集智俱乐部

中山大学 SUN YAT-SEN UNIVERSITY

PMI 物理，机器与智能

大模型上下文学习的统计物理视角

李宇豪

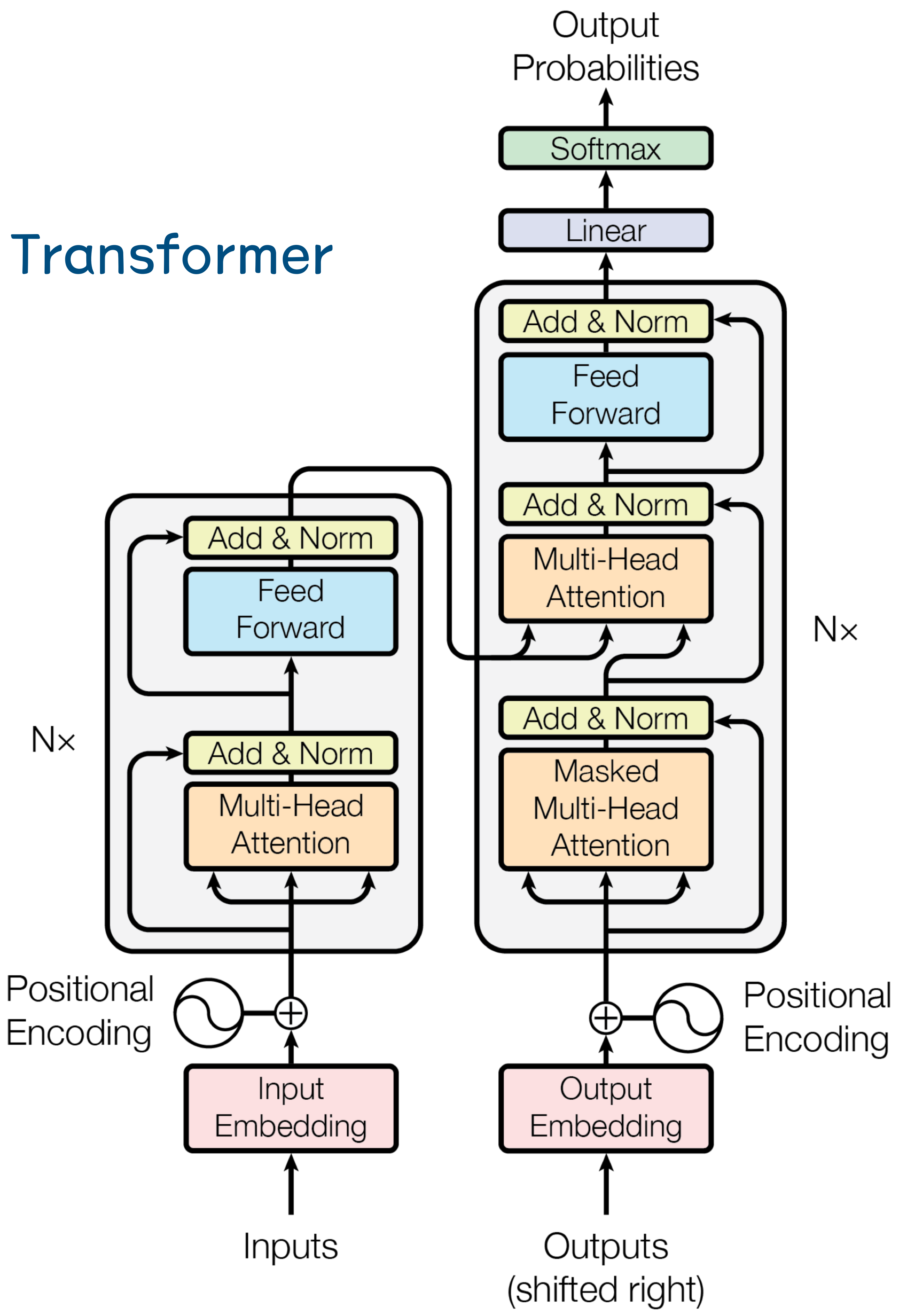
2025年7月24日

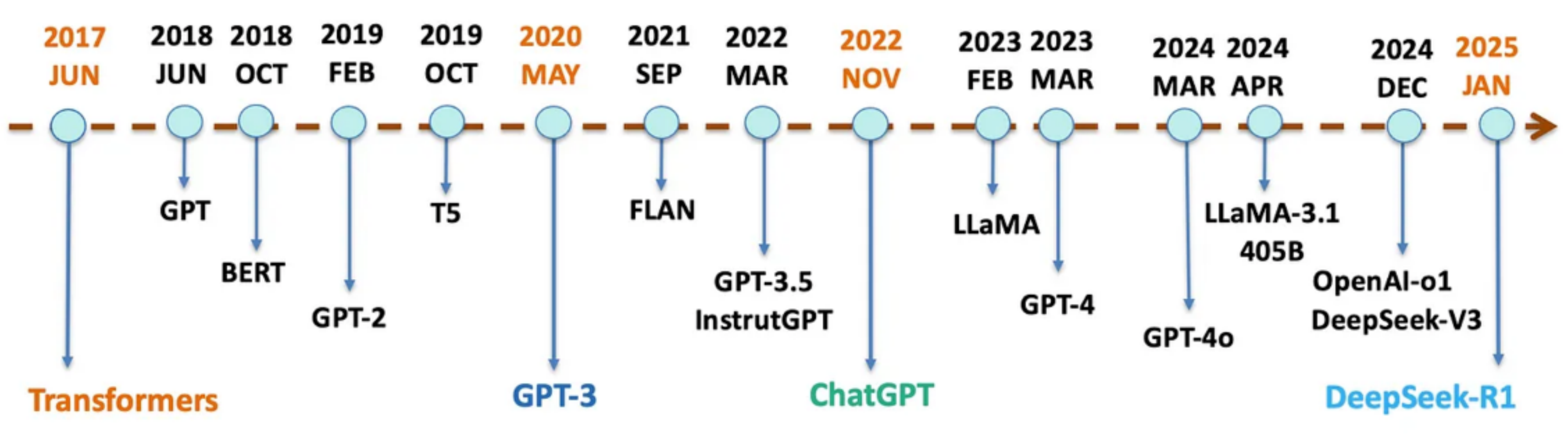
Yu-Hao.Li@outlook.com

https://liyuhao.com.cn

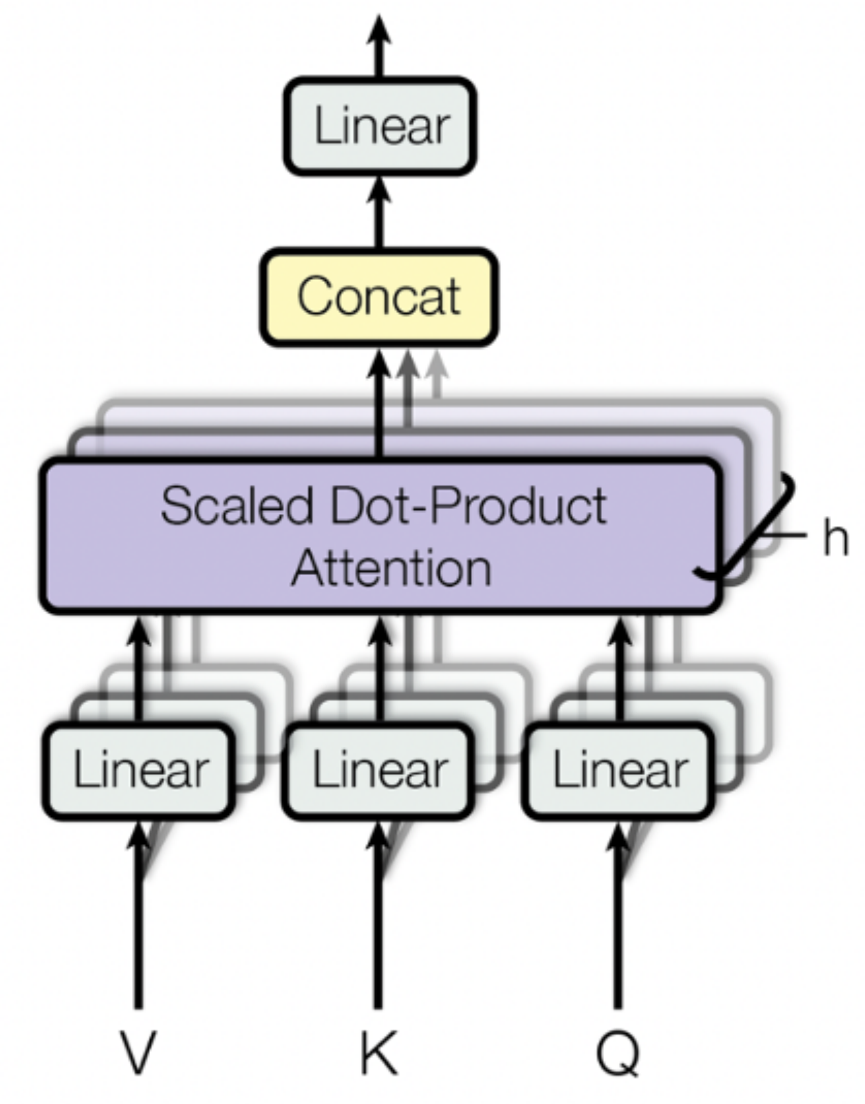
Introduction

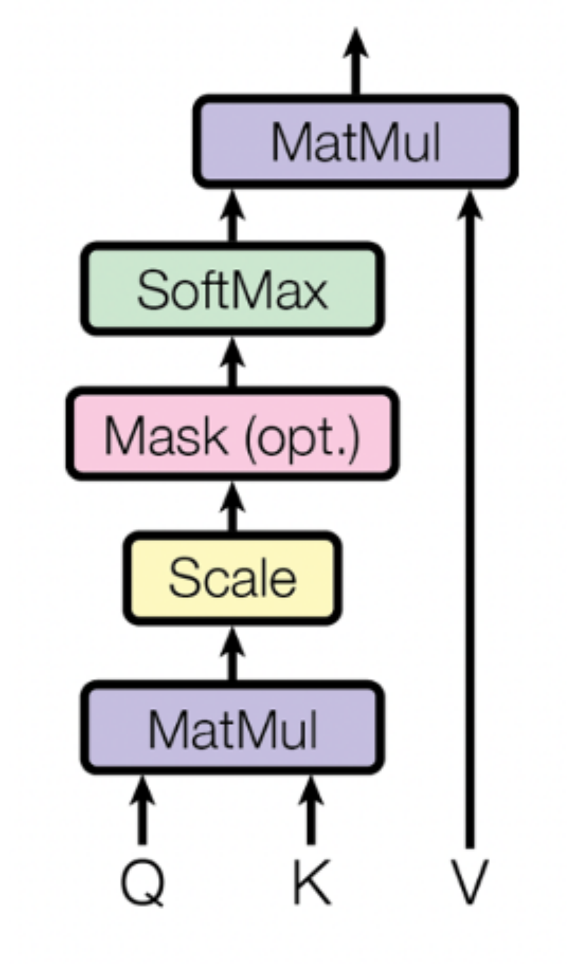
Transformer





LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

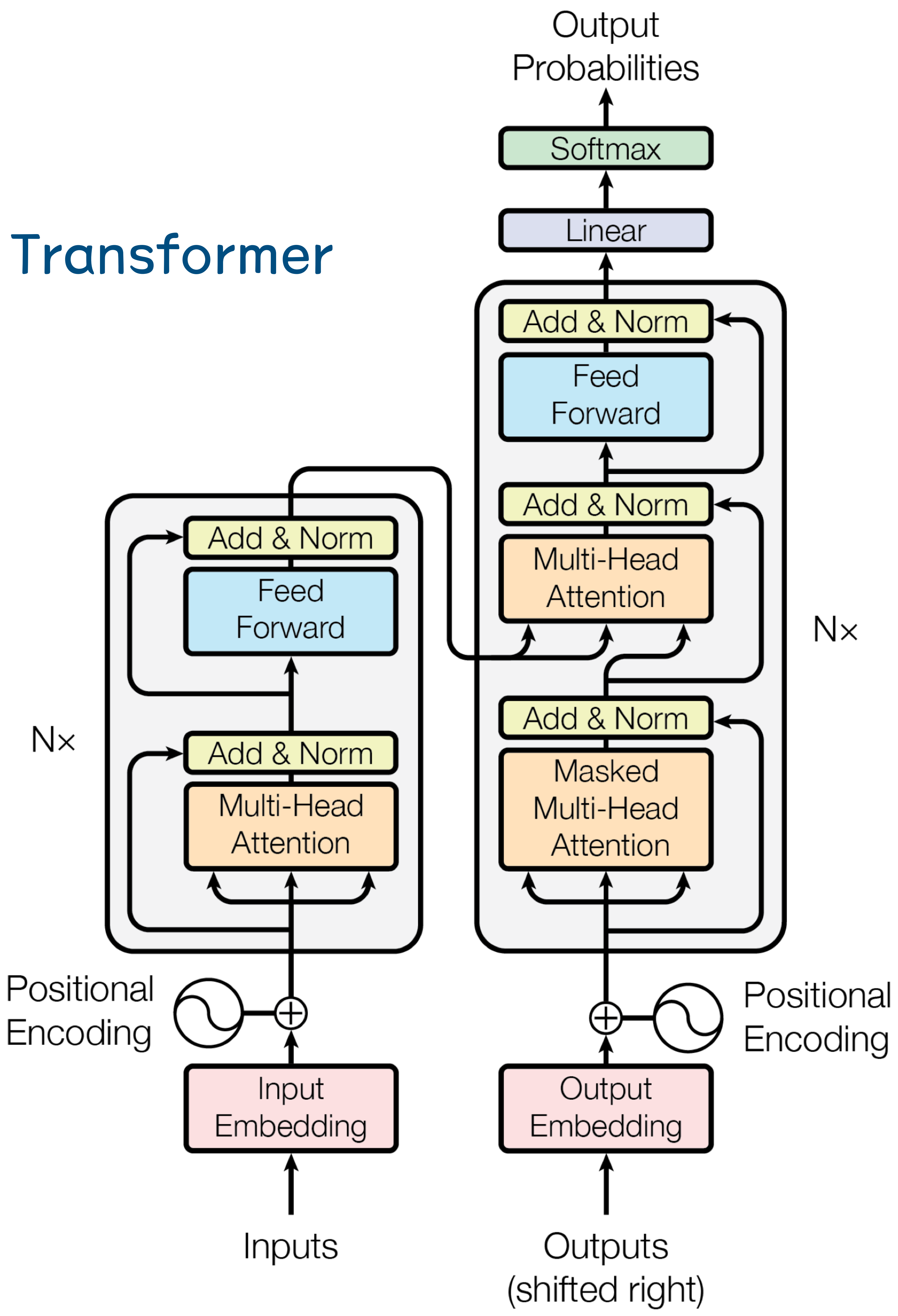




A. Vaswani et al. Attention Is All You Need. arXiv: 1706.03762

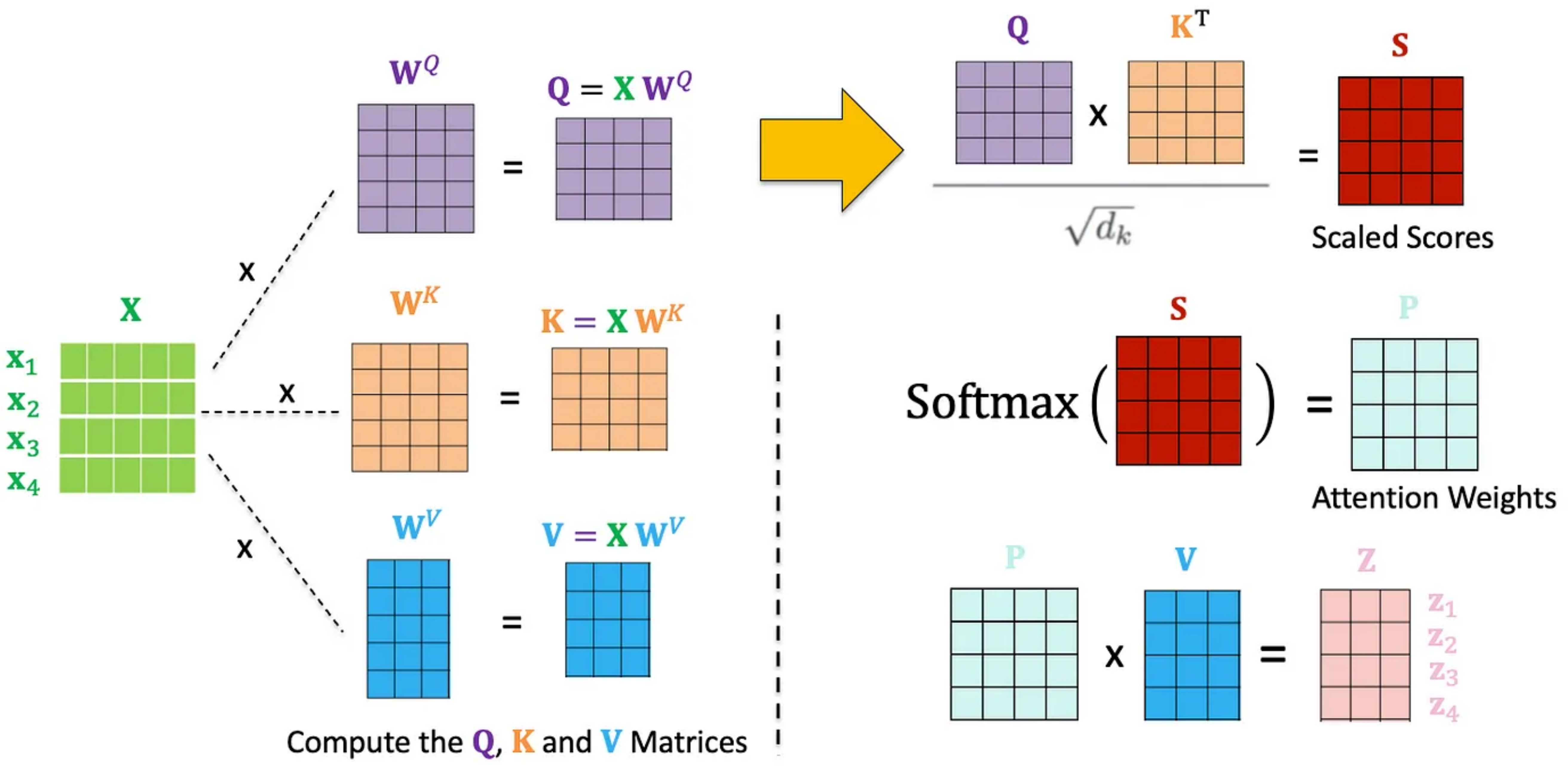
Introduction

Transformer

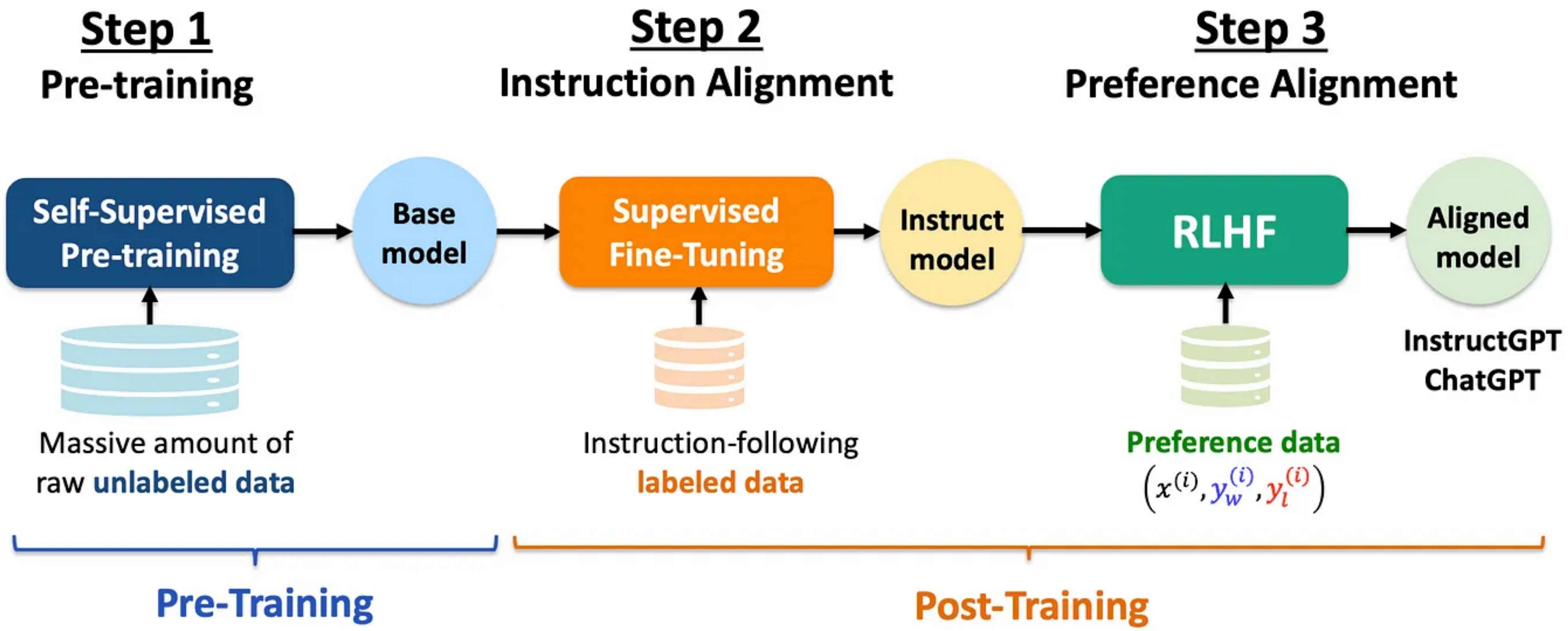


Self-Attention

$$\mathbf{Y}(\mathbf{X}) = \mathbf{W}_V \mathbf{X} \cdot \text{softmax}\left(\frac{(\mathbf{W}_Q \mathbf{X})^\top \mathbf{W}_K \mathbf{X}}{\sqrt{d_k}}\right)$$

