

第七章 复合优化算法

修贤超

<https://xianchaoxiu.github.io>

- 7.1 近似点梯度法
- 7.2 Nesterov 加速算法
- 7.3 近似点算法
- 7.4 分块坐标下降法
- 7.5 对偶算法
- 7.6 交替方向乘子法
- 7.7 随机优化算法

■ 考虑如下复合优化问题

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

□ $f(x)$ 为可微函数 (可能非凸)

□ $h(x)$ 可能为不可微函数

■ 定义 7.1 对于一个凸函数 h , 定义邻近算子为

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

■ 定理 7.1 如果 h 为闭凸函数, 则对任意 x 有 $\text{prox}_h(x)$ 存在且唯一

■ **定理 7.2** 若 h 是适当的闭凸函数, 则

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

证明 若 $u = \text{prox}_h(x)$, 则由最优性条件得 $0 \in \partial h(u) + (u - x)$, 因此 $x - u \in \partial h(u)$. 反之, 若 $x - u \in \partial h(u)$ 则由次梯度的定义可得到

$$h(v) \geq h(u) + (x - u)^\top (v - u), \quad \forall v \in \text{dom } h$$

两边同时加 $\frac{1}{2}\|v - x\|^2$, 即有

$$\begin{aligned} h(v) + \frac{1}{2}\|v - x\|^2 &\geq h(u) + (x - u)^\top (v - u) + \frac{1}{2}\|(v - u) - (x - u)\|^2 \\ &\geq h(u) + \frac{1}{2}\|u - x\|^2, \quad \forall v \in \text{dom } h \end{aligned}$$

根据定义可得 $u = \text{prox}_h(x)$

例 7.1

■ 给定 ℓ_1 范数 $h(x) = \|x\|_1$, 则 $\text{prox}_{th}(x) = \text{sign}(x) \max\{|x| - t, 0\}$

证明 邻近算子 $u = \text{prox}_{th}(x)$ 的最优性条件为

$$x - u \in t\partial\|u\|_1 = \begin{cases} \{t\}, & u > 0 \\ [-t, t], & u = 0 \\ \{-t\}, & u < 0 \end{cases}$$

\Downarrow

$$u = \begin{cases} x - t, & x > t \\ x + t, & x < -t \\ 0, & x \in [-t, t] \end{cases}$$

例 7.1

- 给定 ℓ_2 范数 $h(x) = \|x\|_2$, 则 $\text{prox}_{th}(x) = \begin{cases} (1 - \frac{t}{\|x\|_2})x, & \|x\|_2 \geq t \\ 0, & \text{其他} \end{cases}$

证明 邻近算子 $u = \text{prox}_{th}(x)$ 的最优性条件为

$$x - u \in t\partial\|u\|_2 = \begin{cases} \{\frac{tu}{\|u\|_2}\}, & u \neq 0 \\ \{w : \|w\|_2 \leq t\}, & u = 0 \end{cases}$$

\Downarrow

$$u = \begin{cases} x - \frac{tx}{\|x\|_2}, & \|x\|_2 > t \\ 0, & \|x\|_2 \leq t \end{cases}$$

例 7.2

■ 邻近算子的计算规则

- ▣ 变量的常数倍放缩以及平移 ($\lambda \neq 0$)

$$h(x) = g(\lambda x + a), \quad \text{prox}_h(x) = \frac{1}{\lambda} \left(\text{prox}_{\lambda^2 g}(\lambda x + a) - a \right)$$

- ▣ 函数（及变量）的常数倍放缩 ($\lambda > 0$)

$$h(x) = \lambda g\left(\frac{x}{\lambda}\right), \quad \text{prox}_h(x) = \lambda \text{prox}_{\lambda^{-1}g}\left(\frac{x}{\lambda}\right)$$

- ▣ 加上线性函数

$$h(x) = g(x) + a^\top x, \quad \text{prox}_h(x) = \text{prox}_g(x - a)$$

例 7.2

- 加上二次项 ($u > 0$)

$$h(x) = g(x) + \frac{u}{2}\|x - a\|_2^2, \quad \text{prox}_h(x) = \text{prox}_{\theta g}(\theta x + (1 - \theta)a)$$

其中 $\theta = \frac{1}{1+u}$

- 向量函数

$$h\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \varphi_1(x) + \varphi_2(y), \quad \text{prox}_h\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \text{prox}_{\varphi_1}(x) \\ \text{prox}_{\varphi_2}(y) \end{bmatrix}$$

