Lecture 5: Unconstrained smooth optimization

School of Mathematical Sciences, Xiamen University

1. Taylor's theorem

- Taylor's theorem shows how smooth functions can be locally approximated by low-order (e.g., linear or quadratic) functions.

定理 12.3.1(Taylor 公式) 设 $f(x,y)$ 在点 (x_0,y_0) 的邻域 $U=O((x_0,y_0),r)$ 上具有 $k+1$ 阶连续偏导数，那么对于 U 内每一点 $(x_0+\Delta x,y_0+\Delta y)$ 都成立

$$
\begin{aligned}
& f(x_0+\Delta x, y_0+\Delta y) \\
= & f(x_0,y_0) + \left(\Delta x\frac{\partial}{\partial x}+\Delta y\frac{\partial}{\partial y}\right)f(x_0,y_0) \\
& + \frac{1}{2!}\left(\Delta x\frac{\partial}{\partial x}+\Delta y\frac{\partial}{\partial y}\right)^2 f(x_0,y_0)+\cdots \\
& + \frac{1}{k!}\left(\Delta x\frac{\partial}{\partial x}+\Delta y\frac{\partial}{\partial y}\right)^k f(x_0,y_0)+R_k,
\end{aligned}
$$

其中 $R_k=\dfrac{1}{(k+1)!}\left(\Delta x\dfrac{\partial}{\partial x}+\Delta y\dfrac{\partial}{\partial y}\right)^{k+1} f(x_0+\theta\Delta x, y_0+\theta\Delta y)(0<\theta<1)$ 称为 **Lagrange 余项**.

Theorem 1

Given a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x}+\xi\mathbf{p})^\top \mathbf{p}, \ \textit{for some } \xi \in (0,1),$$

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \int_0^1 \nabla f(\mathbf{x}+t\mathbf{p})^\top \mathbf{p} \mathrm{d}t,$$

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{p} + o(\|\mathbf{p}\|).$$

If f is twice continuously differentiable, we have for some $\xi \in (0,1)$,

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{x}+\xi\mathbf{p})\mathbf{p},$$

and

$$\nabla f(\mathbf{x}+\mathbf{p}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x}+t\mathbf{p})\mathbf{p} \mathrm{d}t,$$

$$f(\mathbf{x}+\mathbf{p}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{x})\mathbf{p} + o(\|\mathbf{p}\|^2).$$

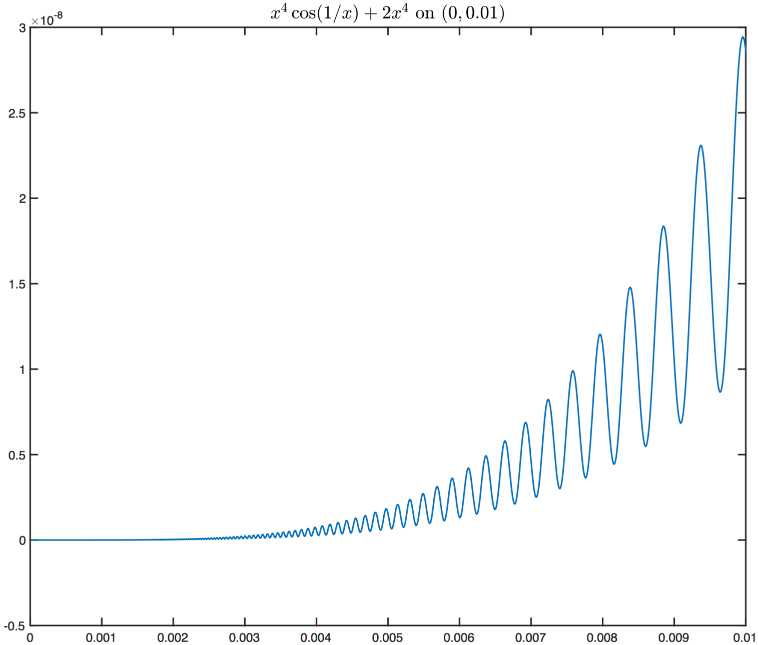
2. Global and local solutions of $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$

