Lecture 15

Newton Method and Self-Concordance

October 23, 2008

Outline

- Self-concordance Notion
- Self-concordant Functions
- Operations Preserving Self-concordance
- Properties of Self-concordant Functions
- Implications for Newton's Method
- Equality Constrained Minimization

Classical Convergence Analysis of Newton's Method

- Given a level set $L_0 = \{x \mid f(x) \le f(x_0)\}$, it requires for all $x, y \in L_0$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x - y\|^2 \qquad mI \preceq \nabla^2 f(x) \preceq MI$$

  for some constants $L > 0$, $m > 0$, and $M > 0$

- Given a desired error level ϵ, we are interested in an ϵ-solution of the problem i.e., a vector $\tilde{x}$ such that $f(\tilde{x}) \le f^* + \epsilon$
- An upper bound on the number of iterations to generate an ϵ-solution is given by

$$\frac{f(x_0) - f^*}{\gamma} + \log_2 \log_2 \left(\frac{\epsilon_0}{\epsilon}\right)$$

  where $\gamma = \sigma\beta\eta^2 \frac{m}{M^2}$, $\eta \in (0, m^2/L)$, and $\epsilon_0 = 2m^3/L^2$

- This follows from

$$\frac{L}{2m^2}\|\nabla f(x_{k+K})\| \le \left(\frac{L}{2m^2}\|\nabla f(x_k)\|\right)^{2^K} \le \left(\frac{1}{2}\right)^{2^K}$$

where k is such that $\|\nabla f(x_k)\| \le \eta$ and η is small enough so that $\eta\frac{L}{2m^2} \le \frac{1}{2}$. By the above relation and strong convexity of f, we have

$$f(x_k) - f^* \le \frac{1}{2m}\|\nabla f(x_k)\|^2 \le \frac{2m^3}{L^2}\left(\frac{1}{2}\right)^{2^K}$$

- The bound is conceptually informative, but not practical
- Furthermore, the constants L, m, and M change with affine transformations of the space, while Newton's method is affine invariant
- Can a bound be obtained in terms of problem data that is affine invariant and, moreover, practically verifiable?

Self-concordance

- Nesterov and Nemirovski introduced a notion of *self-concordance* and a class of *self-concordant functions*
- Importance of the self-concordance:
  - Possesses affine invariant property
  - Provides a new tool for analyzing Newton's method that exploits the affine invariance of the method
  - Results in a practical upper bound on the Newton's iterations
  - Plays a crucial role in performance analysis of interior point method

Def. A function $f : \mathbb{R} \mapsto \mathbb{R}$ is *self-concordant* when f is *convex* and

$$|f'''(x)| \le 2 f''(x)^{3/2} \qquad \text{for all } x \in \operatorname{dom} f$$

The rate of change in curvature of f is bounded by the curvature

Note: One can use a constant κ other than 2 in the definition

Examples

- Linear and quadratic functions are self-concordant $[f'''(x) = 0$ for all $x]$
- Negative logarithm $f(x) = -\ln x$, $x > 0$ is self-concordant:

$$f''(x) = \frac{1}{x^2}, \qquad f'''(x) = -\frac{2}{x^3}, \qquad \frac{|f'''(x)|}{f''(x)^{3/2}} = 2 \qquad \text{for all } x > 0$$

- Exponential function e^x is not self-concordant: $f''(x) = f'''(x) = e^x$,

$$\frac{|f'''(x)|}{f''(x)^{3/2}} = \frac{e^x}{e^{3x/2}} = e^{-x/2}, \qquad \frac{|f'''(x)|}{f''(x)^{3/2}} \to \infty \text{ as } x \to -\infty$$

- Even-power monomial x^{2p}, $p > 2$ is not self-concordant:

$f''(x) = 2p(2p-1)x^{2p-2}$, $f'''(x) = 2p(2p-1)(2p-2)x^{2p-3}$,

$$\frac{|f'''(x)|}{f''(x)^{3/2}} = \frac{p_3|x^{2p-3}|}{p_2 x^{3(p-1)}}, \qquad \frac{|f'''(x)|}{f''(x)^{3/2}} \to \infty \text{ as } x \downarrow 0$$

