An Initial Alignment between Neural Network and Target is Needed for Gradient Descent to Learn

E. Abbé, E. Cornacchia, J. Hązła, C. Marquis

École Polytechnique Fédérale de Lausanne (EPFL)

ICML 2022

Initial Alignment (INAL)

Is a certain amount of "initial alignment" needed between a neural network at initialization and a target function in order for (S)GD to learn?

Initial Alignment (INAL)

Is a certain amount of "initial alignment" needed between a neural network at initialization and a target function in order for (S)GD to learn?

For a target function $f : \mathcal{X} \to \mathcal{Y}$, input distribution $P_{\mathcal{X}}$ and a neural network $\mathrm{NN}_\theta : \mathcal{X} \to \mathcal{Y}$ randomly initialized with $\theta^0 \sim P_0$

$$\mathrm{INAL}(f, \mathrm{NN}) := \max_{v \in \text{neurons}} \mathbb{E}_{\theta^0 \sim P_0} \mathbb{E}_{x \sim P_{\mathcal{X}}} [f(x) \cdot \mathrm{NN}_{\theta^0}^{(v)}(x)]^2,$$

Initial Alignment (INAL)

Is a certain amount of "initial alignment" needed between a neural network at initialization and a target function in order for (S)GD to learn?

For a target function $f : \mathcal{X} \to \mathcal{Y}$, input distribution $P_{\mathcal{X}}$ and a neural network $\mathrm{NN}_\theta : \mathcal{X} \to \mathcal{Y}$ randomly initialized with $\theta^0 \sim P_0$

$$\mathrm{INAL}(f, \mathrm{NN}) := \max_{v \in \text{neurons}} \mathbb{E}_{\theta^0 \sim P_0} \mathbb{E}_{x \sim P_{\mathcal{X}}} [f(x) \cdot \mathrm{NN}_{\theta^0}^{(v)}(x)]^2,$$

Question: Does small $\mathrm{INAL}(f, \mathrm{NN})$ imply that after T steps of GD, $|\mathbb{E} f(x) \, \mathrm{NN}^{(T)}(x)|$ is small (for a reasonable T)?

Setting

Data:

- Boolean target function: $f_n : \{\pm 1\}^n \to \{\pm 1\}$
- Uniform input distribution: $P_{\mathcal{X}} = \mathrm{Unif}\left(\{\pm 1\}^n\right)$
- Assume f_n asymptotically balanced: $\mathbb{P}_X\left(f_n(X) = 1\right) = 1/2 + o_n(1)$

Setting

Data:

- Boolean target function: $f_n : \{\pm 1\}^n \to \{\pm 1\}$
- Uniform input distribution: $P_{\mathcal{X}} = \text{Unif}(\{\pm 1\}^n)$
- Assume f_n asymptotically balanced: $\mathbb{P}_X(f_n(X) = 1) = 1/2 + o_n(1)$

Architecture/algorithm:

- Fully connected neural networks of poly(n) size with iid gaussian initialization with rescaled variance and ReLU activation
- Noisy GD with full batch and gradient precision A [Abbe and Sandon,'20, Abbe et al.,'21]

$$\theta^t = \theta^{t-1} - \gamma_t \mathbb{E}_{x \sim P_{\mathcal{X}}} \left[\nabla_\theta L(\text{NN}_{\theta^{t-1}}(x), f(x)) \right]_A + Z^{(t)},$$

where $Z^{(t)} \stackrel{iid}{\sim} \mathcal{N}(0, \mathbb{I}\sigma^2)$

Main Result

'Extended' function: $\bar{f}_n(x_1, ..., x_n, x_{n+1}, ..., x_{n^2}) = f_n(x_1, ..., x_n)$

Main Result

'Extended' function: $\bar{f}_n(x_1, ..., x_n, x_{n+1}, ..., x_{n^2}) = f_n(x_1, ..., x_n)$

Theorem 1

If $\mathrm{INAL}(f_n, \mathrm{NN}_n) = O(n^{-c})$, *for* $c \geq 1$, *then the noisy GD algorithm after* T *steps of training on* ***any*** *fully connected network of size* E *and* ***any*** *iid initialization, outputs a network* $\mathrm{NN}_n^{(T)}$ *such that*

$$|\mathbb{E}[\bar{f}_n(x) \cdot \mathrm{NN}_n^{(T)}(x)]| = O\left(\frac{\gamma T \sqrt{E} A}{\sigma} \cdot n^{-\frac{c-1}{8}}\right)$$

Main Result

'Extended' function: $\bar{f}_n(x_1, ..., x_n, x_{n+1}, ..., x_{n^2}) = f_n(x_1, ..., x_n)$

