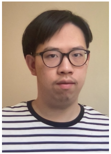
Deep Learning - Parameters and Functions

Implicit Biases

Guido Montúfar
montufar@math.ucla.edu

48th Winter Conference in Statistics, March 2024, Hemavan

Hui Jin



- “Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks”

Intuition

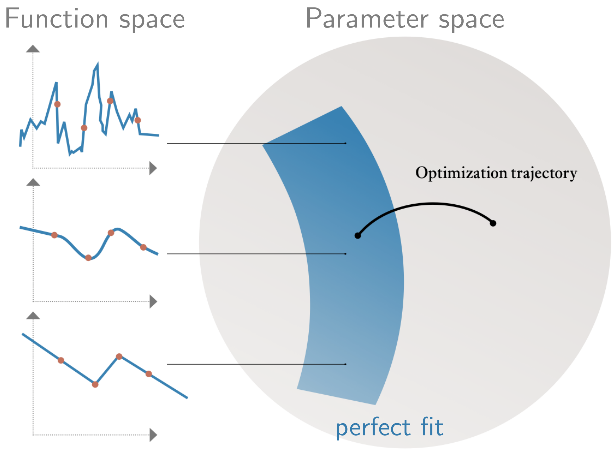


Figure 1: In an overparametrized model, there may be many different parameters and model functions that perfectly fit the training data

Motivation

- Neural networks in practice are often overparameterized
  - Number of model parameters $\gg$ Number of training samples
  - Can fit random labels
- Many global minima fit the training data perfectly
  - Most of them generalize horribly
- Nevertheless, deep models often generalize well, even without any explicit regularization
- The capacity of the hypothesis class alone does not explain this[1]

[1] Zhang et al. 2021.

Motivation

- One possible explanation is that optimization algorithms are implicitly biased towards selecting simple solutions
- Question: What kinds of minima does an optimization algorithm converge to?
  - Examples: maximum margin classifier, smooth interpolation, sparse function, ...
  - Depends on loss function, optimization algorithm, learning model, ...

Factors

- **Loss:** ℓ_2 loss (for regression); logistic loss (for classification).
- **Optimization algorithm:** gradient descent; stochastic gradient descent; mirror descent; steepest descent.
- **Learning model:** linear models; linear neural networks; neural networks; parametrization
- **Hyperparameters:** learning rate; initialization; mini-batch size;

Consider the following linear model with loss function $\ell(y_1, y_2)$:

$$L(w) = \sum_{i=1}^{n} \ell(\langle x_i, w \rangle, y_i),$$

and the gradient descent iterations

$$w_{t+1} = w_t - \eta \nabla L(w_t) = w_t - \eta \sum_{i=1}^{n} \nabla_1 \ell(\langle x_i, w_t \rangle, y_i) \cdot x_i.$$

Theorem 1 (Gunasekar et al. 2018a)

Consider a convex loss function ℓ with a unique finite minimizer ($\ell(y_1, y_2) = 0$ iff $y_1 = y_2$). Assume that the gradient descent iteration converges to the global minimum of $L(w)$ with zero loss, i.e., $L(w_t) \to 0$. Then the algorithm returns the unique solution of following constrained optimization problem:

$$\min_w \|w - w_0\|_2 \quad s.t.\ \langle x_i, w\rangle = y_i, \quad i = 1, \ldots, n. \tag{1}$$

The key idea is that the gradients are restricted to a n-dimensional subspace that is spanned by $\{x_i\}_{i=1}^n$ and is independent of w.

Gradient Descent.

- Let $w_\infty = \lim_{t\to\infty} w_t$. By assumption, $\langle x_i, w\rangle = y_i, i = 1, \dots, n$. The gradient descent iteration gives

$$w_\infty = w_0 - \sum_{t=0}^{\infty} \eta \sum_{i=1}^{n} \nabla_1 \ell(\langle x_i, w_t\rangle, y_i) \cdot x_i$$
$$= w_0 - \eta \sum_{i=1}^{n} x_i \sum_{t=0}^{\infty} \nabla_1 \ell(\langle x_i, w_t\rangle, y_i).$$

