ME491 Lecture 10

4.2 Nonlinear Regression and Gradient Descent

Not every system can be fit to a set of linear equations, more generally, we have nonlinear curve fitting, where we assume the general fitting form

$$f(x) = f(x, \beta)$$

we use $m<n$ fitting coefficients $\beta \in \mathbb{R}^m$ to minimize the error. The root-mean-square error is then defined as

$$E_2(\beta) = \sum_{k=1}^{n} \left(f(x_k, \beta) - y_k \right)^2$$

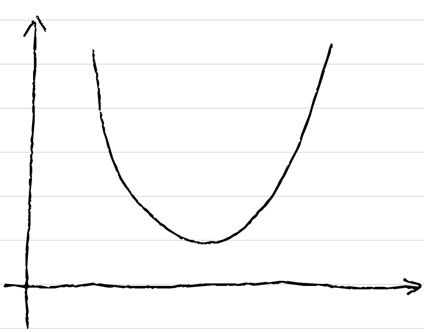
We minimize E_2 by finding zeros in the partial derivatives

$$\frac{\partial E_2}{\partial \beta_j} = 0 \quad \text{for } j = 1, 2, \ldots, m$$

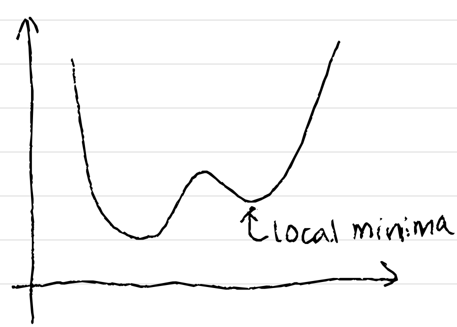
Which gives us a nonlinear set of equations:

$$\sum_{k=1}^{n} \left(f(x_k, \beta) - y_k \right) \frac{\partial f}{\partial \beta_j} = 0, \quad \text{for } j = 1, 2, \ldots, m$$

There is no general method to solve such systems, and normally we use some iterative scheme.



Convex function



nonconvex function

* Gradient Descent

The goal for gradient descent is to find values of x that satisfies $\nabla f(x) = 0$

(Note: in high dimensional systems, we need to test whether the zero gradient term is a maximum or minimum)

To illustrate the procedures of gradient descent, let's consider

$$f(x, y) = x^2 + 3y^2$$

which has a single minimum at $(x, y) = (0, 0)$

The gradient for $f(x, y)$ is

$$\nabla f(x) = \frac{\partial f}{\partial x}\hat{x} + \frac{\partial f}{\partial y}\hat{y} = 2x\hat{x} + 6y\hat{y}$$

Note: gradient does not point at the minimum, but gives a local steepest path to minimize $f(x)$.

To get to the next step, we have

$$x_{k+1}(\delta) = x_k - \delta \nabla f(x_k)$$

Note: δ is not a parameter we choose arbitrarily, rather, we compute it to make sure we are going "down hill" the optimal way.

To compute the optimal δ, we consider

$$F(\delta) = f(x_{k+1}(\delta))$$

Now we find δ that minimizes $F(\delta)$ by solving $\partial F / \partial \delta = 0$

$$\frac{\partial F}{\partial \delta} = -\nabla f(x_{k+1}) \nabla f(x_k) = 0$$

• Geometric interpretation

we want $\nabla f(x_k)$ and $\nabla f(x_{k+1})$ to be orthogonal to each other

This method of finding the optimal path is a special case of gradient descent, and its name is steepest descent.

Going back to the example $f(x,y) = x^2 + 3y^2$, we can compute δ for steepest descent.

$$x_{k+1} = x_k - \delta \nabla f(x_k) = x_k \hat{x} + y_k \hat{y} - \delta(2x_k \hat{x} + 6y_k \hat{y})$$

$$= (1-2\delta)x_k \hat{x} + (1-6\delta) y_k \hat{y}$$

Putting this value in $F(\delta)$, we have

$$F(\delta) = f(x_{k+1}(\delta)) = (1-2\delta)^2 x^2 + 3(1-6\delta)^2 y^2$$

$$\frac{\partial F(\delta)}{\partial \delta} = 2(1-2\delta)\cdot(-2)x^2 + 6(1-6\delta)\cdot(-6)y^2$$

$$= -4x^2 + 8\delta x^2 - 36y^2 + 216\delta y^2 = 0$$

Rearrange the equation, we get

$$(2x^2 + 54y^2)\delta = x^2 + 9y^2$$

$$\delta = \frac{x^2 + 9y^2}{2x^2 + 54y^2}$$

Note: δ is updated for every new pair of (x_k, y_k)