Modern Theory of 2nd-Order Methods

Lecture 1: Global complexity bounds for 2nd-order methods. Systems of equations

Yurii Nesterov (CORE/INMA, UCLouvain)

Minicourse: August 16-20, 2021 (Zinal)

General Contents

Lecture 1: Global complexity bounds for 2nd-order methods. Systems of nonlinear equations.

Lecture 2: Accelerated second-order methods. Lower complexity bounds.

Lecture 3: Universal 2nd-order methods.

Lecture 4: Implementable Tensor Methods.

References, I

1. Yu. Nesterov, B. Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, **108**(1), 177-205 (2006).

2. Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, **112**(1) 159-181 (2008)

3. Yu. Nesterov. Modified Gauss-Newton scheme with worst-case guarantees for its global performance. *Optimization Methods and Software*, **22**(3) 469-483 (2007)

4. Yu. Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, **171**, 311330 (2018)

5. G.N. Grapiglia, Yu. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIOPT*, **27**(1), 478-506 (2017).

References, II

6. H.Lu, R.Freund, and Yu.Nesterov. Relatively smooth convex optimization by first-order methods. *SIOPT*, **28**(1), 333-354 (2018).

7. Yu. Nesterov. Superfast 2nd-order methods for unconstrained convex optimization. CORE Discussion Paper, Dec. 2019.

OR (instead of 1-5): sections 4.1-4.4, and section 6.4.6 of

Yu. Nesterov. *Lecture notes on Convex Optimization*. Springer (2018).

This Lecture

Historical remarks

Trust region methods

Cubic regularization of second-order model

Local and global convergence

Solving the system of nonlinear equations

Numerical experiments

Historical remarks

Problem: $f(x) \quad \to \quad \min: \quad x \in \mathbb{R}^n$

is replaced by a system of non-linear equations $\boxed{f'(x) = 0}$

Linearization: $f'(\bar{x}) + f''(\bar{x})(x_+ - \bar{x}) = 0.$

Newton method: $x_{k+1} = x_k - [f''(x_k)]^{-1} f'(x_k).$

Standard objections:

- The method can brake down (det $f''(x_k) = 0$).
- Possible divergence.
- Possible convergence to a saddle point or even to a local maximum.
- Possible chaotic global behavior.

Pre-History (see Ortega, Rheinboldt [1970])

- *Bennet [1916]:* first use of Newton's method in existence theorem.
- *Levenberg [1944]:* Regularization.

  If $f''(x) \not\succ 0$, then use $d = G^{-1} f'(x)$ with $G = f''(x) + \gamma I \succ 0$.

  (See also *Marquardt [1963]*.)
- *Kantorovich [1948]:* First proof of local quadratic convergence.

  Assumptions:

  a) $f \in C^3(\mathbb{R}^n)$.
  b) $\|f''(x) - f''(y)\| \leq L_2 \|x - y\|$.
  c) $f''(x^*) \succ 0$.
  d) $x_0 \approx x^*$.

Proof: Let $\|f''(x)u\| \geq \mu \|u\|$ for all u and x with $\mu > 0$. Then

$$\begin{aligned} \|f'(x_+)\| &= \|f'(x_+) - f'(x) - f''(x)(x_+ - x)\| \\ &\leq \tfrac{1}{2} L_2 \|x_+ - x\|^2 \leq \tfrac{L_2}{2\mu} \|f'(x)\|^2. \end{aligned}$$

Thus, $x: \ \|f'(x)\| < \frac{2\mu}{L_2}$ is in the *region of quadratic convergence*.

Global analysis

Global convergence: Use line search (good advice).

Global performance: Not addressed.

<u>Complexity for convex functions</u>

Assumptions

- **Strong convexity:** $\nabla^2 f(x) \succeq \mu I$ for all x. Consequence:

  $$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2.$$

  Minimizing this inequality in y, we get

  $$\boxed{\frac{1}{2\mu}\|f'(x)\|^2 \geq f(x) - f^*}$$

- **Lipschitz continuous gradient:** $\|f'(x) - f'(y)\| \leq L_1\|x - y\|$.

