CS 2601 Linear and Convex Optimization

6. Gradient descent (part 2)

Bo Jiang

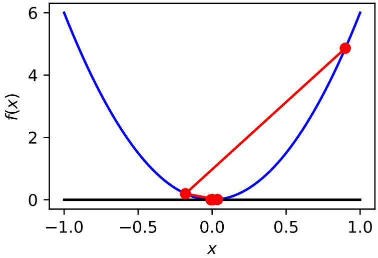
John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

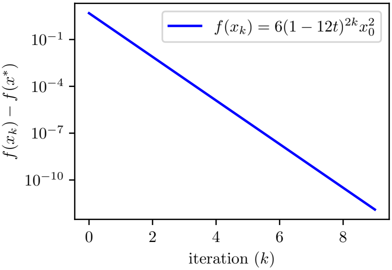
Fall 2022

Fast convergence

The following f is 12-smooth,

$$f(x) = 6x^2$$





For small enough step size t (e.g. 0.1),

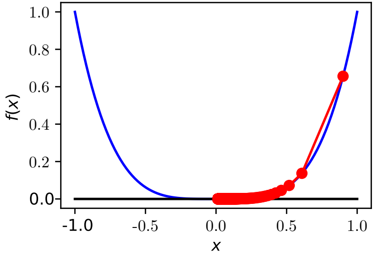
$$f(x_k) = 6x_0^2(1 - 12t)^{2k}$$

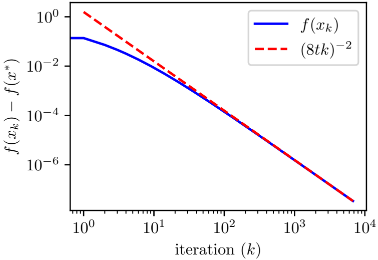
Need $O(\log \frac{1}{\epsilon})$ iterations to get within ϵ from optimal.

Slow convergence

The following f is also 12-smooth,

$$f(x) = \begin{cases} x^4, & \text{if } |x| \leq 1 \\ 4|x| - 3, & \text{if } |x| \geq 1 \end{cases}$$





For $x_0 \in (0, 1)$, small enough step size t (e.g. 0.1), and large k,

$$x_k \sim \frac{1}{\sqrt{8tk}}, \quad f(x_k) \sim \frac{1}{(8tk)^2}$$

Need $O(1/\sqrt{\epsilon})$ iterations to get within ϵ from optimal value (i.e. 0).

Strong convexity

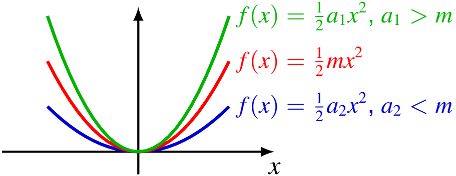
A function f is strongly convex with parameter $m > 0$, or simply m-strongly convex, if

$$\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$$

is convex.

Note. $f(\boldsymbol{x}) = \frac{m}{2}\|\boldsymbol{x}\|^2 + \tilde{f}(\boldsymbol{x})$, i.e. f is $\frac{m}{2}\|\boldsymbol{x}\|^2$ plus an extra convex term. Informally, "m-strongly convex" means at least as "convex" as $\frac{m}{2}\|\boldsymbol{x}\|^2$.

Example. $f(\boldsymbol{x}) = \frac{a}{2}\|\boldsymbol{x}\|^2$ is m-strongly convex iff $a \geq m$



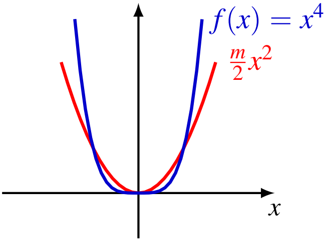
Strong convexity (cont'd) stronger than strict

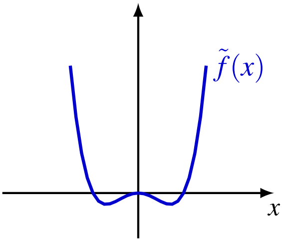
Example. $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}$ is not m-strongly convex for any $m > 0$, as $\tilde{f}(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} - \frac{m}{2}\|\boldsymbol{x}\|^2$ is concave.