- $\mathbf{x}_\star$ is a *local minimizer* of f if there is a neighborhood $\mathcal{N}$ of $\mathbf{x}_\star$ such that $f(\mathbf{x}) \geq f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathcal{N}$.
- $\mathbf{x}_\star$ is a *strict local minimizer* if it is a local minimizer on some neighborhood $\mathcal{N}$ and in addition $f(\mathbf{x}) > f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathcal{N}$ with $\mathbf{x} \neq \mathbf{x}_\star$.
- $\mathbf{x}_\star$ is an *isolated local minimizer* if there is a neighborhood $\mathcal{N}$ of $\mathbf{x}_\star$ such that $f(\mathbf{x}) \geq f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathcal{N}$ and in addition, $\mathcal{N}$ contains no local minimizers other than $\mathbf{x}_\star$.

  Strict local minimizers are not always isolated: for example,

  $$f(x) = x^4 \cos(1/x) + 2x^4, \;\; f(0) = 0.$$

  All isolated local minimizers are strict.
- $\mathbf{x}_\star$ is a *global minimizer* of f if $f(\mathbf{x}) \geq f(\mathbf{x}_\star)$ for all $\mathbf{x} \in \mathbb{R}^n$.


$x^4\cos(1/x) + 2x^4$ on $(0, 0.01)$
$\times 10^{-8}$
3
2.5
2
1.5
1
0.5
0
-0.5
0
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.01

3. Optimality conditions for smooth functions

Theorem 2 (first-order necessary condition)

If $\mathbf{x}_\star$ is a local minimizer of f and f is continuously differentiable in an open neighborhood of $\mathbf{x}_\star$, then $\nabla f(\mathbf{x}_\star) = \mathbf{0}$.

Proof. Suppose for contradiction that $\nabla f(\mathbf{x}_\star) \neq \mathbf{0}$. Define the vector $\mathbf{p} = -\nabla f(\mathbf{x}_\star)$ and note that $\mathbf{p}^\top \nabla f(\mathbf{x}_\star) = -\|\nabla f(\mathbf{x}_\star)\|^2 < 0$. Because ∇f is continuous near $\mathbf{x}_\star$, there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla f(\mathbf{x}_\star + t\mathbf{p}) < 0, \quad \text{for all} \quad t \in [0, T].$$

For any $s \in (0, T]$, we have by Taylor's theorem that

$$f(\mathbf{x}_\star + s\mathbf{p}) = f(\mathbf{x}_\star) + s\mathbf{p}^\top \nabla f(\mathbf{x}_\star + \xi s\mathbf{p}) \quad \text{for some} \quad \xi \in (0, 1).$$

Therefore, $f(\mathbf{x}_\star + s\mathbf{p}) < f(\mathbf{x}_\star)$ for all $s \in (0, T]$. We have found a direction leading away from $\mathbf{x}_\star$ along which f decreases, so $\mathbf{x}_\star$ is not a local minimizer, and we have a contradiction. $\square$

Theorem 3 (second-order necessary conditions)

If $\mathbf{x}_\star$ *is a local minimizer of* f *and* $\nabla^2 f$ *is continuous in an open neighborhood of* $\mathbf{x}_\star$*, then* $\nabla f(\mathbf{x}_\star) = \mathbf{0}$ *and* $\nabla^2 f(\mathbf{x}_\star) \succeq \mathbf{0}$*.*

Proof. We know from Theorem 2 that $\nabla f(\mathbf{x}_\star) = \mathbf{0}$. Assume that $\nabla^2 f(\mathbf{x}_\star)$ is not positive semidefinite. Then we can choose a vector $\mathbf{p}$ such that $\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star)\mathbf{p} < 0$, and because $\nabla^2 f$ is continuous near $\mathbf{x}_\star$, there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + t\mathbf{p})\mathbf{p} < 0, \quad \text{for all} \quad t \in [0, T].$$