  By integration, we get

  $$\boxed{f(y) \leq f(x) + \langle f'(x), y - x \rangle + \frac{L_1}{2}\|x - y\|^2}$$

Convergence rate

Gradient method $\boxed{x_{k+1} = x_k - \frac{1}{L_1} f'(x_k)}$

Thus, at every iteration we have

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L_1} \|f'(x_k)\|^2 \geq \frac{\mu}{2L_1}(f(x_k) - f^*).$$

Newton Method $\boxed{x_{k+1} = x_k - \tau[f''(x_k)]^{-1} f'(x_k)}$

Then

$$f(x_{k+1}) - f(x_k) \leq -\tau \langle f'(x_k), [f''(x_k)]^{-1} f'(x_k) \rangle + \frac{L_1}{2}\tau^2 \|[f''(x_k)]^{-1} f'(x_k)\|^2$$

$$\leq \left[-\tau + \frac{L_1}{2\mu}\tau^2\right] \langle f'(x_k), [f''(x_k)]^{-1} f'(x_k) \rangle.$$

Minimization in τ gives

$$f(x_{k+1}) - f(x_k) \leq -\frac{\mu}{2L_1} \langle f'(x_k), [f''(x_k)]^{-1} f'(x_k) \rangle.$$

Since $\langle f'(x_k), [f''(x_k)]^{-1} f'(x_k) \rangle \geq \frac{1}{L_1} \|f'(x_k)\|^2 \geq \frac{2\mu}{L_1}(f(x_k) - f^*)$,

we get $\quad f(x_k) - f(x_{k+1}) \geq \left(\frac{\mu}{L_1}\right)^2 (f(x_k) - f^*).$

This is worse than for GM!

Modern History (Conn, Gould and Toint [2000])

Main idea: *Trust Region Approach*.

1. Use some norm $\| \cdot \|_k$ for defining a trust region

$$\mathcal{B}_k = \{x \in \mathbb{R}^n : \; \|x - x_k\|_k \leq \Delta_k\}.$$

2. Denote $m_k(x) = f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{2}\langle G_k(x - x_k), x - x_k \rangle$.

Variants: $G_k = f''(x_k)$, $G_k = f''(x_k) + \gamma_k I \succ 0$, etc.

3. Compute the trial point $\hat{x}_k = \arg\min\limits_{x \in \mathcal{B}_k} m_k(x)$.

4. Compute the ratio $\rho_k = \frac{f(x_k) - f(\hat{x}_k)}{f(x_k) - m_k(\hat{x}_k)}$.

5. In accordance to ρ_k, either accept $x_{k+1} = \hat{x}_k$, or decrease the value Δ_k and repeat the steps above.

Comments

Advantages:

- More parameters $\Rightarrow$ Flexibility
- Convergence to a point, which satisfies second-order necessary optimality condition:

$$f'(x^*) = 0, \quad f''(x^*) \succeq 0.$$

Disadvantages:

- Complicated strategies for parameters' coordination.
- For certain $\| \cdot \|_k$, the auxiliary problem is difficult.
- Line search abilities are limited.
- Unselective theory (local convergence).
- Global complexity issues are not addressed, even in convex case.

Trust Region Method with Contraction

Consider the problem: $\min_{x \in Q} f(x)$,

where Q is a bounded closed convex set, and f is a closed convex function.

Assumptions: $L_1(f) < +\infty$, $L_2(f) < +\infty$.

Method:

1. Choose arbitrary $x_0 \in Q$.

2. For $k \geq 0$ **iterate**: Choose $\tau_k \in (0,1)$ and compute

$$\begin{aligned}\min_x \{ \quad & \langle f'(x_k), y - x_k \rangle + \tfrac{1}{2} \langle f''(x_k)(y - x_k), y - x_k \rangle : \\ & y = (1-\tau_k)x_k + \tau_k x, \ x \in Q \}.\end{aligned}$$

Theorem. If $\tau_k = \frac{6(k+1)}{(k+2)(2k+3)}$, $k \geq 0$, then

$$f(x_k) - f^* \leq \frac{18 L_2 D^3}{(k+1)(2k+1)} + \frac{9 L_1 D^2}{2(2k+1)},$$

where $D = \operatorname{diam} Q$.

Development of numerical schemes

Classical style: Problem formulation $\Rightarrow$ Method

Examples:

- Gradient and Newton methods in optimization.
- Runge-Kutta method for ODE, etc.

