优化理论与算法

第八章 投影梯度法
近端梯度法

郭加熠|助理教授

目录

讲　　员

郭加熠，江波，刘慧康

投影梯度法

考虑如下优化问题

$$\min_{x\in C} g(x)$$

其中 $g(x)$ 可微，C 为凸集。

投影梯度下降法的迭代格式为：

$$x^{k+1} = P_C(x^k - t_k \nabla g(x^k)),$$

其中投影算子 $P_C(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ 定义为：对任意 $a \in \mathbb{R}^n$，

$$P_C(x) = \arg\min_{u\in C} \|x-u\|_2 = \arg\min_{u\in C} \frac{1}{2}\|x-u\|_2^2.$$

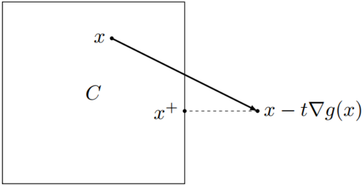
另一个解释：

$$x^{k+1} = \arg\min_{u\in C} g(x^k) + \nabla g(x^k)^T(u-x^k) + \frac{1}{2t_k}\|u-x^k\|^2.$$

示例：带盒子约束的凸二次规划

$$\begin{aligned} \min \quad & g(x) = \frac{1}{2}x^T A x + b^T x \\ \text{s.t.} \quad & 0 \le x \le 1 \end{aligned}$$

$$x^+ = P_C\left(x - t\nabla g(x)\right)$$



仿射集上的投影

- 超平面：$C=\{x|a^Tx=b\}$ 其中（$a\neq 0$）
$$P_C(x)=x+\frac{b-a^Tx}{\|a\|^2}a$$
- 仿射集：$C=\{x|Ax=b\}$ 其中（$A\in\mathbb{R}^{p\times n}$ 且 $\mathrm{rank}(A)=p$）
$$P_C(x)=x+A^T(AA^T)^{-1}(b-Ax)$$
当 $p\ll n$ 或 $AA^T=I$ 时计算高效
- 半空间：$C=\{x|a^Tx\leq b\}$ 其中（$a\neq 0$）
$$P_C(x)=\begin{cases} x+\frac{b-a^Tx}{\|a\|^2}a & \text{if } \ a^Tx>b \\ x & \text{if } \ a^Tx\leq b\end{cases}$$
- 对一般多面体 $C=\{x|Ax\leq b\}$，需通过求解子问题实现投影：
$$\min\|x-u\|_2^2 \quad \text{s.t.} \quad Au\leq b.$$

- 矩形区域：$C = [l, u] = \{x \in \mathbb{R}^n | l \leq x \leq u\}$

$$P_C(x)_i = \begin{cases} l_i & \text{if } \ x_i \leq l_i \\ x_i & \text{if } \ l_i \leq x_i \leq u_i \\ u_i & \text{if } \ x_i \geq u_i \end{cases}$$

特例——l_∞ 范数球：$C = \{x \mid \|x\|_\infty \leq 1\}$（矩形区域）

- 非负象限：$C = \mathbb{R}^n_+$

$$P_C(x) = x_+ = \max\{x, 0\}$$

- 概率单纯形： $C = \{x \in \mathbb{R}^n | e^T x = 1, x \geq 0\}$

$$P_C(x) = (x - \nu e)_+$$

其中 ν 满足方程

$$e^T(x - \nu e)_+ = \sum_{i=1}^{n} \max\{0, x_i - \nu\} = 1$$

- 超平面与矩形交集的投影： $C = \{x \in \mathbb{R}^n | a^T x = b,\ l \leq x \leq u\}$

$$P_C(x) = P_{[l,u]}(x - \nu a)$$

其中，ν 满足

$$a^T P_{[l,u]}(x - \nu a) = b$$

范数球上的投影

- 欧式球：$C = \{x \mid \|x\|_2 \leq 1\}$

$$P_C(x) = x/\|x\|_2\,, \text{if} \quad \|x\|_2 \geq 1, \quad P_C(x) = x; \quad \text{if } \|x\|_2 \leq 1$$

- 1-范数球：$C = \{x \mid \|x\|_1 \leq 1\}$

$$P_C(x)_i = \begin{cases} x_i - \lambda & \text{当} x_i \geq \lambda \\ 0 & \text{当} -\lambda \leq x_i \leq \lambda \\ x_i + \lambda & \text{当} x_i \leq -\lambda \end{cases}$$

