机器学习中的优化算法

Lecture07: 复合优化算法–Nesterov加速算法

张立平

清华大学数学科学系

办公室：理科楼#A302, Tel: 62798531

E-mail: lipingzhang@tsinghua.edu.cn

Contents and Acknowledgement

- 教材：**最优化：建模、算法与理论**

  http://bicmr.pku.edu.cn/ wenzw/bigdata2021.html

- 致谢：北京大学文再文教授

Outline of Lecture07

- FISTA
- 其他加速算法
- 应用: LASSO问题
- 收敛性分析

Nesterov加速算法简史

- 复合优化问题

$$\min_{x\in\mathbb{R}^n} \quad \psi(x) = f(x) + h(x) \tag{1}$$

的近似点梯度法

$$x^{k+1} = \mathrm{prox}_{t_k h}(x^k - t_k \nabla f(x^k))$$

在基本假设下, 步长取常数$t_k = 1/L$时, 收敛速度为$\mathcal{O}(1/k)$.

- Nesterov加速算法:
  - 一个自然的问题是如果仅用梯度信息, 我们能不能取得更快的收敛速度.
  - Nesterov分别在1983年、1988年和2005年提出了三种改进的一阶算法, 收敛速度能达到$\mathcal{O}\left(1/k^2\right)$. 实际上, 这三种算法都可以应用到近似点梯度法上.
  - 在Nesterov加速算法刚提出的时候, 由于牛顿算法有更快的收敛速度, Nesterov加速算法在当时并没有引起太多的关注. 但近年来, 随着数据量

的增大, 牛顿型方法由于其过大的计算复杂度, 不便于有效地应用到实际中, Nesterov加速算法作为一种快速的一阶算法重新被挖掘出来并迅速流行起来.

- Beck和Teboulle就在2008年给出了Nesterov在1983年提出的算法的近似点梯度法版本——FISTA.
- FISTA算法由两步组成: 第一步沿着前两步的计算方向计算一个新点, 第二步在该新点处做一步近似点梯度迭代.

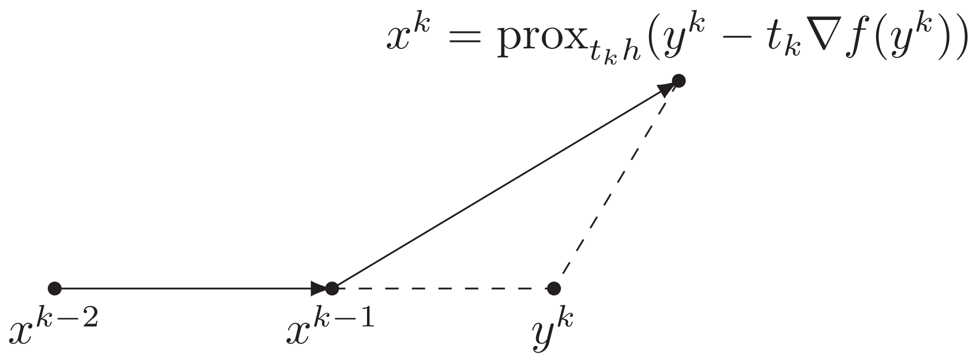


Figure 1: FISTA算法

- FISTA的迭代格式:

$$
\begin{aligned}
y^k &= x^{k-1} + \frac{k-2}{k+1}(x^{k-1} - x^{k-2}), \\
x^k &= \mathrm{prox}_{t_k h}(y^k - t_k \nabla f(y^k)), \quad t_k \in \left(0, \frac{1}{L}\right].
\end{aligned} \tag{2}
$$

- FISTA的等价变形:

$$
\begin{aligned}
y^k &= (1-\gamma_k)x^{k-1} + \gamma_k v^{k-1}, \\
x^k &= \mathrm{prox}_{t_k h}(y^k - t_k \nabla f(y^k)), \\
v^k &= x^{k-1} + \frac{1}{\gamma_k}(x^k - x^{k-1}).
\end{aligned} \tag{3}
$$

- 在(3)中, 当$\gamma_k = \frac{2}{k+1}$, 且取固定步长时, 算法(2)与(3)是等价的.
- 当γ_k采用别的取法时, 算法(3) 将给出另一个版本的加速算法.
- 也就是说, 算法(2)中$\dfrac{k-2}{k+1}$可以取其他值.

