Stationary Points of Shallow Neural Networks with Quadratic Activation Function

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

arXiv:1912.01599

MIT MLTea

July 28, 2020

Overview

Motivation

Motivation

(a) Neural networks are extremely powerful in practice: natural language processing, image recognition, image classification, speech recognition, game playing, etc.

Motivation

(a) Neural networks are extremely powerful in practice: natural language processing, image recognition, image classification, speech recognition, game playing, etc.

(b) A mathematical understanding of this success is largely missing.

Motivation

(a) Neural networks are extremely powerful in practice: natural language processing, image recognition, image classification, speech recognition, game playing, etc.

(b) A mathematical understanding of this success is largely missing.

(c) For instance, training is NP-hard (Blum and Rivest [89]). However, the gradient descent has great empirical success.

Motivation

(a) Neural networks are extremely powerful in practice: natural language processing, image recognition, image classification, speech recognition, game playing, etc.

(b) A mathematical understanding of this success is largely missing.

(c) For instance, training is NP-hard (Blum and Rivest [89]). However, the gradient descent has great empirical success.

(d) Our main motivation is to provide further insights for these networks. In particular, we focus on both **training (through the lens of optimization landscape), initialization**, and **generalization** aspects.

Model

Model

(a) Shallow architecture, one hidden layer of width $m \in \mathbb{Z}^+$, quadratic activation ($\sigma(x) = x^2$). Data $X \in \mathbb{R}^d$ has i.i.d. centered sub-Gaussian coordinates. Realizable model: labels are generated by a teacher network with planted weights $W^* \in \mathbb{R}^{m\times d}$.

Model

(a) Shallow architecture, one hidden layer of width $m \in \mathbb{Z}^+$, quadratic activation ($\sigma(x) = x^2$). Data $X \in \mathbb{R}^d$ has i.i.d. centered sub-Gaussian coordinates. Realizable model: labels are generated by a teacher network with planted weights $W^* \in \mathbb{R}^{m \times d}$.

(b) The network computes, for each $X \in \mathbb{R}^d$, the label $f(W^*; X) = \sum_{1 \le j \le m} \langle W_j^*, X \rangle^2$ (which is $\|W^* X\|_2^2$). Here, $W_j^* \in \mathbb{R}^d$ is the $j^{\rm th}$ row of W^*. Also, $\mathrm{rank}(W^*) = d$.

Model

(a) Shallow architecture, one hidden layer of width $m \in \mathbb{Z}^+$, quadratic activation ($\sigma(x) = x^2$). Data $X \in \mathbb{R}^d$ has i.i.d. centered sub-Gaussian coordinates. Realizable model: labels are generated by a teacher network with planted weights $W^* \in \mathbb{R}^{m \times d}$.

(b) The network computes, for each $X \in \mathbb{R}^d$, the label $f(W^*; X) = \sum_{1 \le j \le m} \langle W_j^*, X \rangle^2$ (which is $\|W^* X\|_2^2$). Here, $W_j^* \in \mathbb{R}^d$ is the $j^{\rm th}$ row of W^*. Also, $\mathrm{rank}(W^*) = d$.

(c) Generate $X_i \in \mathbb{R}^d$, $1 \le i \le N$ i.i.d. Let $Y_i = f(W^*; X_i)$, $1 \le i \le N$ be the associated label. **The empirical risk** for any "candidate" $W \in \mathbb{R}^{m \times d}$:

$$\widehat{\mathcal{L}}(W) \triangleq \frac{1}{N} \sum_{1 \le i \le N} \left(Y_i - f(W^*; X_i)\right)^2 .$$

$\widehat{\mathcal{L}}(\cdot)$ has a highly non-convex nature.

Model

(a) Shallow architecture, one hidden layer of width $m \in \mathbb{Z}^+$, quadratic activation ($\sigma(x) = x^2$). Data $X \in \mathbb{R}^d$ has i.i.d. centered sub-Gaussian coordinates. Realizable model: labels are generated by a teacher network with planted weights $W^* \in \mathbb{R}^{m\times d}$.

(b) The network computes, for each $X \in \mathbb{R}^d$, the label $f(W^*;X) = \sum_{1\le j\le m}\langle W_j^*, X\rangle^2$ (which is $\|W^*X\|_2^2$). Here, $W_j^* \in \mathbb{R}^d$ is the $j^{\rm th}$ row of W^*. Also, $\mathrm{rank}(W^*) = d$.

