Generalization in Deep Learning

Study of the paper: Kenji Kawaguchi , Leslie Pack Kaelbling and Yoshua Bengio, Generalization in Deep Learning, arXiv:1710.05468v9

Maria Oprea
Group meeting, 21 Nov 2024

Overview

- Introduction/ the goals of machine learning
- Classical approaches on generalization
- Overparametrization paradox
- Background
- New approach on generalization
- Theoretical results & proof
- Validation

Goal in Machine Learning

Find a model $f : X \to Y$ that fits the data $S = \{(x_i, y_i\}_{i=1}^{m}$ and generalizes well

Expressive:

true model ∈ available models

f performs well on $x_i \notin \pi_x(S)$

Trainable:

can reach optimal in finite time

Terminology

Assume:

- $(x, y) \sim \mathbb{P}_{xy}$, $L : Y \times Y \to [0,\infty)$ loss , $S_m = \{(x_i, y_i\}_{i=1}^{m}$ with $(x_i, y_i) \sim \mathbb{P}_{xy}$

Define:

- Expected risk: $R[f] = \mathbb{E}_{\mathbb{P}_{xy}}[L(f(X), Y)]$
- Empirical risk: $R_S[f] = \dfrac{1}{m}\sum_{i=1}^{m} L(f(x_i), y_i)$
- Generalization gap (S, f) = $R[f] - R_S[f]$

Goal:

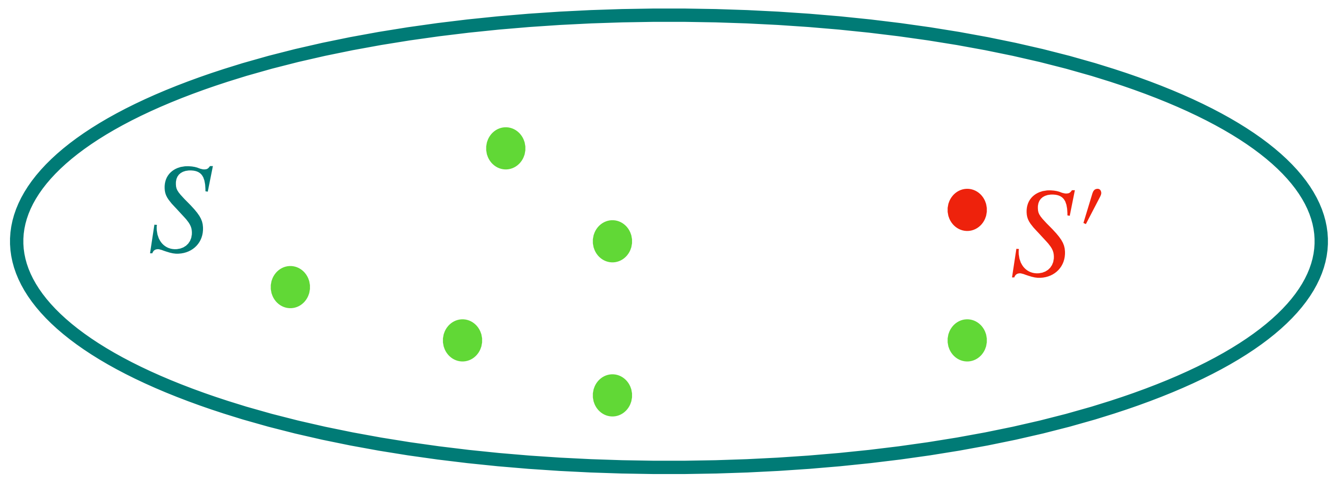
- Find $f_{A(S)} = argmin_{f \in \mathcal{F}} R[f]$, but instead $f_{A(S)} = argmin_{f \in \mathcal{F}} R_S[f]$

