Towards Understanding the Importance of Noise in Training Neural Networks

Mo Zhou[2], Tianyi Liu[1], Yan Li[1], Dachao Lin[2], Enlu Zhou[1], Tuo Zhao[1]
[1] Georgia Institute of Technology [2] Peking University

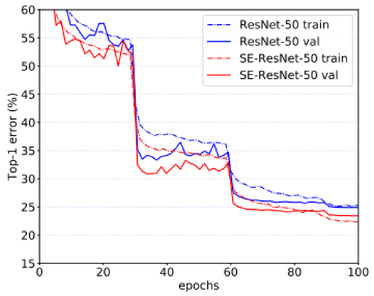
Background

Challenges of Deep Neural Networks (DNNs):

- Nonconvex: Saddlle Points, Spurious Optima;
- Computationally intractable;
- Serious overfitting and curse of dimensionality.

Success of First Order Methods:

- Existing Results:
  · Escape strict saddle and converge to optima:
  - Gradient Descent (GD): Lee et al., 2016, etc.;
  - Stochastic Gradient Descent (SGD): Dauphin et al., 2014, etc.
- Practitioners' Choice: Step Size Annealing (SSA)



- Our Empirical Observations:

	Generalization	Optima	Noise Level
GD	Bad	Sharp	No
SGD w./ very small Step Size	Bad	Sharp	Very Small
SGD w./ SSA	Good	Flat	Stagewise Decreasing

–Not all local optima generalize well.
–Noise can help select good optima .

Question: How does noise help train neural networks in the presence of bad optima?

Challenges

- Beyond Technical Limit:
  – General NNs: Complex Nonconvex Landscape;
  – SGD: Noise Structure: Distribution and Dependency.
- We Study:
  – Two-Layer Nonoverlapping Convolutional NNs (CNNs):
    1. A non-trivial spurious local optimum;
    2. GD gets trapped with constant probability ($\frac{1}{4} \sim \frac{3}{4}$);
  – Perturbed Gradient Descent with Noise Annealing:
    1. Independent injected Uniform noise;
    2. Imitate the behavior of SGD.

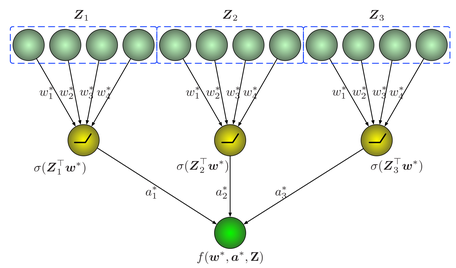
A non-trival example provides new insights!

Two-layer Nonoverlapping CNNs

- Teacher Network Model:

$$f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) = \sum_{j=1}^{k} a_j^* \sigma(\boldsymbol{Z}_j^\top \boldsymbol{w}^*),$$

  – $||\boldsymbol{w}^*||_2 = 1$, $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{a} \in \mathbb{R}^k$,
  – $\mathbf{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_k]$ with $\boldsymbol{Z}_j$'s i.i.d. $N(\mathbf{0}, \mathbf{I})$,
  – $\sigma(\cdot) = \max\{\cdot, 0\}$.



- Nonconvex Optimization:

$$(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{a}}) = \operatorname*{argmin}_{\boldsymbol{w}, \boldsymbol{a}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) \quad \text{subject to} \quad ||\boldsymbol{w}||_2 = 1,$$

where $\mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) = \mathbb{E}_{\mathbf{Z}}(f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) - f(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.
- A nontrivial **spurious local optimum** exists! (Du et al., 2017)

$$\overline{\boldsymbol{w}} = -\boldsymbol{w}^*, \ \overline{\boldsymbol{a}} = (\mathbf{1}\mathbf{1}^\top + (\pi - 1)\mathbf{I})^{-1}(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\boldsymbol{a}^*.$$

Perturbed Gradient Descent (P-GD)

