CS 2601 Linear and Convex Optimization

6. Gradient descent (part 3)

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

Fall 2022

Step size

Gradient descent

$$\boxed{\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)}$$

- constant step size: $t_k = t$ for all k
- exact line search: optimal t_k for each step

$$t_k = \arg\min_s f(\boldsymbol{x}_k - s\nabla f(\boldsymbol{x}_k))$$

- backtracking line search (Armijo's rule): t_k satisfies

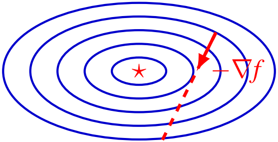
$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)) \geq \alpha t_k \|\nabla f(\boldsymbol{x}_k)\|_2^2$$

for some given $\alpha \in (0, 1)$.

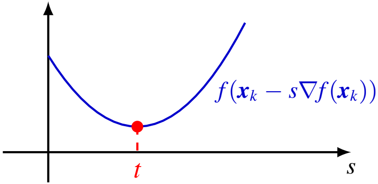
Exact line search

initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in \mathbb{R}^n$
while $\|\nabla f(\boldsymbol{x})\| > \delta$ **do**
    $t \leftarrow \arg\min_s f(\boldsymbol{x} - s\nabla f(\boldsymbol{x}))$
    $\boldsymbol{x} \leftarrow \boldsymbol{x} - t\nabla f(\boldsymbol{x})$
end while
return $\boldsymbol{x}$

Find a t for each iteration,
But this step could cost a lot



level curves of $f(x_1, x_2) = \frac{x_1^2}{4} + x_2^2$



Note. Often impractical; used only if the inner minimization is cheap.

Exact line search for quadratic functions

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}, \quad \boldsymbol{Q} \succ \boldsymbol{O}$$

- gradient at $\boldsymbol{x}_k$ is $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k) = \boldsymbol{Q}\boldsymbol{x}_k + \boldsymbol{b}$
- second-order Taylor expansion is exact for quadratic functions,

$$\begin{aligned} h(t) &= f(\boldsymbol{x}_k - t\boldsymbol{g}_k) \\ &= f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T(-t\boldsymbol{g}_k) + \frac{1}{2}(-t\boldsymbol{g}_k)^T\nabla^2 f(\boldsymbol{x}_k)(-t\boldsymbol{g}_k) \\ &= \left(\frac{1}{2}\boldsymbol{g}_k^T\boldsymbol{Q}\boldsymbol{g}_k\right)t^2 - \boldsymbol{g}_k^T\boldsymbol{g}_k t + f(\boldsymbol{x}_k) \end{aligned}$$

- minimizing $h(t)$ yields best step size

$$t_k = \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T\boldsymbol{Q}\boldsymbol{g}_k}$$

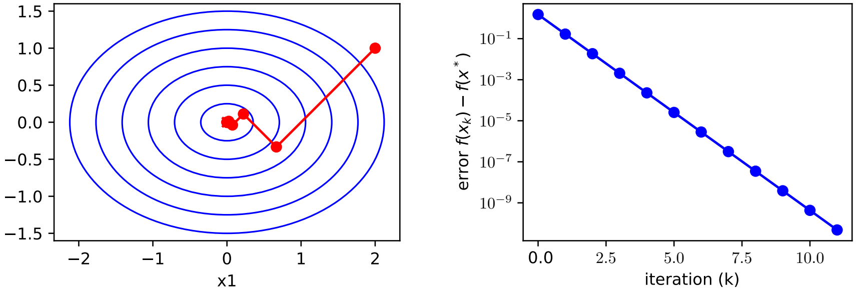
- update step

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - t_k\boldsymbol{g}_k = \boldsymbol{x}_k - \frac{\boldsymbol{g}_k^T\boldsymbol{g}_k}{\boldsymbol{g}_k^T\boldsymbol{Q}\boldsymbol{g}_k}\boldsymbol{g}_k$$

Example

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

Well-conditioned. $\gamma = 0.5$, $\boldsymbol{x}_0 = (2, 1)^T$



Fast convergence.