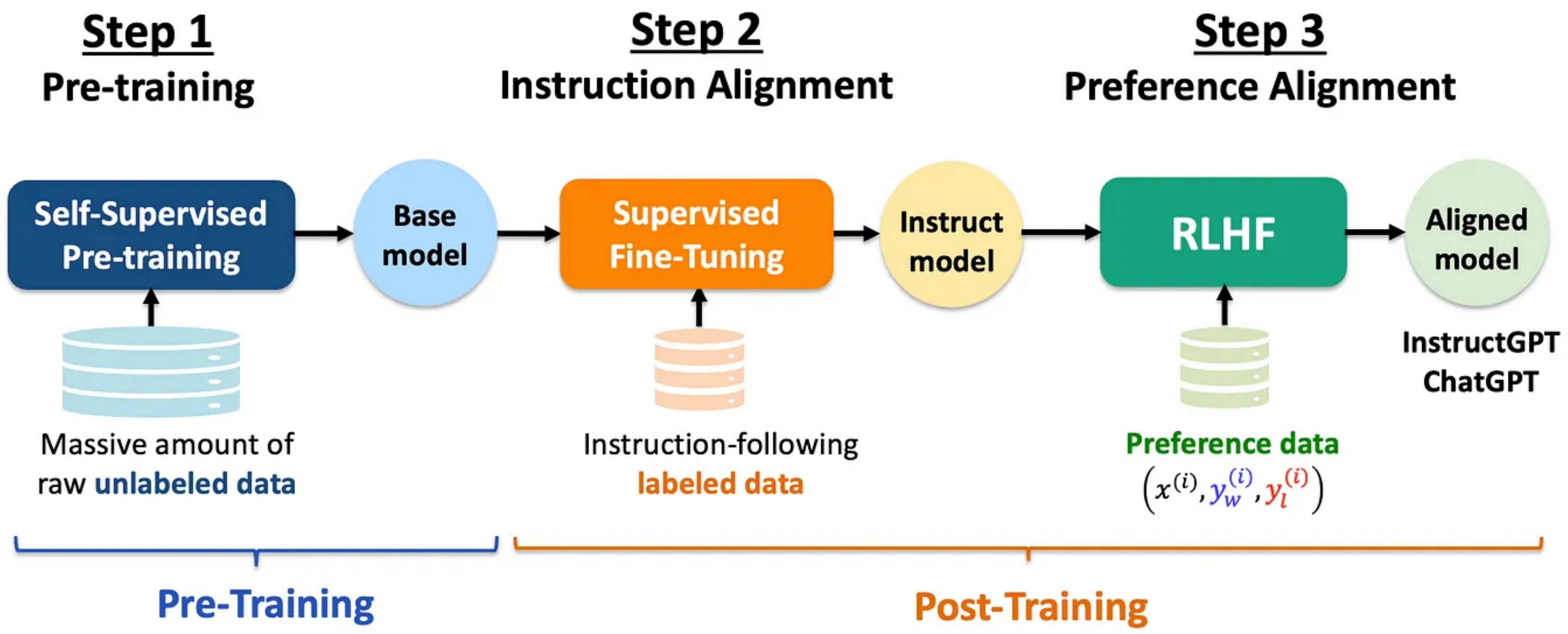


Introduction

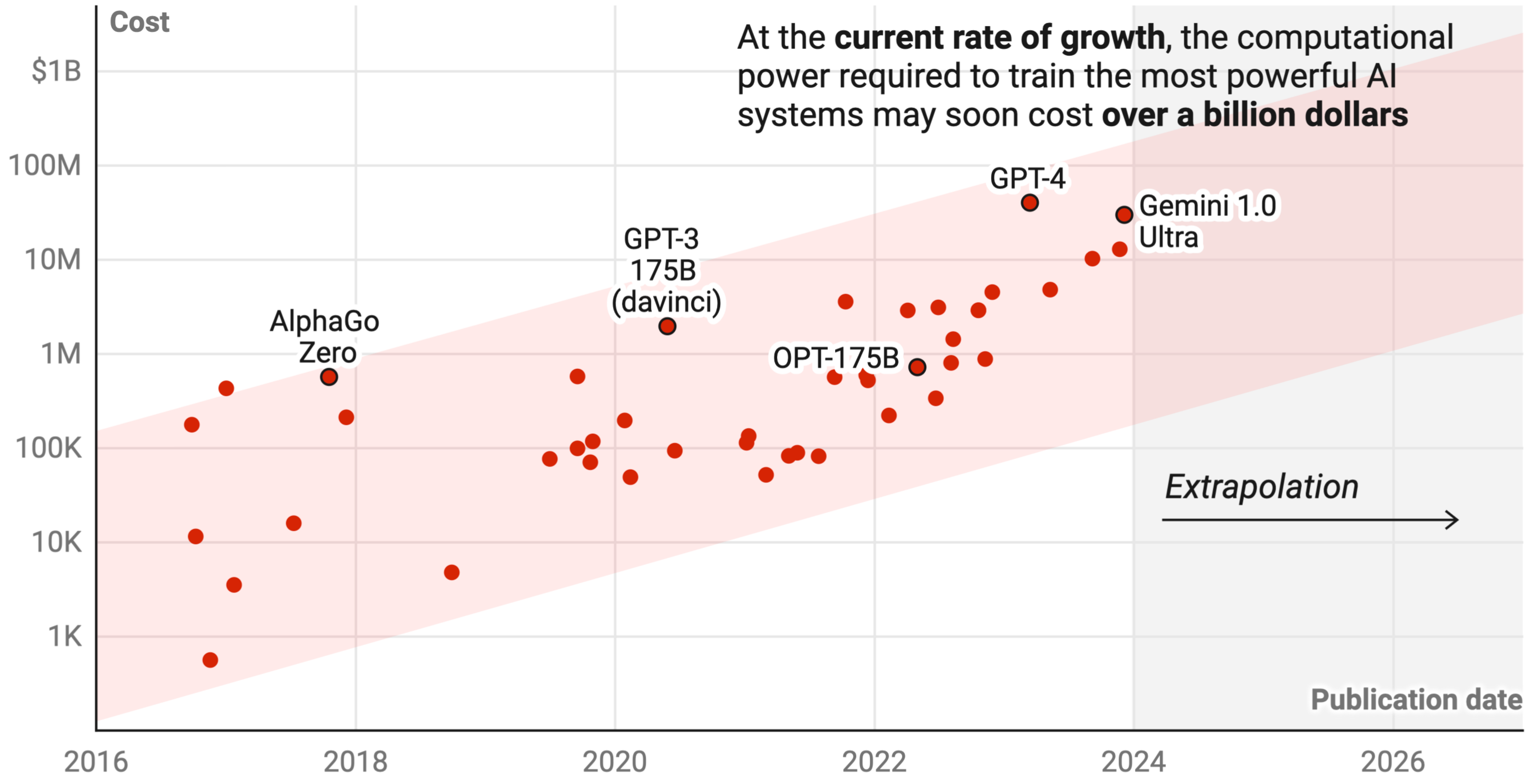


LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

Introduction



Cost of computational power required to train frontier AI systems

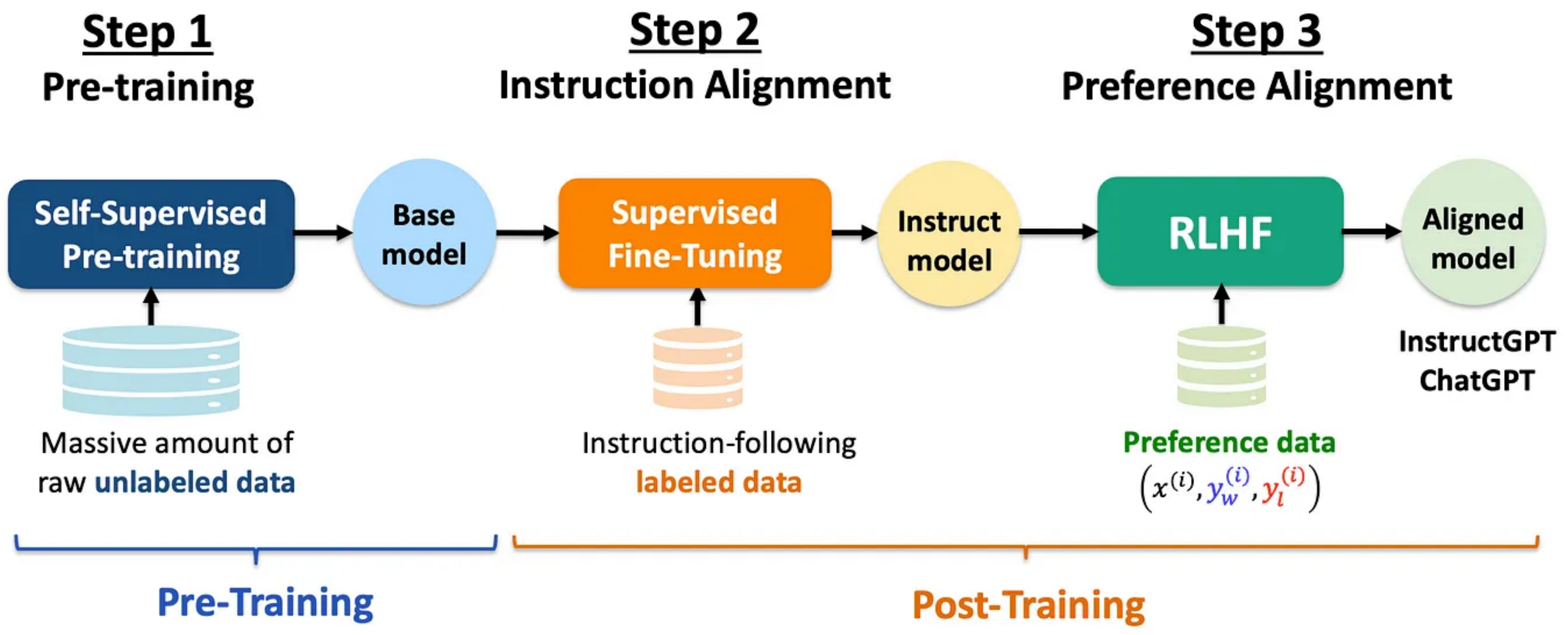


Cost includes amortized hardware acquisition and energy consumption. Red shaded area indicates 95% confidence prediction interval.

Chart: Will Henshall for TIME • Source: Epoch AI • Get the data • Created with Datawrapper

LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

Introduction



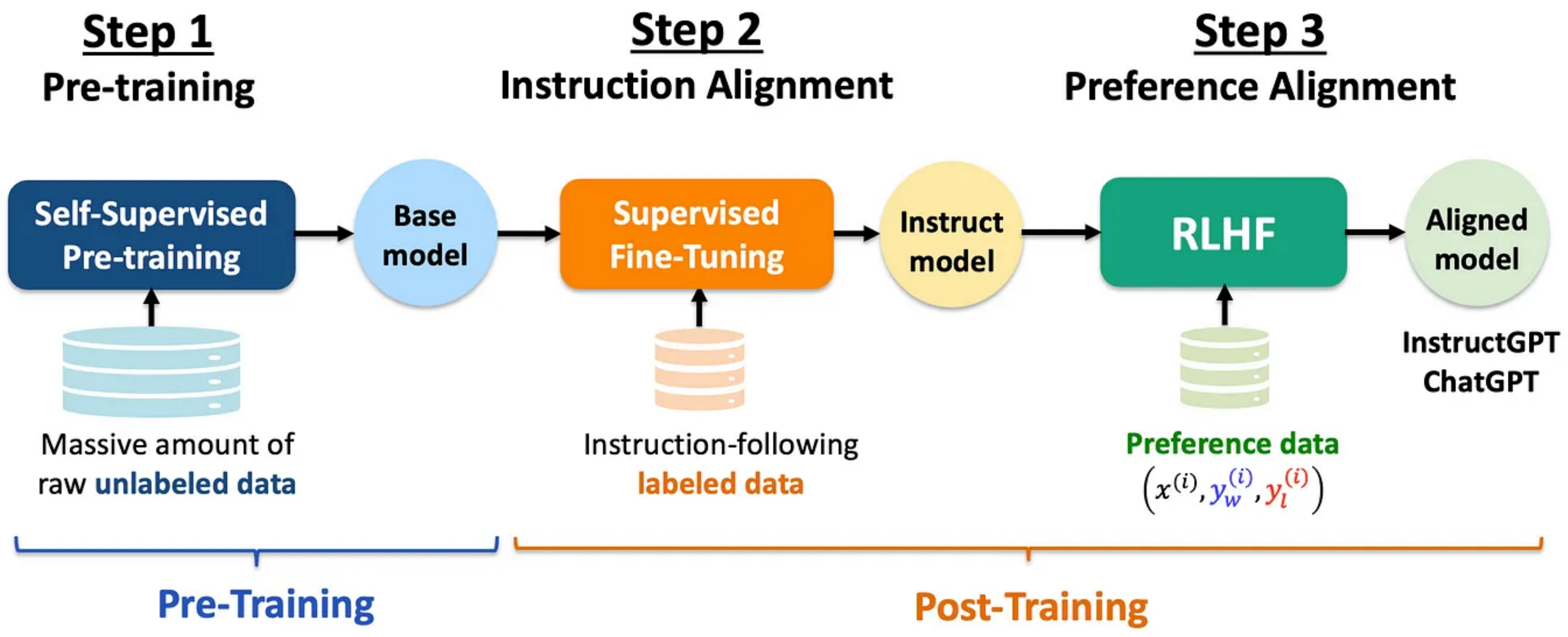
Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

T Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020

Introduction



Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

DeepSeek-V3: Technical report

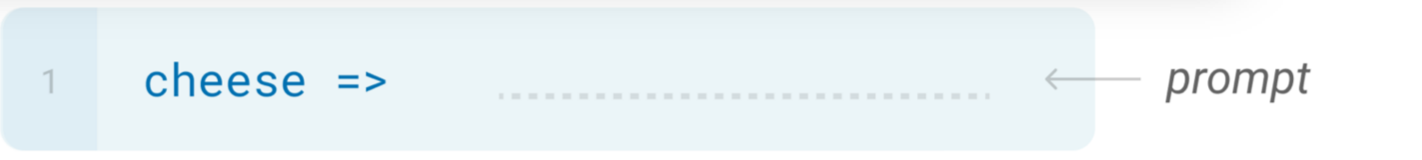
Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

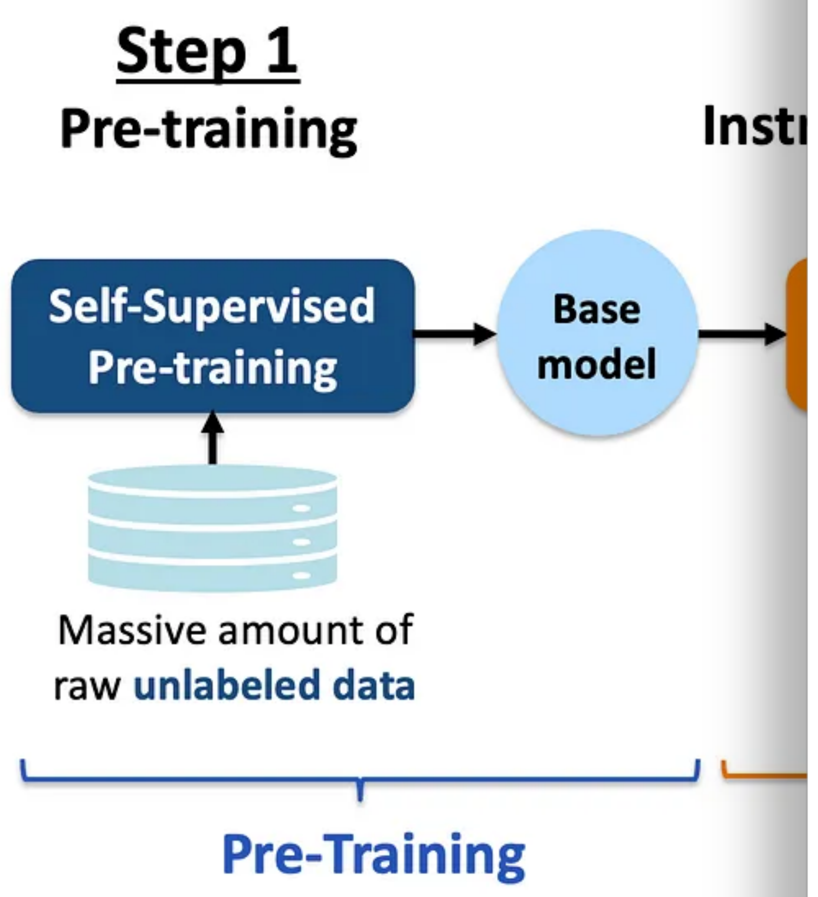
LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

Introduction

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.





DeepSeek-V3: Technical report

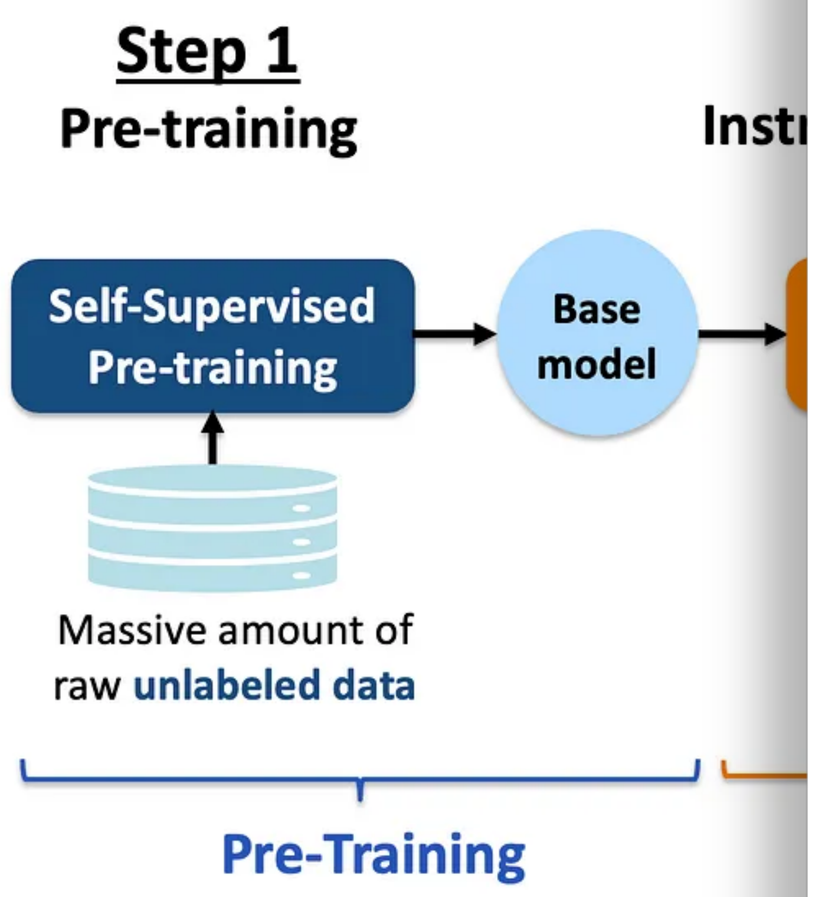
Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

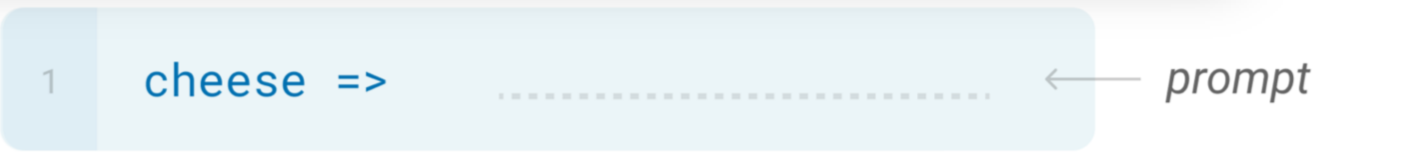
LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

Introduction

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.





DeepSeek-V3: Technical report

Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

LM Po @ Medium, A Brief History of LLMs: From Transformers (2017) to DeepSeek-R1 (2025)

Introduction

In-Context Learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

task description (line 1) · *prompt* (line 2)

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

task description (line 1) · *example* (line 2) · *prompt* (line 3)

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

task description (line 1) · *examples* (lines 2–4) · *prompt* (line 5)

T Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020

Introduction

In-Context Learning

Zero-shot

The model predicts t
description of the tas

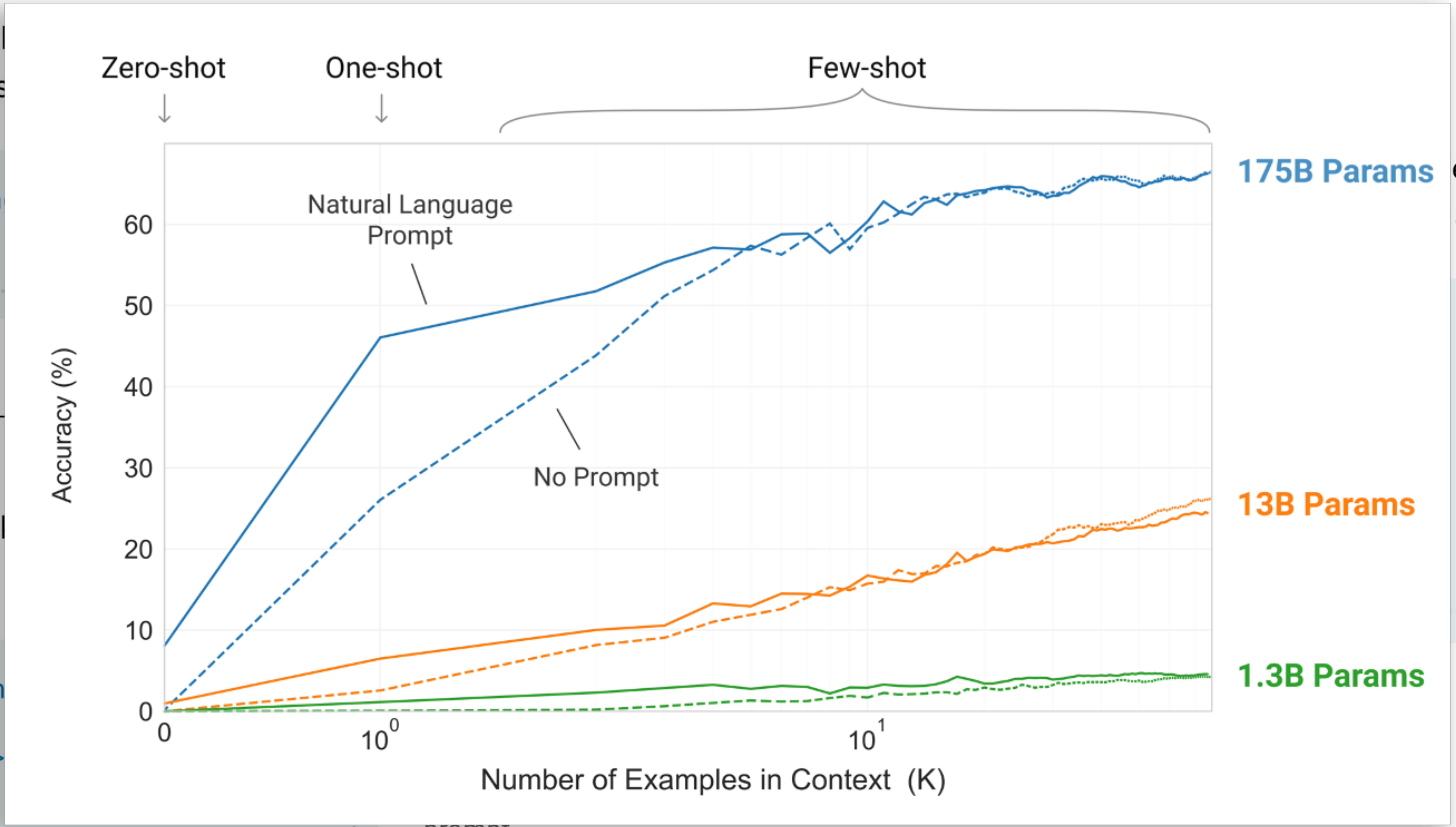
```
Translate En
cheese =>
```

One-shot

In addition to the tasl
example of the task.

```
Translate En
sea otter =>
cheese =>
```

prompt



el sees a few
are performed.

task description
examples
prompt

T Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020

Related Work

emergent abilities of LLMs

[19] Dong Q, Li L, Dai D, et al. A Survey on In-context Learning[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 1107-1128.