其中当 $\|x\|_1 \leq 1$ 有 $\lambda = 0$；否则 λ 满足

$$\sum_{i=1}^{n} \max\{|x_i| - \lambda, 0\} = 1$$

凸锥上的投影

- 二阶锥：$C = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\|_2 \le t\}$

$$P_C(v,s) = \begin{cases} 0 & \|v\| \le -s \\ (v,s) & \|v\| \le s \\ \frac{1}{2}(1+\frac{s}{\|v\|})(v, \|v\|_2) & \|v\| \ge |s| \end{cases}$$

- 半正定锥：$C = \mathbb{S}^n_+$

$$P_C(X) = \sum_{i=1}^n \max\{0, \lambda_i\} q_i q_i^T$$

其中 $X = \sum_{i=1}^n \lambda_i q_i q_i^T$ 为 X 的特征值分解

示性函数

定义集合 C 的示性函数：

$$I_C(x) = \begin{cases} 0 & \text{if } \; x \in C \\ \infty, & \text{Otherwise} \end{cases}$$

当 C 为凸集时，$I_C(x)$ 为凸函数。此时优化问题

$$\min_{x \in C} g(x) \quad \Longleftrightarrow \quad \min g(x) + I_C(x)$$

而投影梯度法的迭代步骤

$$x^{k+1} = \arg\min_{u \in C} g(x^k) + \nabla g(x^k)^T (u - x^k) + \frac{1}{2t_k} \|u - x^k\|^2.$$

可重新表述为

$$x^{k+1} = \arg\min_{u} g(x^k) + \nabla g(x^k)^T (u - x^k) + \frac{1}{2t_k} \|u - x^k\|^2 + I_C(u).$$

目录

讲　　　员

郭加熠，江波，刘慧康

复合函数

考虑如下无约束优化问题：

$$\min f(x) = g(x) + h(x)$$

- g 可微，$\operatorname{dom} g = \mathbb{R}^n$
- h 凸函数但不可微

近端梯度法（PGM）的迭代格式为：

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_u g(x^k) + \nabla g(x^k)^T(u - x^k) + \frac{1}{2t_k}\left\|u - x^k\right\|^2 + h(u) \\ &= \operatorname*{argmin}_u \frac{1}{2t_k}\left\|u - (x^k - t_k\nabla g(x^k))\right\|^2 + h(u) \end{aligned}$$

- 当 $h = 0$ 时，退化为梯度下降法
- 当 $h = I_C$ 时，退化为投影梯度法

近端映射

对于凸函数 $h(\cdot)$，定义近端映射 $\mathrm{prox}_h(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$，即对任意 $a \in \mathbb{R}^n$，

$$\mathrm{prox}_h(x) = \arg\min_u \frac{1}{2}\|x - u\|_2^2 + h(u).$$

- 存在性：$h(u) + \frac{1}{2}\|u - x\|^2$ 是闭函数且具有有界下水平集
- 唯一性：$h(u) + \frac{1}{2}\|u - x\|^2$ 是强凸函数
- 由定义中最优化问题的最优性条件可得：

$$u = \mathrm{prox}_h(x) \Leftrightarrow x - u \in \partial h(u)$$

- 当 $h(\cdot) = I_C(\cdot)$ 时，等同于投影算子

算法解释

使用近端映射，近端梯度法可表示为：

$$
\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_u \frac{1}{2t_k}\left\|u-(x^k-t_k\nabla g(x^k))\right\|^2+h(u) \\
&= \operatorname*{argmin}_u \frac{1}{2}\left\|u-(x^k-t_k\nabla g(x^k))\right\|^2+t_kh(u) \\
&= \mathrm{prox}_{t_kh}\left(x^k-t_k\nabla g(x^k)\right)
\end{aligned}
$$

根据最优性条件，等价于：

$$x^{k+1}\in x^k-t_k\nabla g(x^k)-t_k\partial h(x^{k+1})$$

与次梯度法的区别在于后者形式为：

$$x^{k+1}\in x^k-t_k\nabla g(x^k)-t_k\partial h(x^k)$$

近端映射易解示例

- 二次函数（$A \succeq 0$）

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c,\ \mathrm{prox}_{tf}(x) = (I + tA)^{-1}(x - tb)$$