例 7.3

- 设 C 为闭凸集, 则示性函数 I_C 的邻近算子为点 x 到 C 的投影 $\mathcal{P}_C(x)$

$$\begin{aligned}\operatorname{prox}_{I_C}(x) &= \arg \min_u \left\{ I_C(u) + \frac{1}{2} \|u - x\|^2 \right\} \\ &= \arg \min_{u \in C} \|u - x\|^2 \\ &= \mathcal{P}_C(x)\end{aligned}$$

- 几何意义

$$u = \mathcal{P}_C(x) \quad \Leftrightarrow \quad (x - u)^\top (z - u) \leq 0, \quad \forall z \in C$$

近似点梯度法

■ 考虑复合优化问题

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

■ 对于光滑部分 f 做梯度下降, 对于非光滑部分 h 使用邻近算子

=====

算法 7.1 近似点梯度法

- 1 给定函数 $f(x), h(x)$, 初始点 x^0
- 2 **while** 未达到收敛准则 **do**
- 3 $x^{k+1} = \text{prox}_{t_k h}(x^k - t_k \nabla f(x^k))$
- 4 **end while**

对近似点梯度法的理解

■ 把迭代公式展开

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

\Downarrow

$$\begin{aligned} x^{k+1} &= \arg \min_u \left\{ h(u) + \frac{1}{2t_k} \|u - x^k + t_k \nabla f(x^k)\|^2 \right\} \\ &= \arg \min_u \left\{ h(u) + f(x^k) + \nabla f(x^k)^\top (u - x^k) + \frac{1}{2t_k} \|u - x^k\|^2 \right\} \end{aligned}$$

■ 根据邻近算子与次梯度的关系, 可改写为

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

■ 对光滑部分做显式的梯度下降, 对非光滑部分做隐式的梯度下降

步长选取

- 当 f 为梯度 L -利普希茨连续函数时, 可取固定步长 $t_k = t \leq \frac{1}{L}$
- 当 L 未知时可使用线搜索准则

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{1}{2t_k} \|x^{k+1} - x^k\|^2$$

- BB 步长
- 可构造如下适用于近似点梯度法的非单调线搜索准则

$$\psi(x^{k+1}) \leq C^k - \frac{c_1}{2t_k} \|x^{k+1} - x^k\|^2$$

应用举例: LASSO 问题

- 考虑用近似点梯度法求解 LASSO 问题

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

- 令 $f(x) = \frac{1}{2} \|Ax - b\|^2$, $h(x) = \mu \|x\|_1$, 则

$$\begin{aligned}\nabla f(x) &= A^\top (Ax - b) \\ \text{prox}_{t_k h}(x) &= \text{sign}(x) \max\{|x| - t_k \mu, 0\}\end{aligned}$$

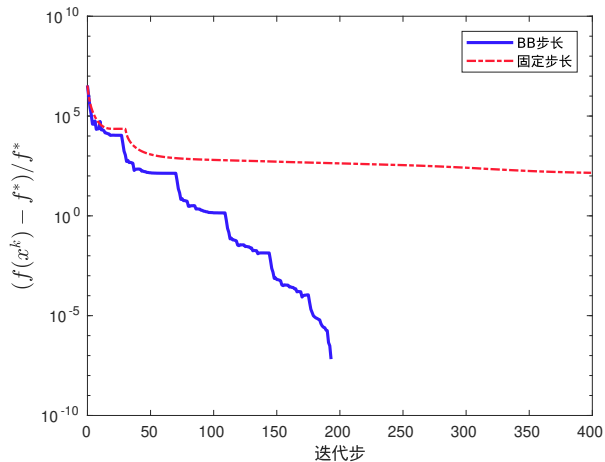
- 相应的迭代格式为

$$\begin{aligned}y^k &= x^k - t_k A^\top (Ax^k - b) \\ x^{k+1} &= \text{sign}(y^k) \max\{|y^k| - t_k \mu, 0\}\end{aligned}$$

即第一步做梯度下降, 第二步做收缩

应用举例: LASSO 问题

■ 使用 BB 步长加速收敛



应用举例：低秩矩阵恢复

■ 考虑低秩矩阵恢复模型

$$\min_{X \in \mathbb{R}^{m \times n}} \mu \|X\|_* + \frac{1}{2} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2$$

■ 令

$$f(X) = \frac{1}{2} \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2, \quad h(X) = \mu \|X\|_*$$

