Flattening the Path to Generalization in LLMs

Ting Han and Prof. Dr. Michael Kamp

Lamarr Institute, TU Dortmund and Institute of Artificial Intelligence in Medicine (IKIM)

Partner institutions:

technische universität dortmund

Fraunhofer IAIS

UNIVERSITÄT BONN

Fraunhofer IML

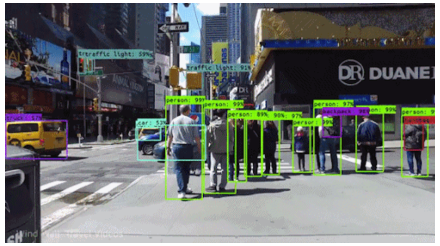
Institutionally funded by:

Bundesministerium für Forschung, Technologie und Raumfahrt

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

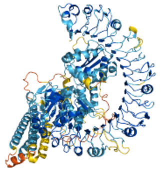
AI/LLMs are EVERYWHERE!

Machine Learning



Ilija Mihajlovic, 2019

Computer Vision algorithms can now surpass 99% human accuracy in image recognition tasks.

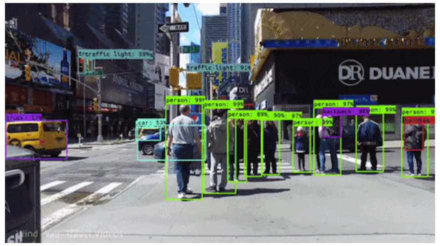


AlphaFold

AlphaFold, developed by DeepMind, has solved the 50-year-old protein folding problem with remarkable accuracy.

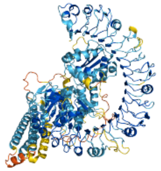
ChatGPT reached 1 million users in 5 days.

Machine Learning



Ilija Mihajlovic, 2019

Computer Vision algorithms can now surpass 99% human accuracy in image recognition tasks.



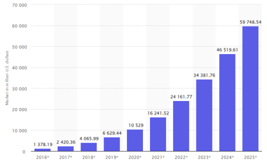
AlphaFold

AlphaFold, developed by DeepMind, has solved the 50-year-old protein folding problem with remarkable accuracy.

ChatGPT reached 1 million users in 5 days.

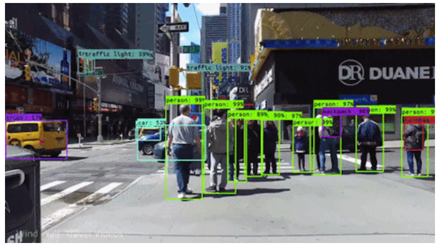
Machine Learning

AI-powered personalization increases sales by 10-15%.



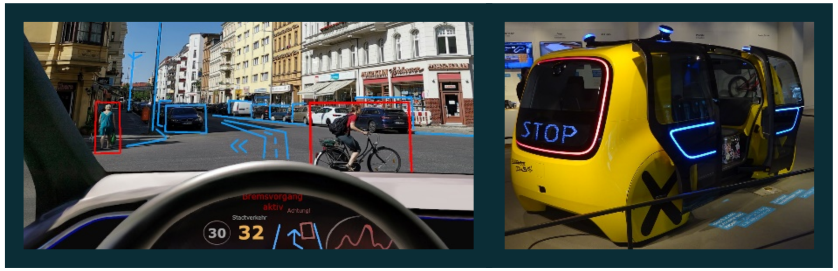
Statista, 2024

In 2022, the global AI market was valued at over $100 billion and is projected to reach $190 billion by 2025.

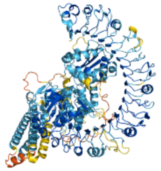


Ilija Mihajlovic, 2019

Computer Vision algorithms can now surpass 99% human accuracy in image recognition tasks.



Machine Learning

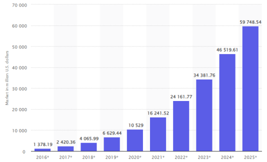


AlphaFold

AlphaFold, developed by DeepMind, has solved the 50-year-old protein folding problem with remarkable accuracy.

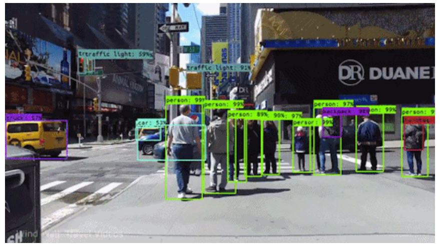
ChatGPT reached 1 million users in 5 days.

AI-powered personalization increases sales by 10-15%.



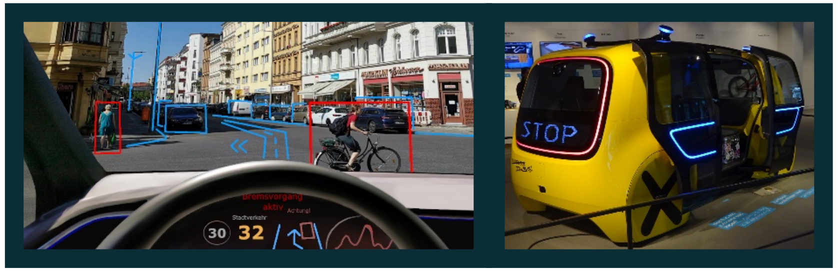
Statista, 2024

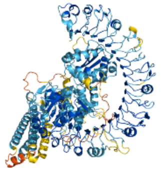
In 2022, the global AI market was valued at over $100 billion and is projected to reach $190 billion by 2025.



Ilija Mihajlovic, 2019

Computer Vision algorithms can now surpass 99% human accuracy in image recognition tasks.





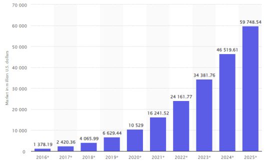
AlphaFold

AlphaFold, developed by DeepMind, has solved the 50-year-old protein folding problem with remarkable accuracy.

ChatGPT reached 1 million users in 5 days.

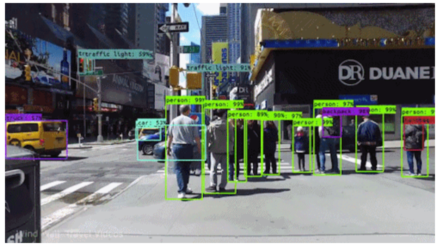
AI-powered personalization increases sales by 10-15%.

Machine Learning



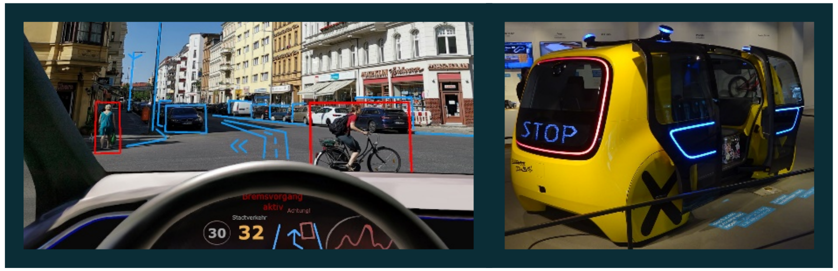
Statista, 2024

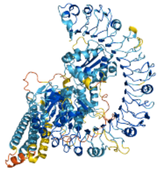
In 2022, the global AI market was valued at over $100 billion and is projected to reach $190 billion by 2025.



Ilija Mihajlovic, 2019

Computer Vision algorithms can now surpass 99% human accuracy in image recognition tasks.



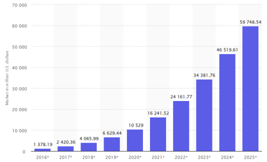


AlphaFold

AlphaFold, developed by DeepMind, has solved the 50-year-old protein folding problem with remarkable accuracy.

ChatGPT reached 1 million users in 5 days.

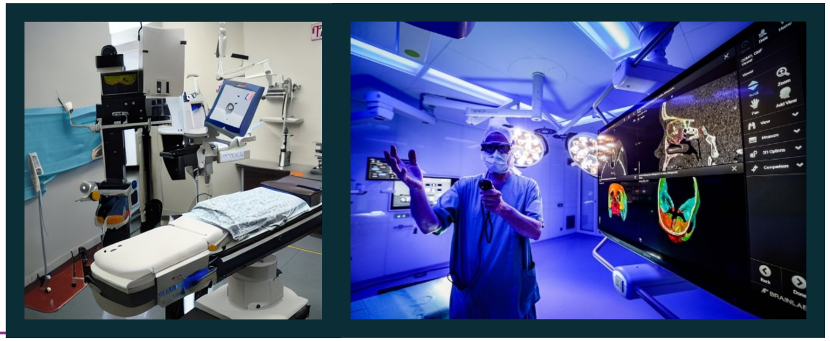
AI-powered personalization increases sales by 10-15%.



Statista, 2024

In 2022, the global AI market was valued at over $100 billion and is projected to reach $190 billion by 2025.

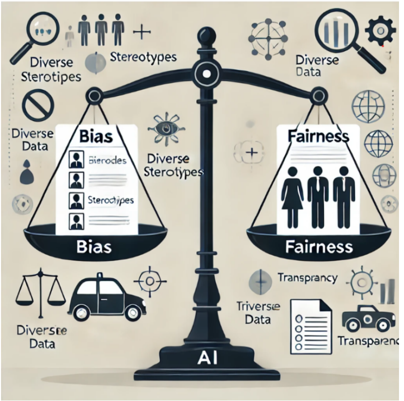
Machine Learning



Are AI/LLMs TRUELY everywhere? Not yet!

Trustwhothiess & Responsibility

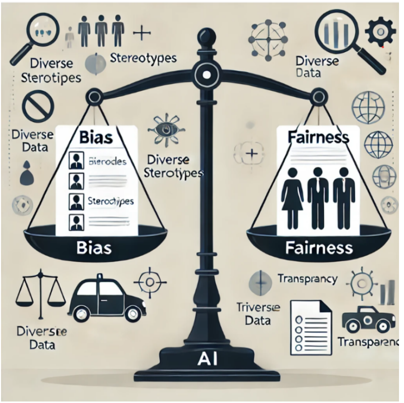
Bias & Fairness



Private data recovery

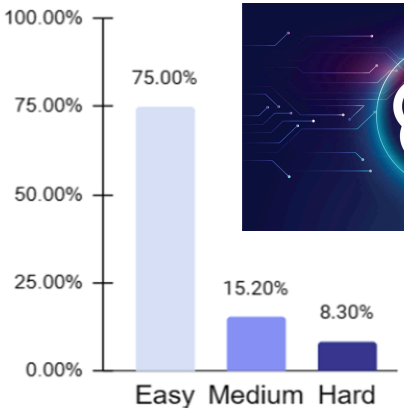
Are AI/LLMs TRUELY everywhere?