By doing a Taylor series expansion around $\mathbf{x}_\star$, we have for all $s \in (0, T]$ and some $\xi \in (0, 1)$ that

$$f(\mathbf{x}_\star + s\mathbf{p}) = f(\mathbf{x}_\star) + s\mathbf{p}^\top \nabla f(\mathbf{x}_\star) + \frac{1}{2}s^2\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi s\mathbf{p})\mathbf{p} < f(\mathbf{x}_\star).$$

As in Theorem 2, we have found a direction from $\mathbf{x}_\star$ along which f is decreasing, and so again, $\mathbf{x}_\star$ is not a local minimizer. $\square$

Theorem 4 (second-order sufficient conditions)

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of $\mathbf{x}_\star$ and that $\nabla f(\mathbf{x}_\star) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}_\star) \succ \mathbf{0}$. Then $\mathbf{x}_\star$ is a strict local minimizer of f.

Proof. Because the Hessian $\nabla^2 f$ is continuous and positive definite at $\mathbf{x}_\star$, we can choose a radius $r > 0$ so that $\nabla^2 f(\mathbf{x})$ remains positive definite for all $\mathbf{x}$ in the open ball $\mathcal{B} = \{\mathbf{z} \mid \|\mathbf{z} - \mathbf{x}_\star\| < r\}$. Taking any nonzero vector $\mathbf{p}$ with $\|\mathbf{p}\| < r$, we have $\mathbf{x}_\star + \mathbf{p} \in \mathcal{B}$ and

$$
\begin{aligned}
f(\mathbf{x}_\star + \mathbf{p}) &= f(\mathbf{x}_\star) + \mathbf{p}^\top \nabla f(\mathbf{x}_\star) + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi\mathbf{p})\mathbf{p} \\
&= f(\mathbf{x}_\star) + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi\mathbf{p})\mathbf{p},
\end{aligned}
$$

for some $\xi \in (0, 1)$. Since $\mathbf{x}_\star + \xi\mathbf{p} \in \mathcal{B}$, we have

$$
\mathbf{p}^\top \nabla^2 f(\mathbf{x}_\star + \xi\mathbf{p})\mathbf{p} > 0,
$$

and therefore $f(\mathbf{x}_\star + \mathbf{p}) > f(\mathbf{x}_\star)$, giving the result. □

- A point $\mathbf{x}$ is called a *stationary point* if

$$\nabla f(\mathbf{x}) = \mathbf{0}.$$

- A stationary point $\mathbf{x}$ is called a *saddle point* if there exist $\mathbf{u}$ and $\mathbf{v}$ such that

$$f(\mathbf{x} + \alpha\mathbf{u}) < f(\mathbf{x}) \quad \text{and} \quad f(\mathbf{x} + \alpha\mathbf{v}) > f(\mathbf{x})$$

  for all sufficiently small $\alpha > 0$.
- Stationary points are not necessarily local minimizers. Stationary points can be *local maximizers* or *saddle points*.
- If $\nabla f(\mathbf{x}) = \mathbf{0}$, and $\nabla^2 f(\mathbf{x})$ has both strictly positive and strictly negative eigenvalues, then $\mathbf{x}$ is a saddle point.
- If $\nabla^2 f(\mathbf{x})$ is positive semidefinite or negative semidefinite, then $\nabla^2 f(\mathbf{x})$ alone is insufficient to classify $\mathbf{x}$.

4. Line search methods

- Consider an iterative method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k, \quad k = 0, 1, 2, \ldots,$$

  where $\mathbf{d}_k$ is the *direction* and $t_k > 0$ is the *stepsize*.
- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. A nonzero vector $\mathbf{d} \in \mathbb{R}^n$ is called a *descent direction* of f at $\mathbf{x}$ if the directional derivative $f'(\mathbf{x}; \mathbf{d})$ is negative, meaning that

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^\top \mathbf{d} < 0.$$

Lemma 5 (descent property of descent directions)

Let f be a continuously differentiable function over an open set U, and let $\mathbf{x} \in U$. Suppose that $\mathbf{d}$ is a descent direction of f at $\mathbf{x}$. Then there exists $\varepsilon > 0$ such that

$$f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x}) \quad \textit{for any} \quad t \in (0, \varepsilon].$$