(c) Generate $X_i \in \mathbb{R}^d$, $1 \le i \le N$ i.i.d. Let $Y_i = f(W^*;X_i)$, $1 \le i \le N$ be the associated label. **The empirical risk** for any "candidate" $W \in \mathbb{R}^{m\times d}$:

$$\widehat{\mathcal{L}}(W) \triangleq \frac{1}{N}\sum_{1\le i\le N}\left(Y_i - f(W^*;X_i)\right)^2.$$

$\widehat{\mathcal{L}}(\cdot)$ has a highly non-convex nature.

(d) **Goal of the learner.** Solve the **empirical risk minimization** problem $\min_{W\in\mathbb{R}^{m\times d}}\widehat{\mathcal{L}}(W)$ and understand its **generalization ability**, as measured by the **population risk**

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

Some Prior Work (and also more motivation!)

Some Prior Work (and also more motivation!)

(a) Such shallow architectures with planted weights and Gaussian input data have been explored previously in literature: Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

Some Prior Work (and also more motivation!)

(a) Such shallow architectures with planted weights and Gaussian input data have been explored previously in literature: Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

(b) Quadratic networks, in particular, were considered by Du and Lee [18], Soltanolkotabi, Javanmard, and Lee [18].

Some Prior Work (and also more motivation!)

(a) Such shallow architectures with planted weights and Gaussian input data have been explored previously in literature: Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

(b) Quadratic networks, in particular, were considered by Du and Lee [18], Soltanolkotabi, Javanmard, and Lee [18].

(c) Quadratic activation function is, admittedly, stylized. However,

Some Prior Work (and also more motivation!)

(a) Such shallow architectures with planted weights and Gaussian input data have been explored previously in literature: Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

(b) Quadratic networks, in particular, were considered by Du and Lee [18], Soltanolkotabi, Javanmard, and Lee [18].

(c) Quadratic activation function is, admittedly, stylized. However,

- Block of quadratic activations can be stacked together to approximate deeper networks with sigmoid activations (Livni et al. [14]).

Some Prior Work (and also more motivation!)

(a) Such shallow architectures with planted weights and Gaussian input data have been explored previously in literature: Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

(b) Quadratic networks, in particular, were considered by Du and Lee [18], Soltanolkotabi, Javanmard, and Lee [18].

(c) Quadratic activation function is, admittedly, stylized. However,
   - Block of quadratic activations can be stacked together to approximate deeper networks with sigmoid activations (Livni et al. [14]).
   - They serve as a second order approximation of general nonlinear activations (Venturi et al. [18]).

Some Prior Work (and also more motivation!)

(a) Such shallow architectures with planted weights and Gaussian input data have been explored previously in literature: Du et al. [17], Li & Yuan [17], Tian [17], Zhong et al. [17], Soltanolkotabi [17], Brutzkus & Globerson [17], ...

(b) Quadratic networks, in particular, were considered by Du and Lee [18], Soltanolkotabi, Javanmard, and Lee [18].

(c) Quadratic activation function is, admittedly, stylized. However,
   - Block of quadratic activations can be stacked together to approximate deeper networks with sigmoid activations (Livni et al. [14]).
   - They serve as a second order approximation of general nonlinear activations (Venturi et al. [18]).

(d) Study of quadratic networks is an attempt to gain further insights on more complex networks.

Overview

Optimization Landscape: An Energy Barrier

Optimization Landscape: An Energy Barrier

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

Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$ be i.i.d. data with centered i.i.d. sub-Gaussian coordinates; and $Y_i = f(W^*; X_i)$ be the associated label. Then, under certain technical assumptions (in particular if $N = d^{O(1)}$), it holds that with high probability,

$$\min_{\substack{W \in \mathbb{R}^{m \times d} \\ \mathrm{rank}(W) \le d-1}} \widehat{\mathcal{L}}(W) = \min_{\substack{W \in \mathbb{R}^{m \times d} \\ \mathrm{rank}(W) \le d-1}} \frac{1}{N} \sum_{1 \le i \le N} (Y_i - f(W; X_i))^2 \ge \frac{1}{2} C_5 \sigma_{\min}(W^*)^4.$$

Here, $C_5 > 0$ is a constant determined only by the (conditional) moments of data distribution.

