Nesterov加速算法

文再文

北京大学北京国际数学研究中心

教材《最优化：建模、算法与理论》配套电子教案

http://bicmr.pku.edu.cn/~wenzw/optbook.html

致谢：本教案由李煦恒协助准备

提纲

1. FISTA算法
2. 其他加速算法
3. 应用举例
   - LASSO问题求解
   - 小波模型求解
4. 收敛性分析

典型问题形式

考虑如下复合优化问题：

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

- $f(x)$是连续可微的凸函数，且梯度是利普西茨连续的：

$$\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\|;$$

- $h(x)$是适当的闭凸函数，且邻近算子容易计算:

$$\mathrm{prox}_h(x)=\operatorname*{argmin}_{u\in \mathrm{dom} h}\left\{h(u)+\frac{1}{2}\|x-u\|^2\right\}$$

- 对于上述问题，近似点梯度法

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

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

Nesterov加速算法简史

- 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$\mathcal{O}\left(\frac{1}{k^2}\right)$. 实际上，这三种算法都可以应用到近似点梯度算法上.
- 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

- 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（如图所示）.

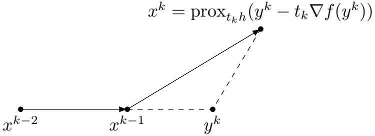

Figure: FISTA算法图示

- FISTA算法：

$$
\begin{aligned}
y^k &= x^{k-1} + \frac{k-2}{k+1}(x^{k-1} - x^{k-2}) \\
x^k &= \text{prox}_{t_k h}(y^k - t_k \nabla f(y^k))
\end{aligned}
\tag{2}
$$

FISTA的等价形式

- FISTA的一个等价变形：

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

- 当$\gamma_k = \frac{2}{k+1}$时，并且取固定步长时，两个版本是等价的；
- 但是当γ_k采用别的取法时，(3) 将给出另一个版本的加速算法.
- 也就是说，(2)中$\frac{k-2}{k+1}$可以取其他值.

FISTA的收敛条件

下面给出算法(3)以$\mathcal{O}\left(\frac{1}{k^2}\right)$的速度收敛的条件：

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

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

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

- 可以看到当取$t_k = \frac{1}{L}$，$\gamma_k = \frac{2}{k+1}$时，以上条件满足.
- γ_k的选取并不唯一，例如我们可以采取

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

线搜索方法一

- 迭代(2)和(3)都要求步长满足$t_k \le \frac{1}{L}$，此时条件(4)满足．
- 对很多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．
- 方法一在(3)的第2行中加入线搜索，并取$\gamma_k = \frac{2}{k+1}$，以回溯的方式找到满足条件(4)的t_k．该算法的具体过程见:

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

- 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无法终止的情况．
- 容易验证其他两个条件(5), (6)在迭代过程中也得到满足．

线搜索方法二

- 第二种线搜索方法同时改变步长t_k和γ_k，所以y^k也随之改变.
- 该算法的具体过程见:

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

线搜索方法二

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

$$\frac{\sqrt{t_{k-1}}}{\gamma_{k-1}} = \frac{\sqrt{(1-\gamma_k)t_k}}{\gamma_k} \le \frac{\sqrt{t_k}}{\gamma_k} - \frac{\sqrt{t_k}}{2},$$

- 反复利用上式可得$\frac{\sqrt{t_k}}{\gamma_k} \ge \sqrt{t_1} + \frac{1}{2}\sum_{i=2}^{k}\sqrt{t_i}$, 因此

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

- 以上的分析说明条件(5)和(6)在(8)的执行中也得到满足.
- (8)的执行过程同时改变了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

- 总的来说，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(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.

提纲

1. FISTA算法
2. 其他加速算法
3. 应用举例
   - LASSO问题求解
   - 小波模型求解
4. 收敛性分析

第二类Nesterov加速算法

- 对于复合优化问题(1)，我们给出第二类Nesterov加速算法：

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

- 和经典FISTA 算法的一个重要区别在于，第二类Nesterov 加速算法中的三个序列$\{x^k\}$，$\{y^k\}$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．

- 第二类Nesterov 加速算法的一步迭代可参考下图.

$$y^k = \mathrm{prox}_{(t_k/\gamma_k)h}(y^{k-1} - (t_k/\gamma_k)\nabla f(z^k))$$

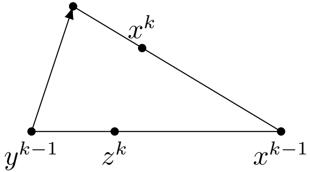

Figure: 第二类Nesterov加速算法的一步迭代

- 针对问题(1)的第三类Nesterov加速算法框架为：

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

- 该算法和第二类Nesterov加速算法(11) 的区别仅仅在于y^k的更新：第三类Nesterov加速算法计算y^k时需要利用全部已有的$\{\nabla f(z^i)\}, i = 1, 2, \cdots, k$.
- 同样地，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.

