第七章 复合优化算法

修贤超

https://xianchaoxiu.github.io

目录

邻近算子

- 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

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

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

邻近算子

- **定理 7.2** 若h 是适当的闭凸函数, 则

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

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

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

两边同时加 $\frac{1}{2}\|v - x\|^2$, 即有

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

根据定义可得 $u =_h (x)$

例 7.1

- ℓ_1 **范数** $h(x) = \|x\|_1, \quad \mathrm{prox}_{th}(x) = \mathrm{sign}(x)\max\{|x| - t, 0\}$

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

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

  当 $x > t$ **时,** $u = x - t$; **当** $x < -t$ **时,** $u = x + t$; **当** $x \in [-t, t]$ **时,** $u = 0$
  因此 $u = \mathrm{sign}(x)\max\{|x| - t, 0\}$

例 7.1

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

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

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

当 $\|x\|_2 > t$ 时, $u = x - \frac{tx}{\|x\|_2}$; 当 $\|x\|_2 \leqslant t$ 时, $u = 0$

例 7.2

- 邻近算子的计算规则
  - 变量的常数倍放缩以及平移 $(\lambda \neq 0)$

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

  - 函数（及变量）的常数倍放缩 $(\lambda > 0)$

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

  - 加上线性函数

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

例 7.2

- **加上二次项** $(u > 0)$

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

  其中 $\theta = \frac{1}{1+u}$

- **向量函数**

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

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

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

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

==========

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

对近似点梯度法的理解

- 把迭代公式展开

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

$$\Downarrow$$

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

- 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 $t_k = t \leqslant \frac{1}{L}$. 当 L 未知时可使用线搜索准则

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

- 利用 BB 步长作为 t_k 的初始估计并用非单调线搜索进行校正

$$\alpha_{\mathrm{BB1}}^k = \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}} \quad \text{或} \quad \alpha_{\mathrm{BB2}}^k = \frac{(s^{k-1})^\top s^{k-1}}{(s^{k-1})^\top y^{k-1}}$$

其中 $s^{k-1} = x^k - x^{k-1}$ 以及 $y^{k-1} = \nabla f(x^k) - \nabla f(x^{k-1})$

- 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

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

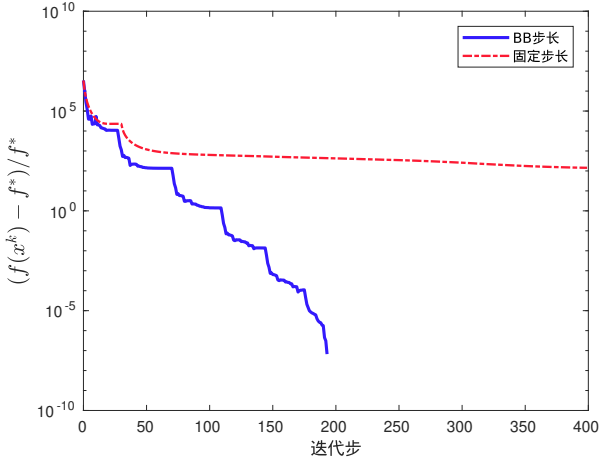
- 相应的迭代格式为

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

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

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

- 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

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

- 得到近似点梯度法的迭代格式

$$Y^k = X^k - t_k P \odot (X^k - M)$$
$$X^{k+1} = \mathrm{prox}_{t_k h}(Y^k)$$

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

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

  - h 是适当的闭凸函数 (因此 $_{th}$ 的定义是合理的)
  - 函数 $\psi(x) = f(x) + h(x)$ 的最小值 ψ^* 是有限的，并且在点 x^* 处可取到 (并不要求唯一)
- **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

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

目录

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

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

- $h(x)$ 是适当的闭凸函数，且临近算子

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

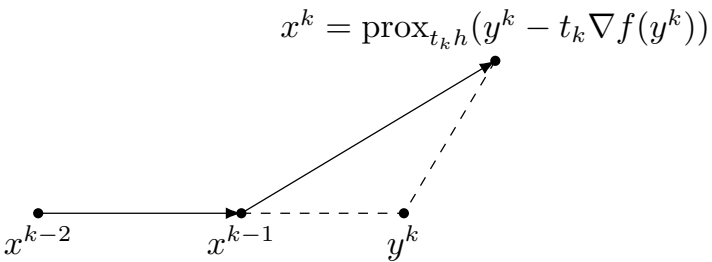
- 近似点梯度法

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

在步长取常数 $t_k = 1/L$ 时，收敛速度为 $(1/k)$

Nesterov 加速算法简史

- 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 $\mathcal{O}\left(\frac{1}{k^2}\right)$
- 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



FISTA 的等价形式

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 $x^k = \mathrm{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$
5 $k \leftarrow k + 1$
6 **end while**

==========

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 $y^k = (1 - \gamma_k)x^{k-1} + \gamma_k v^{k-1}$
4 选取 t_k，计算 $x^k = \mathrm{prox}_{t_k h}(y^k - t_k \nabla f(y^k))$
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**

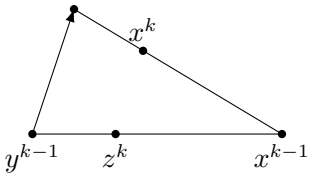
第二类 Nesterov 加速算法

- 第二类 Nesterov 加速算法

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

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

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



第三类 Nesterov 加速算法

- 第三类 Nesterov 加速算法

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

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

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

- 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

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

- 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

应用举例: LASSO 问题求解

- 第二类 Nesterov 加速算法

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

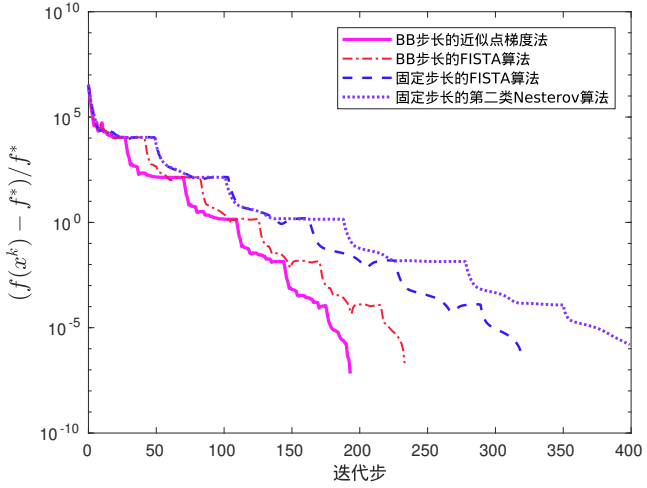
应用举例: LASSO 问题求解

- 第三类 Nesterov 加速算法

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

应用举例: LASSO 问题求解

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



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，初始点 x^0 的选取有关. 特别地，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