Lecture 15

Global Methods

October 31, 2007

Outline

- Path Newton Method
- Differentiability Concepts
- Stationarity Concepts
- Line Search Algorithm

Path Newton Method

Step 0 Select initial vector x^0 and $\gamma \in (0,1)$. Set $k = 0$.

Step 1 If $G(x^k) = 0$, then stop.

Step 2 Select an approximation $A(x^k, \cdot) \in \mathcal{A}(x^k)$ and **the corresponding path** p^k with the domain defined by the scalar $\bar{\tau}_k = \min\{1, \epsilon/\|G(x^k)\|\}$ [cf. comments after Prop. 8.1.6].

Step 2(i) Set $i = 0$.

Step 2(ii) If

$$\|G(p^k(\bar{\tau}_k/2^i))\| \leq \left(1 - \frac{\gamma\,\bar{\tau}_k}{2^i}\right)\|G(x^k)\|,$$

set $t_k = \frac{\bar{\tau}_k}{2^i}$, $i_k = i$, and go to Step 3. Otherwise, set $i =: i+1$ and go to Step 2(ii).

Step 3 Set $x^{k+1} = p^k(t_k)$ and $k =: k+1$, and go to Step 1.

Convergence of the Path Newton Method

Theorem 8.1.10. Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz (continuous) on $\mathbb{R}^n$, and let G have a **nonsingular uniform Newton approximation** $\mathcal{A}$ on $\mathbb{R}^n$. Then, the following statements are true:

(1) The map G has a unique zero.

(2) For any x^0, the sequence $\{x^k\}$ generated by the Path Newton method converges Q-superlinearly to the zero of G.

(3) If $\mathcal{A}$ is **nonsingular uniform strong Newton approximation** on $\mathbb{R}^n$, then $\{x^k\}$ converges **Q-quadratically**.

Recall that due to the "sufficient decrease" condition at Step 2, the algorithm generates decreasing sequence $\{\|G(x^k)\|\}$. Hence, for all k,

$$x^k \in S_0, \qquad S_0 = \left\{ z \in \mathbb{R}^n \mid \|G(z)\| \leq \|G(x^0)\| \right\}$$

This suggests a possibility of relaxing the requirement that G has a nonsingular uniform approximation at every point in $\mathbb{R}^n$.

Convergence for Less Restrictive Condition

Theorem 8.1.11. Let $G : \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz on $\mathbb{R}^n$, and let $x^0 \in \mathbb{R}^n$ be given. Assume that G has a Newton approximation at every $x \in S_0$. Also, assume that the approximations are **uniformly nonsingular**, i.e., there exist ϵ and L' such that that for each $x \in S_0$ and every $A(x, \cdot) \in \mathcal{A}(x)$,

(a) There are two open sets U_x and V_x with $B(0, \epsilon) \subseteq U_x$ and $B(0, \epsilon) \subseteq V_x$

(b) $A(x, \cdot)$ is Lipschitz homeomorphism mapping of U_x onto V_x with L' being the Lipschitz constant for the inverse of $A(x, \cdot)\ |_{U_x}$ (the map restricted to U_x)

Then, the following holds for the iterates generated by the Path Method

(1) The sequence $\{\|G(x^k)\|\}$ converges to 0.

(2) If the set S_0 is bounded, then the results of Theorem 8.1.10 hold.

Comments on Path Newton Algorithm

- It provides a way of globalizing the local Newton Method
- Major deficiency is its restrictive convergence property
  - Requires nonsingular Newton approximation at every point visited by the path
- Alternative approaches of globalizing the local Newton Method exist that circumvent the nonsingularity requirement
- We discuss two of them, both of which are
  - Based on the use of merit functions
  - Minimizing a nonnegative merit function $\theta(x)$ satisfying

$$\theta(x) = 0 \qquad \textbf{if and only if} \qquad G(x) = 0$$

- Distinctive feature of the algorithms considered here:
  - Minimizing a **nonnegative function** with **zero optimal value**

Differentiability Concepts

Let $U \subseteq \mathbb{R}^n$ be open and let $\bar{x} \in U$. Let $\theta : U \to \mathbb{R}$.

