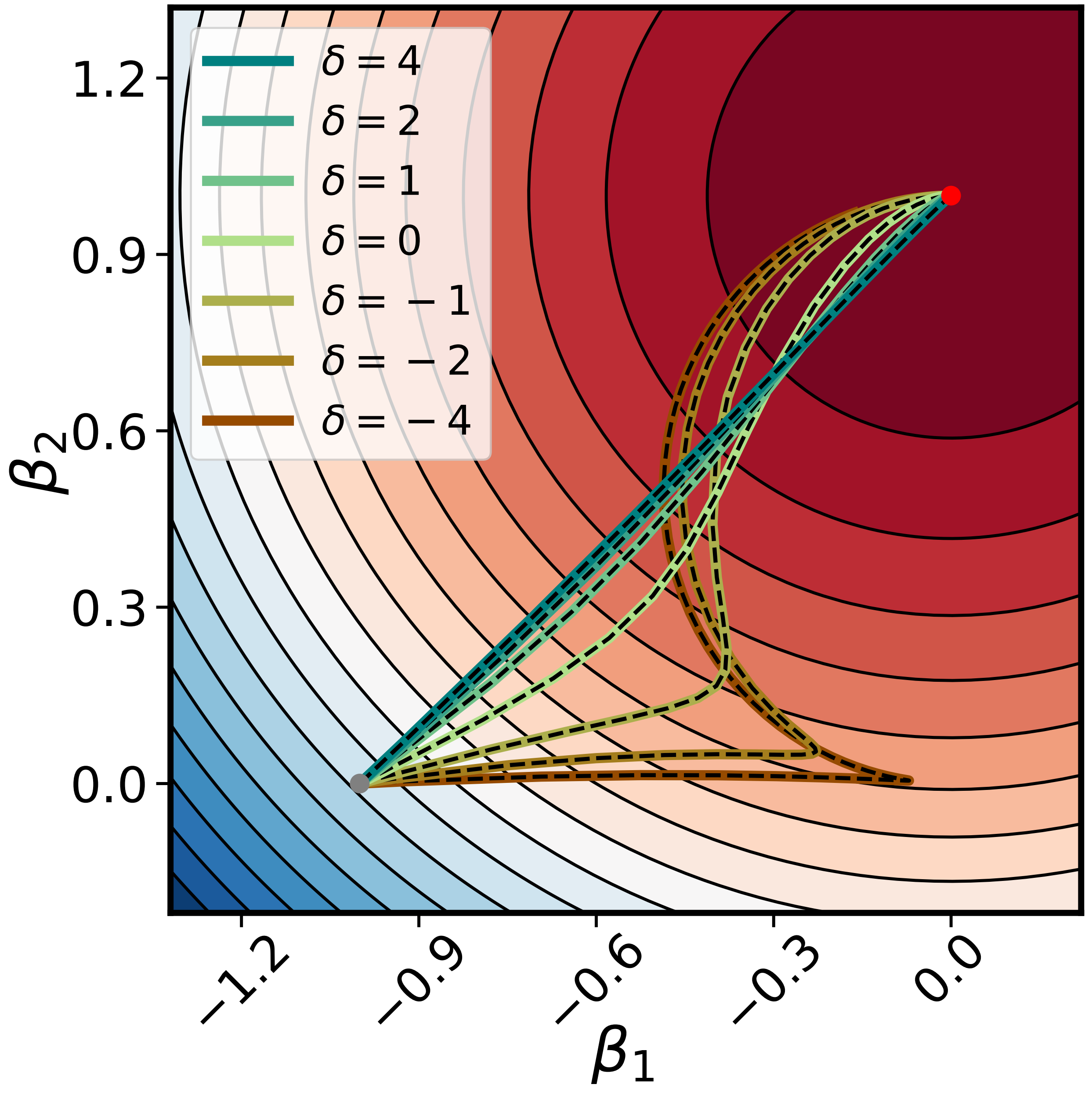


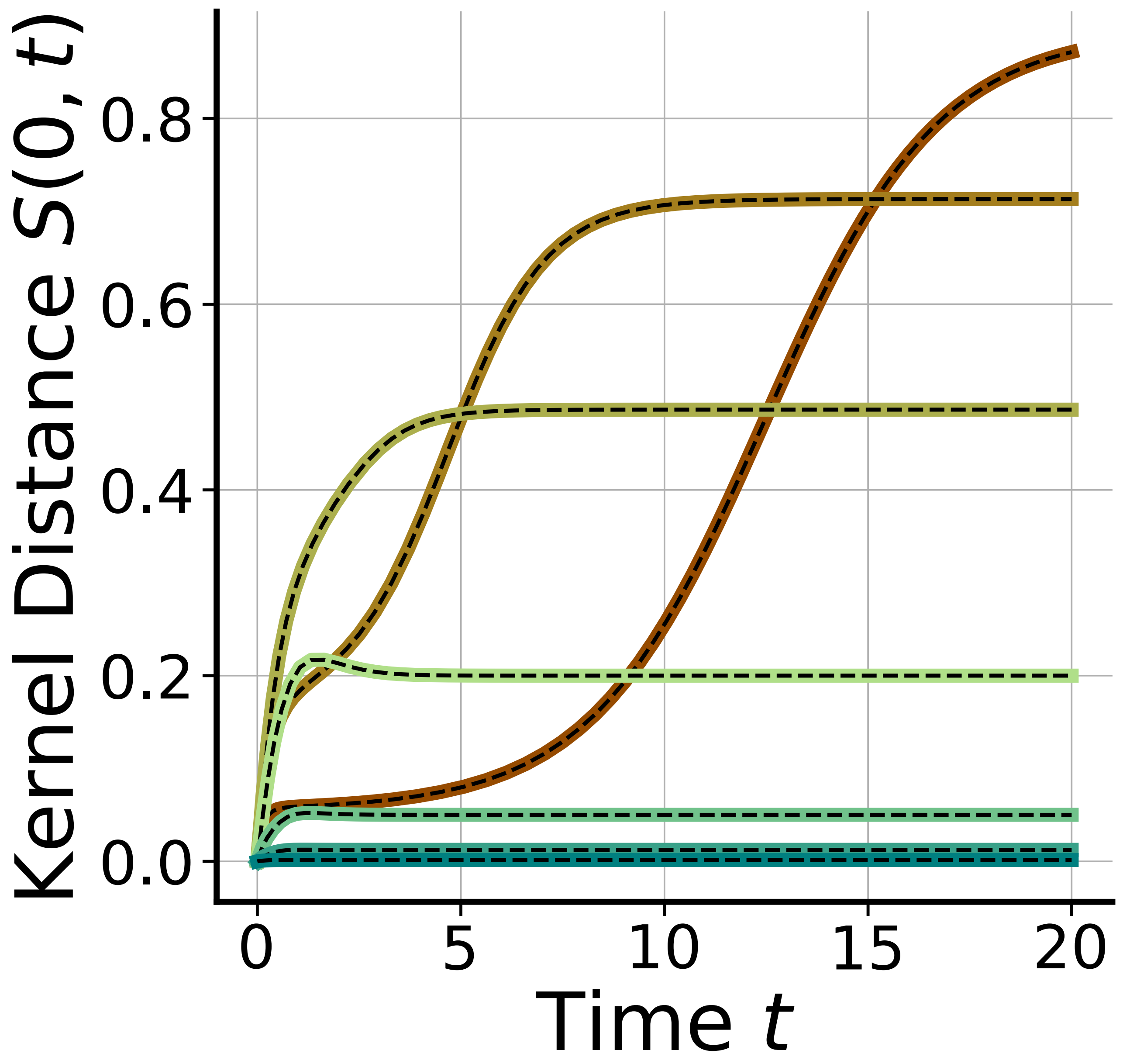
Downstream $\delta < 0$ Balanced $\delta = 0$ Upstream $\delta > 0$

Using the conserved quantity we derive a self-consistent equation for the dynamics of $\beta = aw$ in function space,

$$\dot{\beta} = -\underbrace{\left(\frac{\sqrt{\delta^2 + 4\|\beta\|^2} + \delta}{2} \mathbf{I}_d + \frac{\sqrt{\delta^2 + 4\|\beta\|^2} - \delta}{2} \frac{\beta\beta^\intercal}{\|\beta\|^2} \right)}_{M} \frac{\partial \mathscr{L}}{\partial \beta},$$

which can be interpreted as preconditioned gradient flow on the loss. The preconditioning matrix determines the trajectory and NTK matrix $K = XMX^\intercal$.