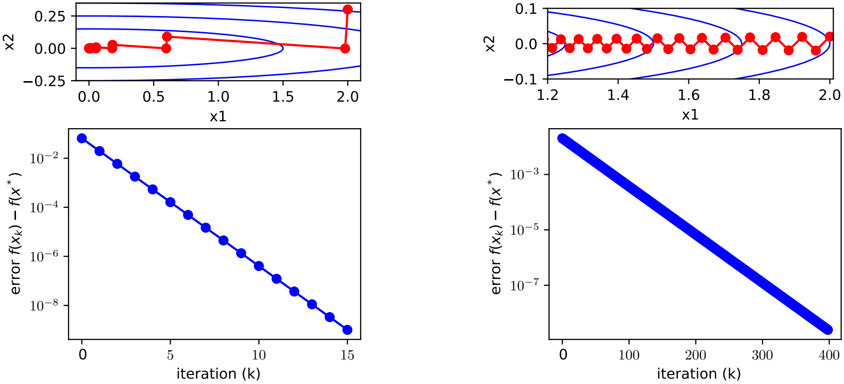
Note. Successive gradient directions are always orthogonal, as

$$0 = h'(t_k) = -\nabla f(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))^T \nabla f(\boldsymbol{x}_k) = -\nabla f(\boldsymbol{x}_{k+1})^T \nabla f(\boldsymbol{x}_k)$$

Example (cont'd)

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

Ill-conditioned. $\gamma = 0.01$, convergence rate depends on initial point



$\boldsymbol{x}_0 = (2, 0.3)$, fast convergence $\qquad$ $\boldsymbol{x}_0 = (2, 0.02)$, slow convergence

Convergence analysis

Theorem. If f is m-strongly convex and L-smooth, and $\boldsymbol{x}^*$ is a minimum of f, then the sequence $\{\boldsymbol{x}_k\}$ produced by gradient descent with exact line search satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{m}{L}\right)^k [f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

Notes.

- $0 \leq 1 - \frac{m}{L} < 1$, so $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}^*)$ exponentially fast
- $\frac{m}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$ by strong convexity, so $\boldsymbol{x}_k \to \boldsymbol{x}^*$ 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.
- The convergence rate depends on the condition number L/m and can be slow if L/m is large. When close to $\boldsymbol{x}^*$, we can estimate L/m by $\kappa(\nabla f^2(\boldsymbol{x}^*))$.

Proof

Similar to slide 12 of §6 part 2, with a modified first step (highlighted).

1. Lower bound the improvement in the k-th iteration
   - By the quadratic upper bound for L-smooth functions,
   $$f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)) - f(\boldsymbol{x}_k) \leq -t(1 - \frac{Lt}{2})\|\nabla f(\boldsymbol{x}_k)\|^2$$
   - Minimize over t on both sides,
   $$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{x}_k)\|^2 \tag{†}$$
2. Upper bound the suboptimality gap (slide 8 of §6 part 2),
$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{1}{2m}\|\nabla f(\boldsymbol{x}_k)\|^2 \tag{‡}$$
3. Eliminate $\|\nabla f(\boldsymbol{x}_k)\|$ from (†) and (‡),
$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{m}{L}\right)[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)]$$

Backtracking line search

Exact line search is often expensive and not worth it. Suffices to find a good enough step size. One way to do so is to use backtracking line search, aka Armijo's rule.