Classical approaches to generalization

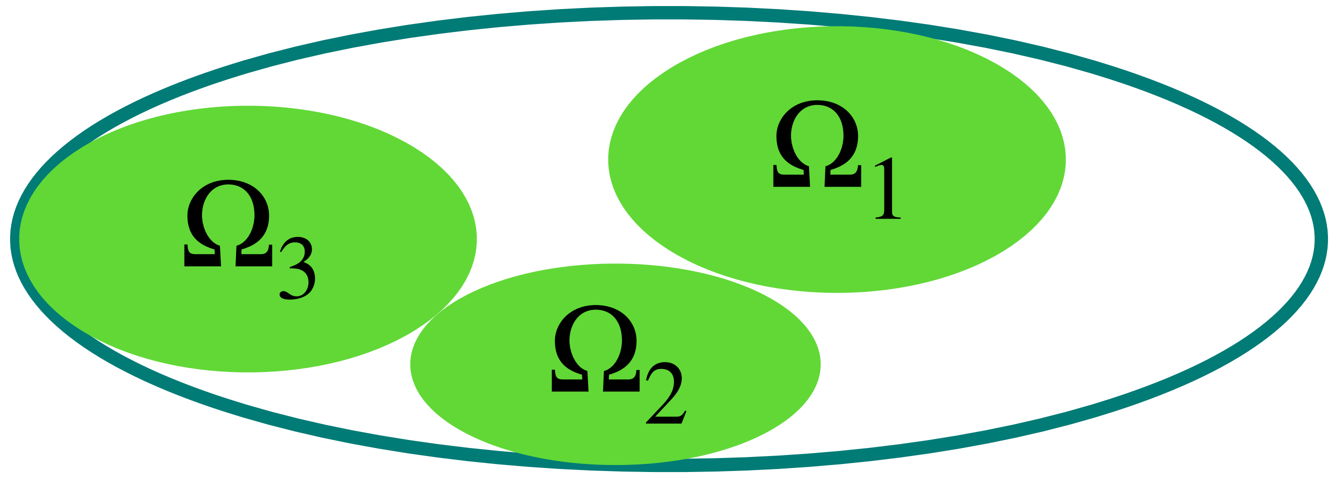
- Hypothesis class complexity → gives guarantees for worst case scenario

$$\sup_{f \in \mathcal{F}} R[f] - R_S[f]$$

- Stability of algorithm A to dataset S → $\Delta S \implies \Delta f_{A(S)}$



- Robustness of A for all possible S → how much $f_{A(S)}$ vary in the input space



Apparent paradox

„Deep neural networks easily fit random labels”[1]

- Same hypothesis class can achieve small errors on true data and fit random data

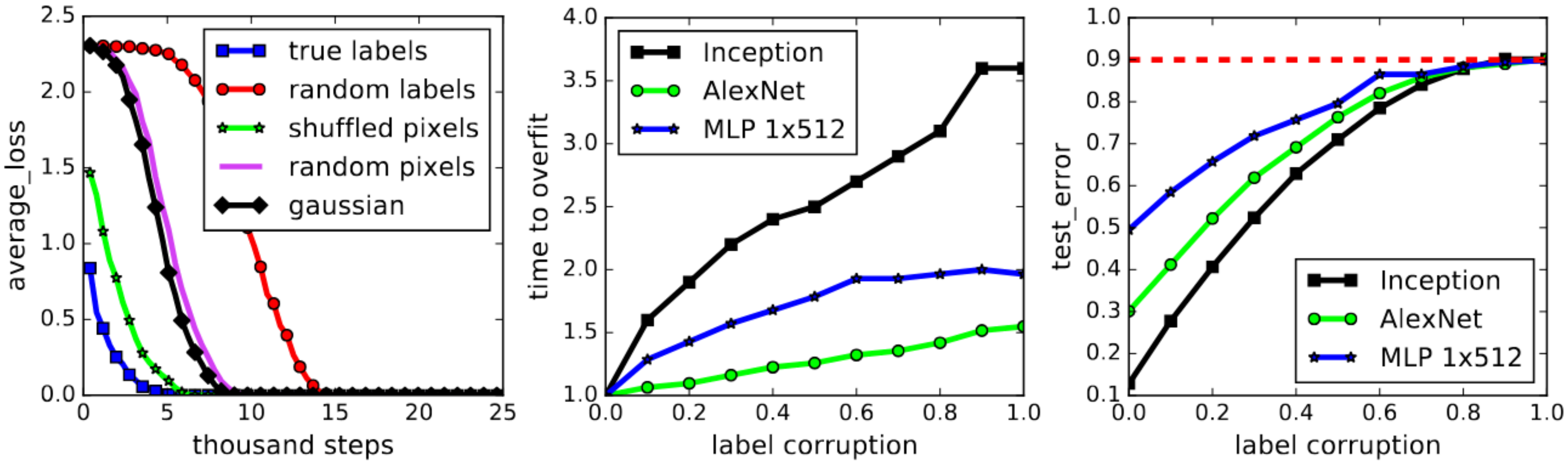
model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78

- Regularization techniques don’t matter very much.

