# 机器学习中的优化算法

## Lecture09: 复合优化算法–半光滑牛顿算法

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 SSNM

- 广义雅可比
- 半光滑性质
- 半光滑牛顿算法
- 应用举例

# 引言和动机

▶一阶算法困难:

- 尽管一阶算法(近似点梯度法，Nesterov加速算法等等)有非常多的优点，比如易于实现，容易并行并且可以很快的计算低精度的解，但是收敛到高精度的解往往很慢.

- 可以考虑应用牛顿方法来得到更快的收敛速度，但应用牛顿方法有很多的困难:

  ① 在很多应用中，问题不可微，海瑟矩阵不存在.

  ② 并不能保证全局收敛性.

  ③ 如何合理控制住计算牛顿方向的计算代价.

- **目标：**如何对带结构不可微问题引入具有全局收敛性的半光滑牛顿算法？

# 问题引入

▶考虑LASSO问题

$$\min_{x} \ \psi(x) \triangleq \mu\|x\|_1 + \frac{1}{2}\|Ax-b\|^2. \tag{1}$$

- $\ell_1$范数不可微，$\psi(x)$的一个次梯度为$A^{\mathrm{T}}(Ax-b)+\mu\mathrm{sgn}(x)$.

如何定义海瑟矩阵？

- 经典牛顿法求解$\min\phi(x)$的更新格式为:

$$x^{k+1} = x^k - \nabla^2\phi(x^k)^{-1}\nabla\phi(x^k). \tag{2}$$

不可微情形如何定义牛顿法？

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

$$\nabla f(x) = A^{\mathrm{T}}(Ax - b),$$

$$\mathrm{prox}_{t_k h}(x) = \mathrm{sgn}(x)\max\{|x| - t_k\mu, 0\}.$$

- LASSO问题的近似点梯度算法：

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

近似点梯度法收敛到不动点方程的解：

$$F(x) = x - \mathrm{prox}_{th}(x - t\nabla f(x)) = 0.$$

如何对上述方程应用牛顿法?

- 困难：由于prox 算子不可微，如何定义雅可比矩阵？如何定义牛顿法？

# 雅可比矩阵

**Definition 1 (梯度).** *给定函数$f:\mathbb{R}^n\to\mathbb{R}$，且$f$在点$x$的一个邻域内有意义，若存在向量$g\in\mathbb{R}^n$ 满足*

$$\lim_{p\to 0}\frac{f(x+p)-f(x)-g^{\mathrm{T}}p}{\|p\|}=0, \tag{3}$$

*就称$f$ 在点$x$ 处**可微**. 此时$g$ 称为$f$ 在点$x$ 处的**梯度**.*

▶雅可比矩阵: 当$f:\mathbb{R}^n\to\mathbb{R}^m$ 是向量值函数时，可以定义它的雅可比(Jacobi)矩阵$J(x)\in\mathbb{R}^{m\times n}$, 它的第$i$行是分量$f_i(x)$**梯度**的转置:

$$J(x)=\begin{pmatrix}\frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_n}\\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_n}\\ \vdots & \vdots & & \vdots\\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \cdots & \frac{\partial f_m(x)}{\partial x_n}\end{pmatrix}.$$

# 广义雅可比

假定$\Omega \subseteq \mathbb{R}^n$是开集，$F:\Omega \to \mathbb{R}^m$是局部利普希茨连续的，根据Rademacher定理，$F$几乎处处可微, 故可引入广义微分的概念.

**Definition 2** (Clark广义雅可比)**.** *设$F:\Omega \to \mathbb{R}^m$是局部利普希茨连续的, $D_F$是$\Omega$中$F$可微的点组成的集合, $F$在$x \in \Omega$ 的B-次微分定义为*

$$\partial_B F(x) := \left\{ \lim_{k\to\infty} \nabla F(x^k) | x^k \in D_F, x^k \to x \right\}.$$

*Clarke广义雅可比定义为B-次微分的凸包*

$$\partial F(x) = \mathrm{conv}(\partial_B F(x)).$$

►注: 如果$F:\Omega \to \mathbb{R}^m$ 上局部利普希茨连续，则对任意$x \in \Omega$，广义雅可比$\partial F(x)$ 是非空紧凸集.

# 广义雅克比的性质

▶ 相差一个零测集意义下，广义雅克比的矩阵向量乘是一样的.

