Stationary Points of Shallow Neural Networks with Quadratic Activation Function

Eren C. Kızıldağ (MIT), joint work with David Gamarnik and Ilias Zadik (MIT)

Simons Institute, Joint CCSI/GMOS student seminar

November 19, 2021

Overview

Motivation

- NN models achieved great practical success:

  Image recognition, image classification, speech recognition, natural language processing, game playing,...

- **Rigorous understanding?** Still an ongoing quest.
- **Example:**
  - Training is (worst-case) **NP-hard** (Blum and Rivest [89]).
  - **Loss function:** In general, highly non-convex.
  - **Gradient descent:** Simple, first order method. Yet, great empirical success.

This Work

Our Motivation

- Provide further insights for these networks.
- Our focus:
  - **Training.** Through the landscape lens. Convergence of GD due to benign landscape.
  - **Initialization.** In the context of random planted weights.
  - **Generalization.**

Setup and Main Assumptions

- One hidden layer, width $m \in \mathbb{N}$. Quadratic activation, $\sigma(x) = x^2$.
- **Realizable Model.** Planted weights $W^* \in \mathbb{R}^{m\times d}$. $j^{\rm th}$ row of W^*, $W_j^* \in \mathbb{R}^d$.
- For $X \in \mathbb{R}^d$, computes the *label*

$$f(W^*;X) = \sum_{1\le j\le m} \langle W_j^*, X\rangle^2 = \|W^*X\|_2^2.$$

Main Assumptions

- $\mathrm{rank}(W^*) = d$. Hence, $m \ge d$.
- Data $X \in \mathbb{R}^d$ has i.i.d. centered **sub-Gaussian** coordinates (can sometimes be relaxed).

Setup and Main Assumptions

- Generate i.i.d. $X_i \in \mathbb{R}^d$, $1 \le i \le N$. Label $Y_i = f(W^*; X_i)$.
- **Learner:** Given training data (X_i, Y_i), $1 \le i \le N$, find a NN with small **training error/empirical risk**:

$$\widehat{\mathcal{L}}(W) \triangleq \frac{1}{N} \sum_{1 \le i \le N} \left(Y_i - \sum_{1 \le j \le m} \langle W_j, X \rangle^2 \right)^2$$

  Run any training algorithm (e.g. GD, SGD, etc.) to solve $\min_{W \in \mathbb{R}^{m \times d}} \widehat{\mathcal{L}}(W)$.
- **Generalization ability.** Use "learned" W to **predict unseen** data. Quantified by **generalization error/population risk**:

$$\mathcal{L}(W) \triangleq \mathbb{E}\left[\left(f(W; X) - f(W^*; X) \right)^2 \right]$$

Prior Work: Planted Weights, Sub-Gaussianity, and Quadratic Networks

- Shallow NN with **planted** weights and **Gaussian** data is popular in literature:

  Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

- **Quadratic networks**, also popular:

  Du and Lee [18]; Soltanolkotabi, Javanmard, and Lee [18]; Mannelli, Vanden-Eijnden, and Zdeborová [20]; and Abbe, Boix-Adsera, Brennan, Bresler, and Nagaraj [21].

- **Quadratic activation:** Admittedly stylized. However,
  - Stack blocks of quadratic networks to approximate deep sigmoid networks (Livni, Shalev-Shwartz, and Shamir [14]).
  - **Second order approximation** to general nonlinearities (Venturi, Bandeira, and Bruna [18]).
- **Quadratic networks**: Provide further insights on complex architectures.

Overview

Optimization Landscape: An Energy Barrier

Theorem (Gamarnik, K.; and Zadik, 2020)

$X_i \in \mathbb{R}^d$, $1 \le i \le N$, i.i.d. data with centered i.i.d. sub-Gaussian coordinates. $Y_i = f(W^*; X_i)$. Then with high probability,

$$\min_{\substack{W \in \mathbb{R}^{m \times d} \\ \mathrm{rank}(W) \le d-1}} \widehat{\mathcal{L}}(W) = \min_{\substack{W \in \mathbb{R}^{m \times d} \\ \mathrm{rank}(W) \le d-1}} \frac{1}{N} \sum_{1 \le i \le N} (Y_i - f(W; X_i))^2 \ge \frac{1}{2} C \sigma_{\min}(W^*)^4.$$

- $C > 0$: absolute constant, depends only on **(conditional) moments of data**.
- **Energy barrier** for $\widehat{\mathcal{L}}(\cdot)$: for $\mathrm{rank}(W) < d$, $\widehat{\mathcal{L}}(W)$ is **bounded away from zero** by an explicit quantity. Analogue result for **population risk,** $\mathcal{L}(W)$.
- **Tight** up to a multiplicative constant.
- **Sub-Gaussianity not essential**: $\mathbb{P}(|X_i(j)| > t) \le \exp(-\Omega(t^\alpha))$ type tail behavior is ok.