We consider the influence of δ on the preconditioning matrix and notice three distinct regimes:

1. **Lazy** — when $\delta \gg 0$, $M \approx \delta \mathbf{I}_d$, akin to linear regression.
2. **Rich** — when $\delta = 0$, $M = \sqrt{\eta_a \eta_w} \|\beta\| (\mathbf{I}_d + \frac{\beta\beta^\intercal}{\|\beta\|^2})$, akin to to silent alignment (Atanasov et al. 2021).
3. **Delayed rich** — when $\delta \ll 0$, $M \approx |\delta| \frac{\beta\beta^\intercal}{\|\beta\|^2}$, initially lazy followed by rich

Part II: Extending analysis to wide, deep, and nonlinear networks

We consider the dynamics of a two-layer linear network, $f(x;\theta) = A^\intercal W x : \mathbb{R}^d \to \mathbb{R}^c$, where the matrix $\Delta = AA^\intercal - WW^\intercal \in \mathbb{R}^{h\times h}$ is conserved throughout gradient flow. By assuming structure on Δ at initialization we derive a self-consistent equation for the dynamics of $\beta = W^\intercal A$.

Theorem 4.2 When $\Delta = \delta \mathbf{I}_h$ and $h = d$ if $\delta < 0$ or $h = c$ if $\delta > 0$,

$$\dot{\beta} = -\frac{\partial \mathcal{L}}{\partial \beta}\sqrt{\beta^\intercal \beta + \frac{\delta^2}{4}\mathbf{I}_c} - \sqrt{\beta\beta^\intercal + \frac{\delta^2}{4}\mathbf{I}_d}\frac{\partial \mathcal{L}}{\partial \beta}$$

Using this expression the dynamics of the singular values of β can be described as a mirror flow with a **hyperbolic entropy** potential, which smoothly interpolates between an ℓ^1 and ℓ^2 penalty on the singular values for the rich ($\delta \to 0$) and lazy ($\delta \to \pm\infty$) regimes respectively.

We consider the dynamics of a two-layer piecewise linear network without biases, $f(x;\theta) = a^\intercal \sigma(Wx) : \mathbb{R}^d \to \mathbb{R}$, where the quantity $\delta_k = a_k^2 - \|w_k\|^2$ for each hidden neuron $k \in [h]$ is conserved through gradient flow.

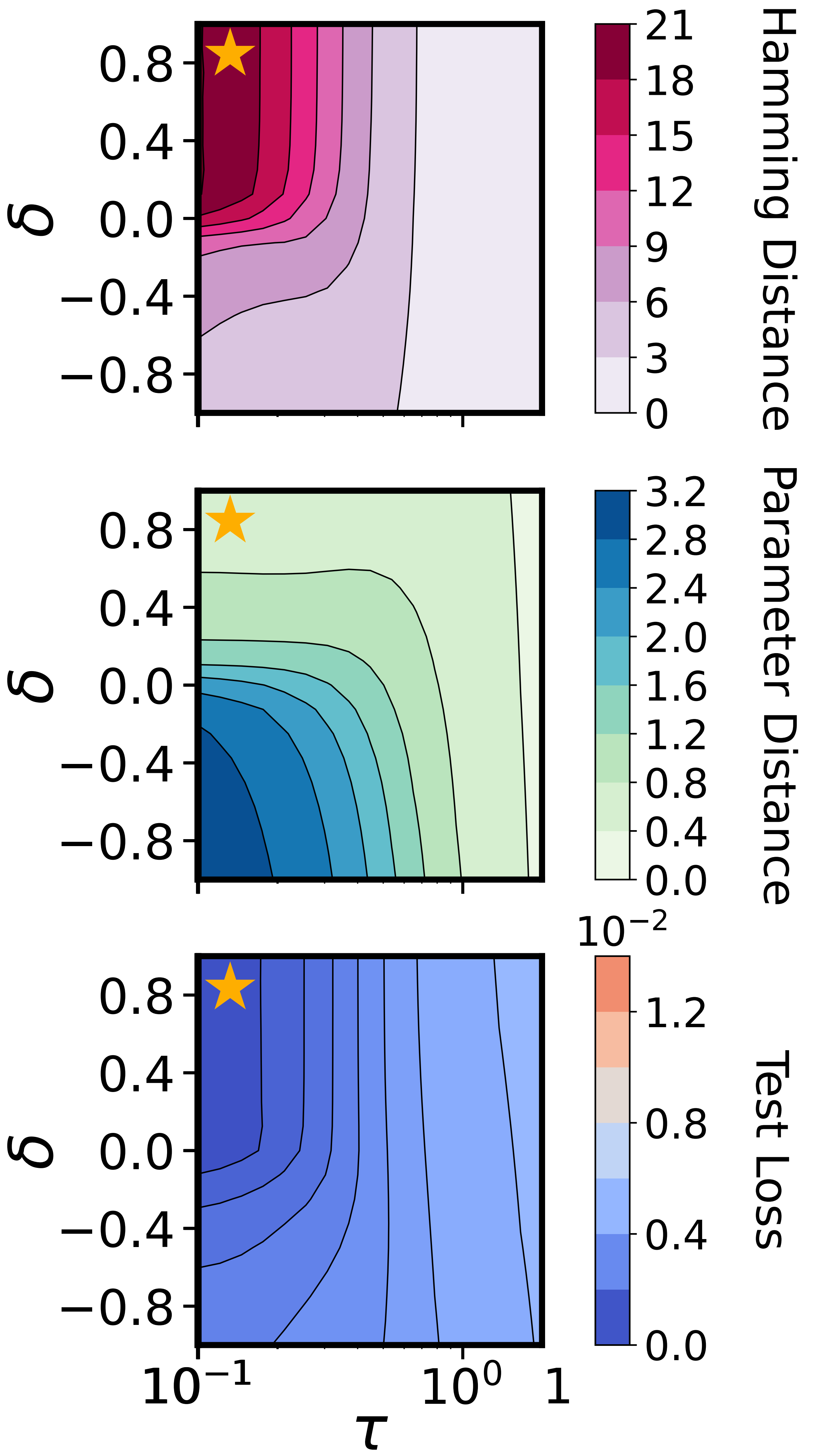
As in linear networks, we attempt to derive a self-consistent equation for the dynamics of each neuron's map $\beta_k = a_k w_k$,

