Towards Understanding the Importance of Noise in Training Neural Networks

Mo Zhou‡, **Tianyi Liu†**, Yan Li†, Dachao Lin‡, Enlu Zhou† and Tuo Zhao†

†Georgia Tech and ‡Peking University

June. 12, 2019
Background: Deep Neural Networks

Great Success

- Speech and image recognition
- Nature language processing
- Recommendation systems

Training Challenges

- Highly nonconvex optimization landscape: Saddle Points, Spurious Optima
- Computationally intractable
- Serious overfitting and curse of dimensionality
Background: Deep Neural Networks

Great Success

- Speech and image recognition
- Nature language processing
- Recommendation systems

Training Challenges

- Highly nonconvex optimization landscape: Saddle Points, Spurious Optima
- Computationally intractable
- Serious overfitting and curse of dimensionality
Background: Deep Neural Networks

Great Success

- Speech and image recognition
- Nature language processing
- Recommendation systems

Training Challenges

- Highly nonconvex optimization landscape: Saddle Points, Spurious Optima
- Computationally intractable
- Serious overfitting and curse of dimensionality
Efficient Training by First Order Algorithms

**Existing Results:** Escape strict saddle and converge to optima:

- Gradient Descent (GD): Lee et al., 2016; Jin et al., 2017; Panageas et al., 2017; Lee et al., 2017;

- Stochastic Gradient Descent (SGD): Dauphin et al., 2014; Ge et al., 2015; Kawaguchi, 2016; Hardt and Ma, 2016; Jin et al., 2017; Jin et al., 2019.

*Still far from being well understood!*
Efficient Training by First Order Algorithms

Existing Results: Escape strict saddle and converge to optima:

- Gradient Descent (GD): Lee et al., 2016; Jin et al., 2017; Panageas et al., 2017; Lee et al., 2017;

- Stochastic Gradient Descent (SGD): Dauphin et al., 2014; Ge et al., 2015; Kawaguchi, 2016; Hardt and Ma, 2016; Jin et al., 2017; Jin et al., 2019.

Still far from being well understood!
Efficient Training by First Order Algorithms

Existing Results: Escape strict saddle and converge to optima:

- Gradient Descent (GD): Lee et al., 2016; Jin et al., 2017; Panageas et al., 2017; Lee et al., 2017;

- Stochastic Gradient Descent (SGD): Dauphin et al., 2014; Ge et al., 2015; Kawaguchi, 2016; Hardt and Ma, 2016; Jin et al., 2017; Jin et al., 2019.

Still far from being well understood!
Practitioners’ Choice: Step Size Annealing

Remark:

- The variance of the noise scales with the step size;
- Noise level: Large $\Rightarrow$ Small.
Practitioners’ Choice: Step Size Annealing

Remark:

- The variance of the noise scales with the step size;
- Noise level: Large $\Rightarrow$ Small.
## Our Empirical Observations

<table>
<thead>
<tr>
<th></th>
<th>Generalization</th>
<th>Optima</th>
<th>Noise Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>Bad</td>
<td>Sharp</td>
<td>No</td>
</tr>
<tr>
<td>SGD w./ very small Step Size</td>
<td>Bad</td>
<td>Sharp</td>
<td>Very Small</td>
</tr>
<tr>
<td>SGD w./ Step Size Annealing</td>
<td>Good</td>
<td>Flat</td>
<td>Stagewise Decreasing</td>
</tr>
</tbody>
</table>

### What We Know:

- Not all optima generalize;
- Noise helps select optima that generalize.
Our Empirical Observations

<table>
<thead>
<tr>
<th></th>
<th>Generalization</th>
<th>Optima</th>
<th>Noise Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>Bad</td>
<td>Sharp</td>
<td>No</td>
</tr>
<tr>
<td>SGD w./ very small</td>
<td>Bad</td>
<td>Sharp</td>
<td>Very Small</td>
</tr>
<tr>
<td>Step Size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGD w./ Step Size</td>
<td>Good</td>
<td>Flat</td>
<td>Stagewise Decreasing</td>
</tr>
<tr>
<td>Annealing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What We Know:

- Not all optima generalize;
- Noise helps select optima that generalize.
A Natural Question:

*How does noise help train neural networks in the presence of bad optima?*
Challenges

**General Neural Networks (NNs)**

- Complex nonconvex landscape;
- Beyond our technical limit.