- The constrained optimization problem (1) is strongly convex. The first order optimality condition is

$$\begin{cases} w - w_0 + \sum_{i=1}^n \lambda_i x_i = 0, \\ \langle x_i, w\rangle = y_i, \quad i = 1, \dots, n. \end{cases} \tag{2}$$

- Setting $\lambda_i = \sum_{t=0}^{\infty} \nabla_1 \ell(\langle x_i, w_t\rangle, y_i)$, one has that w_∞ satisfies (2). So w_∞ is the solution of problem (1). $\square$

Given a strongly convex and differentiable potential ϕ, the mirror descent updates are:

$$w_{t+1} = \arg\min_{w} \eta\langle w, \nabla L(w_t)\rangle + D_\phi(w, w_t),$$

where $D_\phi(w, w') = \phi(w) - \phi(w') - \langle \nabla\phi(w'), w - w'\rangle$ is the Bregman divergence with respect to ϕ.

The first order optimality condition for the parameter update gives

$$\nabla\phi(w_{t+1}) = \nabla\phi(w_t) - \eta\nabla L(w_t).$$

Examples of ϕ:

- ℓ_2 norm: $\phi(w) = \frac{1}{2}\|w\|_2^2$, which leads to gradient descent;
- unnormalized negative entropy: $\phi(w) = \sum_i w_i \log w_i - w_i$.

Theorem 2 (Gunasekar et al. 2018a)

For any strongly convex potential ϕ. Assume that the mirror descent iteration converges to the global minimum of $L(w)$ with zero loss, i.e., $L(w_t) \to 0$. Then the algorithm returns the solution of following constrained optimization problem:

$$\min_w D_\phi(w, w_0) \quad s.t.\ \langle x_i, w \rangle = y_i, \quad i = 1, \ldots, n. \tag{3}$$

The key idea is that $\nabla\phi(w_{t+1})$ (called dual iterates) are restricted to a n-dimensional manifold $\nabla\phi(w_0) + \mathrm{span}(\{x_i\})$.

Proof

Mirror Descent.

The constrained optimization problem (3) is strongly convex.
The first order optimality condition of the problem is

$$\begin{cases} \nabla\phi(w) - \nabla\phi(w_0) + \sum_{i=1}^{n} \lambda_i x_i = 0, \\ \langle x_i, w\rangle = y_i, \quad i = 1, \ldots, n. \end{cases} \tag{4}$$

Since

$$\nabla\phi(w_{t+1}) = \nabla\phi(w_t) - \eta\nabla L(w_t)$$
$$\nabla\phi(w_\infty) = \nabla\phi(w_0) - \eta\sum_{i=1}^{n} x_i \sum_{t=0}^{\infty} \nabla_1\ell(\langle x_i, w_t\rangle, y_i).$$

One sees that w_∞ satisfies (4). So w_∞ is the solution of (3). □

Reparametrization and change of geometry

- D_ϕ can be approximated locally by a quadratic function

$$D_\phi(w, w') = (w - w')^T \nabla^2 \phi(w'')(w - w').$$

- If we use $D_\phi(w, w') = (w - w')^T K (w - w')$, the mirror descent iterations become:

$$w_{t+1} = w_t - \eta K^{-1} \nabla L(w_t).$$

- For mirror descent we have the update rule:

$$w_{t+1} = w_t - \eta (\nabla^2 \phi(w''))^{-1} \nabla L(w_t).$$

- If step size goes to 0, we have the following gradient flow:

$$\dot{w}_t = -(\nabla^2 \phi(w_t))^{-1} \nabla L(w_t).$$

Reparametrization and change of geometry

- Consider the least squares problem and reparametrization:

$$L(w) = \frac{1}{2}\sum_{i=1}^{n}(\langle x_i, u\rangle - y_i)^2 = \frac{1}{2}\|Xu - y\|_2^2,$$

  where $X = [x_1, \ldots, x_n]$ and $u = (w_1^2, \ldots, w_d^2)$ is the entry-wise square of w.
- The gradient flow over w is

$$\dot{w}(t) = -\nabla_w L(w(t)).$$

- If we consider the space of u, the above iteration becomes

$$\begin{aligned}\dot{u}(t) = D_w \cdot \dot{w}(t) &= -D_w \cdot \nabla_w L(w(t)) \\ &= -2D_w \cdot D_w \cdot X^T(Xu - y) \\ &= -2D_u \cdot X^T(Xu - y) \\ &= -2D_u \cdot \nabla_u L(u(t)),\end{aligned}$$

  where $D_u = \text{diag}(u)$ and $D_w = \text{diag}(w)$.