Gradient descent with backtracking line search

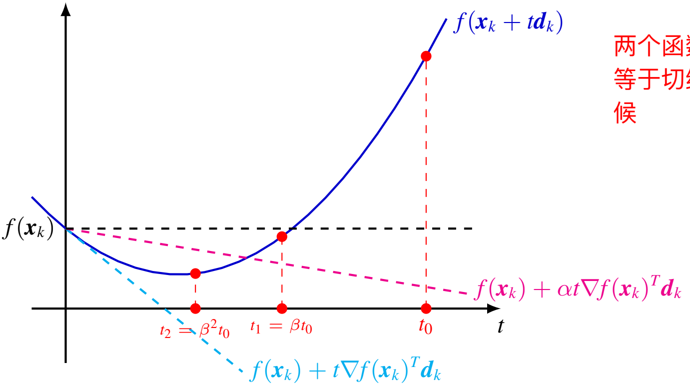
initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in \mathbb{R}^n$
while $\|\nabla f(\boldsymbol{x})\| > \delta$ **do**
    $t \leftarrow t_0$ X_{k+1}   Negative gradient
    while $f(\boldsymbol{x} - t\nabla f(\boldsymbol{x})) > f(\boldsymbol{x}) - \alpha t \|\nabla f(\boldsymbol{x})\|_2^2$ **do**
        $t \leftarrow \beta t$
    end while
    $\boldsymbol{x} \leftarrow \boldsymbol{x} - t\nabla f(\boldsymbol{x})$
end while
return $\boldsymbol{x}$

$\alpha \in (0,1)$ and $\beta \in (0,1)$ are constants. Armijo used $\alpha = \beta = 0.5$

Values suggested in [BV]: $\alpha \in [0.01, 0.3]$, $\beta \in [0.1, 0.8]$

Note. For general $\boldsymbol{d}$, use condition $f(\boldsymbol{x} + t\boldsymbol{d}) > f(\boldsymbol{x}) + \alpha t \nabla f(\boldsymbol{x})^T \boldsymbol{d}$

Backtracking line search (cont'd)



两个函数值的差距等于切线距离的时候

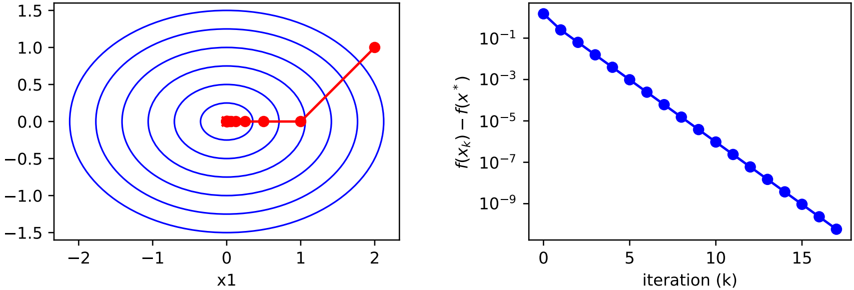
- $\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k < 0$ for descent direction $\boldsymbol{d}_k$
- start from some "large" step size t_0 ([BV] uses $t_0 = 1$)
- reduce step size geometrically until decrease is "large enough"

$$\underbrace{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + t\boldsymbol{d}_k)}_{\text{actual decrease in function value}} \geq \alpha \times \underbrace{t|\nabla f(\boldsymbol{x}_k)^T \boldsymbol{d}_k|}_{\text{decrease along tangent line}}$$

Example

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

Well-conditioned. $\gamma = 0.5$, $\boldsymbol{x}_0 = (2, 1)^T$

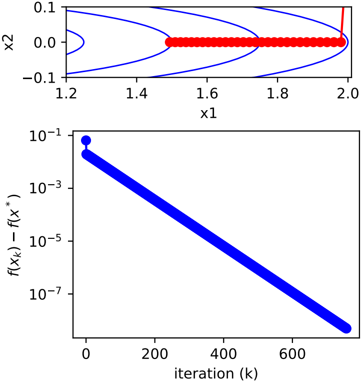


Fast convergence.