**We Study:** Two-Layer Nonoverlapping Convolutional NNs:

- Non-trivial spurious local optimum (does not generalize);
- GD with random initialization gets trapped with constant probability (at least $\frac{1}{4}$, can be $\frac{3}{4}$ in the worst case);
- Simple structure that is technically manageable.
Challenges

General Neural Networks (NNs)

- Complex nonconvex landscape;
- Beyond our technical limit.

We Study: Two-Layer Nonoverlapping Convolutional NNs:

- Non-trivial spurious local optimum (does not generalize);
- GD with random initialization gets trapped with constant probability (at least $\frac{1}{4}$, can be $\frac{3}{4}$ in the worst case);
- Simple structure that is technically manageable.
Challenges

**General Neural Networks (NNs)**

- Complex nonconvex landscape;
- Beyond our technical limit.

**We Study:** Two-Layer Nonoverlapping Convolutional NNs:

- Non-trivial spurious local optimum (does not generalize);
  - GD with random initialization gets trapped with constant probability (at least \( \frac{1}{4} \), can be \( \frac{3}{4} \) in the worst case);
- Simple structure that is technically manageable.
Challenges

**General Neural Networks (NNs)**

- Complex nonconvex landscape;
- Beyond our technical limit.

**We Study:** Two-Layer Nonoverlapping Convolutional NNs:

- Non-trivial spurious local optimum (does not generalize);
- GD with random initialization gets trapped with constant probability (at least $\frac{1}{4}$, can be $\frac{3}{4}$ in the worst case);
- Simple structure that is technically manageable.
Challenges

General Neural Networks (NNs)

- Complex noncovex landscape;
- Beyond our technical limit.

We Study: Two-Layer Nonoverlapping Convolutional NNs:

- Non-trivial spurious local optimum (does not generalize);
- GD with random initialization gets trapped with constant probability (at least $\frac{1}{4}$, can be $\frac{3}{4}$ in the worst case);
- Simple structure that is technically manageable.
Challenges

**General Neural Networks (NNs)**

- Complex nonconvex landscape;
- Beyond our technical limit.

**We Study:** Two-Layer Nonoverlapping Convolutional NNs:

- Non-trivial spurious local optimum (does not generalize);
- GD with random initialization gets trapped with constant probability (at least $\frac{1}{4}$, can be $\frac{3}{4}$ in the worst case);
- Simple structure that is technically manageable.
Challenges

**Stochastic Gradient Descent**

- Complex distribution of noise;
- Dependency on iterates.

**We Study:** Perturbed Gradient Descent with Noise Annealing:

- Independent injected noise;
- Uniform distribution;
- Imitate the behavior of SGD.

A non-trivial example provides new insights!
Challenges

**Stochastic Gradient Descent**

- Complex distribution of noise;
- Dependency on iterates.

**We Study:** Perturbed Gradient Descent with Noise Annealing:

- Independent injected noise;
- Uniform distribution;
- Imitate the behavior of SGD.

A non-trivial example provides new insights!
Challenges

**Stochastic Gradient Descent**

- Complex distribution of noise;
- Dependency on iterates.

**We Study:** Perturbed Gradient Descent with Noise Annealing:

- Independent injected noise;
- Uniform distribution;
- Imitate the behavior of SGD.

A non-trivial example provides new insights!
Challenges

**Stochastic Gradient Descent**

- Complex distribution of noise;
- Dependency on iterates.

**We Study:** Perturbed Gradient Descent with Noise Annealing:

- Independent injected noise;
- Uniform distribution;
- Imitate the behavior of SGD.

A non-trivial example provides new insights!
Challengess

**Stochastic Gradient Descent**

- Complex distribution of noise;
- Dependency on iterates.

**We Study:** Perturbed Gradient Descent with Noise Annealing:

- Independent injected noise;
- Uniform distribution;
- Imitate the behavior of SGD.

A non-trival example provides new insights!
Two-layer Nonoverlapping CNNs

- Teacher Network Model:

\[ f(w^*, a^*, Z) = \sum_{j=1}^{k} a_j^* \sigma(Z_j^\top w^*) \text{ with } \|w^*\|_2 = 1, \]

where \( w \in \mathbb{R}^p, a \in \mathbb{R}^k, Z = [Z_1, ..., Z_k] \) with \( Z_j \)'s are independently sampled from \( N(0, I) \), and \( \sigma(\cdot) = \max\{\cdot, 0\} \).