Optimization Landscape: An Energy Barrier

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

Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$ be i.i.d. data with centered i.i.d. sub-Gaussian coordinates; and $Y_i = f(W^*; X_i)$ be the associated label. Then, under certain technical assumptions (in particular if $N = d^{O(1)}$), it holds that with high probability,

$$\min_{\substack{W \in \mathbb{R}^{m\times d} \\ \mathrm{rank}(W) \le d-1}} \widehat{\mathcal{L}}(W) = \min_{\substack{W \in \mathbb{R}^{m\times d} \\ \mathrm{rank}(W) \le d-1}} \frac{1}{N} \sum_{1 \le i \le N} (Y_i - f(W; X_i))^2 \ge \frac{1}{2} C_5 \sigma_{\min}(W^*)^4.$$

Here, $C_5 > 0$ is a constant determined only by the (conditional) moments of data distribution.

(a) An energy barrier in the landscape of $\widehat{\mathcal{L}}(\cdot)$: if $\mathrm{rank}(W) < d$, then $\widehat{\mathcal{L}}(W)$ is bounded away from zero by an explicit constant. Analogue result for the population risk $\mathcal{L}(W)$.

Optimization Landscape: An Energy Barrier

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

Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$ be i.i.d. data with centered i.i.d. sub-Gaussian coordinates; and $Y_i = f(W^*; X_i)$ be the associated label. Then, under certain technical assumptions (in particular if $N = d^{O(1)}$), it holds that with high probability,

$$\min_{\substack{W \in \mathbb{R}^{m\times d} \\ \mathrm{rank}(W) \le d-1}} \widehat{\mathcal{L}}(W) = \min_{\substack{W \in \mathbb{R}^{m\times d} \\ \mathrm{rank}(W) \le d-1}} \frac{1}{N} \sum_{1\le i\le N} (Y_i - f(W; X_i))^2 \ge \frac{1}{2} C_5 \sigma_{\min}(W^*)^4.$$

Here, $C_5 > 0$ is a constant determined only by the (conditional) moments of data distribution.

(a) An energy barrier in the landscape of $\widehat{\mathcal{L}}(\cdot)$: if $\mathrm{rank}(W) < d$, then $\widehat{\mathcal{L}}(W)$ is bounded away from zero by an explicit constant. Analogue result for the population risk $\mathcal{L}(W)$.

(b) Tight up to a multiplicative constant. Sub-Gaussianity not essential: $\mathbb{P}(|X_i(j)| > t) \le \exp(-Ct^\alpha)$ type tail behavior is ok.

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

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

Suppose $\mathrm{rank}(W) = d$, *and* $\nabla_W \widehat{\mathcal{L}}(W) = 0$. *Then,* $\widehat{\mathcal{L}}(W) = 0$. *If, furthermore, number* N *of data is* $N \geq d(d+1)/2$ *then* $W = QW^*$ *for an orthogonal matrix* $Q \in \mathbb{R}^{m\times m}$.

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

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

Suppose $\mathrm{rank}(W) = d$, *and* $\nabla_W \widehat{\mathcal{L}}(W) = 0$. *Then,* $\widehat{\mathcal{L}}(W) = 0$. *If, furthermore, number* N *of data is* $N \geq d(d+1)/2$ *then* $W = QW^*$ *for an orthogonal matrix* $Q \in \mathbb{R}^{m \times m}$.

(a) An analogue holds also for the population risk.

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

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

Suppose $\mathrm{rank}(W) = d$, *and* $\nabla_W \widehat{\mathcal{L}}(W) = 0$. *Then,* $\widehat{\mathcal{L}}(W) = 0$. *If, furthermore, number* N *of data is* $N \geq d(d+1)/2$ *then* $W = QW^*$ *for an orthogonal matrix* $Q \in \mathbb{R}^{m \times m}$.

(a) An analogue holds also for the population risk.

(b) Corresponding losses $\widehat{\mathcal{L}}(\cdot)$ and $\mathcal{L}(\cdot)$ admit no full-rank saddle points.

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

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

Suppose $\mathrm{rank}(W) = d$, *and* $\nabla_W \widehat{\mathcal{L}}(W) = 0$. *Then,* $\widehat{\mathcal{L}}(W) = 0$. *If, furthermore, number* N *of data is* $N \geq d(d+1)/2$ *then* $W = QW^*$ *for an orthogonal matrix* $Q \in \mathbb{R}^{m \times m}$.