Definition: The (upper) $D(ini)$ **derivative** of θ at $\bar{x}$ along d is given by

$$\theta^D(\bar{x}; d) = \limsup_{t \downarrow 0} \frac{\theta(\bar{x} + td) - \theta(\bar{x})}{t}.$$

- This "limsup" is well defined for any function, but not necessarily finite
- When θ is locally Lipschitz at $\bar{x}$, the D-derivative is finite for every d

Recall that $C(larke)$-**derivative** of θ at $\bar{x}$ along direction d is given by

$$\theta^\circ(\bar{x}; d) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{\theta(y + td) - \theta(y)}{t}$$

Definition: A function θ is $B(ouligand)$-**differentiable** at $\bar{x}$, if θ is locally Lipschitz and directionally differentiable at $\bar{x}$.

- When θ is B-differentiable at $\bar{x}$, we have

$$\theta^D(\bar{x}; d) = \theta'(\bar{x}; d) \qquad \text{for all } d$$

- When θ is C-regular at $\bar{x}$, we have

$$\theta^{\circ}(\bar{x}; d) = \theta'(\bar{x}; d) = \theta^{D}(\bar{x}; d) \qquad \text{for all } d \in \mathbb{R}^n$$

Recall, that by definition θ is C-regular at $\bar{x}$ when

- θ is locally Lipschitz and directionally differentiable at $\bar{x}$ (equivalent to B-differentiable at $\bar{x}$)

- The equality $\theta^{\circ}(\bar{x}; d) = \theta'(\bar{x}; d)$ holds for all $d \in \mathbb{R}^n$

Stationarity Concepts

Stationarity concept here ties to the following minimization problem

$$\begin{array}{ll} \text{minimize} & \theta(x) \\ \text{subject to} & x \in X \end{array} \tag{1}$$

Definition Given a closed convex set $X \subset \mathbb{R}^n$, a function $\theta : X \to \mathbb{R}$, and a vector $\bar{x} \in X$.

- When θ is differentiable at $\bar{x}$ and

$$\nabla\theta(\bar{x})^T(x - \bar{x}) \geq 0 \qquad \text{for all } x \in X,$$

the point $\bar{x}$ is said to be a **stationary point** for problem (1)

- When θ is B-differentiable at $\bar{x}$ and

$$\theta'(\bar{x}; x - \bar{x}) \geq 0 \qquad \text{for all } x \in X,$$

the point $\bar{x}$ is said to be a **B-stationary point**

- When

$$\theta^D(\bar{x}; x - \bar{x}) \geq 0 \qquad \text{for all } x \in X,$$

the point $\bar{x}$ is said to be **D-stationary point**

- When

$$\theta^\circ(\bar{x}; x - \bar{x}) \geq 0 \qquad \text{for all } x \in X,$$

the point $\bar{x}$ is said to be **C-stationary point**

When X is **not convex**, these relations reduce, respectively, to

$$\nabla\theta(\bar{x})^T y \geq 0 \quad \text{for all } y \in T_X(\bar{x}) \quad \textbf{stationary}$$

$$\theta'(\bar{x}; y) \geq 0 \quad \text{for all } y \in T_X(\bar{x}) \quad \textbf{B-stationary}$$

$$\theta^D(\bar{x}; y) \geq 0 \quad \text{for all } y \in T_X(\bar{x}) \quad \textbf{D-stationary}$$

$$\theta^\circ(\bar{x}; y) \geq 0 \quad \text{for all } y \in T_X(\bar{x}) \quad \textbf{C-stationary}$$

- It can be seen that

$$\theta^D(\bar{x}; d) \leq \theta^o(\bar{x}, d) \qquad \text{for all } d \in \mathbb{R}^n$$

- Thus, every D-stationary point is also C-stationary

- When θ is **C-regular**, both D- and C-stationarity reduce to B-stationarity (the stationarity in terms of directional derivatives), i.e.,

$$\theta'(\bar{x}; x - \bar{x}) \geq 0 \qquad \text{for all } x \in X \qquad \textbf{convex } X$$

$$\theta'(\bar{x}; y) \geq 0 \qquad \text{for all } y \in T_X(\bar{x})$$

Local Minima Characterization

Proposition 8.2.1 Let $X \subseteq \mathbb{R}^n$ be a closed set, $x^* \in X$, and $\theta : X \to \mathbb{R}$ be locally Lipschitz at x^*.