FISTA的收敛条件

►如何选取步长t_k和γ_k决定了算法(3)的收敛速度. 算法(3)以$\mathcal{O}\left(1/k^2\right)$的速度收敛的条件:

$$f(x^k) \le f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle + \frac{1}{2t_k}\|x^k - y^k\|_2^2, \tag{4}$$

$$\gamma_1 = 1, \qquad \frac{(1-\gamma_i)t_i}{\gamma_i^2} \le \frac{t_{i-1}}{\gamma_{i-1}^2}, \quad i > 1, \tag{5}$$

$$\frac{\gamma_k^2}{t_k} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{6}$$

- 当取$t_k = \dfrac{1}{L}$, $\gamma_k = \dfrac{2}{k+1}$时, 条件(4)(5)(6)满足.
- γ_k的选取并不唯一, 例如可以选取

$$\gamma_1 = 1, \quad \frac{1}{\gamma_k} = \frac{1}{2}\left(1 + \sqrt{1 + \frac{4}{\gamma_{k-1}}}\right).$$

- 算法(2)和算法(3)都要求步长满足$t_k \le \dfrac{1}{L}$, 此时条件(4)满足.
- 对绝大多数问题我们不知道函数∇f的利普希茨常数. 为了在这种情况下条件(4)依然能满足, 需要使用线搜索来确定合适的t_k.
- 线搜索算法1: 在算法(3)中加入线搜索, 取$\gamma_k = \dfrac{2}{k+1}$, 以回溯的方式找到满足条件(4)的$t_k$:

$$\textbf{重复} \quad \begin{cases} t_k \leftarrow \rho t_k \quad (\rho < 1) \\ x^k \leftarrow \mathrm{prox}_{t_k h}(y^k - t_k \nabla f(y^k)) \end{cases} \quad \textbf{直到(4)满足} \tag{7}$$

- 当t_k足够小时, 条件(4)是一定会得到满足的, 因此不会出现线搜索(7)无法终止的情况.
- 容易验证条件(5)(6)在迭代过程中也得到满足.

▶ 线搜索算法2: 在算法(3)中加入线搜索, 不仅改变步长t_k而且改变γ_k, 从而y^k也随之改变.

$$
\text{重复}\left\{\begin{array}{l}
\text{取}\gamma_k\text{为}t_{k-1}\gamma^2 = t_k\gamma_{k-1}^2(1-\gamma)\text{的正根} \\
y^k \leftarrow (1-\gamma_k)x^{k-1} + \gamma_k v^{k-1} \\
x^k \leftarrow \mathrm{prox}_{t_k h}(y^k - t_k\nabla f(y^k)) \\
t_k \leftarrow \rho t_k
\end{array}\right. \quad \text{直到(4)成立} \tag{8}
$$

- 由算法(8), γ_k满足条件(5)且有$0 < \gamma_k \le 1$, t_k有下界$t_{\min}$.
- 由$\sqrt{1-x}$在点$x=0$处的凹性,

$$
\frac{\sqrt{t_{k-1}}}{\gamma_{k-1}} = \frac{\sqrt{(1-\gamma_k)t_k}}{\gamma_k} \le \frac{\sqrt{t_k}}{\gamma_k} - \frac{\sqrt{t_k}}{2} \quad \Rightarrow \quad \frac{\sqrt{t_k}}{\gamma_k} \ge \sqrt{t_1} + \frac{1}{2}\sum_{i=2}^{k}\sqrt{t_i}
$$

$$
\Rightarrow \quad \frac{\gamma_k^2}{t_k} \le \frac{1}{\left(\sqrt{t_1} + \frac{1}{2}\sum_{i=2}^{k}\sqrt{t_i}\right)^2} \le \frac{4}{t_{min}(k+1)^2} = \mathcal{O}\left(\frac{1}{k^2}\right). \tag{9}
$$

- (9)说明条件(6)在算法(8)的执行中也得到满足.
- 算法(8)的执行过程比算法(7)的复杂. 由于它同时改变了t_k和γ_k, 迭代点x^k和参照点y^k在线搜索的过程中都发生了变化, 点y^k处的梯度也需要重新计算.
- 但此算法给我们带来的好处就是步长t_k不再单调下降, 在迭代后期也可以取较大值, 这会进一步加快收敛.