Bias & Fairness



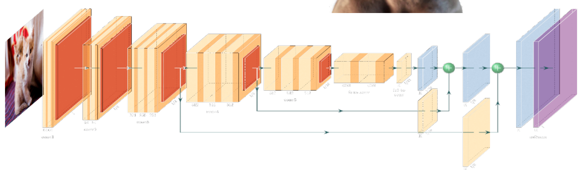
Trustwhothiess & Responsibility

Private data recovery



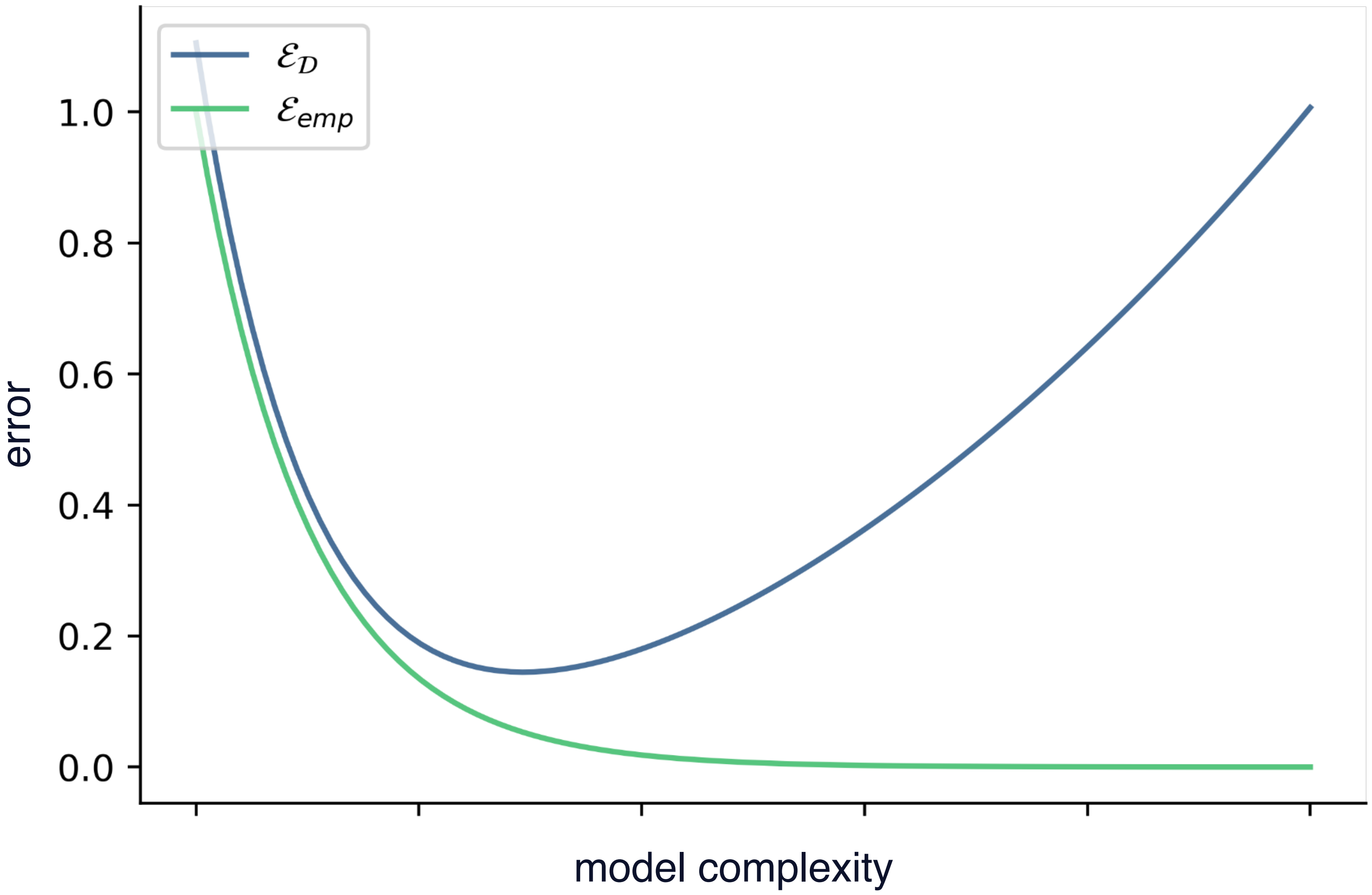
ChatGPT-4V performance on Medical VQA data

Interpretability?

