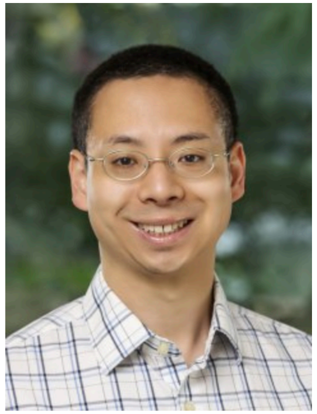
Regularization Matters: Neural Nets v.s. their Induced Kernel

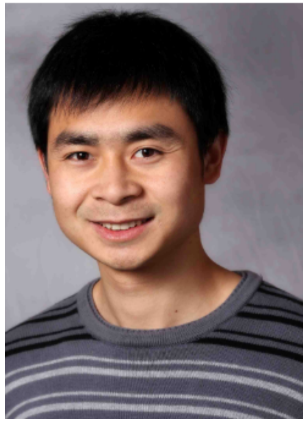
Colin Wei
Stanford University

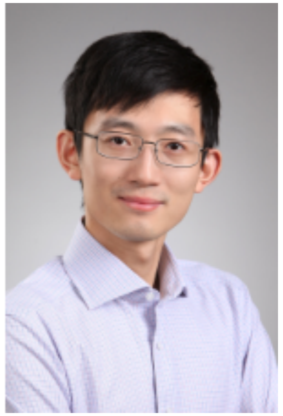
Jason Lee
Princeton University

Qiang Liu
UT Austin

Tengyu Ma
Stanford University







Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning

[Du et. al’18, Li and Liang’18, Jacot et. al’18]

Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning
  - Specific learning rate and initialization scale => gradient descent learns kernel method with features from random initialization

[Du et. al'18, Li and Liang'18, Jacot et. al'18]

Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning
  - Specific learning rate and initialization scale => gradient descent learns kernel method with features from random initialization
- NTK can’t explain generalization 🙁

[Du et. al’18, Li and Liang’18, Jacot et. al’18]

Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning
  - Specific learning rate and initialization scale => gradient descent learns kernel method with features from random initialization
- NTK can't explain generalization
  - Unrealistic learning rate and initialization scale

[Du et. al'18, Li and Liang'18, Jacot et. al'18]

Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning
  - Specific learning rate and initialization scale => gradient descent learns kernel method with features from random initialization
- NTK can't explain generalization 🙁
  - Unrealistic learning rate and initialization scale
  - Neural nets outperform kernel methods in practice

[Du et. al'18, Li and Liang'18, Jacot et. al'18]

Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning
  - Specific learning rate and initialization scale => gradient descent learns kernel method with features from random initialization
- NTK can’t explain generalization 🙁
  - Unrealistic learning rate and initialization scale
  - Neural nets outperform kernel methods in practice
  - Doesn’t allow for ℓ_2 regularization!

[Du et. al’18, Li and Liang’18, Jacot et. al’18]

Can Neural Tangent Kernel (NTK) Explain Deep Learning?

- Neural Tangent Kernel (NTK): recent attempt to explain success of deep learning
  - Specific learning rate and initialization scale => gradient descent learns kernel method with features from random initialization
- NTK can't explain generalization 🙁
  - Unrealistic learning rate and initialization scale
  - Neural nets outperform kernel methods in practice
  - Doesn't allow for ℓ_2 regularization!

Our work: what can we say about optimization/generalization with ℓ_2 regularizer?

[Du et. al'18, Li and Liang'18, Jacot et. al'18]

Main Results I: ℓ_2-regularized NN can generalize much better than NTK

Main Results I: ℓ_2-regularized NN can generalize much better than NTK

Our work: distribution in d dimensions with

NTK: $\boldsymbol{\Omega(d^2)}$ samples to learn $\qquad$ ℓ_2**-regularized logistic loss:** $\boldsymbol{O(d)}$ samples

Main Results I: ℓ_2-regularized NN can generalize much better than NTK

Our work: distribution in d dimensions with

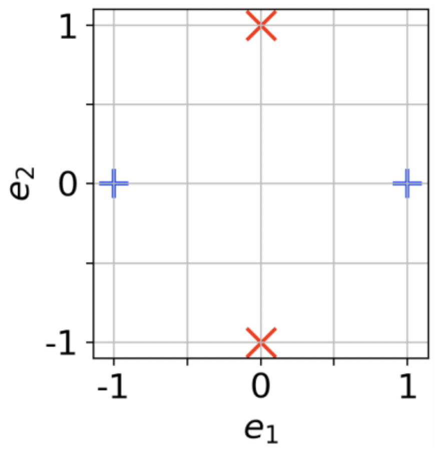
NTK: $\boldsymbol{\Omega(d^2)}$ samples to learn $\qquad$ **ℓ_2-regularized logistic loss:** $\boldsymbol{O(d)}$ samples

- Construction:

  First two coordinates:

  $$y = +1, (x_1, x_2) = (\pm 1, 0) \text{ w.p. } ½$$
  $$y = -1, (x_1, x_2) = (0, \pm 1) \text{ w.p. } ½$$