2. Modern style: $\left.\begin{array}{l}\text{Problem formulation}\\ \text{Problem class}\end{array}\right\} \Rightarrow$ Method

Examples:

- Non-smooth convex minimization.
- Smooth minimization: $\min\limits_{x \in Q} f(x)$, with $f \in C^{1,1}$.

  Gradient mapping (Nemirovsky&Yudin 77):

$$
\begin{aligned}
x_+ \quad &= \quad T(x) \equiv \arg\min_{y \in Q} \; m_1(y), \\
m_1(y) \quad &\equiv \quad f(x) + \langle f'(x), y - x \rangle + \tfrac{L_1}{2} \| y - x \|^2.
\end{aligned}
$$

Justification: $f(y) \leq m_1(y)$ for all $y \in Q$.

Using the second-order model

Problem: $f(x) \quad \min: \; x \in \mathbb{R}^n$.

Assumption: Let $\mathcal{F}$ be an open convex set. Then

$$\|f''(x) - f''(y)\| \leq L_2 \|x - y\| \quad \forall x, y \in \mathcal{F},$$

$$\mathcal{L}(x_0) = \{x \in \mathbb{R}^n : \; f(x) \leq f(x_0)\} \subset \mathcal{F}.$$

Define

$$m_2(x, y) = f(x) + \langle f'(x), y - x \rangle + \tfrac{1}{2} \langle f''(x)(y - x), y - x \rangle,$$

$$m_2'(x, y) = f'(x) + f''(x)(y - x).$$

Lemma 1. For any $x, y \in \mathcal{F}$, $\quad \|f'(y) - m_2'(x, y)\| \leq \frac{1}{2} L_2 \|y - x\|^2$,

$$|f(y) - m_2(x, y)| \quad \leq \quad \tfrac{1}{6} L_2 \|y - x\|^3.$$

Corollary: For any x and y from $\mathcal{F}$,

$$f(y) \leq m_2(x, y) + \tfrac{1}{6} L_2 \|y - x\|^3.$$

Cubic regularization

For $M > 0$, define $\hat{f}_M(x,y) = m_2(x,y) + \frac{1}{6} M \|y-x\|^3$, and

$$T_M(x) \in \operatorname{Arg}\min_y \; \hat{f}_M(x,y),$$

where "Arg" indicates that $T_M(x)$ is the *global* minimum.

Computability: If $\|\cdot\|$ is a Euclidean norm, then $T_M(x)$ can be computed from as a solution of *convex* problem.

$$\begin{aligned}
& \min_{y \in \mathbb{R}^n} \quad \left\{ \langle f'(x), y-x \rangle + \tfrac{1}{2} \langle f''(x)(y-x), y-x \rangle + \tfrac{M}{6} \|y-x\|^3 \right\} \\
= & \min_{y \in \mathbb{R}^n} \max_{r \geq 0} \quad \left\{ \langle f'(x), y-x \rangle + \tfrac{1}{2} \langle f''(x)(y-x), y-x \rangle \right. \\
& \qquad \left. + \tfrac{Mr}{4} \|y-x\|^2 - \tfrac{Mr^3}{12} \right\} \\
\geq & \max_{r \geq 0} \min_{y \in \mathbb{R}^n} \quad \left\{ \langle f'(x), y-x \rangle + \tfrac{1}{2} \langle f''(x)(y-x), y-x \rangle \right. \\
& \qquad \left. + \tfrac{Mr}{4} \|y-x\|^2 - \tfrac{Mr^3}{12} \right\} \\
= & \sup_{r \geq 0} \quad \left\{ -\tfrac{1}{2} \langle f'(x), [f''(x) + \tfrac{Mr}{2} I]^{-1} f'(x) \rangle - \tfrac{Mr^3}{12} : f''(x) + \tfrac{Mr}{2} I \succ 0 \right\}.
\end{aligned}$$

Dual problem

For $r \in \mathcal{D} \equiv \{r \in R : f''(x) + \frac{M}{2} rI \succ 0, r \geq 0\}$,

Denote $v(r) \; = \; -\frac{1}{2}\langle f'(x), [f''(x) + \frac{Mr}{2} I]^{-1} f'(x)\rangle - \frac{M}{12} r^3$.

Lemma. For any $M > 0$, we have

$$\min_{h \in \mathbb{R}^n} \hat{f}_M(x, x+h) \; = \; \sup_{r \in \mathcal{D}} v(r).$$

(No duality gap.)