▶FISTA算法小结:

- 固定步长的FISTA算法对于步长的选取是较为保守的，为了保证收敛，有时不得不选取一个很小的步长，这使得固定步长的FISTA算法收敛较慢.
- 如果采用线搜索，则在算法执行过程中会有很大机会选择符合条件的较大步长，因此线搜索可能加快算法的收敛，但代价就是每一步迭代的复杂度变高.
- 在实际的FISTA算法中，需要权衡固定步长和线搜索算法的利弊，从而选择针对特定问题的高效算法.

▶下降的FISTA算法

- 原始的FISTA算法不是一个下降算法，这里给出一个FISTA的下降算法变形.
- 只需要对算法(3)的第2步进行修改. 在计算邻近算子之后，我们并不立即选取此点作为新的迭代点，而是检查函数值在当前点处是否下降，只有当函数值下降时才更新迭代点.
- 假设经过近似点映射之后的点为u，则对当前点x^k做如下更新：

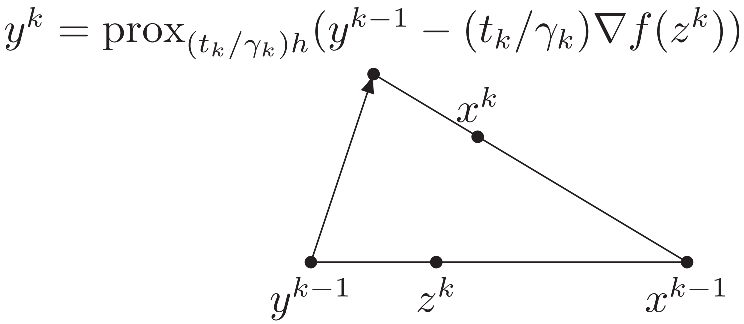
$$x^k = \begin{cases} u, & \psi(u) \le \psi(x^{k-1}), \\ x^{k-1}, & \psi(u) > \psi(x^{k-1}). \end{cases} \tag{10}$$

- 由于步长或γ_k会随着k变化，(10)式中的$\psi(u) > \psi(x^{k-1})$不会一直成立，即算法不会停留在某个x^{k-1}而不进行更新.
- 步长和γ_k的选取只需使用固定步长$t_k \le \frac{1}{L}$, $\gamma_k = \frac{2}{k+1}$或者使用前述的任意一种线搜索方法均可.

第二类Nesterov加速算法

▶复合优化问题(1)的第二类Nesterov加速算法:

$$\begin{cases} z^k = (1-\gamma_k)x^{k-1} + \gamma_k y^{k-1} \\ y^k = \mathrm{prox}_{(t_k/\gamma_k)h}\left(y^{k-1} - \frac{t_k}{\gamma_k}\nabla f(z^k)\right) \\ x^k = (1-\gamma_k)x^{k-1} + \gamma_k y^k \end{cases} \tag{11}$$



- 和FISTA 算法(3)的一个重要区别在于, 第二类Nesterov 加速算法(11)中的三个序列$\{x^k\}$, $\{y^k\}$和$\{z^k\}$都可以保证在定义域内. 而FISTA 算法中的序列$\{y^k\}$不一定在定义域内.

