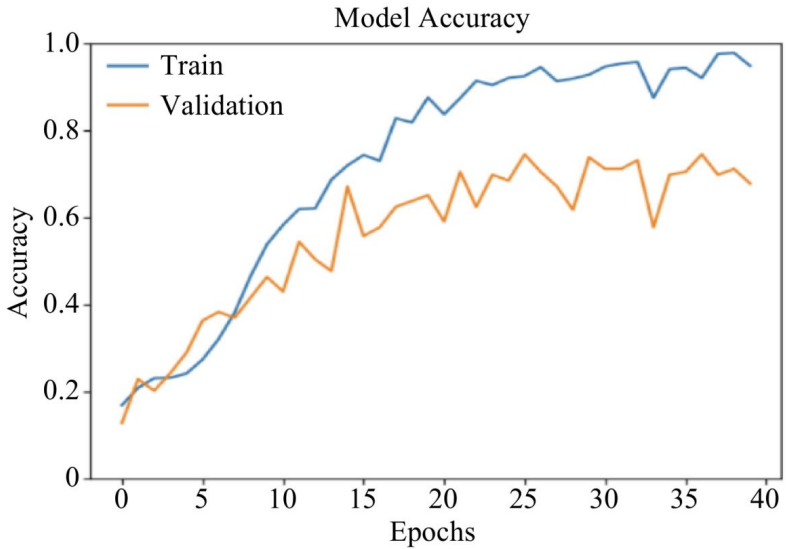
Emergent Abilities of Large Language Models

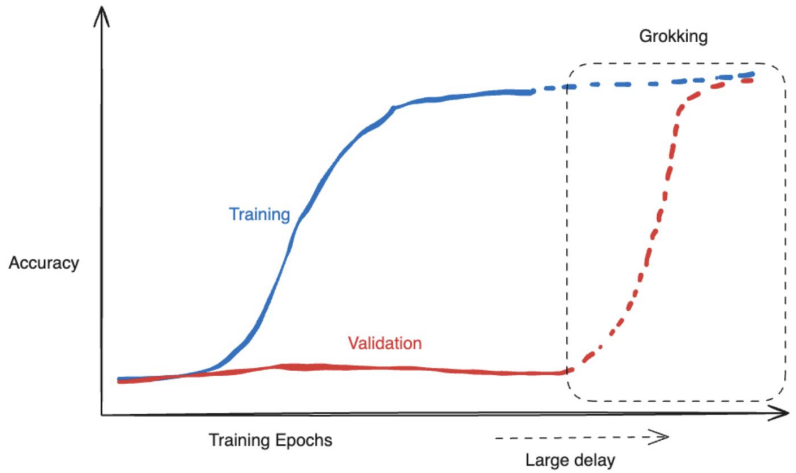
Grigoris Velegkas

Recap

Grokking

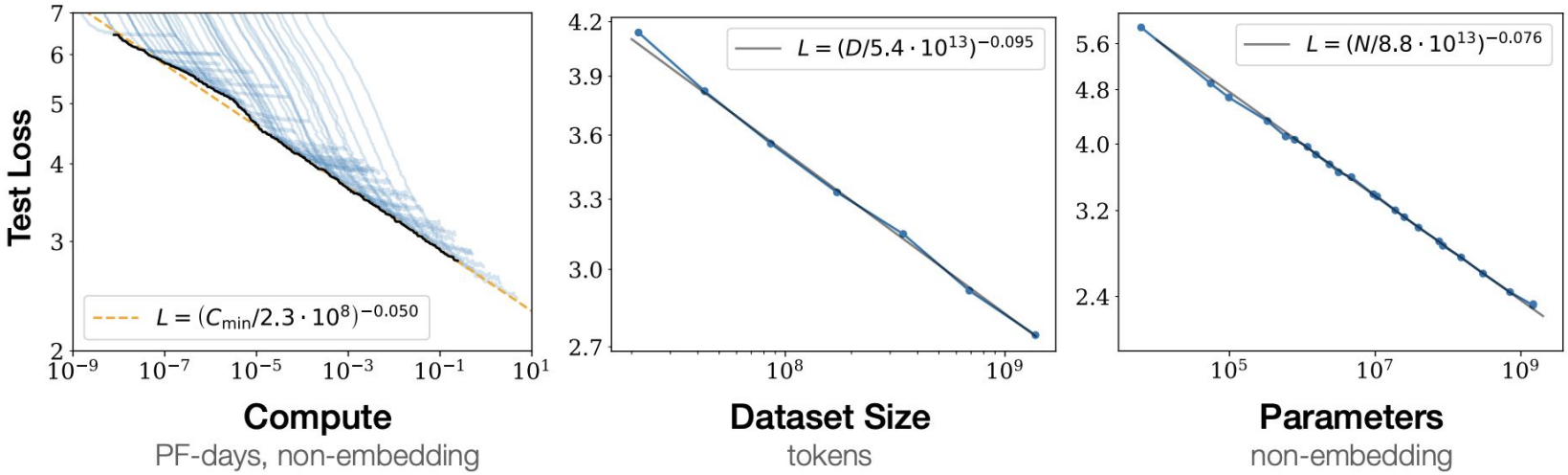


Typical training behavior



Grokking

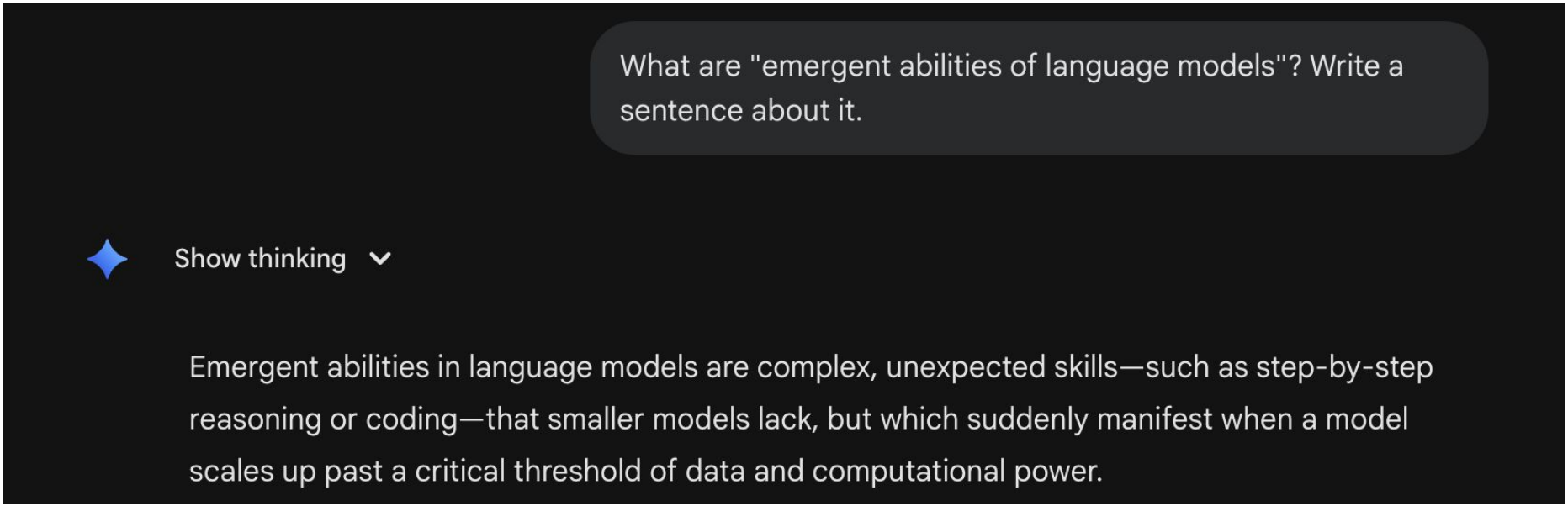
Scaling Laws (of Cross-Entropy Loss)



Cross-entropy loss has a very predictable behavior! (Kaplan et al., 2020)

Emergent Abilities

Emergent Abilities of Language Models



Emergent Abilities of Language Models

“Emergence is when quantitative changes in a system result in qualitative changes in behavior.”
- Wei et al. (2022); Philip Anderson (1972)

Background: Other Emergent Phenomena

Emergent phenomena appear in several different disciplines.