4.1 Choices for stepsize selection rules

- Assume that $\mathbf{d}_k$ is a descent direction. Three popular choices:

  (1) **constant**. $t_k = \bar{t} > 0$ for any k

  (2) **exact line search**. t_k is a minimizer of f along the ray $\mathbf{x}_k + t\mathbf{d}_k$, i.e.,

  $$t_k \in \underset{t \geq 0}{\operatorname{argmin}} f(\mathbf{x}_k + t\mathbf{d}_k)$$

  (3) **backtracking**. Three parameters $s > 0$, $\alpha \in (0,1)$, $\beta \in (0,1)$. First, set $t_k = s$. Then, while

  $$f(\mathbf{x}_k) - f(\mathbf{x}_k + t_k\mathbf{d}_k) < -\alpha t_k \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k,$$

  set $t_k \leftarrow \beta t_k$. In other words, $t_k = s\beta^{i_k}$, where i_k is the smallest nonnegative integer satisfying (the sufficient decrease condition)

  $$f(\mathbf{x}_k) - f(\mathbf{x}_k + s\beta^{i_k}\mathbf{d}_k) \geq -\alpha s\beta^{i_k} \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k.$$

Lemma 6 (validity of the sufficient decrease condition)

Let f be a continuously differentiable function over $\mathbb{R}^n$. Suppose that $\mathbf{0} \neq \mathbf{d} \in \mathbb{R}^n$ is a descent direction of f at $\mathbf{x}$ and let $\alpha \in (0,1)$. Then there exists $\varepsilon > 0$ such that the inequality

$$f(\mathbf{x}) - f(\mathbf{x} + t\mathbf{d}) \geq -\alpha t \nabla f(\mathbf{x})^\top \mathbf{d}$$

holds for all $t \in [0, \varepsilon]$.

Proof. It follows from $\mathbf{d}$ is a descent direction that

$$\lim_{t \to 0^+} \frac{(1-\alpha) t \nabla f(\mathbf{x})^\top \mathbf{d} + o(t)\|\mathbf{d}\|}{t} = (1-\alpha)\nabla f(\mathbf{x})^\top \mathbf{d} < 0.$$

Hence, there exists $\varepsilon > 0$ such that for all $t \in (0, \varepsilon]$ the inequality $(1-\alpha) t \nabla f(\mathbf{x})^\top \mathbf{d} + o(t)\|\mathbf{d}\| < 0$ holds. The statement follows from

$$f(\mathbf{x}) - f(\mathbf{x} + t\mathbf{d}) = -\alpha t \nabla f(\mathbf{x})^\top \mathbf{d} - (1-\alpha) t \nabla f(\mathbf{x})^\top \mathbf{d} - o(t)\|\mathbf{d}\|. \quad \square$$

4.2 The gradient method

- Set $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$, the steepest descent direction.

Proposition 7

Let f be a continuously differentiable function over $\mathbb{R}^n$, and let $\mathbf{x}$ be a nonstationary point ($\nabla f(\mathbf{x}) \neq \mathbf{0}$). Then we have

$$-\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|} = \operatorname*{argmin}_{\mathbf{d}\in\mathbb{R}^n, \|\mathbf{d}\|=1} \nabla f(\mathbf{x})^\top \mathbf{d}.$$

Proposition 8 ("zig-zag")

Let $\{\mathbf{x}_k\}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a continuously differentiable function f. Then for any $k = 0, 1, 2, \ldots,$

$$(\mathbf{x}_{k+2} - \mathbf{x}_{k+1})^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) = 0.$$

- We assume that f is continuously differentiable and that ∇f is Lipschitz continuous over $\mathbb{R}^n$: there exists $L > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

- Notation: $C_L^{1,1}(\mathbb{R}^n)$, $C^{1,1}(\mathbb{R}^n)$, $C_L^{1,1}(D)$, $C^{1,1}(D)$

Theorem 9

Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then $f \in C_L^{1,1}(\mathbb{R}^n) \Leftrightarrow \|\nabla^2 f(\mathbf{x})\| \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$.

