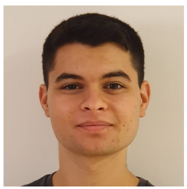
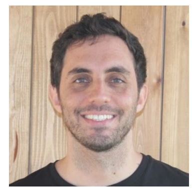
How Uniform Random Weights Induce Non-uniform Bias: Typical Interpolating Neural Networks Generalize with Narrow Teachers


Buzaglo*


Harel*


Nacson*

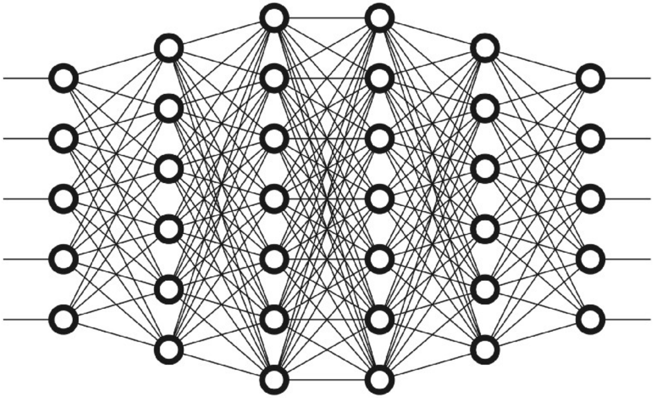

Brutzkus


Srebro


Soudry

Funded by ERC

Question: Why Neural Networks Generalize?



Understanding Generalization in Deep Learning

- SGD 'Implicit Bias' toward generalizing solutions:
  - Gunasekar et al. 2017; Soudry et al. 2018; Arora et al. 2019; Lyu and Li 2020; Chizat and Bach 2020; Vardi 2023

Understanding Generalization in Deep Learning

- SGD 'Implicit Bias' toward generalizing solutions:
  - Gunasekar et al. 2017; Soudry et al. 2018; Arora et al. 2019; Lyu and Li 2020; Chizat and Bach 2020; Vardi 2023
- Randomly sampled interpolating NNs empirically generalize:
  - Valle-Perez et al., 2019; Mingard et al., 2021; Chiang et al., 2023

Understanding Generalization in Deep Learning

- SGD 'Implicit Bias' toward generalizing solutions:
  - Gunasekar et al. 2017; Soudry et al. 2018; Arora et al. 2019; Lyu and Li 2020; Chizat and Bach 2020; Vardi 2023
- Randomly sampled interpolating NNs empirically generalize:
  - Valle-Perez et al., 2019; Mingard et al., 2021; Chiang et al., 2023

Take Home Message

Assumptions

- There exists an underlying **narrow teacher** network
- The weights of all networks are **quantized**

Take Home Message

Assumptions

- There exists an underlying **narrow teacher** network
- The weights of all networks are **quantized**

Our results

We prove that a typical interpolating NN generalizes with

$$\#\text{samples} \approx O\left(\#\,\text{teacher params} + \#\,\text{student neurons}\right)$$

Take Home Message

Assumptions

- There exists an underlying **narrow teacher** network
- The weights of all networks are **quantized**

Our results

We prove that a typical interpolating NN generalizes with

$$\#\text{samples} \approx O\left(\#\,\text{teacher params} + \#\,\text{student neurons}\right)$$

- Usually, $\#\,\text{student neurons} \ll \#\,\text{student params}$

Take Home Message

Assumptions

- There exists an underlying **narrow teacher** network
- The weights of all networks are **quantized**

Our results

We prove that a typical interpolating NN generalizes with

$$\#\text{samples} \approx O\left(\#\,\text{teacher params} + \#\,\text{student neurons}\right)$$

- Usually, $\#\,\text{student neurons} \ll \#\,\text{student params}$
- Results for any depth and activation function, including CNN.

Take Home Message

Assumptions

- There exists an underlying **narrow teacher** network
- The weights of all networks are **quantized**

Our results

We prove that a typical interpolating NN generalizes with

$$\#\text{samples} \approx O\left(\#\,\text{teacher params} + \#\,\text{student neurons}\right)$$

- Usually, $\#\,\text{student neurons} \ll \#\,\text{student params}$
- Results for any depth and activation function, including CNN.
- Relax quantization assumption, for special case (2-layer, LeakyReLU).