- Epidemics
- Biology
- Physics
- Theoretical computer science
  - Average case complexity problems
  - Planted clique detection
  - Solvability of 3-SAT
  - ....

Emergent Abilities of Language Models

Emergent Abilities of Large Language Models

Jason Wei [1]	*jasonwei@google.com*
Yi Tay [1]	*yitay@google.com*
Rishi Bommasani [2]	*nlprishi@stanford.edu*
Colin Raffel [3]	*craffel@gmail.com*
Barret Zoph [1]	*barretzoph@google.com*
Sebastian Borgeaud [4]	*sborgeaud@deepmind.com*
Dani Yogatama [4]	*dyogatama@deepmind.com*
Maarten Bosma [1]	*bosma@google.com*
Denny Zhou [1]	*dennyzhou@google.com*
Donald Metzler [1]	*metzler@google.com*
Ed H. Chi [1]	*edchi@google.com*
Tatsunori Hashimoto [2]	*thashim@stanford.edu*
Oriol Vinyals [4]	*vinyals@deepmind.com*
Percy Liang [2]	*pliang@stanford.edu*
Jeff Dean [1]	*jeff@google.com*
William Fedus [1]	*liamfedus@google.com*

[1] *Google Research* [2] *Stanford University* [3] *UNC Chapel Hill* [4] *DeepMind*

Wei et al. (2022)

Emergent Abilities vs. Scaling Laws

Scaling laws:

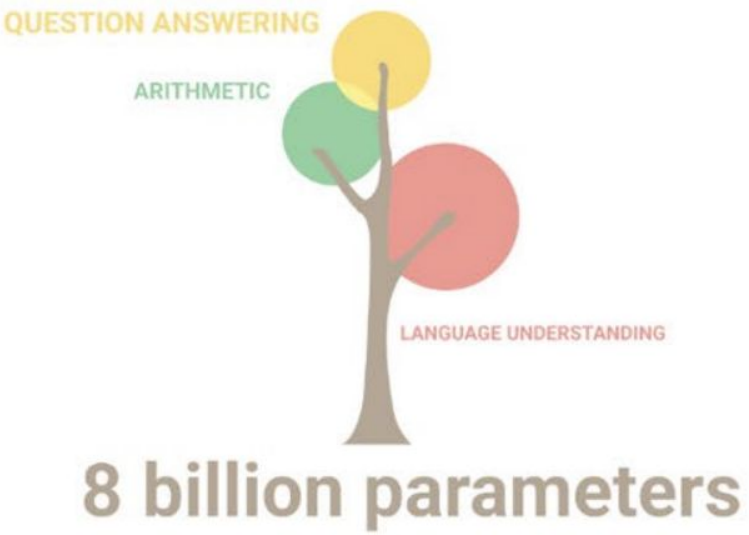
- By observing the performance on low-scale compute / data / model-size we can “extrapolate” the performance on large-scale compute / data / model-size
- Celebrated results that govern decisions for pre-training of LLMs

Emergent abilities

- There are abilities that are present on large models but not on small models
- These abilities cannot be predicted by extrapolating performance on smaller models

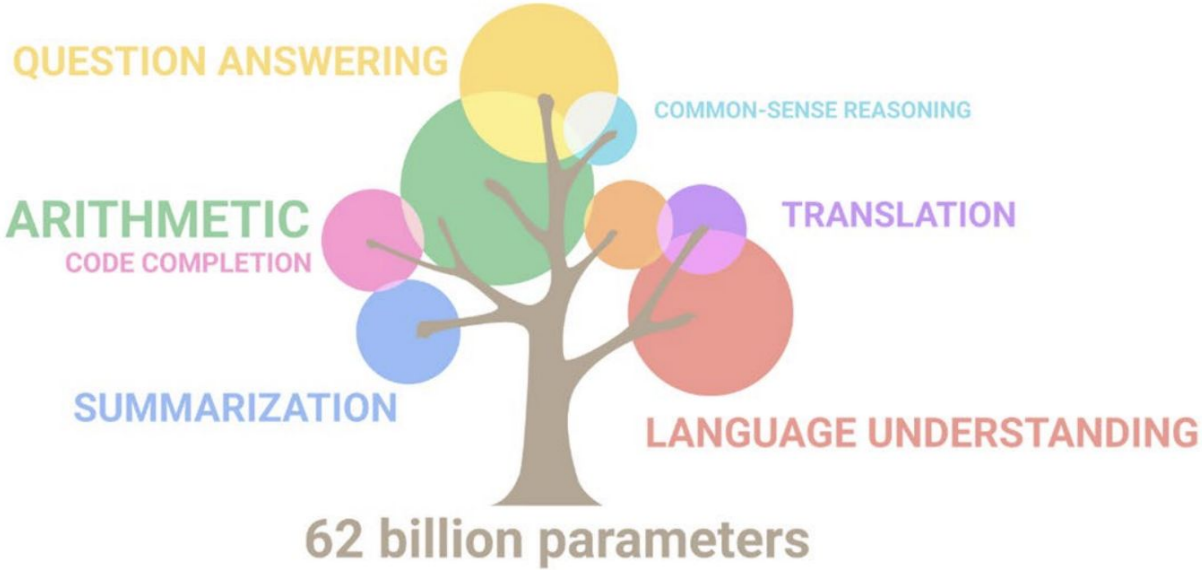
Prelude: Emergent Abilities of PaLM (2022; 540B)

Google Research blog post on how abilities “emerge” with scale



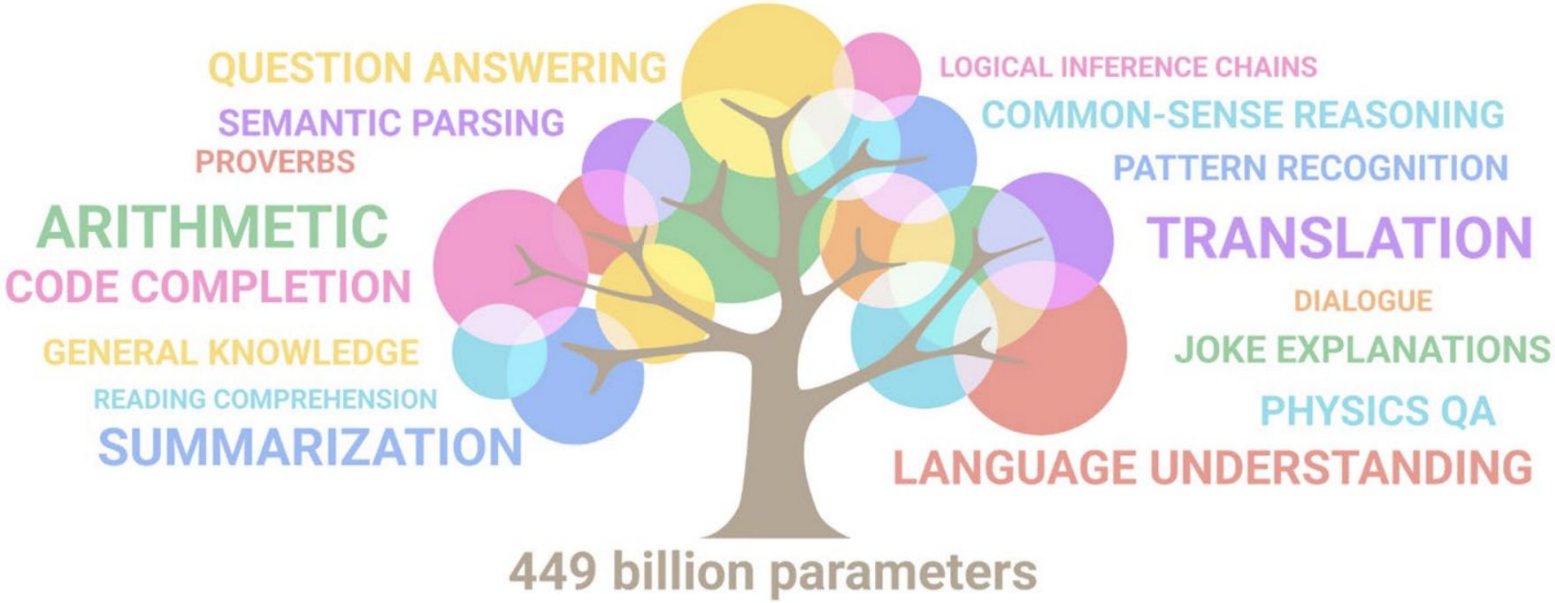
Prelude: Emergent Abilities of PaLM (2022; 540B)