- If x^* is a local minimum of problem (1), then x^* is a D-stationary point
- If θ is B-differentiable at x^* and

$$\theta'(x^*; d) > 0 \qquad \text{for all } d \in T_X(x^*) \text{ with } \ d \neq 0,$$

then x^* is a strict local minimum of problem (1)

Proof: homework 4 assignment

Line Search Methods: Purpose and Idea

- We consider line search methods for solving constrained problem (1)
- Combined with $\theta(x) = \|G(x)\|^2$, they closely relate to solving the constrained system $G(x) = 0$ with $x \in X$
- **Idea** of line search methods
  - Starting with some initial x^0, generate a sequence according to update rule of the form

$$x^{k+1} = x^k + \tau_k \, d^k$$

    - $d^k \neq 0$ is a search direction and $\tau_k > 0$ is a stepsize
  - We want d^k and τ_k such that $\theta(x^k)$ decreases sufficiently with each iteration k
  - This is needed to ensure good convergence properties
  - The **goal** is to force $\{x^k\}$ to converge to a stationary point of θ
  - Additional "regularity" conditions are needed to ensure that a stationary point is a zero of the map G

Algorithm outline for $\theta \in C^1$ and $X = \mathbb{R}^n$

- Given $\gamma \in (0, 1)$
- Given x^k with $\nabla\theta(x^k) \neq 0$, consider the points $x^k + \tau d^k$ with $\tau \in (0, 1]$
- If d^k is a **descent direction**, i.e.,

$$\nabla\theta(x^k)^T d^k < 0$$

we can find $\bar{\tau} \in (0, 1]$ such that

$$\theta(x^k + \tau d^k) \leq \theta(x^k) + \gamma\,\tau\,\nabla\theta(x^k)^T d^k \qquad \text{for all } \tau \in (0, \bar{\tau}]$$

  - Existence of such $\bar{\tau}$ follows from the first-order expansion
- We choose τ_k to be one of such τ's, so that

$$\theta(x^{k+1}) \leq \theta(x^k) - \gamma\,\tau_k\,\sigma(x^k, d^k) \qquad \text{with } \sigma(x, d) = -\nabla\theta(x)^T d$$

- There are other descent directions, such as

$$d^k = -H^k \nabla\theta(x^k)$$

  - H^k is a symmetric positive definite matrix
  - $H^k = I$ corresponds to the classical "steepest-descent" direction, $d^k = -\nabla\theta(x^k)$

- When $\theta \in C^2$ and the Hessian $\nabla^2\theta(x)$ is positive definite at $x = x^k$, we can use

$$H^k = \nabla^2\theta(x^k),$$

  which corresponds to the Newton direction

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

Line Search Method for Nondifferentiable Function

- Similar to the method for a differentiable function
  - At every iteration the function is decreased sufficiently
  - The sufficient decrease is defined in terms of "forcing function" σ
  - The role of the forcing function is to ensure that the limit points of $\{x^k\}$ are stationary points of some sort

- In the presence of the constraint set $X \subset \mathbb{R}^n$, the direction d^k is restricted so that
  - The points $x^k + \tau d^k$ are feasible, i.e.,

$$x^k + \tau d^k \in X \qquad \text{for all } \tau \in (0, 1]$$

  - When X is convex, this is equivalent to the condition

$$d^k \in X - x^k$$

Admissible Directions and Forcing Function

- Line search algorithms have two key ingredients
  - **Admissible directions** playing a role similar to that of approximations in Newton algorithms
  - **Forcing functions** ensuring convergence to desirable points, i.e., stationary points of some type for problem (1)

For a convex X and each $x \in X$, the set of all feasible descent directions is the set of the directions d satisfying

$$d \in X - x, \qquad \theta^D(x; d) < 0$$

A set of admissible directions $\mathcal{D}(x)$ is a subset of feasible descent directions

$$\mathcal{D}(x) \subseteq \{d \mid d \in X - x, \quad \theta^D(x; d) < 0\}$$

This set is empty when x is a D-stationary point of the problem (1).

