Lecture 5: Second Order Methods

- Newtons Method
- Convergence Rate of the Pure Form
- Global Convergence; Variants of Newtons Method
- Trust Region Methods

Newton's Method

$$\boxed{\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha_r \left(\nabla^2 f(\boldsymbol{x}^r)\right)^{-1} \nabla f(\boldsymbol{x}^r)}$$

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

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

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

- 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 Newton's Method

- Consider solution of nonlinear system $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$ where $\boldsymbol{g} : \mathbb{R}^n \mapsto \mathbb{R}^n$, with method
$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \left(\nabla \boldsymbol{g}(\boldsymbol{x}^r)'\right)^{-1} \boldsymbol{g}(\boldsymbol{x}^r)$$
[If $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$, we get pure form of Newton.]

- Quick derivation: Suppose $\boldsymbol{x}^r \to \boldsymbol{x}^*$ with $\boldsymbol{g}(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla \boldsymbol{g}(\boldsymbol{x}^*)$ is invertible. By Taylor expansion
$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{g}(\boldsymbol{x}^r) + \nabla \boldsymbol{g}(\boldsymbol{x}^r)'(\boldsymbol{x}^* - \boldsymbol{x}^r) + o\left(\|\boldsymbol{x}^r - \boldsymbol{x}^*\|\right)$$
Multiplying with $\nabla \boldsymbol{g}(\boldsymbol{x}^r)^{-1}$ yields
$$\boldsymbol{x}^r - \boldsymbol{x}^* - \left(\nabla \boldsymbol{g}(\boldsymbol{x}^r)\right)^{-1} \boldsymbol{g}(\boldsymbol{x}^r) = o\left(\|\boldsymbol{x}^r - x^*\|\right),$$
$$\Longrightarrow \quad \boldsymbol{x}^{r+1} - \boldsymbol{x}^* = o\left(\|\boldsymbol{x}^r - \boldsymbol{x}^*\|\right),$$
implying superlinear (quadratic if $f \in C^2$) convergence and capture.

Convergence Behavior of Newton's Method

Consider $g(x) = e^x - 1$. Initialize $x = -1$.

r	x^r	$g(x^r)$
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

Typically five to ten iterations to converge.

Unfortunately, the pure Newton method can also diverge, and often does!

Modifications of Newton's Method

To ensure the global convergence of Newton's method, we can

- modify the Newton direction when:
  - ⋆ Hessian is not positive definite
  - ⋆ When Hessian is nearly singular (needed to improve performance)
- use a stepsize (damped Newton method) and an Armijo type rule to ensure sufficient descent
- use
$$\boldsymbol{d}^r = -\left(\nabla^2 f(\boldsymbol{x}^r) + \Delta^r\right)^{-1} \nabla f(\boldsymbol{x}^r),$$
whenever the Newton direction does not exist or is not a descent direction. Here Δ^r is a diagonal matrix such that $\nabla^2 f(\boldsymbol{x}^r) + \Delta^r \succ \mathbf{0}$
  - ⋆ Modified Cholesky factorization
  - ⋆ Trust region methods

Trust Region Methods

- Instead of fixing the direction first and then fix stepsize, the trust region methods first limit the stepsize and then determine the direction.

- Let $\boldsymbol{x}^r$ be the current iterate and s be the stepsize. Consider

$$\begin{aligned} &\text{minimize} \quad q(\boldsymbol{x}) = \nabla f(\boldsymbol{x}^r)'(\boldsymbol{x}-\boldsymbol{x}^r) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^r)'\nabla^2 f(\boldsymbol{x}^r)(\boldsymbol{x}-\boldsymbol{x}^r) \\ &\text{subject to} \quad \|\boldsymbol{x}-\boldsymbol{x}^r\| \le s_r. \end{aligned} \tag{1}$$

- The objective function may be nonconvex if $\nabla^2 f(\boldsymbol{x}^r) \not\succeq \mathbf{0}$. But it can be efficiently solved regardless of convexity!

- Stepsize rule: let $\bar{s} > 0$ and $\boldsymbol{x}(s_r)$ be the optimal solution of (1),

$$s_{r+1} := \begin{cases} s_r/2, & \text{if } f(\boldsymbol{x}^r) - f(\boldsymbol{x}(s_r)) < -\frac{1}{2}q(\boldsymbol{x}(s_r)) \\ s_r, & \text{if } -\frac{1}{2}q(\boldsymbol{x}(s_r)) \le f(\boldsymbol{x}^r) - f(\boldsymbol{x}(s_r)) \le -2q(\boldsymbol{x}(s_r)) \\ & \quad \text{or } \|\boldsymbol{x}^r - \boldsymbol{x}(s_r)\| < s_r \\ \min\{2s_r, \bar{s}\}, & \text{if } f(\boldsymbol{x}^r) - f(\boldsymbol{x}(s_r)) > -2q(\boldsymbol{x}(s_r)) \ \&\ \|\boldsymbol{x}^r - \boldsymbol{x}(s_r)\| = s_r \end{cases}$$