Google Research blog post on how abilities “emerge” with scale



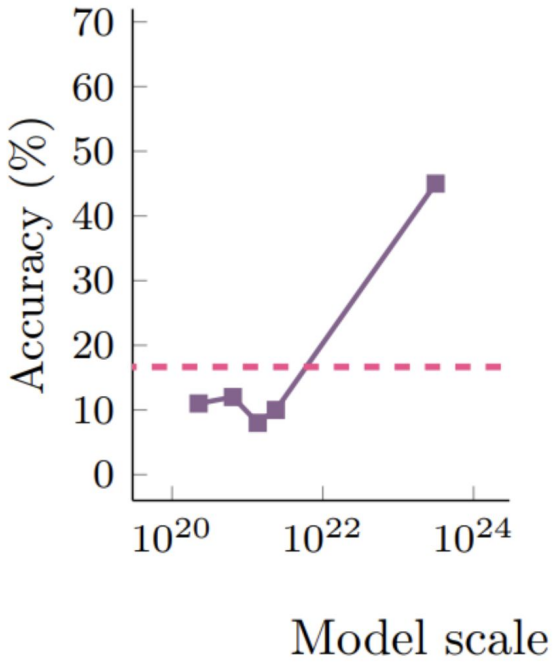
Prelude: Emergent Abilities of PaLM (2022; 540B)

Google Research [blog post] on how abilities “emerge” with scale

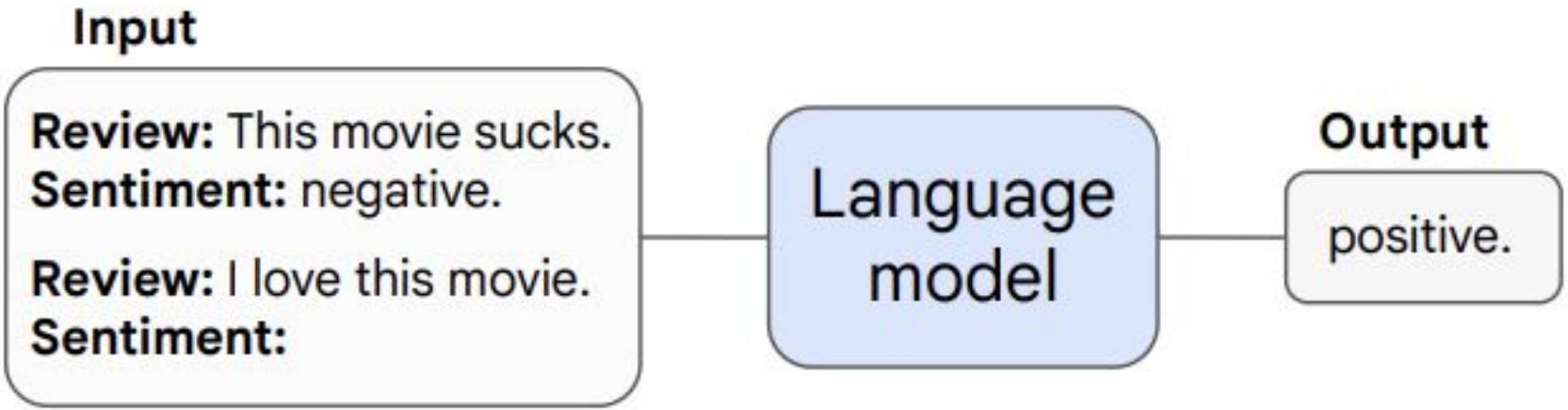


Emergent Abilities (Wei et al. 2022)

When is an emergent ability?