Example. $f(x) = x^4$ is not m-strongly convex for any $m > 0$, as $\tilde{f}(x) = x^4 - \frac{m}{2}x^2$ is not convex,

$$\tilde{f}''(x) = 12x^2 - m < 0$$

for $|x| < \sqrt{m/12}$.

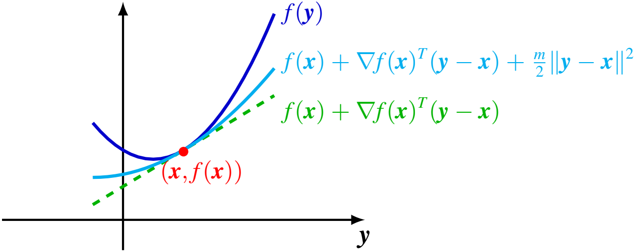




First-order condition for strong convexity

A differentiable f is m-strongly convex iff

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$



- strong convexity $\Longrightarrow$ strict convexity $\Longrightarrow$ convexity
- m-strong convexity and L-smoothness together imply

$$\frac{m}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

Proof

1. By definition,

$$f \text{ is } m\text{-strongly convex} \iff \tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2 \text{ is convex}$$

2. By first-order condition for convexity,

$$\iff \tilde{f}(\boldsymbol{y}) \geq \tilde{f}(\boldsymbol{x}) + \nabla\tilde{f}(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

3. Noting $\nabla\tilde{f}(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - m\boldsymbol{x}$,

$$\iff f(\boldsymbol{y}) - \frac{m}{2}\|\boldsymbol{y}\|^2 \geq f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2 + (\nabla f(\boldsymbol{x}) - m\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}), \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

4. Rearranging and using $\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{x} - 2\boldsymbol{x}^T(\boldsymbol{y}-\boldsymbol{x}) = (\boldsymbol{y}-\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})$,

$$\iff f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-x) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

Second-order condition for strong convexity

A twice continuously differentiable f is m-strongly convex iff

$$\nabla^2 f(\boldsymbol{x}) \succeq m\boldsymbol{I}, \quad \forall \boldsymbol{x}$$

or equivalently, the smallest eigenvalue of $\nabla^2 f(\boldsymbol{x})$ satisfies

$$\lambda_{\min}(\nabla^2 f(\boldsymbol{x})) \geq m, \quad \forall \boldsymbol{x}$$

Proof. $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$ is convex iff $\nabla^2 \tilde{f}(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - m\boldsymbol{I} \succeq \boldsymbol{O}$

Example. With $\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, we obtain $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = \frac{1}{2}x_1^2 + x_2^2$ is 1-strongly convex.

More generally, $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ with $\boldsymbol{Q} \succ \boldsymbol{O}$ is $\lambda_{\min}(\boldsymbol{Q})$-strongly convex, where $\lambda_{\min}(\boldsymbol{Q})$ is the smallest eigenvalue of $\boldsymbol{Q}$.

Bound on suboptimality gap

If f is m-strongly convex, then

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \frac{1}{2m}\|\nabla f(\boldsymbol{x})\|^2$$

Note. When $\nabla f(\boldsymbol{x})$ is small, then $f(\boldsymbol{x})$ is close to optimal. (Do we have this property for general convex functions?) No!

Proof. By the first-order condition,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$$

Minimize over $\boldsymbol{y}$ on both sides,

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{y}} f(\boldsymbol{y}) \geq \min_{\boldsymbol{y}} \left[f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2\right]$$

RHS minimized at $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{m}\nabla f(\boldsymbol{x})$, so

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}) - \frac{1}{2m}\|\nabla f(\boldsymbol{x})\|^2$$

Convergence: 1D example

$f(x) = \frac{1}{2}mx^2$ with $m > 0$ is both m-smooth and m-strongly convex..