[20] Lu S, Bigoulaeva I, Sachdeva R S, et al. Are Emergent Abilities in Large Language Models just In-Context Learning?[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024: 5098-5139.

[25] Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models[J]. Transactions on Machine Learning Research, 2022.

non-transformer models

[22] Lee I, Jiang N, Berg-Kirkpatrick T. Is attention required for ICL? Exploring the Relationship Between Model Architecture and In-Context Learning Ability[A]. 2024. arXiv: 2310.08049.

[23] Park J, Park J, Xiong Z, et al. Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks[C]//Proceedings of the 41st International Conference on Machine Learning. 2024: 39793-39812.

[24] Tong W L, Pehlevan C. MLPs Learn In-Context on Regression and Classification Tasks[A]. 2024. arXiv: 2405.15618.

Related Work

experimental work

[26] Wei J, Bosma M, Zhao V, et al. Finetuned Language Models are Zero-Shot Learners[C]//International Conference on Learning Representations. 2022.

[27] Chan S, Santoro A, Lampinen A, et al. Data distributional properties drive emergent in-context learning in transformers[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022: 18878-18891.

[28] Min S, Lyu X, Holtzman A, et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?[A]. 2022. arXiv: 2202.12837.

[30] Garg S, Tsipras D, Liang P S, et al. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022: 30583-30598.

Related Work

there is still no consensus on how in-context learning works

implicit gradient descent to learn

[31] Von Oswald J, Niklasson E, Randazzo E, et al. Transformers Learn In-Context by Gradient Descent[C]. International Conference on Machine Learning, ICML, 2023.

[32] Dai D, Sun Y, Dong L, et al. Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers[C]//ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models. 2023.

[33] Ahn K, Cheng X, Daneshmand H, et al. Transformers learn to implement preconditioned gradient descent for in-context learning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023: 45614 - 45650.

[34] Cheng X, Chen Y, Sra S. Transformers Implement Functional Gradient Descent to Learn Non-Linear Functions In Context[C]//Proceedings of the 41st International Conference on Machine Learning: Vol. 235. 2024: 8002-8037.

Related Work

there is still no consensus on how in-context learning works

implicit gradient descent to learn

中国人民大学 RENMIN UNIVERSITY OF CHINA　高瓴人工智能学院 Gaoling School of Artificial Intelligence

大模型机理分析

--ICL、数据合成和慢思考

刘 勇

中国人民大学高瓴人工智能学院

$$linearAtten(\boldsymbol{V},\boldsymbol{K},\boldsymbol{q}) = \boldsymbol{V}\boldsymbol{K}^T\boldsymbol{q} = \sum_{i=1}^{N} \boldsymbol{v}_i(\boldsymbol{k}_i^T\boldsymbol{q}) = \left(\sum_{i=1}^{N} \boldsymbol{v}_i \otimes \boldsymbol{k}_i\right)\boldsymbol{q}$$

- 一种直觉是将其视为隐式梯度更新
- 考虑一个简单的线性模型$f(\boldsymbol{x};\boldsymbol{W}) = \boldsymbol{W}\boldsymbol{x}$
  - 训练数据$D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N, \mathcal{L}_D = \frac{1}{n}\sum_{i=1}^N \ell(f(\boldsymbol{x}_i;\boldsymbol{W}), \boldsymbol{y}_i)$　　$\boldsymbol{e}_i = -\eta \frac{\partial \mathcal{L}_D}{\partial f(\boldsymbol{x}_i)}$
  - 更新参数

$$\widehat{\boldsymbol{W}} = \boldsymbol{W} + \Delta\boldsymbol{W} = \boldsymbol{W} - \eta\frac{\partial \mathcal{L}_D}{\partial \boldsymbol{W}} = \boldsymbol{W} - \eta\sum_{i=1}^{N}\frac{\partial \mathcal{L}_D}{\partial f(\boldsymbol{x}_i)}\frac{\partial f(\boldsymbol{x}_i)}{\partial \boldsymbol{W}} = \boldsymbol{W} + \sum_{i=1}^{N} \boldsymbol{e}_i \otimes \boldsymbol{x}_i$$

  - 在新的测试点$\boldsymbol{x}_{test}$:

$$f(\boldsymbol{x}_{test};\widehat{\boldsymbol{W}}) = \boldsymbol{W}\boldsymbol{x}_{test} + \left(\sum_{i=1}^{N} \boldsymbol{e}_i \otimes \boldsymbol{x}_i\right)\boldsymbol{x}_{test} = \boldsymbol{W}\boldsymbol{x}_{test} + linearAtten(\boldsymbol{E},\boldsymbol{X},\boldsymbol{x}_{test})$$

其中 $\boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N), \boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$

输入: $\boldsymbol{H} = [\boldsymbol{H}_D, \boldsymbol{h}_{N+1}]$

ICL输出 → $\hat{\boldsymbol{h}}_{N+1} = W_V\boldsymbol{H} softmax\left(\frac{(W_K\boldsymbol{H})^T W_Q\boldsymbol{h}_{N+1}}{\sqrt{d_{out}}}\right)$

基于梯度下降算法的模型输出 → $\hat{\boldsymbol{y}}_{test} = f(W_Q\boldsymbol{h}_{N+1}) = \widehat{\boldsymbol{W}}\phi(W_Q\boldsymbol{h}_{N+1})$　　$\widehat{\boldsymbol{W}} = \boldsymbol{W} - \eta\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}}$

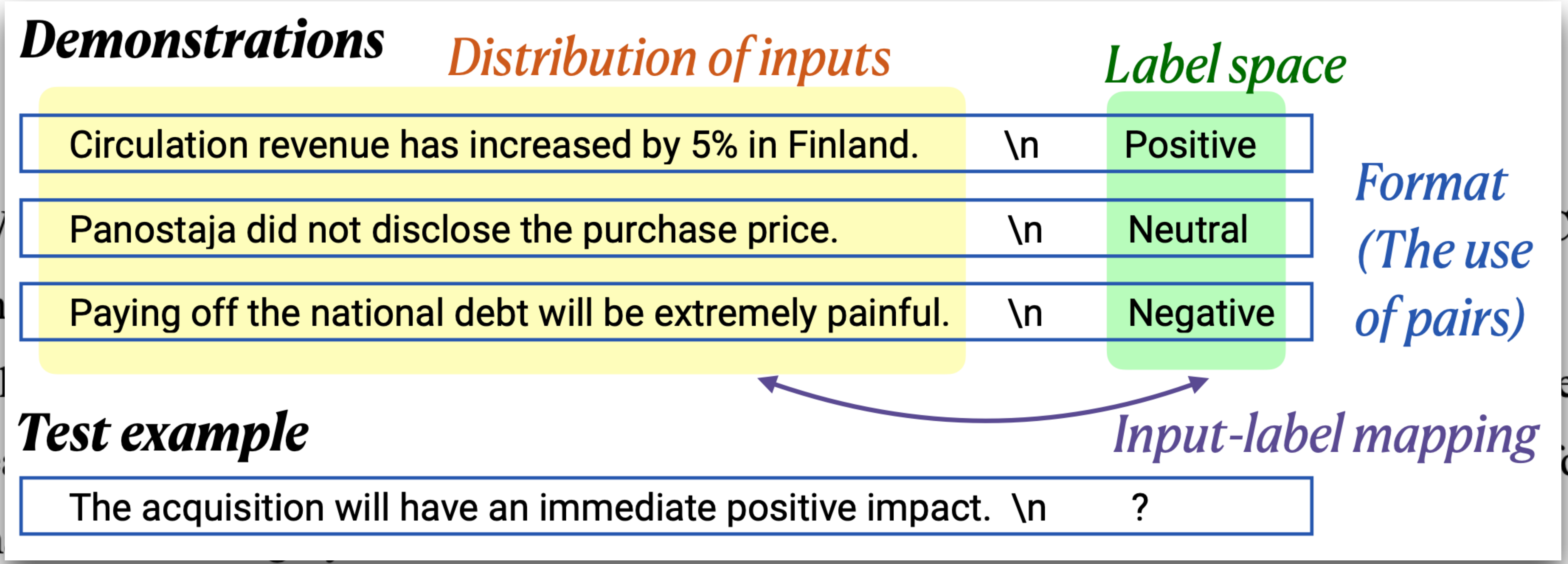
THEOREM 1. *The last token $\hat{\boldsymbol{h}}_{N+1}$ obtained through ICL* https://arxiv.org/abs/2303.07971 *s strictly equivalent to the test prediction $\hat{\boldsymbol{y}}_{test}$ obtained by performing one step of gradient descent on the weight $\boldsymbol{W}$ in the reference model $f(\boldsymbol{x}) = \boldsymbol{W}\phi(\boldsymbol{x})$. The form of the loss function $\mathcal{L}$ is:*

$$\mathcal{L} = -\frac{1}{\eta D}\sum_{i=1}^{N}(W_V\boldsymbol{h}_i)^T \boldsymbol{W}\phi(W_K\boldsymbol{h}_i),$$ ← loss

where η is the learning rate and D is a constant.

Ruifeng Ren, Yong Liu. Towards Understanding How Transformers Learn In-context Through a Representation Learning Lens. In NeurIPS 2024

Related Work



[26] W ...]//
In
[27] Cl ... ext
le ... or-
m
[28] Min S, Lyu X, Holtzman A, et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?[A]. 2022. arXiv: 2202.12837.

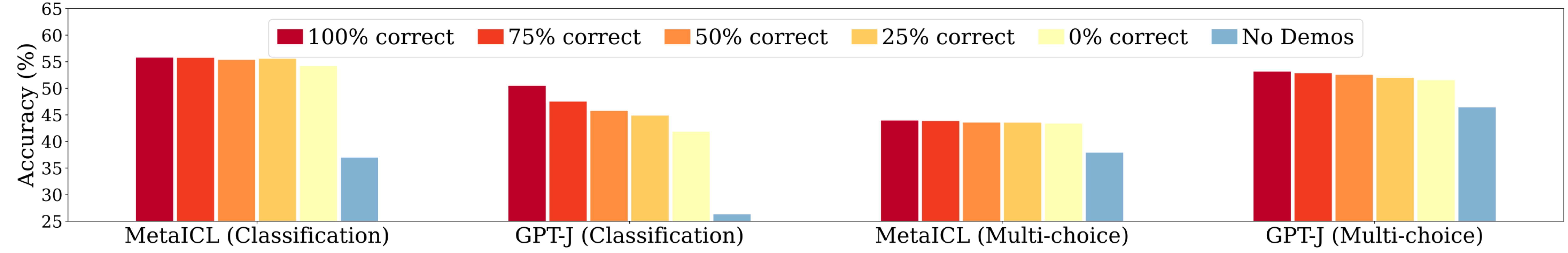


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

Related Work

there is still no consensus on how in-context learning works

implicit Bayesian inference

[37] Xie S M, Raghunathan A, Liang P, et al. An Explanation of In-context Learning as Implicit Bayesian Inference[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2021: 1017-1028.

[38] Panwar M, Ahuja K, Goyal N. In-Context Learning through the Bayesian Prism[C]//The Twelfth International Conference on Learning Representations. 2024.

[39] Wang X, Zhu W, Saxon M, et al. Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023: 15614 - 15638.

[40] Bai Y, Chen F, Wang H, et al. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023: 57125 - 57211.

Related Work

there is still no consensus on how in-context learning works

implicit Bayesian inference

[37] Xie S M, Raghunathan A, Liang P, et al. An Explanation of In-context Learning as Implicit Bayesian Inference[C]//Proceedings of the 2024 Conference of the North American Chapter of … olume 1: L…

[38] Pa… [C]//The Tw…

[39] W… Explain-in… 37th International Conference on Neural Information Processing Systems. 2023: 15614 - 15638.

[40] Bai Y, Chen F, Wang H, et al. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023: 57125 - 57211.

What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning

Jane Pan **Tianyu Gao** **Howard Chen** **Danqi Chen**
Department of Computer Science, Princeton University
{jp7224,tianyug,howardchen,danqic}@cs.princeton.edu

Related Work

there is still no consensus on how in-context learning works

other perspectives:

information theory, optimization algorithms, gradient flow dynamics ···

[41] Hahn M, Goyal N. A Theory of Emergent In-Context Learning as Implicit Structure Induction [A]. 2023. arXiv: 2303.07971.

[42] Li Y, Ildiz M E, Papailiopoulos D, et al. Transformers as Algorithms: Generalization and Stability in In-context Learning[C]//Proceedings of the 40th International Conference on Machine Learning. 2023: 19565 - 19594.

[43] Han C, Wang Z, Zhao H, et al. Explaining Emergent In-Context Learning as Kernel Regression [A]. 2023. arXiv: 2305.12766.

[44] Chen S, Sheen H, Wang T, et al. Unveiling Induction Heads: Provable Training Dynamics and Feature Learning in Transformers[C]//The Thirty-eighth Annual Conference on Neural Information Processing Systems. 2024.

Related Work

simple small models and parameter-controllable synthesis tasks

[45] Akyürek E, Schuurmans D, Andreas J, et al. What learning algorithm is in-context learning? Investigations with linear models[C]//The Eleventh International Conference on Learning Representations. 2023.

[46] Zhang R, Frei S, Bartlett P L. Trained transformers learn linear models in-context[J]. Journal of Machine Learning Research, 2024, 25(49): 1-55.

[47] Kim J, Suzuki T. Transformers Learn Nonlinear Features In Context: Nonconvex Mean-field Dynamics on the Attention Landscape[C]//Proceedings of the 41st International Conference on Machine Learning. 2024: 24527-24561.

linear model

[48] Reddy G. The mechanistic basis of data dependence and abrupt learning in an in-context classification task[C]//12th International Conference on Learning Representations. 2023.

[49] Li H, Wang M, Lu S, et al. How Do Nonlinear Transformers Learn and Generalize in In-Context Learning?[C]//Proceedings of the 41st International Conference on Machine Learning. 2024: 28734 - 28783.

classification task

[50] Raventós A, Paul M, Chen F, et al. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023: 14228-14246.

[51] Huang Y, Cheng Y, Liang Y. In-Context Convergence of Transformers[C]. 2023: 19660-19722.

regression task

Related Work

simple small models and parameter-controllable synthesis tasks

Asymptotic theory of in-context learning by linear attention

Yue M. Lu,[1,*] Mary Letey,[1,**] Jacob A. Zavatone-Veth,[1,2,3,4,†] Anindita Maiti,[5,‡] and Cengiz Pehlevan[1,2,6,§]

[1] *John A. Paulson School of Engineering and Applied Sciences, Harvard University*
[2] *Center for Brain Science, Harvard University*
[3] *Society of Fellows, Harvard University*
[4] *Department of Physics, Harvard University*
[5] *Perimeter Institute for Theoretical Physics*
[6] *Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University*

(Dated: February 6, 2025)

Training Dynamics of Multi-Head Softmax Attention for In-Context Learning: Emergence, Convergence, and Optimality

Siyu Chen Heejune Sheen Tianhao Wang Zhuoran Yang
Department of Statistics and Data Science, Yale University
{siyu.chen.sc3226, heejune.sheen, tianhao.wang, zhuoran.yang}@yale.edu

How Many Pretraining Tasks Are Needed for In-Context Learning of Linear Regression?

Jingfeng Wu
UC Berkeley
uuujf@berkeley.edu

Difan Zou
The University of Hong Kong
dzou@cs.hku.hk

Zixiang Chen
UCLA
chenzx19@cs.ucla.edu

Vladimir Braverman
Rice University
vb21@rice.edu

Quanquan Gu
UCLA
qgu@cs.ucla.edu

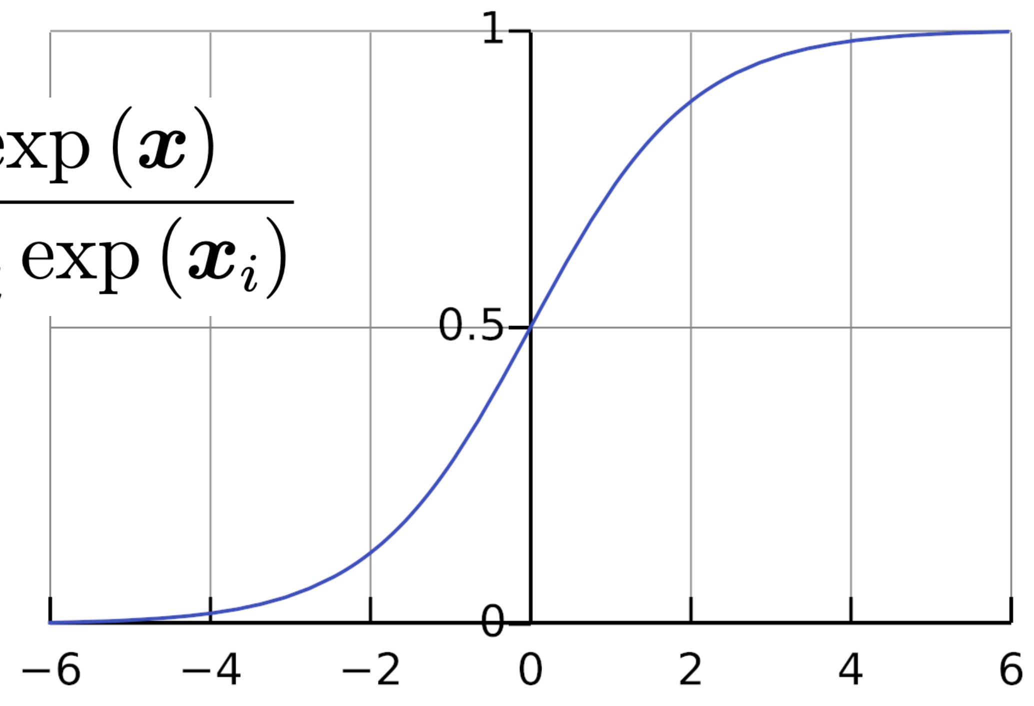
Peter L. Bartlett
Google DeepMind & UC Berkeley
peter@berkeley.edu

A Toy Model of In-Context Learning

Vanilla Attention

$$\mathbf{Y}(\mathbf{X}) = \mathbf{W}_{\mathrm{V}}\mathbf{X} \cdot \mathrm{softmax}\left(\frac{(\mathbf{W}_{\mathrm{Q}}\mathbf{X})^{\top}\mathbf{W}_{\mathrm{K}}\mathbf{X}}{\sqrt{d_k}}\right)$$

$$\mathrm{softmax}(\boldsymbol{x})_i = \frac{\exp(\boldsymbol{x})}{\sum_i \exp(\boldsymbol{x}_i)}$$



Yuhao Li, Ruoran Bai, and Haiping Huang. Spin Glass Model of In-Context Learning. Phys. Rev. E 112, L013301. 2025

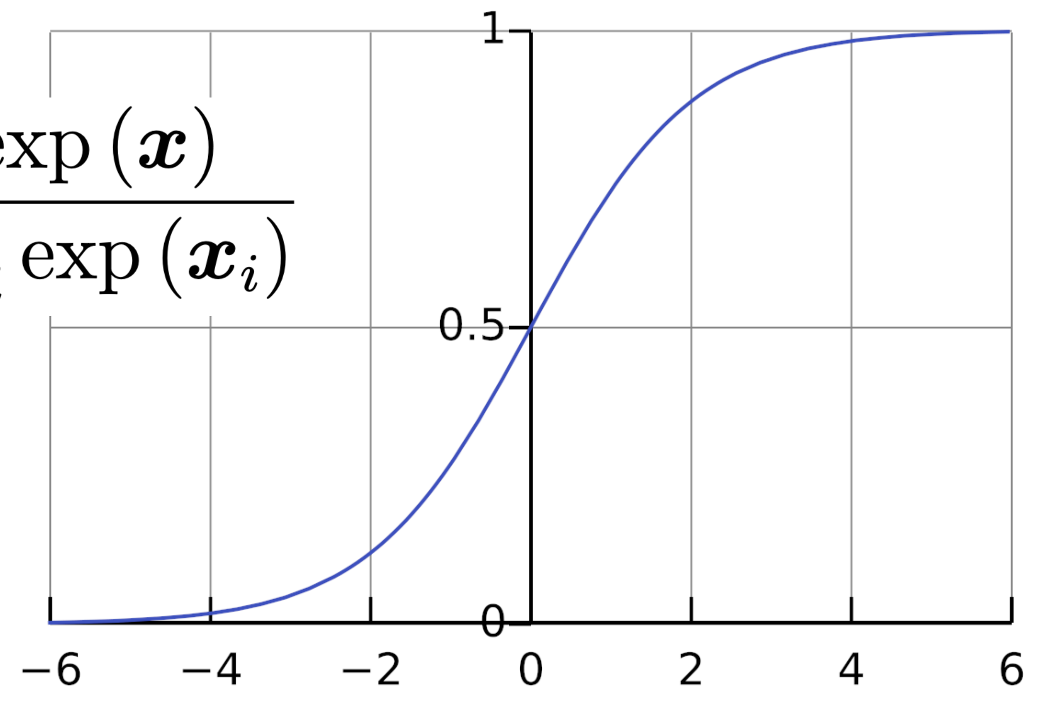
A Toy Model of In-Context Learning

Vanilla Attention

$$\mathbf{Y}(\mathbf{X}) = \mathbf{W}_{\mathrm{V}}\mathbf{X} \cdot \text{softmax}\left(\frac{(\mathbf{W}_{\mathrm{Q}}\mathbf{X})^{\top}\mathbf{W}_{\mathrm{K}}\mathbf{X}}{\sqrt{d_k}}\right)$$

$$\mathbf{Y}(\mathbf{X}) = \frac{1}{N}\mathbf{W}_{\mathrm{V}}\mathbf{X}\mathbf{X}^{\top}\mathbf{W}_{\mathrm{Q}}^{\top}\mathbf{W}_{\mathrm{K}}\mathbf{X}$$

$$\text{softmax}(\boldsymbol{x})_i = \frac{\exp(\boldsymbol{x})}{\sum_i \exp(\boldsymbol{x}_i)}$$



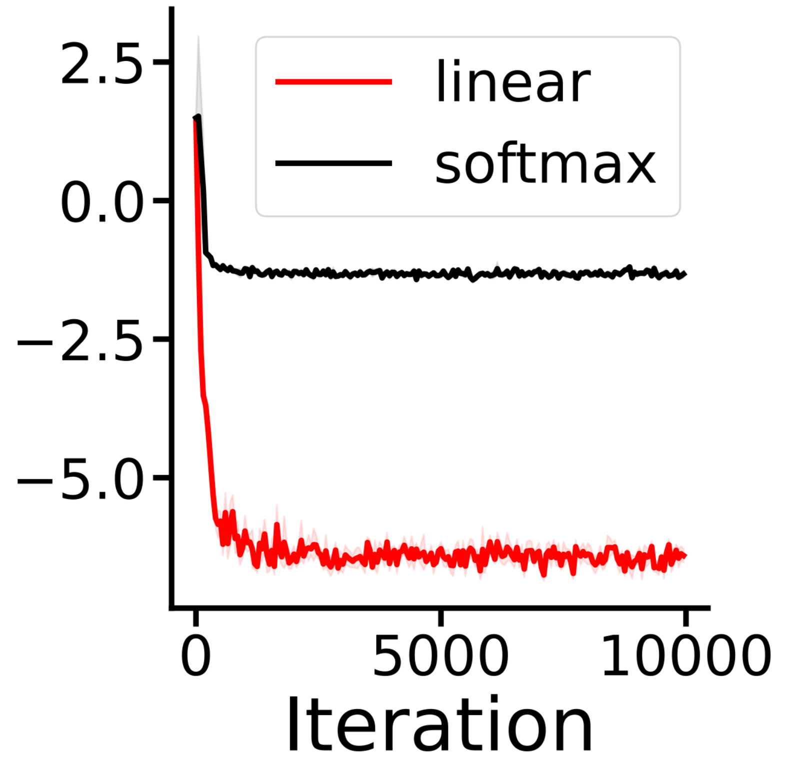


Figure 7: log(loss) against iteration. Comparison between linear attention and softmax attention for the 3-layer Transformers. Note that the loss of linear Transformer decreases much faster.