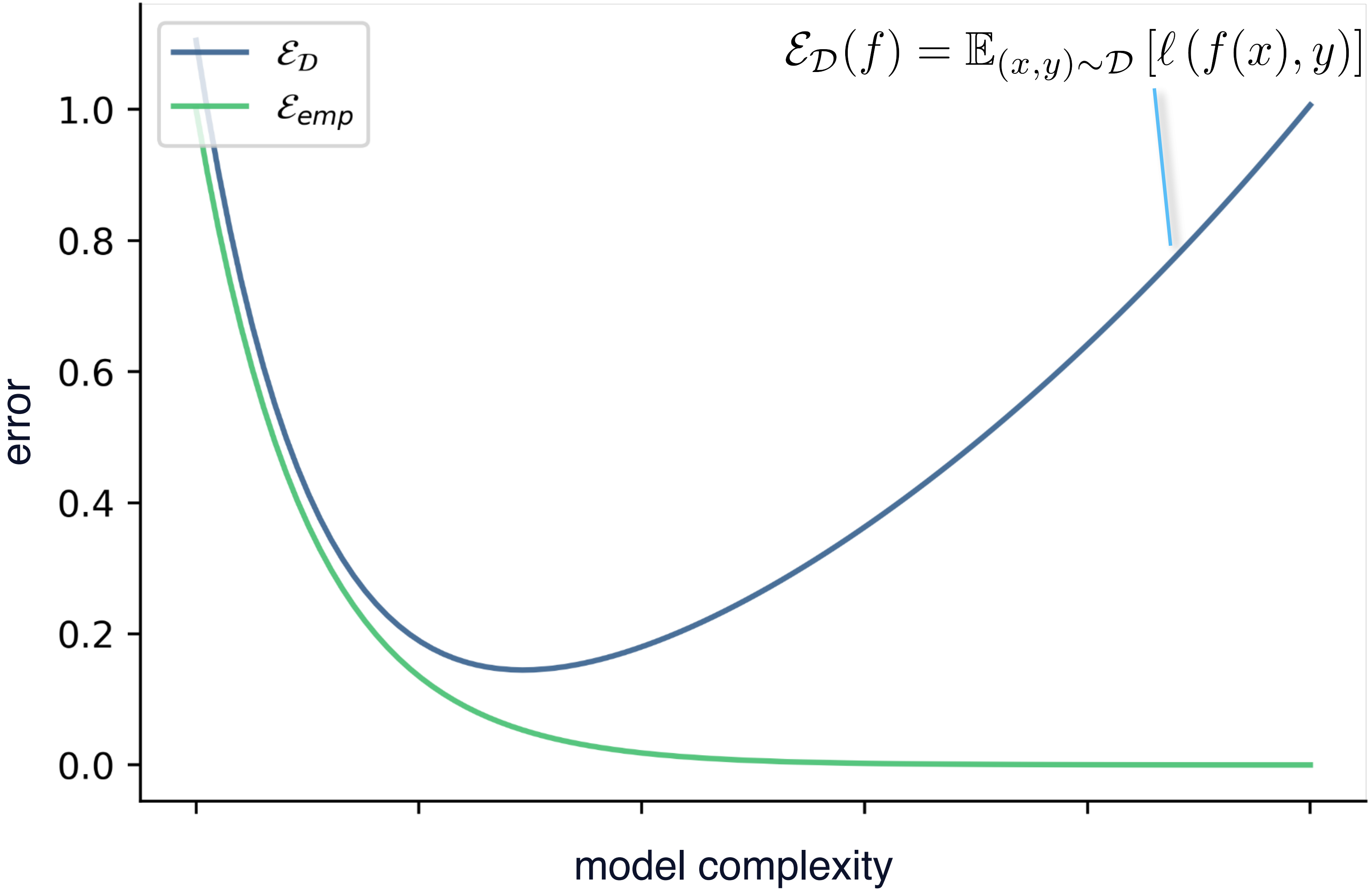


Generalization

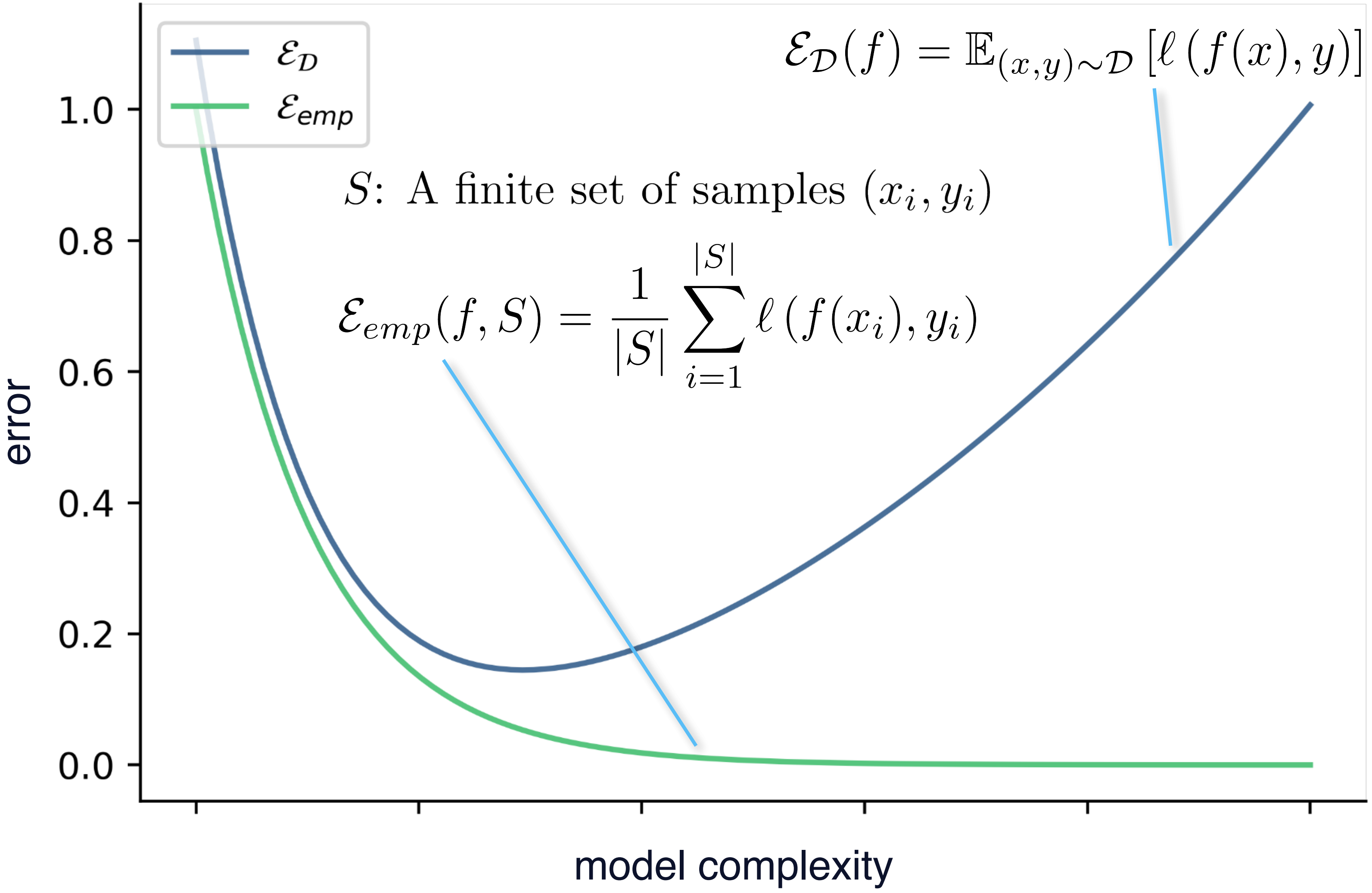
The Generalization Gap



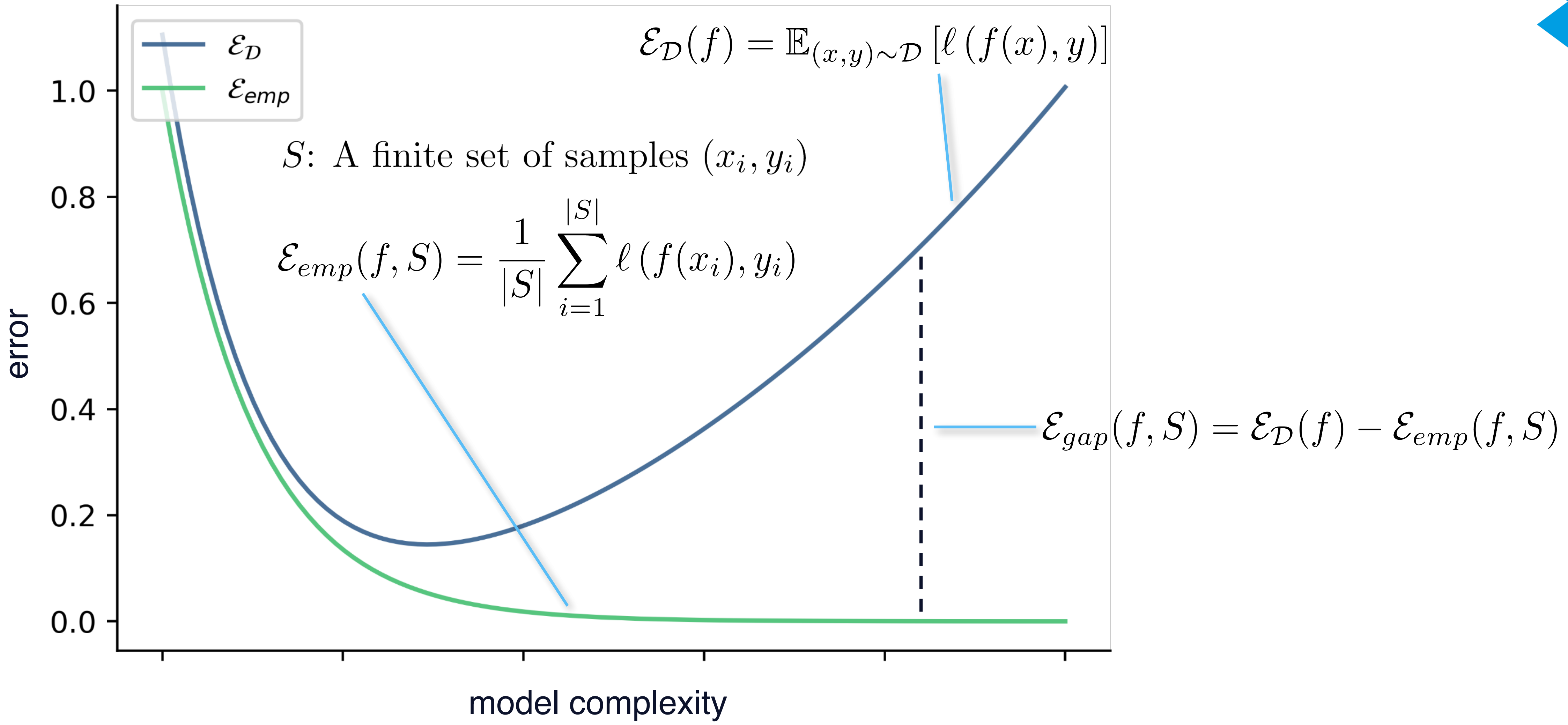
The Generalization Gap



The Generalization Gap



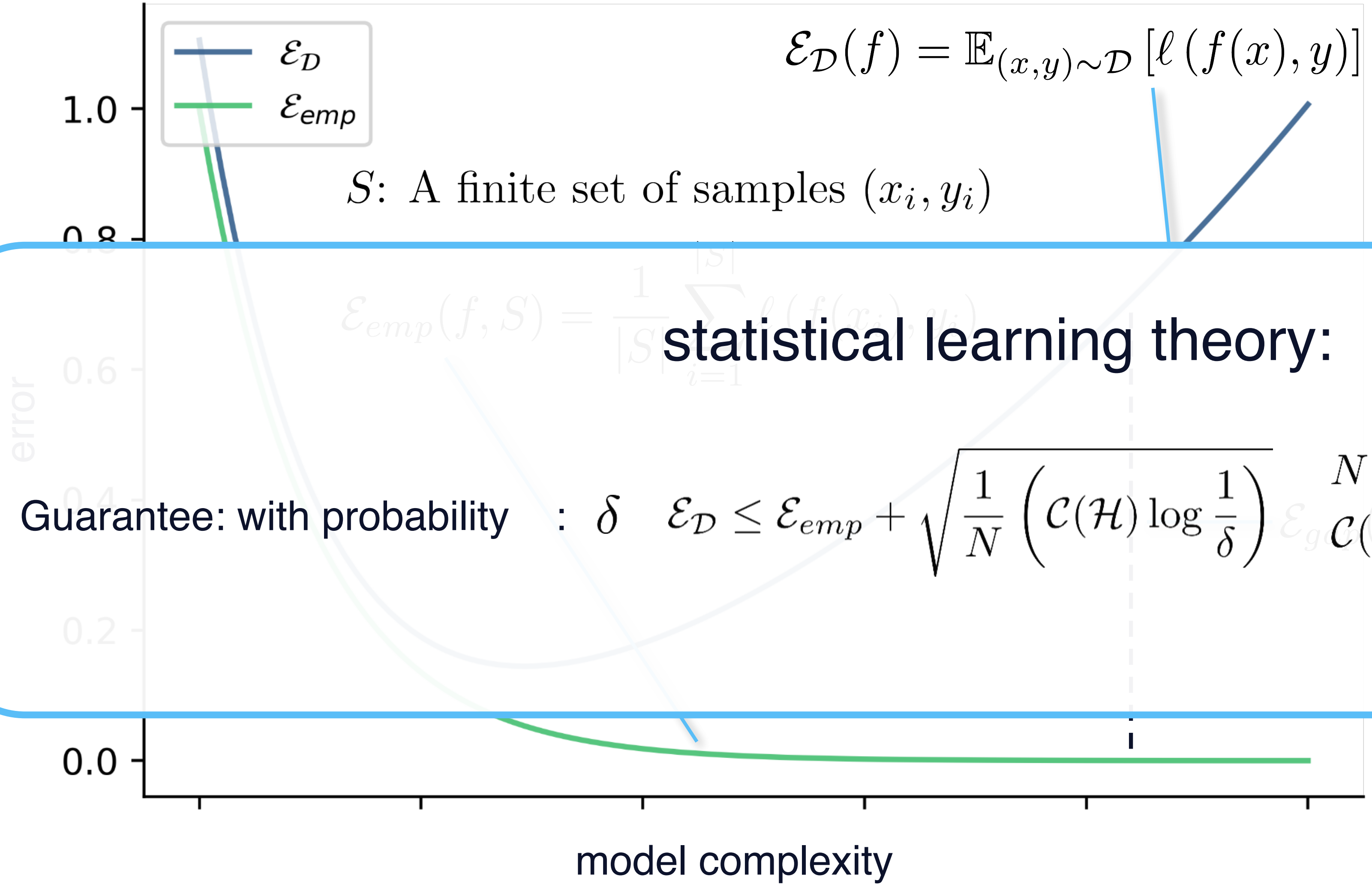
The Generalization Gap



The Generalization Gap

$$\mathcal{E}_{\mathcal{D}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(f(x), y\right)\right] = \mathbb{E}_{x\sim\mathcal{D}_x}\left[\ell\left(f(x), y(x)\right)\right]$$

S: A finite set of samples (x_i, y_i)