- 欧几里得范数：$f(x) = \|x\|_2$

$$\mathrm{prox}_{tf}(x) = \begin{cases} (1 - t/\|x\|_2)x & \text{若 } \|x\|_2 \geq t \\ 0 & \text{否则} \end{cases}$$

- 对数障碍函数

$$f(x) = -\sum_{i=1}^{n} \log x_i,\ \mathrm{prox}_{tf}(x)_i = \frac{x_i + \sqrt{x_i^2 + 4t}}{2}, i = 1, \ldots, n$$

- $h(x) = \|x\|_1$： prox_{th} 为" 软阈值"（收缩）算子

$$\mathrm{prox}_{th}(x)_i = \begin{cases} x_i - t & 若 x_i \geq t \\ 0, & 若 -t \leq x_i \leq t \\ x_i + t & 若 x_i \leq -t \end{cases}$$

证明：

$$\mathrm{prox}_{th}(x) = \underset{u}{\mathrm{argmin}}\, t\,\|u\|_1 + \frac{1}{2}\,\|u - x\|^2$$

问题具有可分离性：

$$\mathrm{prox}_{th}(x)_i = \underset{u_i}{\mathrm{argmin}}\, t|u_i| + \frac{1}{2}(u_i - x_i)^2$$

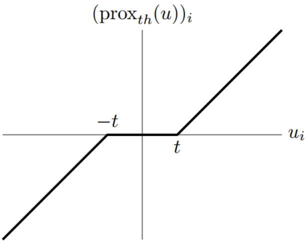
LASSO: 1-范数正则化最小二乘

考虑 LASSO 问题:

$$\min \quad \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1$$

where

$$(\mathrm{prox}_{th}(u))_i = \begin{cases} u_i - t & u_i \ge t \\ 0 & -t \le u_i \le t \\ u_i + t & u_i \le -t \end{cases}$$



近端梯度法迭代格式为:

$$x^{k+1} = \mathrm{prox}_{t\hat{h}}(x^k - tA^T(Ax^k - b)).$$

非扩张性

若 $u = \mathrm{prox}_h(x), v = \mathrm{prox}_h(y)$，则：

$$(u-v)^T(x-y) \geq \|u-v\|^2$$

近端映射 prox_h 是严格非扩张的：

- 证明：由 ∂h 的单调性可得：

$$x-u \in \partial h(u),\ y-v \in \partial h(v) \Rightarrow (x-u-y+v)^T(u-v) \geq 0$$

- 从而推出非扩张性（由 Cauchy-Schwarz 不等式）：

$$\|\mathrm{prox}_h(x) - \mathrm{prox}_h(y)\|_2 \leq \|x-y\|_2$$

近端映射 prox_h 是非扩张的（Lipschitz 常数为 1）

近端梯度法收敛性

目标函数分解为两部分的无约束优化问题:

$$f(x) = g(x) + h(x)$$

近端梯度法步骤:

- 选择初始点 x^0 并重复迭代:

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_{t_k h}(x^k - t_k \nabla g(x^k)), \\ &= \operatorname*{argmin}_x \frac{1}{2t_k} \left\| x - \left(x^k - t_k \nabla g(x^k)\right) \right\|^2 + h(x) \end{aligned}$$

假设条件

- g 为凸函数且 $\operatorname{dom} g = \mathbb{R}^n$
- ∇g 具有 Lipschitz 常数 L：

$$\|\nabla g(x) - \nabla g(y)\|_2 \le L \|x - y\|_2, \forall x, y$$

- $g(x)$ 具有 m 强凸性（$m \ge 0$），即：

$$g(x) \ge g(y) + \nabla g(y)^T (y - x) + \frac{m}{2} \|x - y\|^2$$

  当 $m = 0$ 时退化为凸情形
- h 是闭凸函数（保证 prox_{th} 良定义）
- 最优值 f^* 有限且在 x^* 处可达（不要求唯一）

梯度映射

将近端梯度法改写为:

$$x^+ = \mathrm{prox}_{th}(x - t\nabla g(x)) = x - tG_t(x)$$

即定义梯度映射:

$$G_t(x) = \frac{1}{t}\left(x - \mathrm{prox}_{th}(x - t\nabla g(x))\right).$$

- $G_t(x)$ 不是 $f = g + h$ 的梯度或次梯度
- 根据近端算子的次梯度定义可得:

$$G_t(x) \in \nabla g(x) + \partial h(x - tG_t(x))$$

- $G_t(x) = 0$ 当且仅当 x 是 $f(x) = g(x) + h(x)$ 的极小点

Lipschitz 假设的推论

回忆具有 Lipschitz 连续梯度凸函数 g 的上界估计：

$$g(y) \leq g(x) + \nabla g(x)^T (y - x) + \frac{L}{2} \|y - x\|^2, \forall x, y$$