K. Ahn et al. ICLR 2024 poster

Yuhao Li, Ruoran Bai, and Haiping Huang. Spin Glass Model of In-Context Learning. Phys. Rev. E 112, L013301. 2025

A Toy Model of In-Context Learning

Vanilla Attention

$$\mathbf{Y}(\mathbf{X}) = \mathbf{W}_\mathrm{V}\mathbf{X} \cdot \mathrm{softmax}\left(\frac{(\mathbf{W}_\mathrm{Q}\mathbf{X})^\top \mathbf{W}_\mathrm{K}\mathbf{X}}{\sqrt{d_k}}\right)$$

$$\mathbf{Y}(\mathbf{X}) = \frac{1}{N}\mathbf{W}_\mathrm{V}\mathbf{X}\mathbf{X}^\top\mathbf{W}_\mathrm{Q}^\top\mathbf{W}_\mathrm{K}\mathbf{X}$$

$$\mathbf{W}_\mathrm{V} = \mathbf{I}$$

$$\mathbf{W} \equiv \mathbf{W}_\mathrm{Q}^\top\mathbf{W}_\mathrm{K}$$

$$\mathbf{Y}(\mathbf{X}) = \frac{1}{N}\mathbf{X}\mathbf{X}^\top\mathbf{W}\mathbf{X}$$

YM Lu et al. Asymptotic theory of in-context learning by linear attention. PNAS 2024

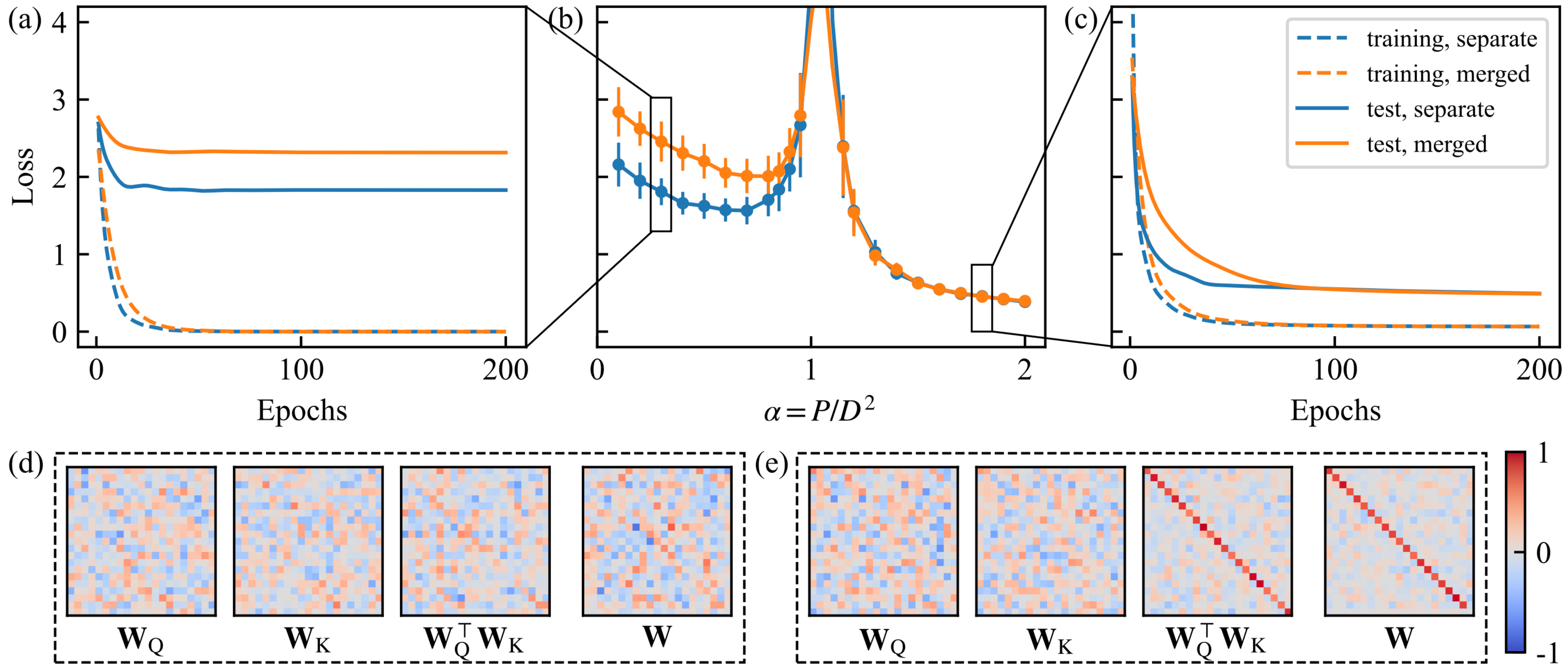
$$A := Z + \frac{1}{\ell} V Z (KZ)^\top (QZ)$$

$$V = \begin{bmatrix} V_{11} & v_{12} \\ v_{21}^\top & v_{22} \end{bmatrix}$$

$$\hat{y} = \langle \Gamma, H_Z \rangle$$

$$\Gamma := v_{22}\left[M_{11}^\top/d \;\; m_{21}\right]$$

$$H_Z := x_{\ell+1}\left[\tfrac{d}{\ell}\sum_{i\le\ell} y_i x_i^\top \;\; \tfrac{1}{\ell}\sum_{i\le\ell} y_i^2\right]$$



Yuhao Li, Ruoran Bai, and Haiping Huang. Spin Glass Model of In-Context Learning. Phys. Rev. E 112, L013301. 2025