- Initialization: $a_0 \in \mathbb{B}_0\left(\frac{|\mathbf{1}^\top a^*|}{\sqrt{k}}\right)$ and $w_0 \in \mathbb{S}_0(1)$.
- At the t-th iteration, we independently sample

$$\boldsymbol{\epsilon}_t \sim \mathrm{Unif}(\mathbb{B}^p(\rho_{\boldsymbol{w}})), \quad \boldsymbol{\xi}_t \sim \mathrm{Unif}(\mathbb{B}^k(\rho_{\boldsymbol{a}})),$$

and further take

$$\widetilde{\boldsymbol{w}}_t = \boldsymbol{w}_t + \boldsymbol{\xi}_t, \quad \widetilde{\boldsymbol{a}}_t = \boldsymbol{a}_t + \boldsymbol{\epsilon}_t.$$

- We then update $\boldsymbol{w}$ and $\boldsymbol{a}$ by

$$\boldsymbol{a}_{t+1} = \boldsymbol{a}_t - \eta \nabla_{\boldsymbol{a}} \mathcal{L}_t(\widetilde{\boldsymbol{w}}_t, \widetilde{\boldsymbol{a}}_t, \mathbf{Z}^{(t)}),$$
$$\boldsymbol{w}_{t+1} = \Pi_{\mathbb{S}(1)}(\boldsymbol{w}_t - \eta(\mathbf{I} - \boldsymbol{w}_t \boldsymbol{w}_t^\top)\nabla_{\boldsymbol{w}} \mathcal{L}_t(\widetilde{\boldsymbol{w}}_t, \widetilde{\boldsymbol{a}}_t, \mathbf{Z}^{(t)})),$$

where $\mathcal{L}(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}) = \mathbb{E}_{\mathbf{Z}}(f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) - f(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.

Noise Annealing:
Noise level schedule: $\{\rho_{\boldsymbol{w}}^{(s)}\}_{s=1}^S$ and $\{\rho_{\boldsymbol{a}}^{(s)}\}_{s=1}^S$
- Multi-Epoch:

$$s\text{-th initialization} \Longleftarrow (s-1)\text{-th output}.$$

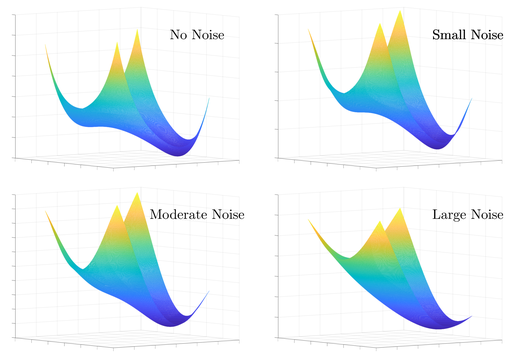
- Noise Annealing:

$$\rho_{\boldsymbol{w}}^{(s)} < \rho_{\boldsymbol{w}}^{(s-1)} \text{ and } \rho_{\boldsymbol{a}}^{(s)} < \rho_{\boldsymbol{a}}^{(s-1)}$$

Convolutional Effects

- P-GD essentially solves:

$$\min_{\boldsymbol{w}, \boldsymbol{a}} \ \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\xi}} \mathcal{L}(\boldsymbol{w} + \boldsymbol{\epsilon}, \boldsymbol{a} + \boldsymbol{\xi}) \quad \text{subject to} \quad ||\boldsymbol{w}||_2 = 1.$$



- Partial Dissipativity Condition: Let $\widetilde{\boldsymbol{w}} = \boldsymbol{w} + \boldsymbol{\epsilon}$ and $\widetilde{\boldsymbol{a}} = \boldsymbol{a} + \boldsymbol{\xi}$. For $(\boldsymbol{w}, \boldsymbol{a}) \in \mathcal{U} \subseteq \mathbb{S}^p(1) \times \mathbb{B}^k(R)$, we have:

$$\mathrm{C1}: \langle -\mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\epsilon}}(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top)\nabla_{\boldsymbol{w}} \mathcal{L}(\widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{a}}), \boldsymbol{w}^* - \boldsymbol{w} \rangle \geq c_{\boldsymbol{w}} ||\boldsymbol{w} - \boldsymbol{w}^*||_2^2 - \gamma_{\boldsymbol{w}},$$

$$\mathrm{C2}: \ \langle -\mathbb{E}_{\boldsymbol{\xi}, \boldsymbol{\epsilon}} \nabla_{\boldsymbol{a}} \mathcal{L}(\widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{a}}), \boldsymbol{a}^* - \boldsymbol{a} \rangle \geq c_{\boldsymbol{a}} ||\boldsymbol{a} - \boldsymbol{a}^*||_2^2 - \gamma_{\boldsymbol{a}}.$$

- Technical Challenges:
  – C1 and C2 do NOT globally hold;
  – C1 and C2 do NOT necessarily hold at the same time;
  – C1 and C2 vary as the noise levels vary.