$\mathcal{D}$ is viewed as multi-valued mapping, $\mathcal{D} : X \to \mathbb{R}^n$.

Descent Lemma

The following lemma identifies the relation between forcing function and admissible directions central to convergence of line search algorithms

Lemma 8.3.1 Let $X \subseteq \mathbb{R}^n$ be convex and $\theta : X \to \mathbb{R}$ be locally Lipschitz. Let $(x, d) \in \mathrm{gph}\mathcal{D}$. Then, for every $\gamma \in (0, 1)$ and any scalar $\sigma > 0$ satisfying

$$\sigma \leq -\theta^D(x; d), \tag{2}$$

there exists $\bar{\tau} \in (0, 1]$ such that

$$\theta(x + \tau d) \leq \theta(x) - \gamma\tau\sigma \qquad \text{for all } \tau \in [0, \bar{\tau}].$$

Proof. To arrive at a contradiction, assume the contrary. Then, there is a sequence $\{\tau_k\}$ such that $\tau_k \downarrow 0$ and

$$\theta(x + \tau_k d) - \theta(x) > -\gamma\tau_k\sigma \qquad \text{for all } k$$

Dividing by τ_k and letting $k \to \infty$, we obtain $\theta^D(x; d) \geq -\gamma\sigma$. Since $\gamma \in (0, 1)$, it follows that $\theta^D(x; d) \geq -\sigma$, contradicting relation (2).

General Line Search Algorithm

Applicable to problem (1) with convex X and locally Lipschitz function θ.

- Each realization of the algorithm requires specification of the **line search map** $\mathcal{D}$ and the **forcing function** $\sigma : \mathbb{R}^n \times \mathbb{R}^n \to (0, \infty)$.

Step 0 Choose $x^0 \in X$ and $\gamma \in (0, 1)$. Set $k = 0$.

Step 1 If x^k is D-stationary, stop.

Step 2 Choose a vector d^k from the set $\mathcal{D}(x^k)$. Set $i = 0$.

Step 2(i) If

$$\theta(x^k + d^k/2^i) \leq \theta(x^k) - \gamma \frac{\sigma(x^k, d^k)}{2^i},$$

set $i_k = i$, and $\tau_k = 2^i$. Otherwise, set $i := i + 1$ and repeat Step 2(i).

Step 3 Set $x^{k+1} = x^k + \tau_k d^k$, $k := k + 1$ and go to Step 1.

Comments

- The algorithms implements Armijo stepsize rule
- As in the Path Algorithm, we can use any backtracking factor $\rho \in (0, 1)$
- Each iteration x^{k+1} stays in the (convex) set X
- Lemma 8.3.1 ensures that Step 2(ii) is exited after a finite number of iterations
- The objective value sequence $\{\theta(x^k)\}$ is decreasing
- When Step 2(ii) is exited with $i_k > 1$, at the preceding substep, we have

$$\theta(x^k + d^k/2^{i_k-1}) > \theta(x^k) - \gamma \frac{\sigma(x^k, d^k)}{2^{i_k-1}}$$

Forcing Function Convergence Properties

Theorem 8.3.3. Let $X \subseteq \mathbb{R}^n$ be a convex set and let θ be locally Lipschitz function on X. Let $\{x^k\}$ be a sequence generated by the Line Search Algorithm. Assume that $\{x^k \mid k \in \mathcal{K}\}$ is a subsequence with the following properties:

(1) The function value sequence $\{\theta(x^k) \mid k \in \mathcal{K}\}$ is bounded below.