Example taken from 18.408 Lecture 4: https://people.csail.mit.edu/moitra/408b.html

Reparametrization and change of geometry

If we let $\phi(u) = \sum_{i=1}^{d}(u_i \log u_i - u_i)$, then we have

$$(\nabla^2 \phi(u(t)))^{-1} = D_u.$$

Then we can show that the following two iterations are equivalent:

1. Gradient descent under square parametrization;
2. Mirror descent under $\phi(u)$.

According to the implicit bias of mirror descent, $u(t)$ converges to the solution of the following optimization problem:

$$\min_u D_\phi(u, u_0) \quad \text{s.t. } \langle x_i, u\rangle = y_i, \quad i = 1, \ldots, n.$$

If $u_0 = \alpha \mathbf{1}$: as $\alpha \to 0$, we have $D_\phi(u, u_0) \to C_\alpha \|u\|_1$.

In classification problems, gradient descent on linear models converges to the ℓ_2 maximum margin solution if training data is linearly separable[2]

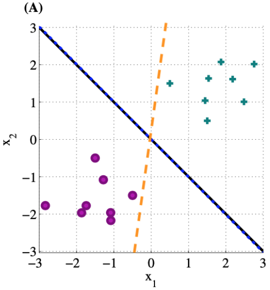


Figure 2: Implicit bias of gradient descent for classification problems.

[2] Soudry et al. 2018.

Classification: Linear Classifier

- Consider binary classification problem $y_i \in \{-1, +1\}$
- Linear decision boundaries $f(x) = \langle x_i, w \rangle$
- Decision rule $\hat{y}(x) = \mathrm{sign}(f(x))$
- Consider the exponential loss function $\ell(y_1, y_2) = \exp(-y_1 y_2)$,

$$L(w) = \sum_{i=1}^{n} \ell(\langle x_i, w \rangle, y_i)$$

- Gradient descent iteration

$$w_{t+1} = w_t - \eta \nabla L(w_t) = w_t - \eta \sum_{i=1}^{n} \ell(\langle x_i, w_t \rangle, y_i)(-x_i y_i)$$

- If the dataset is linearly separable, $L(w) \to 0$ only as $\|w\| \to \infty$.
- Study the limit direction $\bar{w}_\infty = \lim_{t\to\infty} \frac{w_t}{\|w_t\|}$.

Theorem 3 (Soudry et al. 2018)

For any dataset which is linearly separable, suitable learning rate η*, and any starting point* w_0*,* $\frac{w_t}{\|w_t\|}$ *converges to the unique solution of the SVM problem:*

$$\max_w \min_i y_i \langle x_i, w \rangle \quad s.t. \ \|w\|_2 \leq 1.$$

Proof idea.

- Suppose $\frac{w_t}{\|w_t\|}$ converges to some limit $\bar{w}_\infty$, so $w_t = g(t)\bar{w}_\infty + \rho(t)$ with $g(t) \to \infty$ and $\lim_{t\to\infty} \frac{\rho(t)}{g(t)} = 0$.
- The gradient at w_t is given by:

$$\begin{aligned}\nabla L(w_t) &= \sum_{i=1}^{n} \exp(-w_t^T x_i) x_i \\ &= \sum_{i=1}^{n} \exp(-g(t)\bar{w}_\infty^T x_i) \exp(-\rho(t)^T x_i) x_i\end{aligned}$$

- As $g(t) \to \infty$, only those samples with the largest exponents will contribute to the gradient. So w_t are asymptotically dominated by a non-negative linear combination of support vectors. These are precisely the KKT conditions for the SVM problem. □

Classification: Steepest Descent

Steepest descent with respect to a generic norm is given by:

$$w_{t+1} = w_t + \eta_t \Delta w_t, \text{ where } \Delta w_t = \arg\min \langle \nabla L(w_t), v \rangle + \frac{1}{2}\|v\|^2.$$

For classification problem we consider the exponential loss.