- Nonconvex Optimization:

\[ \left( \hat{w}, \hat{a} \right) = \arg \min_{w, a} L(w, a) \text{ subject to } \|w\|_2 = 1, \]

where \( L(w, a) = \mathbb{E}_Z(f(w^*, a^*, Z) - f(w, a, Z))^2 \).

- A nontrivial spurious local optimum \( (\overline{w}, \overline{a}) \) exists!

\[ \overline{w} = -w^*, \overline{a} = (11^\top + (\pi - 1)I)^{-1}(11^\top - I)a^*. \]
Two-layer Nonoverlapping CNNs

- **Teacher Network Model:**
  
  \[
  f(w^*, a^*, Z) = \sum_{j=1}^{k} a^*_j \sigma(Z_j^\top w^*) \quad \text{with} \quad \|w^*\|_2 = 1,
  \]

  where \( w \in \mathbb{R}^p, a \in \mathbb{R}^k, Z = [Z_1, \ldots, Z_k] \) with \( Z_j \)'s are independently sampled from \( N(0, I) \), and \( \sigma(\cdot) = \max\{\cdot, 0\} \).

- **Nonconvex Optimization:**
  
  \[
  (\hat{w}, \hat{a}) = \arg \min_{w, a} \mathcal{L}(w, a) \quad \text{subject to} \quad \|w\|_2 = 1,
  \]

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

- A nontrivial spurious local optimum \((\overline{w}, \overline{a})\) exists!
  
  \[
  \overline{w} = -w^*, \quad \overline{a} = (11^\top + (\pi - 1)I)^{-1}(11^\top - I)a^*.
  \]
Two-layer Nonoverlapping CNNs

- **Teacher Network Model:**

\[
f(w^*, a^*, Z) = \sum_{j=1}^{k} a_j^* \sigma(Z_j^\top w^*) \quad \text{with} \quad \|w^*\|_2 = 1,
\]

where \( w \in \mathbb{R}^p, \ a \in \mathbb{R}^k, \ Z = [Z_1, ..., Z_k] \) with \( Z_j \)'s are independently sampled from \( N(0, I) \), and \( \sigma(\cdot) = \max\{\cdot, 0\} \).

- **Nonconvex Optimization:**

\[
(\hat{w}, \hat{a}) = \arg\min_{w, a} \mathcal{L}(w, a) \quad \text{subject to} \quad \|w\|_2 = 1,
\]

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

- A nontrivial spurious local optimum \((\overline{w}, \overline{a})\) exists!

\[
\overline{w} = -w^*, \quad \overline{a} = (11^\top + (\pi - 1)I)^{-1}(11^\top - I)a^*.
\]
Teacher Network Model
Perturbed Gradient Descent (P-GD)

Initialization: \( a_0 \in \mathbb{B}_0 \left( \frac{1^\top a^*}{\sqrt{k}} \right) \) and \( w_0 \in S_0(1) \).

At the \( t \)-th iteration, we independently sample

\[
\epsilon_t \sim \text{Unif}(\mathbb{B}^p(\rho_w)), \quad \xi_t \sim \text{Unif}(\mathbb{B}^k(\rho_a)),
\]

and \( Z^{(t)} = [Z_1^{(t)}, ..., Z_t^{(t)}] \) with \( Z_j^{(t)} \sim N(0, I) \), and further take

\[
\tilde{w}_t = w_t + \xi_t, \quad \tilde{a}_t = a^{(t)} + \epsilon_t.
\]

We then update \( w \) and \( a \) by

\[
\begin{align*}
a_{t+1} &= a^{(t)} - \eta \nabla_a \mathcal{L}_t(\tilde{w}_t, \tilde{a}_t, Z^{(t)}), \\
w_{t+1} &= \Pi_{S(1)}(w_t - \eta(I - w_t w_t^\top) \nabla_w \mathcal{L}_t(\tilde{w}_t, \tilde{a}_t, Z^{(t)})),
\end{align*}
\]

where \( \ell(w, a, Z) = \frac{1}{2} (f(w^*, a^*, Z) - f(w, a, Z))^2 \).
Perturbed Gradient Descent (P-GD)

Initialization: $a_0 \in \mathbb{B}_0 \left( \frac{1^\top a^*}{\sqrt{k}} \right)$ and $w_0 \in S_0(1)$.