针对非凸问题的Nesterov加速算法

- 仍然考虑问题(1)的形式，这里并不要求f是凸的，但是要求其是可微的且梯度是利普希茨连续的，h与之前的要求相同．
- 非凸复合优化问题的加速梯度法框架:

$$
\begin{aligned}
z^k &= \gamma_k y^{k-1} + (1-\gamma_k)x^{k-1} \\
y^k &= \operatorname{prox}_{\lambda_k h}\left(y^{k-1} - \lambda_k \nabla f(z^k)\right) \\
x^k &= \operatorname{prox}_{t_k h}\left(z^k - t_k \nabla f(z^k)\right)
\end{aligned}
\tag{13}
$$

针对非凸问题的Nesterov加速算法

- 当λ_k和t_k 取特定值时，它等价于第二类Nesterov加速算法.
- 在非凸函数情形下，一阶算法一般只能保证收敛到一个稳定点，并不能保证收敛到最优解，因此无法用函数值与最优值的差来衡量优化算法解的精度. 对于非凸复合函数(1)，我们利用梯度映射

$$G_t(x) = \frac{1}{t}(x - \mathrm{prox}_{th}(x - t\nabla f(x)))$$

来判断算法是否收敛. 注意到$G_t(x) = 0$是优化问题(1)的一阶必要条件，因此利用$\|G_{t_k}(x^k)\|$来刻画(13)的收敛速度.
- 可以证明，当f为凸函数时，(13)的收敛速度与FISTA算法相同，两者都为$\mathcal{O}\left(\frac{1}{k^2}\right)$；当$f$为非凸函数时，(13)也收敛，且收敛速度为$\mathcal{O}\left(\frac{1}{k}\right)$.

提纲

1. FISTA算法
2. 其他加速算法
3. 应用举例
   - LASSO问题求解
   - 小波模型求解
4. 收敛性分析

- LASSO问题为
$$\min_x \quad \frac{1}{2}\|Ax-b\|_2^2+\mu\|x\|_1 \tag{14}$$
- 求解LASSO问题(14)的FISTA算法可以由下面的迭代格式给出：
$$\begin{aligned} y^k &= x^{k-1}+\frac{k-2}{k+1}(x^{k-1}-x^{k-2}),\\ w^k &= y^k-t_kA^T(Ay^k-b),\\ x^k &= \operatorname{sign}(w^k)\max\{|w^k|-t_k\mu,0\}. \end{aligned}$$
- 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．

- 我们也给出第二类Nesterov加速算法：

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

- 和第三类Nesterov加速算法：

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

LASSO问题求解（续）

- 取$\mu = 10^{-3}$，分别利用连续化近似点梯度法、连续化FISTA加速算法、连续化第二类Nesterov算法来求解问题
- 分别取固定步长$t = \frac{1}{L}$，这里$L = \lambda_{\max}(A^TA)$，和结合线搜索的BB步长.
- 结果如下图：

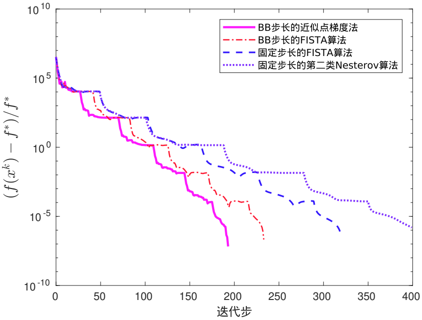

Figure: 使用近似点梯度法以及不同的加速算法求解LASSO 问题

LASSO问题求解（续）

可以看到：

- 就固定步长而言，FISTA算法相较于第二类Nesterov加速算法收敛得略快一些；
- 注意到FISTA算法是非单调算法．
- 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．

