An exactly solvable model for emergence and scaling laws in the multitask sparse parity problem

Yoonsoo Nam*[,a], Nayara Fonseca*[,a], Seok Hyeong Lee[b], Chris Mingard[a,c], and Ard A. Louis[a]

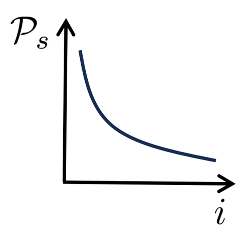
[arXiv:2404.17563]

Question: Can we find an analytically solvable model that exhibits both: 1) Emergence and 2) Scaling Laws?

Setup

- Represent **'skills' as orthogonal functions**, where $\boldsymbol{g_k(i,x)}$ are the **skill basis functions.**
- Apply to the **multitask sparse parity problem** [3], where task frequencies follow a power-law.

Skill idx (I)	Control bits	Skill bits (X)	y	$M(i,x)$	$g_1(i,x)$	$g_2(i,x)$	…	$g_{n_s}(i,x)$
1	1000000	110110000100	S	[1,1,0]	1	0	…	0
1	1000000	100101010001	$-S$	[0,1,0]	-1	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	0100000	001001011011	$-S$	[0,0,1]	0	-1	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n_s	0000001	001010100110	$-S$	[1,1,1]	0	0	…	-1



$$\mathcal{P}_s(I=i) := \frac{i^{-(\alpha+1)}}{\sum_j^{n_s} j^{-(\alpha+1)}} \qquad g_k(i,x) := \begin{cases} (-1)^{\sum_j M_j(i,x)} & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}$$

Each skill (task) is an orthogonal basis function

Target function. $f^*(i,x) := S\sum_{k=1}^{n_s} g_k(i,x)$

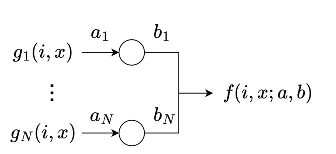
MSE loss. $\mathcal{L}_k := \frac{1}{2}\mathbf{E}_X\left[(f^*(I=k,X)-f(I=k,X))^2\right] \qquad \mathcal{L} = \sum_{k=1}^{n_s}\mathcal{P}_s(I=k)\mathcal{L}_k$

Skill strength. The k^{th} coefficient if a model is expanded in the basis of the skill functions.

$$\mathcal{R}_k(T) := \mathbf{E}_X\left[g_k(I=k,X)f_T(I=k,X)\right]$$

Measures how well the k^{th} skill is learned by the model at time T

Multilinear Model



$$f_T(i,x;a,b) = \sum_{k=1}^{N} \underbrace{a_k(T)b_k(T)}\underbrace{g_k(i,x)}$$

Skill functions as basis functions

Multilinear

$a_k(T)b_k(T) = \mathcal{R}_k(T)$

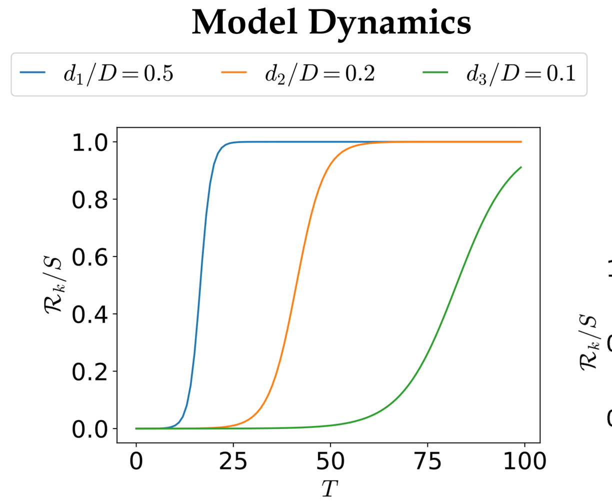
Analytically solvable under gradient flow

$$\frac{\mathcal{R}_k(T)}{S} = \frac{1}{1+\left(\frac{S}{\mathcal{R}_k(0)}-1\right)e^{-2\eta\frac{d_k}{D}ST}}$$

Key Properties

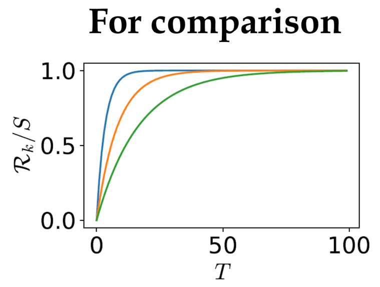
1. **Decoupling** among the skills
2. **Sigmoidal** growth