(2) For every sequence of positive scalars $\{t_k \mid k \in \mathcal{K}\}$ converging to zero, there holds

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} \frac{\theta(x^k + t_k d^k) - \theta(x^k) + t_k \sigma(x^k, d^k)}{t_k} \leq 0$$

Then, we have

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \sigma(x^k, d^k) = 0.$$

Proof. The subsequence $\{\theta(x^k) \mid k \in \mathcal{K}\}$ is bounded below, so it has a limit point $\bar{v} \in \mathbb{R}$. Since the sequence $\{\theta(x^k)\}$ is decreasing, it must converge to $\bar{v}$. Thus,

$$\lim_{k\to\infty} \Big(\theta(x^{k+1}) - \theta(x^k)\Big) = 0.$$

By Step 2(ii) and the positivity of σ, γ, and τ_k, it follows that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \tau_k \, \sigma(x^k, d^k) = 0. \tag{3}$$

To arrive at a contradiction, assume that $\sigma(x^k, d^k)$ does not converge to zero. Then, there exists an index-subset $\mathcal{K}' \subseteq \mathcal{K}$ such that

$$\limsup_{\substack{k\to\infty \\ k\in\mathcal{K}'}} \sigma(x^k, d^k) > 0. \tag{4}$$

In view of Eq. (3), we have $\lim_{\substack{k\to\infty \\ k\in\mathcal{K}'}} \tau_k = 0$, implying that $i_k \to \infty$ over $k \in \mathcal{K}'$. Thus, for sufficiently large $k \in \mathcal{K}'$,

$$\frac{\theta(x^k + \tau'_k d^k) - \theta(x^k) + \tau'_k \sigma(x^k, d^k)}{\tau'_k} > (1-\gamma)\, \sigma(x^k, d^k).$$

Letting $k \to \infty$ over $k \in \mathcal{K}'$, and using the condition in part (2), we conclude that

$$0 \geq (1-\gamma) \limsup_{\substack{k\to\infty \\ k\in\mathcal{K}'}} \sigma(x^k, d^k),$$

contradicting relation(4).

- The preceding result provides us with a "tool" for analyzing the convergence properties of the line search algorithms

- For practical purposes, we need to understand how to ensure that conditions (1) and (2) hold

- Condition (1) holds for example when
  - The function θ is bounded below on X
  - The set X is bounded.
  - The level set $\{x \in X \mid \theta(x) \leq \theta(x^0)\}$ is bounded. Note that since $\theta(x^k)$ is decreasing we have

$$x^k \in \{x \in X \mid \theta(x) \leq \theta(x^0)\} \qquad \text{for all } k$$

    This set is bounded for example when the function θ is **coercive on the set X**, i.e..,

$$\lim_{\substack{\|x\| \to \infty \\ x \in X}} \theta(x) = \infty$$

- Ensuring that condition (2) holds is specific to the problem objective and the choice of forcing function
- We consider some realizations of the algorithm for which we prove that the condition (2) is valid
- The convergence results will fall in two types
  - **Sequential convergence**: establishing that the sequence $\{x^k\}$ converges to a stationary point of some sort
  - **Subsequential convergence**: establishing that all limit points of $\{x^k\}$ are stationary points of some sort
- We start with application of the line search algorithm to a B-differentiable function
- To apply the algorithm, we need to discuss the construction of "feasible descent directions"

Descent Directions: A Related Problem

$$\begin{array}{ll} \text{minimize} & \theta'(x;d) + \frac{1}{2}d^T H d \\ \text{subject to} & d \in X - x \end{array} \tag{5}$$

The problem is convex when θ is differentiable. It has important properties discussed in the following.

Proposition 8.3.4. Let $X \subseteq \mathbb{R}^n$ be closed convex set and $x \in X$. Let θ be B-differentiable and H be symmetric positive definite matrix. Then, for problem (5) the following is valid:

- There is an optimal solution.
- The optimal value is nonpositive.
- The optimal value is zero if and only if the point x is B-stationary

Proof. The objective function is continuous and coercive. Hence, an optimal solution exists. Since the objective value for $d = 0$ is zero, it follows that the optimal value is nonpositive.

