Implicit Bias of Gradient Descent on Linear Convolutional Networks

SeongSik Choi

May, 2021

Seoul National University

Table of Contents

Introduction

Large scale neural networks are highly over-parameterized.

However, specific optimization algorithms take us some special global minima.

Example

Linear regression (under-determined model) - minimum l_2 solution

Linear logistic regression (linearly separable) with gradient descent
- hard margin support vector machine solution

Linear logistic regression (linearly separable) with coordinate descent
- maximum l_1 margin solution

Introduction

Changing to a different parameterization of the same model class changes implicit bias.

(Ex 1. Fully connected, Convolutional)

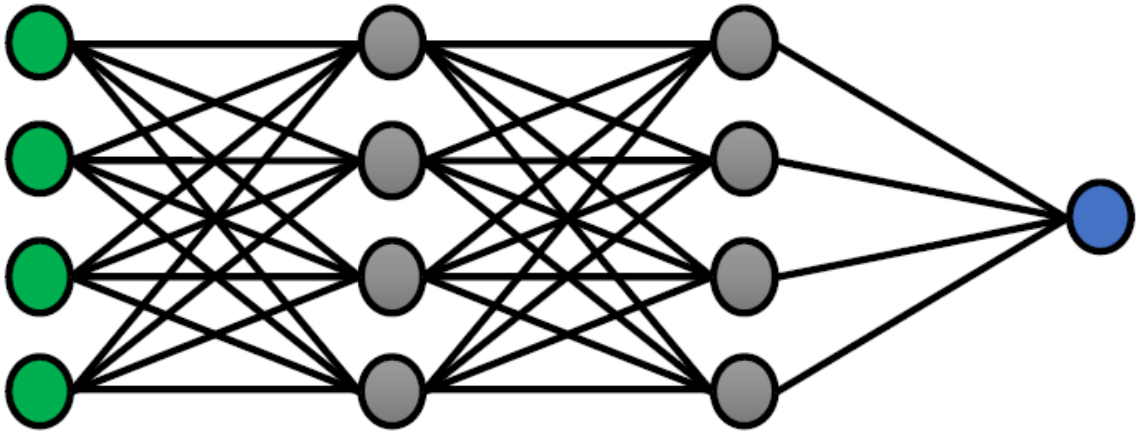
(Ex 2. Optimizing w, β)

Changing the optimization algorithm changes implicit bias.

(Ex. Gradient descent, Coordinate descent)

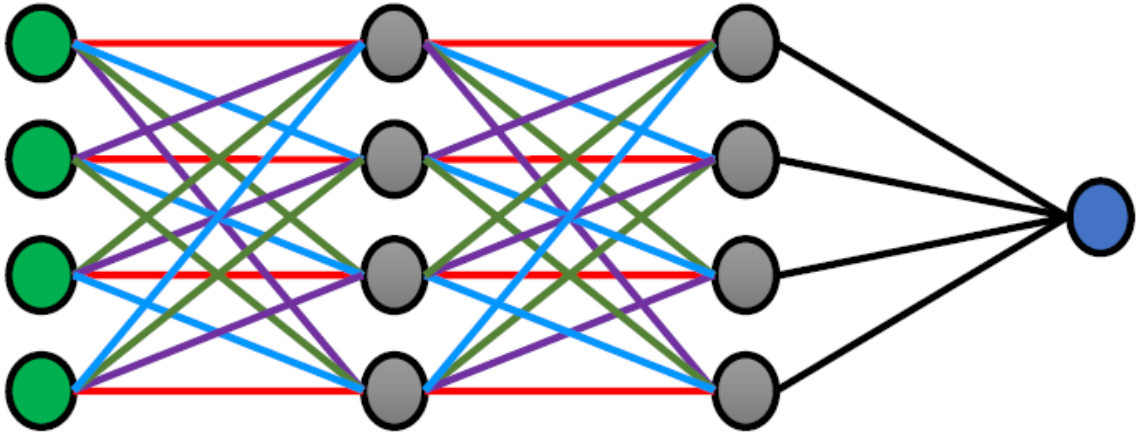
Table of Contents

Multi-layer Linear Networks



(a) Fully connected network of depth L

$\overline{\beta}^{\infty} \propto \underset{\forall n,\, y_n \langle \mathbf{x}_n, \beta \rangle \geq 1}{\operatorname{argmin}} \|\beta\|_2$ (independent of L)