Theorem 4 (Gunasekar et al. 2018a)

For any dataset which is linearly separable, any norm $\|\cdot\|$, suitable learning rate η and any starting point w_0, $\frac{w_t}{\|w_t\|}$ converges to the solution of the optimization problem:

$$\max_w \min_i y_i \langle x_i, w \rangle \quad s.t. \ \|w\| \leq 1.$$

Implicit Bias for Linear Networks

- Deep linear networks can be regarded as parameterizations of linear models.
- Gunasekar et al. 2018b showed that gradient descent on full-width linear convolutional networks of depth L converges to a linear predictor related to the $\ell_{2/L}$ penalty in frequency domain.
- And gradient descent on fully-connected linear networks converges to ℓ_2 maximum margin solution regardless of depth.
- This elucidates the impact of the network architectures.
- The approximation ability may be the same, but the implicit bias of gradient descent is different.

1. Implicit Bias and Algorithmic Regularization

2. Implicit Bias in Wide Shallow ReLU Networks

Implicit bias of gradient descent

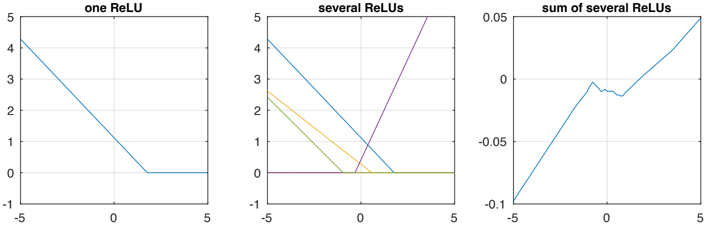
- Consider a shallow ReLU network with n hidden units,

$$f(x,\theta) = \sum_{i=1}^{n} W_i^{(2)}[\langle W_i^{(1)}, x\rangle + b_i^{(1)}]_+ + b^{(2)}.$$

- Initialize the parameters by independent samples of $(\mathcal{W}, \mathcal{B})$.
- For data $\{(x_j, y_j)\}_{j=1}^{M}$, select a function by gradient descent minimization of the squared error $L(\theta) = \sum_{j=1}^{M} \|f(x_j,\theta) - y_j\|^2$.

We discuss results for shallow ReLU networks from Jin and Montufar 2023.

- Consider first the univariate setting, $x \in \mathbb{R}$.
- A rectified linear unit $[w_i x + b_i]_+$ has breakpoint at $c_i = -b_i/w_i$.
- A density $p_{\mathcal{W},\mathcal{B}}$ induces a breakpoint density $p_{\mathcal{C}}$.



Implicit bias of GD in wide ReLU networks

Theorem 5 (Univariate regression)

- *Consider a feedforward network with a single input unit, a hidden layer of n rectified linear units, skip connections, and a single linear output unit.*
- *Assume standard parametrization and that for each hidden unit the input weight and bias are initialized from a sub-Gaussian $(\mathcal{W}, \mathcal{B})$ with joint density $p_{\mathcal{W},\mathcal{B}}$.*
- *Then, for any finite data set $\{(x_j, y_j)\}_{j=1}^{M}$ and sufficiently large n optimization of the MSE by full-batch gradient descent with sufficiently small step size converges to a parameter θ^* for which the output function $f(x, \theta^*)$ attains zero training error.*
- *Moreover,*

Implicit bias of GD in wide ReLU networks

Theorem 5 (Univariate regression)

letting

$$\zeta(x) = \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W, -Wx)\, \mathrm{d}W$$

and $S = \operatorname{supp}(\zeta) \cap [\min_i x_j, \max_i x_j]$, *we have*

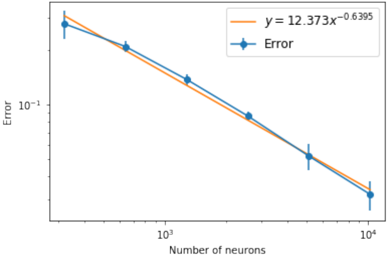
$$\|f(x, \theta^*) - g^*(x)\|_2 = O(n^{-\frac{1}{2}}), \quad x \in S$$