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

Theorem (Gamarnik, K.; and Zadik, 2020)

Let $\mathrm{rank}(W) = d$ *and* $\nabla_W \widehat{\mathcal{L}}(W) = 0$. *Then,* $\widehat{\mathcal{L}}(W) = 0$.
Furthermore, if $N \geq d(d+1)/2$, *then* $W = QW^*$ *for some orthogonal* $Q \in \mathbb{R}^{m \times m}$.

- Analogue result holds for **population risk**.
- **No full-rank saddle points** for $\widehat{\mathcal{L}}(\cdot)$ and $\mathcal{L}(\cdot)$.
- **Benign landscape** below the **energy barrier**:
  recall that whp no rank-deficient $W \in \mathbb{R}^{m \times d}$ below the barrier.

Next. Benign landscape $\implies$ Convergence of GD.

Optimization Landscape: Convergence of Gradient Descent

Theorem (Gamarnik, K.; and Zadik, 2020)

Suppose $\widehat{\mathcal{L}}(W_0) < \frac{1}{2}C_5\sigma_{\min}(W^*)^4$. *Then, there is a high probability event on which:*

- *Running GD (with appropriate step size) generates a full-rank,* $\epsilon-$*approximate stationary point* $W \in \mathbb{R}^{m\times d}$ $(\|\nabla\widehat{\mathcal{L}}(W)\|_F \le \epsilon)$ *in time* $\mathrm{poly}(\epsilon^{-1}, d)$.
- *For this* W, $\widehat{\mathcal{L}}(W) \le C\epsilon\sigma_{\min}(W^*)^{-2}\mathrm{poly}(d)$, $\mathcal{L}(W) \le C'\epsilon\sigma_{\min}(W^*)^{-1}\mathrm{poly}(d)$; *and* $\|W^TW - (W^*)^TW^*\|_F \le C''\epsilon^{\frac{1}{2}}\sigma_{\min}(W^*)^{-1}\mathrm{poly}(d)$. $C, C', C'' > 0$ *constants.*

- GD finds in **polynomial time** an **approx. stationary** W, if **initializated "properly"**.
- W^TW uniformly close to planted $(W^*)^TW^*$: **good generalization**.
- **Technicality.** Control the **condition number** of a certain matrix with i.i.d. rows consisting of **tensorized** $X_i^{\otimes 2}$. Analyze **spectrum of expected covariance** matrix of **tensorized** data.

Remarks

- **Energy barrier, separating rank-deficient points**: only **full rank** W below the barrier.
- **Full-rank stationary points are globally optimal**:
  benign landscape below the barrier, no spurious full-rank stationary points.
- GD, when **initialized properly**, "approximately" minimizes $\widehat{\mathcal{L}}(W)$, and recovers W^* in polynomial time. **Learned** W has **good generalization.**
- **Technicalities.**
  - Covering and concentration arguments.
  - Novel concentration result for matrices having i.i.d. rows with tensorized data $X_i^{\otimes 2}$.
  - Uses tools from our recent work, Emschwiller, Gamarnik, K., and Zadik [20].

,

Next: *"How to initialize properly?"*

Overview

Proper Initialization

- **Recall:** GD is successful provided initialized *properly*.
- **Focus.** Initialization in the context of **random** $W^* \in \mathbb{R}^{m\times d}$:
  - NN with random weights: **initial loss landscape**.
  - Closely related to **random feature methods**, Rahimi & Recht [09].
  - Approximate dynamical systems (Gonon et al. [20]). Also studied for extreme learning machine (Huang et al. [06]), and in random matrix theory (Pennington & Worah [17]).
- **Intuition.**
  - $\widehat{\mathcal{L}}(W)/\mathcal{L}(W)$ determined by **spectrum** of $W^TW - (W^*)^TW^*$ and **data moments**.
  - **Tight concentration** for Wishart spectrum, $(W^*)^TW^*$. **Semicircle law**: Bai & Yin [88,93].

$\implies$ Spectrum of $W^TW - (W^*)^TW^*$ can be controlled by tuning W.

$\implies$ $\widehat{\mathcal{L}}(W)/\mathcal{L}(W)$ can be controlled by tuning W.

Proper Initialization. Main Result.