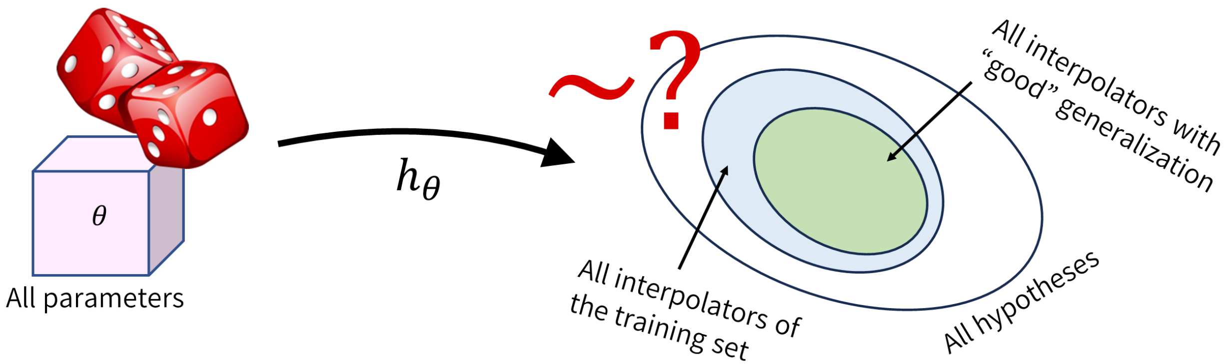
Posterior Sampling

- Prior over functions

$$\mathcal{P}(h) = \mathbb{P}_{\boldsymbol{\theta}}(h_{\boldsymbol{\theta}} = h)$$

- Posterior given training set

$$\mathcal{P}_S(h) = \mathcal{P}(h \mid \mathcal{L}_S(h) = 0)$$



Posterior Sampling Generalizes with Fixed $\tilde{p}$

Teacher Equivalence

The probability to sample a teacher-equivalent model is

$$\tilde{p} \triangleq \mathbb{P}_{h \sim \mathcal{P}}\left(h \equiv h^\star\right) .$$

Lemma (Generalization of Abstract Posterior Sampling, Informal)

$$\mathbb{P}_{S \sim \mathcal{D}^N, h \sim \mathcal{P}_S}\left(\mathcal{L}_\mathcal{D}(h) < \epsilon\right) \geq 1 - \delta \,,$$

if

$$N \geq \frac{-\log\left(\tilde{p}\right) + 3\log\left(\frac{2}{\delta}\right)}{\epsilon} \leftarrow \text{sample complexity}$$

Key Assumptions

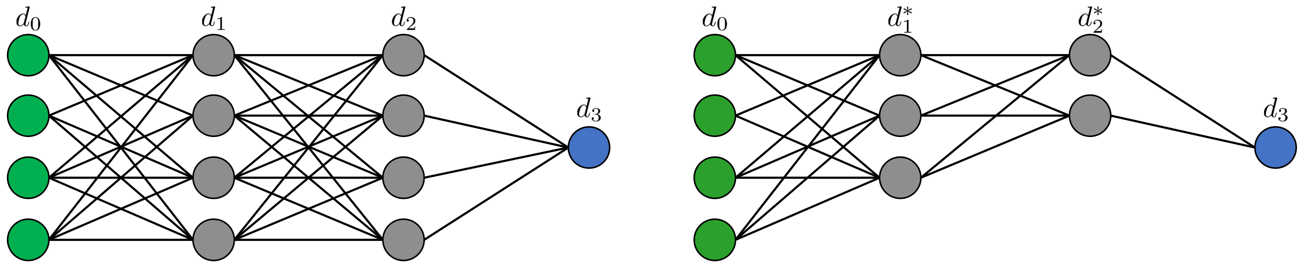
Quantization

We consider Q-**quantized** networks where each of the parameters is chosen from a fixed set $\mathcal{Q} \subset \mathbb{R}$ such that $0 \in \mathcal{Q}$ and $|\mathcal{Q}| \leq Q$.

- e.g., Numbers representable as $\log_2 Q$-bit floats.