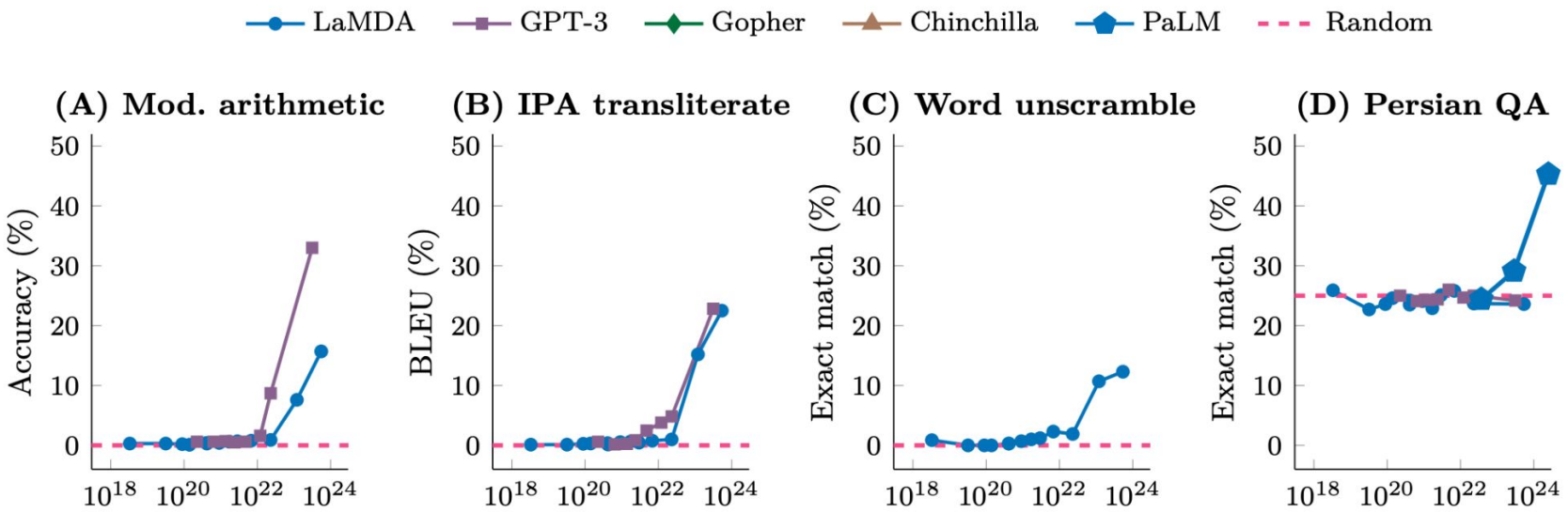


Few-Shot Prompting



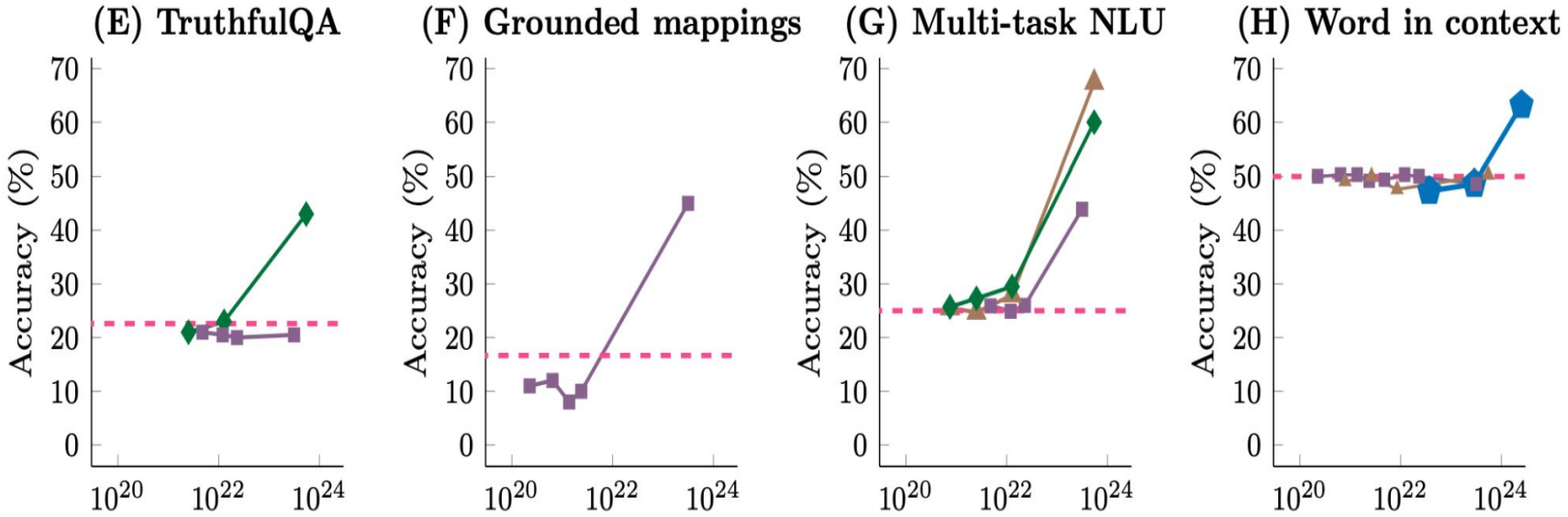
The ability to perform a task via few-shot prompting is **emergent** when a model has **random performance until a certain scale**, after which performance i**ncreases to well-above random**.

Emergent Abilities with Few-Shot Prompting (BIG-Bench)



- **Mod. arithmetic:** 3-digit addition-subtraction; 2-digit multiplication
- **IPA:** transliterating from the international phonetic alphabet
- **Unscramble:** recover a word from scrambled letters
- **Persian QA:** input is Persian text, model needs to respond to question

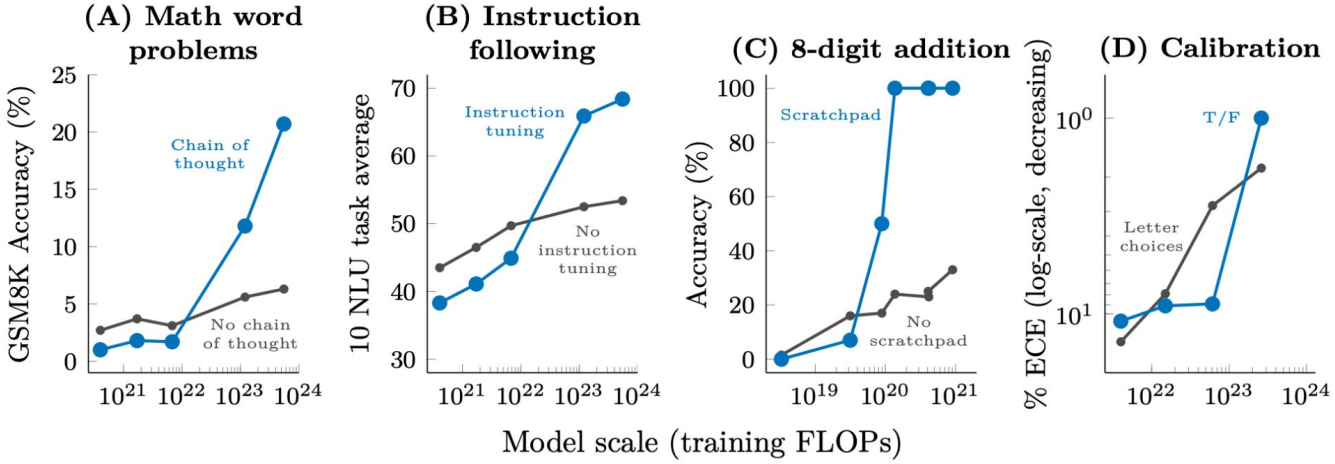
Emergent Abilities with Few-Shot Prompting (BIG-Bench)



- **TrughfulQA:** ability to answer truthfully
- **Grounded mappings:** create mappings from textual instructions
- **NLU:** tests covering a wide range of topics such as math, history, law
- **Word in context:** semantic understanding benchmark

Augmented Prompting Strategies

- Interaction with LLMs beyond few-shot prompting
- A “technique” is emergent if using it on small-scale models is not useful (or even harmful), but it gives important gains beyond a certain threshold



Summary of Experiments

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

Discussion

- Certain abilities of LLMs cannot be predicted by the behavior of small models
  - Enhances the narrative that scaling is crucial
- Explanations of emergence?
  - Need to look at evaluation metrics carefully! (e.g., are rewards binary?)
  - Smaller models might achieve these “emergent” abilities with better training (architecture, data, algorithms)
- Can we predict if a given “ability” that LLMs lack can be obtained by scaling?
- Can we predict if a given “ability” plateaus beyond a certain threshold?

Setup

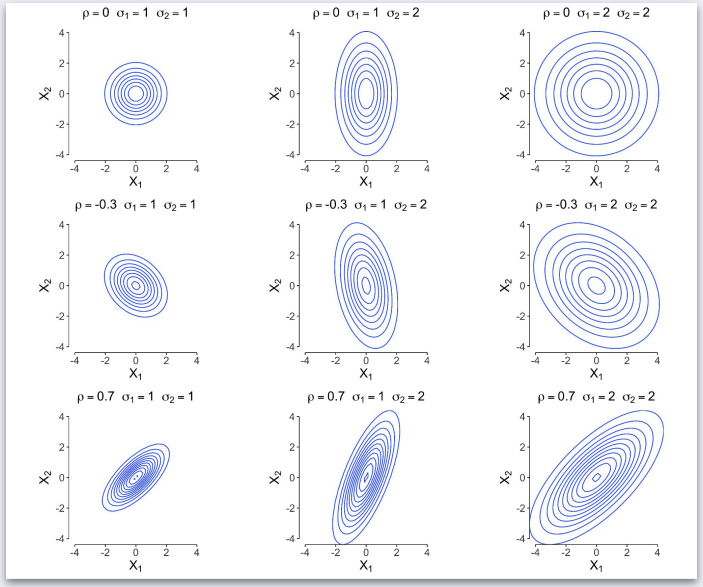
Simple Prediction Setting: Linear Regression

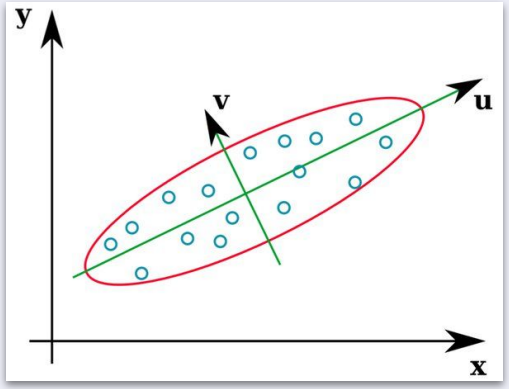
- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad \text{(assume } \lambda_1 \geq \lambda_2 \geq \cdots\text{)}$$

Intuition for eigenvalues + eigenvectors

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad (\text{assume } \lambda_1 \geq \lambda_2 \geq \cdots)$$





Setup

Simple Prediction Setting: Linear Regression

- Covariate $x \in \mathbb{H}$ (Hilbert space); response $y \in \mathbb{R}$.
- (x, y) subgaussian, mean zero, well-specified: $\mathbb{E}[y|x] = x^\top \theta^*$.
- Define:

$$\Sigma := \mathbb{E}xx^\top = \sum_i \lambda_i v_i v_i^\top, \qquad \text{(assume } \lambda_1 \geq \lambda_2 \geq \cdots\text{)}$$

$$\theta^* := \arg\min_\theta \mathbb{E}\left(y - x^\top \theta\right)^2,$$

$$\sigma^2 := \mathbb{E}(y - x^\top \theta^*)^2.$$

A bit more general than linear regression

We can capture **both linear** and **polynomial regression**

- A simple way to do that is to do a **feature expansion**
- + have a linear function over the “expanded” features

Eg.

