The inductive bias of ReLU networks on orthogonally separable data

Mary Phuong, Christoph H. Lampert

mary-phuong.github.io pub.ist.ac.at/~chl

ICLR 2021

IST AUSTRIA

Institute of Science and Technology

Inductive bias

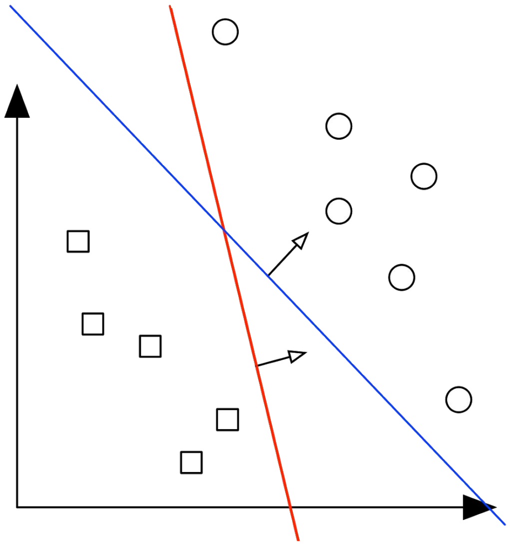
- Many solutions with zero training error, but different generalisation
- Which solution does the algo pick?

Inductive bias

- Many solutions with zero training error, but different generalisation
- Which solution does the algo pick?

Setting

- Binary classification $\{(\mathbf{x}_i, y_i)\} \subset \mathbb{R}^d \times \{\pm 1\}$
- Linearly separable data
- Classifier $\operatorname{sign} f_{\boldsymbol{\theta}}(\mathbf{x})$
- Train by minimising the cross-ent loss by gradient flow

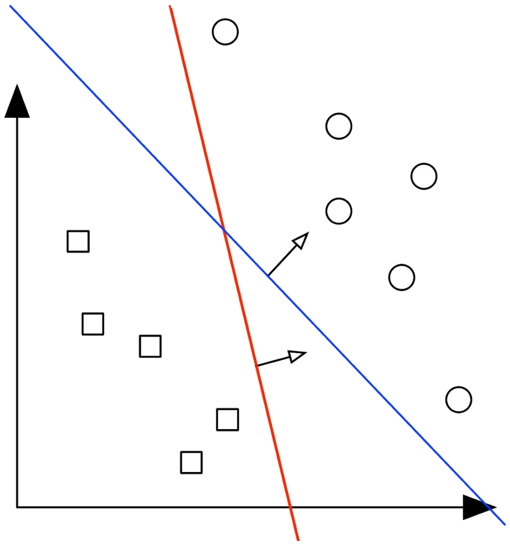


Previous work -- linear models

- Logistic regression [Soudry etal 2017]

$$f(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x}$$

$$\mathbf{w}/\|\mathbf{w}\| \to \text{max-margin direction } \mathbf{w}_{\text{max}}$$



Previous work -- linear models

- Logistic regression [Soudry etal 2017]

$$f(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x}$$

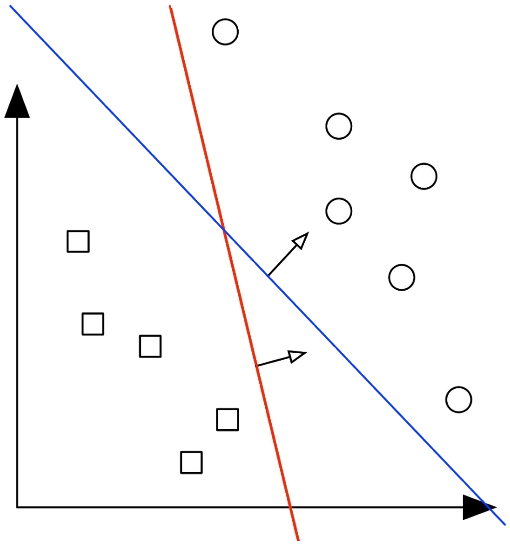
$$\mathbf{w}/\|\mathbf{w}\| \to \text{max-margin direction } \mathbf{w}_{\max}$$

- Deep linear nets [Ji & Telgarsky 2019]

$$f(\mathbf{x}) = \underbrace{\mathbf{W}_L \cdots \mathbf{W}_2 \mathbf{W}_1}_{\mathbf{w}_{\boldsymbol{\theta}}} \mathbf{x}$$

$$\mathbf{w}_{\boldsymbol{\theta}}/\|\mathbf{w}_{\boldsymbol{\theta}}\| \to \text{max-margin direction } \mathbf{w}_{\max}$$

$$\mathbf{W}_1/\|\mathbf{W}_1\| \to \mathbf{u}\mathbf{w}_{\max}^\intercal$$



This work -- ReLU networks

- Orthogonal separability
  - Stronger version of linear separability
- Two-layer ReLU networks

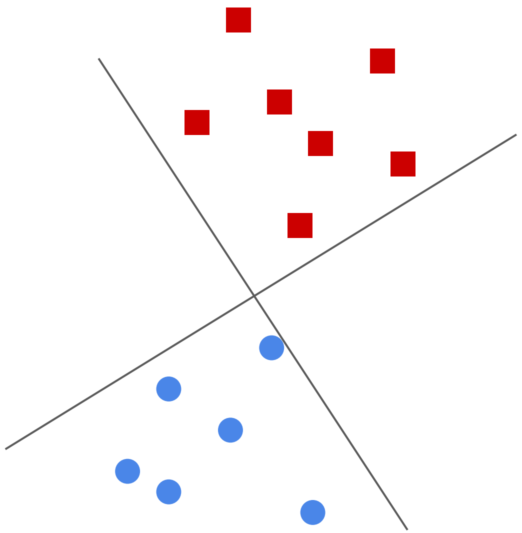
Main result: We characterise what the net converges to.