$$\boldsymbol{x}^{r+1} := \boldsymbol{x}(s_r), \quad \text{if } f(\boldsymbol{x}^r) - f(\boldsymbol{x}(s_r)) > -\tfrac{1}{4}q(\boldsymbol{x}(s_r)), \quad \text{else } \boldsymbol{x}^{r+1} := \boldsymbol{x}^r.$$

Trust Region Subproblem

$$\boxed{\begin{array}{ll} \text{minimize} & f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}'\boldsymbol{x} \\ \text{subject to} & \|\boldsymbol{x}\|^2 \le 1 \end{array}}$$

- This is a constrained optimization problem.

- If $\boldsymbol{b} = \boldsymbol{0}$, then the optimal solution

$$\boldsymbol{x}^* = \begin{cases} \boldsymbol{0} & \text{if } \lambda_{\min}(\boldsymbol{Q}) \ge 0, \\ \text{the eig. vector of } \boldsymbol{Q} \text{ for } \lambda_{\min}(\boldsymbol{Q}), & \text{if } \lambda_{\min}(\boldsymbol{Q}) < 0 \end{cases}$$

  i.e., $\boldsymbol{Q}\boldsymbol{x}^* = \lambda_{\min}(\boldsymbol{Q})\boldsymbol{x}^*, \ \|\boldsymbol{x}^*\| = 1$.

- For general $\boldsymbol{b}$, the optimality condition is

$$\boldsymbol{Q}\boldsymbol{x}^* + \boldsymbol{b} + \lambda^*\boldsymbol{x}^* = \boldsymbol{0}, \quad (\|\boldsymbol{x}^*\|^2 - 1)\lambda^* = 0, \quad \lambda^* \ge 0, \quad \boldsymbol{Q} + \lambda^*\boldsymbol{I} \succeq \boldsymbol{0}. \quad (2)$$

Proof of (2) by the following steps.

- An optimal solution $\boldsymbol{x}^*$ always exists (why?). Suppose $\|\boldsymbol{x}^*\| < 1$. Then $\boldsymbol{x}^*$ is an unconstrained local min of f. Moreover, it must be a global min (why?). The optimality condition (2) holds with $\lambda^* = 0$.

- Suppose $\|\boldsymbol{x}^*\| = 1$. Let $h(\boldsymbol{x}) = \frac{1}{2}\max\{0, (\|\boldsymbol{x}\|^2 - 1)\}$ and $\alpha > 0$. Consider
$$f^k(\boldsymbol{x}) = f(\boldsymbol{x}) + k|h(\boldsymbol{x})|^2 + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2.$$

- Let $\boldsymbol{x}^k$ be a constrained minimizer of f^k over the ball $\|\boldsymbol{x} - \boldsymbol{x}^*\| \le 1$. We will show that $\boldsymbol{x}^k$ is an *unconstrained local min* of f^k for all large k.

- Taking limit $k \to \infty$ of
$$f^k(\boldsymbol{x}^k) = f(\boldsymbol{x}^k) + k|h(\boldsymbol{x}^k)|^2 + \frac{\alpha}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 \le f^k(\boldsymbol{x}^*) = f(\boldsymbol{x}^*),$$

along any convergent subsequence of $\{\boldsymbol{x}^k\}$, we get $h(\bar{\boldsymbol{x}}) = \lim_{k\to\infty} h(\boldsymbol{x}^k) = 0$.

- Furthermore, taking limit of $f(\boldsymbol{x}^k) + \frac{\alpha}{2}\|\boldsymbol{x}^k - \boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*)$ shows

$$f(\bar{\boldsymbol{x}}) + \frac{\alpha}{2}\|\bar{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \le f(\boldsymbol{x}^*)$$

- Since $h(\bar{\boldsymbol{x}}) = 0$, it follows that $f(\boldsymbol{x}^*) \le f(\bar{\boldsymbol{x}})$. Thus, we have $\bar{\boldsymbol{x}} = \boldsymbol{x}^*$ and $f(\boldsymbol{x}^*) = f(\bar{\boldsymbol{x}})$.

- Since $\bar{\boldsymbol{x}}$ is any limit point, we have $\boldsymbol{x}^k \to \boldsymbol{x}^*$, so $\|\boldsymbol{x}^k - \boldsymbol{x}^*\| < 1$ for large k, $\Rightarrow \boldsymbol{x}^k$ is an unconstrained local min of f^k, $\nabla f^k(\boldsymbol{x}^k) = 0$, $\nabla^2 f^k(\boldsymbol{x}^k) \succeq \mathbf{0}$.