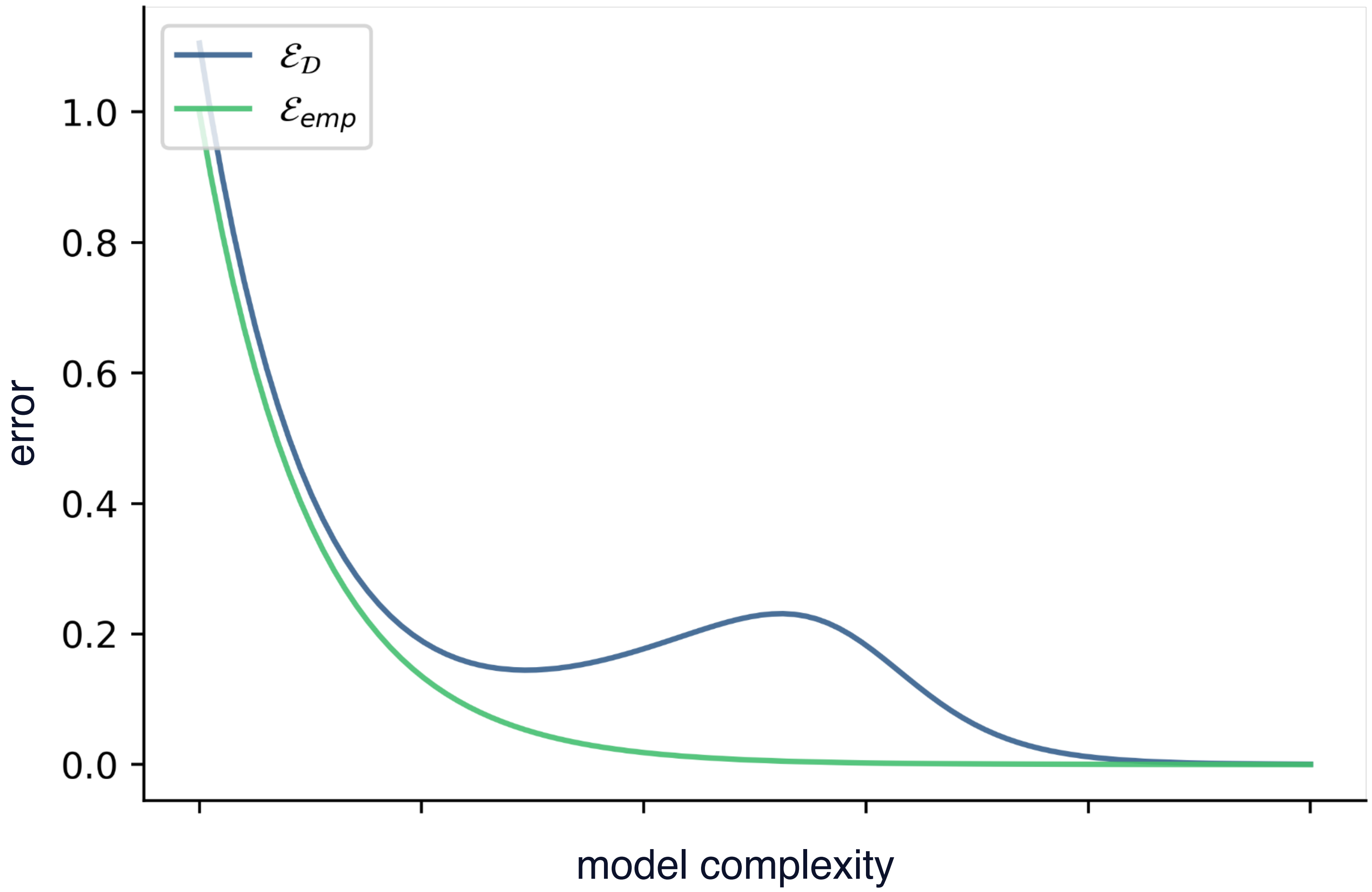
statistical learning theory:

Guarantee: with probability $: \delta$ $\quad \mathcal{E}_{\mathcal{D}} \leq \mathcal{E}_{emp} + \sqrt{\frac{1}{N}\left(\mathcal{C}(\mathcal{H}) \log \frac{1}{\delta}\right)}$

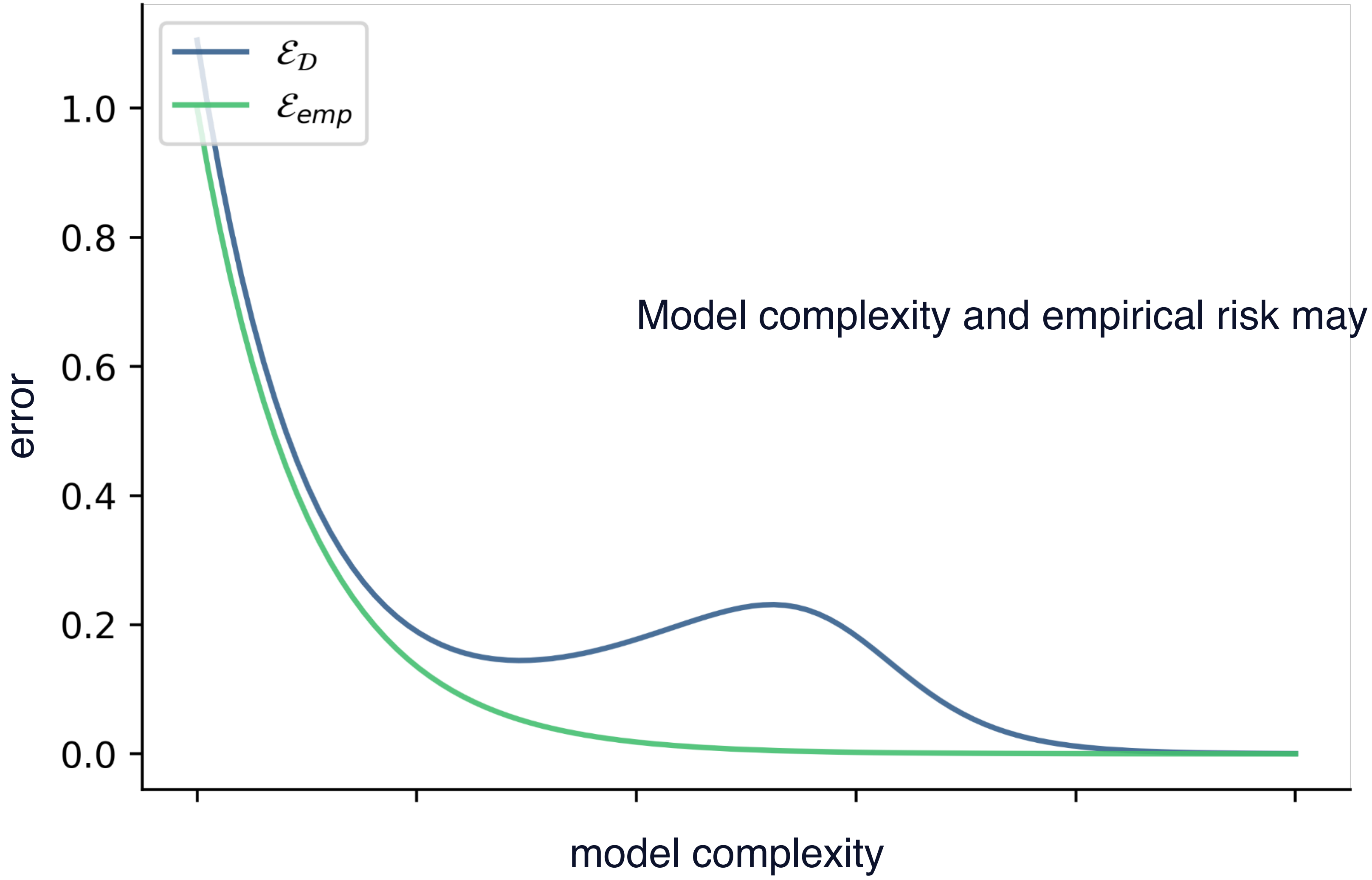
N: number of training samples

$\mathcal{C}(\mathcal{H})$: complexity of model class $\mathcal{H}$

Double Descent

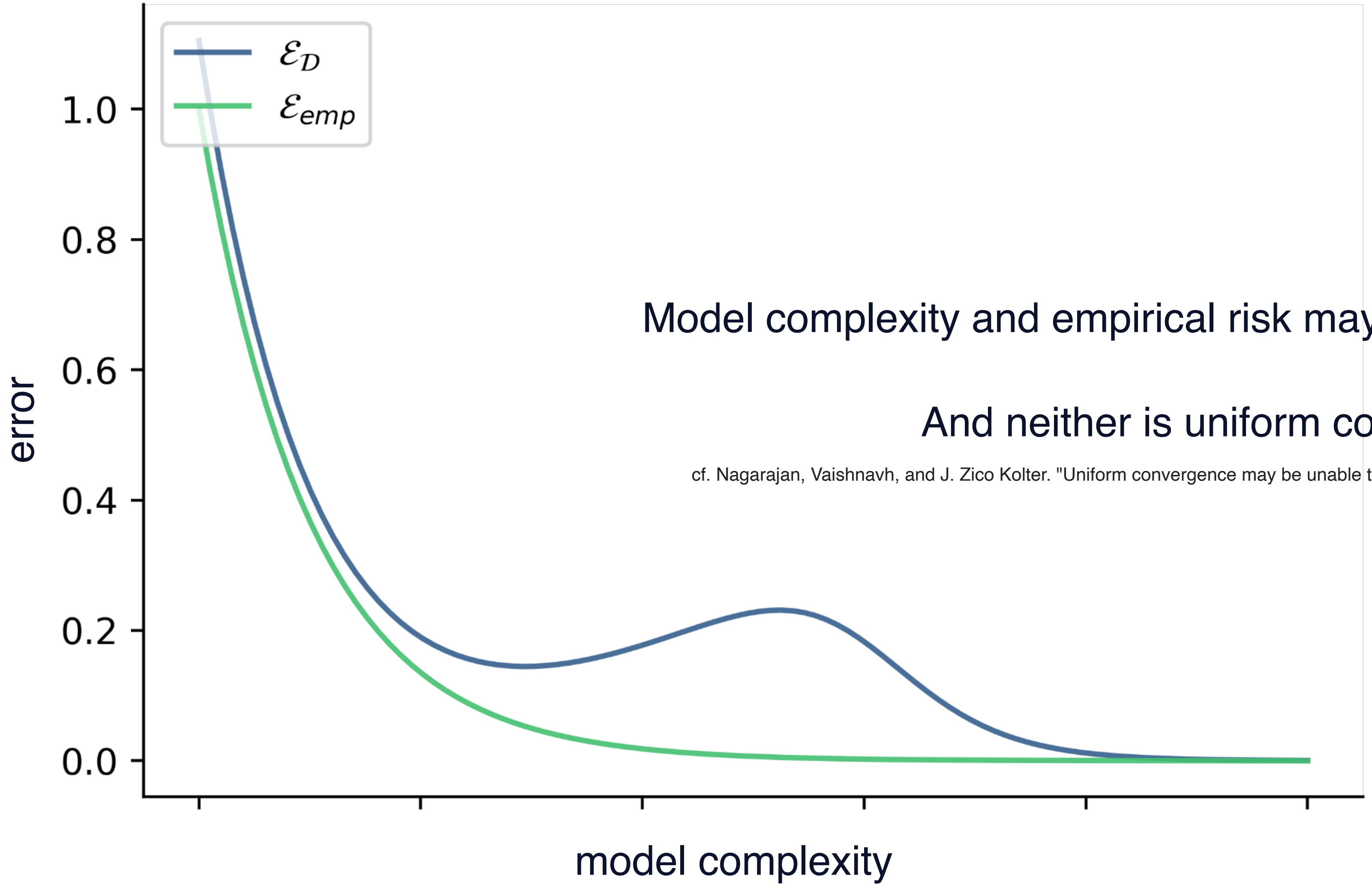


Double Descent



Model complexity and empirical risk may not be informative anymore!