第三类Nesterov加速算法

▶复合优化问题(1)的第三类Nesterov加速算法:

$$\begin{cases} z^k = (1-\gamma_k)x^{k-1} + \gamma_k y^{k-1} \\ y^k = \mathrm{prox}_{(t_k \sum_{i=1}^k 1/\gamma_i)h}\left(-t_k \sum_{i=1}^k \frac{1}{\gamma_i} \nabla f(z^i)\right) \\ x^k = (1-\gamma_k)x^{k-1} + \gamma_k y^k \end{cases} \tag{12}$$

- 第三类Nesterov 加速算法(12)和第二类Nesterov加速算法(11)的区别仅仅在于y^k的更新: 第三类Nesterov 加速算法(12)计算y^k时需要利用全部已有的$\{\nabla f(z^i)\}, i = 1, 2, \cdots, k$.
- 该算法取$\gamma_k = \frac{2}{k+1}$, $t_k = \frac{1}{L}$时, 也有$\mathcal{O}\left(\frac{1}{k^2}\right)$ 的收敛速度.

Nesterov 加速算法应用: LASSO问题

▶ LASSO问题:

$$\min_x \quad \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1. \tag{13}$$

- 求解LASSO问题(13)的FISTA算法:

$$\begin{aligned} y^k &= x^{k-1} + \frac{k-2}{k+1}(x^{k-1} - x^{k-2}), \\ w^k &= y^k - t_k A^T(Ay^k - b), \\ x^k &= \mathsf{sign}(w^k)\max\{|w^k| - t_k\mu, 0\}. \end{aligned}$$

与近似点梯度算法相同, 由于最后一步将w^k中绝对值小于$t_k\mu$的分量置零, 该算法能够保证迭代过程中解具有稀疏结构.

- 求解LASSO问题(13)的第二类Nesterov加速算法:

$$
\begin{aligned}
z^k &= (1-\gamma_k)x^{k-1} + \gamma_k y^{k-1},\\
w^k &= y^{k-1} - \frac{t_k}{\gamma_k} A^T(Az^k - b),\\
y^k &= \mathrm{sign}(w^k)\max\left\{|w^k| - \frac{t_k}{\gamma_k}\mu, 0\right\},\\
x^k &= (1-\gamma_k)x^{k-1} + \gamma_k y^k.
\end{aligned}
$$

- 求解LASSO问题(13)的第三类Nesterov加速算法:

$$
\begin{aligned}
z^k &= (1-\gamma_k)x^{k-1} + \gamma_k y^{k-1},\\
w^k &= -t_k \sum_{i=1}^{k} \frac{1}{\gamma_i} A^T(Az^i - b),\\
y^k &= \mathbf{sign}(w^k)\max\left\{|w^k| - t_k \sum_{i=1}^{k} \frac{1}{\gamma_i}\mu, 0\right\},\\
x^k &= (1-\gamma_k)x^{k-1} + \gamma_k y^k.
\end{aligned}
$$

- 取$\mu = 10^{-3}$, 分别利用近似点梯度法、FISTA算法、第二类Nesterov算法来求解(13), 分别取固定步长$t = \frac{1}{L}$, 其中$L = \lambda_{\max}(A^T A)$, 和结合线搜索的BB步长.

