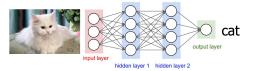
Over-parameterized nonlinear learning: Gradient descent follows the shortest path?

Samet Oymak and Mahdi Soltanolkotabi
Department of Electrical and Computer Engineering

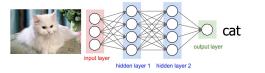


June 2019

Modern learning (e.g. deep learning) involves fitting nonlinear models



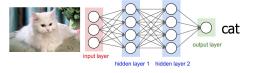
Modern learning (e.g. deep learning) involves fitting nonlinear models



Mystery

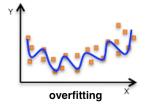
of parameters >> # of training data

Modern learning (e.g. deep learning) involves fitting nonlinear models

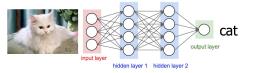


Mystery

of parameters >> # of training data

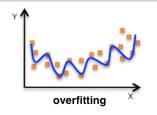


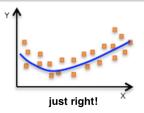
Modern learning (e.g. deep learning) involves fitting nonlinear models



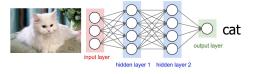
Mystery

of parameters >> # of training data



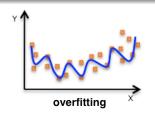


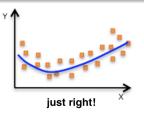
Modern learning (e.g. deep learning) involves fitting nonlinear models



Mystery

of parameters >> # of training data





Challenges

- Optimization: Why can you find a global optima despite nonconvexity?
- Generalization: Why is the global optima any good for prediction?

$$\min_{\pmb{\theta} \in \mathbb{R}^p} \mathcal{L}(\pmb{\theta}) := \frac{1}{2} \left\| \pmb{X} \pmb{\theta} - \pmb{y} \right\|_{\ell_2}^2 \quad \text{with} \quad \pmb{X} \in \mathbb{R}^{n \times p} \quad \text{and} \quad n \leq p.$$

$$\min_{oldsymbol{ heta} \in \mathbb{R}^p} \mathcal{L}(oldsymbol{ heta}) := rac{1}{2} \left\| oldsymbol{X} oldsymbol{ heta} - oldsymbol{y}
ight\|_{\ell_2}^2 \quad ext{with} \quad oldsymbol{X} \in \mathbb{R}^{n imes p} \quad ext{and} \quad n \leq p.$$

Gradient descent starting from $heta_0$ has three properties:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \left\| \boldsymbol{X} \boldsymbol{\theta} - \boldsymbol{y} \right\|_{\ell_2}^2 \quad \text{with} \quad \boldsymbol{X} \in \mathbb{R}^{n \times p} \quad \text{and} \quad n \leq p.$$

Gradient descent starting from θ_0 has three properties:

Global convergence

$$\min_{m{ heta} \in \mathbb{R}^p} \mathcal{L}(m{ heta}) := rac{1}{2} \left\| m{X} m{ heta} - m{y}
ight\|_{\ell_2}^2 \quad ext{with} \quad m{X} \in \mathbb{R}^{n imes p} \quad ext{and} \quad n \leq p.$$

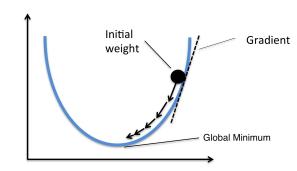
Gradient descent starting from θ_0 has three properties:

- Global convergence
- Converges to closest global optima to $oldsymbol{ heta}_0$

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \ \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \left\| \boldsymbol{X} \boldsymbol{\theta} - \boldsymbol{y} \right\|_{\ell_2}^2 \quad \text{with} \quad \boldsymbol{X} \in \mathbb{R}^{n \times p} \quad \text{and} \quad n \leq p.$$

Gradient descent starting from $heta_0$ has three properties:

- Global convergence
- Converges to closest global optima to θ_0
- Follows a direct trajectory



Over-parametrized nonlinear least-squares

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \, \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \left\| f(\boldsymbol{\theta}) - \boldsymbol{y} \right\|_{\ell_2}^2,$$

where

$$m{y} := egin{bmatrix} m{y}_1 \ m{y}_2 \ dots \ m{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(m{ heta}) := egin{bmatrix} f(m{x}_1; m{ heta}) \ f(m{x}_2; m{ heta}) \ dots \ f(m{x}_n; m{ heta}) \end{bmatrix} \in \mathbb{R}^n, \quad ext{and} \quad n \leq p.$$

Over-parametrized nonlinear least-squares

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \| f(\boldsymbol{\theta}) - \boldsymbol{y} \|_{\ell_2}^2,$$

where

$$m{y} := egin{bmatrix} m{y}_1 \ m{y}_2 \ dots \ m{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(m{ heta}) := egin{bmatrix} f(m{x}_1; m{ heta}) \ f(m{x}_2; m{ heta}) \ dots \ f(m{x}_n; m{ heta}) \end{bmatrix} \in \mathbb{R}^n, \quad ext{and} \quad n \leq p.$$

Run gradient descent: $oldsymbol{ heta}_{ au+1} = oldsymbol{ heta}_{ au} - \eta_{ au}
abla \mathcal{L}(oldsymbol{ heta}_{ au})$