Scaling and Affine Invariance

- In the definition of the self-concordant function, one can use any other κ
- Let f be self-concordant with κ, then $\tilde{f}(x) = \frac{\kappa^2}{4} f(x)$ is self-concordant with $\kappa = 2$:

$$\frac{|\tilde{f}'''(x)|}{\tilde{f}''(x)^{3/2}} = \frac{8\kappa^2}{4\kappa^3} \frac{|f'''(x)|}{f''(x)^{3/2}} \leq 2$$

- Self-concordant function is *affine invariant*: for f self-concordant and $x = ay + b$, we have $\tilde{f}(y) = f(x)$ self-concordant with the same κ:
  - Convexity is preserved: $\tilde{f}$ is convex
  - $\tilde{f}'''(y) = a^3 f'''(x)$, $\tilde{f}''(y) = a^2 f''(x)$

$$\frac{|\tilde{f}'''(x)|}{\tilde{f}''(x)^{3/2}} = \frac{|a|^3}{|a|^3} \frac{|f'''(x)|}{f''(x)^{3/2}} \leq \kappa$$

Self-concordant Function in $\mathbb{R}^n$

Def. A function $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant when it is self-concordant along every line, i.e.,

- f is convex
- $g(t) = f(x + tv)$ is self-concordant for all $x \in \mathrm{dom} f$ and all v,

Note: The constant κ of self-concordance is independent of the choice of x and v, i.e.,

$$\frac{|g'''(t)|}{g''(t)^{3/2}} \leq 2$$

for all $x \in \mathrm{dom} f$, $v \in \mathbb{R}^n$, and $t \in \mathbb{R}$ such that $x + tv \in \mathrm{dom} f$

Operations Preserving Self-concordance

Note: To start with, these operations have to *preserve convexity*

- *Scaling with a positive factor of at least 1*: when f is self-concordant and $a > 1$, then af is also self-concordant
- The *sum* $f_1 + f_2$ of two self-concordant functions is self-concordant (extends to any finite sum)
- Composition with affine mapping: when $f(y)$ is self-concordant, then $f(Ax + b)$ is self-concordant
- Composition with ln-function: Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function with $\mathrm{dom} g = \mathbb{R}_{++}$, and

  $$|g'''(x)| \le 3\,\frac{g''(x)}{x} \qquad \text{for all } x > 0$$

  Then:

  $$f(x) = -\ln(-g(x)) - \ln(x) \qquad \text{over } \{x > 0 \mid g(x) < 0\}$$

  is self-concordant

HW7

Implications

- When $f''(x) > 0$ for all x (then f is strictly convex), the self-concordant condition can be written as

$$\left|\frac{d}{dx}\left(f''(x)^{-\frac{1}{2}}\right)\right| \le 1 \qquad \text{for all } x \in \mathrm{dom} f.$$

- Assuming that for some $t > 0$, the interval $[0, t]$ lies in $\mathrm{dom} f$, we can integrate from 0 to t, and obtain

$$-t \le \int_0^t \frac{d}{dx}\left(f''(x)^{-\frac{1}{2}}\right) dx \le t.$$

- This implies the following lower and upper bounds on $f''(x)$,

$$\frac{f''(0)}{\left(1 + t\sqrt{f''(0)}\right)^2} \le f''(t) \le \frac{f''(0)}{\left(1 - t\sqrt{f''(0)}\right)^2}, \tag{1}$$

where the lower bound is valid for all $t \ge 0$ with $t \in \mathrm{dom} f$, and the upper bound is valid for all $t \in \mathrm{dom} f$ with $0 \le t < f''(0)^{-1/2}$.

Bounds on f

Consider $f : \mathbb{R}^n \to \mathbb{R}$ with $\nabla^2 f(x) > 0$. Let v be a descent direction, i.e., such that $\nabla f(x)^T v < 0$. Consider $g(t) = f(x + tv) - f(x)$ for $t > 0$. Integrating the lower bound of Eq. (1) twice, we obtain

$$g(t) \geq g(0) + tg'(0) + t\sqrt{g''(0)} - \ln\left(1 + t\sqrt{g''(0)}\right)$$

