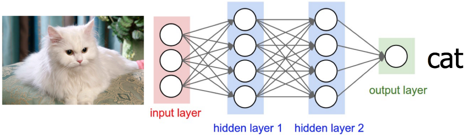
Over-parameterized nonlinear learning: Gradient descent follows the shortest path?

Samet Oymak and Mahdi Soltanolkotabi
Department of Electrical and Computer Engineering

June 2019

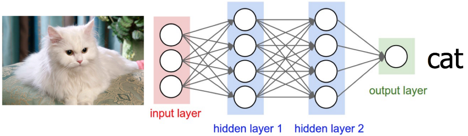
Motivation

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



Motivation

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

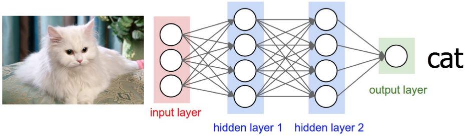


Mystery

# of parameters >> # of training data

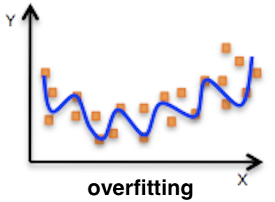
Motivation

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



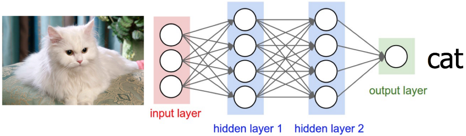
Mystery

$\#$ *of parameters* $>>$ $\#$ *of training data*



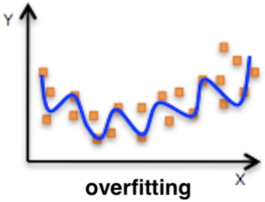
Motivation

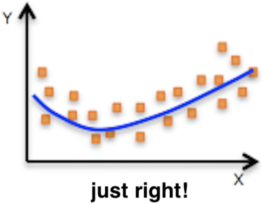
Modern learning (e.g. deep learning) involves fitting nonlinear models



Mystery

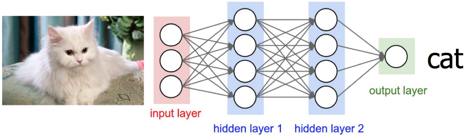
# of parameters $>>$ *# of training data*





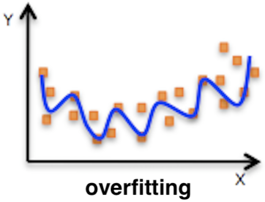
Motivation

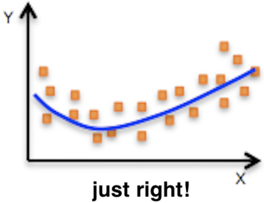
Modern learning (e.g. deep learning) involves fitting nonlinear models



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?*

Prelude: over-parametrized linear least-squares

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

Prelude: over-parametrized linear least-squares

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

Gradient descent starting from $\boldsymbol{\theta}_0$ has three properties:

Prelude: over-parametrized linear least-squares

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

Gradient descent starting from $\boldsymbol{\theta}_0$ has three properties:

- Global convergence

Prelude: over-parametrized linear least-squares

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

Gradient descent starting from $\boldsymbol{\theta}_0$ has three properties:

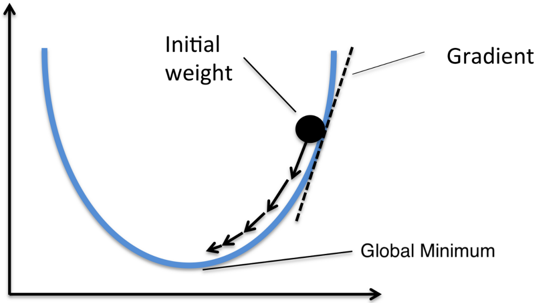
- Global convergence
- Converges to closest global optima to $\boldsymbol{\theta}_0$

Prelude: over-parametrized linear least-squares

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

Gradient descent starting from $\boldsymbol{\theta}_0$ has three properties:

- Global convergence
- Converges to closest global optima to $\boldsymbol{\theta}_0$
- 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

$$\boldsymbol{y} := \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\boldsymbol{\theta}) := \begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ f(\boldsymbol{x}_2;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_n;\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \leq p.$$

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

$$\boldsymbol{y} := \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\boldsymbol{\theta}) := \begin{bmatrix} f(\boldsymbol{x}_1; \boldsymbol{\theta}) \\ f(\boldsymbol{x}_2; \boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_n; \boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \leq p.$$