If the *sup* is attained at $r^* : \; f''(x) + \frac{Mr^*}{2} I \succ 0$, then

$$h^* = -[f''(x) + \frac{Mr^*}{2} I]^{-1} f'(x),$$

where $r^* > 0$ is a unique solution to the equation

$$r = \|[f''(x) + \frac{Mr}{2} I]^{-1} f'(x)\|.$$

Underlying fact: Convexity of *numerical range*.

Theorem. The set $\{u \in \mathbb{R}^2 : \; u^{(1)} = q_1(x), \; u^{(2)} = q_2(x), \; x \in \mathbb{R}^n\}$,

where functions $q_1(\cdot)$ and $q_2(\cdot)$ are quadratic and $n \geq 2$, is *convex*.

Our case: minimize $u^{(1)} + \frac{M}{6}(u^{(2)})^{3/2}$.

Simple properties, I

Denote $r_M(x) = \|x - T_M(x)\|$. Then, by the first-order optimality condition we have

$$(A_1): \quad f'(x) + f''(x)(T_M(x) - x) + \tfrac{Mr_M(x)}{2}(T_M(x) - x) = 0.$$

Moreover, since $r_M(x)$ is dual feasible, we have

$$(B_1): \quad f''(x) + \tfrac{1}{2}Mr_M(x)I \succeq 0.$$

1. We have $\langle f'(x), x - T_M(x)\rangle \geq 0$.

Proof. In view of (A_1) we have

$$\langle f'(x), x - T_M(x)\rangle = \langle [f''(x) + \tfrac{1}{2}Mr_M(x)I](x - T_M(x)), x - T_M(x)\rangle.$$

It is non-negative in view of (B_1). □

2. $f(x) - \bar{f}_M(x) \geq \frac{M}{12} r_M^3(x)$. If $M \geq L$, then $\bar{f}_M(x) \geq f(T_M(x))$. Hence

$$f(x) - f(T_M(x)) \geq \tfrac{M}{12} r_M^3(x). \text{ (Convexity: } \tfrac{M}{3} r_M^3(x).)$$

Proof. For $T = T_M(x)$ and $r = r_M(x)$, by (A_1) and (B_1) we have

$$\begin{aligned} f(x) - \bar{f}_M(x) &= \langle f'(x), x - T\rangle - \tfrac{1}{2}\langle f''(x)(T - x), T - x\rangle - \tfrac{M}{6} r^3 \\ &= \tfrac{1}{2}\langle f''(x)(T - x), T - x\rangle + \tfrac{M}{3} r^3. \end{aligned}$$

□

Simple properties, II

3. For any $M > 0$, we have

$$r_M^2(x) \geq \frac{2}{L_2+M}\|f'(T_M(x))\|.$$

Proof. Indeed, for $T = T_M(x)$ and $r = r_M(x)$, by (A_1) we have

$$\begin{aligned}
\|f'(T)\| &= \|f'(T) - f'(x) - f''(x)(T-x) - \tfrac{M}{2} r (T-x)\| \\
&\leq \|f'(T) - f'(x) - f''(x)(T-x)\| + \tfrac{M}{2} r^2 \leq \tfrac{L_2+M}{2} r^2.
\end{aligned}$$

4. For any $M > 0$ we have

$$\bar{f}_M(x) \leq \min_y \left[f(y) + \tfrac{L_2+M}{6}\|y-x\|^3 \right].$$

Proof. Indeed,

$$\begin{aligned}
\bar{f}_M(x) &= \min_y \quad \{ f(x) + \langle f'(x), y-x\rangle + \tfrac{1}{2}\langle f''(x)(y-x), y-x\rangle \\
&\qquad\quad + \tfrac{M}{6}\|y-x\|^3 \} \\
&\leq \min_y \quad \{ f(y) + \tfrac{L_2+M}{6}\|y-x\|^3 \}. \qquad \square
\end{aligned}$$

Corollary 1: For $M \geq L_2$, we have $f(x_1) - f^* \leq \frac{L_2+M}{6} D^3$.

Cubic regularization of Newton method

Consider the process: $\quad x_{k+1} = T_L(x_k), \quad k = 0, 1, \ldots .$

Note that $f(x_{k+1}) \leq f(x_k)$.

