On the Effective Number of Linear Regions in Shallow Univariate ReLU Networks: Convergence Guarantees and Implicit Bias

Itay Safran

PRINCETON UNIVERSITY

Joint work with Gal Vardi and Jason D. Lee

NeurIPS

December 2022

In a Nutshell...

We study depth 2 ReLU networks in a univariate binary classification setting, showing

- An end-to-end learning guarantee
- Characterization of the implicit bias of Gradient Flow
- Over-parameterization is necessary for optimization

Assumptions

- Network weights and biases are i.i.d. normal random variables
- There exists a teacher network $\mathcal{N}^*$ of width r which determines the labels of the data

Assumptions

- Network weights and biases are i.i.d. normal random variables
- There exists a teacher network $\mathcal{N}^*$ of width r which determines the labels of the data
- The length of the shortest interval on which $\mathcal{N}^*$ doesn't change signs is $\rho > 0$
- Data distribution $\mathcal{D}$ is compactly supported and has bounded density

Main Result

$n =$ sample size, $k =$ width of student network,
$r =$ width of teacher network, $\mathcal{N}^* =$ teacher network,
$\rho =$ length of shortest interval where $\mathcal{N}^*$ doesn't change signs

Theorem

For appropriate scaling of the initialization, given any $\varepsilon, \delta \in (0,1)$, suppose that

$$n \geq \Omega\left(\frac{r\log(1/\varepsilon) + \log(2/\delta)}{\varepsilon}\right), \qquad k \geq \Omega\left(\frac{\log\left(\frac{r}{\delta}\right)}{\rho}\right)$$

Then with probability at least $1-\delta$, GF converges to zero loss, and converges in direction (suitable defined) to $\boldsymbol{\theta}^*$ such that the network $\mathcal{N}_{\boldsymbol{\theta}^*}$ has at most $32r+67$ linear regions and satisfies

$$\mathbb{P}_{x\sim\mathcal{D}}\left[\mathrm{sign}(\mathcal{N}_{\boldsymbol{\theta}^*}(x)) \neq \mathrm{sign}(\mathcal{N}^*(x))\right] \leq \varepsilon$$

Optimization:

- Identify a direction in weight space which strictly decreases the loss
- Show that the loss satisfies the *PL-condition* in a neighborhood of the initialization

Some Key Proof Ideas

Optimization:

- Identify a direction in weight space which strictly decreases the loss
- Show that the loss satisfies the *PL-condition* in a neighborhood of the initialization

Implicit bias:

- GF must converge to a point satisfying the *KKT conditions*
- Too many linear segments of the student in a single linear segment of the teacher violates KKT

Summary

- An end-to-end convergence guarantee for GF in a univariate, binary classification, teacher-student setting
- Over-parameterization is sufficient for successful optimization (also necessary – student must be wide enough)
- Don't be afraid of overfitting – implicit bias of GF guarantees good generalization

Summary

- An end-to-end convergence guarantee for GF in a univariate, binary classification, teacher-student setting
- Over-parameterization is sufficient for successful optimization (also necessary – student must be wide enough)
- Don't be afraid of overfitting – implicit bias of GF guarantees good generalization

Thank you!