Narrowness

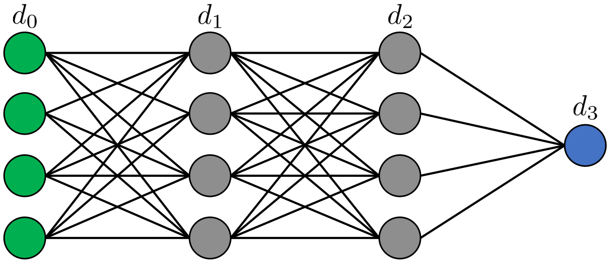
We consider a teacher $h^\star = h_{\boldsymbol{\theta}^\star}$ which is a Q-quantized network of some depth L and small widths $D^\star = (d_1^\star, \ldots, d_L^\star)$, and a wider student of the same depth L but widths $D \gg D^\star$.

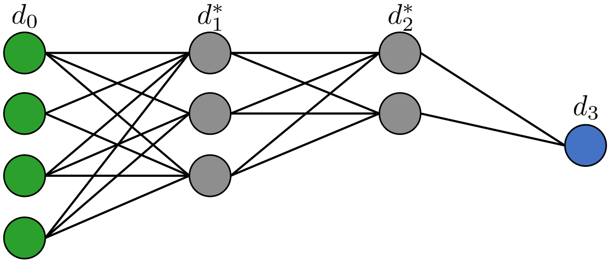


Key Assumptions

Narrowness

We consider a teacher $h^\star = h_{\theta^\star}$ which is a Q-quantized network of some depth L and small widths $D^\star = (d_1^\star, \ldots, d_L^\star)$, and a wider student of the same depth L but widths $D \gg D^\star$.





Uniform Prior

We consider a **uniform prior over Q-quantized parameterizations**.

Main Result

Theorem (Effective Sample Complexities, Informal)

For any depth and activation function we have that:

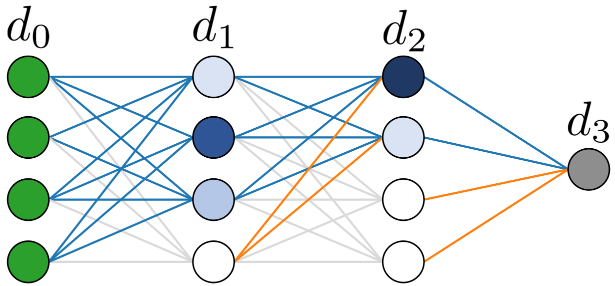
- *For Vanilla Fully Connected Networks:*

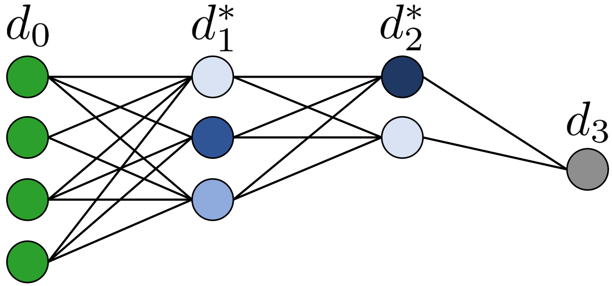
$$\text{sample complexity} = O\left(\log Q \cdot \sum_{l=1}^{L} (d_l^\star d_{l-1} + d_l^\star)\right).$$

- *With "batch-normalization-like scaling":*

$$\text{sample complexity} = O\left(\log Q \cdot \sum_{l=1}^{L} (d_l^\star d_{l-1}^\star + 2d_l)\right).$$

Proof Idea

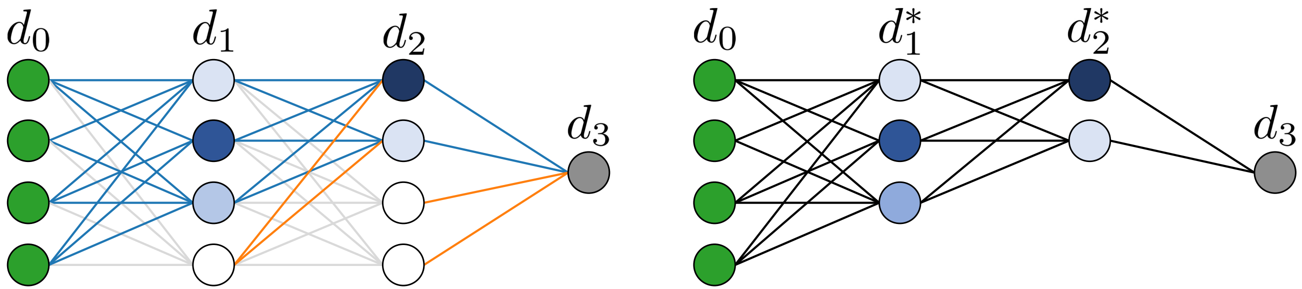




Vanilla Fully Connected Networks

How can a sampled network (left) replicate the teacher (right)?