- Feature expansion: $x \rightarrow (1, x, x^2, x^3, \ldots) = f(x)$
- Polynomial regression := linear regression on f(x)

Our estimator

Overparameterized regime: $n \ll d = p$

Regularized linear regression

$$\min \quad \lambda\|\theta\|^2 + \frac{1}{n}\|X\theta - y\|^2,$$

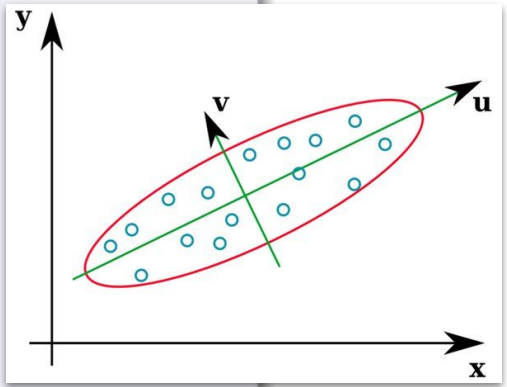
$$\begin{aligned} \min \quad & \|X\theta - y\|^2 \\ \text{s.t.} \quad & \|\theta\| \le b, \end{aligned}$$

$$\begin{aligned} \min \quad & \|\theta\| \\ \text{s.t.} \quad & \frac{1}{n}\|X\theta - y\|^2 \le c. \end{aligned}$$

Quantity of interest

Excess prediction error

$$R(\hat{\theta}) := \mathbb{E}_{(x,y)} \left(y - x^\top \hat{\theta} \right)^2 - \underbrace{\min_\theta \mathbb{E} \left(y - x^\top \theta \right)^2}_{\text{optimal prediction error}}$$



All in one

Overfitting regime: n ≪ d = p

Overfitting regime

- We consider situations where $\min_\beta \|X\beta - y\|^2 = 0$.
- Estimator $\hat{\theta} = (X^\top X)^\dagger X^\top y$ solves

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{s.t.} \quad & \|X\theta - y\|^2 = \min_\beta \|X\beta - y\|^2 = 0. \end{aligned}$$

- Hence, $y_1 = x_1^\top \hat{\theta}, \ldots, y_n = x_n^\top \hat{\theta}$.

Main question

Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

     Problematic.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

(1) Inevitable error that comes from **small training size**

(2) The contribution of **overfitting to noise** (!)