Saddle points. Let $f'(x^*) = 0$ and $f''(x^*) \not\succeq 0$. Then $\exists \epsilon, \delta > 0$ such that

$$\boxed{\|x - x^*\| \leq \epsilon,\ f(x) \geq f(x^*)} \Rightarrow \boxed{f(T_L(x)) \leq f(x^*) - \delta}$$

Example. Let $f'(x) = 0$ and $f''(x) \not\succeq 0$. Then $\bar{f}_M(T) < f(x)$.

Hence, if $M > L_2(f)$, then $f(T) < f(x)$. □

Local rate of convergence: Quadratic.

Proof. Indeed, $\|f'(T)\| \leq \frac{1}{2}(L_2 + M) r_M^2(x)$. At the same time,

$$r_M(x) = \|[f''(x) + \tfrac{1}{2} M r_M(x) I]^{-1} f'(x)\|.$$

□

Behavior of minimizing sequence

Let x_* be a limiting point of the sequence $\{x_k\}_{k \geq 0}$. Then

$$f'(x_*) = 0 \text{ and } f''(x_*) \succeq 0.$$

Proof. Follows from the following facts:

- $f(x_k) - f(x_{k+1}) \geq \frac{M}{12} r_M^3(x_k)$,
- $\|f'(x_{k+1})\| \leq \frac{L_2+M}{2} r_M^2(x_k)$,
- $f''(x_x) + \frac{M}{2} r_M(x_k) I \succeq 0.$ $\square$

Global convergence

Denote $g_k \equiv \min\limits_{1 \leq i \leq k} \|f'(x_i)\|$. Then $\boxed{g_k \leq O\left(\frac{1}{k^{2/3}}\right)}$

NB: For the gradient method, we can guarantee only $g_k \leq O\left(\frac{1}{k^{1/2}}\right)$.

Global performance: Star-convex functions

Def. For any x^*, any $x \in \mathbb{R}^n$, and $\alpha \in [0,1]$, we have

$$f(\alpha x^* + (1-\alpha)x) \leq \alpha f(x^*) + (1-\alpha) f(x).$$

Theorem 1.

1. If $f(x_0) - f^* \geq \frac{3}{2} L_2 D^3$, then $f(x_1) - f^* \leq \frac{1}{2} L_2 D^3$.
2. If $f(x_0) - f^* \leq \frac{3}{2} L_2 D^3$, then $f(x_k) - f^* \leq \frac{3 L_2 D^3}{2(1+\frac{1}{3}k)^2}$.

Proof.

$$\begin{aligned} f(x_{k+1}) &\leq \min_x \left\{ f(x) + \tfrac{L_2+M}{6} \|x - x_k\|^3 \right\} \\ &\leq \min_{\alpha \in [0,1]} \left\{ f(x) - \alpha(f(x) - f^*) + \tfrac{L_2+M}{6} \alpha^3 D^3 \right\}. \end{aligned}$$

Our conditions ensure $\alpha_k^* \in [0,1]$. Then

$$f(x_k) - f(x_{k+1}) \geq O((f(x_k) - f^*)^{3/2}).$$

This means that $f(x_k) - f^* = O(\frac{1}{k^2})$. $\square$

Superlinear convergence

Let the set of optimal solutions X^* be non-degenerate:

$$f(x) - f^* \geq \frac{\gamma}{2}\rho^2(x, X^*).$$

Denote $\bar{\omega} = \frac{1}{L_2^2}(\frac{\gamma}{2})^3$.

Theorem 2. Let k_0 the first number with $f(x_{k_0}) - f^* \leq \frac{4}{9}\bar{\omega}$.

If $k \leq k_0$, then $f(x_k) - f^* \leq \left[(f(x_0) - f^*)^{1/4} - \frac{k}{6}\sqrt{\frac{2}{3}}\bar{\omega}^{1/4} \right]^4$.

For $k \geq k_0$, we have $f(x_{k+1}) - f^* \leq \frac{1}{2}(f(x_k) - f^*)\sqrt{\frac{f(x_k) - f^*}{\bar{\omega}}}$.

Proof. Indeed,

$$\begin{aligned} f(x_{k+1}) &\leq \min_{x \in X^*} \left\{ f(x) + \frac{L_2+M}{6}\|x - x_k\|^3 \right\} \\ &= f^* + \frac{L_2+M}{6}\left[\left(\frac{2}{\gamma}(f(x_k) - f^*) \right)^{1/2} \right]^3. \quad \square \end{aligned}$$

NB The Hessian $f''(x^*)$ can be degenerate!