Suppose the optimal value is zero. Then, we have

$$\theta'(x; d) + \frac{1}{2} d^T H d \geq 0 \qquad \text{for all } d \in X - x$$

By convexity of X, the point $x + \tau(y - x) \in X$ for all $\tau \in [0, 1]$, so that $\tau(y - x) \in X - x$ for all $\tau \in [0, 1]$. Therefore,

$$\tau\theta'(x; y - x) + \frac{\tau^2}{2}(y - x)^T H (y - x) \geq 0 \qquad \text{for all } y \in X$$

Dividing by τ and letting $\tau = 0$, we see that

$$\theta'(x; y - x) \geq 0 \qquad \text{for all } y \in X,$$

showing that x is B-stationary.

Technical Lemma

Lemma 8.3.5. Let $X \subseteq \mathbb{R}^n$ be closed convex set and θ be B-differentiable. Let $\{x^j\}$ be a convergent subsequence of $\{x^k\}$ (consisting of nonstationary points). Let $\{H^j\}$ be a sequence of symmetric uniformly positive definite matrices, i.e.,

$$\inf_j \lambda_{\min}(H^j) > 0.$$

Then the union

$$\cup_j \{ \text{ solutions to problem (5) with } x = x^j \text{ and } H = H^j \}$$

is bounded.

Line Search for B-Differentiable Problems

Forcing function $\sigma(x; d) = -\theta'(x; d)$.

Step 0 Choose $x^0 \in X$ and $\gamma \in (0, 1)$. Set $k = 0$.

Step 1 If x^k is B-stationary point, stop.

Step 2 Choose a symmetric positive definite matrix H^k, and determine a vector d^k solving the problem (5) with $x = x^k$ and $H = H^k$. Set $i = 0$.

Step 2(i) If

$$\theta(x^k + d^k/2^i) \leq \theta(x^k) + \frac{\gamma}{2^i}\,\theta'(x^k; d^k),$$

set $i_k = i$, and $\tau_k = 2^i$. Otherwise, set $i := i + 1$ and repeat Step 2(i).

Step 3 Set $x^{k+1} = x^k + \tau_k d^k$, $k := k + 1$ and go to Step 1.

Convergence Result

Proposition 8.3.7. Let $X \subseteq \mathbb{R}^n$ be closed convex set and θ be B-differentiable on X. Let $\{x^k\}$ be a sequence generated by the Line Search Algorithm. Let $\{x^k \mid k \in \mathcal{K}\}$ be a subsequence such that

(a) There exist positive scalars c_1 and c_2 satisfying for every $k \in \mathcal{K}$,

$$c_1\|y\|^2 \leq y^T H^k y \leq c_2\|y\|^2 \qquad \text{for all } y \in \mathbb{R}^n$$

(b) The subsequence $\{x^k \mid k \in \mathcal{K}\}$ converges to a vector x^*

(c) The function θ has a strong F-derivative at x^*, i.e.,

$$\lim_{\substack{y \neq z \\ (y,z) \to (x^*,x^*)}} \frac{\mathrm{e}(y) - \mathrm{e}(z)}{\|y - z\|} = 0,$$

where $\mathrm{e}(y) = \theta(y) - \theta(x^*) - \nabla\theta(x^*)^T(y - x^*)$

Then, x^* is a B-stationary point of problem (1).

Proof sketch. Filling out the details - homework assignment.

- The strong F-differentiability at x^* implies that the directional derivative $\theta'(x^*; \cdot)$ is continuous function on $\mathbb{R}^n$.

- Using Lemma 8.3.5, we show that the condition (2) of Theorem 8.3.3 holds and that the sequence $\{d^k \mid k \in \mathcal{K}\}$ is bounded.

- By Theorem 8.3.3 where $\sigma(x, d) = -\theta'(x; d)$, we see that

$$\lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \sigma(x^k, d^k) = 0.$$

  Here in particular, we have $d^k \to d^\infty$, so that

$$\nabla\theta(x^*)^T d^\infty = 0.$$

- Using the preceding relation and the property of the matrices $H^k, k \in \mathcal{K}$ given in condition (a), after some algebra, we conclude that

$$\theta'(x^*; y - x^*) \geq 0 \qquad \text{for all } y \in X.$$