- Taking limit of

$$\begin{aligned} \mathbf{0} &= \nabla f(\boldsymbol{x}^k) + 2kh(\boldsymbol{x}^k)\nabla h(\boldsymbol{x}^k) + \alpha(\boldsymbol{x}^k - \boldsymbol{x}^*) \\ &= \boldsymbol{Q}\boldsymbol{x}^k + \boldsymbol{b} + 2kh(\boldsymbol{x}^k)\boldsymbol{x}^k + \alpha(\boldsymbol{x}^k - \boldsymbol{x}^*) \qquad (3) \end{aligned}$$

shows

$$2kh(\boldsymbol{x}^k) = -\frac{(\boldsymbol{x}^k)'(\boldsymbol{Q}\boldsymbol{x}^k + \boldsymbol{b} + \alpha(\boldsymbol{x}^k - \boldsymbol{x}^*))}{\|\boldsymbol{x}^k\|^2} \to -(\boldsymbol{x}^*)'(\boldsymbol{Q}\boldsymbol{x}^* + \boldsymbol{b}) \equiv \lambda^*.$$

Taking limit in (3) yields $\boldsymbol{Q}\boldsymbol{x}^* + \boldsymbol{b} + \lambda^*\boldsymbol{x}^* = \boldsymbol{0}$, $\lambda^* \geq 0$.

- The remaining condition $\boldsymbol{Q} + \lambda^*\boldsymbol{I} \succeq \boldsymbol{0}$ follows from a contra-positive argument: if $\boldsymbol{Q} + \lambda^*\boldsymbol{I} \not\succeq \boldsymbol{0}$, then there exists a $\boldsymbol{u} \neq \boldsymbol{0}$ (e.g., the eigenvector of $\boldsymbol{Q} + \lambda^*\boldsymbol{I}$ corresponding to a negative eigenvalue, perturbed if necessary) s.t.

$$\boldsymbol{u}'(\boldsymbol{Q} + \lambda^*\boldsymbol{I})\boldsymbol{u} < 0, \quad \langle \boldsymbol{u}, \boldsymbol{x}^* \rangle < 0.$$

Then, perturb $\boldsymbol{x}^*$ by $\boldsymbol{u}$ to derive a contradiction to the optimality of $\boldsymbol{x}^*$.

- Finally, check the sufficiency: suppose (2) holds for some $\boldsymbol{x}^*,\ \lambda^*$. Consider the following quadratic function

$$f_{\lambda^*}(\boldsymbol{x}) = f(\boldsymbol{x}) + \frac{1}{2}\lambda^*(\|\boldsymbol{x}\|^2 - 1)$$

Claim: $\boldsymbol{x}^*$ is a global min of f_{λ^*}. So $\boldsymbol{x}^*$ is a global min of f over the sphere $\|\boldsymbol{x}\|^2 \leq 1$.

Solving the Trust Region Subproblem

- Observation: if $\boldsymbol{Q} + \lambda \boldsymbol{I} \succ \mathbf{0}$, then

$$\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b} + \lambda \boldsymbol{x} = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{x}(\lambda) = -(\boldsymbol{Q} + \lambda \boldsymbol{I})^{-1}\boldsymbol{b}.$$

  Moreover, $\|\boldsymbol{x}(\lambda)\|$ is a decreasing function of λ.

- Note that $\boldsymbol{x}(-\lambda_{\min}(\boldsymbol{Q}))$ is not uniquely defined, and $\|\boldsymbol{x}(-\lambda_{\min}(\boldsymbol{Q}))\|$ can be made arbitrarily large (why?).

- Strategy: binary search of λ over $[\lambda_b, \infty)$ with $\lambda_b = \max\{0, -\lambda_{\min}(\boldsymbol{Q})\}$.
  - ⋆ Check if $\|\boldsymbol{x}(\lambda_b)\| \leq 1$. If yes, then $\boldsymbol{x}(\lambda_b)$ is an optimal solution (why?). Stop.
  - ⋆ Else, we must have $\|\boldsymbol{x}(\lambda_b)\| > 1$. Let $\lambda_a = \max\{1, \lambda_b\}$, and check if $\|\boldsymbol{x}(\lambda_a)\| \leq 1$. If not, update $\lambda_a := 2\lambda_a$ until $\|\boldsymbol{x}(\lambda_a)\| \leq 1$.
  - ⋆ Let $\lambda = (\lambda_a + \lambda_b)/2$. If $\|\boldsymbol{x}(\lambda)\| \leq 1$, then update $\lambda_a = \lambda$. Else, we update $\lambda_b = \lambda$. Stop when $|\lambda_b - \lambda_a| \leq \epsilon$.