**Theorem 1.** *设*$F:\Omega\to\mathbb{R}^m$*是局部利普希茨连续的,* $D_F$*是*$\Omega$*中*$F$*可微的点组成的集合, 取*$S$*是一个零测集，定义*

$$\partial_S F(x)=\operatorname{conv}\left(\left\{\lim_{k\to\infty}\nabla F(x^k)|x^k\in D_F, x^k\notin S, x^k\to x\right\}\right).$$

*则对于任意的*$v\in\mathbb{R}^n$ *和*$w\in\mathbb{R}^m$*，有*

$$\partial F(x)v=\partial_S F(x)v,\qquad \partial F(x)^*w=\partial_S F(x)^*w,$$

*其中*$*$*表示集合中每一个元素都转置.*

▶ 复合函数的广义雅克比的运算法则,只有当外层函数可微时，链式法则才是成立的.

**Theorem 2.** *设*$f=g\circ F$*，其中*$F:\mathbb{R}^n\to\mathbb{R}^m$*，并且在*$x$*附近是利普希茨连续的，*$g:\mathbb{R}^m\to\mathbb{R}$*在*$F(x)$*附近是利普希茨的. 那么*$f$*在*$x$*附近是利普希茨的且*

$$\partial f(x)\subset\operatorname{conv}\left\{\partial g(F(x))\partial F(x)\right\}.$$

*如果$g$在$F(x)$点是可微的，那么等式成立，即有*

$$\partial f(x) = \partial g(F(x))\partial F(x).$$

▶复合映射的广义雅克比的运算法则是在矩阵向量乘意义下的.

**Theorem 3.** *假设$F : \mathbb{R}^n \to \mathbb{R}^m$在$x$附近是利普希茨的，$G : \mathbb{R}^m \to \mathbb{R}^k$在$F(x)$附近是利普希茨的. 那么我们有*

$$\partial(G \circ F)(x)v \subset \operatorname{conv}\{\partial G(F(x))\partial F(x)v\}.$$

*如果$G$在$F(x)$附近是连续可微的，那么有*

$$\partial(G \circ F)(x)v = \partial G(F(x))\partial F(x)v.$$

REF: FRANK H. CLARKE, Optimization and Nonsmooth Analysis. New York: Wiley, 1983.

# 单调映射的广义雅克比

▶ 单调映射的广义雅克比中每一个广义雅克比矩阵都是半正定的:

**Theorem 4.** *对于利普希茨连续映射$F:\mathbb{R}^n\to\mathbb{R}^n$, 如果$F$ 是单调的，那么对任意$x\in\mathbb{R}^n,\partial_B F(x)$ 中的每个元素都是半正定的.*

*Proof.* 首先用反证法证明对于任何的可微点$\bar{x}$, $\nabla F(\bar{x})$是半正定的. 假设存在常数$a>0$ 和单位向量$d\in\mathbb{R}^n$使得

$$\langle d,\nabla F(\bar{x})d\rangle=-a.$$

对任意的常数$t>0$, 定义函数$\Phi(t):=F(\bar{x}+td)-F(\bar{x})-t\nabla F(\bar{x})d$. 由$F$ 在点$\bar{x}$处可微知, $\|\Phi(t)\|=o(t)$. 由映射$F$的单调性可知,

$$\begin{aligned}0&\le\langle td,F(\bar{x}+td)-F(\bar{x})\rangle=\langle td,t\nabla F(\bar{x})d+\Phi(t)\rangle\\&\le -at^2+t\|d\|\|\Phi(t)\|=-at^2+o(t^2).\end{aligned}$$

当$t$充分小时, 有$-at^2+o(t^2)<0$, 矛盾! 故所有可微点的雅克比矩阵都是半正定的.

由B-次微分的定义知，对$\forall x\in\mathbb{R}^n,\forall J\in\partial_B F(x)$, 存在一个收敛到$x$的可微点序列$x^k\to x$ 使得$\nabla F(x^k)\to J$. 因为每一个$\nabla F(x^k)$ 都是半正定的，故$J$ 也是半正定的. □

# 邻近算子的广义雅克比

**Theorem 5.** *设$g$ 是$\mathbb{R}^n$ 上的适当闭凸函数, $x \in \mathbb{R}^n$, $\gamma > 0$. 则对任意的$J \in \partial(\mathbf{prox}_{\gamma g}(x))$, $J$ 是对称半正定矩阵且$\|J\|_2 \leq 1$.*

**Proposition 1.** *设$g : \mathbb{R}^n \to \mathbb{R}$ 是(块)可分的，即$g$可表示成: $g(x) = \sum_{i=1}^{n} g_i(x_i)$, ($g(x) = \sum_{i=1}^{k} g_i(x_i)$, 其中$x_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^{k} n_i = n$ 是$x$ 的所有分量的块划分), 则$\partial_{\mathsf{B}}(\mathbf{prox}_{\gamma g}(x))$ 和$\partial(\mathbf{prox}_{\gamma g}(x))$ 中的所有元素均为(块)对角矩阵.*

►注: 可分函数的邻近算子的广义雅克比具有对角结构，这对算法中降低运算量很有意义.