Main result

Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if

Important object

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

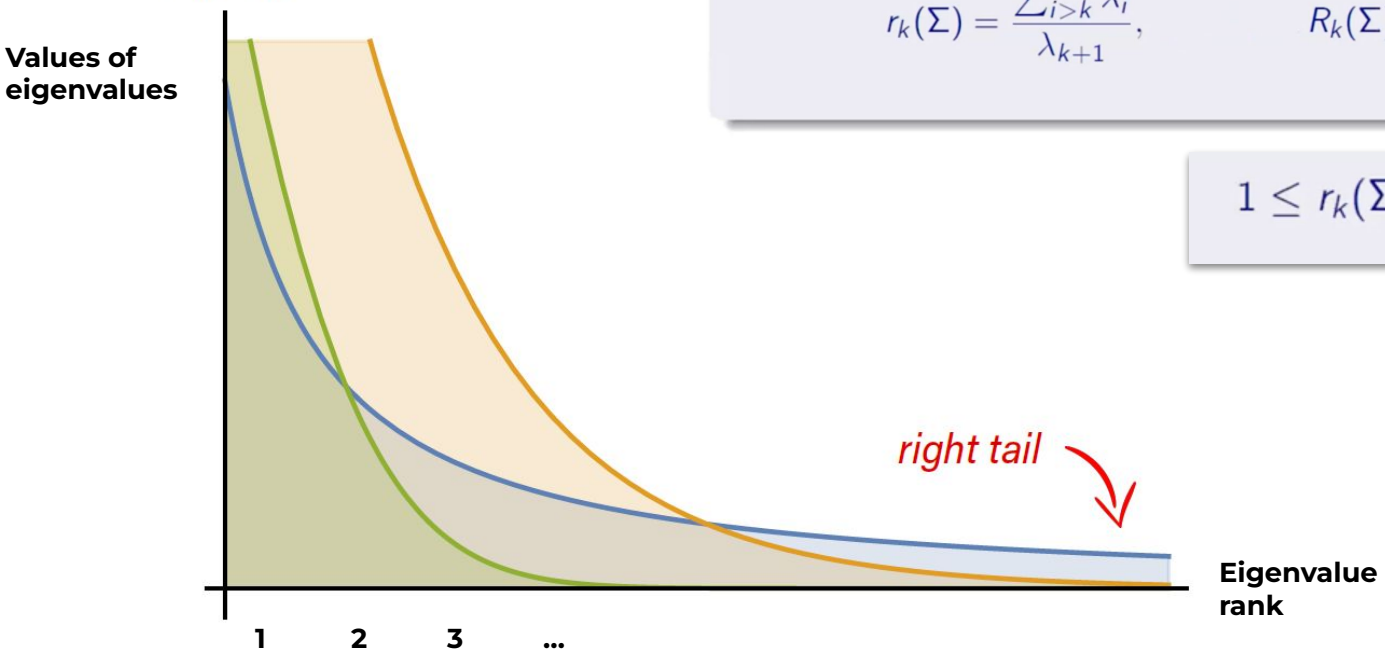
Example in picture

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

$$1 \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$



Example in picture

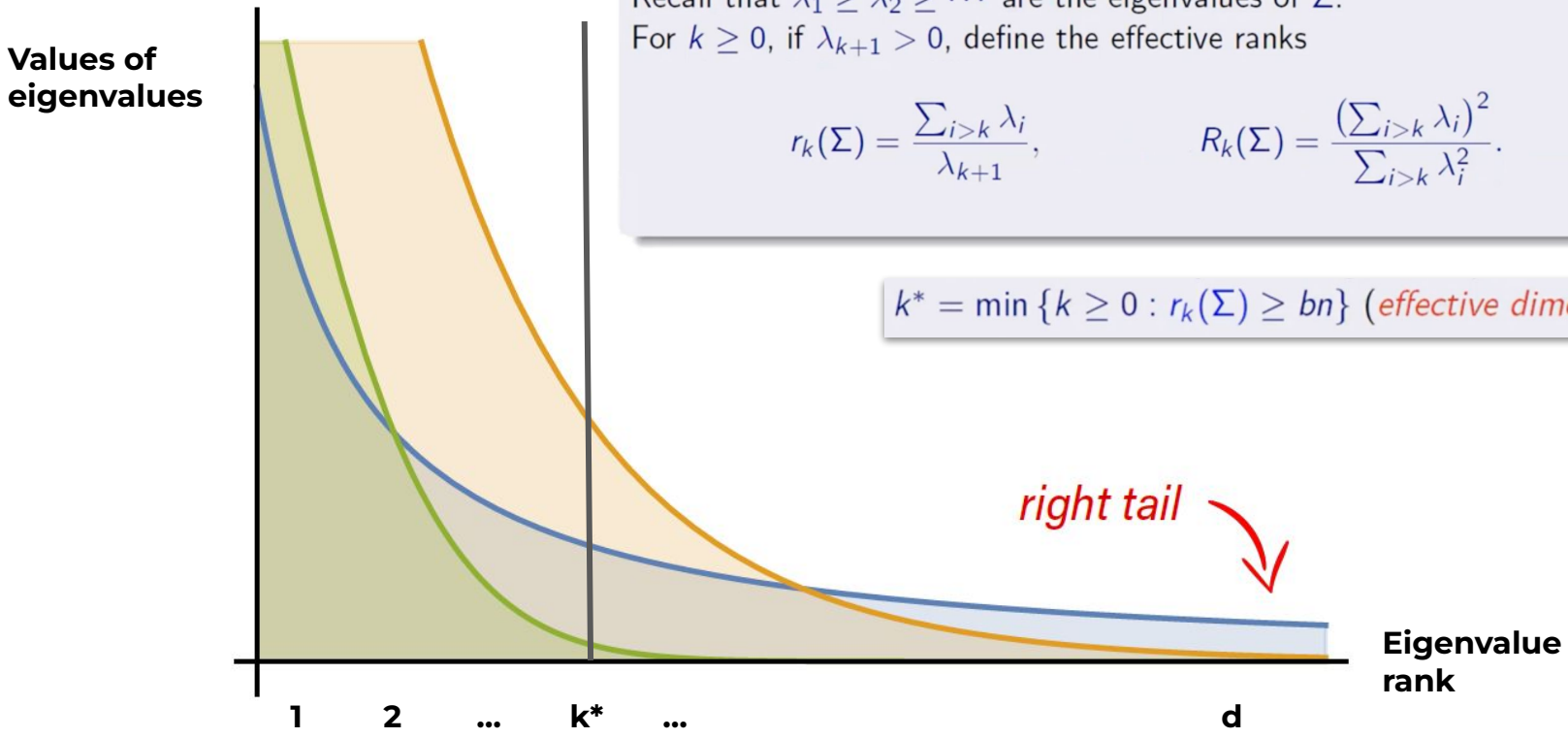
Overfitting regime: n ≪ d = p

Definition (Effective Ranks)

Recall that $\lambda_1 \geq \lambda_2 \geq \cdots$ are the eigenvalues of Σ.
For $k \geq 0$, if $\lambda_{k+1} > 0$, define the effective ranks

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*)



Theorem

For universal constants b, c, and any linear regression problem $(\theta^*, \sigma^2, \Sigma)$ with $\lambda_n > 0$, if $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ (*effective dimension*),

1. With high probability,

$$R(\hat{\theta}) \leq c\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right),$$

2. With some independence properties,

$$\mathbb{E}R(\hat{\theta}) \geq \frac{1}{c}\left(\text{bias}(\theta^*, \Sigma, n) + \sigma^2 \min\left\{\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}, 1\right\}\right).$$

$$\text{bias}(\theta^*, \Sigma, n) = \|\theta^*_{k+1:\infty}\|^2_{\Sigma_{k+1:\infty}} + \|\theta^*_{1:k}\|^2_{\Sigma^{-1}_{1:k}}\left(\frac{\sum_{i>k}\lambda_i}{n}\right)^2.$$

Intuition

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

Intuition

We say $\{\Sigma_n\}$ is *asymptotically benign* if

Eigenvalues should decay fast so that their sum is o(n)

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

Small number of large eigenvectors

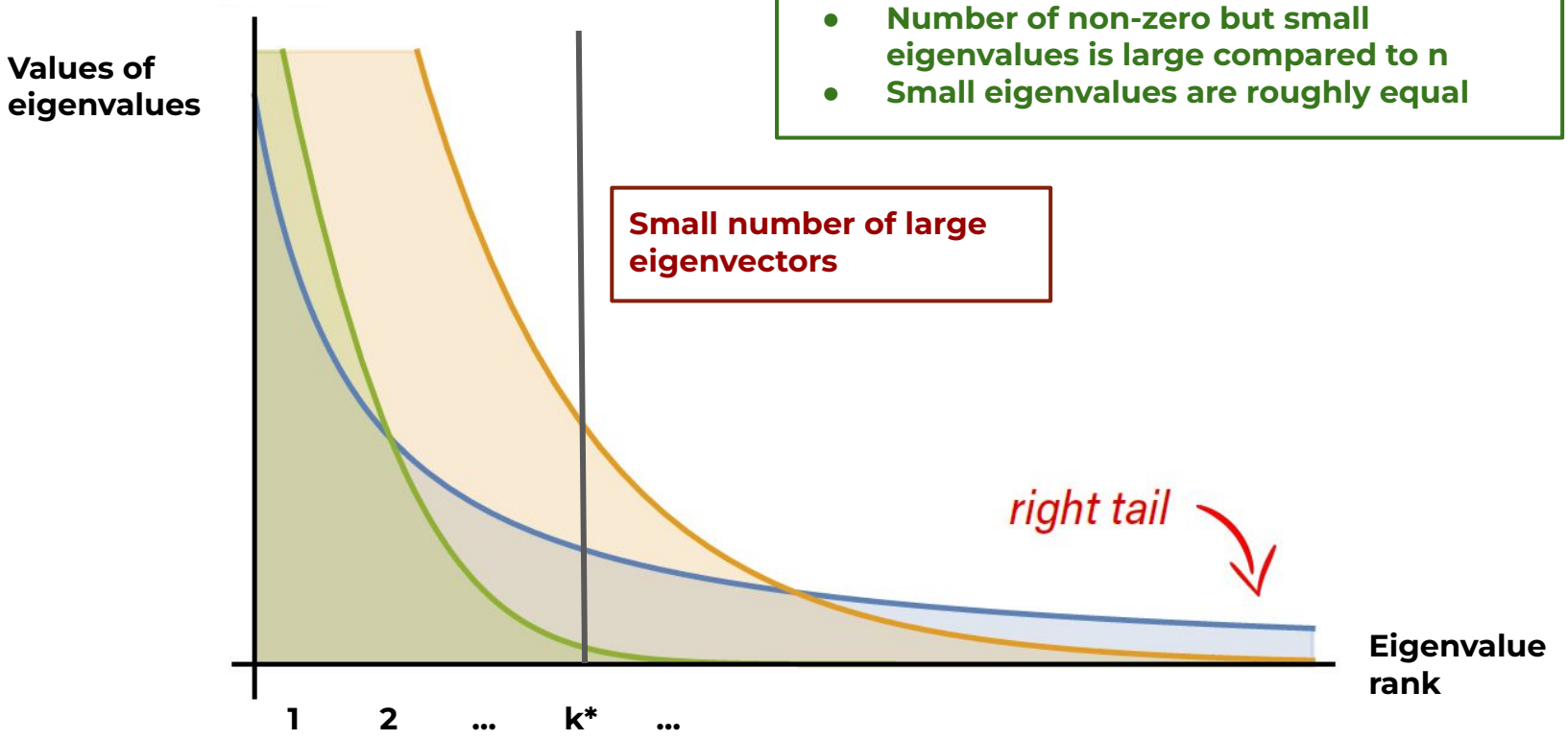
- **Number of non-zero but small eigenvalues is large compared to n**
- **Small eigenvalues are roughly equal**

Sum of eigenvalues must **almost** diverge

Example in picture

Eigenvalues should decay fast so that their sum is o(n)

- Number of non-zero but small eigenvalues is large compared to n
- Small eigenvalues are roughly equal



Interpolation for linear prediction

- Excess expected loss, has two components: (corresponding to $x^\top \theta^*$ and $y - x^\top \theta^*$)
  1. $\hat{\theta}$ is a distorted version of θ^*, because the sample $x_1, \ldots, x_n$ distorts our view of the covariance of x.