Consider now $v = Newton's\ direction$, i.e., $v = [\nabla^2 f(x)]^{-1} \nabla f(x)$. Let $h(t) = f(x + tv)$. We have

$$h'(0) = \nabla f(x)v = -\lambda^2(x), \qquad h''(0) = v^T \nabla^2 f(0)v = \lambda^2(x).$$

Integrating the lower bound of Eq. (1) twice, we obtain for $0 \leq t < \frac{1}{\lambda(x)}$

$$h(t) \leq h(0) - t\lambda^2(x) - t\lambda(x) - \ln\left(1 - t\lambda(x)\right) \tag{2}$$

Newton Direction for Self-concordant Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be self-concordant and $\nabla^2 f(x) \succ 0$ for all $x \in L_0$

- Using self-concordance it can be seen that
  - For Newton's decrement $\lambda(x)$ and any $v \in \mathbb{R}^n$:

$$\lambda(x) = \sup_{v \neq 0} \frac{-v^T \nabla f(x)}{\left(v^T \nabla^2 f(x) v\right)^{1/2}}$$

with the supremum attainment at $v = -[\nabla^2 f(x)]^{-1} \nabla f(x)$

HW7

Self-Concordant Functions: Newton Method Analysis

- Consider Newton's method, started at x_0, with backtracking line search.
- Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant and strictly convex.
  (If $\mathrm{dom} f \neq \mathbb{R}^n$, assume the level set $\{x \mid f(x) \le f(x_0)\}$ is closed.
- Also, assume that f is bounded from below.

Under the preceding assumptions an optimizer x^* of f over $\mathbb{R}^n$ exists. HW7

- Analysis is similar to the classic one except
  - Self-concordance replaces strict convexity and Lipschitz Hessian assumptions
  - Newton decrement replaces the role of the gradient norm

Convergence Result

Theorem 1 *Assume that f is self-concordant strictly convex function that is bounded below over $\mathbb{R}^n$. Then, there exist $\eta \in (0, 1/4)$ and γ such that*

- **Damped phase** *If $\lambda_k > \eta$, we have $f(x_{k+1}) - f(x_k) \le -\gamma$.*
- **Pure Newton** *If $\lambda_k \le \eta$, the backtracking line search selects $\alpha = 1$ and*

$$2\lambda_{k+1} \le (2\lambda_k)^2$$

where $\lambda_k = \lambda(x_k)$.

Proof

Let $h(\alpha) = f(x_k + \alpha d_k)$. As seen in Eq. (2), we have for $0 \le \alpha < \frac{1}{\lambda_k}$

$$h(\alpha) \le h(0) - \alpha\lambda_k^2 - \alpha\lambda_k - \ln(1 - \alpha\lambda_k) \tag{3}$$

We use this relation to show that the stepsize $\widehat{\alpha} = \frac{1}{1+\lambda_k}$ satisfies the exit condition of the line search. In particular, letting $\alpha = \widehat{\alpha}$, we have

$$\begin{aligned} h(\widehat{\alpha}) &\le h(0) - \widehat{\alpha}\lambda_k^2 - \widehat{\alpha}\lambda_k - \ln(1 - \widehat{\alpha}\lambda_k) \\ &= h(0) - \lambda_k + \ln(1 + \lambda_k) \end{aligned}$$

Using the inequality $-t + \ln(1+t) \le -\frac{t^2}{2(1+t)}$ for $t \ge 0$, and $\sigma \le 1/2$, we obtain

$$h(\widehat{\alpha}) \le h(0) - \sigma \frac{\lambda_k^2}{1+\lambda_k} = h(0) - \sigma\widehat{\alpha}\lambda_k^2.$$

Therefore, at the exit of the line search, we have $\alpha_k \ge \frac{\beta}{1+\lambda_k}$, implying

$$h(\alpha_k) \le h(0) - \sigma\beta\frac{\lambda_k^2}{1+\lambda_k}.$$

When $\lambda_k > \eta$, since $h(\alpha_k) = f(x_{k+1})$ and $h(0) = f(x_k)$, we have

$$f(x_{k+1}) \le f(x_k) - \gamma, \qquad \gamma = \sigma\beta\frac{\eta^2}{1+\eta}.$$

Assume that $\lambda_k \le \frac{1-2\sigma}{2}$, from Eq. (3) and the inequality

$$-x - \ln(1-x) \le x^2/2 + x^3 \qquad \text{for } 0 \le x \le 0.81,$$

we have

$$h(1) \leq h(0) - \frac{1}{2}\lambda_k^2 + \lambda_k^3 \leq h(0) - \sigma\lambda_k^2,$$

where the last inequality follows from $\lambda_k \leq \frac{1-2\sigma}{2}$.
Thus, the unit step satisfies the sufficient decrease condition (no damping).