证： 根据邻近算子的定义，可分函数$g$ 的邻近算子具有可分的结构:

$\mathbf{prox}_{\gamma g}(x) = (\mathbf{prox}_{\gamma g_1}(x_1), \ldots, \mathbf{prox}_{\gamma g_n}(x_n)) \Rightarrow \partial_{\mathsf{B}}(\mathbf{prox}_{\gamma g})(x)$ 由对角矩阵组成

**Proposition 2.** *设$g$ 是$\mathbb{R}^n$ 上适当的闭凸函数，$g^*$ 为其共轭函数，则*

$$\partial_B(\mathbf{prox}_{\gamma g^*}(x)) = I - \partial_B(\mathbf{prox}_{g/\gamma}(x/\gamma)), \quad \partial(\mathbf{prox}_{\gamma g^*}(x)) = I - \partial(\mathbf{prox}_{g/\gamma}(x/\gamma)).$$

# 常见凸函数的邻近算子的广义雅可比

- 超平面: $D=\{x|Ax=b\}$，其中$A\in\mathbb{R}^{m\times n}$. $\mathcal{P}_D(x)=x-A^{\dagger}(Ax-b)$,

$$\partial(\mathcal{P}_D(x))=\partial_{\mathrm{B}}(\mathcal{P}_D(x))=\nabla\mathcal{P}_D(x)=\{I-A^{\dagger}A\}.$$

- 半空间: $D=\{x|a^{\top}x\le b\}$.

$$\mathcal{P}_D(x)=x-\frac{(a^{\top}x-b)_+}{\|a\|_2^2}a,$$

$$\partial(\mathcal{P}_D(x))=\begin{cases}\left\{I-\dfrac{aa^{\top}}{\|a\|_2^2}\right\} & \textbf{if } a^{\top}x>b,\\ \{I\} & \textbf{if } a^{\top}x<b,\\ \operatorname{conv}\left\{I,\ I-\dfrac{aa^{\top}}{\|a\|_2^2}\right\} & \textbf{if } a^{\top}x=b.\end{cases}$$

- 单位球: $B = \{x \mid \|x\|_2 = 1\}$.

$$\mathcal{P}_B(x) = \begin{cases} \dfrac{x}{\|x\|_2} & \text{if } \|x\|_2 > 1, \\ x & \text{if } \|x\|_2 \le 1. \end{cases}$$

定义$w = \dfrac{x}{\|x\|_2}$，则

$$\partial(\mathcal{P}_B(x)) = \begin{cases} \left\{ \dfrac{I - ww^\top}{\|x\|_2} \right\}, & \text{if } \|x\|_2 > 1, \\ \{I\}, & \text{if } \|x\|_2 < 1, \\ \operatorname{conv}\left\{ \dfrac{I - ww^\top}{\|x\|_2}, I \right\}, & \text{if } \|x\|_2 = 1. \end{cases}$$

- $\ell_2$ 范数: 设$g = \|x\|_2$，则

$$\mathbf{prox}_{\gamma g}(x) = \begin{cases} \left(1 - \dfrac{\gamma}{\|x\|_2}\right) x, & \text{if } \|x\|_2 \ge \gamma, \\ 0, & \text{if } \|x\|_2 < \gamma. \end{cases}$$

因$\mathbf{prox}_{\gamma g}(x)$ 是分片光滑的，故其B-次微分可以通过分片求其雅可比矩阵得到. 令$w = \dfrac{x}{\|x\|_2}$，则

$$\partial_{\mathrm{B}}(\mathbf{prox}_{\gamma g}(x)) = \begin{cases} \left\{ I - \dfrac{\gamma}{\|x\|_2}(I - ww^\top) \right\}, & \text{if } \|x\|_2 \geq \gamma, \\ \{0\}, & \text{if } \|x\|_2 < \gamma, \\ \left\{ I - \dfrac{\gamma}{\|x\|_2}(I - ww^\top),\ 0 \right\}, & \text{if } \|x\|_2 = \gamma. \end{cases}$$