     Not a problem, even in high dimensions ($p > n$).
  2. $\hat{\theta}$ is corrupted by the noise in $y_1, \ldots, y_n$.

     Problematic.
- When can the label noise be hidden in $\hat{\theta}$ without hurting predictive accuracy?

One sentence answer

In **high-dimensional linear regression**,
perfectly fitting noisy data can still yield vanishing test error
(benign overfitting)
when **many comparable weak features (eigenvalues)** exists

Proof Intuition

1. Start with a bias–variance decomposition for the min-norm interpolator.
2. Variance term involves Tr(Σ) that captures how noise affects prediction accuracy.
3. Then use concentration (this is where *sub-gaussian assumption is useful)*
   + spectral inequalities to bound each term.

Lemma 7. *The excess risk of the minimum norm estimator satisfies*

$$R(\hat{\theta}) \leq 2\theta^{*\top} B\theta^* + c\sigma^2 \log\frac{1}{\delta}\operatorname{tr}(C)$$

with probability at least $1-\delta$ *over* ϵ, *and*

$$\mathbb{E}_{\boldsymbol{\epsilon}} R(\hat{\theta}) \geq \theta^{*\top} B\theta^* + \sigma^2 \operatorname{tr}(C),$$

where

$$B = \left(I - X^\top \left(XX^\top\right)^{-1} X\right)\Sigma\left(I - X^\top \left(XX^\top\right)^{-1} X\right),$$
$$C = \left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}.$$

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}} + \frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)}\right) = 0,$$

where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$.

Proof Intuition

Inevitable error that comes from small training size

The contribution of **overfitting to noise** in the direction of **important** eigenvectors

The contribution of **overfitting to noise** in the direction of **unimportant** eigenvectors

We say $\{\Sigma_n\}$ is *asymptotically benign* if

$$\lim_{n\to\infty}\left(\|\Sigma_n\|\sqrt{\frac{r_0(\Sigma_n)}{n}}+\frac{k_n^*}{n}+\frac{n}{R_{k_n^*}(\Sigma_n)}\right)=0,$$

where $k_n^*=\min\{k\geq 0: r_k(\Sigma_n)\geq bn\}$.

Example: *Finite dimension*, fast λ_i decay, plus isotropic noise

If

$$\lambda_{k,n} = \begin{cases} e^{-k} + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise,} \end{cases}$$

then Σ_n is benign iff

- $p_n = \omega(n)$,
- $\epsilon_n p_n = o(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$. $(n \ge 40 \implies ne^{-n} < 2^{-52})$

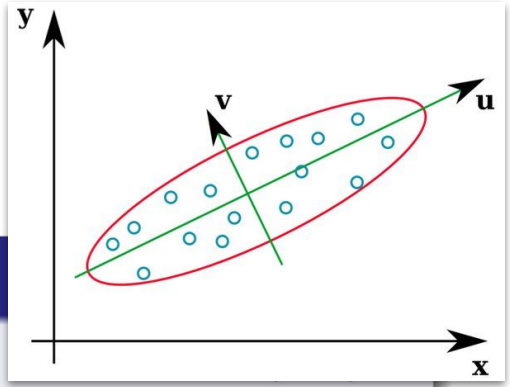
Furthermore, for $p_n = \Omega(n)$ and $\epsilon_n p_n = \omega(ne^{-n})$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Generic phenomenon:

quickly converging λ_i plus noise in all directions, $p_n \gg n$.

Very brittle for **adversarial** noise



Label noise appears in $\hat{\theta}$

We can find a unit norm Δ
such that perturbing an input x by Δ changes the output enormously:
even if $\Delta^\top \theta^* = 0$,

$$\left\|(x+\Delta)^\top \hat{\theta} - x^\top \hat{\theta}\right\|^2 \geq \frac{\sigma}{\sqrt{\lambda_{k^*+1}}} \geq \sqrt{\frac{n}{\mathrm{tr}(\Sigma)}}\sigma.$$

Benign overfitting leads to huge sensitivity.

Conclusion

Far from the regime of a **tradeoff** between fit to training data and complexity.

In linear regression, **a long, flat tail of the covariance eigenvalues** is ***necessary*** and ***sufficient*** for the **minimum norm interpolant** to predict well:

The noise is hidden in many unimportant directions.

Questions

Beyond **minimum euclidean norm interpolant**?

What is the approximately equivalent approach for analyzing **deep neural networks**?

Thank you!

Presentation on

Beyond Benign Overfitting in Nadaraya-Watson Interpolators

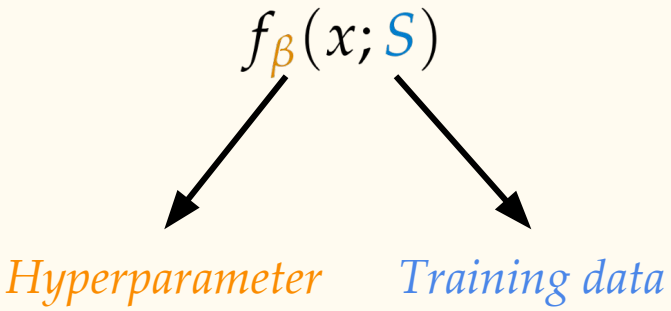
Daniel Barzilai* Guy Kornowski* Ohad Shamir
Weizmann Institute of Science
{daniel.barzilai,guy.kornowski,ohad.shamir}@weizmann.ac.il

October 22, 2025

by Anay Mehrotra

Nadaraya–Watson (NW) estimator

Nadaraya–Watson (NW) estimator [Nadaraya, 1964; Watson, 1964]



Nadaraya–Watson (NW) estimator

Nadaraya–Watson (NW) estimator [Nadaraya, 1964; Watson, 1964]

$$f_\beta(x; S) := \begin{cases} \operatorname{sign}\left(\sum_i \frac{y_i}{\|x - x_i\|^\beta}\right) & \text{if } x \notin S\,, \\ y_i & \text{if } x \in S \text{ and } x = x_i\,. \end{cases}$$

Hyperparameter *Training data*

➔ Nearest-neighbour based classification rule

➔ Displays benign overfitting (details next) [Devroye, Györfi, and Krzyżak, 1998]

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^{\star}(x)$

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^{\star}(x)$

Noisy training data: Each label y is flipped to $1 - y$ with probability p

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^\star(x)$

Noisy training data: Each label y is flipped to $1 - y$ with probability p

Testing error: Error is evaluated on *clean data*

$$\mathrm{Err}(f(S_p)) := \Pr_{x \sim \mathcal{D}}[f(x; S_p) \neq f^\star(x)]$$

Benign Overfitting in Classification

Sample distribution: Features $x \sim D$ *and* labels are $y = f^\star(x)$

Noisy training data: Each label y is flipped to $1 - y$ with probability p

Testing error: Error is evaluated on *clean data*

$$\mathrm{Err}(f(S_p)) := \Pr_{x \sim \mathcal{D}}[f(x; S_p) \neq f^\star(x)]$$

Theorem [Devroye, Györfi, and Krzyżak, 1998] If D has d-dimensional support. Then, the NW-estimator with $\beta = d$, despite noise $p \in [0, 1/2)$ achieves:

$$\mathrm{Err}(f; S_p) \to 0 \qquad \text{as} \qquad \left|S_p\right| \to \infty$$

Benign Overfitting in Classification

Theorem [Devroye, Györfi, and Krzyżak, 1998] If D has d-dimensional support. Then, the NW-estimator with $\beta = d$, despite noise $p \in [0, 1/2)$ achieves:

$$\mathrm{Err}(f_d; S_p) \to 0 \qquad \text{as} \qquad |S_p| \to \infty$$

- ➔ Even though ***f*** *fits the noise in the data exactly,* it still *generalizes to clean data*
- ➔ Benign over-fitting is *not an entirely new observation!*
- ➔ Similar estimators also analyzed for, e.g., *regression*. *Do they benignly overfit too?*