Theorem (Gamarnik, K.; and Zadik, 2020)

$W^* \in \mathbb{R}^{m\times d}$ has centered i.i.d. entries with unit variance, finite fourth moment.
Data $X_i \in \mathbb{R}^d$, $1 \le i \le N$ has i.i.d. centered sub-Gaussian coordinates.
Initialize W_0 so that $W_0^T W_0 = mI_{d\times d}$. Then, whp

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2} C \sigma_{\min}(W^*)^4,$$

provided $m > C'd^2$ for a sufficiently large constant $C' > 0$.

- **Deterministic initialization.** Below the energy barrier, provided the NN is sufficiently **overparameterized**, $m = \Omega(d^2)$. Based on the semicircle law.
- Analogous result for the **population risk**.
- For W^* with i.i.d. standard normal entries, **non-asymptotic** guarantees available.

Overview

Sample Complexity

Main question.

"What is the smallest number of samples required to claim that small empirical risk also controls the generalization error?"

Theorem (Gamarnik, K.; and Zadik, 2020)

$X_i \in \mathbb{R}^d$, $1 \le i \le N$ *be data (not necessarily random).* $\mathcal{S} \triangleq \{A \in \mathbb{R}^{d\times d} : A^T = A\}$.

- *Suppose* $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$, *and* $\widehat{m} \in \mathbb{N}$ *arbitrary. Then, for any* $W \in \mathbb{R}^{\widehat{m}\times d}$ *"interpolating" the data* ($f(W;X_i) = f(W^*;X_i)$, $1 \le i \le N$), $W^T W = (W^*)^T W^*$. *Thus,* W *generalizes well:* $\mathcal{L}(W) = 0$.
- *Suppose* $\mathrm{span}(X_i X_i^T : 1 \le i \le N) \subsetneq \mathcal{S}$. *Then for any* $\widehat{m} \in \mathbb{N}$, *there exists a* $W \in \mathbb{R}^{\widehat{m}\times d}$ *such that while* W *interpolates the data* ($f(W;X_i) = f(W^*;X_i)$ *for every* i), $W^T W \ne (W^*)^T W^*$. *In particular,* $\mathcal{L}(W) > 0$ *(where* $\mathcal{L}$ *is defined w.r.t. any jointly continuous distribution on* $\mathbb{R}^d$*).*

Sample Complexity: Remarks.

- If $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of $\widehat{\mathcal{L}}(\cdot)$ has necessarily **zero generalization error**.
- **Not retrospective:** $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$ can be checked beforehand.
- **No randomness.** Purely geometrical, necessary and sufficient condition.
- If W has **non-zero but small** $\widehat{\mathcal{L}}(W)$, earlier results allow bounding $\|W^TW - (W^*)^TW^*\|_F$, and $\mathcal{L}(W)$.
- **Parameter** $\widehat{m} \in \mathbb{N}$**:** Interpolating NN need **not** have the same width m.
- Provided the span condition holds, **any** interpolant (potentially overparameterized) **generalize well**.

Theorem

As soon as $N \ge d(d+1)/2$, $\mathbb{P}\big[\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}\big] = 1$.

Sample Complexity Bound for Planted Network.

Theorem (Gamarnik, K.; and Zadik, 2020)

$X_i \in \mathbb{R}^d$, $1 \le i \le N$, i.i.d. with a jointly continuous distribution. Let $W^* \in \mathbb{R}^{m\times d}$ with $\mathrm{rank}(W^*) = d$ and $Y_i = f(W^*; X_i) = \sum_{1\le j\le m}\langle W_j^*, X_i\rangle^2$.

- *Suppose $N \ge d(d+1)/2$, and $\widehat{m} \in \mathbb{N}$. Then, with probability one over X_i, $1 \le i \le N$ the following holds: if $f(W; X_i) = f(W^*; X_i)$, $1 \le i \le N$, then $f(W; x) = f(W^*; x)$ for every $x \in \mathbb{R}^d$.*
- *Suppose X_i has centered i.i.d. coordinates with variance μ_2 and (finite) fourth moment μ_4, and $N < d(d+1)/2$. Then, there exists a $W \in \mathbb{R}^{m\times d}$ such that while $\widehat{\mathcal{L}}(W) = 0$ (namely $f(W; X_i) = f(W^*; X_i)$ for $1 \le i \le N$),*

$$\mathcal{L}(W) \ge \min\{\mu_4 - \mu_2^2, 2\mu_2^2\}\sigma_{\min}(W^*)^4.$$

Lower bound in second part: coincides with energy barrier.

Thank you!