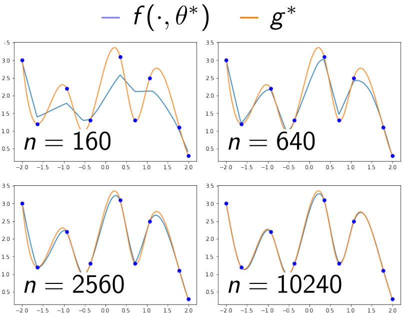
with high probability over the random initialization θ_0, *where* g^* *solves following variational problem:*

$$\begin{aligned} \min_{g \in C^2(S)} \quad & \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2 \, \mathrm{d}x \\ \text{subject to} \quad & g(x_j) = y_j, \quad j = 1, \ldots, M. \end{aligned} \tag{5}$$

Numerical illustration

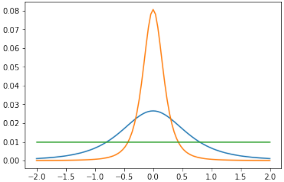
Uniform error between g^* and $f(\cdot, \theta^*)$



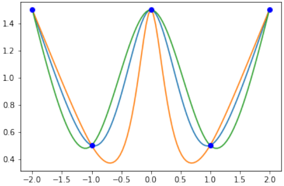


The curvature penalty function

ζ



Solution g^*



- The reciprocal curvature penalty is $\zeta(x) = \int_{\mathbb{R}} |W|^3 p_{\mathcal{W},\mathcal{B}}(W, -Wx)\, \mathrm{d}W$.
- We obtain the explicit form of ζ for various initialization procedures.
- We obtain parameter initialization procedures leading to any desired ζ.

Explicit form of the curvature penalty

Theorem 6 (Curvature penalty for various initializations)

1. *Gaussian initialization. Assume that* $\mathcal{W}$ *and* $\mathcal{B}$ *are independent,* $\mathcal{W} \sim \mathcal{N}(0, \sigma_w^2)$ *and* $\mathcal{B} \sim \mathcal{N}(0, \sigma_b^2)$*. Then* ζ *is given by* $\zeta(x) = \frac{2\sigma_w^3 \sigma_b^3}{\pi(\sigma_b^2 + x^2\sigma_w^2)^2}$.
2. *Binary-uniform initialization. Assume that* $\mathcal{W}$ *and* $\mathcal{B}$ *are independent,* $\mathcal{W} \in \{-1, 1\}$ *and* $\mathcal{B} \sim \mathcal{U}(-a_b, a_b)$ *with* $a_b \geq L$*. Then* ζ *is constant on* $[-L, L]$.
3. *Uniform initialization. Assume that* $\mathcal{W}$ *and* $\mathcal{B}$ *are independent,* $\mathcal{W} \sim \mathcal{U}(-a_w, a_w)$ *and* $\mathcal{B} \sim \mathcal{U}(-a_b, a_b)$ *with* $\frac{a_b}{a_w} \geq L$*. Then* ζ *is constant on* $[-L, L]$.

Gaussian initialization

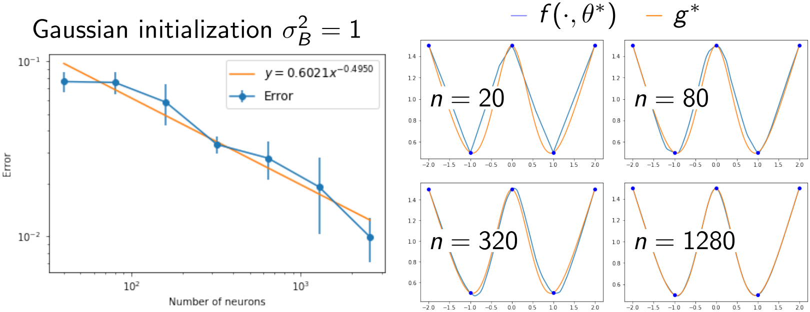


Figure 3: Initialization $W \sim N(0, 1)$ and $B \sim N(0, 1)$.