(b) Convolutional network of depth L

$\overline{\beta}^{\infty} \propto$ first order stationary point of $\underset{\forall n,\, y_n \langle \mathbf{x}_n, \beta \rangle \geq 1}{\operatorname{argmin}} \|\widehat{\beta}\|_{2/L}$

Linear convolutional network

We consider 1-dim circular conv network. Each non-output layer has D units (same as the input dimensionality)
Circular convolutional operation parameterized by full width filters with weights $\left[\mathrm{w}_l \in \mathbb{R}^D\right]_{l=1}^{L-1}$.

$$\mathrm{h}_l[d] = \frac{1}{\sqrt{D}} \sum_{k=0}^{D-1} \mathrm{w}_l[k] \mathrm{h}_{l-1}[(d+k) \bmod D] := \left(\mathrm{h}_{l-1} \star \mathrm{w}_l\right)[d]$$

The output layer is fully connected.

Multi-layer Linear Networks

A mapping $\mathcal{P} : \mathcal{W} \to \mathbb{R}^D$ that maps the input parameters $\mathrm{w} \in \mathcal{W}$ to a linear predictor in $\mathbb{R}^D$, such that the output of the network is given by $f_{\mathrm{w}}(\mathrm{x}) = \langle \mathrm{x}, \mathcal{P}(\mathrm{w}) \rangle$.

For fully connected networks, $\mathcal{P}_{\text{full}}(\mathrm{w}) = \mathrm{w}_1 \mathrm{w}_2 \ldots \mathrm{w}_L$,

For convolutional networks, $\mathcal{P}_{\text{conv}}(\mathrm{w}) = \left(\left(\left(\mathrm{w}_L^{\downarrow} \star \mathrm{w}_{L-1}\right) \star \mathrm{w}_{L-2}\right) \ldots \star \mathrm{w}_1\right)^{\downarrow}$, where $\mathrm{w}^{\downarrow}$ denotes the flipped vector corresponding to w given by $\mathrm{w}^{\downarrow}[k] = \mathrm{w}[D - k - 1]$ for $k = 0, 1, \ldots, D - 1$.

Multi-layer Linear Networks

Objective for training network

$$\min_{\mathsf{w} \in \mathcal{W}} \mathcal{L}_{\mathcal{P}}(\mathsf{w}) := \sum_{n=1}^{N} \ell\left(\left\langle \mathsf{x}_n, \mathcal{P}(\mathsf{w}) \right\rangle, y_n\right) \tag{1}$$

is equivalent to the following optimization over $\boldsymbol{\beta} = \mathcal{P}(\mathsf{w})$

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^D} \mathcal{L}(\boldsymbol{\beta}) := \sum_{n=1}^{N} \ell\left(\left\langle \mathsf{x}_n, \boldsymbol{\beta} \right\rangle, y_n\right) \tag{2}$$

But optimizing over w leads to different classifiers compared to optimizing over β directly.

Multi-layer Linear Networks

Consider problem (1) and (2) on a linearly separable dataset using the logistic loss.

Global infimum of $L(\beta)$ is 0, but not attainable by any finite β.

Want to find the direction $\overline{\boldsymbol{\beta}}^{\infty} = \lim_{t \to \infty} \frac{\boldsymbol{\beta}^{(t)}}{\left\| \boldsymbol{\beta}^{(t)} \right\|}$.

If this limit exist we say that $\boldsymbol{\beta}^{(t)}$ converges in direction to the limit direction $\overline{\boldsymbol{\beta}}^{\infty}$.

Multi-layer Linear Networks

Soudry et al. studied this implicit bias of gradient descent on (2) over the direct parameterization of β.

In this paper, we study the behavior of gradient descent on (1) for linear fully connected or convolutional networks.

Want to find the direction $\overline{\boldsymbol{\beta}}^{\infty} = \lim_{t \to \infty} \frac{\mathcal{P}\left(\mathrm{w}^{(t)}\right)}{\left\|\mathcal{P}\left(\mathrm{w}^{(t)}\right)\right\|}$

Table of Contents

Main Results

Assumptions.