Lemma 10 (descent lemma)

Let $D \subseteq \mathbb{R}^n$ and $f \in C_L^{1,1}(D)$ for some $L > 0$. Then for any $\mathbf{x}, \mathbf{y} \in D$ satisfying $[\mathbf{x}, \mathbf{y}] \subseteq D$ it holds that

$$-\frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Lemma 11 (sufficient decrease lemma)

Suppose that $f \in C_L^{1,1}(\mathbb{R}^n)$. *Then for any* $\mathbf{x} \in \mathbb{R}^n$ *and* $t > 0$, *we have*

$$f(\mathbf{x}) - f(\mathbf{x} - t\nabla f(\mathbf{x})) \geq t(1 - tL/2)\|\nabla f(\mathbf{x})\|^2.$$

Lemma 12 (sufficient decrease of the gradient method)

Let $f \in C_L^{1,1}(\mathbb{R}^n)$. *Let* $\{\mathbf{x}_k\}$ *be the sequence generated by the gradient method for solving* $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$ *with one of the following stepsize strategies: constant stepsize* $\bar{t} \in (0, 2/L)$, *exact line search, backtracking procedure with parameters* $s > 0$, $\alpha \in (0,1)$, *and* $\beta \in (0,1)$. *Then*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq M\|\nabla f(\mathbf{x}_k)\|^2,$$

where

$$M = \begin{cases} \bar{t}(1 - \bar{t}L/2), & \textit{constant stepsize}, \\ 1/(2L), & \textit{exact line search}, \\ \alpha \min\{s, 2(1-\alpha)\beta/L\}, & \textit{backtracking}. \end{cases}$$

Theorem 13 (convergence of the gradient method)

Let $f \in C_L^{1,1}(\mathbb{R}^n)$*. Let* $\{\mathbf{x}_k\}$ *be the sequence generated by the gradient method for solving*

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies: constant stepsize $\bar{t} \in (0, 2/L)$*, exact line search, backtracking procedure with parameters* $s > 0$*,* $\alpha \in (0,1)$*, and* $\beta \in (0,1)$*. Assume that* f *is bounded below over* $\mathbb{R}^n$*, that is, there exists* $m \in \mathbb{R}$ *such that* $f(\mathbf{x}) \geq m$ *for all* $\mathbf{x} \in \mathbb{R}^n$*. Then we have the following:*

(a) *The sequence* $\{f(\mathbf{x}_k)\}$ *is nonincreasing. In addition, for any* $k > 0$*,* $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ *unless* $\nabla f(\mathbf{x}_k) = 0$*.*

(b) *The sequence* $\{f(\mathbf{x}_k)\}$ *converges, and* $\nabla f(\mathbf{x}_k) \to \mathbf{0}$ *as* $k \to \infty$*.*

(c) *Let* $f_\star = \lim_{k\to\infty} f(\mathbf{x}_k)$*. Then*

$$\min_{k=0,1,\dots,n} \|\nabla f(\mathbf{x}_k)\| \leq \sqrt{\frac{f(\mathbf{x}_0) - f_\star}{M(n+1)}}.$$

4.3 The scaled gradient method

- Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be nonsingular. Consider the equivalent problem

$$\min\{g(\mathbf{y}) \equiv f(\mathbf{S}\mathbf{y}) : \mathbf{y} \in \mathbb{R}^n\}.$$

We have $\nabla g(\mathbf{y}) = \mathbf{S}^\top \nabla f(\mathbf{S}\mathbf{y}) = \mathbf{S}^\top \nabla f(\mathbf{x})$. The gradient method takes the form

$$\mathbf{y}_{k+1} = \mathbf{y}_k - t_k \mathbf{S}^\top \nabla f(\mathbf{S}\mathbf{y}_k).$$

Multiplying by $\mathbf{S}$ from the left and using the notation $\mathbf{x}_k = \mathbf{S}\mathbf{y}_k$ and $\mathbf{D} = \mathbf{S}\mathbf{S}^\top$ yield

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{D} \nabla f(\mathbf{x}_k).$$

The direction $-\mathbf{D}\nabla f(\mathbf{x}_k)$ is a descent direction.