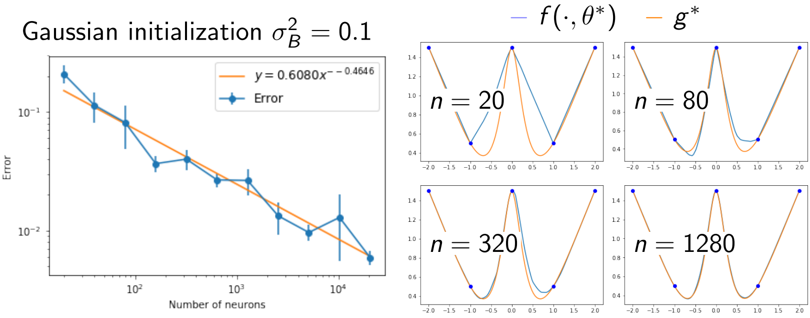


Figure 4: Initialization $W \sim N(0, 1)$ and $B \sim N(0, 0.1)$. In this case ζ that is more peaked at $x = 0$. Solutions more curvy around $x = 0$.

Exploiting the initialization

- With the presented bias description we can formulate heuristics for parameter initialization either to ease optimization or also to induce specific smoothness priors on the solutions.
- In particular, any curvature penalty $1/\zeta$ can be implemented by an appropriate choice of the initialization distribution.

Proposition 7 (Constructing any curvature penalty)

Given any function $\varrho\colon \mathbb{R} \to \mathbb{R}_{>0}$*, satisfying* $Z = \int_{\mathbb{R}} \frac{1}{\varrho} < \infty$*, if we set the density of* $\mathcal{C}$ *as* $p_{\mathcal{C}}(x) = \frac{1}{Z}\frac{1}{\varrho(x)}$ *and make* $\mathcal{W}$ *independent of* $\mathcal{C}$ *with non-vanishing second moment, then*

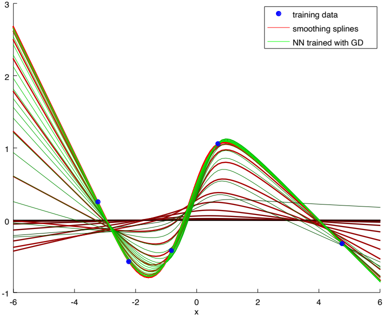
$$(\mathbb{E}(W^2|C = x)p_{\mathcal{C}}(x))^{-1} = (\mathbb{E}(W^2)p_{\mathcal{C}}(x))^{-1} \propto \varrho(x), \quad x \in \mathbb{R}.$$

Optimization trajectory in function space

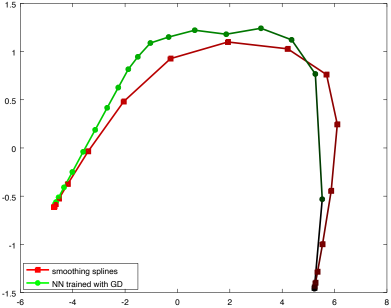
- Optimization trajectory described by smoothing splines

$$\min_{g \in C^2(S)} \quad \sum_{j=1}^{M} [g(x_j) - y_j]^2 + \frac{1}{\bar{\eta} t} \int_S \frac{1}{\zeta(x)} (g''(x) - f''(x, \theta_0))^2 \, \mathrm{d}x.$$

Trajectories of functions



2D PCA of the trajectories



Early stopping and spectral bias

- The result can be interpreted in combination with early stopping.
- The training trajectory is approximated by a smoothing spline, meaning that the network will filter out high frequencies which are usually associated to noise in the training data.
- This behavior is sometimes referred to as a spectral bias[3].

[3] Rahaman et al. 2019.