  Remaining $d - 2$ coordinates are noise



Main Results I: ℓ_2-regularized NN can generalize much better than NTK

Our work: distribution in d dimensions with

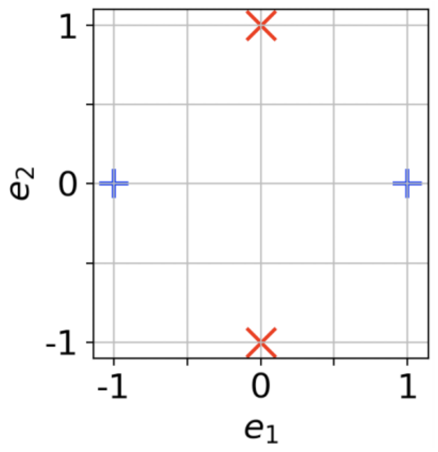
NTK: $\boldsymbol{\Omega(d^2)}$ samples to learn $\qquad$ **ℓ_2-regularized logistic loss:** $\boldsymbol{O(d)}$ samples

- Construction:

  First two coordinates:

  $y = +1, (x_1, x_2) = (\pm 1, 0)$ w.p. ½
  $y = -1, (x_1, x_2) = (0, \pm 1)$ w.p. ½

  Remaining $d - 2$ coordinates are noise



- NTK overfits to noise, whereas regularized neural net focuses on signal in x_1, x_2

Main Results I: ℓ_2-regularized NN can generalize much better than NTK

Our work: distribution in d dimensions with

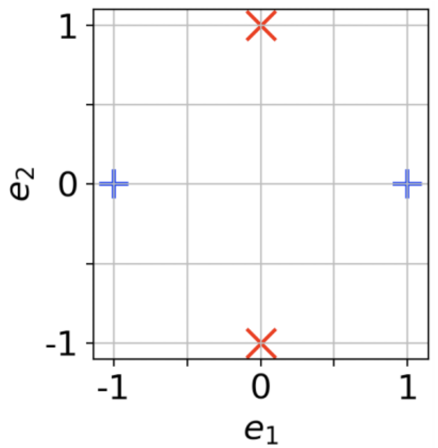
NTK: $\boldsymbol{\Omega(d^2)}$ samples to learn ℓ_2**-regularized logistic loss:** $\boldsymbol{O(d)}$ samples

- Construction:

  First two coordinates:

  $$y = +1, (x_1, x_2) = (\pm 1, 0) \text{ w.p. } 1/2$$
  $$y = -1, (x_1, x_2) = (0, \pm 1) \text{ w.p. } 1/2$$

  Remaining $d - 2$ coordinates are noise



- NTK overfits to noise, whereas regularized neural net focuses on signal in x_1, x_2

Takeaway: ℓ_2 regularization can **adaptively choose important features**, whereas NTK can't

Main Results II: global-minimizer of regularized logistic loss $\approx$ max-margin neural net

Main Results II: global-minimizer of regularized logistic loss ≈ max-margin neural net

- Training objective:

$$\text{logistic loss} + \lambda \cdot ||\text{parameters}||^2$$

Main Results II: global-minimizer of regularized logistic loss ≈ max-margin neural net

- Training objective:

$$\text{logistic loss} + \lambda \cdot ||\text{parameters}||^2$$

Our result: If network is homogeneous, global minimizer approaches max-margin solution as $\lambda \to 0$

Main Results II: global-minimizer of regularized logistic loss ≈ max-margin neural net

- Training objective:

$$\text{logistic loss} + \lambda \cdot ||\text{parameters}||^2$$

Our result: If network is homogeneous, global minimizer approaches max-margin solution as $\lambda \to 0$

- Holds regardless of depth, e.g. for any feedforward relu network

Main Results II: global-minimizer of regularized logistic loss ≈ max-margin neural net

- Training objective:

$$\text{logistic loss} + \lambda \cdot ||\text{parameters}||^2$$

> **Our result:** If network is homogeneous, global minimizer approaches max-margin solution as $\lambda \to 0$

- Holds regardless of depth, e.g. for any feedforward relu network
- [Golowich et. al'17] => generalization of global min bounded by inverse max-margin
- Max-margin non-decreasing with width => increasing network size improves bound

Main Results III: Optimization

- Previous slide: global min of regularized logistic loss => good statistical properties

Main Results III: Optimization

- Previous slide: global min of regularized logistic loss => good statistical properties

How do we obtain a global optimizer of the loss?

Main Results III: Optimization

- Previous slide: global min of regularized logistic loss => good statistical properties

$$\boxed{\text{How do we obtain a global optimizer of the loss?}}$$

- For infinite-width two-layer neural net, **noisy** gradient descent converges to global optimizer in polynomial iterations

Conclusion

Conclusion

- Statistical properties of ℓ_2-regularized neural net:
  - Can generalize much better than NTK

Conclusion

- Statistical properties of ℓ_2-regularized neural net:
  - Can generalize much better than NTK
  - Global minimizer of regularized loss $\approx$ max-margin solution

Conclusion

- Statistical properties of ℓ_2-regularized neural net:
  - Can generalize much better than NTK
  - Global minimizer of regularized loss $\approx$ max-margin solution
- Optimization:
  - For infinite-width nets, noisy gradient descent finds global optimizers in poly iterations

Conclusion

- Statistical properties of ℓ_2-regularized neural net:
  - Can generalize much better than NTK
  - Global minimizer of regularized loss $\approx$ max-margin solution
- Optimization:
  - For infinite-width nets, noisy gradient descent finds global optimizers in poly iterations
- **Future work:** optimization for regularized finite-size neural nets?

Conclusion

- Statistical properties of ℓ_2-regularized neural net:
  - Can generalize much better than NTK
  - Global minimizer of regularized loss $\approx$ max-margin solution
- Optimization:
  - For infinite-width nets, noisy gradient descent finds global optimizers in poly iterations
- **Future work:** optimization for regularized finite-size neural nets?

Come find our poster: 05:00 -- 07:00 PM @ East Exhibition Hall B + C #236!