(a) An analogue holds also for the population risk.

(b) Corresponding losses $\widehat{\mathcal{L}}(\cdot)$ and $\mathcal{L}(\cdot)$ admit no full-rank saddle points.

(c) Landscape of $(\widehat{\mathcal{L}}(\cdot)$ and $\mathcal{L}(\cdot))$ has fairly benign properties below the aforementioned energy barrier (**recall:** below the barrier no rank-deficient $W \in \mathbb{R}^{m \times d}$, w.h.p.).

Optimization Landscape: Global Optimality of Full-Rank Stationary Points

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

Suppose $\mathrm{rank}(W) = d$*, and* $\nabla_W \widehat{\mathcal{L}}(W) = 0$*. Then,* $\widehat{\mathcal{L}}(W) = 0$*. If, furthermore, number* N *of data is* $N \geq d(d+1)/2$ *then* $W = QW^*$ *for an orthogonal matrix* $Q \in \mathbb{R}^{m \times m}$*.*

(a) An analogue holds also for the population risk.

(b) Corresponding losses $\widehat{\mathcal{L}}(\cdot)$ and $\mathcal{L}(\cdot)$ admit no full-rank saddle points.

(c) Landscape of ($\widehat{\mathcal{L}}(\cdot)$ and $\mathcal{L}(\cdot)$) has fairly benign properties below the aforementioned energy barrier (**recall:** below the barrier no rank-deficient $W \in \mathbb{R}^{m \times d}$, w.h.p.).

(d) **Next.** Benign landscape $\implies$ Convergence of gradient descent.

Optimization Landscape: Convergence of Gradient Descent

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 gradient descent (with a certain step size) generates a full-rank,* $\epsilon-$*approximate stationary point* $W \in \mathbb{R}^{m\times d}$ *(namely* $\|\nabla\widehat{\mathcal{L}}(W)\|_F \le \epsilon$*) 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.*

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 gradient descent (with a certain step size) generates a full-rank,* $\epsilon-$*approximate stationary point* $W \in \mathbb{R}^{m\times d}$ *(namely* $\|\nabla\widehat{\mathcal{L}}(W)\|_F \le \epsilon$*) 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.*

(a) If the initialization is "nice", GD finds an approx. stat. point W in polynomial time.

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 gradient descent (with a certain step size) generates a full-rank,* $\epsilon-$*approximate stationary point* $W \in \mathbb{R}^{m\times d}$ *(namely* $\|\nabla\widehat{\mathcal{L}}(W)\|_F \le \epsilon$*) 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.*

(a) If the initialization is "nice", GD finds an approx. stat. point W in polynomial time.

(b) For this W, the weights W^TW are (uniformly) close to planted weights $(W^*)^TW^*$, thus has well-controlled generalization error.

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 gradient descent (with a certain step size) generates a full-rank, $\epsilon-$approximate stationary point $W \in \mathbb{R}^{m\times d}$ (namely $\|\nabla\widehat{\mathcal{L}}(W)\|_F \le \epsilon$) 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.*

(a) If the initialization is "nice", GD finds an approx. stat. point W in polynomial time.

(b) For this W, the weights W^TW are (uniformly) close to planted weights $(W^*)^TW^*$, thus has well-controlled generalization error.

(c) From a technical point: control the condition number of a certain matrix with i.i.d. rows consisting of tensorized data $X_i^{\otimes 2}$. Uses results from a very recent work of us analyzing the spectrum of expected covariance matrices of tensorized data.

Remarks.

Remarks.

(a) An energy barrier separating rank-deficient points. Below this barrier, only full-rank points survive.

Remarks.

(a) An energy barrier separating rank-deficient points. Below this barrier, only full-rank points survive.

(b) Any full-rank stationary point of the risk is globally optimal. Thus, the landscape below the barrier is actually benign. No spurios stationary points within the set of full-rank matrices.

Remarks.

(a) An energy barrier separating rank-deficient points. Below this barrier, only full-rank points survive.

(b) Any full-rank stationary point of the risk is globally optimal. Thus, the landscape below the barrier is actually benign. No spurios stationary points within the set of full-rank matrices.

(c) Gradient descent, provided **initialized properly**, “approximately” minimizes the empirical risk, and recovers the planted weights in polynomial time.

Remarks.

(a) An energy barrier separating rank-deficient points. Below this barrier, only full-rank points survive.