■ 定义矩阵

$$P_{ij} = \begin{cases} 1, & (i,j) \in \Omega \\ 0, & \text{其他} \end{cases}$$

则

$$f(X) = \frac{1}{2} \|P \odot (X - M)\|_F^2$$

应用举例：低秩矩阵恢复

- 进一步可以得到

$$\begin{aligned}\nabla f(X) &= P \odot (X - M) \\ \text{prox}_{t_k h}(X) &= U \text{Diag}(\max\{|d| - t_k \mu, 0\}) V^\top\end{aligned}$$

- 得到近似点梯度法的迭代格式

$$\begin{aligned}Y^k &= X^k - t_k P \odot (X^k - M) \\ X^{k+1} &= \text{prox}_{t_k h}(Y^k)\end{aligned}$$

收敛性分析

■ 假设 7.1 为了保证近似点梯度算法的收敛性

□ f 在 \mathbb{R}^n 上是凸的; ∇f 为 L -利普希茨连续, 即

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

□ h 是适当的闭凸函数

□ 函数 $\psi(x) = f(x) + h(x)$ 的最小值 ψ^* 是有限的, 并且在点 x^* 处取到

■ 定理 7.3 在假设 7.1 下, 取定步长为 $t_k = t \in (0, \frac{1}{L}]$, 设 $\{x^k\}$ 为迭代产生序列, 则

$$\psi(x^k) - \psi^* \leq \frac{1}{2kt} \|x^0 - x^*\|^2$$

- 7.1 近似点梯度法
- 7.2 Nesterov 加速算法
- 7.3 近似点算法
- 7.4 分块坐标下降法
- 7.5 对偶算法
- 7.6 交替方向乘子法
- 7.7 随机优化算法

典型问题形式

■ 考虑如下复合优化问题

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

□ $f(x)$ 是连续可微的凸函数, 且梯度是利普西茨连续的

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

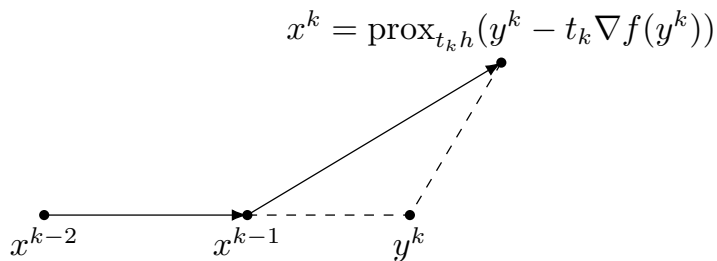
□ $h(x)$ 是适当的闭凸函数, 且邻近算子

$$\text{prox}_h(x) = \arg \min_{u \in \text{dom} h} \left\{ h(u) + \frac{1}{2} \|x - u\|^2 \right\}$$

■ 步长取常数 $t_k = 1/L$ 时, 近似点梯度法的收敛速度为 $\mathcal{O}(1/k)$

Nesterov 加速算法简史

- Nesterov 在 1983、1988、2005 提出了三种改进的一阶算法，收敛速度 $\mathcal{O}\left(\frac{1}{k^2}\right)$
- Beck 和 Teboulle 在 2008 年提出了 FISTA 算法，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



[引用] A method for solving the convex programming problem with convergence rate $O(1/k^2)$

[Y Nesterov](#) - Dokl akad nauk Sssr, 1983 - [cir.nii.ac.jp](#)

A **method** for **solving** the **convex programming problem** with convergence rate $o(1/k^2)$ | CiNii Research ... A **method** for **solving** the **convex programming problem** with convergence rate $o(1/k^2)$...

☆ 保存 ㊄ 引用 被引用次数: 5901 相关文章 所有 5 个版本 ㊄

A fast iterative shrinkage-thresholding algorithm for linear inverse problems

[A Beck](#), [M Teboulle](#) - SIAM journal on imaging sciences, 2009 - SIAM

... algorithm **FISTA** and ... that **FISTA** can be even faster than the proven theoretical rate and can outperform ISTA by several orders of magnitude, thus showing the potential promise of **FISTA**...