- Although random labels = propriety of data

[1]Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, ICLM 2017

Apparent paradox



(a) learning curves (b) convergence slowdown (c) generalization error growth

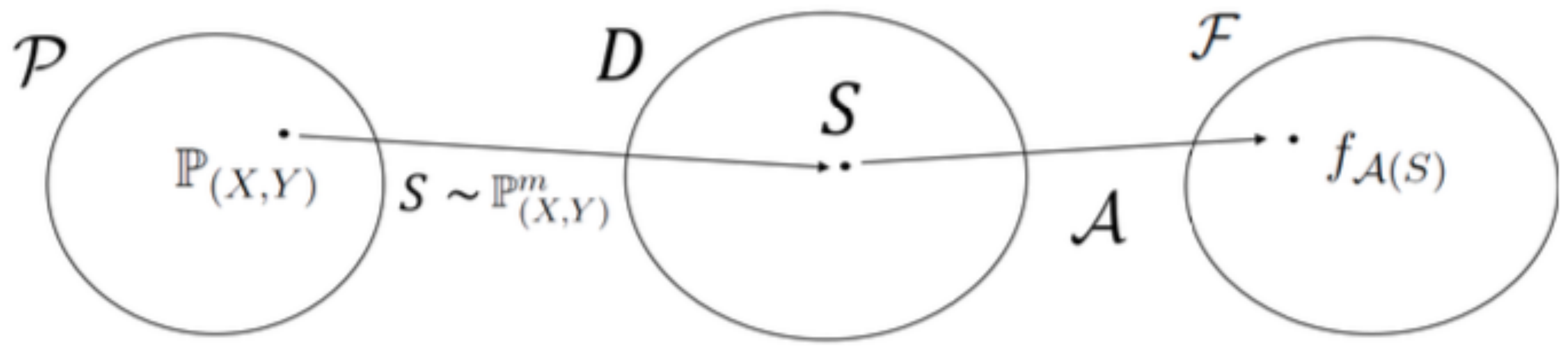
- Random labels $\Longrightarrow$ nothing to learn $\Longrightarrow$ slow convergence

[1]Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, ICLM 2017

Different approach to generalization

<u>Open Problem</u>: Characterize the expected risk $R[f]$ with a sufficiently deep hypothesis space $\mathscr{F}$ producing theoretical insights and distinguishing between the cases of „natural" problem instances $(\mathbb{P}_{xy}, S)$ and „artificial" instances $(\mathbb{P}'_{xy}, S')$

- Problem instance: $(\mathbb{P}_{xy}, S)$ fixed!



Comparison to statistical learning

- In statistical learning

$$p \implies q$$

If hypothesis space complexity is small

The the loss function is bounded on a partition of $X \times Y$

$\implies$

$\cancel{\impliedby}$

The generalization gap is bounded

- Statements about $(\mathscr{P}(X \times Y), \mathscr{D})$
- Example: the no free lunch theorem $\exists \mathbb{P}_{xy}^{bad}$

Theoretical results

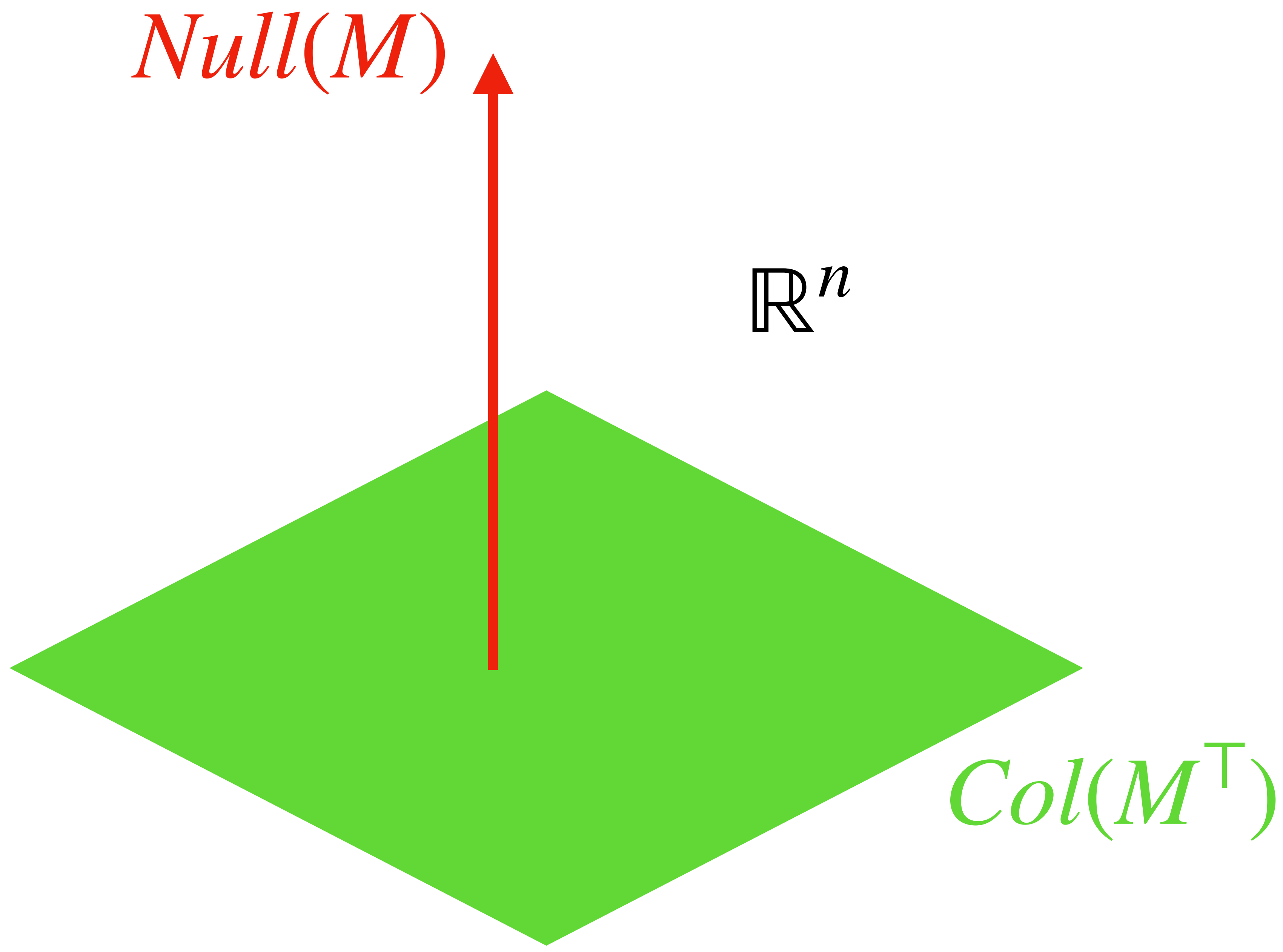
- For fixed problem instance $\mathbb{P}_{xy}, S$
- Intuitively: the hypothesis space of overparametrized linear models can learn any data and reduce train and test errors to 0 even when parameters are arbitrarily far from the ground truth.