Assumptions

1. Orthogonal separability

Dataset $\{(\mathbf{x}_i, y_i)\} \subset \mathbb{R}^d \times \{\pm 1\}$ is orthogonally separable, if for all (i,j),

$$\mathbf{x}_i^\intercal \mathbf{x}_j > 0 \text{ if } y_i = y_j$$
$$\mathbf{x}_i^\intercal \mathbf{x}_j \leq 0 \text{ if } y_i \neq y_j$$

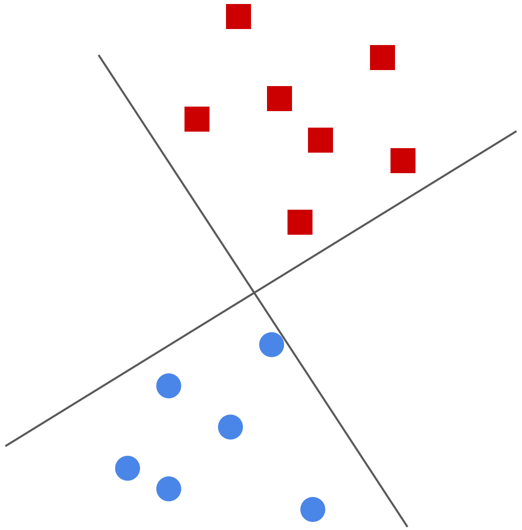


Assumptions

1. Orthogonal separability

Dataset $\{(\mathbf{x}_i, y_i)\} \subset \mathbb{R}^d \times \{\pm 1\}$ is orthogonally separable, if for all (i,j),

$$\mathbf{x}_i^\intercal \mathbf{x}_j > 0 \text{ if } y_i = y_j$$
$$\mathbf{x}_i^\intercal \mathbf{x}_j \leq 0 \text{ if } y_i \neq y_j$$

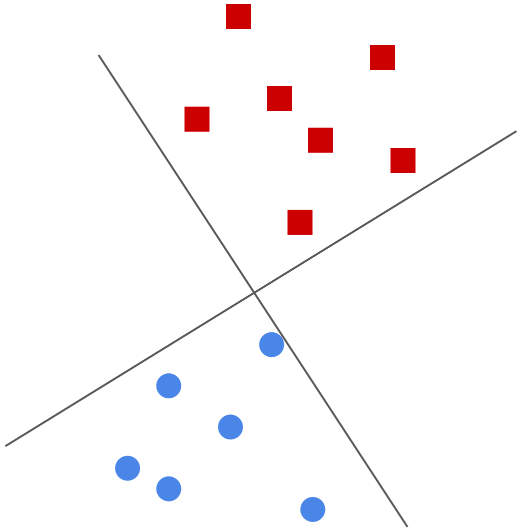


2. Two-layer ReLU networks

$$f_{\boldsymbol{\theta}}(\mathbf{x}) \triangleq \mathbf{a}^\intercal \rho(\mathbf{W}\mathbf{x})$$

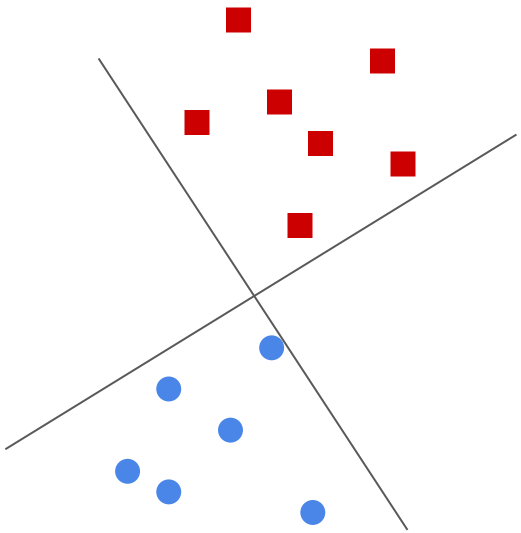
Assumptions

1. Orthogonal separability
2. Two-layer ReLU networks
3. Trained by gradient flow with cross-ent loss



Assumptions

1. Orthogonal separability
2. Two-layer ReLU networks
3. Trained by gradient flow with cross-ent loss
4. <u>Near-zero initialisation</u>



Main result

- Positive / negative max-margin direction:

$$\mathbf{w}_+ = \arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i \geq 1 \text{ for } i : y_i = 1,$$

$$\mathbf{w}_- = \arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{x}_i \geq 1 \text{ for } i : y_i = -1$$

Main result

- Positive / negative max-margin direction:

$$\mathbf{w}_+ = \arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to} \quad \mathbf{w}^\intercal \mathbf{x}_i \geq 1 \ \text{ for } \ i : y_i = 1,$$

$$\mathbf{w}_- = \arg\min_{\mathbf{w}} \|\mathbf{w}\|^2 \quad \text{subject to} \quad \mathbf{w}^\intercal \mathbf{x}_i \geq 1 \ \text{ for } \ i : y_i = -1$$

- <u>Main result</u>

$$\frac{\mathbf{W}(t)}{\|\mathbf{W}(t)\|_F} \to \mathbf{u}\mathbf{w}_+^\intercal + \mathbf{z}\mathbf{w}_-^\intercal \qquad \frac{\mathbf{a}(t)}{\|\mathbf{a}(t)\|} \to \mathbf{u}\|\mathbf{w}_+\| - \mathbf{z}\|\mathbf{w}_-\|$$

$$\mathbf{u}, \mathbf{z} \in \mathbb{R}_+^p \ \textit{such that either } u_i = 0 \textit{ or } z_i = 0$$