A Toy Model of In-Context Learning

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese => ......................    ← prompt
```

0.1 x + 0.2 y = z

1	5	==>	1.1
2	6	==>	1.4
3	9	==>	2.1
4	1	==>	?

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} & \tilde{x}_1 \\ x_{21} & x_{22} & \cdots & x_{2N} & \tilde{x}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{D1} & x_{D2} & \cdots & x_{DN} & \tilde{x}_D \\ y_1 & y_2 & \cdots & y_N & 0 \end{bmatrix}$$

question

answer need to be predicted

N examples

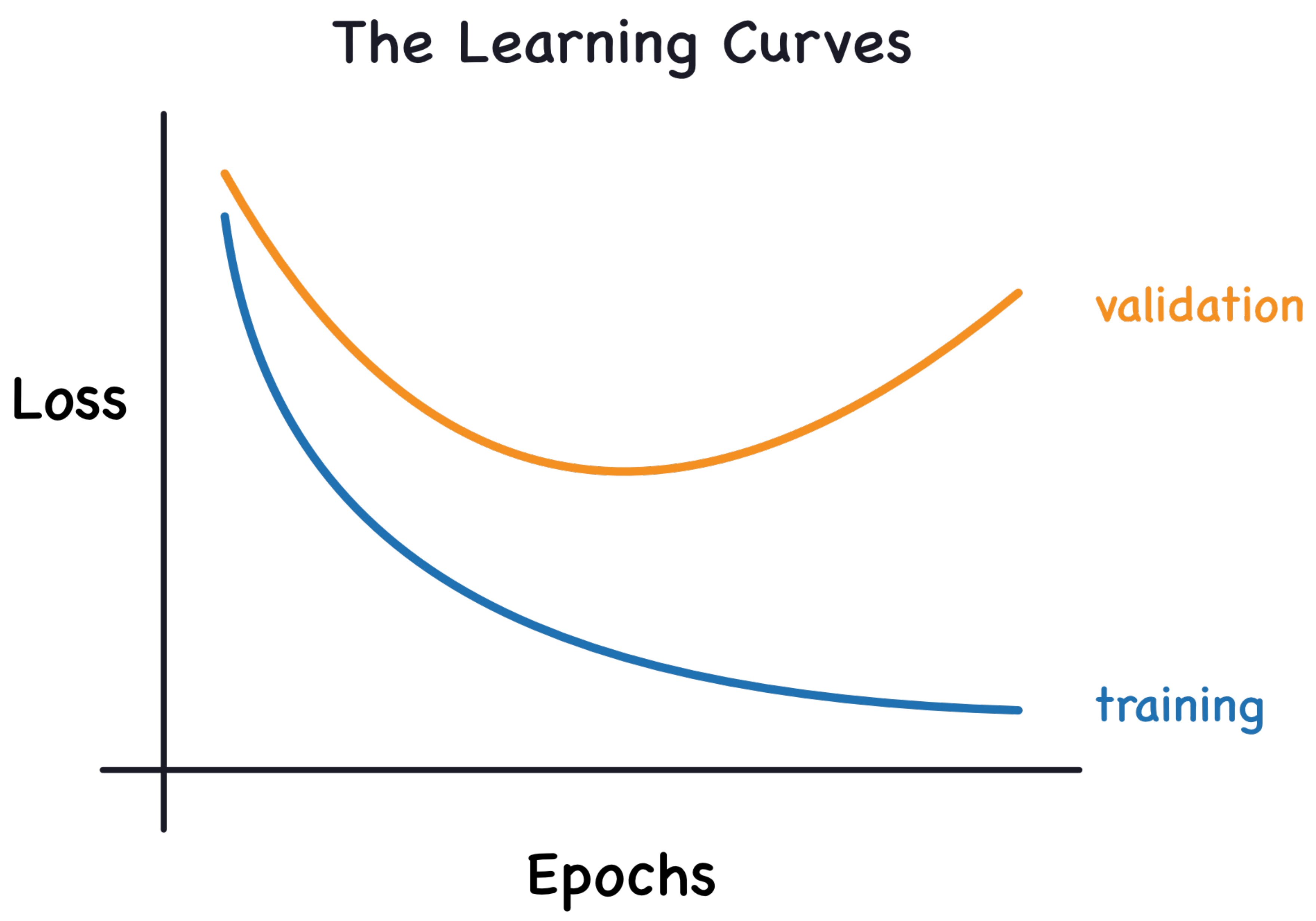
label

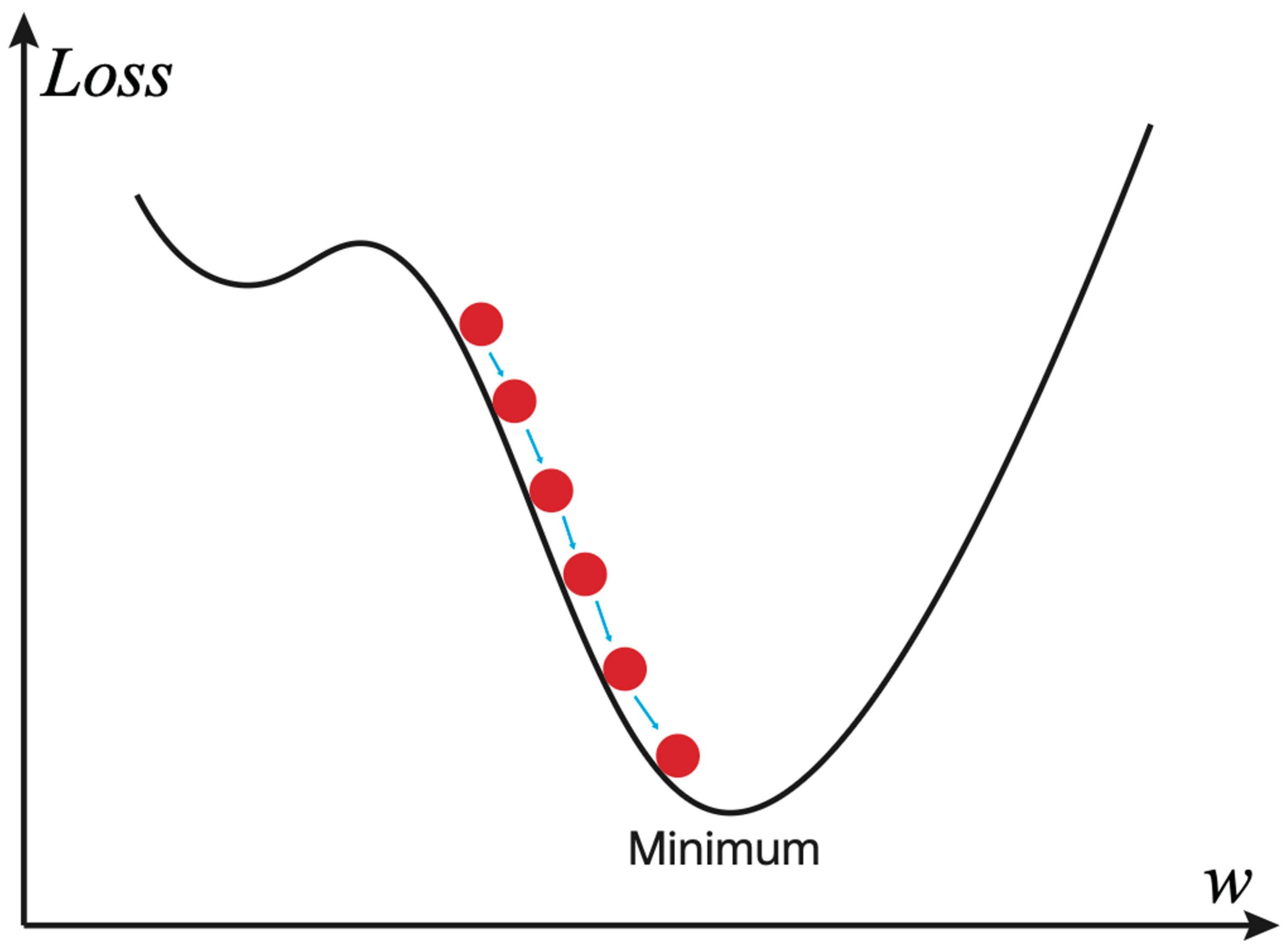
$$y = \frac{1}{\sqrt{D}} \sum_i x_i w_i \quad x_i \sim \mathcal{N}(0,1) \quad w_i \sim \mathcal{N}(0,1)$$

prediction

$$\hat{y} = \mathbf{Y}_{D+1,N+1}$$

Mapping to Spin Glass Model





Loss Function

$$\mathcal{L} = \frac{1}{2P}\sum_{\mu=1}^{P}(\hat{y}^{\mu} - \tilde{y}^{\mu})^2 + \frac{\lambda}{2}\|\mathbf{W}\|^2$$

Data ⟷ Interaction & field

Parameter ⟷ Spin

Training Loss ⟷ Hamiltonian

Mapping to Spin Glass Model

the prediction $\hat{y}^{\mu} = \mathbf{Y}_{D+1,N+1} = \frac{1}{N}\sum_{m,n,i} \mathbf{X}^{\mu}_{D,i}\mathbf{X}^{\mu}_{m,i}\mathbf{W}_{m,n}\mathbf{X}^{\mu}_{n,N+1}$

data

weight

$$\mathbf{S}^{\mu}_{mn} = \frac{1}{N}\sum_{i} \mathbf{X}^{\mu}_{D+1,i}\mathbf{X}^{\mu}_{m,i}\mathbf{X}^{\mu}_{n,N+1}$$

Vectorization

Vectorization

$$\boldsymbol{s}^{\mu} = \mathrm{vec}\left(\mathbf{S}^{\mu}\right) \qquad \boldsymbol{\sigma} = \mathrm{vec}(\mathbf{W})$$

rewrite as $\hat{y}^{\mu} = \sum_i \sigma_i s^{\mu}_i$

Mapping to Spin Glass Model

the prediction $\hat{y}^\mu = \sum_i \sigma_i s_i^\mu$

loss function

$$\mathcal{L} = \frac{1}{2P}\sum_{\mu=1}^{P}(\hat{y}^\mu - \tilde{y}^\mu)^2 + \frac{\lambda}{2}\|\mathbf{W}\|^2$$

$$= \frac{1}{2P}\sum_{\mu,i,j} s_i^\mu s_j^\mu \sigma_i\sigma_j - \frac{1}{P}\sum_{\mu,i}\tilde{y}^\mu s_i^\mu \sigma_i + \frac{1}{2}\lambda\sum_i \sigma_i^2$$

interaction

external field

regularization factor

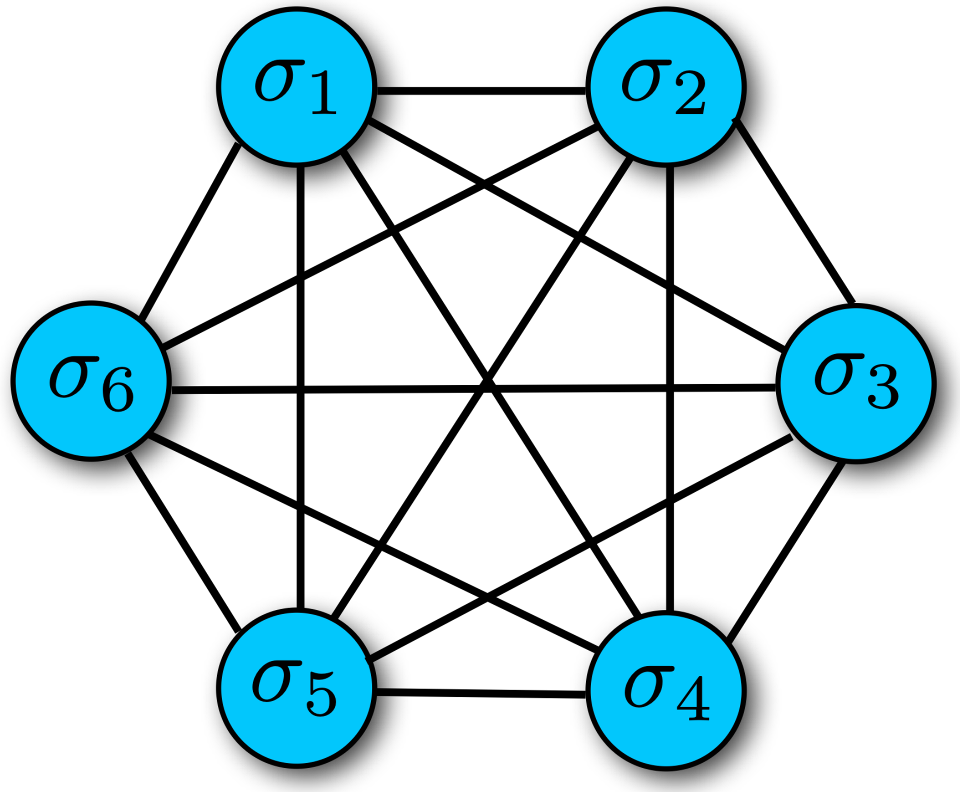
$$J_{ij} = -\frac{1}{P}\sum_\mu s_i^\mu s_j^\mu \qquad h_i = \frac{1}{P}\sum_\mu \tilde{y}^\mu s_i^\mu \qquad \lambda_i = \frac{1}{P}\sum_\mu (\lambda - J_{ii}^\mu)$$

Hamiltonian

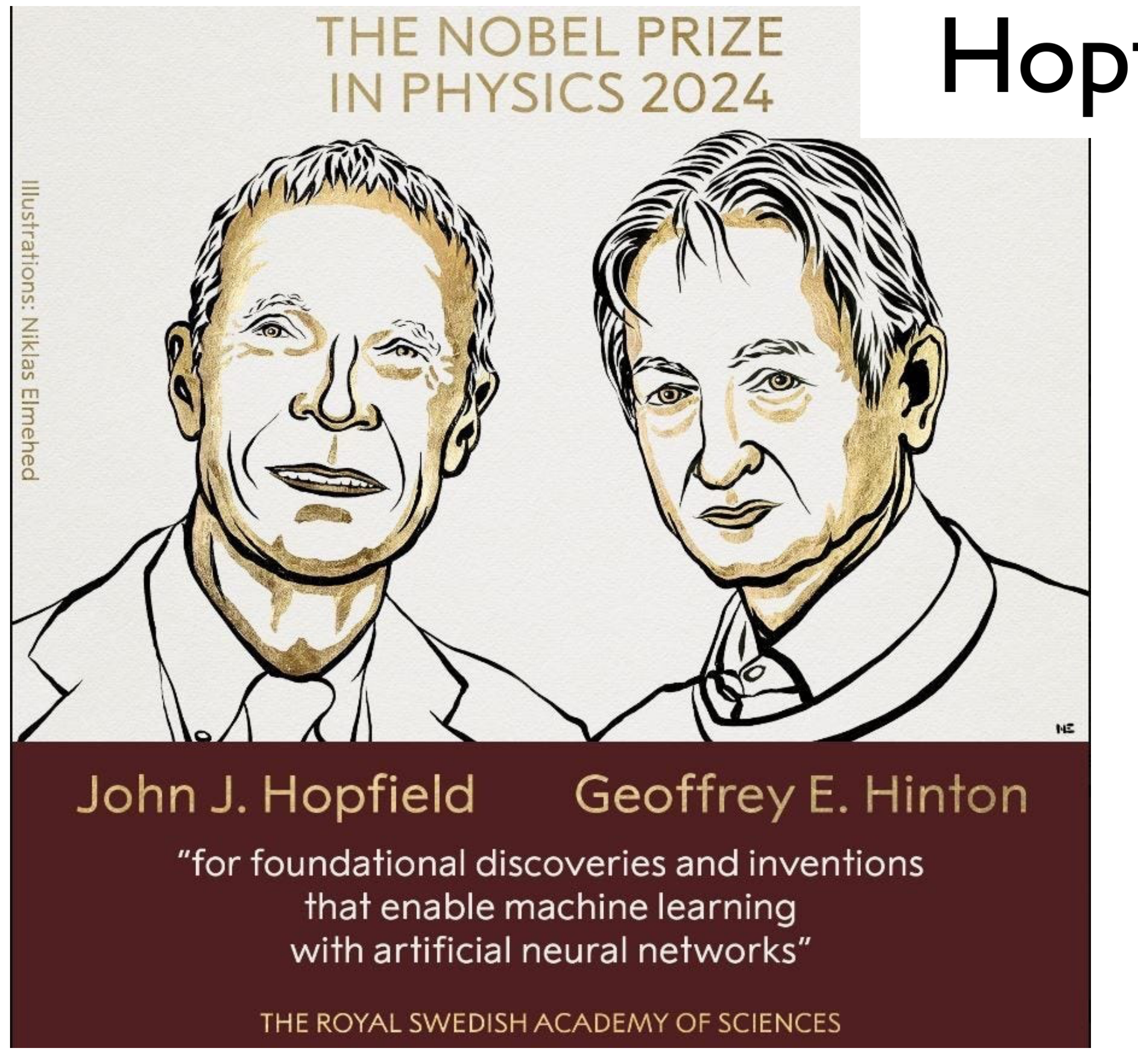
$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i + \frac{1}{2}\sum_i \lambda_i\sigma_i^2$$

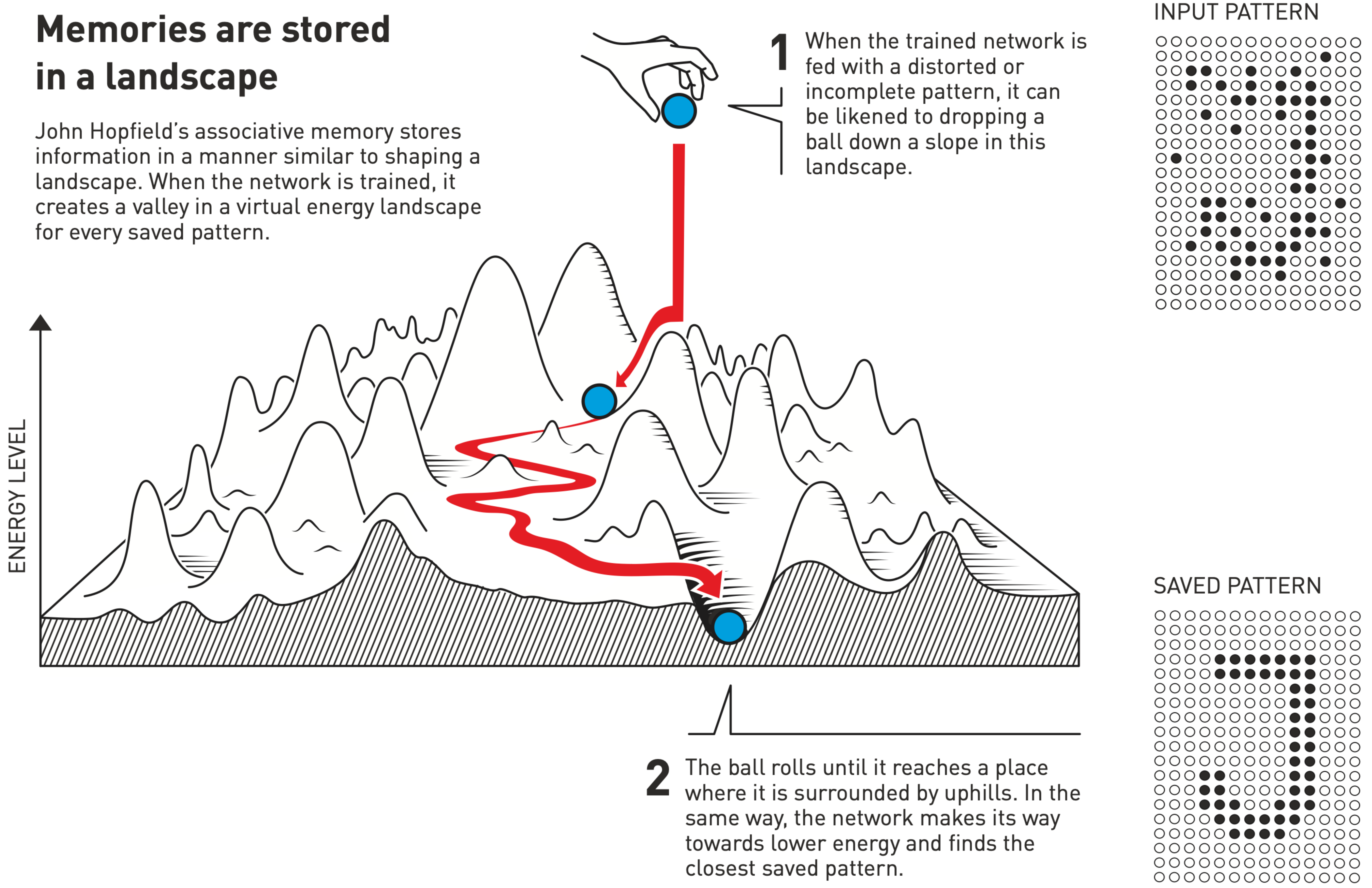
Mapping to Spin Glass Model

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i + \frac{1}{2}\sum_i \lambda_i\sigma_i^2$$



Hopfield Network

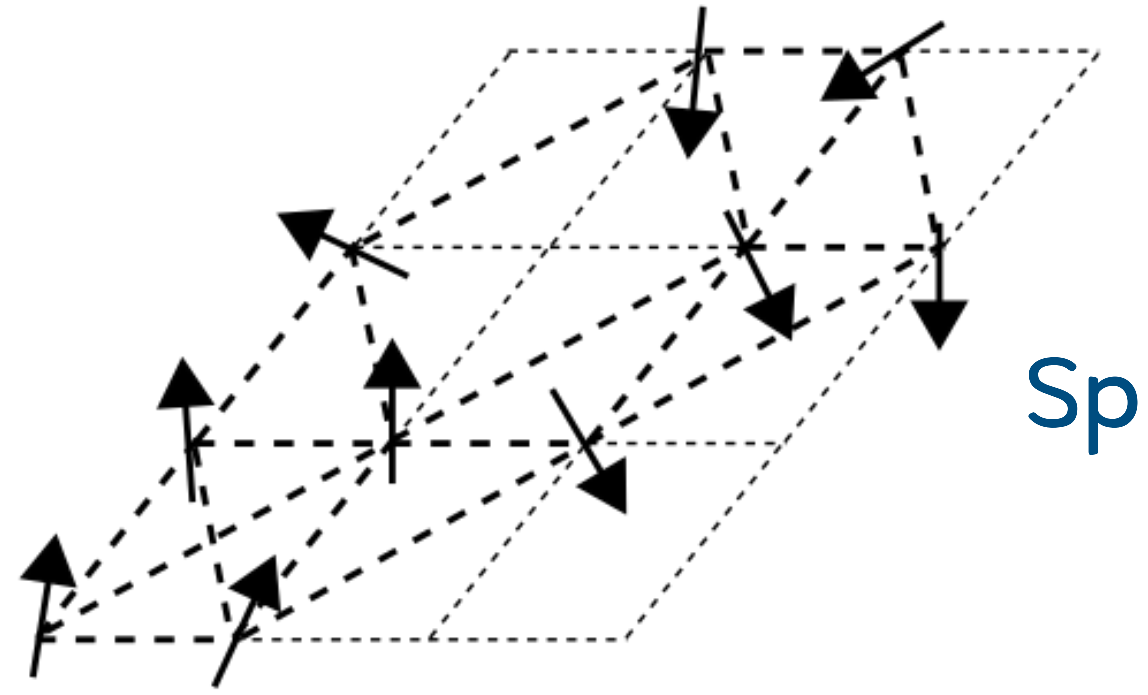




Popular Science Background @ THE NOBEL PRIZE IN PHYSICS 2024

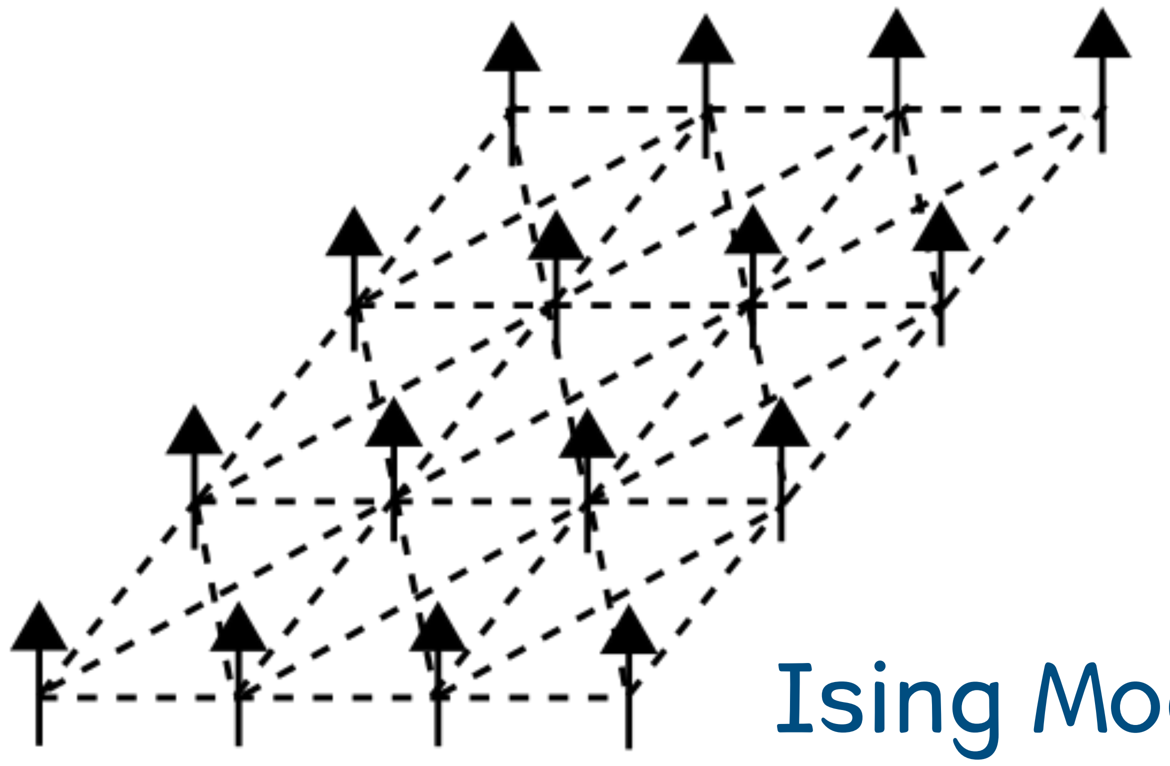
Mapping to Spin Glass Model

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i + \frac{1}{2}\sum_i \lambda_i\sigma_i^2$$