小波模型

- 合成小波模型为

$$\min_{\alpha \in \mathbb{R}^m} \quad \|\lambda \odot \alpha\|_1 + \frac{1}{2}\|AW^\top \alpha - b\|_2^2 \tag{15}$$

- 平衡小波模型为

$$\min_{\alpha \in \mathbb{R}^m} \quad \|\lambda \odot \alpha\|_1 + \frac{1}{2}\|AW^\top \alpha - b\|_2^2 + \frac{\kappa}{2}\|(I - WW^\top)\alpha\|_2^2 \tag{16}$$

合成小波模型求解

- 针对合成小波模型求解的FISTA算法为：

$$
\begin{aligned}
y^k &= d^{k-1} + \frac{k-2}{k+1}(d^{k-1} - d^{k-2}), \\
w^k &= y^k - t_k W A^T (A W^T y^k - b), \\
d^k &= \operatorname{sign}(w^k) \max\{|w^k| - t_k\lambda, 0\}.
\end{aligned}
$$

- 针对合成小波模型的第二类Nesterov加速算法为：

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

平衡小波模型求解

- 平衡小波模型求解的FISTA算法可以为：

$$
\begin{aligned}
y^k &= \alpha^{k-1} + \frac{k-2}{k+1}(\alpha^{k-1} - \alpha^{k-2}), \\
w^k &= y^k - t_k(\kappa(I - WW^T)y^k + WA^T(AW^Ty^k - b)), \\
\alpha^k &= \operatorname{sign}(w^k)\max\{|w^k| - t_k\lambda, 0\},
\end{aligned}
$$

- 而相应的第二类Nesterov加速算法的格式为

$$
\begin{aligned}
z^k &= (1-\gamma_k)\alpha^{k-1} + \gamma_k y^{k-1}, \\
w^k &= y^{k-1} - \frac{t_k}{\gamma_k}(\kappa(I - WW^T)z^k + WA^T(AW^Tz^k - b)), \\
y^k &= \operatorname{sign}(w^k)\max\left\{|w^k| - \frac{t_k}{\gamma_k}\lambda, 0\right\}, \\
\alpha^k &= (1-\gamma_k)\alpha^{k-1} + \gamma_k y^k.
\end{aligned}
$$

提纲

1. FISTA算法
2. 其他加速算法
3. 应用举例
   - LASSO问题求解
   - 小波模型求解
4. 收敛性分析

收敛性假设

- f 在其定义域$\textbf{dom}\, f = \mathbb{R}^n$内为凸的，$\nabla f$在常数$L$意义下利普西茨连续，即

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall x, y;$$

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

定理 (固定步长FISTA算法收敛速度)

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

$$\psi(x^k) - \psi(x^*) \leq \frac{2L}{(k+1)^2}\|x^0 - x^*\|^2. \tag{17}$$

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

$$t_k h(x) \geq t_k h(x^k) + \langle -x^k + y^k - t_k \nabla f(y^k), x - x^k \rangle. \tag{18}$$

- 由f 的凸性、梯度利普希茨连续和$t_k = \frac{1}{L}$可以得到

$$f(x^k) \leq f(y^k) + \langle \nabla f(y^k), x^k - y^k \rangle + \frac{1}{2t_k} \|x^k - y^k\|^2. \tag{19}$$

- 结合以上两个不等式，对于任意的x有

$$\begin{aligned}
\psi(x^k) &= f(x^k) + h(x^k) \\
&\leq h(x) + f(y^k) + \langle \nabla f(y^k), x - y^k \rangle + \frac{1}{t_k}\langle x^k - y^k, x - x^k \rangle \\
&\quad + \frac{1}{2t_k}\left\|x^k - y^k\right\|^2 \\
&\leq h(x) + f(x) + \frac{1}{t_k}\langle x^k - y^k, x - x^k \rangle + \frac{1}{2t_k}\left\|x^k - y^k\right\|^2 \\
&= \psi(x) + \frac{1}{t_k}\langle x^k - y^k, x - x^k \rangle + \frac{1}{2t_k}\left\|x^k - y^k\right\|^2.
\end{aligned} \tag{20}$$