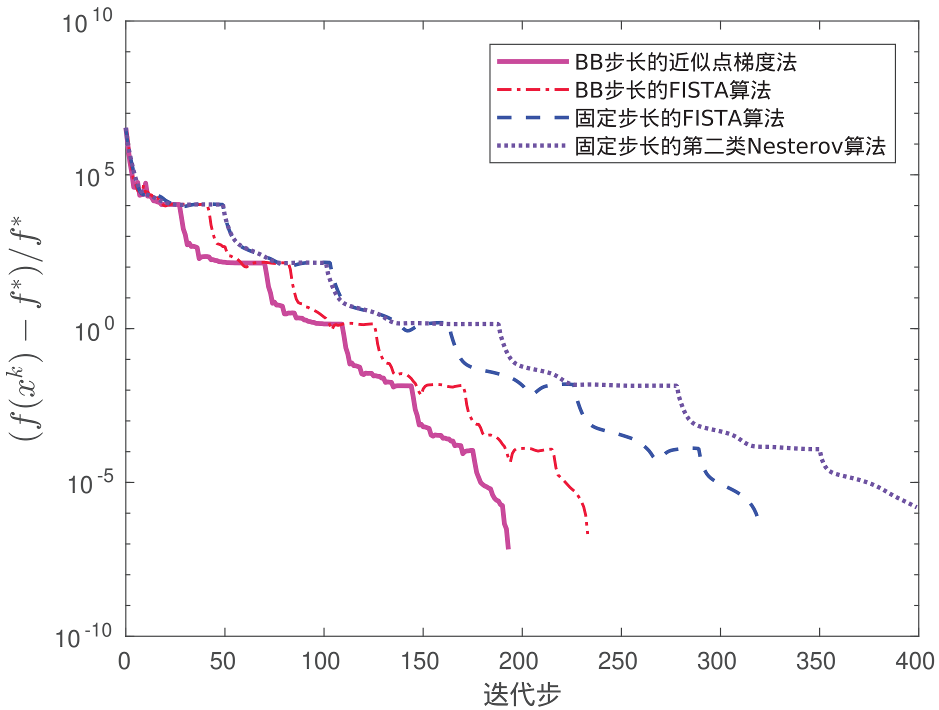


Figure 2: 固定步长, FISTA算法相较于第二类Nesterov加速算法收敛得略快一些; FISTA算法是非单调的; BB步长和线搜索技巧可以加速算法的收敛速度; 带线搜索的近似点梯度法可以比带线搜索的FISTA算法更快收敛.

固定步长FISTA算法收敛速度

收敛性假设:

- f 在其定义域$\mathbf{dom}\ f = \mathbb{R}^n$内凸, 且是$L$-光滑的

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \qquad \forall x, y \in \mathbf{dom}\ f.$$

- h是适当的闭凸函数.
- $\psi(x)$的最小值ψ^*是有限的, 且在点x^*处可以取到.

▶固定步长近似点梯度法的收敛速度为$\mathcal{O}(1/k)$, 而固定步长FISTA算法则可以加速到$\mathcal{O}(1/k^2)$.

Theorem 1. *在收敛性假设的条件下，当用FISTA算法*(3)*求解凸复合优化问题*(1) *时, 若取固定步长$t_k = \frac{1}{L}$, 则*

$$\psi(x^k) - \psi(x^*) \le \frac{2L}{(k+1)^2}\|x^0 - x^*\|^2. \tag{14}$$

定理1的证明

- 根据$x^k = \mathrm{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$, 可知

$$-x^k + y^k - t_k \nabla f(y^k) \in t_k \partial h(x^k).$$

故对于任意的x, 有

$$t_k h(x) \geq t_k h(x^k) + \langle -x^k + y^k - t_k \nabla f(y^k), x - x^k \rangle. \tag{15}$$

- 由f 的凸性、L-光滑和$t_k = \frac{1}{L}$可得,

$$f(x^k) \leq f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle + \frac{1}{2t_k} \|x^k - y^k\|^2. \tag{16}$$

- 结合(15)和(16), 对于任意的x有

$$\begin{aligned}
\psi(x^k) &= f(x^k) + h(x^k) \\
&\le h(x) + f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{1}{t_k}\langle x^k - y^k, x - x^k \rangle \\
&+ \frac{1}{2t_k}\|x^k - y^k\|^2 \\
&\le h(x) + f(x) + \frac{1}{t_k}\langle x^k - y^k, x - x^k \rangle + \frac{1}{2t_k}\|x^k - y^k\|^2 \\
&= \psi(x) + \frac{1}{t_k}\langle x^k - y^k, x - x^k \rangle + \frac{1}{2t_k}\|x^k - y^k\|^2.
\end{aligned} \tag{17}$$

- 在(17)式中分别取$x = x^{k-1}$和$x = x^*$, 并记$\psi(x^*) = \psi^*$, 再分别乘$1 - \gamma_k$和γ_k并相加得到

$$\begin{aligned}
&\psi(x^k) - \psi^* - (1 - \gamma_k)(\psi(x^{k-1}) - \psi^*) \\
\le &\frac{1}{t_k}\langle x^k - y^k, (1 - \gamma_k)x^{k-1} + \gamma_k x^* - x^k \rangle + \frac{1}{2t_k}\|x^k - y^k\|^2.
\end{aligned} \tag{18}$$

