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.

![Diagram of a neural network with an input layer, hidden layer 1, hidden layer 2, and an output layer. The output is labeled 'cat'.]
Motivation

Modern learning (e.g. deep learning) involves fitting nonlinear models. The number of parameters is significantly greater than the number of training data.

Mystery

\[ \# \text{ of parameters} \gg \# \text{ of training data} \]
Motivation

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

Mystery

\[ \text{# of parameters} \gg \text{# of training data} \]

Challenges

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

\[
\min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta) := \frac{1}{2} \| X \theta - y \|_{\ell_2}^2 \quad \text{with} \quad X \in \mathbb{R}^{n \times p} \quad \text{and} \quad n \leq p.
\]
Prelude: over-parametrized linear least-squares

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

Gradient descent starting from \( \theta_0 \) has three properties:
Prelude: over-parametrized linear least-squares

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

Gradient descent starting from \( \theta_0 \) has three properties:

- Global convergence
Prelude: over-parametrized linear least-squares

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

Gradient descent starting from \( \theta_0 \) has three properties:

- Global convergence
- Converges to closest global optima to \( \theta_0 \)
Prelude: over-parametrized linear least-squares

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

Gradient descent starting from \( \theta_0 \) has three properties:

- Global convergence
- Converges to closest global optima to \( \theta_0 \)
- Follows a direct trajectory
Over-parametrized nonlinear least-squares

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

where

\[
y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\theta) := \begin{bmatrix} f(x_1; \theta) \\ f(x_2; \theta) \\ \vdots \\ f(x_n; \theta) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \leq p.
\]
Over-parametrized nonlinear least-squares

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

where

\[
y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\theta) := \begin{bmatrix} f(x_1; \theta) \\ f(x_2; \theta) \\ \vdots \\ f(x_n; \theta) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \leq p.
\]

Run gradient descent: \[\theta_{\tau+1} = \theta_{\tau} - \eta_{\tau} \nabla \mathcal{L}(\theta_{\tau})\]
Over-parametrized nonlinear least-squares

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

where

\[
y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad f(\theta) := \begin{bmatrix} f(x_1; \theta) \\ f(x_2; \theta) \\ \vdots \\ f(x_n; \theta) \end{bmatrix} \in \mathbb{R}^n, \quad \text{and} \quad n \leq p.
\]

Run gradient descent:

\[
\theta_{\tau+1} = \theta_\tau - \eta_\tau \nabla \mathcal{L}(\theta_\tau)
\]

Gradient and Jacobian

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

- \( \mathcal{J}(\theta) = \frac{\partial f(\theta)}{\partial \theta} \in \mathbb{R}^{n \times p} \) is the Jacobian matrix

- **Intuition:** Jacobian replaces the feature matrix \( X \)
Gradient descent trajectory

Assumptions

- **minimum singular value at initialization:** \( \sigma_{\text{min}}(J(\theta_0)) \geq 2\alpha \)
- **maximum singular value:** \( \|J(\theta)\| \leq \beta \)
- **Jacobian smoothness:** \( \|J(\theta_2) - J(\theta_1)\| \leq L \|\theta_2 - \theta_1\|_{\ell_2} \)
- **Initial error:** \( \|f(\theta_0) - y\|_{\ell_2} \leq \frac{\alpha^2}{4L} \)
Gradient descent trajectory

Assumptions

- **minimum singular value at initialization:** \( \sigma_{\text{min}}(J(\theta_0)) \geq 2\alpha \)
- **maximum singular value:** \( \|J(\theta)\| \leq \beta \)
- **Jacobian smoothness:** \( \|J(\theta_2) - J(\theta_1)\| \leq L \|\theta_2 - \theta_1\|_2 \)
- **Initial error:** \( \|f(\theta_0) - y\|_2 \leq \frac{\alpha^2}{4L} \)

Theorem (Oymak and Soltanolkotabi 2018)

Assume above over a ball of radius \( R = \frac{\|f(\theta_0) - y\|_2}{\alpha} \) around \( \theta_0 \) and set \( \eta = \frac{1}{\beta^2} \).

- **Global convergence:**
  \[
  \|f(\theta_\tau) - y\|_2^2 \leq \left(1 - \frac{1}{2} \frac{\alpha^2}{\beta^2}\right)^\tau \|f(\theta_0) - y\|_2^2
  \]

- **Converges to near closest global minima to initialization:**
  \[
  \|\theta_\tau - \theta_0\|_2 \leq \frac{\beta}{\alpha} \|\theta^* - \theta_0\|_2
  \]

- **Takes an approximately direct route**
Concrete example: One-hidden layer neural net

- Training data:
  \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)
- Loss:
  \[\mathcal{L}(v, W) := \sum_{i=1}^{n} (v^T \phi(Wx_i) - y_i)^2\]
- 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: \(\mathcal{L}(v_\tau, W_\tau) \leq (1 - \rho)^\tau \mathcal{L}(v_0, W_0)\)
- 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
(Stochastic) gradient descent has three intriguing properties:

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

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