Double Descent



Model complexity and empirical risk may not be informative anymore!

And neither is uniform convergence…

cf. Nagarajan, Vaishnavh, and J. Zico Kolter. "Uniform convergence may be unable to explain generalization in deep learning." NeurIPS, 2019.

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

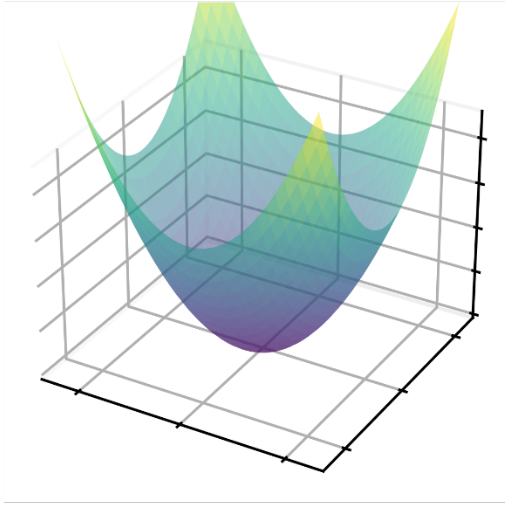
Feature Robustness and Relative Flatness

Flatness of the Loss Curve

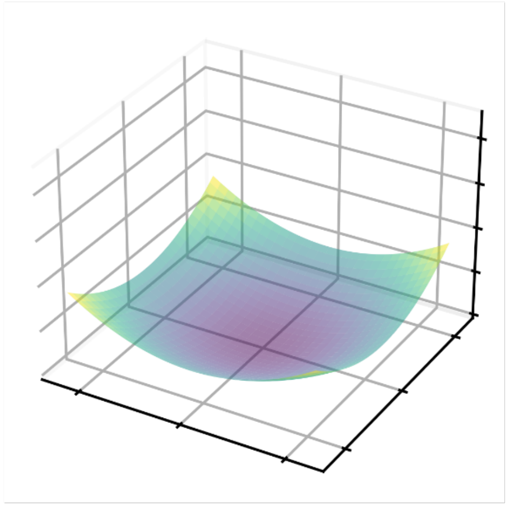
“Large region in weight space with the property that each weight vector from that region leads to similar small error“

Hochreiter and Schmidhuber. “Simplifying neural nets by discovering flat minima.“ NIPS, 1995.

$$\mathcal{E}_{emp}(\mathbf{w} + \delta\mathbf{v}, S) \approx \mathcal{E}_{emp}(\mathbf{w}, S)$$

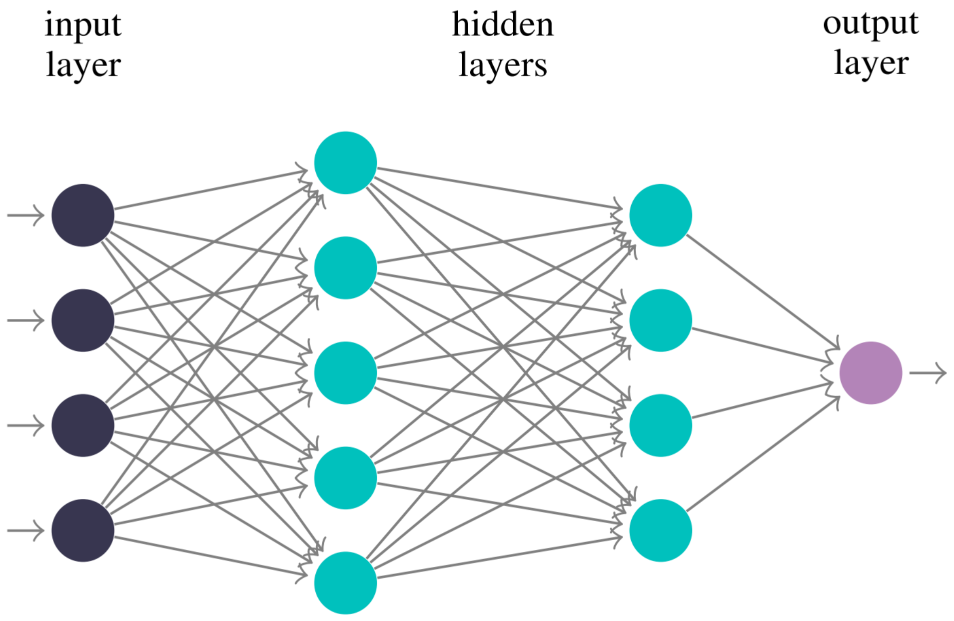


sharp



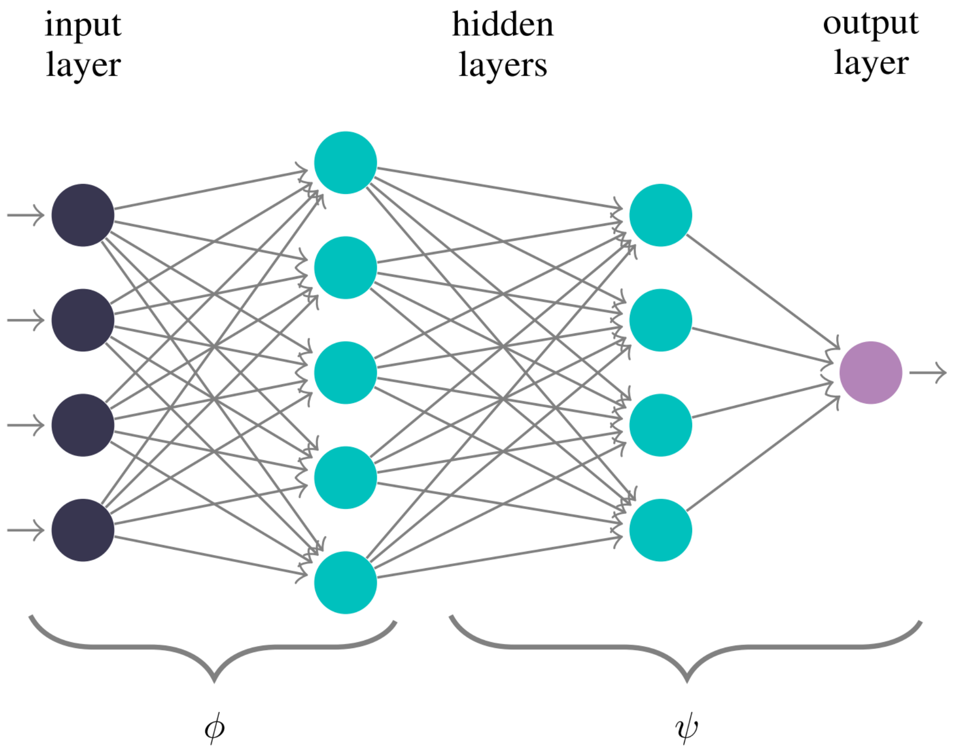
flat

Weights and Feature Space



Weights and Feature Space

$$f(x) = \psi(\mathbf{w}, \phi(x)) = g(\mathbf{w}\phi(x))$$



feature extractor

predictor

Feature Robustness

key insight:

$$g((\mathbf{w} + \delta\mathbf{v})\phi(x))$$

Feature Robustness

key insight:

$$g((\mathbf{w} + \delta\mathbf{v})\phi(x)) = g((\mathbf{w} + \delta\mathbf{w}A)\phi(x))$$

$$\phi(x) \in \mathbb{R}^m, \;\; A \in \mathbb{R}^{m \times m}$$

Feature Robustness

key insight:

$$
\begin{aligned}
g((\mathbf{w} + \delta \mathbf{v})\phi(x)) &= g((\mathbf{w} + \delta \mathbf{w} A)\phi(x)) \\
&= g(\mathbf{w}(\phi(x) + \delta A \phi(x)))
\end{aligned}
$$

$$\phi(x) \in \mathbb{R}^m, \;\; A \in \mathbb{R}^{m \times m}$$

Feature Robustness

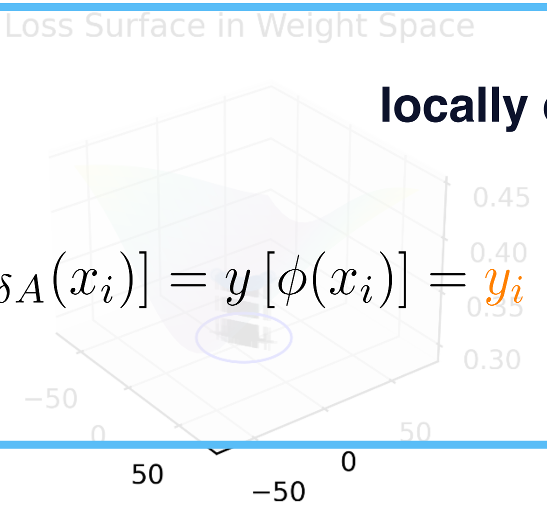
$$f(x) = \psi(\mathbf{w}, \phi(x)) = g(\mathbf{w}\phi(x))$$

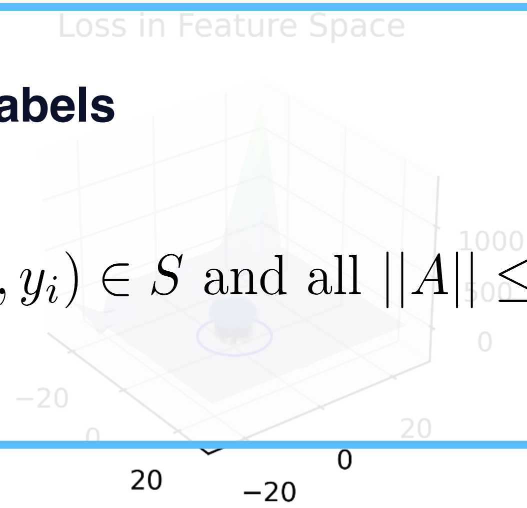
$$g\left(\left(\mathbf{w} \; {\color{red}+\delta \mathbf{w} A}\right) \phi(x)\right) = g\left(\mathbf{w}\left(\phi(x) \; {\color{red}+\delta A \phi(x)}\right)\right)$$

$$\mathcal{E}_{emp}(\mathbf{w} + \delta \mathbf{w} A, \phi(S)) - \mathcal{E}_{emp}(\mathbf{w}, \phi(S)) = \frac{1}{|S|} \sum_{i=1}^{|S|} \ell(\psi(\mathbf{w}, \phi_{\delta A}(x_i)), y_i) - \ell(\psi(\mathbf{w}, \phi(x_i)), y_i)$$

flatness

feature robustness





locally constant labels

$$y\left[\phi_{\delta A}(x_i)\right] = y\left[\phi(x_i)\right] = {\color{orange}y_i} \text{ for all } (x_i, y_i) \in S \text{ and all } ||A|| \leq 1$$