Theorem 1

If $\mathrm{INAL}(f_n, \mathrm{NN}_n) = O(n^{-c})$*, for* $c \geq 1$*, then the noisy GD algorithm after* T *steps of training on* ***any*** *fully connected network of size* E *and* ***any*** *iid initialization, outputs a network* $\mathrm{NN}_n^{(T)}$ *such that*

$$|\mathbb{E}[\bar{f}_n(x) \cdot \mathrm{NN}_n^{(T)}(x)]| = O\left(\frac{\gamma T \sqrt{E} A}{\sigma} \cdot n^{-\frac{c-1}{8}}\right)$$

- INAL characterizes if f_n is *weakly* learnable on Gaussian ReLU networks

Main Result

'Extended' function: $\bar{f}_n(x_1, ..., x_n, x_{n+1}, ..., x_{n^2}) = f_n(x_1, ..., x_n)$

Theorem 1

If $\mathrm{INAL}(f_n, \mathrm{NN}_n) = O(n^{-c})$*, for* $c \geq 1$*, then the noisy GD algorithm after* T *steps of training on* ***any*** *fully connected network of size* E *and* ***any*** *iid initialization, outputs a network* $\mathrm{NN}_n^{(T)}$ *such that*

$$|\mathbb{E}[\bar{f}_n(x) \cdot \mathrm{NN}_n^{(T)}(x)]| = O\left(\frac{\gamma T \sqrt{E} A}{\sigma} \cdot n^{-\frac{c-1}{8}}\right)$$

- INAL characterizes if f_n is *weakly* learnable on Gaussian ReLU networks
- Hardness for *any* iid initialization, activation

Main Result

'Extended' function: $\bar{f}_n(x_1, ..., x_n, x_{n+1}, ..., x_{n^2}) = f_n(x_1, ..., x_n)$

Theorem 1

If $\mathrm{INAL}(f_n, \mathrm{NN}_n) = O(n^{-c})$*, for* $c \geq 1$*, then the noisy GD algorithm after* T *steps of training on* ***any*** *fully connected network of size* E *and* ***any*** *iid initialization, outputs a network* $\mathrm{NN}_n^{(T)}$ *such that*

$$|\mathbb{E}[\bar{f}_n(x) \cdot \mathrm{NN}_n^{(T)}(x)]| = O\left(\frac{\gamma T \sqrt{E} A}{\sigma} \cdot n^{-\frac{c-1}{8}}\right)$$

- INAL characterizes if f_n is *weakly* learnable on Gaussian ReLU networks
- Hardness for *any* iid initialization, activation
- We obtain hardness only for the 'extension' of f_n

Proof Outline

Fourier-Walsh transform of f:

$$f_n(x) = \sum_{S \in [n]} \hat{f}_n(S)\chi_S(x), \quad \chi_S(x) := \prod_{i \in S} x_i, \quad \hat{f}_n(S) := \mathbb{E}_x[f_n(x)\chi_S(x)]$$

- **Step 1:** *If* $\mathrm{INAL}(f_n, \mathrm{NN}_n)$ *is small,* f_n *is high-degree.*

  Specifically: $\underbrace{W^{\leq k}[f_n]}_{\sum_{S:|S|\leq k} \hat{f}_n(S)^2} \leq O\left(n^{k+1}\right) \cdot \mathrm{INAL}(f_n, \mathrm{NN}_n)$, for any k.

Proof Outline

Fourier-Walsh transform of f:

$$f_n(x) = \sum_{S \in [n]} \hat{f}_n(S)\chi_S(x), \quad \chi_S(x) := \prod_{i \in S} x_i, \quad \hat{f}_n(S) := \mathbb{E}_x[f_n(x)\chi_S(x)]$$

- **Step 1:** *If* $\mathrm{INAL}(f_n, \mathrm{NN}_n)$ *is small,* f_n *is high-degree.*

  Specifically: $\underbrace{W^{\leq k}[f_n]}_{\sum_{S:|S|\leq k} \hat{f}_n(S)^2} \leq O\left(n^{k+1}\right) \cdot \mathrm{INAL}(f_n, \mathrm{NN}_n)$, for any k.

- **Step 2:** *High-degree functions are hard to learn for noisy GD on fully connected neural networks.*

Proof Outline

Fourier-Walsh transform of f:

$$f_n(x) = \sum_{S \in [n]} \hat{f}_n(S) \chi_S(x), \quad \chi_S(x) := \prod_{i \in S} x_i, \quad \hat{f}_n(S) := \mathbb{E}_x[f_n(x) \chi_S(x)]$$

- **Step 1:** *If* $\mathrm{INAL}(f_n, \mathrm{NN}_n)$ *is small,* f_n *is high-degree.*

  Specifically: $\underbrace{W^{\leq k}[f_n]}_{\sum_{S:|S| \leq k} \hat{f}_n(S)^2} \leq O\left(n^{k+1}\right) \cdot \mathrm{INAL}(f_n, \mathrm{NN}_n)$, for any k.

- **Step 2:** *High-degree functions are hard to learn for noisy GD on fully connected neural networks.*

Thank you.