(b) Any full-rank stationary point of the risk is globally optimal. Thus, the landscape below the barrier is actually benign. No spurios stationary points within the set of full-rank matrices.

(c) Gradient descent, provided **initialized properly**, "approximately" minimizes the empirical risk, and recovers the planted weights in polynomial time.

(d) **On a technical level.** Covering number and concentration arguments. The proximity above is established using a novel concentration result of us for matrices with i.i.d. rows obtained from tensorized data $X_i^{\otimes 2}$.

Remarks.

(a) An energy barrier separating rank-deficient points. Below this barrier, only full-rank points survive.

(b) Any full-rank stationary point of the risk is globally optimal. Thus, the landscape below the barrier is actually benign. No spurios stationary points within the set of full-rank matrices.

(c) Gradient descent, provided **initialized properly**, "approximately" minimizes the empirical risk, and recovers the planted weights in polynomial time.

(d) **On a technical level.** Covering number and concentration arguments. The proximity above is established using a novel concentration result of us for matrices with i.i.d. rows obtained from tensorized data $X_i^{\otimes 2}$.

(e) **Next.** *"How to initialize properly?"*

Overview

Proper Initialization

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).
- Well approximate dynamical systems (Gonon et al. [20])

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).
- Well approximate dynamical systems (Gonon et al. [20])
- Also studied in the context of extreme learning machine (Huang et al. [06]), and in random matrix theory (Pennington & Worah [17]).

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).
- Well approximate dynamical systems (Gonon et al. [20])
- Also studied in the context of extreme learning machine (Huang et al. [06]), and in random matrix theory (Pennington & Worah [17]).

(c) **Intuition.**

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).
- Well approximate dynamical systems (Gonon et al. [20])
- Also studied in the context of extreme learning machine (Huang et al. [06]), and in random matrix theory (Pennington & Worah [17]).

(c) **Intuition.**

- Value of loss function is determined by spectrum of $W^T W - (W^*)^T W^*$ and data moments.

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).
- Well approximate dynamical systems (Gonon et al. [20])
- Also studied in the context of extreme learning machine (Huang et al. [06]), and in random matrix theory (Pennington & Worah [17]).

(c) **Intuition.**

- Value of loss function is determined by spectrum of $W^T W - (W^*)^T W^*$ and data moments.
- Spectrum of Wishart matrices $(W^*)^T W^*$ are tightly concentrated. Semicircle law by Bai & Yin [88,93].

Proper Initialization

(a) **Recall.** GD is successful provided initialized properly.

(b) **Focus.** Initialization in the context of randomly generated planted $W^* \in \mathbb{R}^{m\times d}$:

- Networks with random weights is an active research area, define initial loss landscape.
- Closely related to random feature methods (Rahimi & Recht [09]).
- Well approximate dynamical systems (Gonon et al. [20])
- Also studied in the context of extreme learning machine (Huang et al. [06]), and in random matrix theory (Pennington & Worah [17]).

(c) **Intuition.**

- Value of loss function is determined by spectrum of $W^T W - (W^*)^T W^*$ and data moments.
- Spectrum of Wishart matrices $(W^*)^T W^*$ are tightly concentrated. Semicircle law by Bai & Yin [88,93].
- Thus the spectrum of $W^T W - (W^*)^T W^*$ can be controlled by choosing W appropriately.

Proper Initialization. Main Result.

Proper Initialization. Main Result.

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

Suppose $W^* \in \mathbb{R}^{m\times d}$ *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, with high probability,*

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4,$$

provided $m > Cd^2$*, for a large constant* $C > 0$.

Proper Initialization. Main Result.

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

Suppose $W^* \in \mathbb{R}^{m \times d}$ 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, with high probability,

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4,$$

provided $m > Cd^2$, for a large constant $C > 0$.

(a) A deterministic initialization beats the energy barrier, provided the network is sufficiently overparametrized ($m = \Omega(d^2)$). Based on the semicircle law.

Proper Initialization. Main Result.

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

Suppose $W^* \in \mathbb{R}^{m\times d}$ 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, with high probability,

$$\widehat{\mathcal{L}}(W_0) < \frac{1}{2} C_5 \sigma_{\min}(W^*)^4,$$

provided $m > Cd^2$, for a large constant $C > 0$.