- 由$v^k = x^{k-1} + \frac{1}{\gamma_k}(x^k - x^{k-1}), \quad y^k = (1-\gamma_k)x^{k-1} + \gamma_k v^{k-1}$ 知, 不等式(18)可以化为

$$\begin{aligned}
&\psi(x^k) - \psi^* - (1-\gamma_k)(\psi(x^{k-1}) - \psi^*) \\
\le& \frac{1}{2t_k}(\|y^k - (1-\gamma_k)x^{k-1} - \gamma_k x^*\|^2 - \|x^k - (1-\gamma_k)x^{k-1} - \gamma_k x^*\|^2) \\
=& \frac{\gamma_k^2}{2t_k}(\|v^{k-1} - x^*\|^2 - \|v^k - x^*\|^2).
\end{aligned} \tag{19}$$

- t_k, γ_k的取法满足不等式$\frac{1-\gamma_k}{\gamma_k^2} t_k \le \frac{1}{\gamma_{k-1}^2} t_{k-1}$, 因此,

$$\frac{t_k}{\gamma_k^2}(\psi(x^k) - \psi^*) + \frac{1}{2}\|v^k - x^*\|^2 \le \frac{t_{k-1}}{\gamma_{k-1}^2}(\psi(x^{k-1}) - \psi^*) + \frac{1}{2}\|v^{k-1} - x^*\|^2. \tag{20}$$

- 根据(20),

$$\frac{t_k}{\gamma_k^2}(\psi(x^k)-\psi^*)+\frac{1}{2}\|v^k-x^*\|^2 \le \frac{t_1}{\gamma_1^2}(\psi(x^1)-\psi^*)+\frac{1}{2}\|v^1-x^*\|^2. \quad (21)$$

- 对$k=1$, 注意到$\gamma_1=1, v^0=x^0$, 再次利用(19)可得

$$\begin{aligned}&\frac{t_1}{\gamma_1^2}(\psi(x^1)-\psi^*)+\frac{1}{2}\|v^1-x^*\|^2\\ \le&\frac{(1-\gamma_1)t_1}{\gamma_1^2}(\psi(x^0)-\psi^*)+\frac{1}{2}\|v^0-x^*\|^2=\frac{1}{2}\|x^0-x^*\|^2.\end{aligned} \quad (22)$$

由(21)和(22)可得(14)成立.

▶**Remark:**

- 定理1的证明关键的一步在于建立(20), 而建立这个递归关系并不需要$t=1/L, \gamma_k=2/(k+1)$这一具体条件, 只需要保证条件(4)和条件(5)成立即可.
- 条件(4)主要依赖于$f(x)$的L-光滑性, (5)的成立依赖于γ_k和t_k的选取. 条件(6)的成立保证了算法(3)的收敛速度达到$\mathcal{O}\left(\frac{1}{k^2}\right)$. 如果将条件(4)-(6)作为算法收敛的假设条件, 则可以证明一大类FISTA算法的变形都具有$\mathcal{O}\left(\frac{1}{k^2}\right)$的收敛速度.

第二类Nesterov加速算法的收敛速度

Theorem 2. *利用算法(11)求解问题(1), 取$\gamma_k = \frac{2}{k+1}$ 和$t_k = \frac{1}{L}$, 设$\{x^k\}$是算法(11)产生的迭代序列, 则在收敛性假设成立的条件下, 有*

$$\psi(x^k) - \psi(x^*) \le \frac{2L}{(k+1)^2}\|x^0 - x^*\|^2. \tag{23}$$

Proof. 根据$y^k = \operatorname{prox}_{(t_k/\gamma_k)h}\left(y^{k-1} - \left(\frac{t_k}{\gamma_k}\right)\nabla f(z^k)\right)$,

$$\gamma_k(y^{k-1} - y^k) - t_k\nabla f(z^k) \in t_k\partial h(y^k),$$

故对于任意的x,有

$$t_k h(x) \ge t_k h(y^k) + \langle \gamma_k(y^{k-1} - y^k) - t_k\nabla f(z^k), x - y^k\rangle.$$