Global performance: Gradient-dominated functions

Definition. For any $x \in \mathcal{F}$, and $x^* \in X^*$, we have

$$f(x) - f(x^*) \le \tau_f \|f'(x)\|^p$$

with $\tau_f > 0$ and $p \in [1,2]$ (*degree of domination*).

Example 1. *Convex functions*:

$$f(x) - f^* \le \langle f'(x), x - x^* \rangle \le R\|f'(x)\|$$

for $\|x - x^*\| \le R$. Thus, $p = 1$, $\tau_f = \frac{1}{2}D$.

Example 2. *Strongly convex functions*: $\forall x, y \in \mathbb{R}^n$

$$f(x) \le f(y) + \langle f'(y), x - y \rangle + \frac{1}{2\gamma}\|f'(x) - f'(y)\|^2.$$

Thus, $f(x) - f^* \le \frac{1}{2\gamma}\|f'(x)\|^2 \quad \Rightarrow \quad p = 2,\ \tau_f = \frac{1}{2\gamma}$.

Gradient dominated functions, II

Example 3. *Sum of squares.* Consider the system

$$g(x) = 0 \in \mathbb{R}^m, \quad x \in \mathbb{R}^n,$$

which has a solution x^*, $g(x^*) = 0$.

Assume that $m \le n$ and the Jacobian $J(x) = (g_1'(x), \dots, g_m'(x))$ is *uniformly non-degenerate*:

$$\sigma \equiv \inf_{x \in \mathbb{R}^n} \lambda_{\min}(J^T(x)J(x)) > 0.$$

Theorem 3. Consider the function $f(x) = \sum\limits_{i=1}^{m} g_i^2(x)$. Then

$$f(x) - f^* \le \tfrac{1}{2\sigma} \|f'(x)\|^2.$$

Thus, $p = 2$ and $\tau_f = \frac{1}{2\sigma}$.

Proof. Indeed, $f'(x) = J(x)g(x)$. Therefore,

$$\begin{aligned} \|f'(x)\|^2 &= \langle (J^T(x)J(x))g(x), g(x) \rangle \\ &\ge \sigma \|g(x)\|^2 = 2\sigma(f(x) - f^*). \end{aligned}$$

$\square$

Gradient dominated functions: convergence rate

Theorem 3. Let $p = 1$. Denote $\hat{\omega} = \frac{2}{3}L(6\tau_f)^3$.

Let k_0 be defined as $f(x_{k_0}) - f^* \le \xi^2\hat{\omega}$ for some $\xi > 1$.

Then for $k \le k_0$ we have

$$\ln\left(\tfrac{1}{\hat{\omega}}(f(x_k) - f^*)\right) \le \left(\tfrac{2}{3}\right)^k \ln\left(\tfrac{1}{\hat{\omega}}(f(x_0) - f^*)\right).$$

Otherwise, $f(x_k) - f^* \le \hat{\omega} \cdot \frac{\xi^2(2+\frac{3}{2}\xi)^2}{(2+(k+\frac{3}{2})\cdot\xi)^2}$.

Proof. Indeed, we have

- $f(x_k) - f(x_{k+1}) \ge c_1\|f'(x_{k+1})\|^{3/2}$,
- $\|f'(x_{k+1})\| \ge \frac{1}{\tau_f}(f(x_{k+1}) - f^*)$.

Therefore, $f(x_k) - f^* = O(\frac{1}{k^2})$. □

Superlinear rate of convergence

Theorem 4. Let $p = 2$. Denote $\tilde{\omega} = \frac{1}{(144L)^2 \tau_f^3}$.

Let k_0 be defined as $f(x_{k_0}) - f^* \leq \tilde{\omega}$.

Then for $k \leq k_0$ we have $f(x_k) - f^* \leq (f(x_0) - f^*) \cdot e^{-k\sigma}$

with $\sigma = \frac{\tilde{\omega}^{1/4}}{\tilde{\omega}^{1/4} + (f(x_0) - f^*)^{1/4}}$.

Otherwise, $f(x_{k+1}) - f^* \leq \tilde{\omega} \cdot \left(\frac{f(x_k) - f^*}{\tilde{\omega}} \right)^{4/3}$.