Recall the gradient descent step is

$$x_{k+1} = x_k - tf'(x_k) = (1 - mt)x_k$$

and $x_k \to x^* = 0$ iff $t \in (0, \frac{2}{m})$.

If $t = \frac{1}{m}$, it gets to x^* in one step.

For $t \in (0, \frac{1}{m}) \cup (\frac{1}{m}, \frac{2}{m})$,

$$x_k = (1 - mt)^k x_0$$

so both $x_k \to x^*$ and $f(x_k) \to f(x^*)$ exponentially fast,

$$|x_k - x^*| = (1 - mt)^k \cdot |x_0 - x^*|$$

$$|f(x_k) - f(x^*)| = \frac{m(1 - mt)^{2k}}{2} |x_0 - x^*|^2$$

Convergence analysis

Theorem. If f is m-strongly convex and L-smooth, and $\boldsymbol{x}^*$ is a minimum of f, then for step size $t \in (0, \frac{1}{L}]$, the sequence $\{\boldsymbol{x}_k\}$ produced by the gradient descent algorithm satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{L(1-mt)^k}{2} \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$$

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leq (1-mt)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$$

Notes.

- $0 \leq 1 - \frac{m}{L} \leq 1 - mt < 1$, so $\boldsymbol{x}_k \to \boldsymbol{x}^*$ and $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}^*)$ exponentially fast
- The number of iterations to reach $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \epsilon$ is $O(\log \frac{1}{\epsilon})$. For $\epsilon = 10^{-p}$, $k = O(p)$, linear in the number of significant digits!
- Since $\nabla f(\boldsymbol{x}^*) = 0$, the bounds on slide 5 yield

$$\frac{m}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{L}{2} \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

  relating the bounds on $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$ and those on $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$

Proof

Similar to proof without strong convexity, with difference highlighted.

1. By the basic gradient step $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)$,
$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k) - \boldsymbol{x}^*\|^2 \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + t^2\|\nabla f(\boldsymbol{x}_k)\|^2 + 2t\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}^* - \boldsymbol{x}_k)\end{aligned}$$
2. By L-smoothness, the second term is upper bounded by
$$t^2\|\nabla f(\boldsymbol{x}_k)\|^2 \le 2t[f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1})]$$
3. By m-strong convexity,
$$\nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x}^* - \boldsymbol{x}_k) \le f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) - \frac{m}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$
4. Plugging 2 and 3 into 1,
$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le (1-mt)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 + 2t[f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k+1})]$$
5. Since $f(\boldsymbol{x}^*) \le f(\boldsymbol{x}_{k+1})$,
$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \le (1-mt)\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

Convergence: 2D quadratic function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}, \quad \boldsymbol{Q} = \begin{pmatrix} m & 0 \\ 0 & L \end{pmatrix}$$

where $L > m > 0$. f is L-smooth and m-strongly convex. $\boldsymbol{x}^* = \boldsymbol{0}$.

The gradient descent step is

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k) = (\boldsymbol{I} - t\boldsymbol{Q})\boldsymbol{x}_k$$

so

$$\boldsymbol{x}_k = (\boldsymbol{I} - t\boldsymbol{Q})^k\boldsymbol{x}_0 = \begin{bmatrix} (1-mt)^k x_{01} \\ (1-Lt)^k x_{02} \end{bmatrix}$$

and

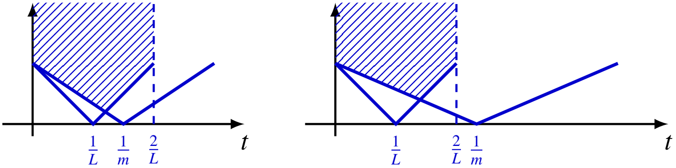
$$f(\boldsymbol{x}_k) = \frac{m}{2}(1-mt)^{2k}x_{01}^2 + \frac{L}{2}(1-Lt)^{2k}x_{02}^2$$

To ensure convergence, $t < \frac{2}{L}$. The convergence rate is determined by the slower of $(1-Lt)^{2k}$ and $(1-mt)^{2k}$.