(a) A deterministic initialization beats the energy barrier, provided the network is sufficiently overparametrized ($m = \Omega(d^2)$). Based on the semicircle law.

(b) An analogous result for the population risk. In the case of W^* having i.i.d. standard normal entries, non-asymptotic guarantees are available.

Overview

Sample Complexity

Sample Complexity

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

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)

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

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)

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

- *Suppose* $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$, *and* $\widehat{m} \in \mathbb{N}$ *arbitrary. Then, for any* $W \in \mathbb{R}^{\widehat{m}\times d}$ *"interpolating" the data* ($f(W;X_i) = f(W^*;X_i)$, $1 \le i \le N$), $W^TW = (W^*)^TW^*$. *Thus,* W *generalizes well:* $\mathcal{L}(W) = 0$.

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)

Let $X_i \in \mathbb{R}^d$, $1 \le i \le N$ *be the 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), $W^T W \neq (W^*)^T W^*$. *In particular,* $\mathcal{L}(W) > 0$ *(where* $\mathcal{L}$ *is defined w.r.t. any jointly continuous distribution on* $\mathbb{R}^d$*).*

Sample Complexity

Sample Complexity

Remarks.

Sample Complexity

Remarks.

(a) If $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ has necessarily zero generalization error.

Sample Complexity

Remarks.

(a) If $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ has necessarily zero generalization error.

(b) The condition is not retrospective in manner: the condition $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$ can be checked ahead of optimization task.

Sample Complexity

Remarks.

(a) If $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ has necessarily zero generalization error.

(b) The condition is not retrospective in manner: the condition $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$ can be checked ahead of optimization task.

(c) No randomness. A purely geometric, necessary and sufficient geometric condition.

Sample Complexity

Remarks.

(a) If $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ has necessarily zero generalization error.

(b) The condition is not retrospective in manner: the condition $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$ can be checked ahead of optimization task.

(c) No randomness. A purely geometric, necessary and sufficient geometric condition.

(d) In case W has non-zero but small training loss $(\widehat{\mathcal{L}}(W))$, earlier results allow bounding $\|W^T W - (W^*)^T W^*\|_F$, and $\mathcal{L}(W)$.

Sample Complexity

Remarks.

(a) If $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ has necessarily zero generalization error.

(b) The condition is not retrospective in manner: the condition $\mathrm{span}(X_i X_i^T : 1 \le i \le N) = \mathcal{S}$ can be checked ahead of optimization task.

(c) No randomness. A purely geometric, necessary and sufficient geometric condition.

(d) In case W has non-zero but small training loss ($\widehat{\mathcal{L}}(W)$), earlier results allow bounding $\|W^T W - (W^*)^T W^*\|_F$, and $\mathcal{L}(W)$.

(e) The parameter $\widehat{m} \in \mathbb{N}$: the interpolating network need not have the same width m. Provided the span condition holds, **any** interpolant, albeit being over-parametrized, generalize well (i.e. in this case, overparametrization does **not** hurt generalization).

Sample Complexity

Remarks.

(a) If $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$, then any minimizer W of the empirical risk $\widehat{\mathcal{L}}(\cdot)$ has necessarily zero generalization error.

(b) The condition is not retrospective in manner: the condition $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$ can be checked ahead of optimization task.

(c) No randomness. A purely geometric, necessary and sufficient geometric condition.

(d) In case W has non-zero but small training loss ($\widehat{\mathcal{L}}(W)$), earlier results allow bounding $\|W^TW - (W^*)^TW^*\|_F$, and $\mathcal{L}(W)$.

(e) The parameter $\widehat{m} \in \mathbb{N}$: the interpolating network need not have the same width m. Provided the span condition holds, **any** interpolant, albeit being over-parametrized, generalize well (i.e. in this case, overparametrization does **not** hurt generalization).

(f) **Theorem.** As soon as $N \ge d(d+1)/2$, random data $X_i \in \mathbb{R}^d$, $1 \le i \le N$ enjoys $\mathrm{span}(X_iX_i^T : 1 \le i \le N) = \mathcal{S}$, with probability one.

Sample Complexity Bound for Planted Network.

Sample Complexity Bound for Planted Network.

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

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

- *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 $1 \le i \le N$),*

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

Sample Complexity Bound for Planted Network.

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

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

- *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* $1 \le i \le N$*),*

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

The lower bound in second part coincides with the (earlier) energy barrier.

Thank you!