Run gradient descent: $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta_\tau \nabla \mathcal{L}(\boldsymbol{\theta}_\tau)$

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

$$\boldsymbol{y} := \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\boldsymbol{\theta}) := \begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\theta}) \\ f(\boldsymbol{x}_2;\boldsymbol{\theta}) \\ \vdots \\ f(\boldsymbol{x}_n;\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \le p.$$

Run gradient descent: $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta_\tau \nabla\mathcal{L}(\boldsymbol{\theta}_\tau)$

Gradient and Jacobian

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

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

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:* $\|\mathcal{J}(\boldsymbol{\theta}_2) - \mathcal{J}(\boldsymbol{\theta}_1)\| \leq L \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{\ell_2}$
- *Initial error:* $\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \leq \frac{\alpha^2}{4L}$

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:* $\|\mathcal{J}(\boldsymbol{\theta}_2) - \mathcal{J}(\boldsymbol{\theta}_1)\| \leq L\left\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\right\|_{\ell_2}$
- *Initial error:* $\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2} \leq \frac{\alpha^2}{4L}$

Theorem (Oymak and Soltanolkotabi 2018)

Assume above over a ball of radius $R = \frac{\|f(\boldsymbol{\theta}_0) - \boldsymbol{y}\|_{\ell_2}}{\alpha}$ *around* $\boldsymbol{\theta}_0$ *and Set* $\eta = \frac{1}{\beta^2}$.

- *Global convergence:*

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

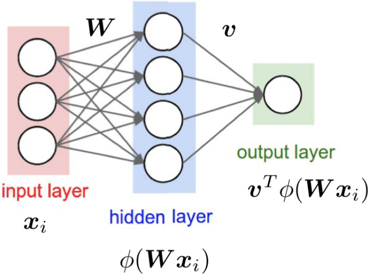
- *Converges to near closest global minima to initialization:*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \leq \frac{\beta}{\alpha}\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_{\ell_2}$$

- *Takes an approximately direct route*

Concrete example: One-hidden layer neural net

- Training data: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$
- Loss: $\mathcal{L}(\boldsymbol{v}, \boldsymbol{W}) := \sum_{i=1}^{n} \left(\boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}_i) - y_i\right)^2$
- Algorithm: gradient descent with random Gaussian initialization



Theorem (Oymak and Soltanolkotabi 2019)

As long as

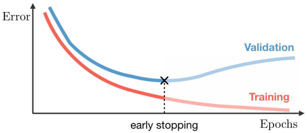
$$\#\text{parameters} \gtrsim (\#\text{of training data})^2$$

Then, with high probability

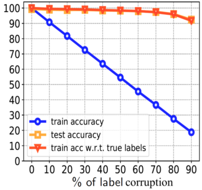
- *Zero training error:* $\mathcal{L}(\boldsymbol{v}_\tau, \boldsymbol{W}_\tau) \leq (1-\rho)^\tau \mathcal{L}(\boldsymbol{v}_0, \boldsymbol{W}_0)$
- *Iterates remain close to initialization*

Further results and applications

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



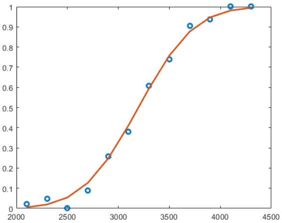
Early stopping



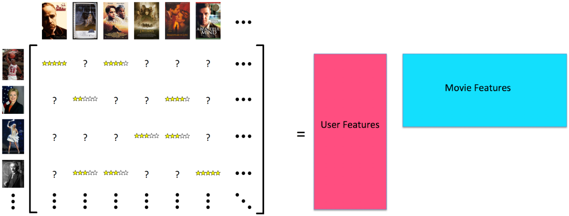
Robustness to label noise

Generalization

- Other applications



Fitting generalized linear models

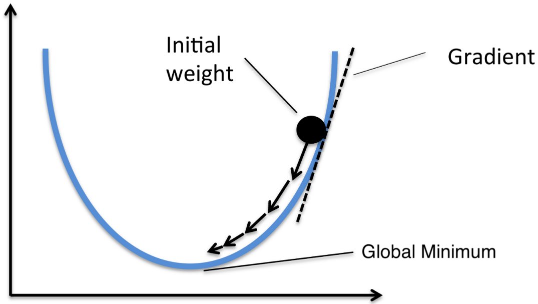


Low-rank matrix recovery

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