Convergence: 2D quadratic function (cont'd)

To maximize convergence rate, solve

$$\begin{aligned} \min_t \quad & \max\{|1 - Lt|, |1 - mt|\} \\ \text{s.t.} \quad & 0 < t < 2/L \end{aligned}$$



Maximum rate achieved by $1 - mt = Lt - 1 \implies t = \frac{2}{m+L}$, in which case

$$\boldsymbol{x}_k = \left(\frac{L-m}{L+m}\right)^k \begin{bmatrix} x_{01} \\ (-1)^k x_{02} \end{bmatrix} \implies \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2 = \left(\frac{L-m}{L+m}\right)^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2$$

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) = \left(\frac{L-m}{L+m}\right)^{2k} [f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

Depends on $\kappa(\boldsymbol{Q}) = \frac{\lambda_{\max}(\boldsymbol{Q})}{\lambda_{\min}(\boldsymbol{Q})} = \frac{L}{m}$, the condition number of $\boldsymbol{Q}$

Condition number

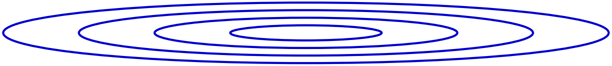
For a matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ s.t. $\boldsymbol{Q} \succ \boldsymbol{O}$, its condition number[1] is defined as

$$\kappa(\boldsymbol{Q}) = \frac{\lambda_{\max}(\boldsymbol{Q})}{\lambda_{\min}(\boldsymbol{Q})}$$

It characterizes how stretched the level curves of $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ are.

Example. $\boldsymbol{Q} = \text{diag}\{\gamma, 1\}, f(x_1, x_2) = \frac{\gamma}{2}x_1^2 + \frac{1}{2}x_2^2$

$\boldsymbol{Q} = \text{diag}\{1, 1\}$
$\kappa(\boldsymbol{Q}) = 1$



$\boldsymbol{Q} = \text{diag}\{0.01, 1\}$
$\kappa(\boldsymbol{Q}) = 100$

Nondiagonal case reduces to diagonal case in eigenbasis of $\boldsymbol{Q}$.

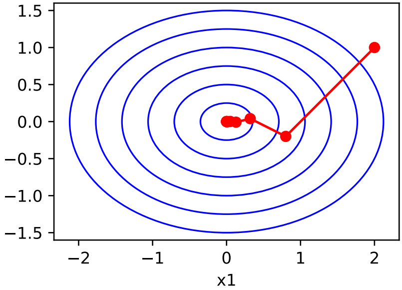
For nonquadratic case, $\kappa(\nabla^2 f(\boldsymbol{x}))$ plays a similar role.

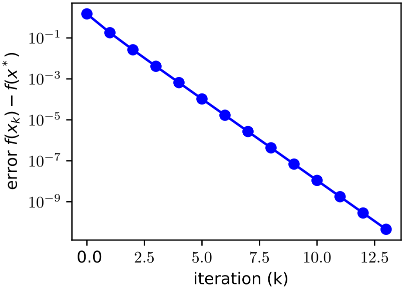
[1] For a general nonsingular matrix, the condition number is the ratio between its largest and smallest singular values, $\kappa(\boldsymbol{A}) = \sigma_{\max}(\boldsymbol{A})/\sigma_{\min}(\boldsymbol{A})$.

Well-conditioned Problem

The problem $\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}$ is well-conditioned if $\kappa(\boldsymbol{Q})$ is small.

