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

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

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

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

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

  - 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 $g : \Re^n \mapsto \Re^n$, with method

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

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

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

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

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

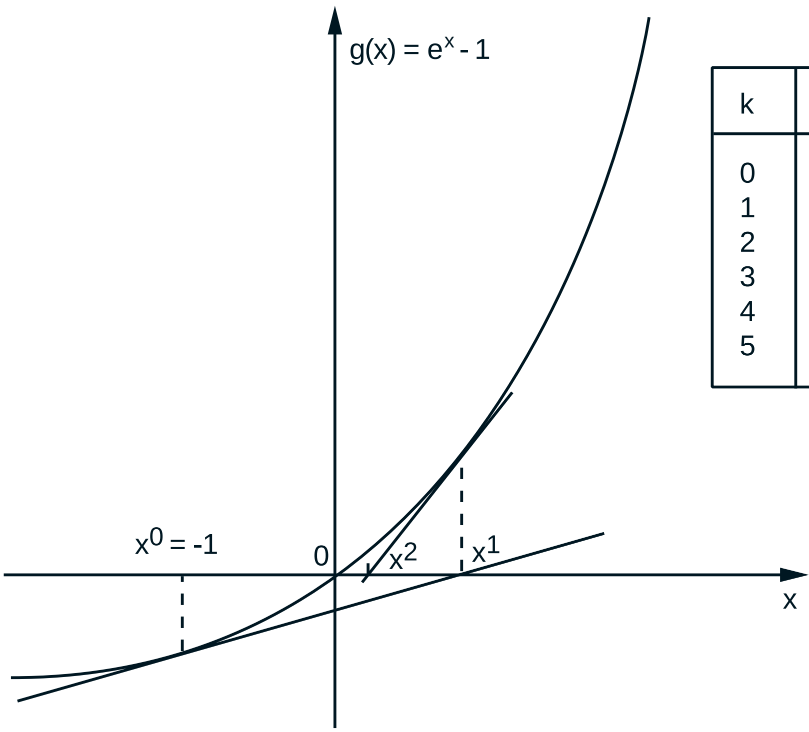
$$x^k - x^* - \big(\nabla g(x^k)'\big)^{-1} g(x^k) = o\big(\|x^k - x^*\|\big),$$

so

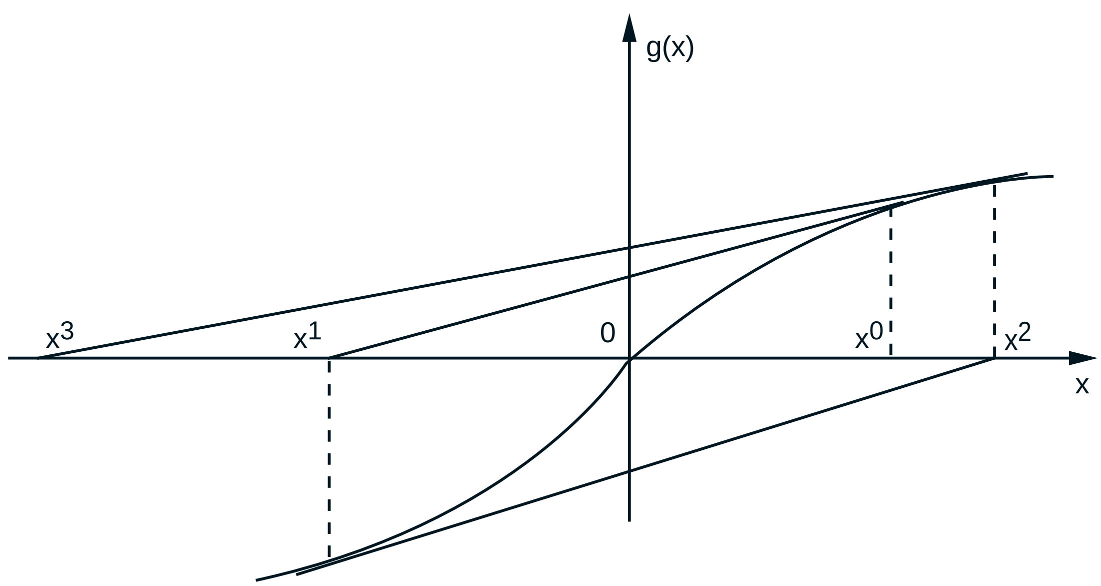
$$x^{k+1} - x^* = o\big(\|x^k - x^*\|\big),$$

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

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

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

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

  - Modified Cholesky factorization
  - Trust region methods

LEAST-SQUARES PROBLEMS

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

$$\text{subject to} \quad x \in \Re^n,$$

where $g = (g_1, \ldots, g_m)$, $g_i : \Re^n \to \Re^{r_i}$.

- Many applications:
  - Model Construction – Curve Fitting
  - Neural Networks
  - Pattern Classification

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 $\tilde{g}$:

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

- The direction

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

is a descent direction since

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

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

MODIFICATIONS OF THE GAUSS-NEWTON

- Similar to those for Newton's method:

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

where α^k is a stepsize and Δ^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 ψ_0 (an estimate of x)
  - Update ψ using a $single$ component of g

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

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), $i = 1, \ldots, m$, 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 $1, \ldots, s$, 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

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

- 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

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

where

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