- 在(20)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

$$\begin{aligned}
&\psi(x^k) - \psi^* - (1 - \gamma_k)(\psi(x^{k-1}) - \psi^*) \\
\leq &\frac{1}{t_k}\langle x^k - y^k, (1 - \gamma_k)x^{k-1} + \gamma_k x^* - x^k \rangle + \frac{1}{2t_k}\left\|x^k - y^k\right\|^2.
\end{aligned} \tag{21}$$

- 结合迭代式

$$
\begin{aligned}
v^k &= x^{k-1} + \frac{1}{\gamma_k}(x^k - x^{k-1}), \\
y^k &= (1-\gamma_k)x^{k-1} + \gamma_k v^{k-1},
\end{aligned}
$$

不等式(21)可以化为

$$
\begin{aligned}
&\psi(x^k) - \psi^* - (1-\gamma_k)(\psi(x^{k-1}) - \psi^*) \\
\le &\frac{1}{2t_k}(\|y^k - (1-\gamma_k)x^{k-1} - \gamma_k x^*\|^2 - \|x^k - (1-\gamma_k)x^{k-1} - \gamma_k x^*\|^2) \\
= &\frac{\gamma_k^2}{2t_k}(\|v^{k-1} - x^*\|^2 - \|v^k - x^*\|^2).
\end{aligned}
\tag{22}
$$

- t_k, γ_k的取法满足不等式

$$\frac{1-\gamma_k}{\gamma_k^2} t_k \le \frac{1}{\gamma_{k-1}^2} t_{k-1}, \tag{23}$$

可以得到一个有关相邻两步迭代的不等式

$$\frac{t_k}{\gamma_k^2}(\psi(x^k)-\psi^*)+\frac{1}{2}\|v^k-x^*\|^2 \le \frac{t_{k-1}}{\gamma_{k-1}^2}(\psi(x^{k-1})-\psi^*)+\frac{1}{2}\|v^{k-1}-x^*\|^2. \tag{24}$$

- 反复利用(24)式，我们有

$$\frac{t_k}{\gamma_k^2}(\psi(x^k)-\psi^*)+\frac{1}{2}\|v^k-x^*\|^2 \le \frac{t_1}{\gamma_1^2}(\psi(x^1)-\psi^*)+\frac{1}{2}\|v^1-x^*\|^2. \tag{25}$$

- 对$k=1$，注意到$\gamma_1=1, v^0=x^0$，再次利用(22)式可得

$$\begin{aligned}&\frac{t_1}{\gamma_1^2}(\psi(x^1)-\psi^*)+\frac{1}{2}\|v^1-x^*\|^2\\ \le&\frac{(1-\gamma_1)t_1}{\gamma_1^2}(\psi(x^0)-\psi^*)+\frac{1}{2}\|v^0-x^*\|^2=\frac{1}{2}\|x^0-x^*\|^2.\end{aligned} \tag{26}$$

- 结合(25)式和(26)式可以得到(17)式.

证明思路

- 证明中关键的一步在于建立(24)式，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，我们只需要保证条件(4)和条件(5)成立即可.
- 条件(4)主要依赖于$f(x)$的梯度利普希茨连续性；而(5)的成立依赖于γ_k和t_k的选取.
- 条件(6)的成立保证了(3)的收敛速度达到$\mathcal{O}\left(\frac{1}{k^2}\right)$.
- 如果抽取条件(4)-(6) 作为算法收敛的一般条件，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.

一般FISTA算法的收敛速度

推论 (一般FISTA算法的收敛速度)

在收敛性假设条件下，当用(3)求解凸复合优化问题(1)时，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

$$\psi(x^k) - \psi(x^*) \leq \frac{C}{k^2}, \tag{27}$$

其中C 仅与函数f，初始点x^0的选取有关．特别地，采用线搜索版本(7)和(8)的*FISTA* 算法具有$\mathcal{O}\left(\frac{1}{k^2}\right)$的收敛速度．

虽然已经抽象出了t_k, γ_k满足的条件，但我们无法再找到其他的t_k, γ_k来进一步改善FISTA算法的收敛速度，即$\mathcal{O}\left(\frac{1}{k^2}\right)$是FISTA算法所能达到的最高的收敛速度．

第二类Nesterov加速算法收敛速度

定理 (第二类Nesterov加速算法收敛速度)

取$\gamma_k = \frac{2}{k+1}$ 和$t_k = \frac{1}{L}$，利用(11)求解问题(1)有如下收敛性结果：

$$\psi(x^k) - \psi(x^*) \le \frac{2L}{(k+1)^2} \left\| x^0 - x^* \right\|^2. \tag{28}$$