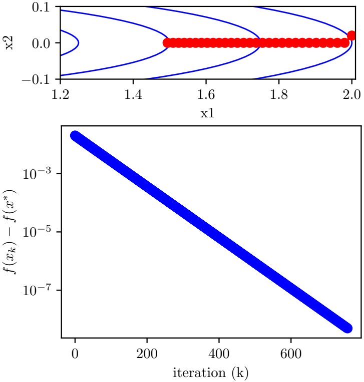
Example (cont'd)

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

Ill-conditioned. $\gamma = 0.01$



$\boldsymbol{x}_0 = (2, 0.3)$, slow convergence



$\boldsymbol{x}_0 = (2, 0.02)$, slow convergence

Convergence analysis

Theorem. If f is m-strongly convex and L-smooth, and $\boldsymbol{x}^*$ is a minimum of f, then the sequence $\{\boldsymbol{x}_k\}$ produced by gradient descent with backtracking line search satisfies

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq c^k [f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

where

$$c = 1 - \min\left\{2m\alpha t_0, \frac{4m\beta\alpha(1-\alpha)}{L}\right\}$$

Notes.

最大特征值大于最小特征值

- $c \in (0,1)$, as
$$\frac{4m\beta\alpha(1-\alpha)}{L} \leq \frac{\beta m}{L} \leq \beta < 1$$
so $\boldsymbol{x}_k \to \boldsymbol{x}^*$ and $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}^*)$ exponentially fast
- 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.

Proof

The inner loop terminates with a step size bounded from below.

1. By the quadratic upper bound for L-smooth functions,

$$f(\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)) \le f(\boldsymbol{x}_k) - t(1 - \frac{Lt}{2})\|\nabla f(\boldsymbol{x}_k)\|^2$$

2. The inner loop terminates for sure if

$$-t(1 - \frac{Lt}{2})\|\nabla f(\boldsymbol{x}_k)\|^2 \le -\alpha t\|\nabla f(\boldsymbol{x}_k)\|^2 \implies t \le \frac{2(1-\alpha)}{L}$$

3. The step size in backtracking line search satisfies

initial value

$$t_k \ge \eta \triangleq \min\left\{t_0, \frac{2\beta(1-\alpha)}{L}\right\}$$

- $t_k = t_0$ if Armijo's condition is satisfied by t_0
- otherwise, $\frac{t_k}{\beta} > \frac{2(1-\alpha)}{L}$, since the inner loop did not terminate at $\frac{t_k}{\beta}$

上一轮循环中需要满足的条件

Proof (cont'd)

Now we look at the outer loop

4. Lower bound the improvement in the k-th iteration
   - By Armijo's condition in the inner loop,

   $$f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)) \leq f(\boldsymbol{x}_k) - \alpha t_k \|\nabla f(\boldsymbol{x}_k)\|^2$$

   - Since $t_k \geq \eta$ by step 3,

   $$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) - \alpha\eta \|\nabla f(\boldsymbol{x}_k)\|^2$$

5. Upper bound the suboptimality gap by (‡) of slide 7,

   $$\|\nabla f(\boldsymbol{x}_k)\|^2 \geq 2m[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)]$$

6. Eliminate $\|\nabla f(\boldsymbol{x}_k)\|$ from steps 4 and 5,

   $$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq (1 - 2m\alpha\eta)[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)] = c[f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)]$$

   so

   $$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq c^k[f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*)]$$

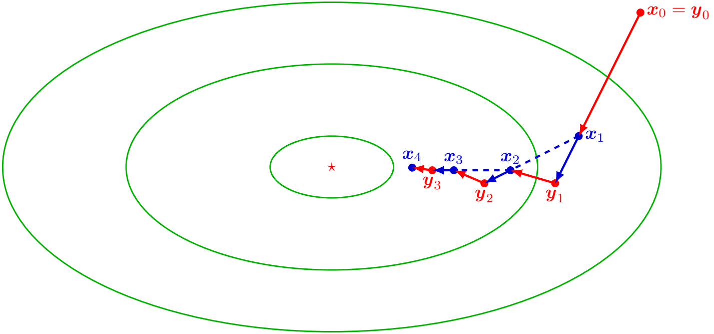
Nesterov's accelerated gradient descent (AGD)

Suppose f is L-smooth and m-strongly convex ($m \geq 0$)

1: initialize $\boldsymbol{x}_0 = \boldsymbol{y}_0$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad \boldsymbol{x}_{k+1} = \boldsymbol{y}_k - \frac{1}{L}\nabla f(\boldsymbol{y}_k)$
4: $\quad \boldsymbol{y}_{k+1} = \boldsymbol{x}_{k+1} + \beta_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$
5: **end for**

? ? ? ? ? ?