Theoretical results

<u>Theorem</u>: Training prediction $\hat{Y}(w) = \Phi w$ and test prediction $\hat{Y}_{test}(w) = \Phi_{test} w$. Let $M^\top = [\Phi^\top, \Phi_{test}^\top]$ with $\Phi \in \mathbb{R}^{m \times n}$, $\Phi_{test} \in \mathbb{R}^{m_{test} \times n}$, $w \in \mathbb{R}^{n \times d_y}$. If $rank(\Phi) = m, rank(M) < n$ and $m < n$ then

1. For any $Y \in \mathbb{R}^{m \times d_y}$ $\exists w'$ such that $\hat{Y}(w') = Y$
2. If there is a ground truth w^* with $Y = \Phi w^*,\ Y_{test} = \Phi_{test} w^*$ then $\forall \epsilon, \delta\ \exists w$ such that
   A. $\hat{Y}(w) = Y + \epsilon A$ with a matrix A such that $\|A\|_F \leq 1$.
   B. $\hat{Y}_{test}(w) = Y_{test} + \delta B$ with a matrix B such that $\|B\|_F \leq 1$.
   C. $\|w\|_F \geq \delta$ and $\|w - w^*\|_F \geq \delta$

Proof of the theorem



Validation

- Setting: After training $\rightarrow$ candidate models in $\mathcal{F}$
- Goal: Given validation set $S_{m_{val}}$ find $f \in \mathcal{F}$
- Example: $\mathcal{F}$ = all models that achieve a 99.5% accuracy after each epoch.
- Intuition for theorem: small validation error $\implies$ good hypothesis independent of capacity

<u>Theorem</u>: Let $\kappa_{f,i} = R[f] - L(f(x_i), y_i)$. Suppose $\mathbb{E}[\kappa_{f,i}] \leq \gamma^2$ and $|\kappa_{f,i}| < C$ a.s.
Then for all δ, with probability at least $1 - \delta$:

$$R[f] - R_{S_{val}}[f] \leq \frac{2C \log \frac{|\mathcal{F}|}{\delta}}{3m_{val}} + \sqrt{\frac{2\gamma^2 \log \frac{|\mathcal{F}|}{\delta}}{m_{val}}}$$

Validation

Proposition: Let $\kappa_{f,i} = R[f] - L(f(x_i), y_i)$. Suppose $\mathbb{E}[\kappa_{f,i}] \leq \gamma^2$ and $|\kappa_{f,i}| < C$ a.s. Then for all δ, with probability at least $1 - \delta$:

$$R[f] - R_{S_{val}}[f] \leq \frac{2C \log \frac{|\mathcal{F}|}{\delta}}{3m_{val}} + \sqrt{\frac{2\gamma^2 \log \frac{|\mathcal{F}|}{\delta}}{m_{val}}}$$

- Reasonable bound: $m_{val} = 10^3, |\mathcal{F}| = 10^9, C = \gamma = 1, \delta = 0.1 \implies$

$$\mathbb{P}[|R[f] - R_{S_{val}}[f]| < 6.95\%] > 0.9$$

- Difference to classical setting: $\mathcal{F}$ does not depend on S_{val} (only on training data)

Proof of the proposition