Model Dynamics



Less frequent skills have a more delayed growth

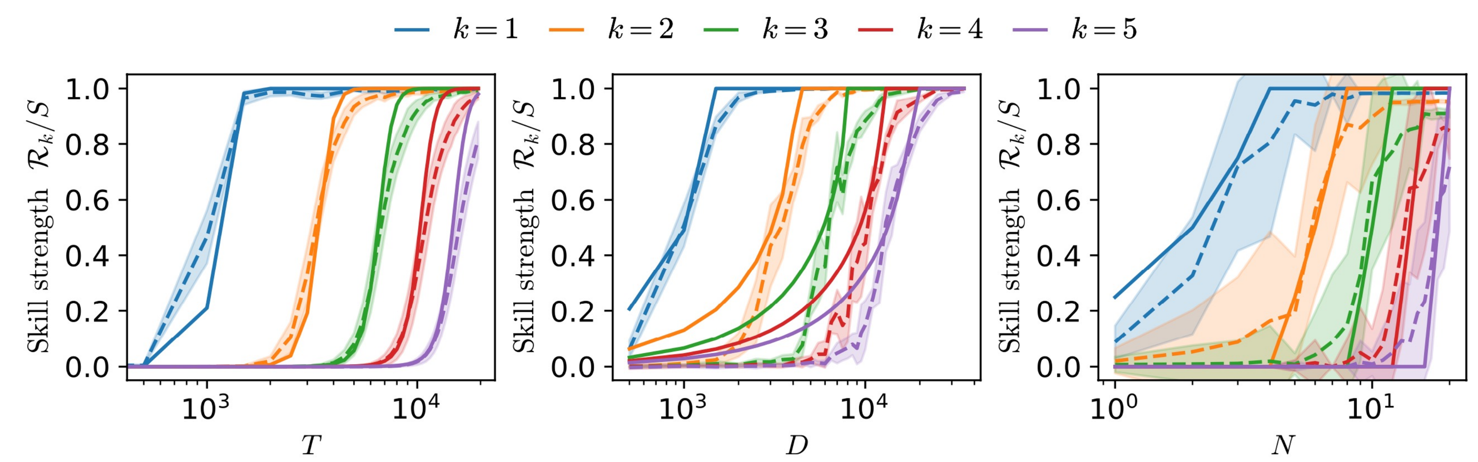
Linear model For comparison



Predicting Emergence

2-layer MLP

Minimally extending the multilinear model (using a single fit parameter), we can predict, the time, data, and parameter emergence of a 2-layer MLP by **calibrating on the first skill** (blue).

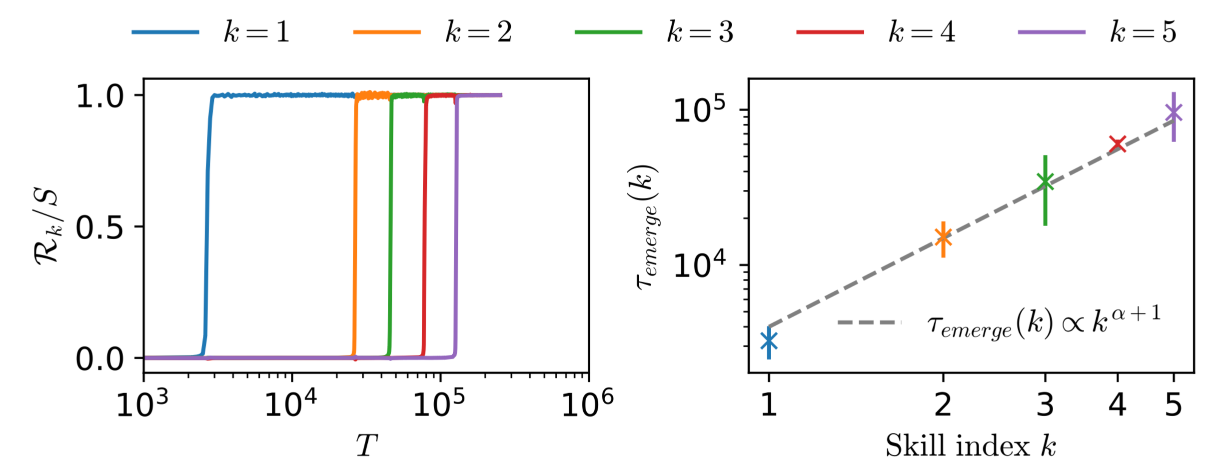


(a) Time emergence (b) Data emergence (c) Parameter emergence

$\mathcal{R}_k$ is normalized by the target scale S such that $\mathcal{R}_k/S=1$ means zero skill loss