- $\ell_1$ 范数: 设$g = \|x\|_1$，则

$$\left(\mathrm{prox}_{\gamma g}(x)\right)_i = \mathrm{sgn}(x_i)\max(|x_i| - \gamma, 0), \quad 1 \leq i \leq n.$$

因$\mathbf{prox}_{\gamma g}(x)$ 是可分的，故$\partial_{\mathrm{B}}(\mathbf{prox}_{\gamma g}(x))$ 中的每个元素均为对角矩阵，则

$$J \in \partial_{\mathrm{B}}(\mathbf{prox}_{\gamma g}(x)), \quad J_{ii} = \begin{cases} 1, & \text{if } i \in \{i \mid |x_i| > \gamma\}, \\ [0,1], & \text{if } i \in \{i \mid |x_i| = \gamma\}, \\ 0, & \text{if } i \in \{i \mid |x_i| < \gamma\}. \end{cases}$$

# 谱函数的邻近算子的广义雅可比

- 谱函数: $F:\mathbb{S}^n\to\mathbb{R}\cup\{\infty\}$

$$F(X)=f(\lambda(X)),\quad X\in\mathbb{S}^n, \tag{4}$$

其中$\lambda:\mathbb{S}^n\to\mathbb{R}^n$ 为对应矩阵的特征值(从大到小排列),
$f:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$是一个适当的闭凸函数, 且绝对对称:

$$f(x)=f(Px),\quad \forall x\in\mathbb{R}^n,\forall\ \text{置换矩阵}P\in\mathbb{R}^{n\times n}.$$

- 设实对称矩阵$X$ 的谱分解为$X=Q\mathbf{diag}(\lambda(X))Q^\top$. 如果$f$是可微的，那么$F$也是可微的，且有

$$\nabla F(X)=Q\nabla f(\lambda(X))Q^T.$$

$F$的邻近似算子为：

$$\mathbf{prox}_{\gamma F}(X)=Q\mathbf{diag}(\mathbf{prox}_{\gamma f}(\lambda(X)))Q^\top.$$

- 设$f$有形式$f(x)=g(x_1)+\cdots+g(x_n)$，则$F$的邻近似算子可以写为

$$\mathbf{prox}_{\gamma F}(X)=Q\mathbf{diag}(\mathbf{prox}_{\gamma g}(\lambda_1(X)),\dots,\mathbf{prox}_{\gamma g}(\lambda_n(X)))Q^\top. \tag{5}$$

**Theorem 6.** *设$h:\mathbb{R}\to\mathbb{R}$是局部利普希茨连续的，假定实对称矩阵$X$的特征分解为$X=Q\mathbf{diag}(\lambda_1,\ldots,\lambda_n)Q^\top$，算子$H:\mathbb{S}^n\to\mathbb{S}^n$定义为:*

$$H(X)=Q(h(\lambda_1),\ldots,h(\lambda_n))Q^\top.$$

*则对于任意的$X\in\mathbb{S}^n$，B-次微分$\partial_B H$存在且非空，且对于任意的$J\in\partial_B H$，有*

$$J(S)=Q(\Omega\odot(Q^\top SQ))Q^\top,\quad \forall S\in\mathbb{S}^n,$$

*其中$\odot$表示Hadamard积，而矩阵$\Omega\in\mathbb{R}^{n\times n}$的各个元素定义如下:*

$$\Omega_{ij}\begin{cases}=\dfrac{h(\lambda_i)-h(\lambda_j)}{\lambda_i-\lambda_j}, & \textbf{if } \lambda_i\neq\lambda_j,\\ \in\partial h(\lambda_i), & \textbf{if } \lambda_i=\lambda_j.\end{cases}$$


- CHEN X, QI H, TSENG P. *Analysis of nonsmooth symmetricmatrix- valued functions with applications to semidefinite complementarity problems*. SIAM Journal on Optimization, 2003, 13(4): 960–985.
- DING C, SUN D, SUN J. *Spectral operators of matrices*. Mathematical Programming, 2018, 168(1-2): 509-531.
- DING C, SUN D, SUN J. *Spectral operators of matrices: semismoothness and characterizations of the generalized jacobian*. SIAM Journal on Optimization, 2020, 30(1): 630-659.

