On Nonlinear Least Squares and Gradient Descent

CS189/289A: Introduction to Machine Learning

Stella Yu

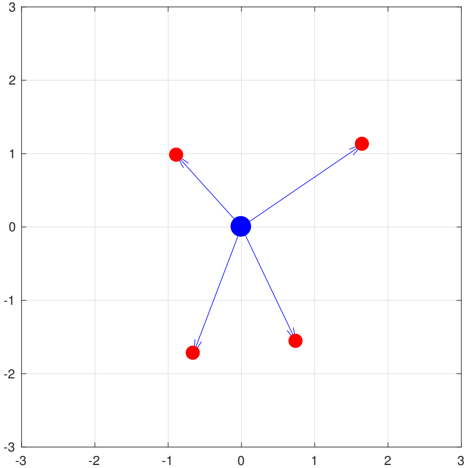
UC Berkeley

28 September 2017

Why Nonlinear Least Squares (NLS) ?

- MLE/MAP estimations are often nonlinear
- Example 1: Sensor localization from range measurements
- Example 2: Orthogonal distance regression

Sensor Localization from Range Measurements



- Consider unknown target location θ on the ground plane.
- There are n sensors providing noisy range measurements Y of this unknown target to their locations X:

$$Y_i = \|X_i - \theta\| + N_i, \quad N_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n \tag{1}$$

Sensor Location: Maximum Likelihood Estimate

- Nonlinear range data generation model:

$$Y_i = f(X_i;\theta) + N_i = \|X_i - \theta\| + N_i, \quad N_i \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

$$Y_i \sim \mathcal{N}(f(X_i;\theta), \sigma^2) = \mathcal{N}(\|X_i - \theta\|, \sigma^2), \quad i = 1, \ldots, n \tag{3}$$

- Maximum likelihood estimate:

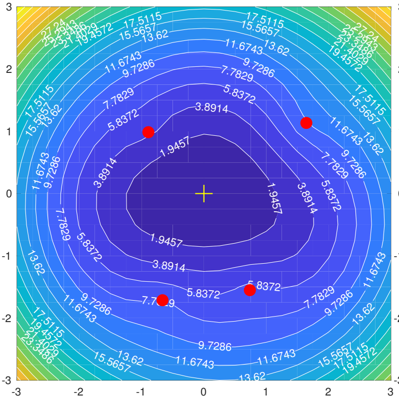
$$\arg\max_\theta \log P(y_1, \ldots, y_n | x_1, \ldots, x_n; \theta, \sigma) \tag{4}$$

$$= \arg\min_\theta \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 \tag{5}$$

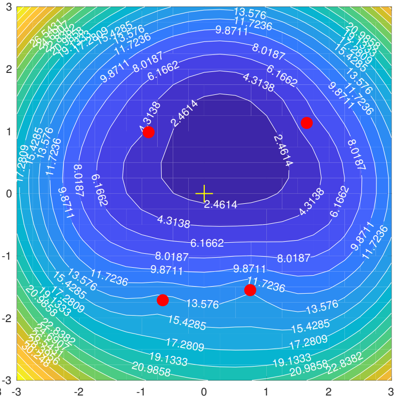
$$= \arg\min_\theta \sum_{i=1}^{n} (y_i - \|x_i - \theta\|)^2 \tag{6}$$

- Nonlinear underlying function $f(x;\theta)$
- Least square error metric for fitting the model