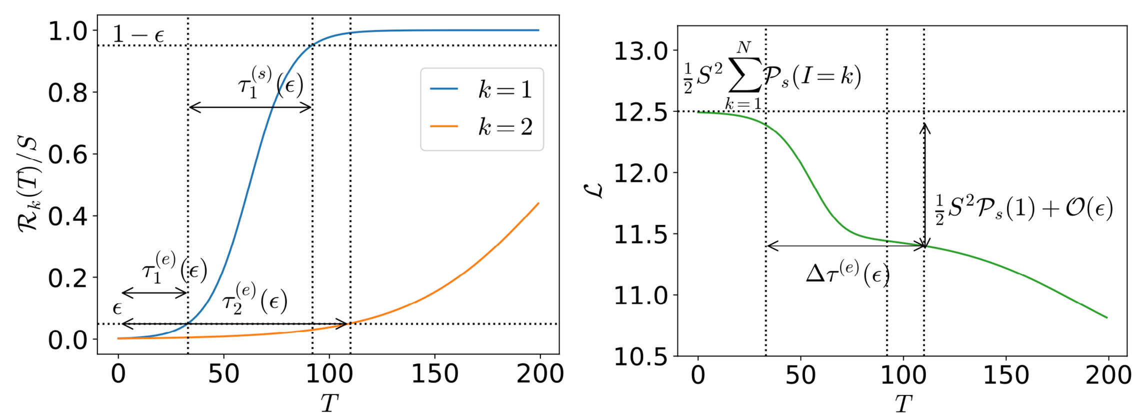
Transformer

Time emergence in a transformer on the multitask sparse parity task with $\alpha=0.9$.



One block decoder transformer (embedding layer with output dimension 512) and four attention heads.

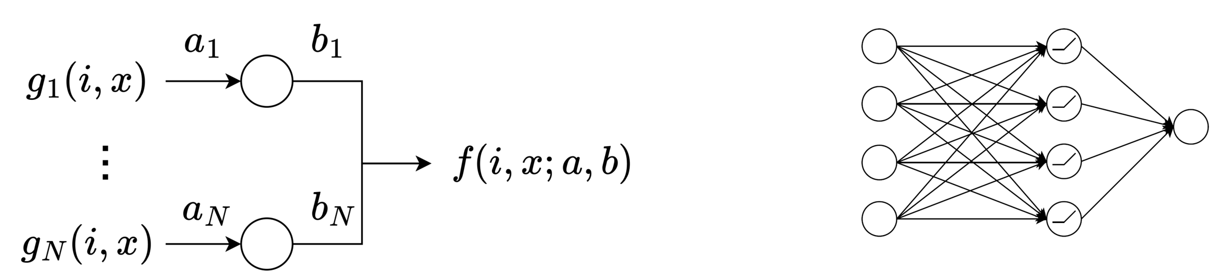
Stage-like training



$$\frac{\Delta\mathcal{L}}{\mathcal{L}} \approx \frac{\mathcal{P}_s(I=k)}{\sum_{j=k}^{\infty}\mathcal{P}_s(I=j)}$$

$$\frac{\Delta T}{T} = \frac{\tau_{k+1}^{(e)}(\epsilon)-\tau_k^{(e)}(\epsilon)}{\tau_k^{(e)}(\epsilon)}$$

$$\frac{\Delta\mathcal{L}}{\mathcal{L}} = -\frac{\alpha}{\alpha+1}\frac{\Delta T}{T}$$