Over-parametrized nonlinear least-squares

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2} \left\| f(\boldsymbol{\theta}) - \boldsymbol{y} \right\|_{\ell_2}^2,$$

where

$$m{y} := egin{array}{c} m{y}_1 \ m{y}_2 \ dots \ m{y}_n \ \end{array} \in \mathbb{R}^n, \quad f(m{ heta}) := egin{array}{c} f(m{x}_1; m{ heta}) \ f(m{x}_2; m{ heta}) \ dots \ f(m{x}_n; m{ heta}) \ \end{array} \in \mathbb{R}^n, \quad ext{and} \quad n \leq p.$$

Run gradient descent:
$$oldsymbol{ heta}_{ au+1} = oldsymbol{ heta}_{ au} - \eta_{ au}
abla \mathcal{L}(oldsymbol{ heta}_{ au})$$

Gradient and Jacobian

$$\nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})^T (f(\boldsymbol{\theta}) - \boldsymbol{u}).$$

- $\mathcal{J}(\theta) = \frac{\partial f(\theta)}{\partial \mathbf{Q}} \in \mathbb{R}^{n \times p}$ is the Jacobian matrix
- ullet Intuition: Jacobian replaces the feature matrix X

Gradient descent trajectory

Assumptions

- minimum singular value at initialization: $\sigma_{\min}\left(\mathcal{J}(\boldsymbol{\theta}_0)\right) \geq 2\alpha$
- maximum singular value: $\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta$
- Jacobian smoothness: $\|\mathcal{J}(m{ heta}_2) \mathcal{J}(m{ heta}_1)\| \leq L \, \|m{ heta}_2 m{ heta}_1\|_{\ell_2}$
- Initial error: $\|f(m{ heta}_0) m{y}\|_{\ell_2} \leq rac{lpha^2}{4L}$

Gradient descent trajectory

Assumptions

- minimum singular value at initialization: $\sigma_{\min}(\mathcal{J}(\boldsymbol{\theta}_0)) \geq 2\alpha$
- maximum singular value: $\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta$
- Jacobian smoothness: $\|\mathcal{J}(m{ heta}_2) \mathcal{J}(m{ heta}_1)\| \leq L \, \|m{ heta}_2 m{ heta}_1\|_{\ell_2}$
- Initial error: $\|f(oldsymbol{ heta}_0) oldsymbol{y}\|_{\ell_2} \leq rac{lpha^2}{4L}$

Theorem (Oymak and Soltanolkotabi 2018)

Assume above over a ball of radius $R=rac{\|f(m{ heta}_0)-m{y}\|_{\ell_2}}{lpha}$ around $m{ heta}_0$ and Set $\eta=rac{1}{eta^2}.$

• Global convergence:

$$\left\|f(\boldsymbol{\theta}_{ au}) - \boldsymbol{y}\right\|_{\ell_2}^2 \le \left(1 - \frac{1}{2} \frac{\alpha^2}{\beta^2}\right)^{\tau} \left\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\right\|_{\ell_2}^2$$

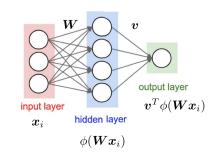
• Converges to near closest global minima to initialization:

$$\left\|oldsymbol{ heta}_{ au} - oldsymbol{ heta}_{0}
ight\|_{\ell_{2}} \leq rac{eta}{lpha} \left\|oldsymbol{ heta}^{*} - oldsymbol{ heta}_{0}
ight\|_{\ell_{2}}$$

Takes an approximately direct route

Concrete example: One-hidden layer neural net

- Training data: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Loss: $\mathcal{L}(oldsymbol{v}, oldsymbol{W}) := \sum_{i=1}^n \left(oldsymbol{v}^T \phi(oldsymbol{W} oldsymbol{x}_i) y_i
 ight)^2$
- Algorithm: gradient descent with random Gaussian initialization



Theorem (Oymak and Soltanolkotabi 2019)

As long as

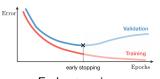
$$\#$$
parameters $\gtrsim (\#$ of training data) 2

Then, with high probability

- Zero training error: $\mathcal{L}(oldsymbol{v}_{ au}, oldsymbol{W}_{ au}) \leq (1ho)^{ au} \mathcal{L}(oldsymbol{v}_{0}, oldsymbol{W}_{0})$
 - Iterates remain close to initialization

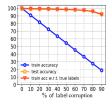
Further results and applications

- Extensions to SGD and other loss functions
- Theoretical justification for



Early stopping

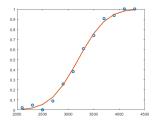
Other applications

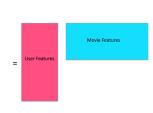


Robustness to label noise



Generalization





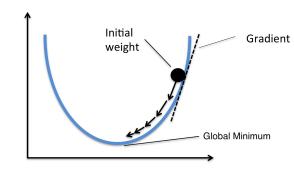
Fitting generalized linear models

Low-rank matrix recovery

Conclusion

(Stochastic) gradient descent has three intriguing properties

- Global convergence
- Converges to near closest global optima to init.
- Follows a direct trajectory



Thanks!

Poster

Thursday, 6:30 PM, # 95

References

- Over-parametrized nonlinear learning: Gradient descent follows the shortest path? S. Oymak and M. Soltanolkotabi
- Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. S. Oymak and M. Soltanolkotabi
- Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks. M. Li, M. Soltanolkotabi, and S. Oymak
- Generalization Guarantees for Neural Networks via Harnessing the Low-rank Structure of the Jacobian. S. Oymak, Z. Fabian, M. Li, and M. Soltanolkotabi