由h的凸性,

$$h(x^k) \le (1-\gamma_k)h(x^{k-1}) + \gamma_k h(y^k),$$

消去$h(y^k)$得到,

$$\begin{aligned} h(x^k) \le (1-\gamma_k)h(x^{k-1}) \\ +\gamma_k \left[h(x) - \left\langle \frac{\gamma_k}{t_k}(y^{k-1}-y^k) - \nabla f(z^k), x - y^k \right\rangle \right]. \end{aligned} \tag{24}$$

利用f 的凸性和L-光滑性,

$$\begin{aligned} f(x^k) &\le f(z^k) + \langle \nabla f(z^k), x^k - z^k \rangle + \frac{L}{2}\|x^k - z^k\|^2 \\ &= f(z^k) + \langle \nabla f(z^k), x^k - z^k \rangle + \frac{1}{2t_k}\|x^k - z^k\|^2. \end{aligned} \tag{25}$$

由算法(11)知,

$$x^k - z^k = \gamma_k(y^k - y^{k-1}), \qquad x^k = (1-\gamma_k)x^{k-1} + \gamma_k y^k.$$

于是根据(25),

$$f(x^k) \le f(z^k) + \langle \nabla f(z^k), (1-\gamma_k)x^{k-1} + \gamma_k y^k - z^k \rangle + \frac{\gamma_k^2}{2t_k}\|y^k - y^{k-1}\|^2. \quad (26)$$

注意到

$$\begin{aligned} & f(z^k) + \langle \nabla f(z^k), (1-\gamma_k)x^{k-1} + \gamma_k y^k - z^k \rangle \\ =& (1-\gamma_k)[f(z^k) + \langle \nabla f(z^k), x^{k-1} - z^k \rangle] + \gamma_k [f(z^k) + \langle \nabla f(z^k), y^k - z^k \rangle] \\ \le& (1-\gamma_k) f(x^{k-1}) + \gamma_k [f(z^k) + \langle \nabla f(z^k), y^k - z^k \rangle], \end{aligned} \quad (27)$$

由(26)和(27)得,

$$f(x^k) \le (1-\gamma_k) f(x^{k-1}) + \gamma_k [f(z^k) + \langle \nabla f(z^k), y^k - z^k \rangle] + \frac{\gamma_k^2}{2t_k}\|y^k - y^{k-1}\|^2. \quad (28)$$

将(24)与(28)相加, 并结合$f(x) \geq f(z^k) + \langle \nabla f(z^k), x - z^k \rangle$, 再取$x = x^*$, 可得

$$\begin{aligned}
&\psi(x^k) - (1-\gamma_k)\psi(x^{k-1}) \\
\leq &\gamma_k \left[h(x^*) + f(x^*) - \frac{\gamma_k}{t_k}\langle y^{k-1} - y^k, x^* - y^k \rangle \right] + \frac{\gamma_k^2}{2t_k}\|y^k - y^{k-1}\|^2 \qquad (29) \\
\leq &\gamma_k \psi(x^*) + \frac{\gamma_k^2}{2t_k}(\|y^{k-1} - x^*\|^2 - \|y^k - x^*\|^2).
\end{aligned}$$

这个不等式和(19)的形式完全相同, 因此后续过程可按照定理1 进行推导, 可以得到(23). □

▶一般第二类Nesterov加速算法的收敛速度: 用算法(11)求解凸复合优化问题(1), 若迭代点x^k, y^k, 步长t_k以及组合系数γ_k满足条件(4)—(6), 则

$$\psi(x^k) - \psi(x^*) \leq \frac{C}{k^2}, \qquad (30)$$

其中C 仅和函数f, 初始点x^0的选取有关.