At the $t$-th iteration, we independently sample

$$\epsilon_t \sim \text{Unif}(\mathbb{B}^p(\rho_w)), \quad \xi_t \sim \text{Unif}(\mathbb{B}^k(\rho_a)),$$

and $Z^{(t)} = [Z_1^{(t)}, ..., Z_t^{(t)}]$ with $Z_j^{(t)} \sim N(0, I)$, and further take

$$\tilde{w}_t = w_t + \xi_t, \quad \tilde{a}_t = a^{(t)} + \epsilon_t.$$

We then update $w$ and $a$ by

$$a_{t+1} = a^{(t)} - \eta \nabla_a \mathcal{L}_t(\tilde{w}_t, \tilde{a}_t, Z^{(t)}),$$
$$w_{t+1} = \Pi_{S(1)}(w_t - \eta(I - w_tw_t^\top)\nabla_w \mathcal{L}_t(\tilde{w}_t, \tilde{a}_t, Z^{(t)})),$$

where $\ell(w, a, Z) = \frac{1}{2} (f(w^*, a^*, Z) - f(w, a, Z))^2$. 
Perturbed Gradient Descent (P-GD)

Initialization: \(a_0 \in B_0 \left( \left| \mathbf{1}^\top a^* \right| / \sqrt{k} \right)\) and \(w_0 \in S_0(1)\).

At the \(t\)-th iteration, we independently sample \(\epsilon_t \sim \text{Unif}(B^p(\rho_w))\), \(\xi_t \sim \text{Unif}(B^k(\rho_a))\), and \(Z^{(t)} = [Z_1^{(t)}, \ldots, Z_t^{(t)}]\) with \(Z_j^{(t)} \sim \mathcal{N}(0, I)\), and further take

\[
\tilde{w}_t = w_t + \xi_t, \quad \tilde{a}_t = a^{(t)} + \epsilon_t.
\]

We then update \(w\) and \(a\) by

\[
a_{t+1} = a^{(t)} - \eta \nabla_a \mathcal{L}_t(\tilde{w}_t, \tilde{a}_t, Z^{(t)}),
\]

\[
w_{t+1} = \Pi_{S(1)}(w_t - \eta (I - w_t w_t^\top) \nabla w \mathcal{L}_t(\tilde{w}_t, \tilde{a}_t, Z^{(t)})),
\]

where \(\ell(w, a, Z) = \frac{1}{2} (f(w^*, a^*, Z) - f(w, a, Z))^2\).
Noise Annealing

Noise level schedule: $\{\rho_w^{(s)}\}_{s=1}^{S}$ and $\{\rho_a^{(s)}\}_{s=1}^{S}$

- Multi-Epoch: At the $s$-th epoch, we initialize using the output solution of the $(s - 1)$-th epoch.

- Then we apply P-GD with

$$
\epsilon_t \sim \text{Unif}(\mathbb{B}^p(\rho_w^{(s)})), \quad \xi_t \sim \text{Unif}(\mathbb{B}^k(\rho_a^{(s)})),
$$

where $\rho_w^{(s)} < \rho_w^{(s-1)}$ and $\rho_a^{(s)} < \rho_a^{(s-1)}$
Noise Annealing

Noise level schedule: \( \{\rho_w^{(s)}\}_{s=1}^S \) and \( \{\rho_a^{(s)}\}_{s=1}^S \)

- Multi-Epoch: At the \( s \)-th epoch, we initialize using the output solution of the \( (s - 1) \)-th epoch.

- Then we apply P-GD with

\[
\epsilon_t \sim \text{Unif}(\mathbb{B}^p(\rho_w^{(s)})), \quad \xi_t \sim \text{Unif}(\mathbb{B}^k(\rho_a^{(s)})),
\]

where \( \rho_w^{(s)} < \rho_w^{(s-1)} \) and \( \rho_a^{(s)} < \rho_a^{(s-1)} \)
Nonasymptotic Convergence Analysis

**Theorem (Informal)**

The theory considers two epochs of P-GD:

- **Epoch I.** With *large noise* and a properly chosen step size, **P-GD escapes the local optimum**, while approaching the basin of attraction of the global optimum in polynomial time with high probability;

- **Epoch II.** With *small noise* and a properly chosen step size, **P-GD converges to the global optimum in polynomial time with high probability.**

**“With High Probability”:** \( \mathbb{P}(A) \geq 1 - O(\exp(-1/\eta)) \).
Convolutional Effects

Remark: P-GD essentially solves:

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

For \((w, a) \in \mathcal{U} \subseteq S_p(1) \times B_k(R)\), we have:

\[\begin{align*}
C1 : \quad & -\mathbb{E}_{\xi, \epsilon}(I -ww^\top)\nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \geq c_w \|w - w^*\|^2 - \gamma_w, \\
C2 : \quad & -\mathbb{E}_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \geq c_a \|a - a^*\|^2 - \gamma_a,
\end{align*}\]

where \(\tilde{w} = w + \epsilon\) and \(\tilde{a} = a + \xi\).