Generalization to multivariate regression

Theorem 8 (Multivariate regression)

- *Use the same network setting as in Theorem 5 except that the number of input units changes to d.*
- *Assume that for each hidden unit the input weight and bias are initialized from a sub-Gaussian $(\boldsymbol{\mathcal{W}}, \mathcal{B})$ where $\boldsymbol{\mathcal{W}}$ is a d-dimensional random vector and $\mathcal{B}$ is a random variable.*
- *Let $\mathcal{U} = \|\boldsymbol{\mathcal{W}}\|_2$, $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{W}}/\|\boldsymbol{\mathcal{W}}\|_2$, $\mathcal{C} = -\mathcal{B}/\|\boldsymbol{\mathcal{W}}\|_2$, and let $p_{\boldsymbol{\mathcal{V}},\mathcal{C}}$ be the joint density of $(\boldsymbol{\mathcal{V}}, \mathcal{C})$.*
- *Then, for any finite data set $\{(\mathbf{x}_j, y_j)\}_{i=1}^{M}$ and sufficiently large n optimization of the MSE by full-batch gradient descent with sufficiently small step size converges to a parameter θ^* for which $f(\cdot, \theta^*)$ attains zero training error.*
- *Moreover,*

Generalization to multivariate regression

Theorem 8 (Multivariate regression)

letting $\zeta(\boldsymbol{V}, c) = p_{\boldsymbol{\mathcal{V}},\mathcal{C}}(\boldsymbol{V}, c)\mathbb{E}(\mathcal{U}^2|\boldsymbol{\mathcal{V}} = \boldsymbol{V}, \mathcal{C} = c)$*, we have*

$$\|f(\mathbf{x}, \theta^*) - g^*(\mathbf{x})\|_2 = O(n^{-\frac{1}{2}}), \quad \mathbf{x} \in \mathbb{R}^d$$

(the 2-norm over $\mathbb{R}^d$*) whp over* θ_0*, where* g^* *solves*

$$\begin{aligned} \min_{g \in C(\mathbb{R}^d)} \quad & \int_{\mathrm{supp}(\zeta)} \frac{1}{\zeta(\boldsymbol{V}, c)} \left(\mathcal{R}\{(-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0))\}(\boldsymbol{V}, c)\right)^2 \, \mathrm{d}\boldsymbol{V}\mathrm{d}c \\ \text{s.t.} \quad & g(\mathbf{x}_j) = y_j, \quad j = 1, \ldots, M, \\ & \mathcal{R}\{(-\Delta)^{(d+1)/2}(g - f(\cdot, \theta_0))\}(\boldsymbol{V}, c) = 0, \quad (\boldsymbol{V}, c) \notin \mathrm{supp}(\zeta). \end{aligned}$$

Here $\mathcal{R}$ *is the Radon transform,* $\mathcal{R}\{f\}(\boldsymbol{\omega}, b) := \int_{\langle \boldsymbol{\omega}, \mathbf{x}\rangle = b} f(\mathbf{x})\mathrm{d}s(\mathbf{x})$*, and the power of the negative Laplacian* $(-\Delta)^{(d+1)/2}$ *is the operator defined in Fourier domain by* $\widehat{(-\Delta)^{(d+1)/2} f}(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|^{d+1}\widehat{f}(\boldsymbol{\xi})$*.*

Generalization to other activation functions

Theorem 9 (Different activation functions)

- *Use the same setting as in Theorem 5 except that we use the activation function* ϕ *instead of ReLU.*
- *Suppose that* ϕ *is a Green's function of a linear operator* L*, i.e.* $\mathrm{L}\phi = \delta$*, where* δ *denotes the Dirac delta function.*
- *Assume that the activation function* ϕ *is homogeneous of degree* k*, i.e.* $\phi(ax) = a^k \phi(x)$ *for all* $a > 0$*.*
- *Then we can find a function* p *satisfying* $\mathrm{L}p \equiv 0$ *and adjust training data* $\{(x_j, y_j)\}_{j=1}^M$ *to* $\{(x_j, y_j - p(x_j)\}_{j=1}^M$*.*

Generalization to other activation functions

Theorem 9 (Different activation functions)

After that, the statement in Theorem 5 holds with

$$\min_{g \in C^2(S)} \quad \int_S \frac{1}{\zeta(x)} [\mathrm{L}(g(x) - f(x, \theta_0))]^2 \, \mathrm{d}x$$
$$\text{s.t.} \quad g(x_j) = y_j - p(x_j), \quad j = 1, \dots, M,$$

where

$$\zeta(x) = p_{\mathcal{C}}(x) \mathbb{E}(\mathcal{W}^{2k} | \mathcal{C} = x)$$

and $S = \mathrm{supp}(\zeta) \cap [\min_i x_i, \max_i x_i]$.