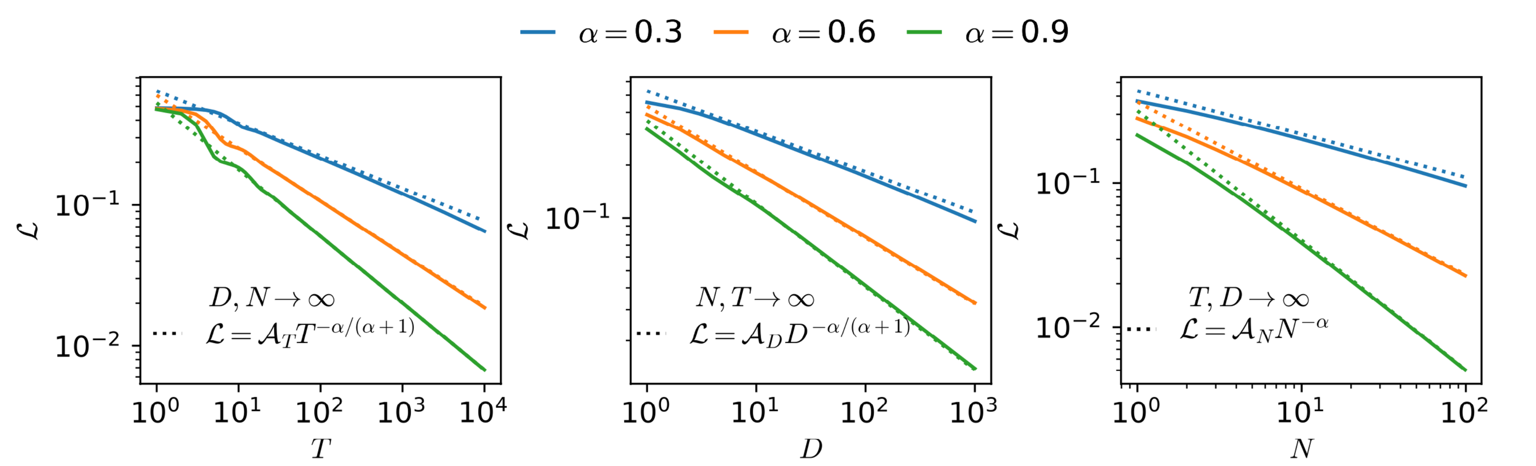
Multilinear model **2-layer MLP**

The layerwise structure and large differences in skill frequency lead to stage-like training, where the model learns more frequent skills before the next skill emerges.

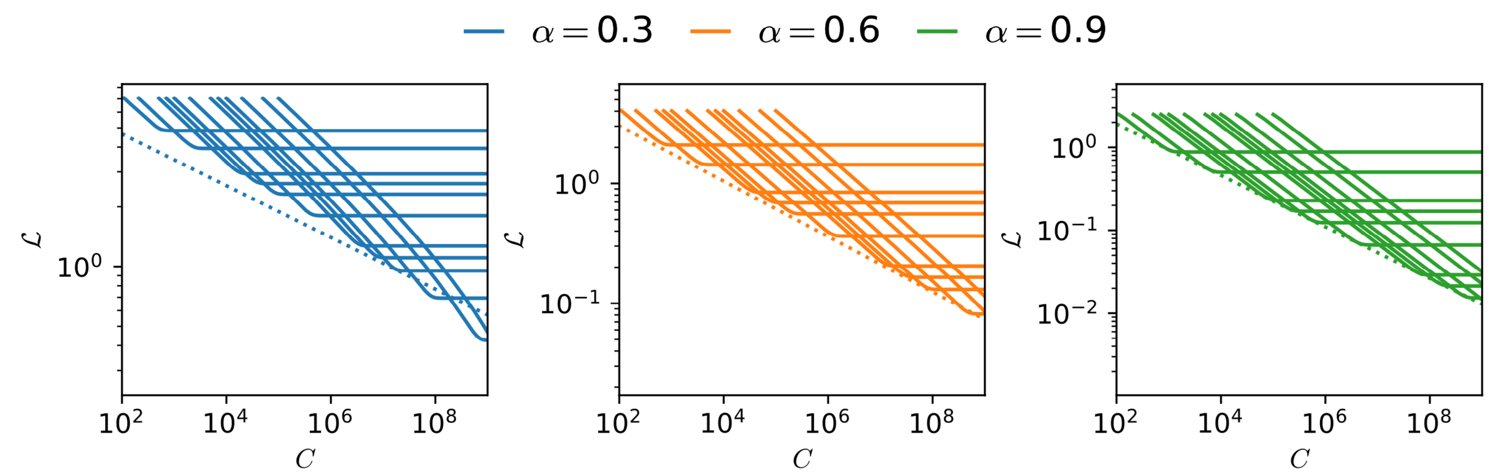
Stage-like training helps us understand the scaling laws and explains why a multilinear model well approximates the emergence in neural networks.

Scaling Laws

Using the decoupled dynamics, we can **analytically** derive the time, data, parameter, and compute scaling laws for the MSE total loss (including prefactors).



(a) Time scaling (b) Data scaling (c) Parameter scaling



The solid lines are the learning curves of the model as a function of compute $C=T\times N$ with varying parameters N from 10 (top plateau) to 10^4 (bottom plateau). Dotted lines are optimal compute scaling laws with exponent $-\alpha/(\alpha+2)$.

Discussion: Why would an MLP behave like a multilinear model with fixed skill functions and decoupled dynamics?

Multilinear	MLP
Fixed basis functions	Feature learning
Decoupled dynamics	No decoupling
Layerwise structure	

Power-law in skill frequency + sigmoidal dynamics = stage-like training = effective decoupling of skills

References

[1] Kaplan, McCandlish, et al., Scaling laws for neural language models. *arXiv:2001.08361, 2020.*
[2] Wei et al., Emergent abilities of large language models. *TMLR 2022.*
[3] Michaud et al., The Quantization Model of Neural Scaling, *NeurIPS 2023.*
[4] Saxe et al., Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, *ICLR 2014.*

a Rudolf Peierls Centre for Theoretical Physics, University of Oxford
b Center for Quantum Structures in Modules and Spaces, Seoul National University
c Physical and Theoretical Chemistry Laboratory, University of Oxford

(*) Equal contribution; yoonsoo.nam@physics.ox.ac.uk, nayara.fonsecadesa@physics.ox.ac.uk