Inexact Solution of the Trust Region Subproblem

- The Cauchy point $\boldsymbol{x}^c$ (corresponding to the linear version of the TR subproblem):

$$\boldsymbol{z} := \underset{\boldsymbol{x}:\|\boldsymbol{x}-\boldsymbol{x}^r\|\le s_r}{\arg\min} \langle \nabla f(\boldsymbol{x}^r), \boldsymbol{x}-\boldsymbol{x}^r\rangle = \boldsymbol{x}^r - s_r \frac{\nabla f(\boldsymbol{x}^r)}{\|\nabla f(\boldsymbol{x}^r)\|}$$

and

$$\boldsymbol{x}^c := \underset{0\le\tau\le1}{\arg\min} \left\{ \langle \nabla f(\boldsymbol{x}^r), \tau(\boldsymbol{z}-\boldsymbol{x}^r)\rangle + \frac{\tau^2}{2}(\boldsymbol{z}-\boldsymbol{x}^r)'\nabla^2 f(\boldsymbol{x}^r)(\boldsymbol{z}-\boldsymbol{x}^r) \right\}.$$

- It can be checked that

$$\boldsymbol{x}^c = \boldsymbol{x}^r - \tau_r s_r \frac{\nabla f(\boldsymbol{x}^r)}{\|\nabla f(\boldsymbol{x}^r)\|}$$

where

$$\tau_r = \begin{cases} 1, & \text{if } \nabla f(\boldsymbol{x}^r)'\nabla^2 f(\boldsymbol{x}^r)\nabla f(\boldsymbol{x}^r) \le 0 \\ \min\left(\frac{\|\nabla f(\boldsymbol{x}^r)\|^3}{s_r \nabla f(\boldsymbol{x}^r)'\nabla^2 f(\boldsymbol{x}^r)\nabla f(\boldsymbol{x}^r)}, 1 \right), & \text{else} \end{cases}$$

Inexact Solution of the Trust Region Subproblem

- Quality of the Cauchy point (sufficient descent): there exists some $c_1 \in (0, 1]$ s.t.

$$q(\boldsymbol{x}^c) \leq -c_1 \|\nabla f(\boldsymbol{x}^r)\| \min \left(s_r, \frac{\|\nabla f(\boldsymbol{x}^r)\|}{\|\nabla^2 f(\boldsymbol{x}^r)\|} \right) \tag{4}$$

- The curvature information only affects the length of Cauchy step, not its direction. For fast convergence, the curvature information should be used to determine the direction.
- One way is to further improve the Cauchy point $\boldsymbol{x}^c$ by consider the two-dimensional minimization problem

$$\min_{\substack{\boldsymbol{x}: \|\boldsymbol{x}-\boldsymbol{x}^r\| \leq s_r \\ \boldsymbol{x}-\boldsymbol{x}^r \in \boldsymbol{V}}} \langle \nabla f(\boldsymbol{x}^r), \boldsymbol{x} - \boldsymbol{x}^r \rangle + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^r)' \nabla^2 f(\boldsymbol{x}^r)(\boldsymbol{x} - \boldsymbol{x}^r)$$

where $\boldsymbol{V}$ is the two-dimensional space

$$\boldsymbol{V} := \text{span}\{\nabla f(\boldsymbol{x}^r), (\nabla^2 f(\boldsymbol{x}^r))^{-1} \nabla f(\boldsymbol{x}^r)\}$$

- The above subproblem is simple to solve (reducible to finding the root of a 4th order polynomial - closed form solution)

Convergence of Trust Region Methods

- Suppose the level set $S := \{\boldsymbol{x} \mid f(\boldsymbol{x}) \le f(\boldsymbol{x}^0)\}$ is bounded, $\nabla^2 f(\boldsymbol{x})$ is bounded on S.

- Assume sufficient descent: $\boldsymbol{x}^r$ is at least as good as the Cauchy point (cf. (4))

$$q(\boldsymbol{x}^r) \le -c_1 \|\nabla f(\boldsymbol{x}^r)\| \min\left(s_r, \frac{\|\nabla f(\boldsymbol{x}^r)\|}{\|\nabla^2 f(\boldsymbol{x}^r)\|}\right)$$

- Then $\nabla f(\boldsymbol{x}^r) \to \mathbf{0}$ as $r \to \infty$.

- In practice, $\nabla^2 f(\boldsymbol{x}^r)$ can be replaced by any matrix $\boldsymbol{B}^r$ (uniformly bounded).