In the following theorems, we characterize the limiting predictor $\overline{\boldsymbol{\beta}}^{\infty} = \lim_{t\to\infty} \frac{\boldsymbol{\beta}^{(t)}}{\left\|\boldsymbol{\beta}^{(t)}\right\|}$ under the following assumptions:

1. $\mathsf{w}^{(t)}$ minimize the objective, i.e., $\mathcal{L}_{\mathcal{P}}\left(\mathsf{w}^{(t)}\right) \to 0$.
2. $\mathsf{w}^{(t)}$, and consequently $\boldsymbol{\beta}^{(t)} = \mathcal{P}\left(\mathsf{w}^{(t)}\right)$, converge in direction to yield a separator $\overline{\boldsymbol{\beta}}^{\infty} = \lim_{t\to\infty} \frac{\boldsymbol{\beta}^{(t)}}{\left\|\boldsymbol{\beta}^{(t)}\right\|}$ with positive margin, i.e., $\min_n y_n \left\langle \mathsf{x}_n, \overline{\boldsymbol{\beta}}^{\infty} \right\rangle > 0$.
3. $\nabla_{\boldsymbol{\beta}} \mathcal{L}\left(\boldsymbol{\beta}^{(t)}\right)$ converge in direction.

(+) The phase of the Fourier coefficients $\widehat{\boldsymbol{\beta}}^{(t)}$ of the linear predictors $\boldsymbol{\beta}^{(t)}$ converge coordinate-wise.

Main Results

Theorem 1 (Linear fully connected networks).

For any depth L, almost all linearly separable datasets $\{\mathrm{x}_n, y_n\}_{n=1}^N$, almost all initializations $\mathrm{w}^{(0)}$, and any bounded sequence of step sizes $\{\eta_t\}_t$, consider the sequence gradient descent iterates $\mathrm{w}^{(t)}$ for minimizing $\mathcal{L}_{\mathcal{P}_{\text{full}}}(\mathrm{w})$ in (1) with exponential loss over L -layer fully connected linear networks.

$$\overline{\boldsymbol{\beta}}^{\infty} = \lim_{t\to\infty} \frac{\mathcal{P}_{\text{full}}\left(\mathrm{w}^{(t)}\right)}{\left\|\mathcal{P}_{\text{full}}\left(\mathrm{w}^{(t)}\right)\right\|} = \frac{\boldsymbol{\beta}^*_{\ell_2}}{\left\|\boldsymbol{\beta}^*_{\ell_2}\right\|},$$

$$\text{where } \boldsymbol{\beta}^*_{\ell_2} := \operatorname*{argmin}_{w} \|\boldsymbol{\beta}\|_2^2 \text{ s.t. } \forall n, y_n \left\langle \mathrm{x}_n, \boldsymbol{\beta} \right\rangle \geq 1$$

Main Results

Theorem 2a (Linear Convolutional Networks of any Depth).

For any depth L, the limit direction $\overline{\beta}^{\infty} = \lim_{t\to\infty} \frac{\mathcal{P}_{\text{conv}}\left(\mathrm{w}^{(t)}\right)}{\left\|\mathcal{P}_{\text{conv}}\left(\mathrm{w}^{(t)}\right)\right\|}$ is a scaling of a first order stationary point of the following optimization problem,

$$\min_{\beta} \|\widehat{\beta}\|_{2/L} \text{ s.t. } \forall n, y_n \langle \beta, \mathrm{x}_n \rangle \geq 1$$

where $\hat{\beta} \in C^D$ denote the Fourier coefficients of β, and the ℓ_p penalty given by $\|z\|_p = \left(\sum_{i=1}^{D} |z[i]|^p\right)^{1/p}$ is a norm for $p = 1$ and a quasi-norm for $p < 1$.

Table of Contents

Discussion

- Merely changing to a convolutional parameterization introduces radically different bias.
- For convenience, we studied one dimensional circular convolutions.
- These results can be directly extended to higher dimensional input signals and convolutions.
- When using convolutions as part of a larger network, with multiple parallel filters, max pooling, and non-linear activations, the situation is of course more complex, and we do not expect to get the exact same bias.
- Another important direction for future study is understanding the implicit bias for networks with multiple outputs.

Q & A

Thank you for listening.

btd63@snu.ac.kr