- $F$的邻近似算子$\mathbf{prox}_{\gamma F}$ (5)的$B$-次微分:

  对于任意$X \in \mathbb{S}^n$ 和$P \in \partial_{\mathrm{B}}(\mathbf{prox}_{\gamma F})(X)$，有

$$P(S) = Q(\Omega \odot (Q^\top S Q))Q^\top, \quad \forall S \in \mathbb{S}^n, \tag{6}$$

  其中矩阵$\Omega \in \mathbb{R}^{n\times n}$ 的各个元素按如下方式定义

$$\Omega_{ij} \begin{cases} = \dfrac{\mathbf{prox}_{\gamma g}(\lambda_i) - \mathbf{prox}_{\gamma g}(\lambda_j)}{\lambda_i - \lambda_j}, & \text{if } \lambda_i \neq \lambda_j, \\ \in \partial(\mathbf{prox}_{\gamma g}(\lambda_i), & \text{if } \lambda_i = \lambda_j. \end{cases}$$

- 半正定锥的指示函数:

  在(5)中, 令$g = \delta_{\mathbb{R}_+}$， 则$F$ 可以看作是正定锥$\mathbb{S}^n_+$ 的指示函数, 且

$$\mathbf{prox}_{\gamma g}(x) = \mathcal{P}_{\mathbb{R}_+}(x) = \max\{x, 0\} = (x)_+.$$

  于是, 由(5)得

$$\mathcal{P}_{\mathbb{S}^n_+}(X) = Q\mathbf{diag}((\lambda_1)_+, \ldots, (\lambda_n)_+)Q^\top.$$

定义

$$\alpha = \{i | \lambda_i > 0\}, \quad \bar{\alpha} = \{i | \lambda_i \le 0\},$$

则对任意的$P \in \partial_B \mathcal{P}_{\mathbb{S}^n_+}(X)$有

$$P(S) = Q(\Omega \odot (Q^\top S Q))Q^\top, \quad \forall S \in \mathbb{S}^n,$$

其中

$$\Omega = \begin{pmatrix} \Omega_{\alpha\alpha} & \Omega_{\alpha\bar{\alpha}} \\ \Omega_{\alpha\bar{\alpha}}^\top & 0 \end{pmatrix},$$

$\Omega_{\alpha\alpha} \in \mathbb{R}^{|\alpha| \times |\alpha|}$ 的元素全为1，而$\Omega_{\alpha\bar{\alpha}} \in \mathbb{R}^{|\alpha| \times |\bar{\alpha}|}$ 且其第$(i, j)$元素为

$$\frac{\lambda_i}{\lambda_i - \lambda_j}, \quad i \in \alpha, j \in \bar{\alpha}.$$


- LEWIS A S, SENDOV H S. *Twice differentiable spectral functions*. SIAM Journal on Matrix Analysis and Applications, 2001, 23(2): 368-386.
- QI H, YANG X. *Semismoothness of spectral functions*. SIAM Journal on Matrix Analysis and Applications, 2003, 25(3): 766-783.

# 半光滑性

**Definition 3.** 设$F:\Omega\to\mathbb{R}^m$是局部利普希茨连续的，称$F$在$x$处是半光滑的(强半光滑)，如果满足

(a) $F$在$x$点是方向可微的;

(b) 对于任意的$d$和$J\in\partial F(x+d)$, 有

$$\|F(x+d)-F(x)-Jd\|=o(\|d\|),\quad \text{as } d\to 0. \qquad \text{(semismooth)}$$

$$\|F(x+d)-F(x)-Jd\|=O(\|d\|^2),\quad \text{as } d\to 0. \qquad \text{(strongly semismooth)}$$

- 半光滑性和强半光滑性在数乘、求和和复合运算下都是封闭的.
- 光滑函数、所有的凸函数、分段连续可微的函数都是半光滑的.
- 具有利普希茨连续梯度的可微函数、$p$范数$\|\cdot\|_p$和分段线性函数是强半光滑的.
- 一个向量值函数是半光滑的(或强半光滑的)当且仅当每个分量函数是半光滑的(或强半光滑的).

- 如果一个函数是分段$C^1$ 的，则它是半光滑的；如果一个函数是分段$C^2$ 的，则它是强半光滑的.
- 很多函数的邻近算子具有半光滑性和和强半光滑性:

① $\|x\|_1$ 与$\|x\|_\infty$ 的邻近算子是强半光滑的.