The proof that $2\lambda_{k+1} \leq (2\lambda_k)^2$ left for HW7.
The preceding material is from Chapter 9.6 of the book by Boyd and Vandenberghe.

Newton's Method for Self-concordant Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be self-concordant and $\nabla^2 f(x) \succ 0$ for all $x \in L_0$

- The analysis of Newton's method with backtracking line search:
  - For an ϵ-solution, i.e., $\tilde{x}$ with $f(\tilde{x}) \le f^* + \epsilon$
  - An upper bound on the number of iterations is given by

$$\frac{f(x_0) - f^*}{\Gamma} + \log_2 \log_2 \frac{1}{\epsilon}, \qquad \Gamma = \sigma\beta \frac{\eta^2}{1+\eta}, \quad \eta = \frac{1-2\sigma}{4}$$

  - Explicitly:

$$\frac{f(x_0) - f^*}{\Gamma} + \log_2 \log_2 \frac{1}{\epsilon} = \frac{20 - 8\sigma}{\sigma\beta(1-2\sigma)^2} \, [f(x_0) - f^*] + \log_2 \log_2 \frac{1}{\epsilon}$$

- The bound is practical (when f^* can be underestimated)

Equality Constrained Minimization

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & Ax = b \end{array}$$

Assumption:

- The function f is convex, twice continuously differentiable
- The matrix $A \in \mathbb{R}^{p \times n}$ has $\mathrm{Rank}\ A = p$
- Optimal value f^* is finite and attained
  - If $Ax = b$ for some $x \in \mathrm{relint}(\mathrm{dom} f)$, there is no duality gap and there exists an optimal dual solution
- **KKT Optimality Conditions** state x^* is optimal if and only if there exists λ^* such that

$$Ax^* = b, \qquad \nabla f(x^*) + A^T \lambda^* = 0$$

- Solving the problem is equivalent to solving the KKT conditions: $n + p$ equations in $n + p$ variables, $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^p$

Equality Constrained Quadratic Minimization

minimize $\frac{1}{2}x^T Px + q^T x + r$ with $P \in \mathcal{S}^n_+$
subject to $Ax = b$

Important in extending the Newton's method to equality constraints

KKT Optimality Conditions

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

- Coefficient matrix is referred to as $KKT\ matrix$
- The KKT matrix is nonsingular if and only if

$$Ax = 0,\ x \neq 0 \quad \Longrightarrow \quad x^T Px > 0 \qquad [P \text{ is psd over null space of } A]$$

- Equivalent conditions for nonsingularity of A:
  - $P + A^T A \succ 0$
  - $\mathcal{N}(A) \cap \mathcal{N}(P) = \{0\}$

Newton's Method with Equality Constraints

Almost the same as Newton's method, except for two differences:

- The **initial iterate x_0 has to be feasible**, $Ax_0 = b$
- The **Newton's directions d_k have to be feasible**, $Ad_k = 0$

Given iterate x_k, Newton's direction d_k is determined as a solution to

$$\begin{array}{ll} \text{minimize} & f(x_k) + \nabla f(x_k) d + \frac{1}{2} d^T \nabla^2 f(x_k) d \\ \text{subject to} & A(x_k + d) = b \end{array}$$

- The KKT conditions:(w_k is optimal dual for the quadratic minimization)

$$\begin{bmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d_k \\ w_k \end{bmatrix} = \begin{bmatrix} -\nabla f(x_k) \\ 0 \end{bmatrix}$$

Newton's Method with Equality Constraints

Given starting point $x \in \mathrm{dom} f$ with $Ax = b$ and tolerance $\epsilon > 0$.

Repeat

1. Compute Newton's direction $d_N(x)$ by solving the corresponding KKT system.
2. Compute the decrement $\lambda(x)$
3. *Stopping criterion.* **Quit** if $\lambda^2/2 \le \epsilon$.
4. *Line search.* Choose stepsize α by backtracking line search.
5. *Update.* $x := x + \alpha d_N(x)$.