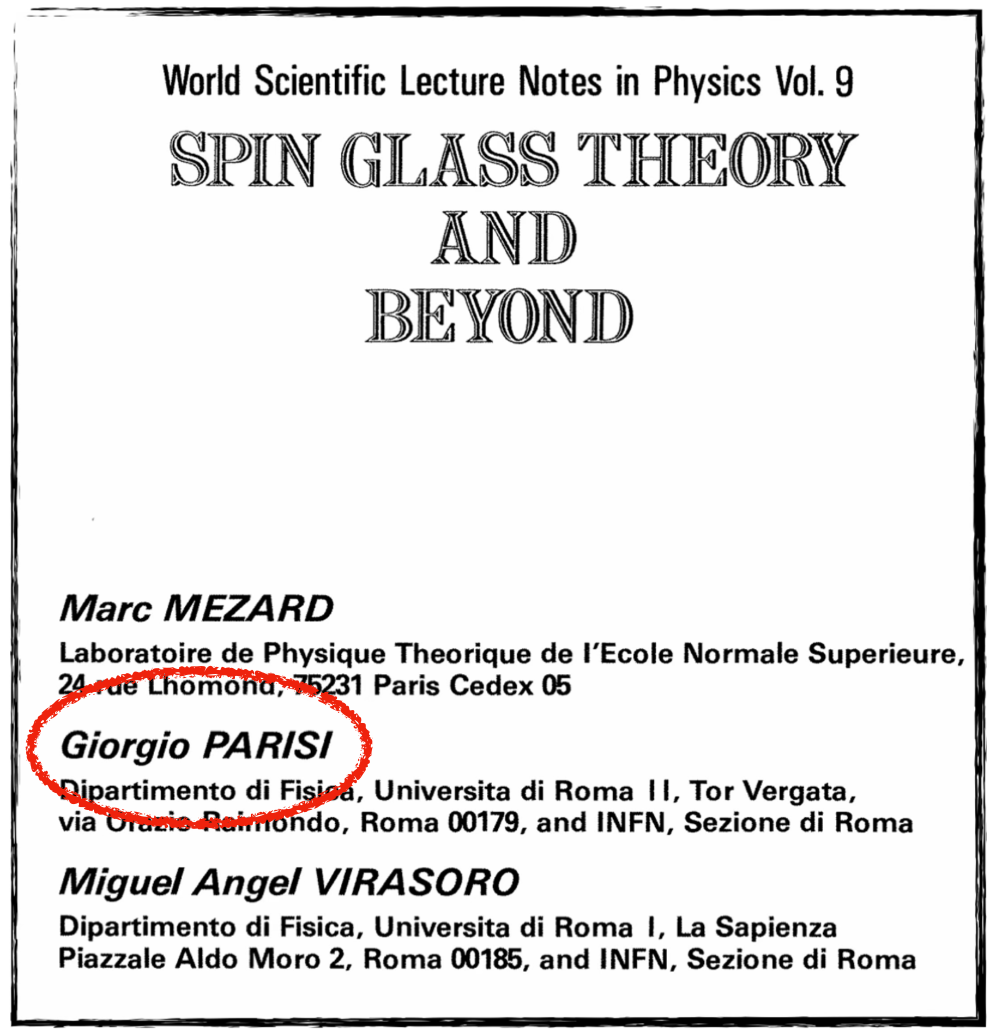


Spin Glass Model

(SK Model)



Ising Model



World Scientific Lecture Notes in Physics Vol. 9

SPIN GLASS THEORY AND BEYOND

Marc MEZARD

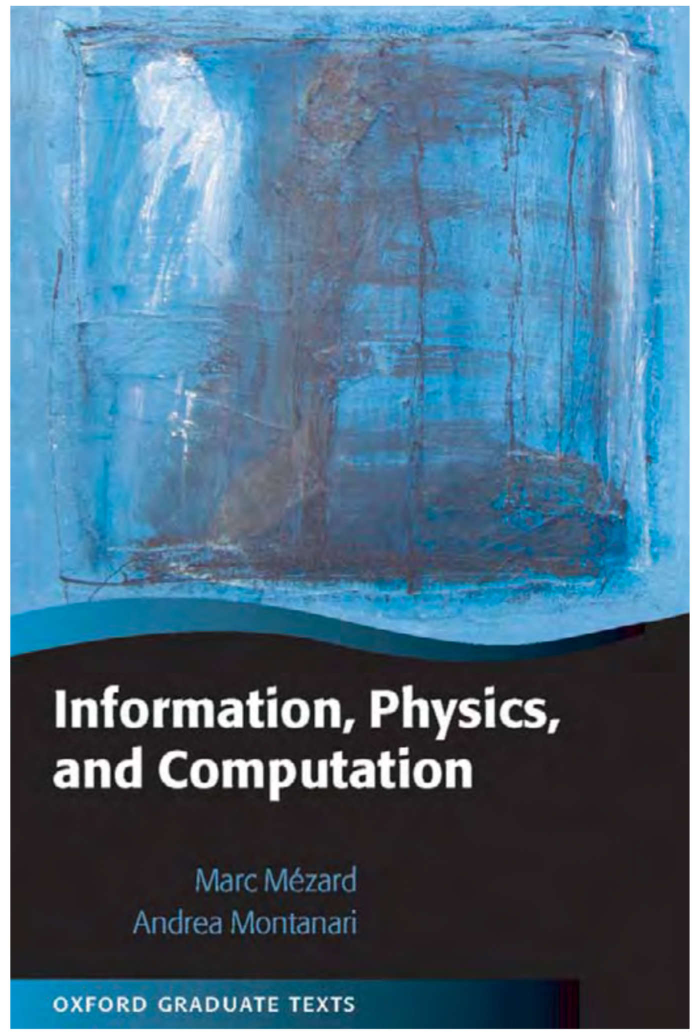
Laboratoire de Physique Theorique de l'Ecole Normale Superieure, 24 rue Lhomond, 75231 Paris Cedex 05

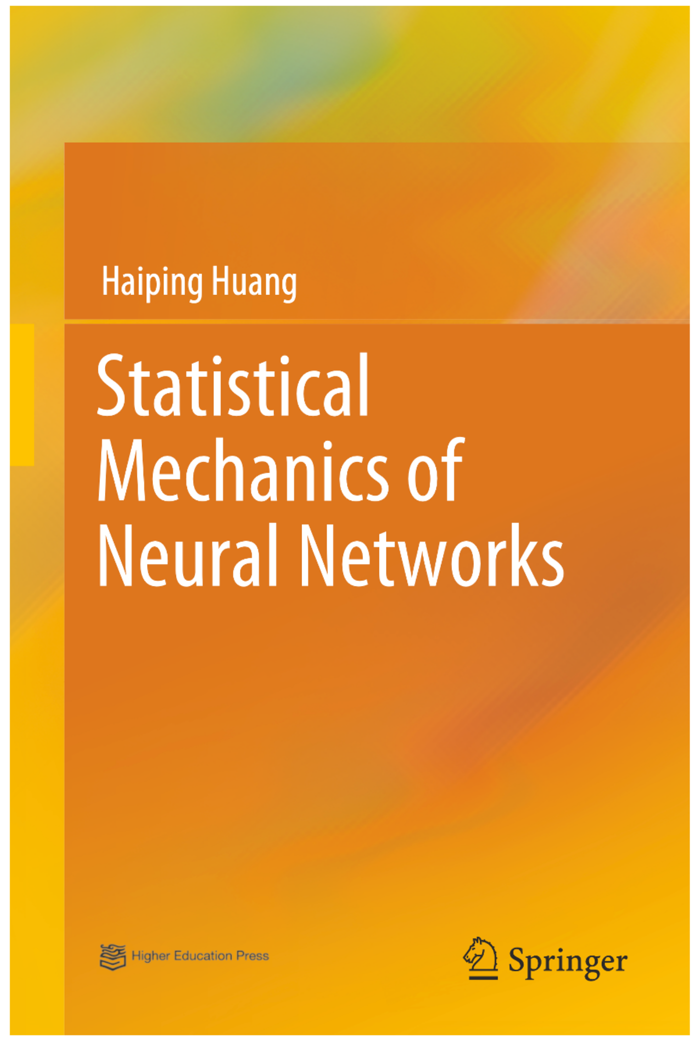
Giorgio PARISI

Dipartimento di Fisica, Universita di Roma II, Tor Vergata, via Orazio Raimondo, Roma 00179, and INFN, Sezione di Roma

Miguel Angel VIRASORO

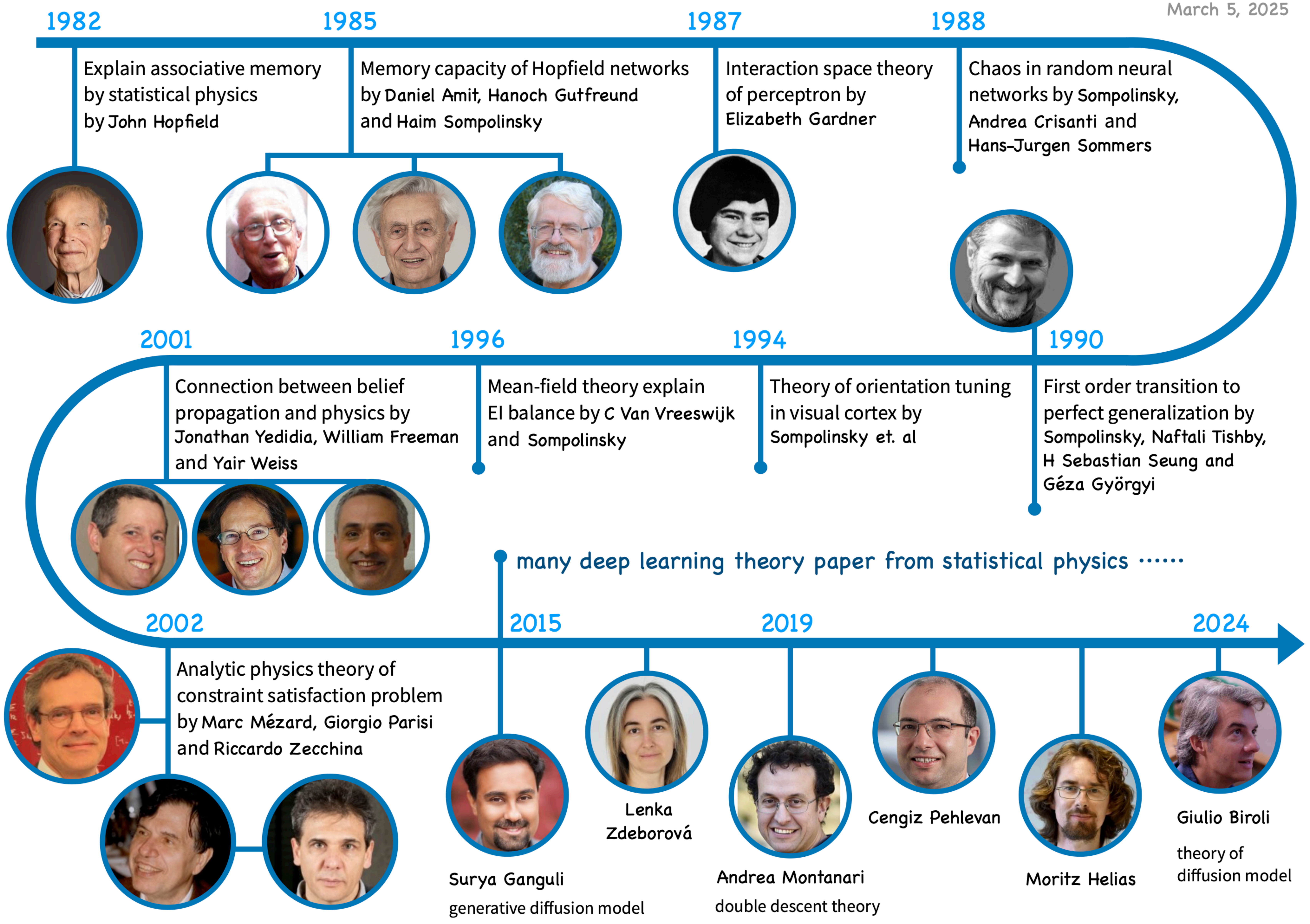
Dipartimento di Fisica, Universita di Roma I, La Sapienza Piazzale Aldo Moro 2, Roma 00185, and INFN, Sezione di Roma





Timeline for Statistical Mechanics of Neural Networks

Yu-Hao Li @ PMI Lab
March 5, 2025



Statistical Physics Approach

Hamiltonian

$$\mathcal{H}(\boldsymbol{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i + \frac{1}{2}\sum_i \lambda_i\sigma_i^2$$

Boltzmann distribution

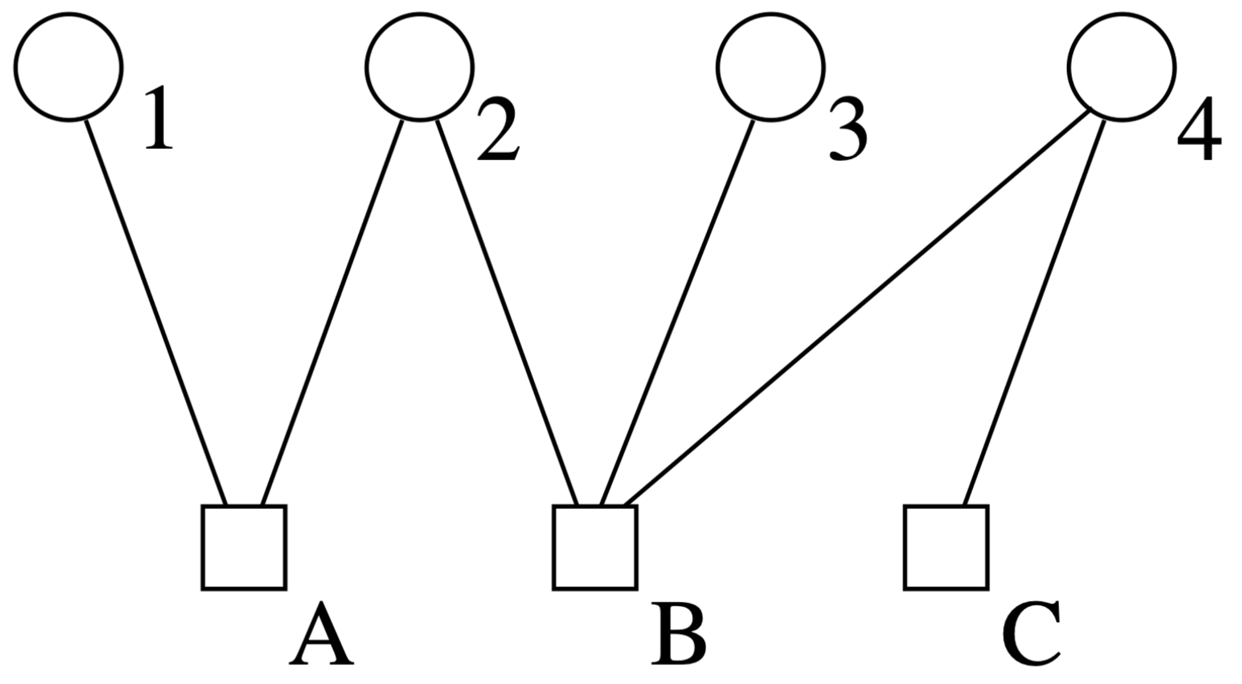
$$P(\boldsymbol{\sigma}) = \frac{1}{Z}e^{-\beta H(\boldsymbol{\sigma})} = \frac{1}{Z}\prod_i e^{\beta h_i\sigma_i - \frac{\beta\lambda}{2}\sigma_i^2}\prod_{i<j} e^{\beta J_{ij}\sigma_i\sigma_j}$$

Cavity Method

⟷

Belief Propagation
(Message Passing)

factor graph

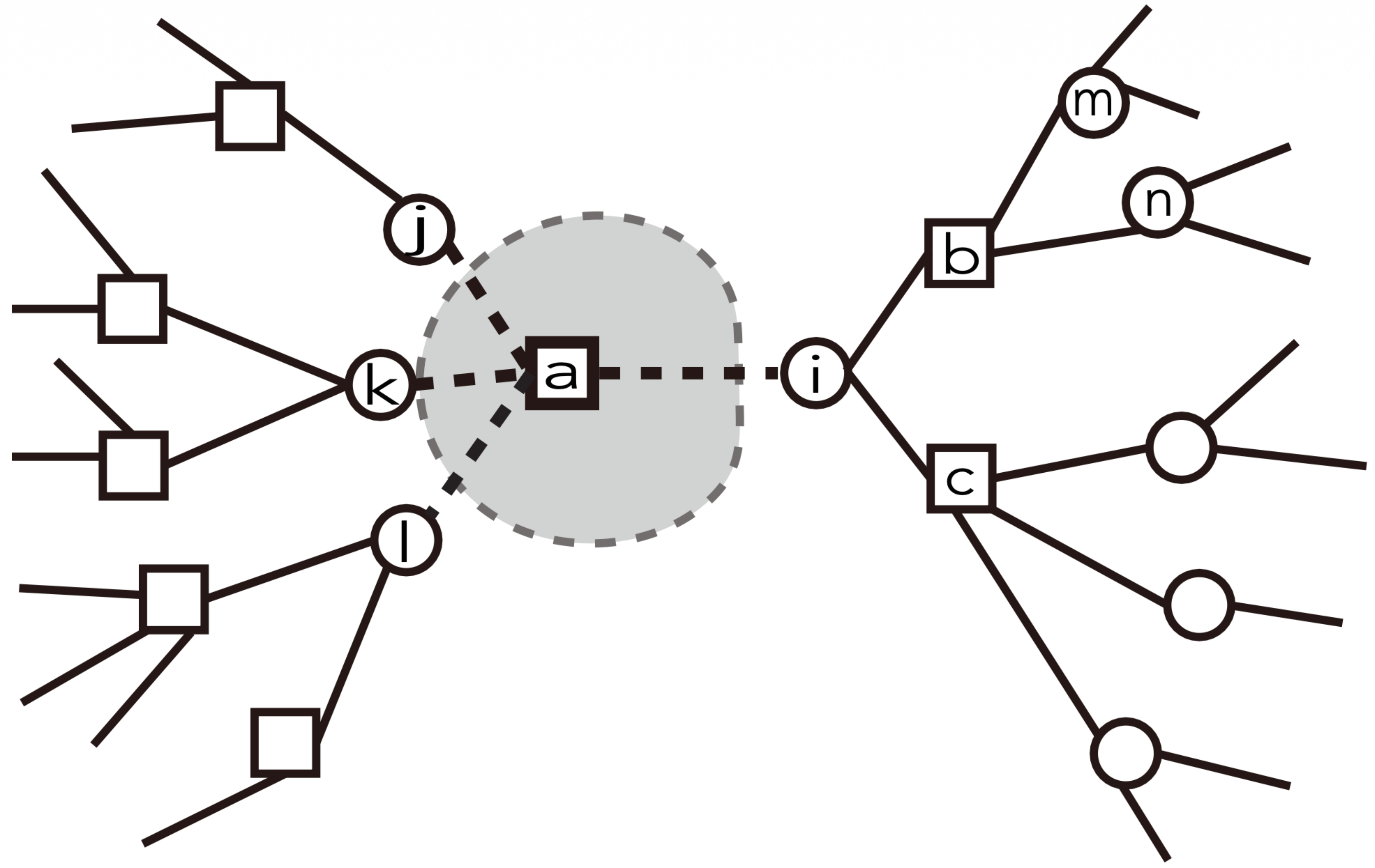


$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z} f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4)$$

J. S. Yedidia et al. IEEE Transactions on Information Theory, vol. 51, no. 7. 2005.

Statistical Physics Approach

A simple example $$H(\boldsymbol{\sigma}) = -\sum_{a=1}^{M} J_a \prod_{i\in\partial a} \sigma_i$$



Addition of the function node a to original system

partition function

$$Z^{\text{new}} = \sum_{\{\sigma_i\}_{i=1}^N} \exp\left(-\beta H^{\text{old}} + \beta J_a \prod_{i\in\partial a} \sigma_i\right)$$

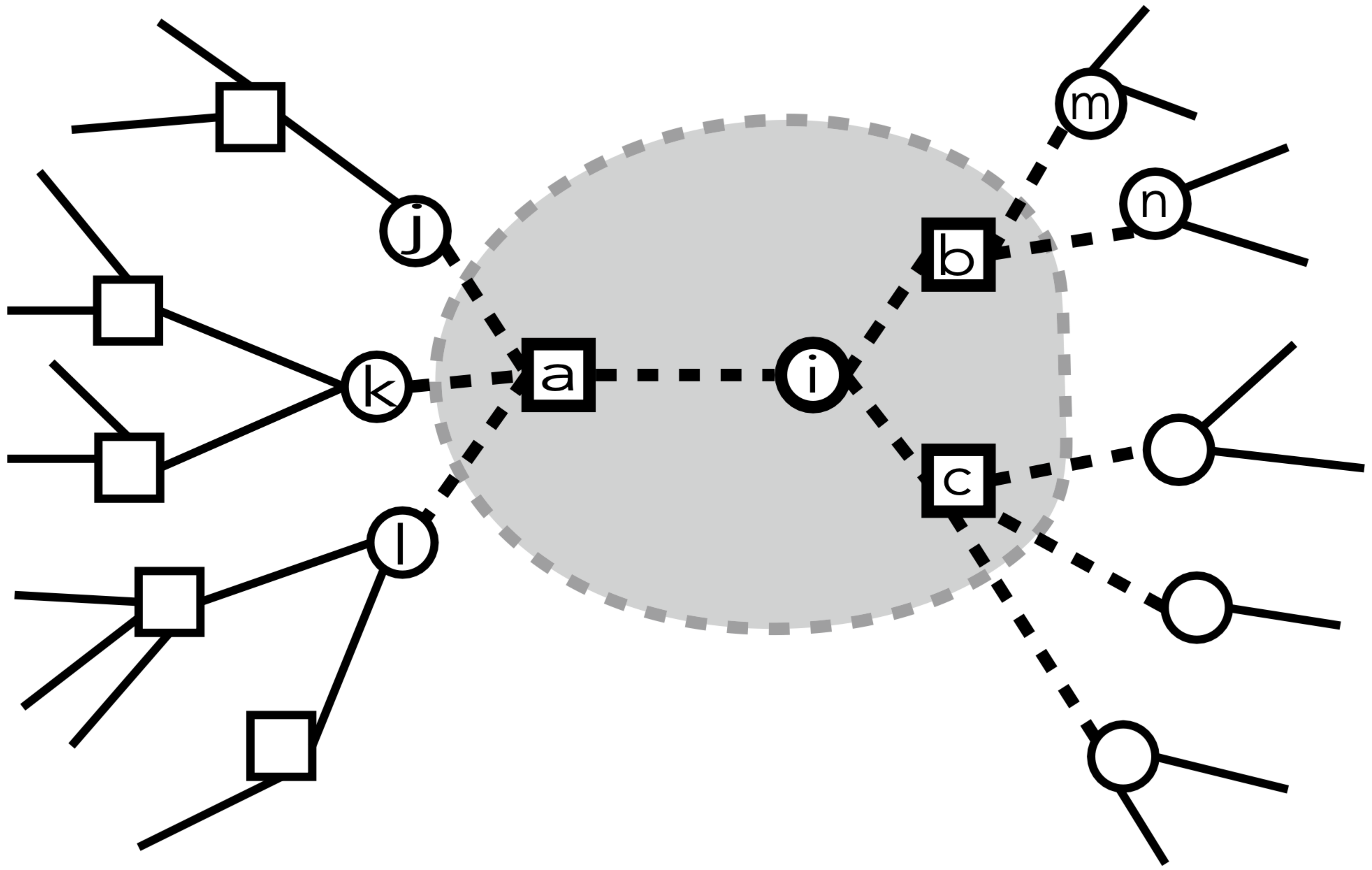
$$= Z^{\text{old}} \sum_{\{\sigma_i\}_{i=1}^N} \frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}} \exp\left(\beta J_a \prod_{i\in\partial a} \sigma_i\right)$$

free energy shift

$$-\beta \Delta F_a = \ln \frac{Z^{\text{new}}}{Z^{\text{old}}} = \ln\left[\cosh(\beta J_a)\left(1 + \tanh(\beta J_a) \prod_{i\in\partial a} m_{i\to a}\right)\right]$$

Statistical Physics Approach

A simple example $H(\boldsymbol{\sigma}) = -\sum_{a=1}^{M} J_a \prod_{i\in\partial a} \sigma_i$



Addition of the variable node a to original system

(together with its neighboring function nodes)

partition function

$$Z^{\text{new}} = \sum_{\boldsymbol{\sigma}^{\text{old}}} \sum_{\sigma_i} \exp\left(-\beta H^{\text{old}} + \beta \sum_{b\in\partial i} J_b \prod_{j\in\partial b} \sigma_j\right)$$

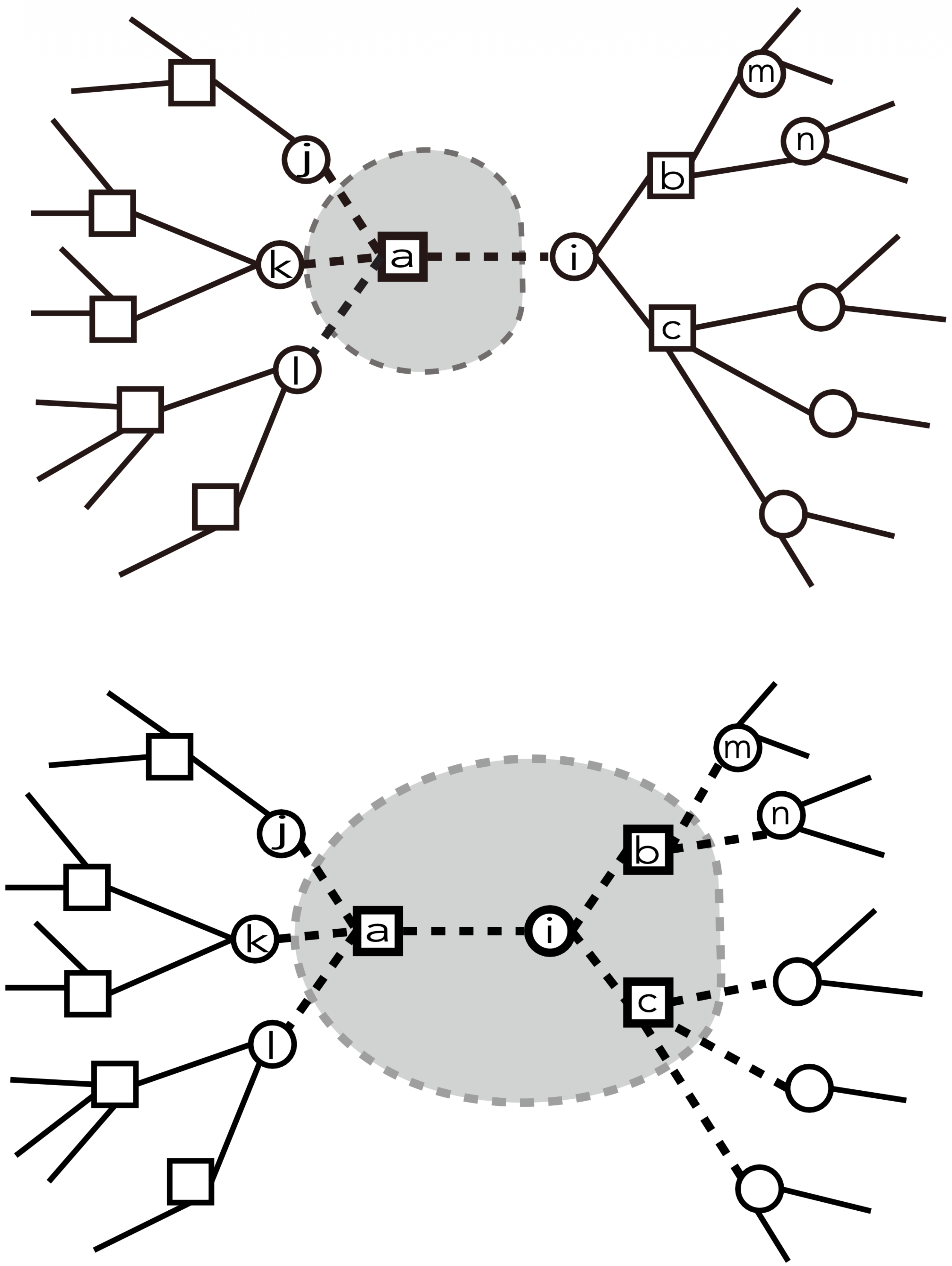