Feature Robustness

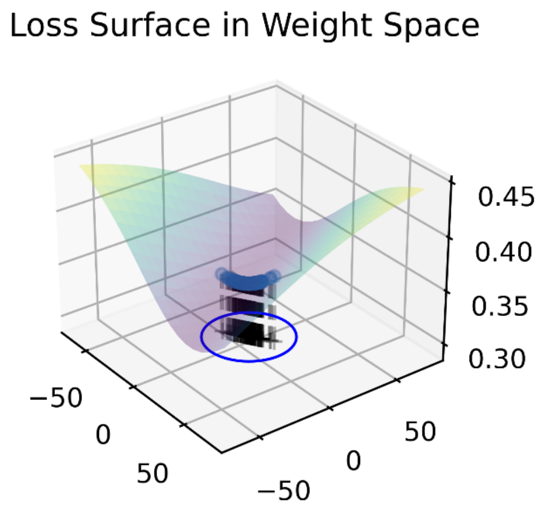
$$f(x) = \psi(\mathbf{w}, \phi(x)) = g(\mathbf{w}\phi(x))$$

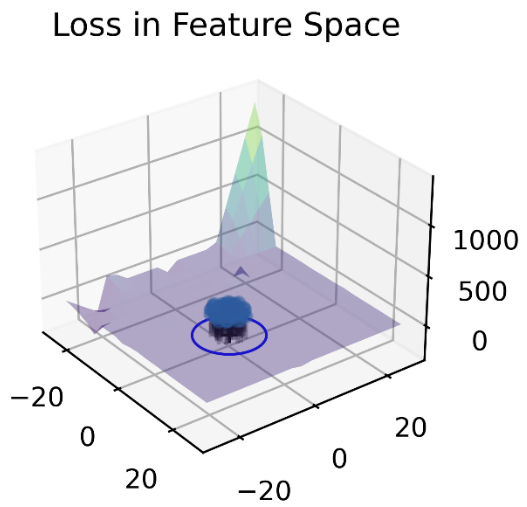
$$g\left(\left(\mathbf{w} \color{red}{+\delta \mathbf{w} A}\right) \phi(x)\right) = g\left(\mathbf{w}\left(\phi(x) \color{red}{+\delta A \phi(x)}\right)\right)$$

$$\mathcal{E}_{emp}(\mathbf{w} + \delta \mathbf{w} A, \phi(S)) - \mathcal{E}_{emp}(\mathbf{w}, \phi(S)) = \frac{1}{|S|} \sum_{i=1}^{|S|} \ell(\psi(\mathbf{w}, \phi_{\delta A}(x_i)), y_i) - \ell(\psi(\mathbf{w}, \phi(x_i)), y_i)$$

flatness

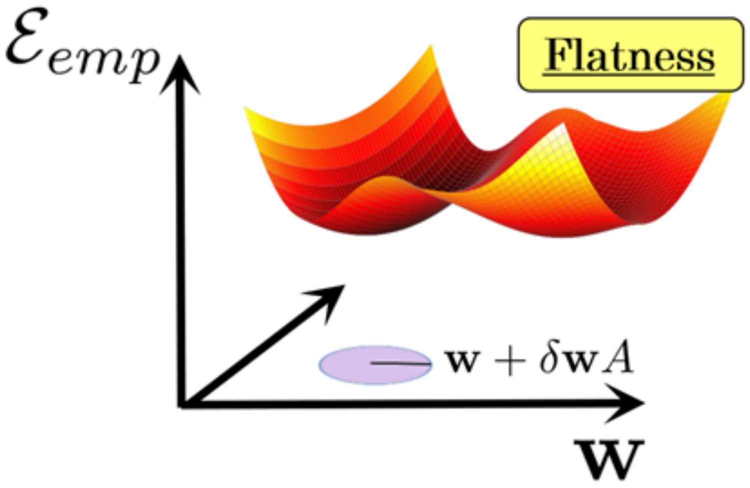
feature robustness





Relative Flatness

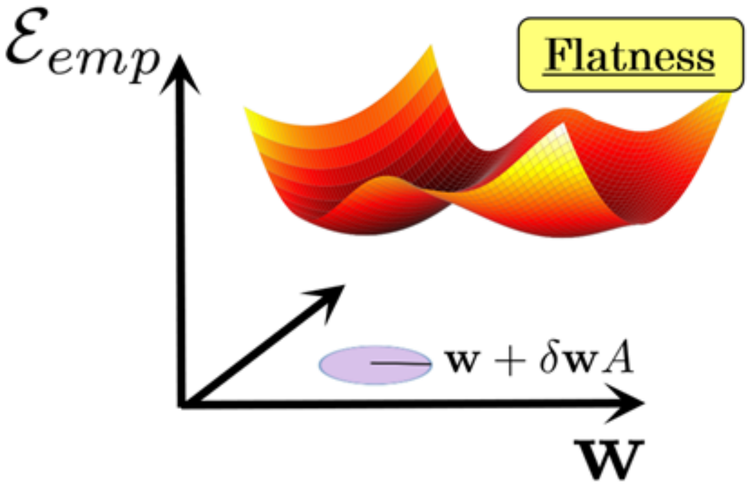
$$\mathcal{E}^{\phi}_{\mathcal{F}}(f, S, \mathcal{A}) = \mathbb{E}_{A\sim\mathcal{A}}\left[\frac{1}{|S|}\sum_{i=1}^{|S|}\ell(\psi(\phi_{\delta A}(x_i)), y(\phi_{\delta A}(x_i)) - \ell(f(x_i), y_i)\right]$$



Relative Flatness

$$\mathcal{E}^{\phi}_{\mathcal{F}}(f, S, \mathcal{A}) = \mathbb{E}_{A\sim\mathcal{A}}\left[\frac{1}{|S|}\sum_{i=1}^{|S|}\ell(\psi(\phi_{\delta A}(x_i)), y(\phi_{\delta A}(x_i)) - \ell(f(x_i), y_i)\right]$$

Second order **Taylor decomposition** of **feature robustness** and taking expectation over a sensibly chosen distribution on feature matrices results in a Hessian based flatness measure:



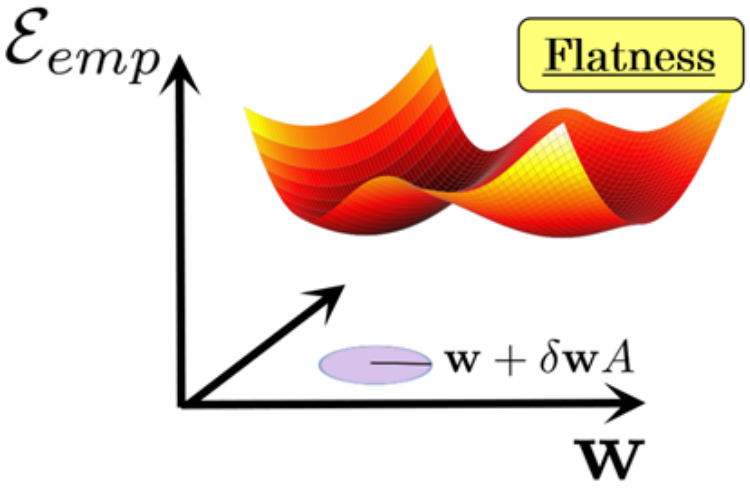
Relative Flatness

$$\mathcal{E}^{\phi}_{\mathcal{F}}(f, S, \mathcal{A}) = \mathbb{E}_{A\sim\mathcal{A}}\left[\frac{1}{|S|}\sum_{i=1}^{|S|}\ell(\psi(\phi_{\delta A}(x_i)), y(\phi_{\delta A}(x_i)) - \ell(f(x_i), y_i)\right]$$

Second order **Taylor decomposition** of **feature robustness** and taking expectation over a sensibly chosen distribution on feature matrices results in a Hessian based flatness measure:

$$H_{s,s'}(\mathbf{w}, \phi(S)) = \left[\frac{\partial^2 \mathcal{E}_{emp}(\mathbf{w}, \phi(S))}{\partial w_{s,t}\partial w_{s',t'}}\right]_{1\le t,t'\le m}$$

$$\mathbf{w}_s = (w_{s,t})_t \in \mathbb{R}^{1\times m}$$



Relative Flatness

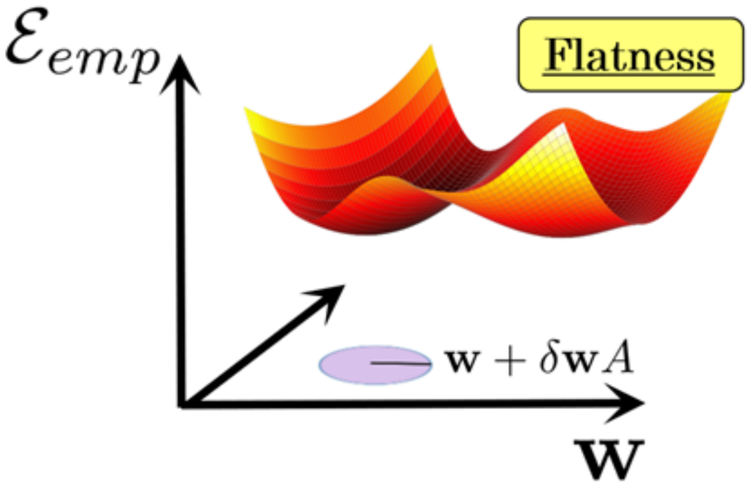
$$\mathcal{E}^{\phi}_{\mathcal{F}}(f, S, \mathcal{A}) = \mathbb{E}_{A\sim\mathcal{A}}\left[\frac{1}{|S|}\sum_{i=1}^{|S|}\ell(\psi(\phi_{\delta A}(x_i)), y(\phi_{\delta A}(x_i)) - \ell(f(x_i), y_i)\right]$$

Second order **Taylor decomposition** of **feature robustness** and taking expectation over a sensibly chosen distribution on feature matrices results in a Hessian based flatness measure:

$$H_{s,s'}(\mathbf{w}, \phi(S)) = \left[\frac{\partial^2 \mathcal{E}_{emp}(\mathbf{w}, \phi(S))}{\partial w_{s,t}\partial w_{s',t'}}\right]_{1\leq t,t'\leq m}$$

$$\mathbf{w}_s = (w_{s,t})_t \in \mathbb{R}^{1\times m}$$

$$\boxed{K^{\phi}_{Tr}(\mathbf{w}) = \sum_{s,s'=1}^{d}\langle \mathbf{w}_s, \mathbf{w}_{s'}\rangle \cdot Tr\left(H_{s,s'}\left(\mathbf{w}, \phi(S)\right)\right)}$$