Benign Overfitting in Classification *Continued*

Theorem **[Barzilai, Kornowski, Shamir, NeurIPS'25]** If D has d-dimensional support. Then, the NW-estimator, under noise $p \in [0, 1/2)$, as $\left|S_p\right| \to \infty$

- If $\beta < d$, then $\mathrm{Err}(f_\beta; S_p) \to \Omega(1)$

Benign Overfitting in Classification *Continued*

Theorem [Barzilai, Kornowski, Shamir, NeurIPS'25] If D has d-dimensional support. Then, the NW-estimator, under noise $p \in [0, 1/2)$, as $\left|S_p\right| \to \infty$

- If $\beta < d$, then $\mathrm{Err}(f_\beta; S_p) \to \Omega(1)$
- If $\beta = d$, then $\mathrm{Err}(f_d; S_p) \to 0$

Benign Overfitting in Classification *Continued*

Theorem [Barzilai, Kornowski, Shamir, NeurIPS'25] If D has d-dimensional support. Then, the NW-estimator, under noise $p \in [0, 1/2)$, as $\left|S_p\right| \to \infty$

- If $\beta < d$, then $\mathrm{Err}(f_\beta; S_p) \to \Omega(1)$
- If $\beta = d$, then $\mathrm{Err}(f_d; S_p) \to 0$
- If $\beta > d$, then $\mathrm{Err}(f_\beta; S_p) \to [p^{O(1)}, O(p)]$

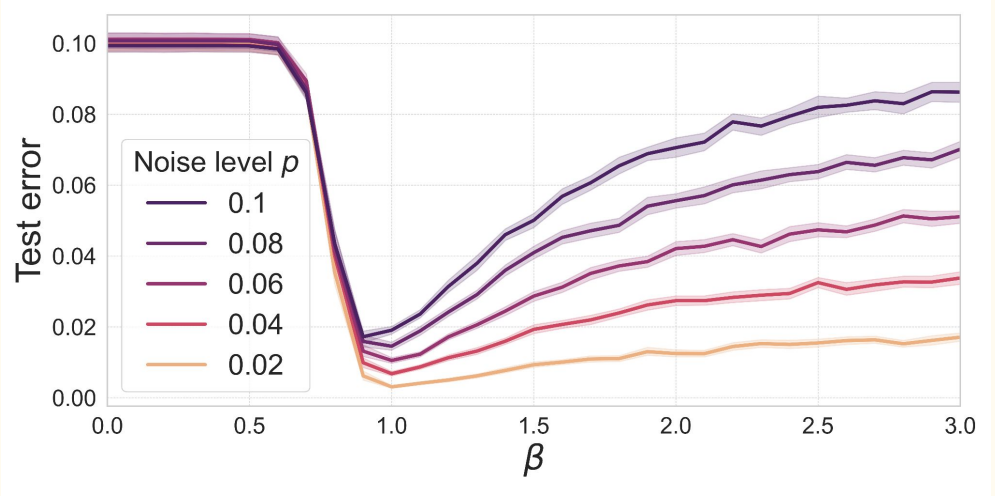
➔ Benign overfitting for the NW estimator is *fragile…*

➔ Right hyperparameter choice depends on the *ambient data dimension*

Empirical Results: 1-Dimensional Data

$\mathcal{D} := \text{Uniform}[0, 1]$

$f^\star(x) := \mathbb{1}\{x \in [0, 1/4]\}$

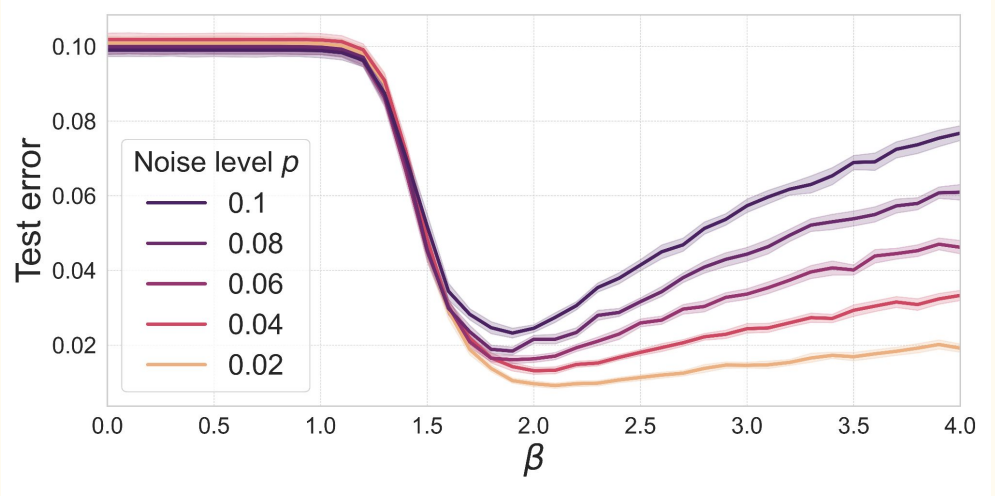


Empirical Results: 2-Dimensional Data

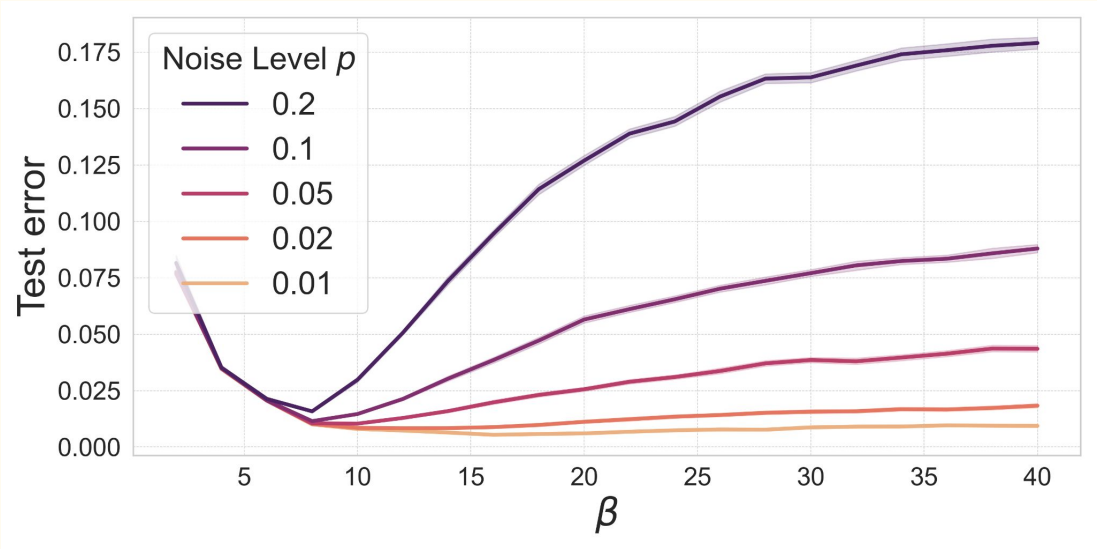
$$A := \left\{ x = (x_1, x_2, x_3) \in \mathbb{S}^2 \;\middle|\; x_3 > \frac{\sqrt{3}}{2} \right\},$$

$$\mathcal{D} = \frac{1}{10} \cdot \mathrm{Unif}(A) + \frac{9}{10} \cdot \mathrm{Unif}\left(\mathbb{S}^2 \setminus A\right)$$

$$f^\star(x) := \mathbb{1}\{x \notin A\}$$



Empirical Results: MNIST



➔ [Pope et al., ICLR'21] estimated MNIST's intrinsic dimension to be in [8, 15]!