定理2的证明

- 首先根据$y^k = \mathrm{prox}_{(t_k/\gamma_k)h}\left(y^{k-1} - \left(\frac{t_k}{\gamma_k}\right)\nabla f(z^k)\right)$，可知

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

故对于任意的x，有

$$t_k h(x) \geq t_k h(y^k) + \langle \gamma_k(y^{k-1} - y^k) - t_k \nabla f(z^k), x - y^k \rangle. \tag{29}$$

- 再由h的凸性，

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

- 消去$h(y^k)$得到

$$\begin{aligned} h(x^k) \leq\ & (1-\gamma_k)h(x^{k-1}) \\ & + \gamma_k \left[h(x) - \left\langle \frac{\gamma_k}{t_k}(y^{k-1} - y^k) - \nabla f(z^k), x - y^k \right\rangle \right]. \end{aligned} \tag{30}$$

- 利用f 的凸性和梯度利普希茨连续的性质，我们有

$$
\begin{aligned}
f(x^k) &\leq f(z^k) + \langle \nabla f(z^k), x^k - z^k \rangle + \frac{L}{2}\|x^k - z^k\|^2 \\
&= f(z^k) + \langle \nabla f(z^k), x^k - z^k \rangle + \frac{1}{2t_k}\|x^k - z^k\|^2.
\end{aligned} \tag{31}
$$

- 用迭代步3 减去迭代步1 有$x^k - z^k = \gamma_k(y^k - y^{k-1})$，将此等式与

$$
x^k = (1-\gamma_k)x^{k-1} + \gamma_k y^k
$$

代入上式右端得

$$
f(x^k) \leq f(z^k) + \langle \nabla f(z^k), (1-\gamma_k)x^{k-1} + \gamma_k y^k - z^k \rangle + \frac{\gamma_k^2}{2t_k}\|y^k - y^{k-1}\|^2. \tag{32}
$$

- 注意到

$$
\begin{aligned}
& f(z^k) + \langle \nabla f(z^k), (1-\gamma_k)x^{k-1} + \gamma_k y^k - z^k \rangle \\
=& (1-\gamma_k)[f(z^k) + \langle \nabla f(z^k), x^{k-1} - z^k \rangle] + \gamma_k [f(z^k) + \langle \nabla f(z^k), y^k - z^k \rangle] \\
\le& (1-\gamma_k) f(x^{k-1}) + \gamma_k [f(z^k) + \langle \nabla f(z^k), y^k - z^k \rangle],
\end{aligned}
\tag{33}
$$

- 结合不等式(32) (33)得到

$$
f(x^k) \le (1-\gamma_k) f(x^{k-1}) + \gamma_k [f(z^k) + \langle \nabla f(z^k), y^k - z^k \rangle] + \frac{\gamma_k^2}{2t_k} \|y^k - y^{k-1}\|^2. \tag{34}
$$

- 将(30)式与(34)式相加，并结合

$$f(x) \geq f(z^k) + \langle \nabla f(z^k), x - z^k \rangle,$$

再取$x = x^*$，

$$\begin{aligned} &\psi(x^k) - (1-\gamma_k)\psi(x^{k-1}) \\ \leq &\gamma_k \left[h(x^*) + f(x^*) - \frac{\gamma_k}{t_k} \langle y^{k-1} - y^k, x^* - y^k \rangle \right] + \frac{\gamma_k^2}{2t_k} \|y^k - y^{k-1}\|^2 \\ \leq &\gamma_k \psi(x^*) + \frac{\gamma_k^2}{2t_k} (\|y^{k-1} - x^*\|_2^2 - \|y^k - x^*\|^2). \end{aligned} \tag{35}$$

- 这个不等式和(22)式的形式完全相同，因此后续过程可按照定理1进行推导，最终我们可以得到(28)式．

推导的关键步骤仍为条件(4)—(6)．因此对采用线搜索步长的第二类Nesterov加速算法，我们仍然有相同的收敛结果．

推论 (一般第二类Nesterov加速算法的收敛速度)

当用算法11求解凸复合优化问题(1)时，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

$$\psi(x^k) - \psi(x^*) \leq \frac{C}{k^2}, \tag{36}$$

其中C仅和函数f，初始点x^0的选取有关．