LS Error Metric: Center Target Location

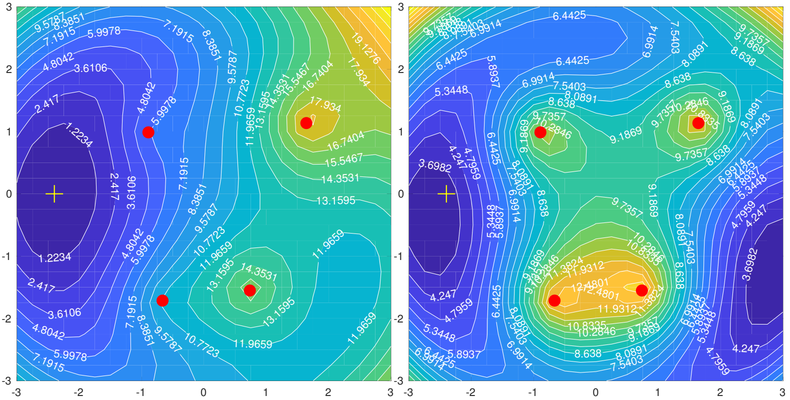


$\sigma = 0$



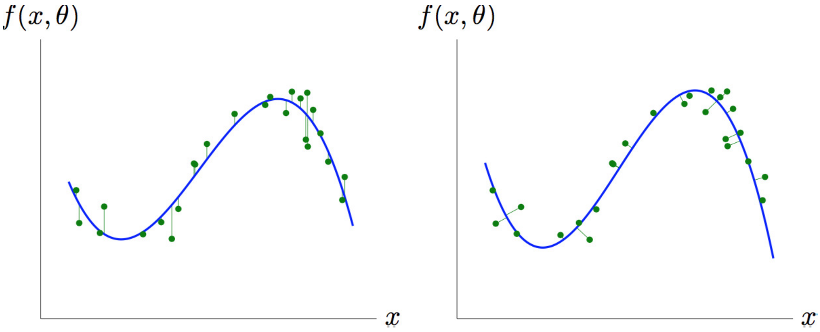
$\sigma = 1$

LS Error Metric: Side Target Location



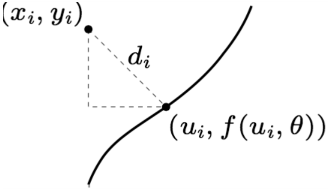
$\sigma = 0$ $\sigma = 1$

Orthogonal Distance Regression



$$f(x;\theta) = \theta_1 + \theta_2 x + \ldots + \theta_p x^{p-1} \qquad (7)$$

Orthogonal Distance Regression – Nonlinear TLS



$$d_i^2 = (f(u_i, \theta) - y_i)^2 + \|u_i - x_i\|^2$$

$$\min \varepsilon_{TLS}(u, \theta) = = \sum_{i=1}^{n} \left\| \begin{bmatrix} u_i \\ f(u_i; \theta) \end{bmatrix} - \begin{bmatrix} x_i \\ y_i \end{bmatrix} \right\|^2 \tag{8}$$

$$= \sum_{i=1}^{n} (f(u_i; \theta) - y_i)^2 + \|u_i - x_i\|^2 \tag{9}$$

- The i-th term is the distance of (x_i, y_i) to point $(u_i, f(u_i; \theta))$
- Optimize over model parameter θ and n points u_i
- Minimizing over u_i gives the distance to a given curve
- Minimizing over u and θ fits TLS to the model curve

Nonlinear Least Squares Model Fitting

$$\min \varepsilon_{LS}(\theta) = \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 = \sum_{i=1}^{n} r_i^2 \tag{10}$$

$$r_i = y_i - f(x_i;\theta) \tag{11}$$

- Model $f(x;\theta)$ has parameter θ of p dimensions
- $(x_1, y_1), \ldots, (x_n, y_n)$ are n data points
- Data x is of d dimensions
- Minimization over model parameter θ
- In linear regression we consider model f linear in θ:

$$f(x;\theta) = \theta_1 f_1(x) + \ldots + \theta_p f_p(x) \tag{12}$$

  Here we allow $f(x;\theta)$ to be a nonlinear function of θ, e.g.

$$f(x;\theta) = \|x - \theta\| \tag{13}$$

Nonlinear Least Squares Model Fitting

- The functional part of the model is not linear with respect to the unknown parameters
- The method of least squares is used to estimate the values of the unknown parameters
- Often, the function is smooth with respect to the unknown
- Broad application: Many processes are nonlinear models
- Good estimates with relatively small datasets
- Iterative optimization, good initialization required
- Sensitivity to outliers, just like linear LS model fitting.

Gradient of A Scalar Differentiable Function

- Scalar differentiable function $g(x)$, where x has d dimensions

$$\nabla g(x) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \vdots \\ \frac{\partial g}{\partial x_d} \end{bmatrix} \tag{14}$$

- Linear approximation of $g(x)$ near fixed point z by the first-order Taylor polynomial expansion:

$$g(x) \approx g(z) + \nabla g(z)'(x - z) \tag{15}$$

- Both $g(z)$ and $\nabla g(z)$ are constant

Jacobian of A Vector Differentiable Function

- Vector differentiable function $g(x)$, where x has d dimensions

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{bmatrix} \tag{16}$$

- Derivative matrix or Jacobian matrix:

$$J = \nabla g(x) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_d} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_d} \end{bmatrix} = \begin{bmatrix} \nabla g_1(x)' \\ \nabla g_2(x)' \\ \vdots \\ \nabla g_n(x)' \end{bmatrix} \tag{17}$$

- Linear approximation of $g(x)$ near fixed point z by the first-order Taylor polynomial expansion:

$$g(x) \approx g(z) + \nabla g(z)(x - z) \tag{18}$$

- Both $g(z)$ and $\nabla g(z)$ are constant

Optimality Condition of Nonlinear Least Squares

- Necessary condition for optimality: gradient must be 0

$$\min \varepsilon_{LS}(\theta) = \sum_{i=1}^{n} r_i^2, \quad r_i = y_i - f(x_i;\theta) \tag{19}$$

$$\frac{1}{2}\frac{\partial \varepsilon_{LS}}{\partial \theta_j} = \sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \theta_j} = 0, \quad j = 1,\ldots,p \tag{20}$$

- For linear models $f(x;\theta) = \theta' x$, θ has a linear solution.

$$\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \theta_j} = \sum_{i=1}^{n} (y_i - \theta' x_i) x_{ij} = 0 \tag{21}$$

- The zero-gradient condition is not sufficient for optimality.