☆ 保存 ㊄ 引用 被引用次数: 13863 相关文章 所有 27 个版本

算法 7.2 FISTA 算法

- 1 输入 $x^0 = x^{-1} \in \mathbb{R}^n, k \leftarrow 1$
- 2 **while** 未达到收敛准则 **do**
- 3 计算 $y^k = x^{k-1} + \frac{k-2}{k+1}(x^{k-1} - x^{k-2})$
- 4 选取 $t_k = t \in (0, 1/L]$, 计算 $x^k = \text{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$
- 5 $k \leftarrow k + 1$
- 6 **end while**

FISTA 的等价形式

算法 7.3 FISTA 算法的等价变形

- 1 输入 $v^0 = x^0 \in \mathbb{R}^n, k \leftarrow 1$
- 2 **while** 未达到收敛准则 **do**
- 3 计算 $y^k = (1 - \gamma_k)x^{k-1} + \gamma_k v^{k-1}$
- 4 选取 t_k , 计算 $x^k = \text{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$
- 5 计算 $v^k = x^{k-1} + \frac{1}{\gamma_k}(x^k - x^{k-1})$
- 6 $k \leftarrow k + 1$
- 7 **end while**

第二类 Nesterov 加速算法

■ 第二类 Nesterov 加速算法

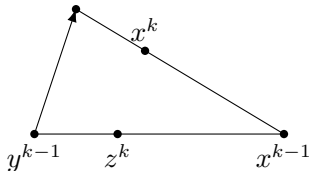
$$z^k = (1 - \gamma_k)x^{k-1} + \gamma_k y^{k-1}$$

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

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

■ 三个序列 $\{x^k\}$, $\{y^k\}$ 和 $\{z^k\}$ 都可以保证在定义域内

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



第三类 Nesterov 加速算法

■ 第三类 Nesterov 加速算法

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

$$y^k = \text{prox}_{(t_k \sum_{i=1}^k 1/\gamma_i)h} \left(-t_k \sum_{i=1}^k \frac{1}{\gamma_i} \nabla f(z^i) \right)$$

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

■ 计算 y^k 时需要利用全部已有的 $\{\nabla f(z^i)\}, i = 1, 2, \dots, k$

■ 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

- 考虑 $f(x)$ 是非凸函数，但可微且梯度是利普希茨连续
- 非凸复合优化问题的加速梯度法框架

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

- 当 λ_k 和 t_k 取特定值时，它等价于第二类 Nesterov 加速算法
- 当 f 为凸函数，收敛速度为 $\mathcal{O}\left(\frac{1}{k^2}\right)$
- 当 f 为非凸函数，收敛速度为 $\mathcal{O}\left(\frac{1}{k}\right)$

应用举例: LASSO 问题求解

- 考虑 LASSO 问题

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

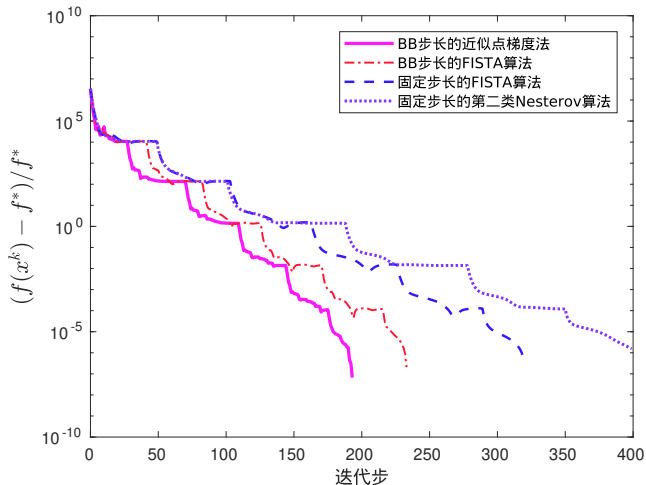
- FISTA 算法可以由下面的迭代格式给出

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

- 第二类和第三类 Nesterov 加速算法

应用举例: LASSO 问题求解

- 取 $\mu = 10^{-3}$, 步长 $t = \frac{1}{L}$, 其中 $L = \lambda_{\max}(A^T A)$



收敛性分析

- **定理 7.5** 在假设 7.1 下, 取定步长 $t_k = t \in (0, 1/L]$. 设 $\{x^k\}$ 是由近似点梯度法迭代产生的序列, 则

$$\psi(x^k) - \psi^* \leq \frac{1}{2kt} \|x^0 - x^*\|^2$$

- **推论 7.1** 在假设 7.1 下, 当用 FISTA 算法求解凸复合优化问题时, 若迭代点 x^k, y^k , 步长 t_k 以及组合系数 γ_k 满足一定条件, 则

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

其中 C 仅与函数 f 和初始点 x^0 的选取有关

- 采用线搜索的 FISTA 算法具有 $\mathcal{O}\left(\frac{1}{k^2}\right)$ 的收敛速度

Q&A

Thank you!

感谢您的聆听和反馈