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.
Partially Dissipative Conditions

For \((w, a) \in \mathcal{U} \subseteq \mathbb{S}_p(1) \times \mathbb{B}_k(R)\), we have:

\begin{align*}
C1 : \quad \langle -E_\xi, \epsilon (I - ww^\top) \nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \rangle &\geq c_\omega \|w - w^*\|_2^2 - \gamma_\omega, \\
C2 : \quad \langle -E_\xi, \epsilon \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle &\geq c_a \|a - a^*\|_2^2 - \gamma_a,
\end{align*}

where \(\tilde{w} = w + \epsilon\) and \(\tilde{a} = a + \xi\).

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.
Partially Dissipative Conditions

For \((w, a) \in \mathcal{U} \subseteq \mathbb{S}_p(1) \times \mathbb{B}_k(R)\), we have:

\[
\begin{align*}
\text{C1} & : \quad \langle -\mathbb{E}_\xi, \epsilon (I - ww^\top) \nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), \; w^* - w \rangle \geq c_w \|w - w^*\|^2 - \gamma_w, \\
\text{C2} & : \quad \langle -\mathbb{E}_\xi, \epsilon \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), \; a^* - a \rangle \geq c_a \|a - a^*\|^2 - \gamma_a,
\end{align*}
\]

where \(\tilde{w} = w + \epsilon\) and \(\tilde{a} = a + \xi\).

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.
Partially Dissipative Conditions

For \((w, a) \in U \subseteq S_p(1) \times B_k(R)\), we have:

\[ C1 : \quad \langle -E \xi, \epsilon (I - ww^\top) \nabla_w L(\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \| w - w^* \|_2^2 - \gamma_w, \]
\[ C2 : \quad \langle -E \xi, \epsilon \nabla_a L(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \| a - a^* \|_2^2 - \gamma_a, \]

where \(\tilde{w} = w + \epsilon\) and \(\tilde{a} = a + \xi\).

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 the Spurious Local Optimum

C1:  \[\langle -E_{\xi, \epsilon}(I - w w^\top) \nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \|w - w^*\|_2^2 - \gamma_w,\]

C2:  \[\langle -E_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \|a - a^*\|_2^2 - \gamma_a,\]

where \(\tilde{w} = w + \epsilon\) and \(\tilde{a} = a + \xi\).

With large noise,

- C2 holds and C1 does not hold around the initialization. \(\Rightarrow\) P-GD reduces the optimization error of \(a\).
- Reducing error of \(a\). \(\Rightarrow\) C1 holds. \(\Rightarrow\) P-GD improves \(w\).
- The output solution is far away from \((w^*, a^*)\).
Epoch I: Escaping the Spurious Local Optimum

C1: $\langle -\mathbb{E}_{\xi, \epsilon} (I - w w^\top) \nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \|w - w^*\|_2^2 - \gamma_w,$

C2: $\langle -\mathbb{E}_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \|a - a^*\|_2^2 - \gamma_a,$

where $\tilde{w} = w + \epsilon$ and $\tilde{a} = a + \xi.$

With large noise,

- C2 holds and C1 does not hold around the initialization. $\Rightarrow$ P-GD reduces the optimization error of $a.$
- Reducing error of $a. \Rightarrow$ C1 holds. $\Rightarrow$ P-GD improves $w.$
- The output solution is far away from $(w^*, a^*).$
Epoch I: Escaping the Spurious Local Optimum

\[ C1 : \langle -\mathbb{E}_{\xi, \epsilon} (I - \omega \omega^\top) \nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), \omega^* - \omega \rangle \geq c_w \| \omega - \omega^* \|_2^2 - \gamma_w, \]

\[ C2 : \langle -\mathbb{E}_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \| a - a^* \|_2^2 - \gamma_a, \]

where \( \tilde{w} = \omega + \epsilon \) and \( \tilde{a} = a + \xi \).

