Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models

Mor Shpigel Nacson[1] Suriya Gunasekar[2] Jason D. Lee[3]

Nathan Srebro[2] Daniel Soudry[1]

[1] Technion, Israel

[2] TTI Chicago, USA

[3] USC Los Angeles, USA

ICML, 2019

Motivation

- Deep neural networks have multiple global minima.
- Each minimum has different generalization properties.

Motivation

- Deep neural networks have multiple global minima.
- Each minimum has different generalization properties.
- Empirically, training deep neural networks we get <u>specific solutions</u> that generalize well.

Motivation

- Deep neural networks have multiple global minima.
- Each minimum has different generalization properties.
- Empirically, training deep neural networks we get specific solutions that generalize well.

Main Goal

We would like to understand the minima selection process in training deep neural networks.

Setting

- Empirical loss:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} e^{-f_n(\boldsymbol{\theta})}$$

$f_n(\boldsymbol{\theta})$ - the prediction function, N - number of samples.

Setting

- Empirical loss:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} e^{-f_n(\boldsymbol{\theta})}$$

  $f_n(\boldsymbol{\theta})$ - the prediction function, N - number of samples.

- We examine overparameterized realizable problems i.e., where it is possible to perfectly classify the training data.

Setting

- Empirical loss:

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} e^{-f_n(\boldsymbol{\theta})}$$

  $f_n(\boldsymbol{\theta})$ - the prediction function, N - number of samples.

- We examine overparameterized realizable problems i.e., where it is possible to perfectly classify the training data.
- The <u>inductive bias</u> introduced in our learning process affects which specific global minimizer is chosen.

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 \qquad (1)$$

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 \qquad (1)$$

– Empirically, using small, and even vanishing λ can improve generalization.

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 \qquad (1)$$

- Empirically, using small, and even vanishing λ can improve generalization.
- What happens at the limit of the regularization path, when $\lambda \to 0$?

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 \qquad (1)$$

- Empirically, using small, and even vanishing λ can improve generalization.
- What happens at the limit of the regularization path, when $\lambda \to 0$?

2) **Constrained path:**

$$\Theta_c(\rho) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \text{ s.t. } \left\|\boldsymbol{\theta}\right\|_2 \leq \rho$$

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 \qquad (1)$$

- Empirically, using small, and even vanishing λ can improve generalization.
- What happens at the limit of the regularization path, when $\lambda \to 0$?

2) **Constrained path:**

$$\Theta_c(\rho) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \text{ s.t. } \|\boldsymbol{\theta}\|_2 \leq \rho$$

- Previously related to problem (1).

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2^2 \qquad (1)$$

- Empirically, using small, and even vanishing λ can improve generalization.
- What happens at the limit of the regularization path, when $\lambda \to 0$?

2) **Constrained path:**

$$\Theta_c(\rho) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \text{ s.t. } \|\boldsymbol{\theta}\|_2 \leq \rho$$

- Previously related to problem (1).
- What happens at the limit of the constrained path, when $\rho \to \infty$?

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 \qquad (1)$$

– Empirically, using small, and even vanishing λ can improve generalization.

– What happens at the limit of the regularization path, when $\lambda \rightarrow 0$?

2) **Constrained path:**

$$\Theta_c(\rho) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \text{ s.t. } \left\|\boldsymbol{\theta}\right\|_2 \leq \rho$$

– Previously related to problem (1).

– What happens at the limit of the constrained path, when $\rho \rightarrow \infty$?

3) **Optimization path:**

$$\bar{\boldsymbol{\theta}}(t) = \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|}, \quad \Delta\boldsymbol{\theta}(t) = -\eta \nabla \mathcal{L}\left(\boldsymbol{\theta}\left(t\right)\right)$$

Inductive Bias Sources

1) **Regularization path:**

$$\Theta_r(\lambda) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \left\|\boldsymbol{\theta}\right\|_2^2 \qquad (1)$$

- Empirically, using small, and even vanishing λ can improve generalization.
- What happens at the limit of the regularization path, when $\lambda \to 0$?

2) **Constrained path:**

$$\Theta_c(\rho) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \text{ s.t. } \left\|\boldsymbol{\theta}\right\|_2 \leq \rho$$

- Previously related to problem (1).
- What happens at the limit of the constrained path, when $\rho \to \infty$?