Proof Idea

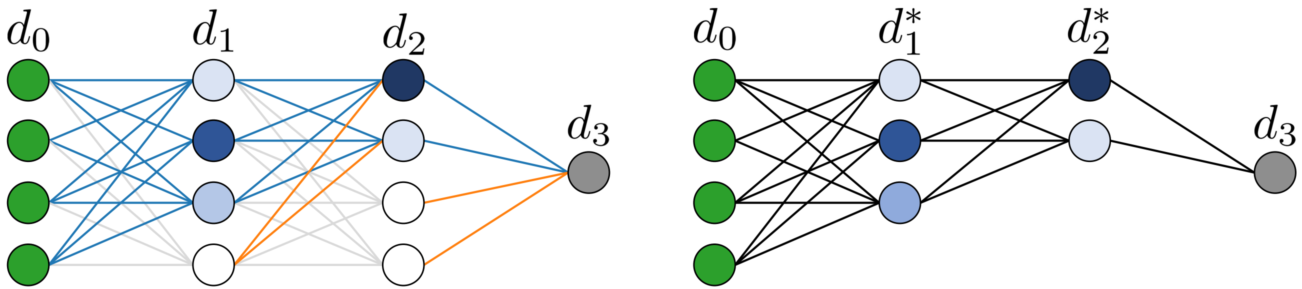


Vanilla Fully Connected Networks

How can a sampled network (left) replicate the teacher (right)?

- Having a sub-network identical to the teacher

Proof Idea

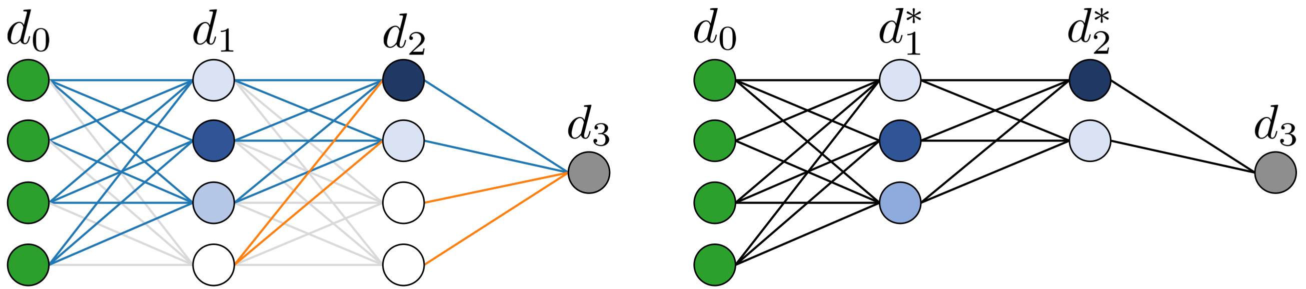


Vanilla Fully Connected Networks

How can a sampled network (left) replicate the teacher (right)?

- Having a sub-network identical to the teacher
- Then zero-out outgoing weights of redundant neurons

Proof Idea



Vanilla Fully Connected Networks

How can a sampled network (left) replicate the teacher (right)?

- Having a sub-network identical to the teacher
- Then zero-out outgoing weights of redundant neurons
- Then the weights entering can be **arbitrary**

Summary

We showed

Posterior Sampling generalizes, assuming underlying narrow teacher.

Summary

We showed

Posterior Sampling generalizes, assuming underlying narrow teacher.

In the paper

- Analogous results for CNN
- Removing Quantization Assumption (2-Layer)
- Beyond interpolators

Summary

We showed

Posterior Sampling generalizes, assuming underlying narrow teacher.

In the paper

- Analogous results for CNN
- Removing Quantization Assumption (2-Layer)
- Beyond interpolators

Future directions

- Random interpolators conditioned on specific implicit bias
- Connections to SGD