

# **6.252 NONLINEAR PROGRAMMING**

## **LECTURE 6**

### **NEWTON AND GAUSS-NEWTON METHODS**

#### **LECTURE OUTLINE**

- Newton's Method
- Convergence Rate of the Pure Form
- Global Convergence
- Variants of Newton's Method
- Least Squares Problems
- The Gauss-Newton Method

# NEWTON'S METHOD

$$x^{k+1} = x^k - \alpha^k (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

assuming that the Newton direction is defined and is a direction of descent

- Pure form of Newton's method (stepsize = 1)

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

- Very fast when it converges (how fast?)
- May not converge (or worse, it may not be defined) when started far from a nonsingular local min
- Issue: How to modify the method so that it converges globally, while maintaining the fast convergence rate

# CONVERGENCE RATE OF PURE FORM

- Consider solution of nonlinear system  $g(x) = 0$  where  $g : \mathbb{R}^n \mapsto \mathbb{R}^n$ , with method

$$x^{k+1} = x^k - (\nabla g(x^k)')^{-1} g(x^k)$$

– If  $g(x) = \nabla f(x)$ , we get pure form of Newton

- **Quick derivation:** Suppose  $x^k \rightarrow x^*$  with  $g(x^*) = 0$  and  $\nabla g(x^*)$  is invertible. By Taylor

$$0 = g(x^*) = g(x^k) + \nabla g(x^k)'(x^* - x^k) + o(\|x^k - x^*\|).$$

Multiply with  $(\nabla g(x^k)')^{-1}$ :

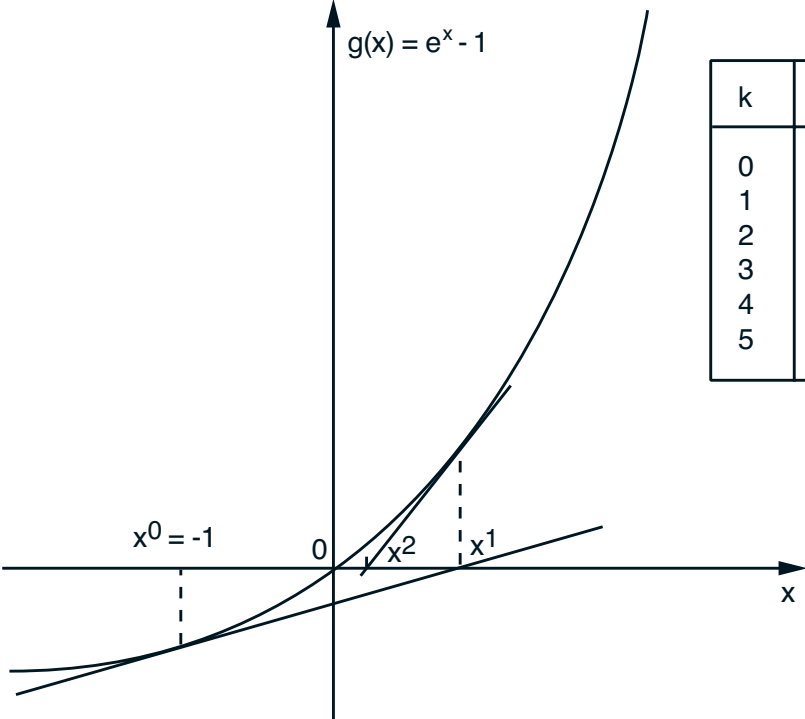
$$x^k - x^* - (\nabla g(x^k)')^{-1} g(x^k) = o(\|x^k - x^*\|),$$

so

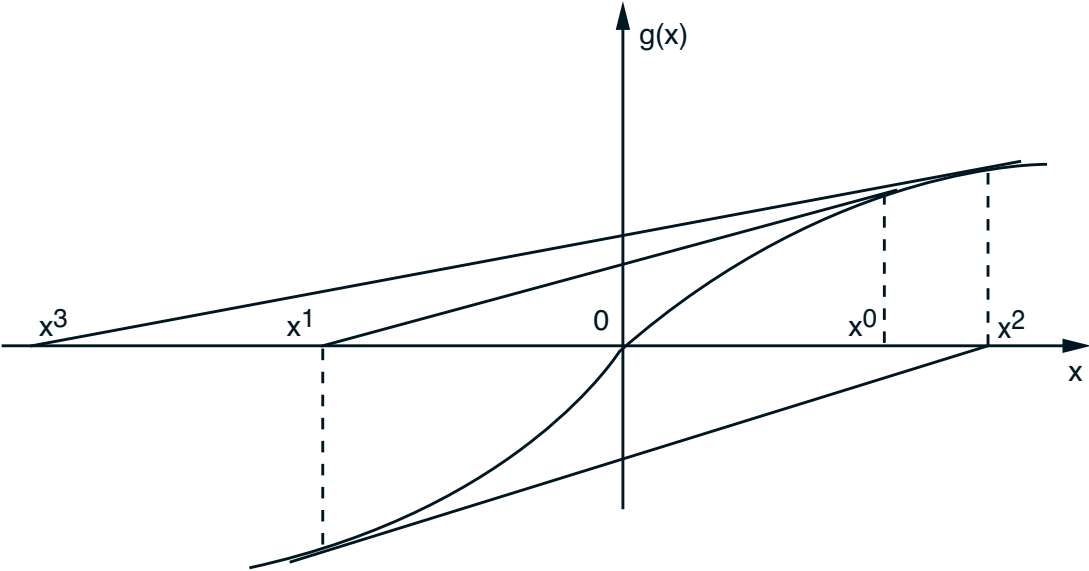
$$x^{k+1} - x^* = o(\|x^k - x^*\|),$$

implying superlinear convergence and capture.