令 $y = x - tG_t(x)$ 代入可得：

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t^2 L}{2} \|G_t(x)\|^2$$

当 $0 \leq t \leq 1/L$ 时：

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2} \|G_t(x)\|^2 \tag{1}$$

全局不等式

若不等式(1)成立，则对所有 z 有：

$$f(x - tG_t(x)) \leq f(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|^2 - \frac{m}{2}\|x - z\|^2 \quad (2)$$

证明：定义 $v = G_t(x) - \nabla g(x)$，

$$\begin{aligned} f(x - tG_t(x)) &\leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|^2 + h(x - tG_t(x)) \\ &\leq g(z) + \nabla g(x)^T(x - z) - \frac{m}{2}\|x - z\|^2 - t\nabla g(x)^T G_t(x) \\ &\quad + \frac{t}{2}\|G_t(x)\|^2 + h(z) + v^T(x - z - tG_t(x)) \\ &= g(z) + h(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|^2 - \frac{m}{2}\|x - z\|^2 \end{aligned}$$

第一个不等式使用 (1)，第二个不等式利用了 g 的 m-强凸性和 h 的凸性，以及 $v \in \partial h(x - tG_t(x))$，第三行等式使用 $G_t(x) = \nabla g(x) + v$。

单步迭代进展

令 $x^+ = x - tG_t(x)$，则：

- 取 $z = x$ 时不等式 (2) 表明算法是下降方法：

$$f(x^+) \le f(x) - \frac{t}{2}\|G_t(x)\|^2$$

- 取 $z = x^*$ 时不等式 (2) 可得：

$$\begin{aligned}
f(x^+) - f^* &\le G_t(x)^T(x - x^*) - \frac{t}{2}\|G_t(x)\|^2 - \frac{m}{2}\|x - x^*\|^2 \\
&= \frac{1}{2t}\left(\|x - x^*\|^2 - \|x - x^* - tG_t(x)\|^2\right) - \frac{m}{2}\|x - x^*\|^2 \\
&= \frac{1}{2t}\left((1 - mt)\|x - x^*\|^2 - \left\|x^+ - x^*\right\|^2\right) \qquad (3) \\
&\le \frac{1}{2t}\left(\|x - x^*\|^2 - \left\|x^+ - x^*\right\|^2\right) \qquad (4)
\end{aligned}$$

（因此 $\|x^+ - x^*\|_2 \le \|x - x^*\|_2$，即到最优解集的距离单调递减）

固定步长分析 (凸函数)

对 $x = x_{i-1}$, $x^+ = x_i$, $t = t_{i-1} \in (0, 1/L)$ 累加不等式 (4):

$$
\begin{aligned}
\sum_{i=1}^{k}(f(x_i) - f^*) &\leq \frac{1}{2t}\sum_{i=1}^{k}(\|x_{i-1} - x^*\|^2 - \|x_i - x^*\|^2) \\
&= \frac{1}{2t}(\|x_0 - x^*\|^2 - \|x_k - x^*\|^2) \\
&\leq \frac{1}{2t}\|x_0 - x^*\|^2
\end{aligned}
$$

由于 $f(x_i)$ 单调不增,

$$
f(x_k) - f^* \leq \frac{1}{k}\sum_{i=1}^{k} f(x_i) - f^* \leq \frac{1}{2kt}\|x_0 - x^*\|^2
$$

结论: 经过 $O(1/\epsilon)$ 次迭代可达 $f(x_k) - f^* \leq \epsilon$

到最优解集的距离 (强凸函数)

- 由 (3) 式和 $f(x^+) \geq f^*$ 可知到最优解集的距离不增加：

$$\|x^+ - x^*\|^2 \leq (1 - mt)\|x - x^*\|^2$$

- 对固定步长 $t_k = 1/L$ 有：

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{m}{L}\right)^k \|x_0 - x^*\|^2,$$