With large noise,

- C2 holds and C1 does not hold around the initialization. \( \Rightarrow \) P-GD reduces the optimization error of \( a \).
- Reducing error of \( a \). \( \Rightarrow \) C1 holds. \( \Rightarrow \) P-GD improves \( \omega \).
- The output solution is far away from \( (\omega^*, a^*) \).
Epoch I: Escaping the Spurious Local Optimum

C1: \[ \langle -\mathbb{E}_{\xi, \epsilon}(I - \omega w^\top)\nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \|w - w^*\|^2_2 - \gamma_w, \]

C2: \[ \langle -\mathbb{E}_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \|a - a^*\|^2_2 - \gamma_a, \]

where \( \tilde{w} = w + \epsilon \) and \( \tilde{a} = a + \xi \).

With large noise,

- C2 holds and C1 does not hold around the initialization. \( \Rightarrow \) P-GD reduces the optimization error of \( a \).
- Reducing error of \( a \). \( \Rightarrow \) C1 holds. \( \Rightarrow \) P-GD improves \( w \).
- The output solution is far away from \( (w^*, a^*) \).
Epoch I: Escaping the Spurious Local Optimum

**Theorem**

Suppose $\rho_0^w = C_0^w kp^2 \geq 1$ and $\rho_0^a = C_0^a$. 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

$$T_1 = O \left( \frac{pk}{\eta \log(1/\eta) \log(1/\delta) \|a^*\|_2^2} \right).$$

**Remark:** (1) is in the basin of attraction of the global optimum.
**Epoch I: Escaping the Spurious Local Optimum**

**Theorem**

Suppose $\rho_w^0 = C_w^0 k p^2 \geq 1$ 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

$$T_1 = O \left( \frac{pk}{\eta} \log(1/\eta) \log(1/\delta) \|a^*\|_2^2 \right).$$

**Remark:** (1) is in the basin of attraction of the global optimum.
Epoch II: Converging to the Global Optimum

\[ C_1 : \langle -\mathbb{E}_{\xi, \epsilon}(I - w w^\top) \nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \| w - w^* \|^2_2 - \gamma_w, \]

\[ C_2 : \langle -\mathbb{E}_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \| a - a^* \|^2_2 - \gamma_a, \]

where \( \tilde{w} = w + \epsilon \) and \( \tilde{a} = a + \xi \).

With small noise,

- C1 and C2 jointly hold;
- \( \gamma_w = 0 \) and \( \gamma_a \) decreases.
Epoch II: Converging to the Global Optimum

\[
\begin{align*}
C_1 : & \quad \langle -E_{\xi, \epsilon}(I - ww^\top)\nabla_w \mathcal{L}(\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \|w - w^*\|_2^2 - \gamma_w, \\
C_2 : & \quad \langle -E_{\xi, \epsilon} \nabla_a \mathcal{L}(\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \|a - a^*\|_2^2 - \gamma_a,
\end{align*}
\]

where \(\tilde{w} = w + \epsilon\) and \(\tilde{a} = a + \xi\).

With small noise,

\begin{itemize}
  \item C1 and C2 jointly hold;
  \item \(\gamma_w = 0\) and \(\gamma_a\) decreases.
\end{itemize}
Epoch II: Converging to the Global Optimum

\[ C_1 : \langle -\mathbb{E}_{\xi, \epsilon} (I - w w^\top) \nabla_w \mathcal{L} (\tilde{w}, \tilde{a}), w^* - w \rangle \geq c_w \| w - w^* \|^2_2 - \gamma_w, \]

\[ C_2 : \langle -\mathbb{E}_{\xi, \epsilon} \nabla_a \mathcal{L} (\tilde{w}, \tilde{a}), a^* - a \rangle \geq c_a \| a - a^* \|^2_2 - \gamma_a, \]

where \( \tilde{w} = w + \epsilon \) and \( \tilde{a} = a + \xi \).

With small noise,

- C1 and C2 jointly hold;
- \( \gamma_w = 0 \) and \( \gamma_a \) decreases.
Epoch II: Converging to the Global Optimum

C1: \[
\langle -\mathbb{E}_{\xi, \epsilon}(\mathbf{I} - \mathbf{w}\mathbf{w}^\top) \nabla_{\mathbf{w}} \mathcal{L}(\tilde{\mathbf{w}}, \tilde{a}), \mathbf{w}^* - \mathbf{w} \rangle \geq c_w \| \mathbf{w} - \mathbf{w}^* \|_2^2 - \gamma_w,
\]

C2: \[
\langle -\mathbb{E}_{\xi, \epsilon} \nabla_{\mathbf{a}} \mathcal{L}(\tilde{\mathbf{w}}, \tilde{a}), \mathbf{a}^* - \mathbf{a} \rangle \geq c_a \| \mathbf{a} - \mathbf{a}^* \|_2^2 - \gamma_a,
\]

where \( \tilde{\mathbf{w}} = \mathbf{w} + \epsilon \) and \( \tilde{\mathbf{a}} = \mathbf{a} + \xi \).

With small noise,

- C1 and C2 jointly hold;
- \( \gamma_w = 0 \) and \( \gamma_a \) decreases.
Epoch II: Converging to the Global Optimum

**Theorem**

For any $\gamma > 0$, we choose $\rho_w^1 \leq C_w^1 \frac{\gamma}{k^p} < 1$ and $\rho_a \leq M_a$ for some constant $C_w^1$. 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

$$
T_2 = O \left( \frac{p}{\eta} \log \frac{1}{\gamma} \log \frac{1}{\delta} \| a^* \|_2^2 \right).
$$
Experiments: Success Rates ($p=6$)

(a) $1^\top a^*/||a^*||_2^2 = 0$

(b) $1^\top a^*/||a^*||_2^2 = 1$

(c) $1^\top a^*/||a^*||_2^2 = 4$

(d) $1^\top a^*/||a^*||_2^2 = 9$

**Graphs:**
- **GD**
- **P-GD**
- **SGD**
Experiments: Empirical Convergence

Phase II starts

P-GD with Noise Annealing

SGD with Step Size Annealing

\[ L(w_t, a_t) \]
Discussions

- The noise helps escape the local optimum.

- The step size annealing in practice has a similar effect on controlling the noise level.

- P-GD behaves differently for training convolutional weight $w$ and output weight $a$ in the early stage.

- For general deep neural networks, there exist many bad global optima, which cannot generalize. Does SGD escape from them for the same reason?

- To the best of our knowledge, this is the first theoretical result towards justifying the effect of noise in training NNs by SGD-type algorithms in the presence of the spurious optima.
Discussions

- The noise helps escape the local optimum.
- The step size annealing in practice has a similar effect on controlling the noise level.
- P-GD behaves differently for training convolutional weight $w$ and output weight $a$ in the early stage.
- For general deep neural networks, there exist many bad global optima, which cannot generalize. Does SGD escape from them for the same reason?
- To the best of our knowledge, this is the first theoretical result towards justifying the effect of noise in training NNs by SGD-type algorithms in the presence of the spurious optima.
Discussions

- The noise helps escape the local optimum.
- The step size annealing in practice has a similar effect on controlling the noise level.
- P-GD behaves differently for training convolutional weight $w$ and output weight $a$ in the early stage.
- For general deep neural networks, there exist many bad global optima, which cannot generalize. Does SGD escape from them for the same reason?
- To the best of our knowledge, this is the first theoretical result towards justifying the effect of noise in training NNs by SGD-type algorithms in the presence of the spurious optima.
Discussions

- The noise helps escape the local optimum.
- The step size annealing in practice has a similar effect on controlling the noise level.
- P-GD behaves differently for training convolutional weight $w$ and output weight $a$ in the early stage.
- For general deep neural networks, there exit many bad global optima, which cannot generalize. Does SGD escape from them for the same reason?
- To the best of our knowledge, this is the first theoretical result towards justifying the effect of noise in training NNs by SGD-type algorithms in the presence of the spurious optima.
Discussions

- The noise helps escape the local optimum.
- The step size annealing in practice has a similar effect on controlling the noise level.
- P-GD behaves differently for training convolutional weight $w$ and output weight $a$ in the early stage.
- For general deep neural networks, there exit many bad global optima, which cannot generalize. Does SGD escape from them for the same reason?
- To the best of our knowledge, this is the first theoretical result towards justifying the effect of noise in training NNs by SGD-type algorithms in the presence of the spurious optima.
Our Paper:

Poster: Jun. 12 Wed 6:30-9:00 PM Pacific Ballroom No.26
Thank You! Questions?