- When f is strictly convex and self-concordant, **the bound on the number of iterations required to achieve an ϵ-accuracy** is the same as in the "unconstrained case"

Newton's Method with Infeasible Points

Useful when determining an initial feasible iterate x_0 is difficult

Linearizing optimality conditions $Ax^* = b$ and $\nabla f(x^*) + A^T\lambda^*$ at some $x + d \approx x^*$, with x is possibly infeasible, and using $w \approx \lambda^*$

- We have

$$\nabla f(x^*) \approx \nabla f(x+d) \approx \nabla f(x) + \nabla^2 f(x) d$$

- Approximate KKT

$$A(x+d) = b, \qquad \nabla f(x) + \nabla^2 f(x) d + A^T w = 0$$

- At **infeasible** x_k, the set of linear inequalities determining d_k and w_k:

$$\begin{bmatrix} \nabla^2 f(x_k) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d_k \\ w_k \end{bmatrix} = - \begin{bmatrix} \nabla f(x_k) \\ Ax_k - b \end{bmatrix}$$

Note: When x_k is feasible, the system of equations is the same as in the feasible point Newton's method

Equality Constrained Analytic Centering

$$\begin{array}{ll} \text{minimize} & f(x) = -\sum_{i=1}^{n} \ln x_i \\ \text{subject to} & Ax = b \end{array}$$

Feasible point Newton's method: $g = \nabla f(x)$, $H = \nabla^2 f(x)$

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ w \end{bmatrix} = \begin{bmatrix} -g \\ 0 \end{bmatrix}, \quad g = \begin{bmatrix} -\frac{1}{x_1} \\ \vdots \\ -\frac{1}{x_n} \end{bmatrix}, \quad H = \operatorname{diag}\left[\frac{1}{x_1^2}, \ldots, \frac{1}{x_n^2}\right]$$

- The Hessian is positive definite
- KKT matrix first row: $Hd + A^T w = -g \;\Rightarrow\; d = -H^{-1}(g + A^T w)$ (1)
- KKT matrix second row, $Ad = 0$, and Eq. (1) $\Rightarrow\; AH^{-1}(g + A^T w) = 0$
- The matrix A has full row rank, thus $AH^{-1}A^T$ is invertible, hence

$$w = -\left(AH^{-1}A^T\right)^{-1} AH^{-1} g, \qquad H^{-1} = \operatorname{diag}\left[x_1^2, \ldots, x_n^2\right]$$

- The matrix $-AH^{-1}A^T$ is known as *Schur complement of H* (any H)

Network Flow Optimization

$$\begin{array}{ll} \text{minimize} & \sum_{l=1}^{n} \phi_l(x_l) \\ \text{subject to} & Ax = b \end{array}$$

- Directed graph with n arcs and $p+1$ nodes
- Variable x_l: flow through arc l
- Cost ϕ_l: cost flow function for arc l, with $\phi_l''(t) > 0$
- Node-incidence matrix $\tilde{A} \in \mathbb{R}^{(p+1)\times n}$ defined as

$$\tilde{A}_{il} = \begin{cases} 1 & \text{arc } j \text{ originates at node } i \\ -1 & \text{arc } j \text{ ends at node } i \\ 0 & \text{otherwise} \end{cases}$$

- Reduced node-incidence matrix $A \in \mathbb{R}^{p\times n}$ is $\tilde{A}$ with last row removed
- $b \in \mathbb{R}^p$ is (reduced) source vector
- Rank $A = p$ when the graph is *connected*

KKT system for infeasible Newton's method

$$\begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} d \\ w \end{bmatrix} = - \begin{bmatrix} g \\ h \end{bmatrix}$$

where $h = Ax - b$ is a measure of infeasibility at the current point x

- $g = [\phi_1'(x_1), \ldots, \phi_n(x_n)]^T$
- $H = \mathrm{diag}\,[\phi_1''(x_1), \ldots, \phi_n''(x_n)]$ with positive diagonal entries
- Solve via elimination:

$$w = (AH^{-1}A^T)^{-1}[h - AH^{-1}g], \qquad d = -H^{-1}(g + A^T w)$$