即当 g 强凸时 $(m > 0)$ 具有线性收敛速率

线搜索

- 固定步长分析基于初始不等式:

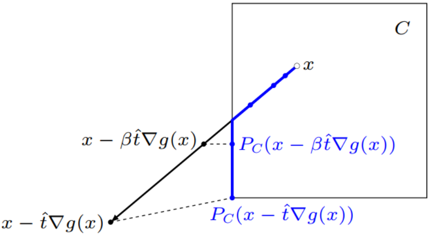
$$g(x - tG_t(x)) \le g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\left\|G_t(x)\right\|^2 \quad (*)$$

- 当 L 未知时，可通过回溯线搜索满足 $(*)$ 式：从初始 $t := \hat{t} > 0$ 开始回溯（$t := \beta t$）直至满足 $(*)$
- 线搜索选取的步长满足 $t \ge t_{\min} = \min\{\hat{t}, \beta/L\}$
- 每次线搜索需计算一次 g 和 prox_{th}
- 收敛速率与固定步长情形类似

示例

投影梯度法（又称梯度投影法）的线搜索过程

$$x^+ = P_C\left(x - t\nabla g(x)\right) = x - tG_t(x)$$



回溯直至投影点 $P_C(x - t\nabla g(x))$ 满足充分下降不等式 $(*)$

近端梯度法总结

1. 适用于可微与不可微凸函数之和的优化问题：

$$f(x) = g(x) + h(x)$$

2. 要求不可微项 h 具有高效近端算子
3. 收敛性质与标准梯度法（$h(x) = 0$ 时）相似
4. 虽不如次梯度法通用，但收敛速度更快

目录

讲　　员

郭加熠，江波，刘慧康

FISTA (基本版本)

目标函数分解为两个部分的无约束优化问题:

$$\min f(x) = g(x) + h(x)$$

- g 为凸可微函数，定义域 $\mathrm{dom} g = \mathbb{R}^n$
- h 为凸但不可微函数，具有高效近端算子 prox_{th}

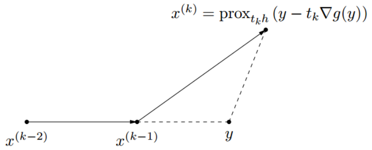
算法步骤: 初始化 $x^0 = x^{-1}$; 对于 $k \geq 1$，重复执行

$$\begin{aligned} y &= x^{k-1} + \frac{k-2}{k+1}\left(x^{k-1} - x^{k-2}\right) \\ x^k &= \mathrm{prox}_{t_k h}(y - t_k \nabla g(y)) \end{aligned}$$

- 项 $x^{k-1} - x^{k-2}$ 称为动量项
- 缩写全称为" 快速迭代收缩阈值算法 (Fast Iterative Shrinkage-Thresholding Algorithm)"

算法解释

- 第一步迭代 ($k=1$) 是在 $y=x^0$ 处执行近端梯度步骤
- 后续迭代在预测点 y 处执行近端梯度步骤



注意 x^k 是可行点 (属于 $\mathrm{dom}\, h$)，但 y 可能在 $\mathrm{dom}\, h$ 之外

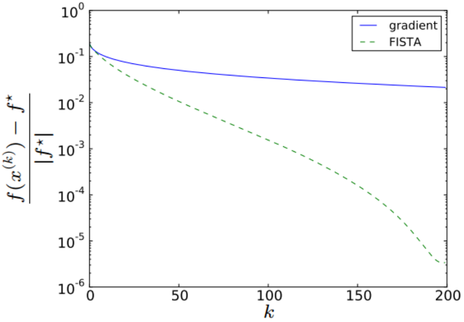
- 当 $h(\cdot)=0$ 时退化为加速梯度法

$$
\begin{aligned}
y &= x^{k-1} + \frac{k-2}{k+1}\left(x^{k-1} - x^{k-2}\right) \\
x^k &= y - t_k \nabla g(y)
\end{aligned}
$$