Neural Collapse

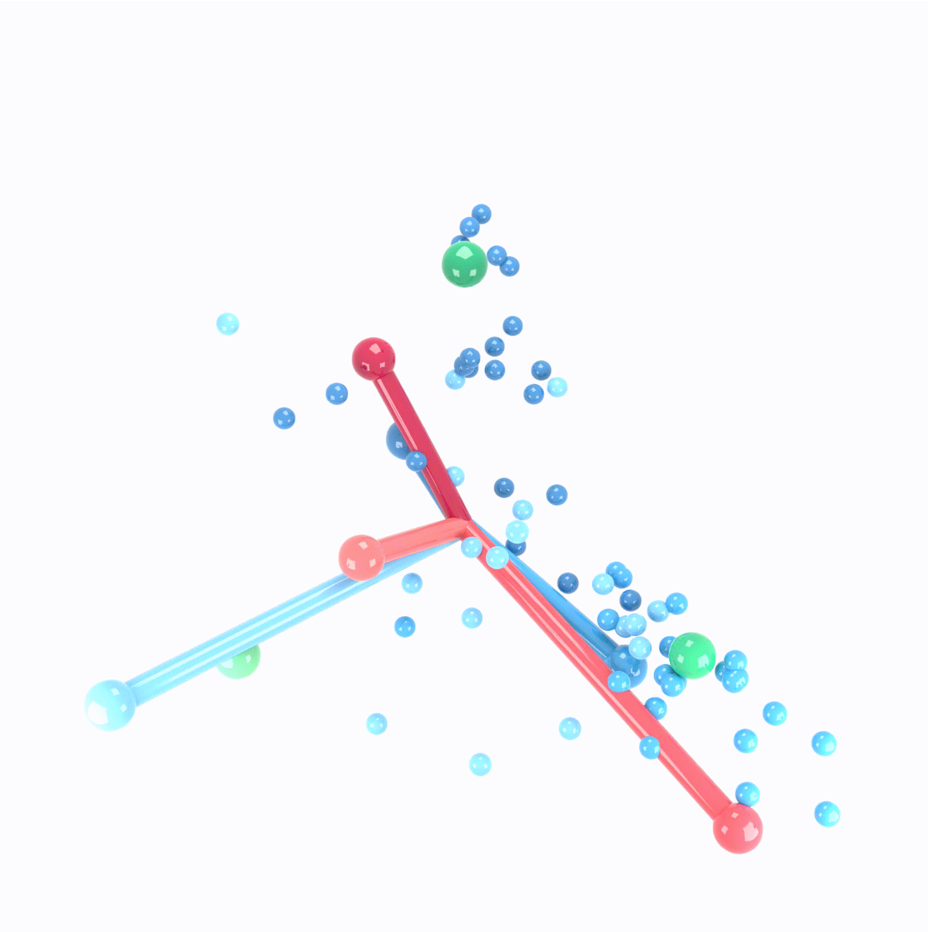
LAMARR
INSTITUTE FOR
MACHINE LEARNING
AND ARTIFICIAL
INTELLIGENCE

Neural Collapse

Neural Collapse is a phenomenon at the end of training where last-layer features and classifiers form a simple, symmetric structure.

- NC1: Variability Collapse: **Features within each class collapse tightly to their class mean.**
- NC2: Convergence to Simplex ETF
- NC3: Convergence to self-duality
- NC4: Nearest Class-Center Simplification

$$\mathrm{NCC} := \sum_{c \neq c'} \frac{V_c + V_{c'}}{2\|\mu_c - \mu_{c'}\|^2}$$

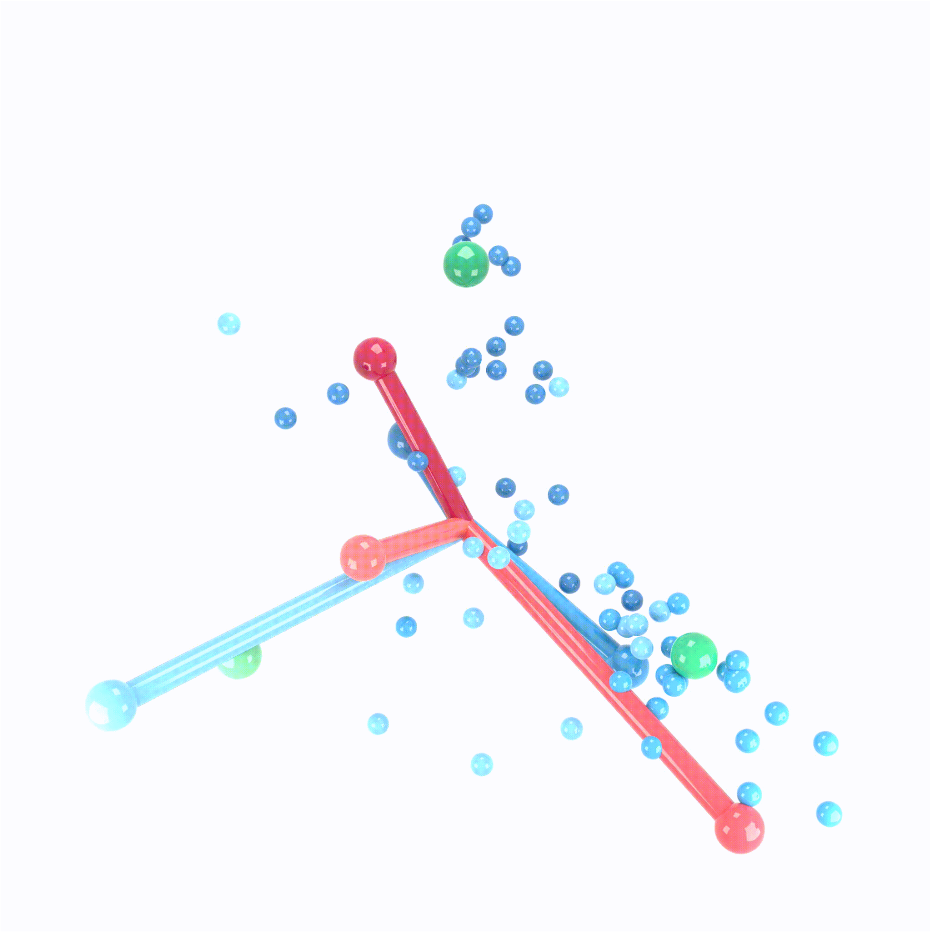


Neural Collapse

Neural Collapse is a phenomenon at the end of training where last-layer features and classifiers form a simple, symmetric structure.

- NC1: Variability Collapse: **Features within each class collapse tightly to their class mean.**
- NC2: Convergence to Simplex ETF
- NC3: Convergence to self-duality
- NC4: Nearest Class-Center Simplification

$$\mathrm{NCC} := \sum_{c \neq c'} \frac{V_c + V_{c'}}{2\|\mu_c - \mu_{c'}\|^2}$$



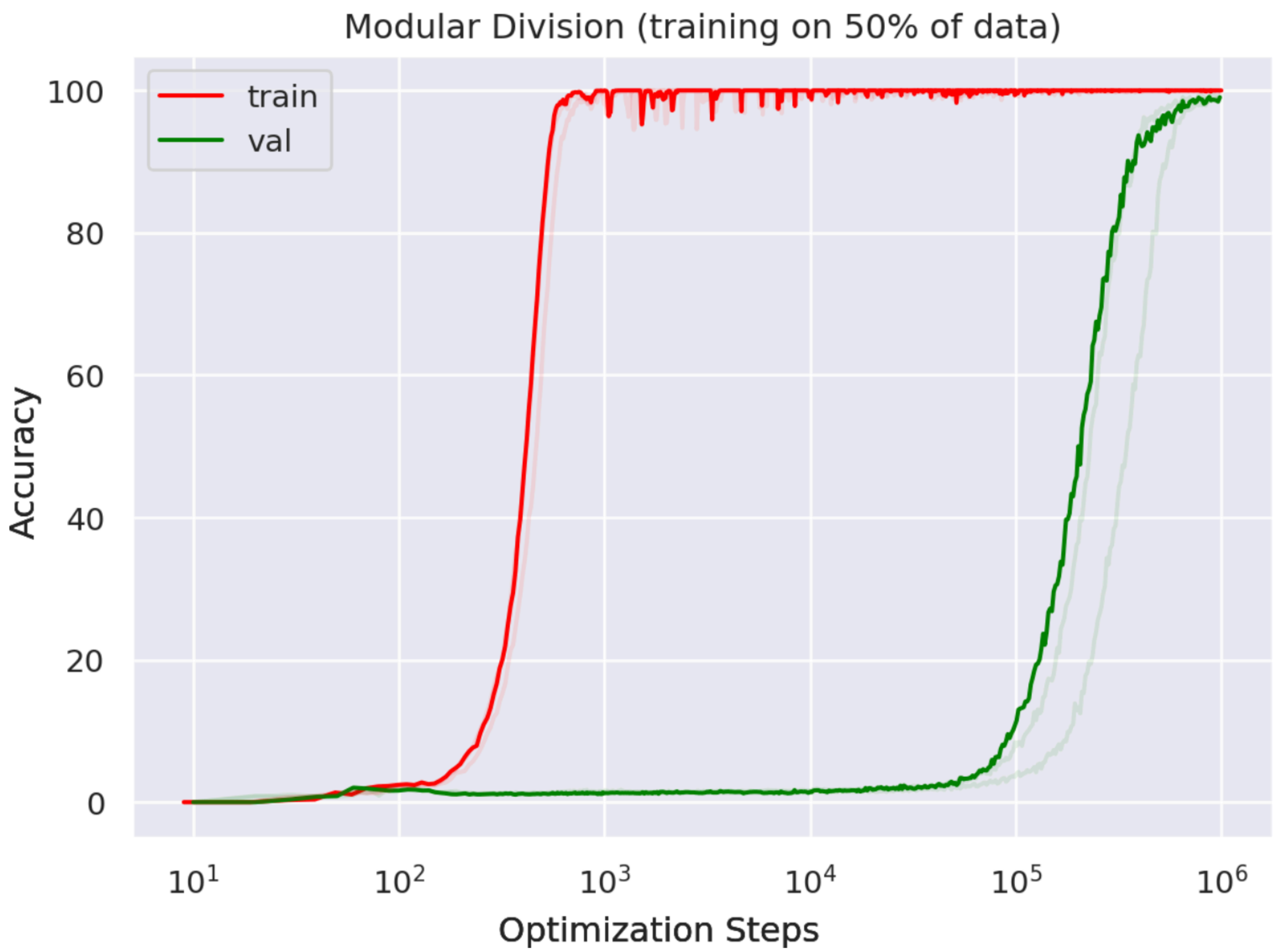
Neural Collapse and Generalization

?