Proof. Indeed

$$
\begin{aligned}
f(x_k) - f(x_{k+1}) &\geq c_1 \| f'(x_{k+1}) \|^{3/2} \\
&\geq c_1 \left[\frac{1}{\tau_f} (f(x_{k+1}) - f^*) \right]^{3/4}.
\end{aligned}
$$

$\square$

NB: Superlinear convergence without direct nondegeneracy assumption for the Hessian.

Transformations of convex functions

Let $u(x) : \mathbb{R}^n \to \mathbb{R}^n$ be non-degenerate. Denote by $v(u)$ its inverse:

$$v(u(x)) \equiv x.$$

Consider the function $f(x) = \phi(u(x))$, where $\phi(u)$ is a convex function. Denote

$$\sigma = \max_u \{ \| v'(u) \| : \ \phi(u) \leq f(x_0) \},$$

$$D = \max_u \{ \| u - u^* \| : \ \phi(u) \leq f(x_0) \}.$$

Theorem 5.

1. If $f(x_0) - f^* \geq \frac{3}{2} L(\sigma D)^3$, then $f(x_1) - f^* \leq \frac{1}{2} L(\sigma D)^3$.
2. If $f(x_0) - f^* \leq \frac{3}{2} L(\sigma D)^3$, then $f(x_k) - f^* \leq \frac{3L(\sigma D)^3}{2(1+\frac{1}{3}k)^2}$.

Proof. It is based on non-degeneracy of $u'(\cdot)$ and the reasoning for star-convex functions. □

Example. For *arbitrary* functions $\phi_i(\cdot)$, $i = 1, \dots, n-1$, define

$$\begin{array}{rcl} u_1(x) & = & x_1, \quad u_2(x) \ = \ x_2 + \phi_1(x_1), \quad \dots, \\ u_n(x) & = & x_n + \phi_{n-1}(x_1, \dots, x_{n-1}). \end{array}$$

Solving the systems of nonlinear equations

1. Standard Gauss-Newton method

Problem: Find $x \in \mathbb{R}^n$ satisfying the system $F(x) = 0 \in \mathbb{R}^m$.

Assumption: $\forall x, y \in \mathbb{R}^n \quad \|F'(x) - F'(y)\| \leq L\|x - y\|$.

Gauss-Newton method: Choose a merit function $\phi(u) \geq 0$, $\phi(0) = 0$, $u \in \mathbb{R}^m$.

Compute $x_+ \in \operatorname{Arg}\min\limits_y [\phi(F(x) + F'(x)(y - x))]$.

Usual choice: $\phi(u) = \sum\limits_{i=1}^m u_i^2$. (Justification: *Why not?*)

Remarks

- Local quadratic convergence ($m \geq n$, non-degeneracy and $F(x^*) = 0$ (?)).
- If $m < n$, then the method is not well-defined.
- No global complexity results.

Modified Gauss-Newton method

Lemma. For all $x, y \in \mathbb{R}^n$, we have

$$\|F(y) - F(x) - F'(x)(y-x)\| \leq \tfrac{1}{2} L \|y - x\|^2.$$

Corollary. Denote $f(y) = \|F(y)\|$. Then

$$f(y) \leq \|F(x) + F'(x)(y-x)\| + \tfrac{1}{2} L \|y - x\|^2.$$

Modified method:

$$x_{k+1} = \arg\min_y \left[\, \|F(x_k) + F'(x_k)(y - x_k)\| + \tfrac{1}{2} L \|y - x_k\|^2 \, \right].$$

Remarks

- The merit function is non-smooth.
- Nevertheless, $f(x_{k+1}) < f(x_k)$ unless x_k is a stationary point.
- Quadratic convergence for non-degenerate solutions.
- Global efficiency bounds.
- Problem of finding x_{k+1} is convex.
- Different norms in $\mathbb{R}^n$ and $\mathbb{R}^m$ can be used.