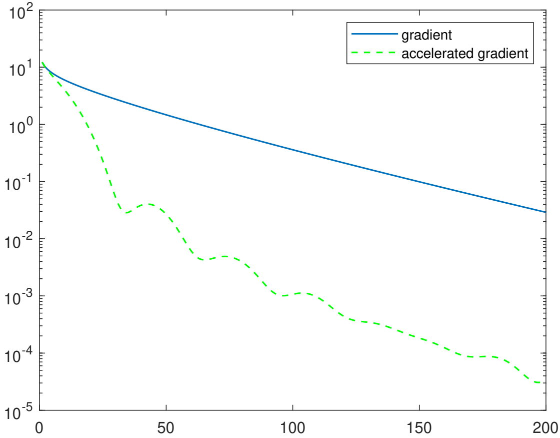
数值算例

$$\min \log \sum_{i=1}^{m} \exp(a_i^T x + b_i)$$

随机生成数据 $(m = 2000, n = 1000)$，使用固定步长



另一个示例 (Nesterov 震荡现象)



FISTA 不是单调下降算法

FISTA 的收敛性

假设条件

- g 为凸可微函数，定义域 $\mathrm{dom} g = \mathbb{R}^n$；梯度 ∇g 满足 Lipschitz 连续性，常数为 L:

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L \|x - y\|_2$$

- h 为闭凸函数 (保证 prox_{th} 算子良定义)
- 最优值 f^* 有限且在 x^* 处可达 (不要求唯一性)

收敛结果: $f(x^k) - f^*$ 的衰减速率不低于 $O(1/k^2)$

- 取步长 $t_k = 1/L$
- 配合适当线搜索策略

FISTA 的等价重构

定义参数 $\theta_k = 2/(k+1)$ 并引入中间变量 v^k

算法步骤: 初始化 $x_0 = v_0$; 对于 $k \geq 0$，重复执行

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

将 v^k 的表达式代入 y 的公式即可还原标准 FISTA 形式

关键不等式

参数选择: 序列 $\theta_k = \frac{2}{k+1}$ 满足 $\theta_1 = 1$ 且

$$\frac{1-\theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2}, \quad k \geq 2$$

Lipschitz 性质导出的 g 上界

$$g(u) \leq g(z) + \nabla g(z)^T (u-z) + \frac{L}{2} \|u-z\|_2^2 \;\; \forall u, z$$

近端算子定义导出的 h 上界

$$h(u) \leq h(z) + \frac{1}{t}(w-u)^T (u-z), \;\; \forall w, u = \mathrm{prox}_{th}(w), z$$

单步迭代进展

记 $x = x_{i-1}, x^+ = x_i, v = v_{i-1}, v^+ = v_i, t = t_i, \theta = \theta_i$

- Lipschitz 性质导出上界 (令 $u = x^+, z = y$): 当 $0 < t \le 1/L$ 时,

$$g(x^+) \le g(y) + \nabla g(y)^T(x^+ - y) + \frac{1}{2t}\left\|x^+ - y\right\|_2^2 \quad (1)$$

- 近端算子导出的上界 (令 $u = x^+$, $w = y - t\nabla g(y)$):

$$h(x^+) \le h(z) + \nabla g(y)^T(z - x^+) + \frac{1}{t}(x^+ - y)^T(z - x^+) \ \forall z$$

- 加和上界并利用 g 的凸性 $g(z) \ge g(y) + \nabla g(y)^T(z - y)$ 得出

$$f(x^+) \le f(z) + \frac{1}{t}(x^+ - y)^T(z - x^+) + \frac{1}{2t}\left\|x^+ - y\right\|_2^2 \ \forall z \quad (2)$$

- 考虑下式

$$
\begin{aligned}
& f(x^+) - f^* - (1-\theta)(f(x) - f^*) \\
= \quad & f(x^+) - \theta f^* - (1-\theta) f(x) \le f(x^+) - f(\theta x^* + (1-\theta)x) \\
\overset{(2)}{\le} \quad & \frac{1}{t}(x^+ - y)^T(\theta x^* + (1-\theta)x - x^+) + \frac{1}{2t}\left\|x^+ - y\right\|_2^2 \\
= \quad & \frac{1}{2t}(\|y - (1-\theta)x - \theta x^*\|_2^2 - \left\|x^+ - (1-\theta)x - \theta x^*\right\|_2^2) \\
= \quad & \frac{\theta^2}{2t}(\|v - x^*\|_2^2 - \left\|v^+ - x^*\right\|_2^2)
\end{aligned}
$$