$$\dot{\beta}_k = -M_k \underbrace{\sum_{i=1}^n c_{ki} x_i \left(f(x_i;\theta) - y_i\right)}_{\frac{\partial \mathscr{L}}{\partial \beta_k}}.$$

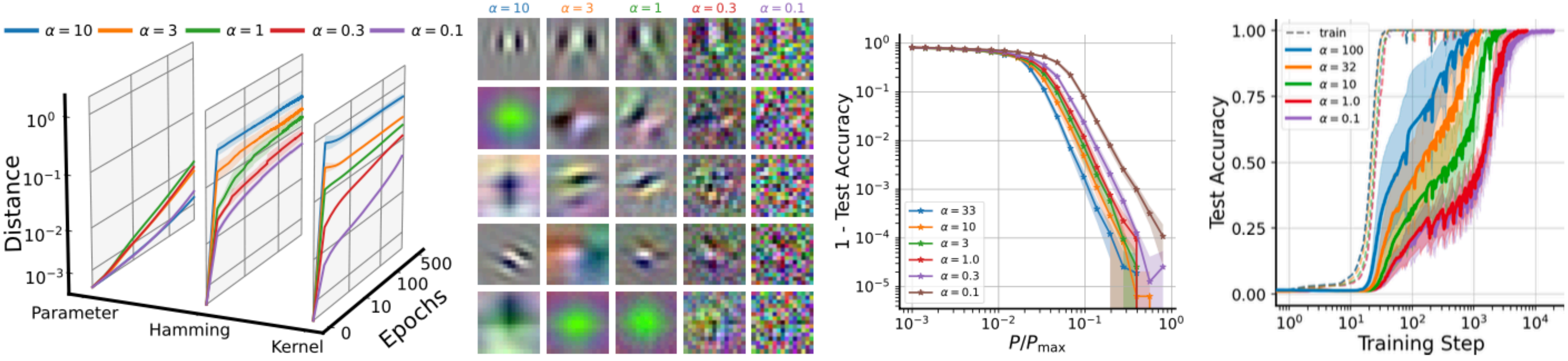
However, unlike the linear setting, the gradient $\partial\mathscr{L}/\partial\beta_k$ driving the dynamics is not shared for all neurons because of its dependence on the activation patterns $c_{ki} = \sigma'(w_k^\intercal x_i)$.

Nevertheless, we can understand the influence of δ_k on the learning dynamics by considering the radial and directional dynamics of β_k — the activation patterns c_{ki} only depend on the direction of β_k.

Rapid feature learning ★ is due to a large change in activation patterns, but a small change in parameter space.



Part III: Observing the influence of relative scale in practice



(a) Kernel Distance (b) Convolutional Filters (c) Random Hierarchy (d) Transformer Grokking

We regulate the first layer's learning speed relative to the rest of the network by dividing its initialization by α — $\alpha = 1$ represents standard parameterization, while $\alpha \gg 1$ and $\alpha \ll 1$ corresponds to upstream and downstream initializations, respectively.

Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning

NeurIPS 2024
Spotlight

Daniel Kunin*, Allan Raventós*, Clémentine Dominé, Feng Chen,
David Klindt, Andrew Saxe, Surya Ganguli

Come to our poster
to learn more!