- It is often beneficial to choose the scaling matrix $\mathbf{D}$ differently at each iteration:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k \mathbf{D}_k \nabla f(\mathbf{x}_k).$$

4.4 Newton's method

- We assume that f is twice continuously differentiable. Given $\mathbf{x}_k$,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k).$$

  If $\nabla^2 f(\mathbf{x}_k)$ is positive definite, then $\mathbf{x}_{k+1}$ is the minimizer of the following quadratic approximation of f around $\mathbf{x}_k$:

$$f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k).$$

- Damped Newton's method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k (\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k),$$

  where t_k is the stepsize.
- Hybrid gradient-Newton method:

$$\mathbf{d}_k = \begin{cases} -(\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k), & \text{if } \nabla^2 f(\mathbf{x}_k) \text{ is pd}, \\ -\nabla f(\mathbf{x}_k), & \text{otherwise}. \end{cases}$$

Theorem 14 (quadratic local convergence of Newton's method)

Suppose $f(\mathbf{x})$ *is twice Lipschitz continuously differentiable with Lipschitz constant* $M > 0$, *i.e.,*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|.$$

Suppose that (the second-order sufficient conditions)

$$\nabla f(\mathbf{x}_\star) = \mathbf{0}, \quad \textit{and} \ \ \nabla^2 f(\mathbf{x}_\star) \succeq \gamma \mathbf{I} \ \ \textit{for some} \ \ \gamma > 0,$$

which ensure that $\mathbf{x}_\star$ *is a local minimizer of* $f(\mathbf{x})$. *If*

$$\|\mathbf{x}_0 - \mathbf{x}_\star\| \leq \frac{\gamma}{2M},$$

then the sequence $\{\mathbf{x}_k\}_0^\infty$ *in Newton's method converges to* $\mathbf{x}_\star$ *at a quadratic rate, with*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_\star\| \leq \frac{M}{\gamma}\|\mathbf{x}_k - \mathbf{x}_\star\|^2, \quad k = 0, 1, 2, \ldots.$$

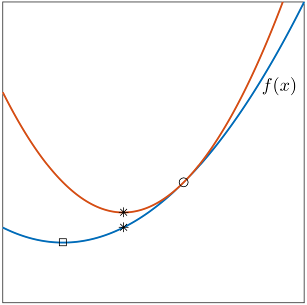
4.4.1 Geometric intuitions via quadratic approximations

- Gradient method:

$$f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2t_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2$$

- Newton's method:

$$f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)$$



4.4.2 Steepest descent, CG, and Newton's method

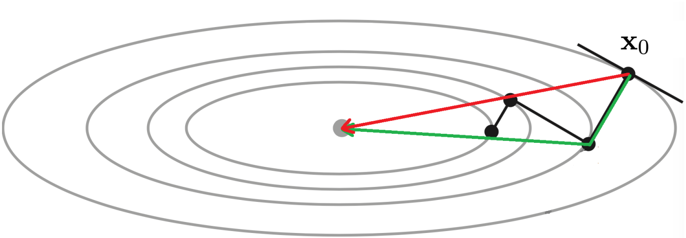
- Given an SPD matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$:

$$\mathbf{A}^{-1}\mathbf{b} = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

- Steepest descent:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{(\mathbf{A}\mathbf{x}_k - \mathbf{b})^\top(\mathbf{A}\mathbf{x}_k - \mathbf{b})}{(\mathbf{A}\mathbf{x}_k - \mathbf{b})^\top \mathbf{A}(\mathbf{A}\mathbf{x}_k - \mathbf{b})}(\mathbf{A}\mathbf{x}_k - \mathbf{b})$$

- Newton's method: $\mathbf{x}_1 = \mathbf{x}_0 - \mathbf{A}^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b})$.



Steepest Descent　Conjugate Gradients　Newton's Method

5. Further reading

- Jorge Nocedal and Stephen J. Wright

  Numerical Optimization

  Second Edition, Springer, 2006

- Amir Beck

  Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with Python and MATLAB

  Second Edition, SIAM, 2023

- Amir Beck

  First-Order Methods in Optimization

  SIAM, 2017