Implementation for Euclidean norm

$$
\begin{aligned}
& \min_{y} \quad \left[\, \|F(x_k) + F'(x_k)(y - x_k)\| + \tfrac{1}{2} L \|y - x_k\|^2 \, \right] \\
= \quad & \min_{y} \min_{\tau > 0} \quad \left[\, \tfrac{1}{2\tau} \|F(x_k) + F'(x_k)(y - x_k)\|^2 + \tfrac{1}{2}\tau^2 + \tfrac{1}{2} L \|y - x_k\|^2 \, \right] \\
= \quad & \min_{\tau > 0} \quad [\, \tfrac{1}{2\tau} \|F(x_k)\|^2 + \tfrac{1}{2}\tau^2 \\
& \qquad - \tfrac{1}{2\tau} \langle F'(x_k)^T F(x_k), [F'(x_k) F'(x_k)^T + \tau L I]^{-1} F'(x_k)^T F(x_k) \rangle \,].
\end{aligned}
$$

This is a convex univariate function.

Testing CNM: Chebyshev oscilator

Consider $f(x) = \frac{1}{4}(1 - x^{(1)})^2 + \sum\limits_{i=1}^{n-1} \left(x^{(i+1)} - p_2(x^{(i)})\right)^2$,

with $p_2(\tau) = 2\tau^2 - 1$.

Note that p_2 is a Chebyshev polynomial: $\quad p_k(\tau) = \cos(k \arccos(\tau))$.

Hence, the equations for the "central path" is

$$x^{(i+1)} = p_2(x^{(i)}) = p_4(x^{(i-1)}) = \ldots = p_{2^i}(x^{(1)}).$$

This is an exponential oscillation!

However, all coefficients in function and derivatives are small.

NB: $\quad f(x)$ is unimodular and $x^* = (1, \ldots, 1)$.

In our experiments we usually take $x_0 = (-1, 1, \ldots, 1)$.

Drawback: $x_0 - 2\nabla f(x_0) = x^*$. Hence, sometimes we use $x_0 = (-1, 0.9. \ldots, 0.9)$.

Solving Chebyshev oscilator by CN: $\|\nabla f(x)\|_{(2)} \leq 10^{-8}$

n	Iter	DF	GNorm	NumF	Time (s)
2	14	$7.0 \cdot 10^{-19}$	$4.2 \cdot 10^{-09}$	18	0.032
3	33	$1.1 \cdot 10^{-24}$	$7.5 \cdot 10^{-12}$	51	0.031
4	82	$1.7 \cdot 10^{-20}$	$9.3 \cdot 10^{-10}$	148	0.047
5	207	$4.5 \cdot 10^{-19}$	$1.2 \cdot 10^{-09}$	395	0.078
6	541	$1.0 \cdot 10^{-17}$	$5.6 \cdot 10^{-09}$	1062	0.266
7	1490	$1.4 \cdot 10^{-18}$	$2.9 \cdot 10^{-09}$	2959	0.609
8	4087	$2.7 \cdot 10^{-17}$	$9.1 \cdot 10^{-09}$	8153	1.782
9	11205	$1.6 \cdot 10^{-16}$	$9.6 \cdot 10^{-09}$	22389	5.922
10	30678	$2.7 \cdot 10^{-15}$	$9.6 \cdot 10^{-09}$	61335	18.89
11	79292	$7.7 \cdot 10^{-14}$	$1.0 \cdot 10^{-08}$	158563	57.813
12	171522	$9.7 \cdot 10^{-13}$	$9.9 \cdot 10^{-09}$	343026	144.266
13	385353	$1.3 \cdot 10^{-11}$	$9.9 \cdot 10^{-09}$	770691	347.094
14	938758	$2.1 \cdot 10^{-11}$	$1.0 \cdot 10^{-08}$	1877500	1232.953
15	2203700	$7.8 \cdot 10^{-11}$	$1.0 \cdot 10^{-08}$	4407385	3204.359

Other methods

	Trust	region	Knitro	Minos	5.5	Snopt	
n	Inner	Iter	Iter	Iter	NFG	Iter$^{\#}$	NFG
3	129	50	30	44	120	106	78
4	431	123	80	136	309	268	204
5	1310	299	203	339	793	647	509
6	3963	722	531	871	2022	1417	1149*
7	12672	1921	1467	2291	5404	$***$	
8	40036	5234	4040	6109	14680		
9	120873	13907	11062	11939	28535		
10	358317	36837	29729*	$***$			
11	842368	78854	$***$				
12	2121780	182261					

Notation: * early termination, $(***)$ numerical difficulties/ inaccurate solution, $^{\#}$ needs an alternative starting point.

Trust region: very reliable, but $T(12) = 2577$ sec (Matlab), $T(n) = Const * (4.5)^n$.