$$= \sum_{\boldsymbol{\sigma}^{\text{old}}} \sum_{\sigma_i} \exp\left(-\beta H^{\text{old}} + \beta \sum_{b\in\partial i} J_b \sigma_i \prod_{j\in\partial b\setminus i} \sigma_j\right)$$

$$= Z^{\text{old}} \sum_{\boldsymbol{\sigma}^{\text{old}}} \sum_{\sigma_i} \frac{\exp(-\beta H^{\text{old}})}{Z^{\text{old}}} \exp\left(\beta \sum_{b\in\partial i} J_b \sigma_i \prod_{j\in\partial b\setminus i} \sigma_j\right)$$

free energy shift

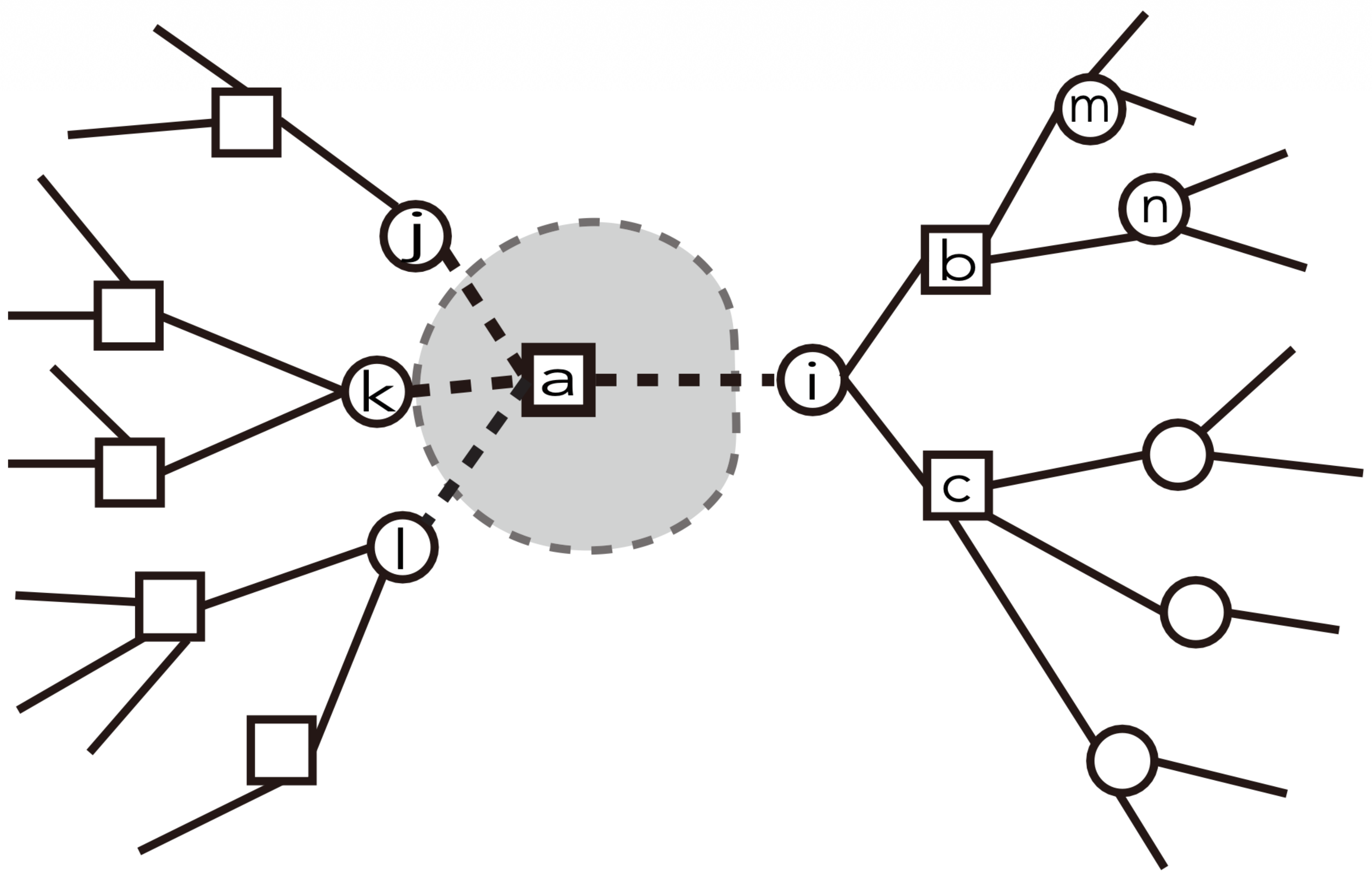
$$-\beta \Delta F_i = \ln \frac{Z^{\text{new}}}{Z^{\text{old}}} = \ln\left[\prod_{b\in\partial i} \Lambda^{+}_{b\to i} + \prod_{b\in\partial i} \Lambda^{-}_{b\to i}\right]$$

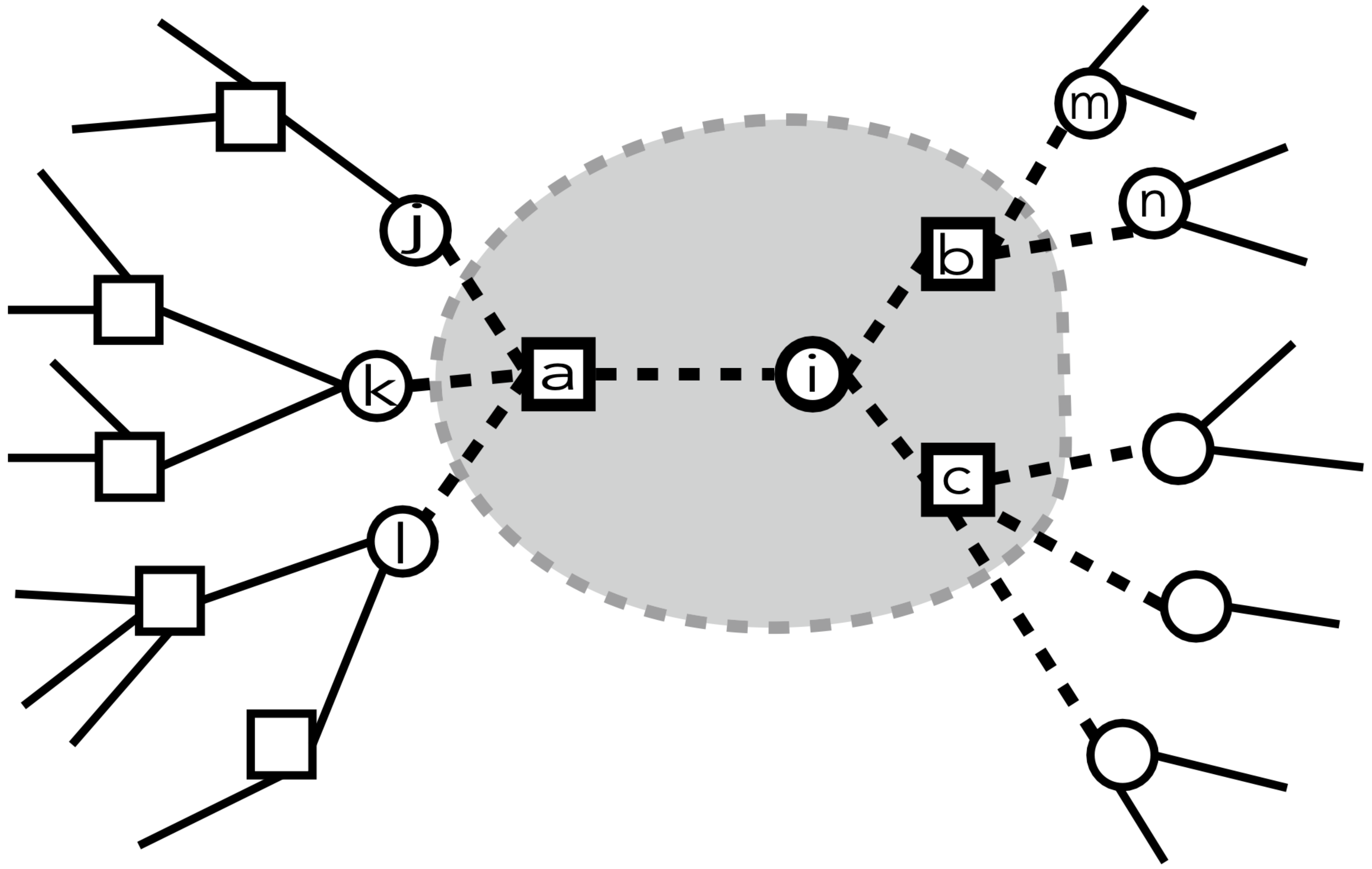
Statistical Physics Approach



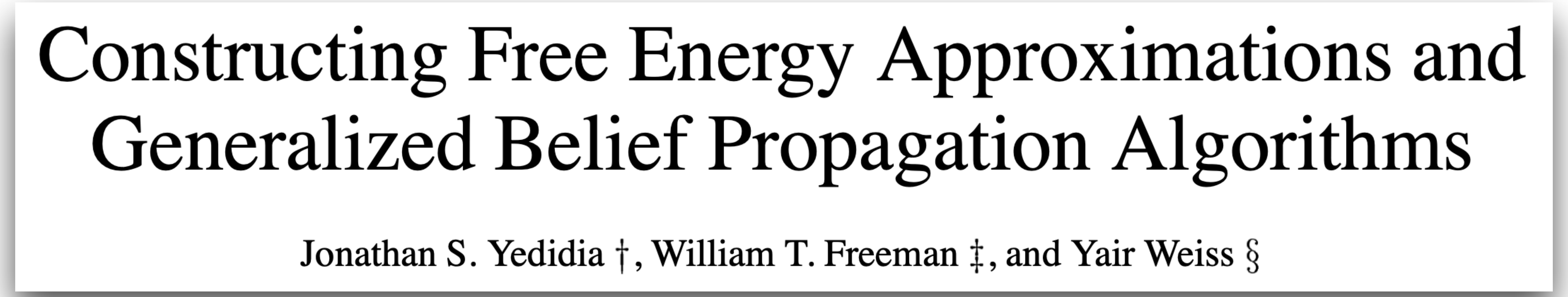
$$F = \sum_i \Delta F_i + \sum_a \Delta F_a - \sum_a |\partial a| \Delta F_a$$

Statistical Physics Approach



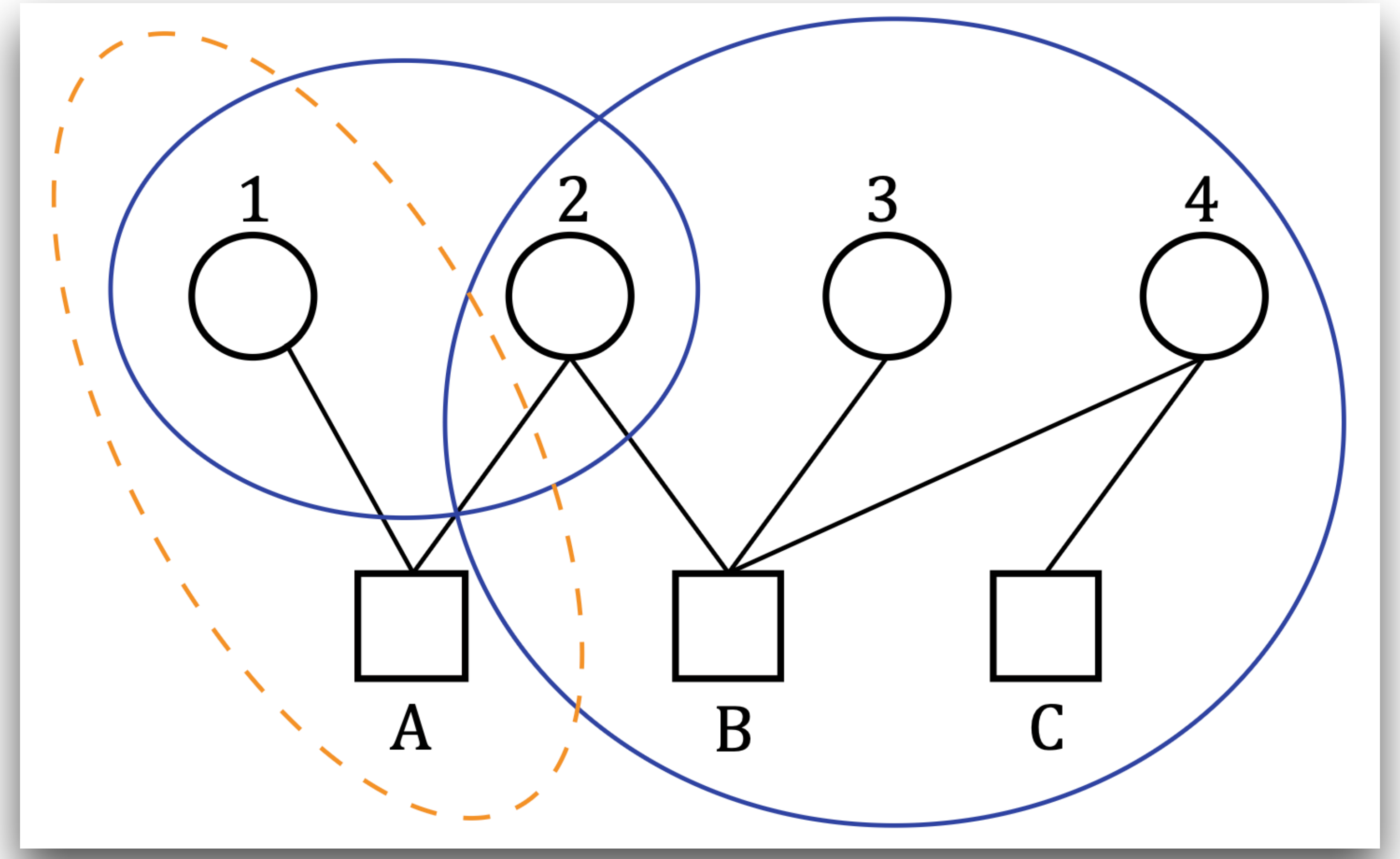


$$F = \sum_i \Delta F_i + \sum_a \Delta F_a - \sum_a |\partial a| \Delta F_a$$

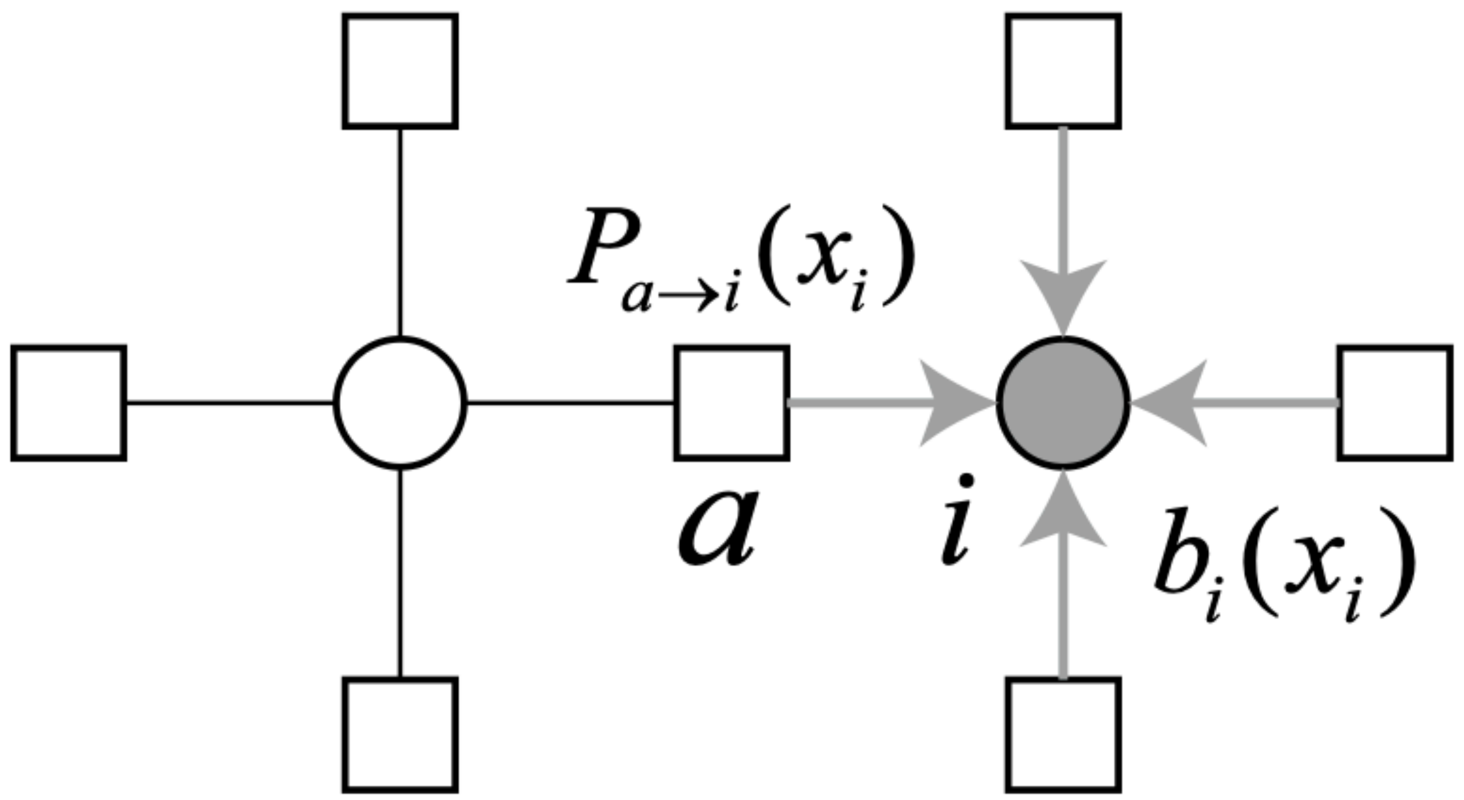


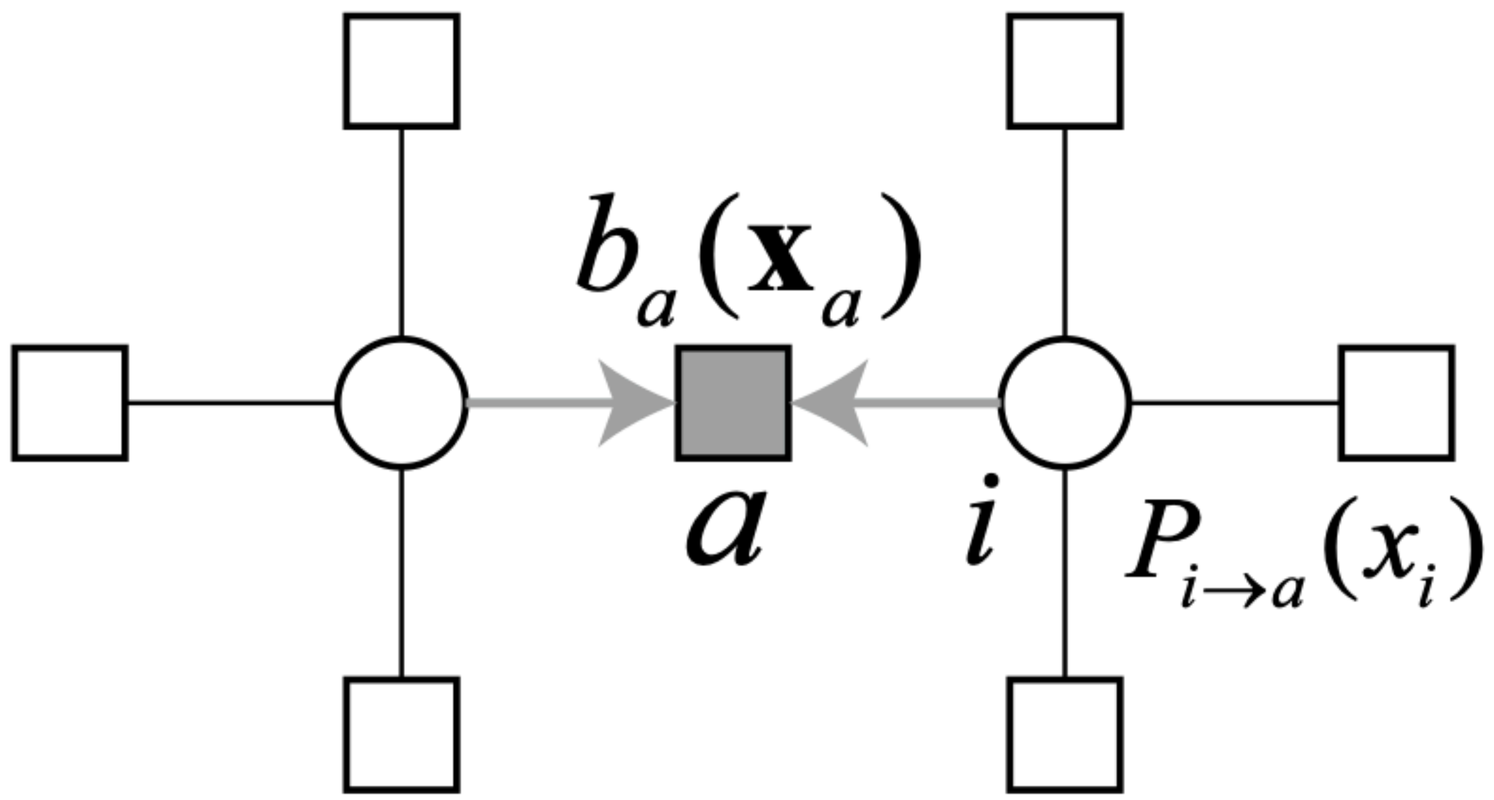
Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms

Jonathan S. Yedidia †, William T. Freeman ‡, and Yair Weiss §



Statistical Physics Approach





Belief Propagation (Message Passing)

$$b_i(x_i) = \frac{1}{Z_i} \prod_{a \in \partial i} P_{a\to i}(x_i),$$

$$b_a(\boldsymbol{x}_a) = \frac{1}{Z_a} f_a(\boldsymbol{x}_a) \prod_{i \in \partial a} \prod_{b \in \partial i \setminus a} P_{b\to i}(x_i)$$

$$P_{a\to i}(x_i) = \sum_{\boldsymbol{x}_j : j \in \partial a \setminus i} f_a(\boldsymbol{x}_a) \prod_{j \in \partial a \setminus i} P_{j\to a}(x_j),$$

$$P_{i\to a}(x_i) = \frac{1}{Z_{i\to a}} \prod_{b \in \partial i \setminus a} P_{b\to i}(x_i).$$

Statistical Physics Approach

cavity equations

$$\eta_{i\to ij}(\sigma_i) = \frac{1}{z_{i\to ij}} \exp\left(\beta h_i \sigma_i - \frac{\beta\lambda}{2}\sigma_i^2\right) \prod_{ik\in\partial i\setminus ij} \eta_{ik\to i}(\sigma_b),$$

$$\eta_{ij\to i}(\sigma_i) = \frac{1}{z_{ij\to i}} \int \prod_{j\in\partial ij\setminus i} \mathrm{d}\sigma_j \; \eta_{j\to ij}(\sigma_j) \; \exp\left(\beta J_{ij}\sigma_i\sigma_j\right).$$

for two-body interaction

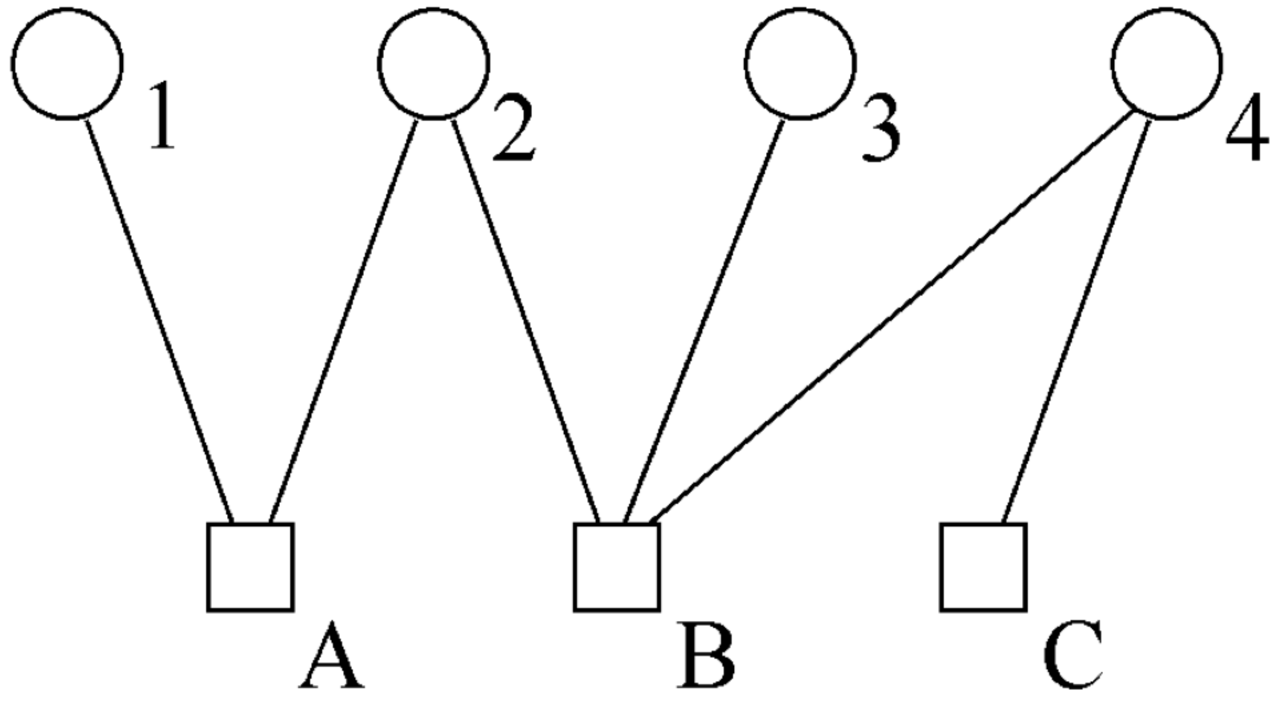
$$\eta_{i\to j}(\sigma_i) = \frac{1}{z_{i\to j}} \exp\left(\beta h_i\sigma_i - \frac{1}{2}\beta\lambda_i\sigma_i^2\right) \prod_{k\neq i,j} \left[\int \mathrm{d}\sigma_k \; \eta_{k\to i}(\sigma_k) \; \exp\left(\beta J_{ik}\sigma_i\sigma_k\right)\right]$$

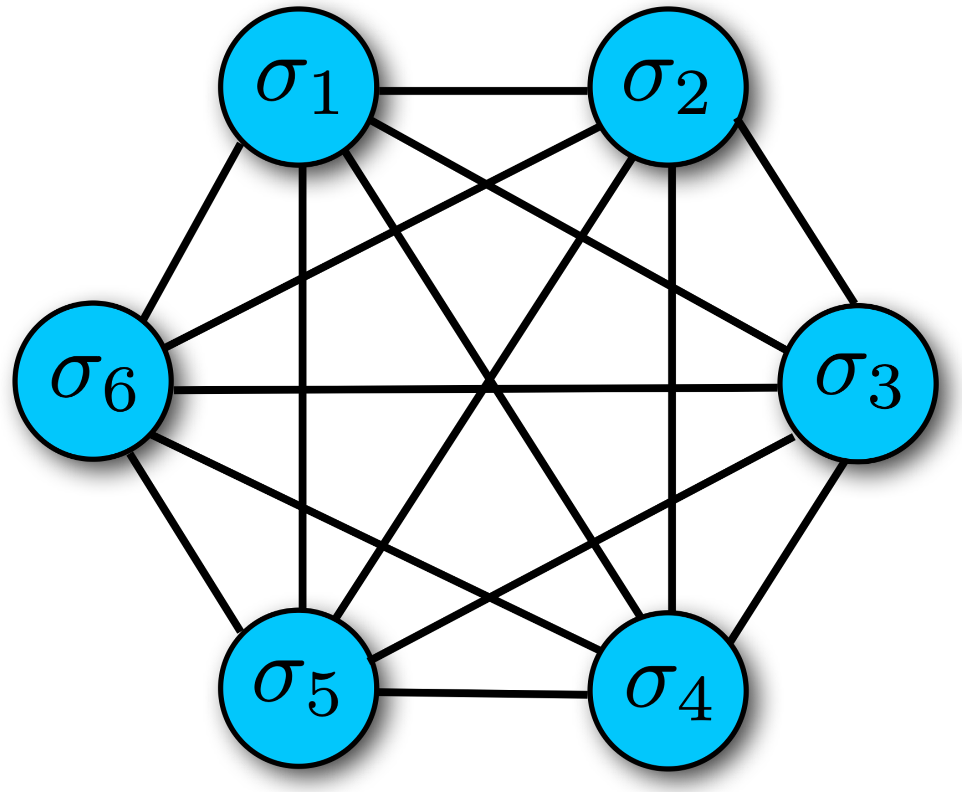
marginal probability distribution

$$P_i(\sigma_i) = \frac{1}{z_i} \exp\left(\beta h_i\sigma_i - \frac{1}{2}\beta\lambda_i\sigma_i^2\right) \prod_{j\neq i} \int \mathrm{d}\sigma_j \; \eta_{j\to i}(\sigma_j) \; \exp\left(\beta J_{ij}\sigma_i\sigma_j\right)$$

Boltzmann distribution