Example. $\boldsymbol{Q} =$ diag$\{0.5, 1\}, f(x_1, x_2) = \frac{1}{4}x_1^2 + \frac{1}{2}x_2^2$, $\kappa(\boldsymbol{Q}) = 2$





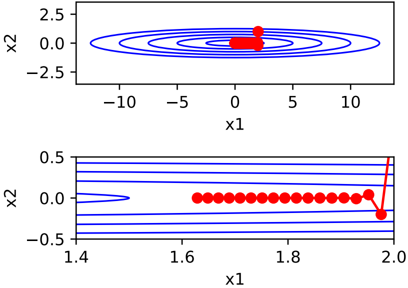
Fast convergence: for $\boldsymbol{x}_0 = (2, 1)^T, t = 1.2$, and large k,

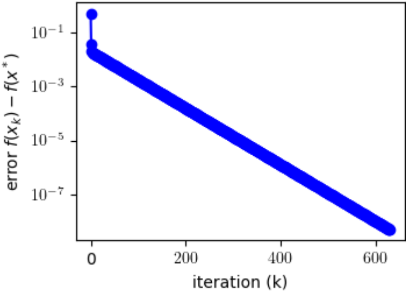
$$f(\boldsymbol{x}_k) \sim \frac{m}{2}(1 - mt)^{2k}x_{01}^2 = (0.4)^{2k}$$

Ill-conditioned problem

The problem $\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q}\boldsymbol{x}$ is ill-conditioned if $\kappa(\boldsymbol{Q})$ is large.

Example. $\boldsymbol{Q} = \text{diag}\{0.01, 1\}, f(x_1, x_2) = \frac{1}{200}x_1^2 + \frac{1}{2}x_2^2,\ \kappa(\boldsymbol{Q}) = 100$





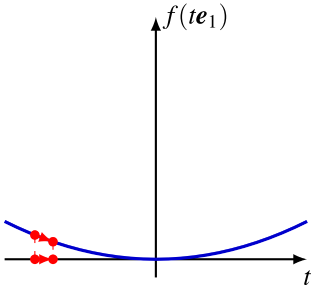
Slow convergence (relatively): for $\boldsymbol{x}_0 = (2,1)^T, t = 1.2$, and large k,

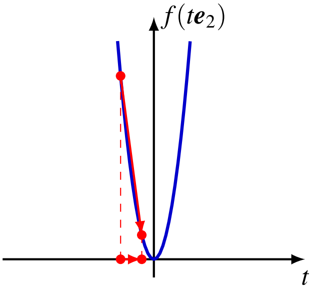
$$f(\boldsymbol{x}_k) \sim \frac{m}{2}(1 - mt)^{2k} x_{01}^2 = \frac{1}{50}(0.988)^{2k}$$

Ill-conditioned problem (cont'd)

$$f(x_1, x_2) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = \frac{1}{200}x_1^2 + \frac{1}{2}x_2^2, \quad \boldsymbol{Q} = \text{diag}\{0.01, 1\}, \ \kappa(\boldsymbol{Q}) = 100$$

- 1-smooth $\implies$ To guarantee convergence, step size[2] $t < 2$
- This limit is imposed by movement along $\boldsymbol{e}_2$ direction
- Too pessimistic along other directions, e.g. along $\boldsymbol{e}_1$, can use $t < 200$





[2]We proved convergence for $t \in (0, 1/L]$. The proofs can be modified slightly to show convergence for $t \in (0, 2/L)$.

Ill-condition problem (cont'd)

The negative gradient direction is far away from the "ideal" direction for ill-conditioned problem.

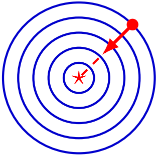
For $\boldsymbol{Q} = \text{diag}\{\gamma, 1\}$, $f(x_1, x_2) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} = \frac{\gamma}{2}x_1^2 + \frac{1}{2}x_2^2$,

negative gradient direction

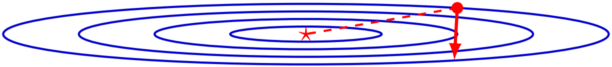
$$-\nabla f(\boldsymbol{x}) = -\boldsymbol{Q}\boldsymbol{x} = (-\gamma x_1, -x_2)^T$$

"ideal" direction

$$\boldsymbol{d} = -\boldsymbol{x}$$



$\boldsymbol{Q} = \text{diag}\{1, 1\}$
$\kappa(\boldsymbol{Q}) = 1$



$\boldsymbol{Q} = \text{diag}\{0.01, 1\}$
$\kappa(\boldsymbol{Q}) = 100$