Epoch I: Escaping Spurious Local Optima

Partial Dissipativity Condition: With **large** noise,

- C2 holds and C1 does not hold around the initialization. ⇒ P-GD reduces the optimization error of a.
- Reducing error of a. ⇒ C1 holds. ⇒ P-GD improves w.
- The output solution is far away from (w^*, a^*).

Theorem 1. *Suppose $\rho_w^0 = C_w^0 kp^2 \geq 1$ and $\rho_a^0 = C_a^0$. For any $\delta \in (0, 1)$, we choose step size*

$$\eta = O\left(\left(k^4 p^6 \cdot \max\left\{1, p \log \frac{1}{\delta}\right\}\right)^{-1}\right).$$

Then with probability at least $1 - \delta$, we have

$$0 < m_a \leq a_t^\top a^* \leq M_a \quad \text{and} \quad \angle(w_t, w^*) \leq \frac{5}{12}\pi$$

for all $T_1 \leq t \leq O(\eta^{-2})$, where m_a, M_a are some constants, and

$$T_1 = O\Big(pk/\eta \log(1/\eta) \log(1/\delta) ||a^*||_2^2\Big).$$

Epoch II: Converging to Global Optima

Partial Dissipativity Condition: With **small** noise,

- C1 and C2 jointly hold;
- $\gamma_{\boldsymbol{w}} = 0$ and $\gamma_{\boldsymbol{a}}$ decreases.

Theorem 2. *For any $\gamma > 0$, we choose $\rho_w^1 \leq C_w^1 \frac{\gamma}{kp} < 1$ and $\rho_a^1 \leq C_a^1$ for some constants C_w^1 and C_a^1. For any $\delta \in (0, 1)$, we choose step size*

$$\eta = O\left(\left(\max\left\{k^4 p^6, \frac{k^2 p}{\gamma}\right\} \max\left\{1, p \log \frac{1}{\gamma} \log \frac{1}{\delta}\right\}\right)^{-1}\right).$$

Then with probability at least $1 - \delta$, we have

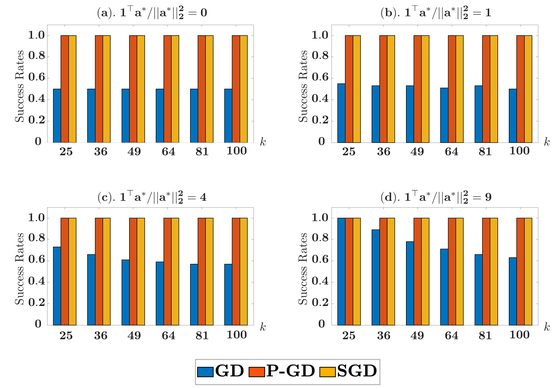
$$||w_t - w^*||_2^2 \leq \gamma \quad \text{and} \quad ||a_t - a^*||_2^2 \leq \gamma$$

for any t's such that $T_2 \leq t \leq T = O(\eta^{-2})$, where

$$T_2 = O\Big(p/\eta \log 1/\gamma \log 1/\delta ||\boldsymbol{a}^*||_2^2\Big).$$

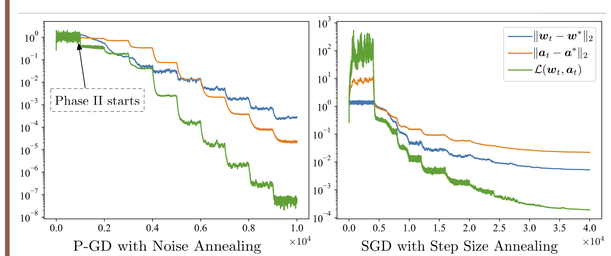
Experments

- Success Rates with $p = 6$ and Varing k and a^* :



  – P-GD, SGD 100% converge to the global optimum.
  – GD can get trapped with probability 50%.

- Empirical Convergence:



  – SGD shows **similar** patterns to P-GD.