结论: 若第 i 次迭代满足不等式 (1)，则上述式子等价于下式成立

$$
\begin{aligned}
& \frac{t_i}{\theta_i^2}(f(x_i) - f^*) + \frac{1}{2}\|v_i - x^*\|_2^2 \\
\le \quad & \frac{(1-\theta_i)t_i}{\theta_i^2}(f(x_{i-1}) - f^*) + \frac{1}{2}\|v_{i-1} - x^*\|_2^2 \quad (3)
\end{aligned}
$$

固定步长收敛性分析

对于 $i \geq 1$, 若选固定补充 $t_i = t \in (0, 1/L]$，那么重复使用公式 (3) 时，采用 $(1-\theta_i)/\theta_i^2 \leq 1/\theta_{i-1}^2$ 对于任意 $i \geq 2$ 成立这个关系，有下式成立:

$$
\begin{aligned}
& \frac{t}{\theta_k^2}(f(x^k) - f^*) + \frac{1}{2}\left\|v^k - x^*\right\|_2^2 \\
\leq \quad & \frac{(1-\theta_1)t}{\theta_1^2}(f(x_0) - f^*) + \frac{1}{2}\left\|v_0 - x^*\right\|_2^2 \\
\overset{\theta_1=1}{=} \quad & \frac{1}{2}\left\|x_0 - x^*\right\|_2^2
\end{aligned}
$$

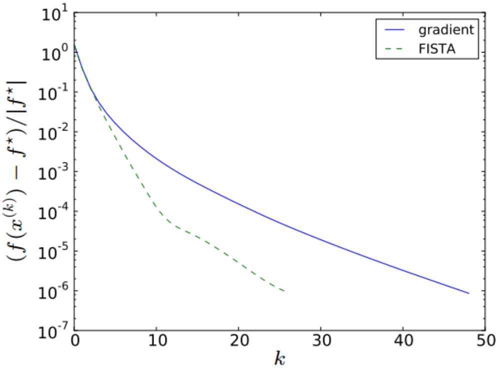
因此

$$
f(x^k) - f^* \leq \frac{\theta_k^2}{2t}\left\|x_0 - x^*\right\|_2^2 = \frac{2}{t(k+1)^2}\left\|x_0 - x^*\right\|_2^2
$$

结论：达到 $f(x^k) - f^* \leq \epsilon$ 需要 $O(1/\sqrt{\epsilon})$ 次迭代

算例：带盒约束的二次规划

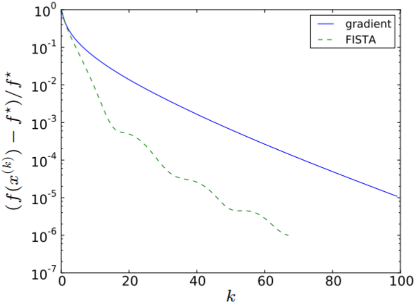
$$\begin{array}{ll} \min & \frac{1}{2}x^T A x + b^T x \\ \text{s.t.} & 0 \le x \le 1 \end{array}$$



$n = 3000$; fixed step size $t = 1/\lambda_{\max}(A)$

采用 1 范数正则化的最小二乘

$$\min \frac{1}{2}\|Ax - b\|_2^2 + \|x\|_1$$



randomly generated $A \in \mathbf{R}^{2000\times 1000}$; step $t_k = 1/L$ with $L = \lambda_{\max}(A^T A)$

FISTA 分析的关键步骤

- 起始点是不等式

$$g(x^+) \le g(y) + \nabla g(y)^T (x^+ - y) + \frac{1}{2t} \left\| x^+ - y \right\|_2^2 \quad (1)$$

该不等式在 $0 \le t \le 1/L$ 时成立

- 若 (1) 成立，则第 i 次迭代的进展由下式界定

$$\begin{aligned} & \frac{t_i}{\theta_i^2}(f(x_i) - f^*) + \frac{1}{2} \left\| v_i - x^* \right\|_2^2 \\ \le \quad & \frac{(1-\theta_i)t_i}{\theta_i^2}(f(x_{i-1}) - f^*) + \frac{1}{2} \left\| v_{i-1} - x^* \right\|_2 \quad (2) \end{aligned}$$