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} e^{-\beta H(\boldsymbol{\sigma})} = \frac{1}{Z} \prod_i e^{\beta h_i\sigma_i - \frac{\beta\lambda}{2}\sigma_i^2} \prod_{i<j} e^{\beta J_{ij}\sigma_i\sigma_j}$$





Statistical Physics Approach

approximate message passing equation

$$\eta_{i\to j}(\sigma_i) = \frac{1}{z_{i\to j}} \exp\left(\beta h_i \sigma_i - \frac{1}{2}\beta\lambda_i\sigma_i^2\right) \prod_{k\neq i,j}\left[\int d\sigma_k\, \eta_{k\to i}(\sigma_k)\, \exp\left(\beta J_{ik}\sigma_i\sigma_k\right)\right]$$

↓ Gaussian

$$\eta_{i\to j}(\sigma_i) \sim \mathcal{N}(m_{i\to j}, v_{i\to j})$$

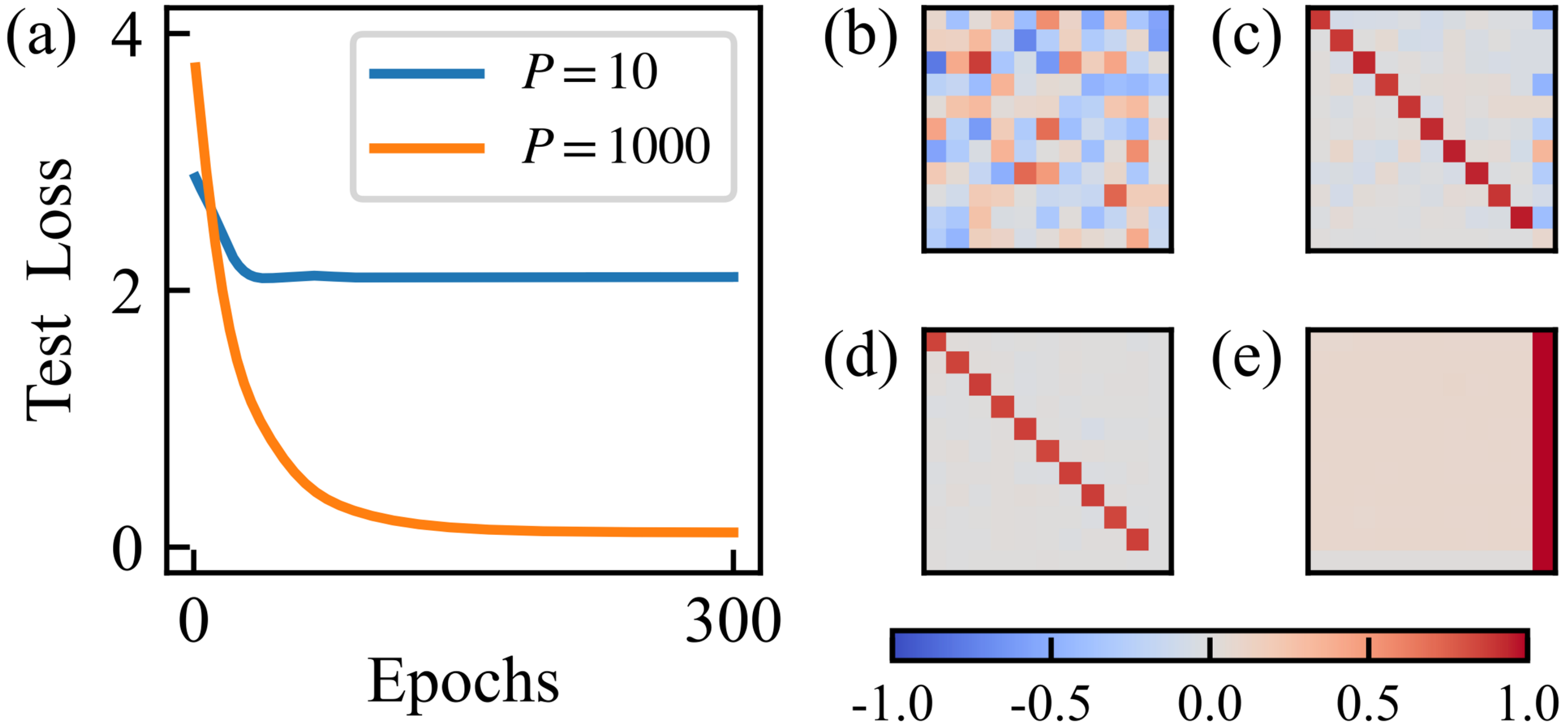
↓

$$m_{i\to j} = \frac{\beta h_i + \beta\sum_{k\neq i,j} J_{ik} m_{k\to i}}{\beta\lambda_i - \beta^2 \sum_{k\neq i,j} J_{ik}^2 v_{k\to i}}$$

$$v_{i\to j} = \frac{1}{\beta\lambda_i - \beta^2 \sum_{k\neq i,j} J_{ik}^2 v_{k\to i}}$$

↓ $m_i - m_{i\to j} \sim \mathcal{O}\left(\frac{1}{D^2}\right)$ $\quad v_i - v_{i\to j} \sim \mathcal{O}\left(\frac{1}{D^4}\right)$

$$m_i = \frac{\beta h_i + \beta\sum_{j\neq i} J_{ij} m_j}{\beta\lambda_i - \beta^2 \sum_{j\neq i} J_{ij}^2 v_j} \qquad v_i = \frac{1}{\beta\lambda_i - \beta^2 \sum_{j\neq i} J_{ij}^2 v_j} \longrightarrow \mathbf{W} = \mathrm{vec}^{-1}(\boldsymbol{\sigma})$$



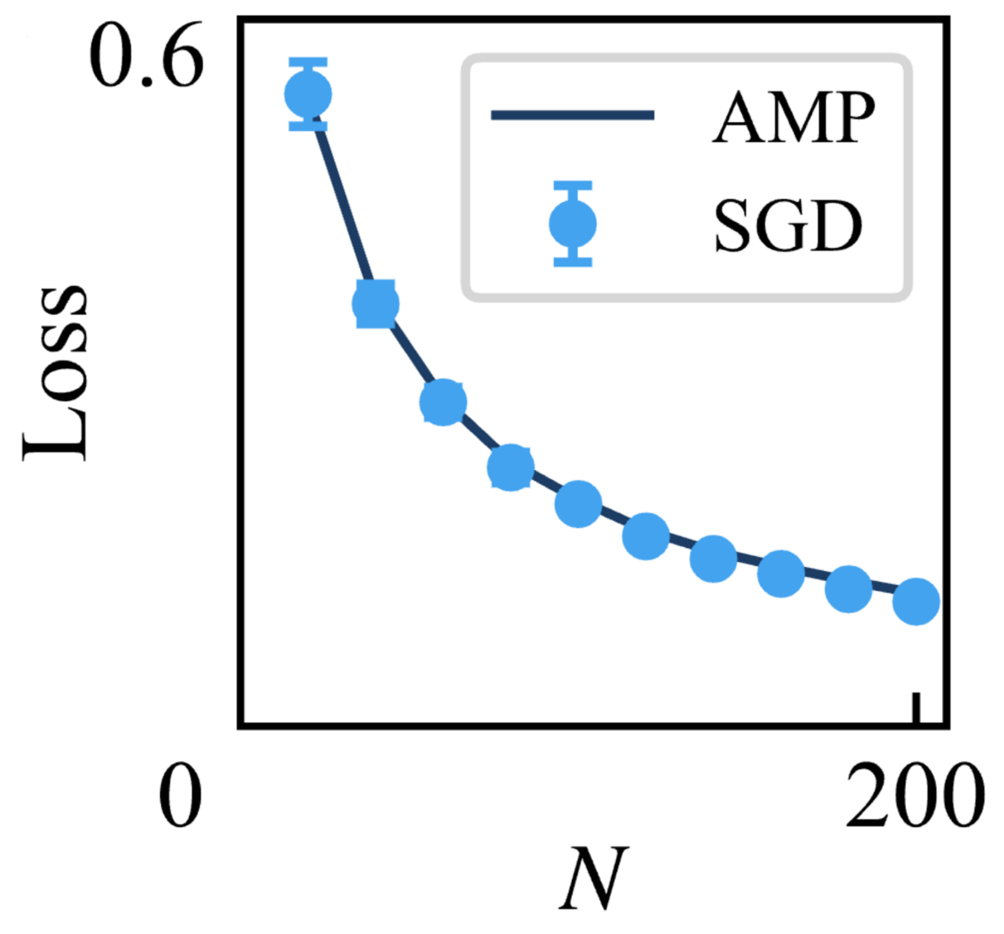
Comparison between Theory and Experiment

$$m_i = \frac{\beta h_i + \beta \sum_{j \neq i} J_{ij} m_j}{\beta \lambda_i - \beta^2 \sum_{j \neq i} J_{ij}^2 v_j}$$

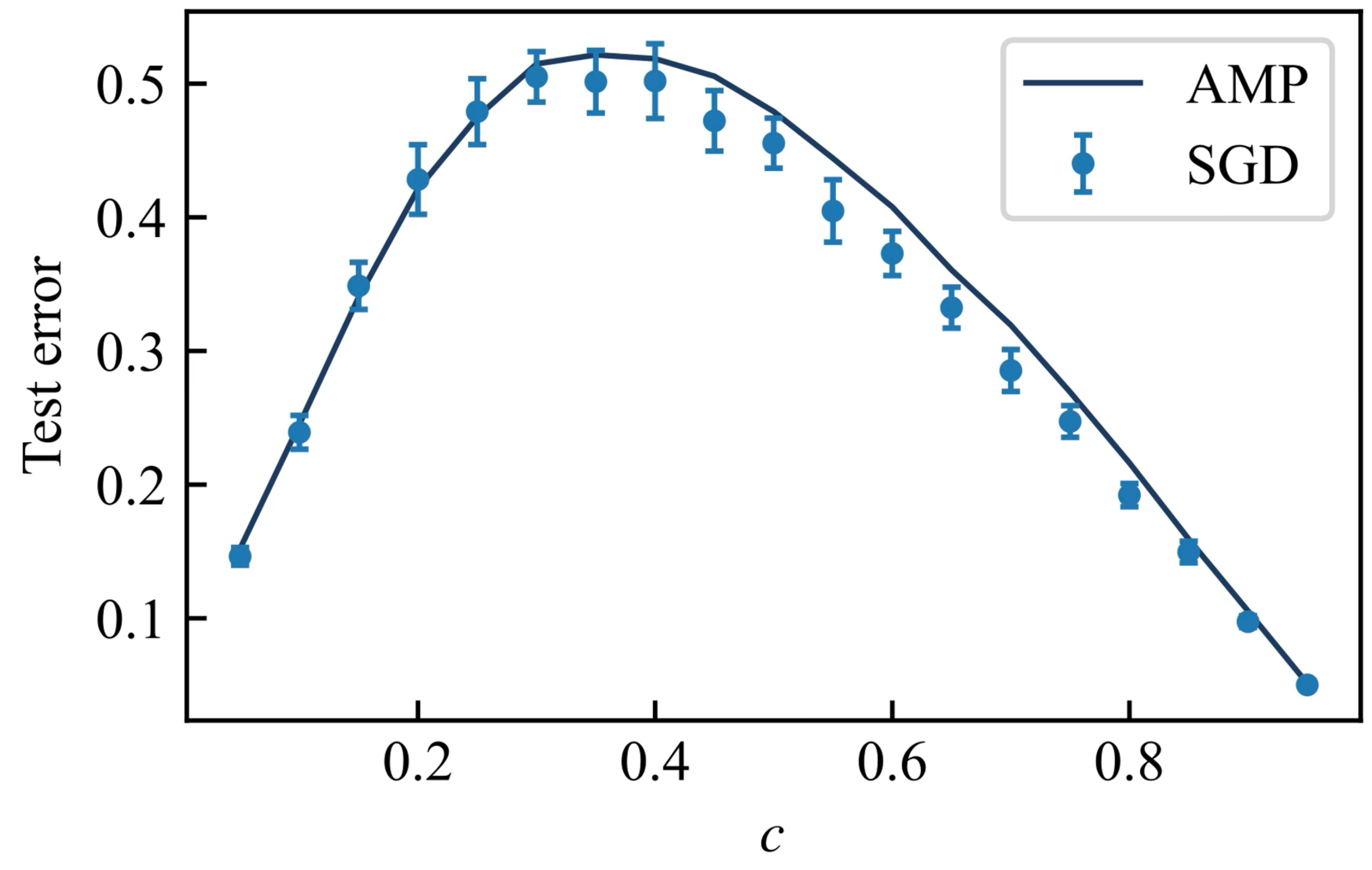
$$\mathbf{W} = \mathrm{vec}^{-1}(\boldsymbol{\sigma})$$

$$\mathbf{Y}(\mathbf{X}) = \frac{1}{N}\mathbf{X}\mathbf{X}^\top \mathbf{W}\mathbf{X}$$

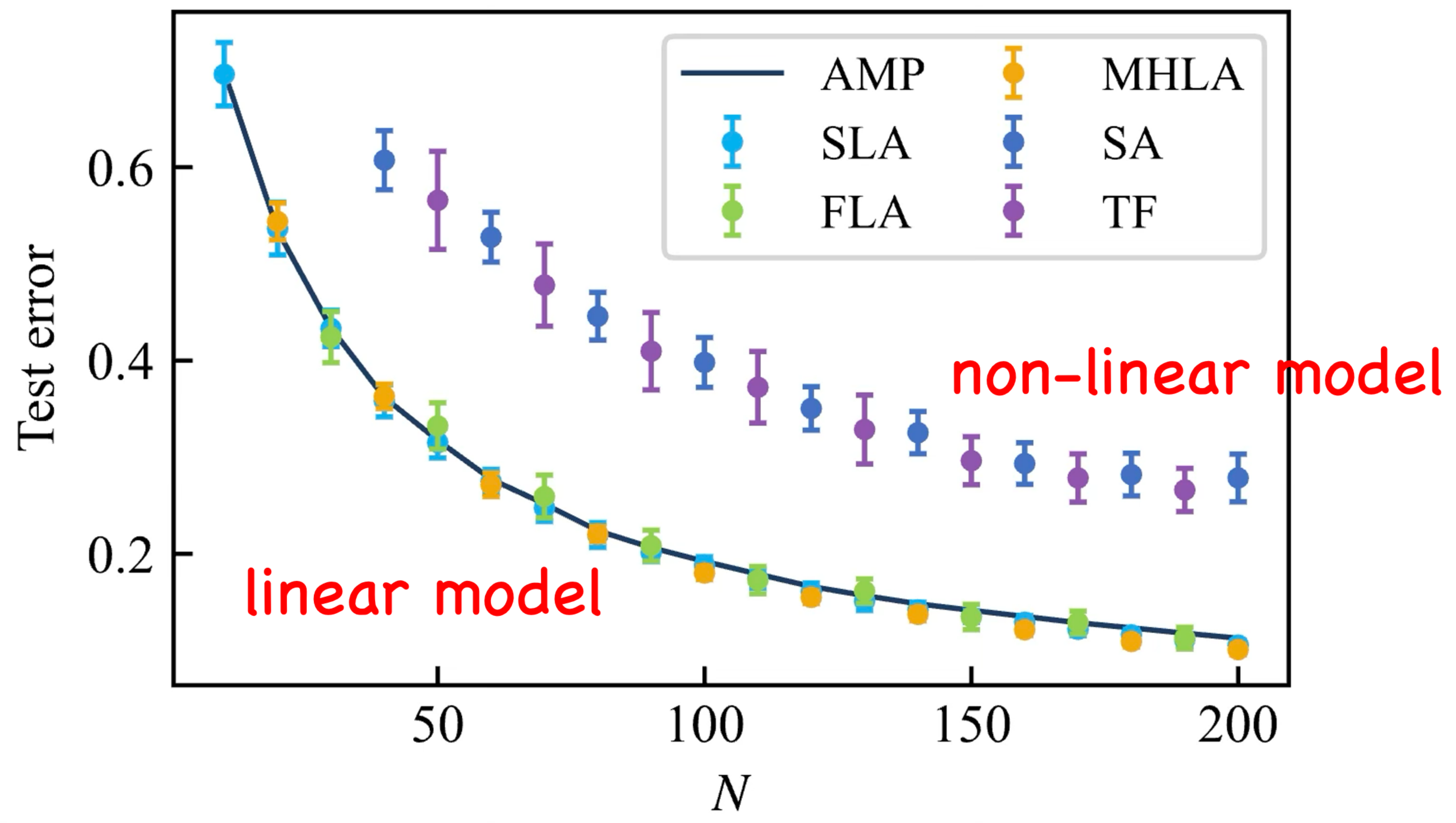
test error with the number of examples



for more complex task



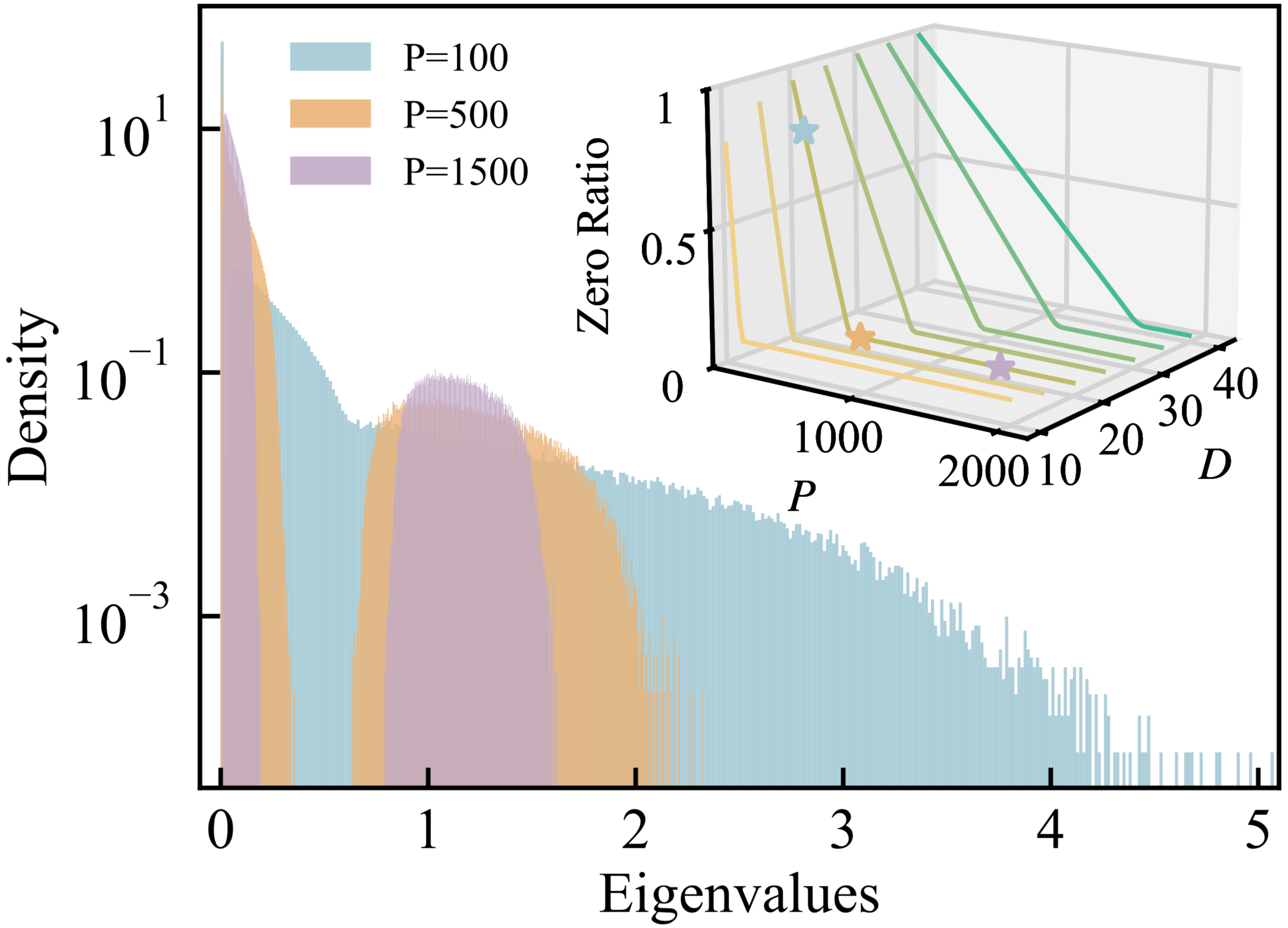
for more complex models



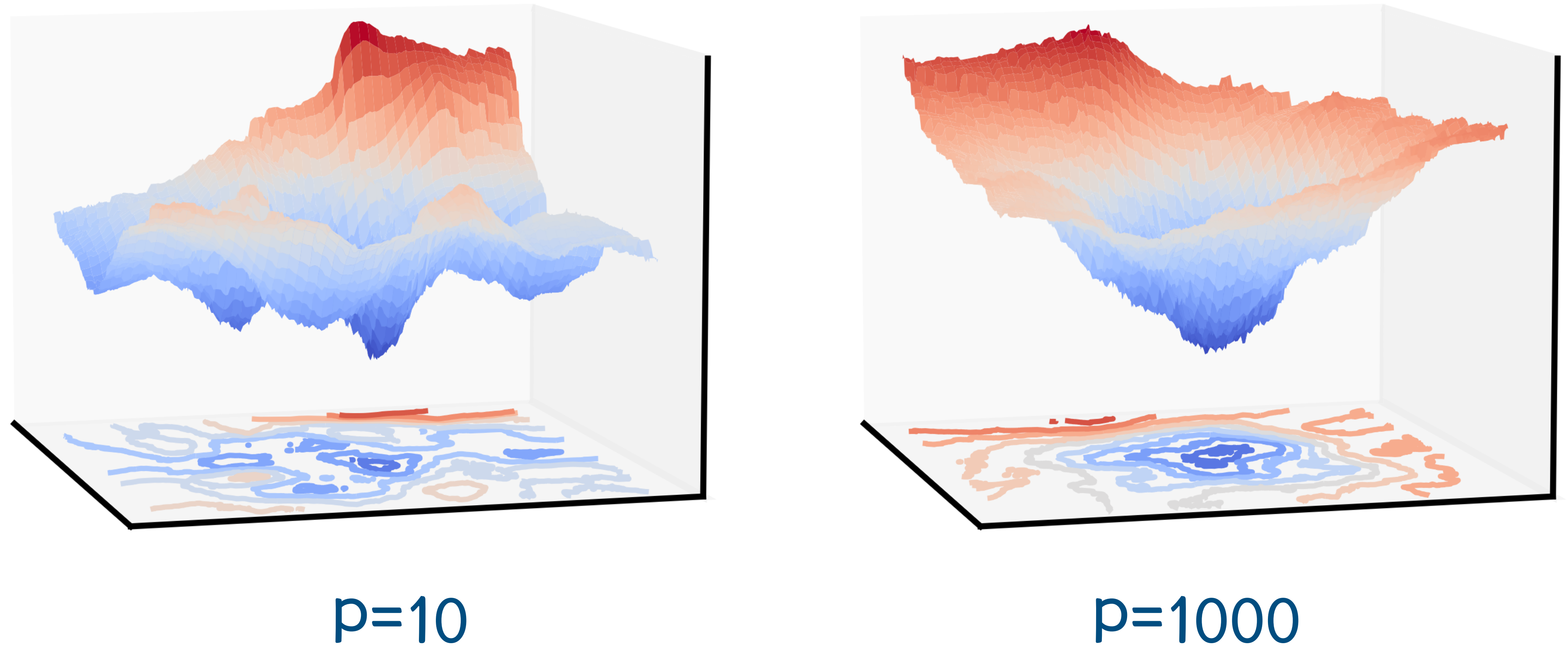
The Properties of Equilibrium State

Hessian matrix

$$\mathbf{H}_{ij} = \frac{\partial^2 \mathcal{L}}{\partial \sigma_i \partial \sigma_j} = \frac{1}{P} \sum_{\mu=1}^{P} s_i^\mu s_j^\mu$$



energy landscape



Summary

- A direct connection between ICL and physical model
- Mean field method to solve the physical model
- A theoretical perspective to explain how task diversity drives the emergence of ICL capabilities

Summary

- A direct connection between ICL and physical model
- Mean field method to solve the physical model
- A theoretical perspective to explain how task diversity drives the emergence of ICL capabilities

Prospect

- A unified and effective theoretical framework to explain ICL
- For theoretical work: more complex models, more practical tasks, ……
  learning (gradient flow) dynamics and non-equilibrium statistical physics
- For applications: multimodal, persistence, personalization, ……