Grokking

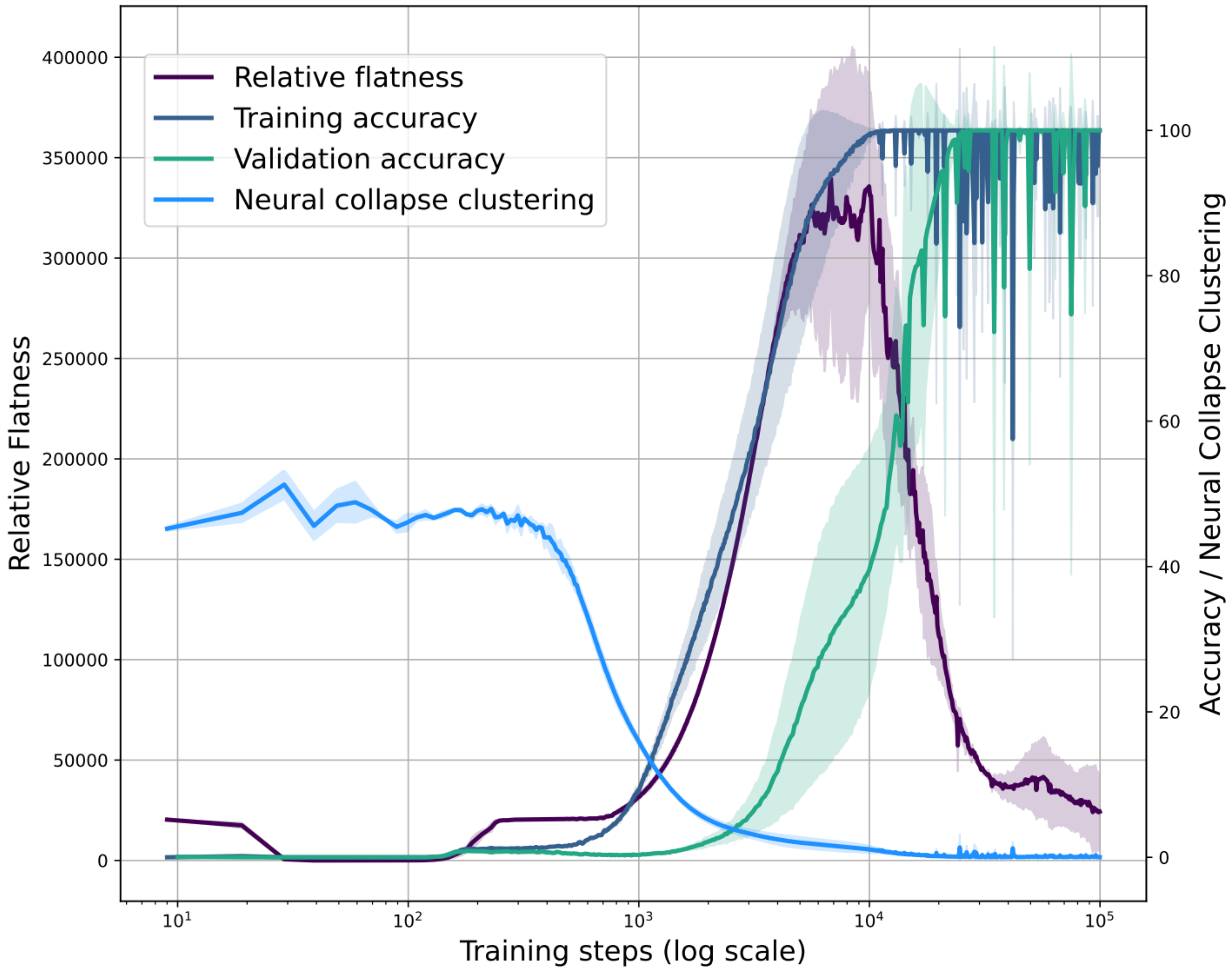
Grokking

Algorithmic modular task: <x> <op> <y> = <x ○ y>, e.g., (40 + 8) mod 97 = 48



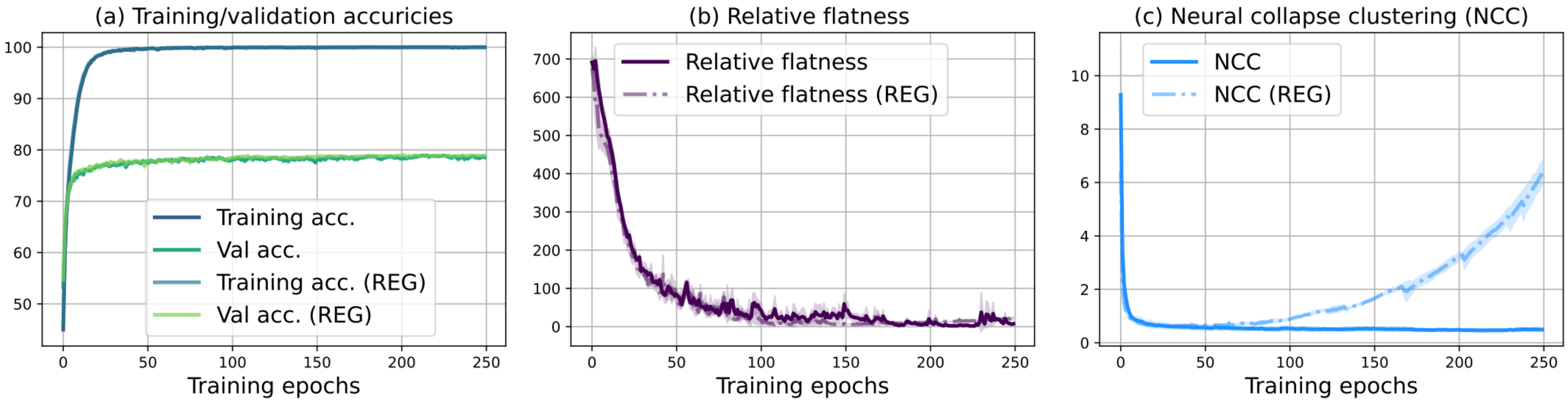
Grokking

Algorithmic m



Causal Interventions

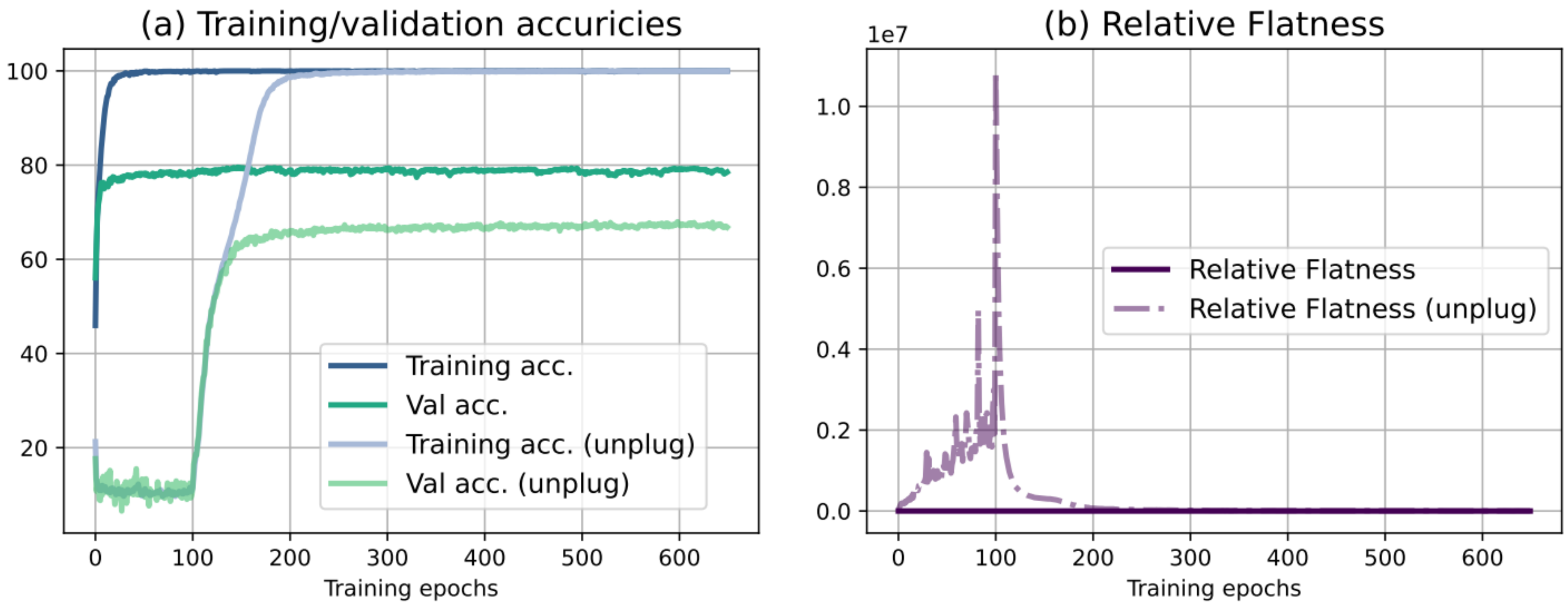
Neural Collapse



$$\mathcal{L}_{\text{NC\_REG}} = \mathcal{L}_{\text{CE}} - \lambda \cdot \text{NCC}$$

$$\text{NCC} := \sum_{c \neq c'} \frac{V_c + V_{c'}}{2\|\mu_c - \mu_{c'}\|^2}$$

Relative Flatness



$$\mathcal{L}_{\text{NC\_RF}} = \mathcal{L}_{\text{CE}} - \lambda * \kappa^{\phi}_{Tr}(\mathbf{w})$$

$$\kappa^{\phi}_{\text{Tr}}(\mathbf{w}) := \sum_{s,s'=1}^{d} \langle \mathbf{w}_s, \mathbf{w}_{s'} \rangle \cdot \text{Tr}(H_{s,s'}(\mathbf{w}, \phi(S)))$$

Conclude:

Relative flatness, rather than Neural Collapse, is necessary to Generalization.

Relative flatness + LLMs

1. LLM intervention:

    - Relative Flatness-Regularized Training, Placement, Diagnostic tool...

2. Cross lingual Generalization

    -A, B and C?

3. Relative flatness + alignment

    -D, E and F?

LAMARR
INSTITUTE FOR MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Thanks and QA?

Partner institutions:

technische universität dortmund · Fraunhofer IAIS · UNIVERSITÄT BONN · Fraunhofer IML

Institutionally funded by:

Bundesministerium für Forschung, Technologie und Raumfahrt · Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen