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!

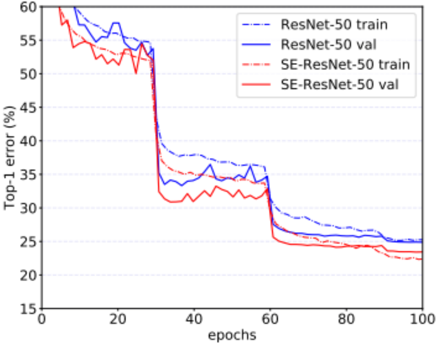
Efficient Training by First Order Algorithms

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!

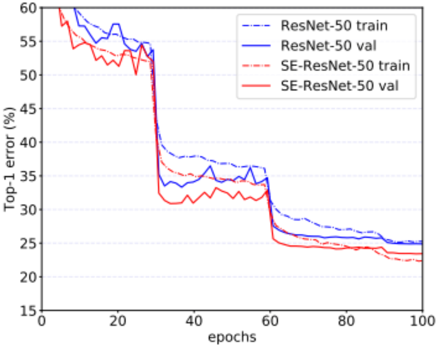
Practitioners' Choice: Step Size Annealing



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

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

What We Know:

- Not all optima generalize;
- Noise helps select optima that generalize.

Our Empirical Observations

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

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 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 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 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 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 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 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.

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!

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!

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!

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!

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:

$$f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) = \sum_{j=1}^{k} a_j^* \sigma(\boldsymbol{Z}_j^\top \boldsymbol{w}^*) \quad \text{with} \quad \|\boldsymbol{w}^*\|_2 = 1,$$

where $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{a} \in \mathbb{R}^k$, $\mathbf{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_k]$ with $\boldsymbol{Z}_j$'s are independently sampled from $N(\mathbf{0}, \mathbf{I})$, and $\sigma(\cdot) = \max\{\cdot, 0\}$.

- Nonconvex Optimization:

$$(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{a}}) = \arg\min_{\boldsymbol{w}, \boldsymbol{a}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) \quad \text{subject to} \quad \|\boldsymbol{w}\|_2 = 1,$$

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 $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{a}})$ exists!

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

Two-layer Nonoverlapping CNNs

- Teacher Network Model:

$$f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) = \sum_{j=1}^{k} a_j^* \sigma(\boldsymbol{Z}_j^\top \boldsymbol{w}^*) \quad \text{with} \quad \|\boldsymbol{w}^*\|_2 = 1,$$

where $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{a} \in \mathbb{R}^k$, $\mathbf{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_k]$ with $\boldsymbol{Z}_j$'s are independently sampled from $N(\mathbf{0}, \mathbf{I})$, and $\sigma(\cdot) = \max\{\cdot, 0\}$.

- Nonconvex Optimization:

$$(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{a}}) = \arg\min_{\boldsymbol{w}, \boldsymbol{a}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) \quad \text{subject to} \quad \|\boldsymbol{w}\|_2 = 1,$$

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 $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{a}})$ exists!

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

Two-layer Nonoverlapping CNNs

- Teacher Network Model:

$$f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) = \sum_{j=1}^{k} a_j^* \sigma(\boldsymbol{Z}_j^\top \boldsymbol{w}^*) \quad \text{with} \quad \|\boldsymbol{w}^*\|_2 = 1,$$

where $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{a} \in \mathbb{R}^k$, $\mathbf{Z} = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_k]$ with $\boldsymbol{Z}_j$'s are independently sampled from $N(\mathbf{0}, \mathbf{I})$, and $\sigma(\cdot) = \max\{\cdot, 0\}$.

- Nonconvex Optimization:

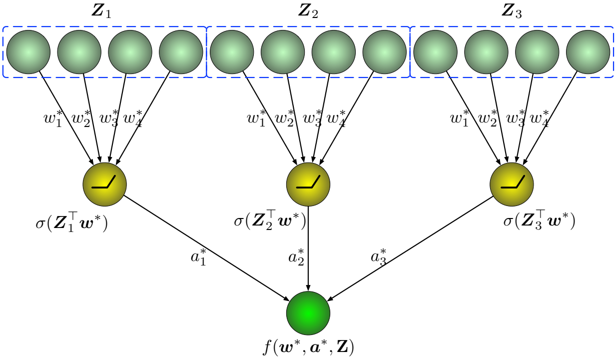
$$(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{a}}) = \arg\min_{\boldsymbol{w}, \boldsymbol{a}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{a}) \quad \text{subject to} \quad \|\boldsymbol{w}\|_2 = 1,$$

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 $(\overline{\boldsymbol{w}}, \overline{\boldsymbol{a}})$ exists!

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

Teacher Network Model



Perturbed Gradient Descent (P-GD)

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

At the t-th iteration, we independently sample

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

and $\mathbf{Z}^{(t)} = [\boldsymbol{Z}_1^{(t)}, ..., \boldsymbol{Z}_t^{(t)}]$ with $\boldsymbol{Z}_j^{(t)} \sim N(0, \mathbf{I})$, 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 $\ell(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}) = \frac{1}{2}(f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) - f(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.

Perturbed Gradient Descent (P-GD)

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

At the t-th iteration, we independently sample

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

and $\mathbf{Z}^{(t)} = [\boldsymbol{Z}_1^{(t)}, ..., \boldsymbol{Z}_t^{(t)}]$ with $\boldsymbol{Z}_j^{(t)} \sim N(0, \mathbf{I})$, 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 w and 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 $\ell(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}) = \frac{1}{2}(f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) - f(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.

Perturbed Gradient Descent (P-GD)

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

At the t-th iteration, we independently sample

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

and $\mathbf{Z}^{(t)} = [\boldsymbol{Z}_1^{(t)}, ..., \boldsymbol{Z}_t^{(t)}]$ with $\boldsymbol{Z}_j^{(t)} \sim N(0, \mathbf{I})$, 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 $\ell(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}) = \frac{1}{2}(f(\boldsymbol{w}^*, \boldsymbol{a}^*, \mathbf{Z}) - f(\boldsymbol{w}, \boldsymbol{a}, \mathbf{Z}))^2$.

Noise Annealing

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

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

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

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

Noise Annealing

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

- 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

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

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

Algorithmic Behaviors

Nonasymptotic Convergence Analysis

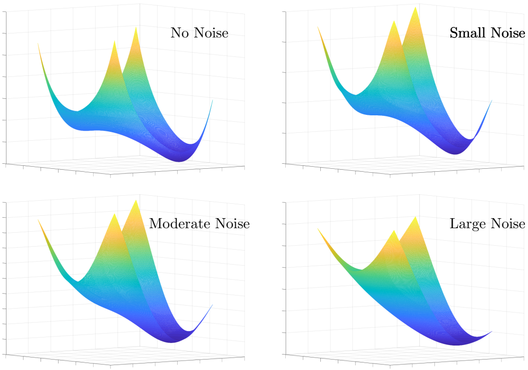
Theorem (Informal)

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": $\mathbb{P}(A) \geq 1 - \mathcal{O}(\exp(-1/\eta))$.

Convolutional Effects



Remark: P-GD essentially solves:

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

Partially Dissipative Conditions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\rho_w^0 = C_w^0 kp^2 \geq 1$ *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*

$$0 < m_a \leq a_t^\top a^* \leq M_a \quad \textit{and} \quad \angle(w_t, w^*) \leq \frac{5}{12}\pi \tag{1}$$

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

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

Remark: (1) is in the basin of attraction of the global optimum.

Epoch I: Escaping the Spurious Local Optimum

Theorem

Suppose $\rho_w^0 = C_w^0 kp^2 \geq 1$ *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*

$$0 < m_a \leq a_t^\top a^* \leq M_a \quad \text{and} \quad \angle(w_t, w^*) \leq \frac{5}{12}\pi \tag{1}$$

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

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

Remark: (1) is in the basin of attraction of the global optimum.

Epoch II: Converging to the Global Optimum

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

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

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

With **small** noise,

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

Epoch II: Converging to the Global Optimum

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

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

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

With **small** noise,

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

Epoch II: Converging to the Global Optimum

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

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

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

With **small** noise,

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

Epoch II: Converging to the Global Optimum

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

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

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

With **small** noise,

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

Epoch II: Converging to the Global Optimum

Theorem

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

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

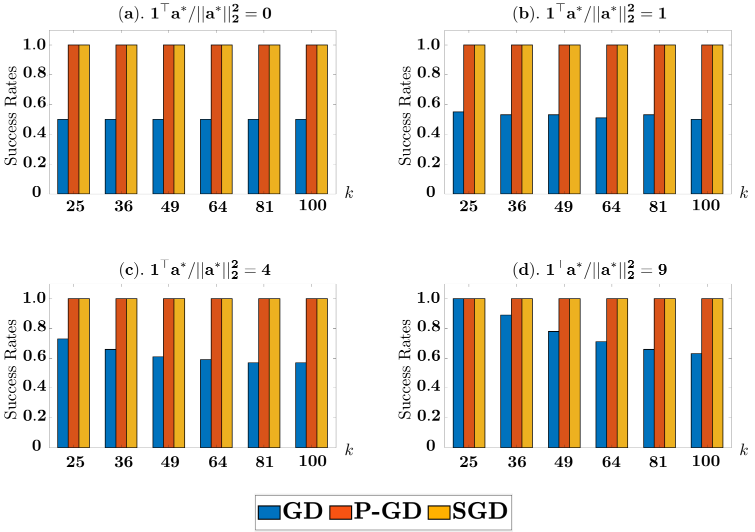
Then with probability at least $1-\delta$, *we have*

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

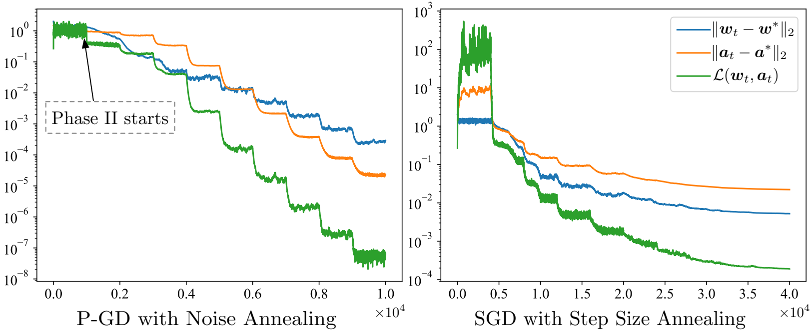
for any t*'s such that* $T_2 \leq t \leq T = O(\eta^{-2})$, *where*

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

Experiments: Success Rates (p=6)



Experiments: Empirical Convergence



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.

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 $\boldsymbol{w}$ and output weight $\boldsymbol{a}$ 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 $\boldsymbol{w}$ and output weight $\boldsymbol{a}$ 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 $\boldsymbol{w}$ and output weight $\boldsymbol{a}$ 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?