From Theorem 5 we can extract generalization results such as

- In the univariate noisless model for a target g_0 on $[a, b]$, if ζ uniform, then
$$\|g^* - g_0\|_\infty \leq C\|g_0^{(4)}\|_\infty h^4$$
where $g^{(4)}$ is the fourth derivative of g_0 and $h = \max_i x_{i+1} - x_i$.
- For univariate noisy models with $y_j = g_0(x_j) + \epsilon_j$, ϵ_j independent zero mean with variance σ^2, if x_i uniform partition and ζ uniform, using early stopping with $t = \Theta(M^{4/5})$, then
$$\mathbb{E}\|g^* - g_0\|_2^2 = O(M^{-4/5})$$
- Similar observations can be obtained in more general settings such as non-uniform training inputs, non-constant ζ

Related works

- Zhang et al. 2019 described the implicit bias of gradient descent in the kernel regime as minimizing a kernel norm from initialization, subject to fitting the training data.
- Savarese et al. 2019 showed infinite-width networks with 2-norm weight regularization represent functions with smallest 1-norm of the second derivative, an example of which are linear splines.
- Williams et al. 2019 showed a similar result for univariate shallow ReLU nets training only the output layer from zero initialization.

Summary

- Gradient descent training of overparametrized ReLU networks is biased towards functions with low curvature.
- The parameter initialization procedure determines the curvature penalty function $1/\zeta$.
- Generalizations to multivariate ReLU networks, different activation functions, and optimization trajectories.

Further topics

- Spectral bias
- Implicit bias in mildly overparametrized nets
- Other optimizers and stability
- Role of the data

References I

- Arora, Sanjeev et al. (2019). "Implicit regularization in deep matrix factorization". In: *Advances in Neural Information Processing Systems* 32.
- Bowman, Benjamin and Guido Montufar (2022). "Implicit Bias of MSE Gradient Optimization in Underparameterized Neural Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=VLgmhQDVBV.
- Gunasekar, Suriya et al. (2017). "Implicit regularization in matrix factorization". In: *Advances in Neural Information Processing Systems* 30.
- Gunasekar, Suriya et al. (2018a). "Characterizing Implicit Bias in Terms of Optimization Geometry". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR, pp. 1832–1841. URL: http://proceedings.mlr.press/v80/gunasekar18a.html.

- Gunasekar, Suriya et al. (2018b). "Implicit bias of gradient descent on linear convolutional networks". In: *Advances in Neural Information Processing Systems* 31.
- Jin, Hui and Guido Montufar (2023). "Implicit Bias of Gradient Descent for Mean Squared Error Regression with Two-Layer Wide Neural Networks". In: *Journal of Machine Learning Research* 24.137, pp. 1–97. URL: http://jmlr.org/papers/v24/21-0832.html.
- Li, Yuanzhi, Tengyu Ma, and Hongyang Zhang (2018). "Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations". In: *Conference On Learning Theory*. PMLR, pp. 2–47.
- Rahaman, Nasim et al. (2019). "On the spectral bias of neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 5301–5310.

- Razin, Noam and Nadav Cohen (2020). "Implicit regularization in deep learning may not be explainable by norms". In: *Advances in neural information processing systems* 33, pp. 21174–21187.
- Savarese, Pedro et al. (2019). "How do infinite width bounded norm networks look in function space?" In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, pp. 2667–2690. URL: http://proceedings.mlr.press/v99/savarese19a.html.
- Soudry, Daniel et al. (2018). "The implicit bias of gradient descent on separable data". In: *Journal of Machine Learning Research* 19.1, pp. 2822–2878.
- Williams, Francis et al. (2019). "Gradient dynamics of shallow univariate ReLU networks". In: *Advances in Neural Information Processing Systems*, pp. 8378–8387.

References IV

Zhang, Chiyuan et al. (2021). "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3, pp. 107–115.

Zhang, Yaoyu et al. (2019). "A type of generalization error induced by initialization in deep neural networks". In: *arXiv preprint arXiv:1905.07777*.