可以把时间复杂度指数前的常数开根号



m-strongly convex L-smooth

$$1-x \le e^{-x}$$

$m=0$ $\quad O\left(\frac{1}{k}\right) \quad k = O\left(\frac{1}{\varepsilon}\right)$

$\quad O\left(\frac{1}{k^2}\right) \quad k = O\left(\frac{1}{\sqrt{\varepsilon}}\right)$

$m>0$ $\quad O\left(\left(1-\frac{m}{L}\right)^k\right) \quad k = O\left(\frac{L}{m}\log\frac{1}{\varepsilon}\right)$

$\quad O\left(\left(1-\sqrt{\frac{m}{L}}\right)^k\right) \quad k = O\left(\sqrt{\frac{L}{m}}\log\frac{1}{\varepsilon}\right)$

Accelerated gradient descent (cont'd)

Theorem. Suppose f is L-smooth and m-strongly convex. Let $q = m/L$, $\alpha_0 \in [\sqrt{q}, 1)$, $\alpha_{k+1} = \frac{\sqrt{(\alpha_k^2 - q)^2 + 4\alpha_k^2} + q - \alpha_k^2}{2}$ for $k \geq 0$. If $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2 + \alpha_{k+1}}$ for $k \geq 0$, then

$$f(\boldsymbol{x}_k) - f^* \leq \min\left\{(1-\sqrt{q})^k, \frac{4}{(2+k\sqrt{\gamma_0})^2}\right\}\left(f(\boldsymbol{x}_0) - f^* + \frac{\gamma_0}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2\right),$$

where $\gamma_0 = \alpha_0(\alpha_0 - q)/(1-\alpha_0)$.

Note. The convergence rate of AGD is $O(1/k^2)$ if $m = 0$ and $O((1-\sqrt{q})^k)$ if $m > 0$. Recall the rate of GD is $O(1/k)$ if $m = 0$ and $O((1-q)^k)$ if $m > 0$.

Note. Nesterov also proved lower bounds for first-order methods, i.e. there exists an L-smooth f_1, and an L-smooth and m-strongly convex f_2 s.t.

$$f_1(\boldsymbol{x}_k) - f^* = \Omega\left(\frac{1}{k^2}\right), \quad f_2(\boldsymbol{x}_k) - f^* = \Omega\left(\left(\frac{1-\sqrt{q}}{1+\sqrt{q}}\right)^{2k}\right).$$

Nonconvex functions

GD can also be applied to nonconvex functions, but with no guarantee for optimality. It only finds an approximately stationary point.

只能找到驻点

Theorem. If f is L-smooth, then for step size $t \in (0, \frac{1}{L}]$, the sequence $\{\boldsymbol{x}_k\}$ produced by GD satisfies

$$\min_{0 \le i \le k} \|\nabla f(\boldsymbol{x}_i)\| \le \sqrt{\frac{2(f(\boldsymbol{x}_0) - f^*)}{t(k+1)}}$$

Proof. By slide 15 of §6 part 1,

$$\min_{0 \le i \le k} \|\nabla f(\boldsymbol{x}_i)\|^2 \le \|\nabla f(\boldsymbol{x}_i)\|^2 \le \frac{2}{t}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_{i+1}))$$

Summing over i from 0 to k completes the proof.

Note. The convergence rate is $O(1/\sqrt{k})$, which turns out to be optimal for deterministic algorithms for finding stationary points of functions with Lipschitz gradients.