- 为递归组合这些不等式，需要满足

$$\frac{(1-\theta_i)t_i}{\theta_i^2} \le \frac{t_{i-1}}{\theta_{i-1}^2} \quad (i \ge 2) \quad (3)$$

- 若 $\theta_1 = 1$，将 (2) 式从 $i = 1$ 到 k 累加可得上界

$$f(x_k) - f^* \le \frac{\theta_k^2}{2t_k} \|x_0 - x^*\|_2^2$$

结论: 当 (1) 和 (3) 满足 $\frac{\theta_k^2}{t_k} = O(\frac{1}{k^2})$ 时，保证 $O(1/k^2)$ 收敛速率

固定步长 FISTA

$$t_k = 1/L,\ \theta_k = \frac{2}{k+1}$$

这些参数满足 (1) 和 (3) 且

$$\frac{\theta_k^2}{t_k} = \frac{4L}{(k+1)^2}$$

带线搜索的 FISTA

将第 k 次迭代中 x 的更新替换为:

- $t := t_{k-1}$ (定义 $t_0 = \hat{t} > 0$)
- $x := \mathrm{prox}_{th}(y - t\nabla g(y))$
- while $g(x) > g(y) + \nabla g(y)^T(x - y) + \frac{1}{2t}\|x - y\|_2^2$ 时

  1. $t := \beta t$
  2. $x := \mathrm{prox}_{th}(y - t\nabla g(y))$

  end

带线搜索的 FISTA

- 回溯退出条件自然满足不等式 (1)
- 因 (1) 成立，故不等式 (2) 成立
- 当 $\theta_k = 2/(k+1)$ 且 $t_k \le t_{k-1}$ 时不等式 (3) 成立（可以让 $\hat{t} = t_{k-1}$）。
- ∇g 的 Lipschitz 连续性保证 $t_k \ge t_{\min} = \min\{\hat{t}, \beta/L\}$
- 因 $\theta_k^2/t_k = O(1/k^2)$ 保持 $1/k^2$ 收敛速率

FISTA 的下降版本

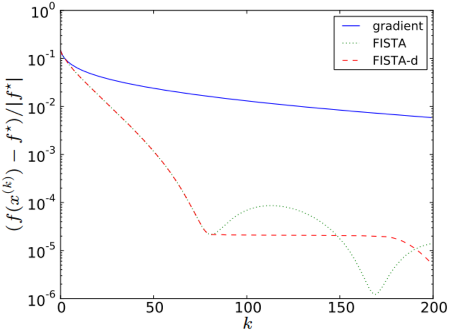
初始化 $x_0 = v_0$; 对 $k \geq 1$ 重复以下步骤:

- $y = (1-\theta_k)x^{k-1} + \theta_k v^{k-1}$
- $u := \mathrm{prox}_{th}(y - t\nabla g(y))$
- $x^k = \begin{cases} u & \text{若} f(u) \leq f(x^{k-1}) \\ x^{k-1} & \text{否则} \end{cases}$
- $v^k = x^{k-1} + \frac{1}{\theta_k}(u - x^{k-1})$

结果:

- 步骤 3 保证 $f(x^k) \leq f(x^{k-1})$
- 使用 $\theta_k = 2/(k+1)$ 和 $t_k = 1/L$，或任一线搜索方法
- 与原始 FISTA 具有相同迭代复杂度

算例



gradient
FISTA
FISTA-d
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0
50
100
150
200
k
$(f(x^{(k)}) - f^\star)/|f^\star|$

重启策略

实践中，当 $\theta_k/\theta_{k-1} \to 1$ 时 FISTA 会退化

此时，$\{x^k\}_{k=1}^{\infty}$ 和 $\{y_k\}_{k=1}^{\infty}$ 会在最优解附近振荡

为避免振荡并加速收敛，可采用重启策略：将当前迭代点作为新初始点

何时重启 FISTA?

- 简单方法：每 T 次迭代重启 (如 $T = 50$)
- O'donoghue 和 Candes (2015) 提出以下重启准则:

$$f(x^k) > f(x^{k-1})$$

或

$$G(y_{k-1})^T(x^k - x^{k-1}) > 0 \quad \Leftrightarrow \quad (y_{k-1} - x_k)^T(x^k - x^{k-1}) > 0$$

比较

$$\min_{x} x^T A x + 2b^T x \text{ 其中 } A \succ 0$$