3) **Optimization path:**

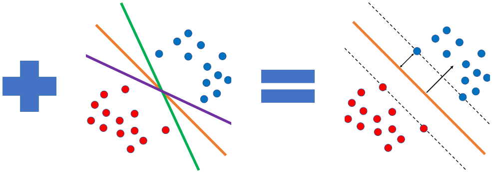
$$\bar{\boldsymbol{\theta}}(t) = \frac{\boldsymbol{\theta}(t)}{\|\boldsymbol{\theta}(t)\|}, \quad \Delta\boldsymbol{\theta}(t) = -\eta\nabla\mathcal{L}\left(\boldsymbol{\theta}\left(t\right)\right)$$

- What happens at the limit of the optimization path, when $t \to \infty$?

Previous Results

- For linear prediction functions:
  - Optimization path $\Rightarrow$ Max-Margin solution.

$$\Delta\boldsymbol{\theta}(t) = -\eta\nabla\mathcal{L}\left(\boldsymbol{\theta}\right)$$



Soudry et al. (2018), Gunasekar et al. (2018), Rosset et al. (2004), Wei et al. (2018).

Previous Results

- For linear prediction functions:
  - Optimization path $\Rightarrow$ Max-Margin solution.
  - Regularization and Constrained paths $\Rightarrow$ Max-Margin solution.

Soudry et al. (2018), Gunasekar et al. (2018), Rosset et al. (2004), Wei et al. (2018).

Previous Results

- For linear prediction functions:
  - Optimization path $\Rightarrow$ Max-Margin solution.
  - Regularization and Constrained paths $\Rightarrow$ Max-Margin solution.
- For homogeneous prediction functions, e.g., ReLU networks:
  - Regularization path $\Rightarrow$ Max-Margin solution.

Soudry et al. (2018), Gunasekar et al. (2018), Rosset et al. (2004), Wei et al. (2018).

We study how infinitesimal regularization or gradient descent optimization lead to margin maximizing solutions in both homogeneous and non-homogeneous models.

Main Contributions - Non-Homogeneous Models

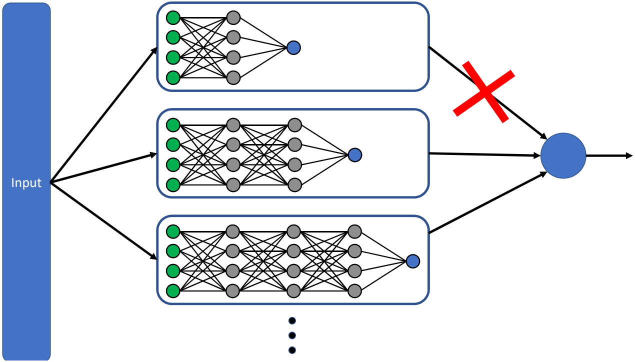
- For $f_n(\boldsymbol{\theta}) =$ sum of homogeneous functions of different orders: we characterized the constrained path asymptotic solution.

Main Contributions - Non-Homogeneous Models

- For $f_n(\boldsymbol{\theta}) =$ sum of homogeneous functions of different orders: we characterized the constrained path asymptotic solution.

Implication

In an ensemble of homogeneous neural networks, e.g., feedforward ReLU networks, the ensemble will aim to discard the most shallow network.



Main Contributions - Homogeneous Models

Q: In non-linear homogeneous models:

1) Are optimization and constrained paths still equivalent?
2) Does the optimization path still leads to max-margin solutions?

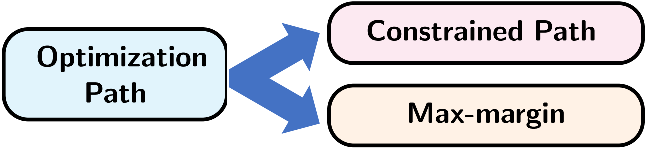
Main Contributions - Homogeneous Models

Q: In non-linear homogeneous models:

1) Are optimization and constrained paths still equivalent?
2) Does the optimization path still leads to max-margin solutions?

A: Yes, we find general conditions under which the optimization path converges to:

1) stationary points of the constrained path.
2) max-margin solutions.



Main Contributions - Homogeneous Models

- Refined characterization:
  - For non-convex prediction functions the max-margin solution is not necessarily unique.
  - We show that the constrained path converges to a specific type of max-margin solution.

Main Contributions - Homogeneous Models

- Refined characterization:
  - For non-convex prediction functions the max-margin solution is not necessarily unique.
  - We show that the constrained path converges to a specific type of max-margin solution.

Q: Is margin maximization all that we do?

Main Contributions - Homogeneous Models

- Refined characterization:
  - For non-convex prediction functions the max-margin solution is not necessarily unique.
  - We show that the constrained path converges to a specific type of max-margin solution.

Q: Is margin maximization all that we do?

A: No. After maximizing the distance to the closest data point (max-margin), we also maximize the distance to the second closest data point, and so on.

Thank You!

Poster –
Pacific Ballroom #72