Nonlinear Optimality Condition

- Known sensor $x_i = \begin{bmatrix} a_i \\ b_i \end{bmatrix}$, unknown target $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$,

$$r_i = y_i - \|x_i - \theta\| = y_i - \sqrt{(a_i - \theta_1)^2 + (b_i - \theta_2)^2} \tag{22}$$

$$\frac{\partial r_i}{\partial \theta_1} = \frac{(a_i - \theta_1)}{\sqrt{(a_i - \theta_1)^2 + (b_i - \theta_2)^2}} \tag{23}$$

$$\frac{\partial r_i}{\partial \theta_2} = \frac{(b_i - \theta_2)}{\sqrt{(a_i - \theta_1)^2 + (b_i - \theta_2)^2}} \tag{24}$$

- The optimality condition leads to a set of nonlinear equations:

$$\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \theta_1} = \sum_{i=1}^{n} \frac{r_i(a_i - \theta_1)}{\sqrt{(a_i - \theta_1)^2 + (b_i - \theta_2)^2}} = 0 \tag{25}$$

$$\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \theta_2} = \sum_{i=1}^{n} \frac{r_i(b_i - \theta_2)}{\sqrt{(a_i - \theta_1)^2 + (b_i - \theta_2)^2}} = 0 \tag{26}$$

- Unlike linear LS model fitting, θ has to be solved iteratively.

NLS in A Compact Matrix Format

error: $$\varepsilon(\theta) = R'R \tag{27}$$

residue: $$R = Y - F(\theta) \tag{28}$$

measurement: $$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{29}$$

prediction: $$F(\theta) = \begin{bmatrix} f(x_1;\theta) \\ f(x_2;\theta) \\ \vdots \\ f(x_n;\theta) \end{bmatrix}, \quad J(\theta) = \nabla F(\theta)_{n\times p} \tag{30}$$

optimality: $$\frac{1}{2}\frac{\partial E}{\partial \theta} = -\nabla F(\theta)'R = \nabla F(\theta)'(F(\theta) - Y) = 0 \tag{31}$$

$$J(\theta)'F(\theta) = J(\theta)'Y \tag{32}$$

Iterative Gradient Descent Optimization Method

- Start with an initial guess. At every iteration, an adjustment is made to θ with shift vector $\Delta\theta$:

$$\theta^{(k+1)} \approx \theta^{(k)} + \Delta\theta \tag{33}$$

- Linearize model with a first-order Taylor expansion:

$$F(\theta) \approx F(\theta^{(k)}) + \nabla F(\theta^{(k)})\Delta\theta = F(\theta^{(k)}) + J(\theta^{(k)})\Delta\theta \tag{34}$$

- The Jacobian is a function of constants, and it changes from one iteration to the next.
- The optimality condition in terms of the linearized model:

$$J(\theta^{(k)})' \cdot (F(\theta^{(k)}) + J(\theta^{(k)})\Delta\theta) = J(\theta^{(k)})'Y \tag{35}$$

$$J(\theta^{(k)})'J(\theta^{(k)})\Delta\theta = J(\theta^{(k)})'(Y - F(\theta^{(k)})) \tag{36}$$

- The normal equations:

$$(J'J)\Delta\theta = J'\Delta Y, \quad \Delta Y = Y - F(\theta^{(k)}) \tag{37}$$

Iterative Optimization of NLS

1. Initialization: $\theta = \theta^{(k)}$, $k = 0$.

2. Compute Jacobian: $J = \nabla F(\theta^{(k)})$.

3. Compute prediction error: $\Delta Y = Y - F(\theta^{(k)})$.

4. Update parameter: $\theta^{(k+1)} = \theta^{(k)} + (J'J)^{-1}J'\Delta Y$.

5. Convergence test:

$$\left| \frac{\varepsilon^{(k+1)} - \varepsilon^{(k)}}{\varepsilon^{(k)}} \right| < \text{threshold}, \quad \text{or} \tag{38}$$

$$\max_j \left| \frac{\Delta\theta_j}{\theta_j^{(k)}} \right| < \text{threshold} \tag{39}$$

6. If convegent, stop; otherwise, $k := k + 1$, go to Step 2.

Summary

- The underlying model is nonlinear with respect to parameters
- The method of least squares is used to estimate the parameters
- NLS has broad applications
- Good estimates with relatively small datasets
- Iterative optimization by gradient descent
- Pratical issues: initialization, regularization, convergence etc
- Sensitivity to outliers, just like linear LS model fitting
- Gradient descent applicable to nonlinear optimization in general, e.g. minimizing robust LS functions instead of quadratic LS errors.