# CONVERGENCE BEHAVIOR OF PURE FORM



k	$x^k$	$g(x^k)$
0	- 1.00000	- 0.63212
1	0.71828	1.05091
2	0.20587	0.22859
3	0.01981	0.02000
4	0.00019	0.00019
5	0.00000	0.00000



# MODIFICATIONS FOR GLOBAL CONVERGENCE

- Use a stepsize
- Modify the Newton direction when:
  - Hessian is not positive definite
  - When Hessian is nearly singular (needed to improve performance)
- Use

$$d^k = -(\nabla^2 f(x^k) + \Delta^k)^{-1} \nabla f(x^k),$$

whenever the Newton direction does not exist or is not a descent direction. Here  $\Delta^k$  is a diagonal matrix such that

$$\nabla^2 f(x^k) + \Delta^k > 0$$

- Modified Cholesky factorization
- Trust region methods

# LEAST-SQUARES PROBLEMS

$$\text{minimize } f(x) = \frac{1}{2} \|g(x)\|^2 = \frac{1}{2} \sum_{i=1}^m \|g_i(x)\|^2$$

subject to  $x \in \mathbb{R}^n$ ,

where  $g = (g_1, \dots, g_m)$ ,  $g_i : \mathbb{R}^n \rightarrow \mathbb{R}^{r_i}$ .

- Many applications:
  - Solution of systems of  $n$  nonlinear equations with  $n$  unknowns
  - Model Construction – Curve Fitting
  - Neural Networks
  - Pattern Classification

# PURE FORM OF THE GAUSS-NEWTON METHOD

- Idea: Linearize around the current point  $x^k$

$$\tilde{g}(x, x^k) = g(x^k) + \nabla g(x^k)'(x - x^k)$$

and minimize the norm of the linearized function  $\tilde{g}$ :

$$\begin{aligned} x^{k+1} &= \arg \min_{x \in \mathbb{R}^n} \frac{1}{2} \|\tilde{g}(x, x^k)\|^2 \\ &= x^k - (\nabla g(x^k) \nabla g(x^k)')^{-1} \nabla g(x^k) g(x^k) \end{aligned}$$

- The direction

$$-(\nabla g(x^k) \nabla g(x^k)')^{-1} \nabla g(x^k) g(x^k)$$

is a descent direction since

$$\nabla g(x^k) g(x^k) = \nabla ((1/2) \|g(x)\|^2)$$

$$\nabla g(x^k) \nabla g(x^k)' > 0$$

# MODIFICATIONS OF THE GAUSS-NEWTON

- Similar to those for Newton's method:

$$x^{k+1} = x^k - \alpha^k (\nabla g(x^k) \nabla g(x^k)' + \Delta^k)^{-1} \nabla g(x^k) g(x^k)$$

where  $\alpha^k$  is a stepsize and  $\Delta^k$  is a diagonal matrix such that

$$\nabla g(x^k) \nabla g(x^k)' + \Delta^k > 0$$

- Incremental version of the Gauss-Newton method:
  - Operate in cycles
  - Start a cycle with  $\psi_0$  (an estimate of  $x$ )
  - Update  $\psi$  using a *single* component of  $g$

$$\psi_i = \arg \min_{x \in \mathbb{R}^n} \sum_{j=1}^i \|\tilde{g}_j(x, \psi_{j-1})\|^2, \quad i = 1, \dots, m,$$

where  $\tilde{g}_j$  are the linearized functions

$$\tilde{g}_j(x, \psi_{j-1}) = g_j(\psi_{j-1}) + \nabla g_j(\psi_{j-1})'(x - \psi_{j-1})$$



# MODEL CONSTRUCTION

- Given set of  $m$  input-output data pairs  $(y_i, z_i)$ ,  $i = 1, \dots, m$ , from the physical system
- Hypothesize an input/output relation  $z = h(x, y)$ , where  $x$  is a vector of unknown parameters, and  $h$  is known
- Find  $x$  that matches best the data in the sense that it minimizes the sum of squared errors

$$\frac{1}{2} \sum_{i=1}^m \|z_i - h(x, y_i)\|^2$$

- Example of a linear model: Fit the data pairs by a cubic polynomial approximation. Take

$$h(x, y) = x_3 y^3 + x_2 y^2 + x_1 y + x_0,$$

where  $x = (x_0, x_1, x_2, x_3)$  is the vector of unknown coefficients of the cubic polynomial.

# NEURAL NETS

- Nonlinear model construction with multilayer perceptrons
- $x$  of the vector of weights
- Universal approximation property

# PATTERN CLASSIFICATION

- Objects are presented to us, and we wish to classify them in one of  $s$  categories  $1, \dots, s$ , based on a vector  $y$  of their features.
- Classical maximum posterior probability approach: Assume we know

$$p(j|y) = P(\text{object w/ feature vector } y \text{ is of category } j)$$

Assign object with feature vector  $y$  to category

$$j^*(y) = \arg \max_{j=1, \dots, s} p(j|y).$$

- If  $p(j|y)$  are unknown, we can estimate them using functions  $h_j(x_j, y)$  parameterized by vectors  $x_j$ . Obtain  $x_j$  by minimizing

$$\frac{1}{2} \sum_{i=1}^m (z_j^i - h_j(x_j, y_i))^2,$$

where

$$z_j^i = \begin{cases} 1 & \text{if } y_i \text{ is of category } j, \\ 0 & \text{otherwise.} \end{cases}$$