② 多面体的投影是强半光滑的.

③ 对称锥上的投影是强半光滑的，因此半定锥和二阶锥上的投影都是强半光滑的.

④ 在许多应用中, 邻近算子是逐段$C^1$的, 因此是半光滑的.

⑤ 对于一般的凸函数$f$，它的邻近算子不一定是半光滑的. 凸函数$f$的邻近算子是(强) 半光滑的当且仅当其上方图epi$f$ 上的投影算子是(强) 半光滑的. ▶ MENG F, SUN D, ZHAO G. *Semismoothness of solutions to generalized equations and Moreau-Yosida regularization*. Mathematical programming, 2005, 104(2): 561-581.

▶ $\ell_1$范数$\|x\|_1$的邻近算子$\phi(x) = \mathrm{sgn}(x)\max(|x| - \mu t, 0)$ 是强半光滑的.

*Proof.* 考察一维的情形, $\phi(x)$的广义雅克比为

$$\partial\phi(x) = \begin{cases} \{1\}, & \textbf{if } |x| > \mu t, \\ [0,1], & \textbf{if } |x| = \mu t, \\ \{0\}, & \textbf{if } |x| < \mu t. \end{cases}$$

当$|x| \neq \mu t$ 时, 函数$\phi(x)$是可微的, 因此是强半光滑的. 故只需要验证在两个不可微点, 强半光滑性是成立的. 对于$x = \mu t$，如果$d > 0$，则$x + d > \mu t$，因此其广义雅克比为$\partial\phi(x) = \{1\}$，则对于任意的$J \in \partial\phi(x)$，我们有

$$|\phi(x+d) - \phi(x) - Jd| = 0.$$

如果$-2\mu t < d < 0$，则$-\mu t < x + d < \mu t$，因此其广义雅克比为$\partial\phi(x) = \{0\}$，则对于任意的$J \in \partial\phi(x)$，我们有

$$|\phi(x+d) - \phi(x) - Jd| = 0.$$

因此可得在$x = \mu t$处，$\phi(x)$是强半光滑的. 类似的，可以证明$x = -\mu t$也是强半光滑的. 故$\phi(x)$是强半光滑的. □

# 半光滑牛顿算法

- 许多算子分裂算法(近似点梯度法、DRS 算法)，等价于一个不动点迭代，其可以诱导一个非线性方程组

$$F(z) = 0. \tag{7}$$

求解非线性方程组(7) 即能求解原始的复合优化问题.

- 求解(7)的半光滑牛顿算法: 假定$F$是局部利普希茨连续的, 则其广义雅克比存在. 取$F$在$z^k$点任意的广义雅克比矩阵$J_k \in \partial F(z^k)$，若$J_k$可逆，则半光滑牛顿算法的迭代格式为

$$z^{k+1} = z^k - J_k^{-1} F(z^k). \tag{8}$$

- 半光滑性也能保证牛顿型算法具有超线性收敛或二次收敛.

- BD正则： 若所有的$J \in \partial_B F(z)$ 都是非奇异的，则称$F$ 在$z$ 点是BD-正则的. BD-正则性是一个非光滑方法局部收敛性分析的普遍假设.

▶ 半光滑牛顿法的局部收敛性:

**Assumption 1.** *定义在(7) 中的映射$F$在最优点$z^*$是半光滑的和BD-正则的.*

**Lemma 1.** *如果假设1 成立，则存在常数$c > 0$，$\kappa > 0$ 和一个小邻域$N(z^*, \varepsilon_0)$ 使得对于任意的$y \in N(z^*, \varepsilon_0)$ 和$J \in \partial_B F(y)$，下面的结论成立：*

① *$z^*$ 是一个孤立解；*

② *$J$ 是非奇异的并且$\|J^{-1}\| \leq c$；*

③ *局部误差界条件对于$F(z)$ 在邻域$N(z^*, \varepsilon_0)$ 上成立: $\|y - z^*\| \leq \kappa \|F(y)\|$.*

**Theorem 7.** *设假设1 成立且$z^*$是$F(z) = 0$的解, 则存在一个小邻域$N(z^*, \epsilon)$，使得迭代(8) 是良定义的，且对任意的$z^k \in N(z^*, \epsilon)$，迭代(8) 是超线性收敛的．如果$F$是强半光滑的，迭代(8) 是二次收敛的.*

*Proof.* 由引理1，迭代(8) 是良定义的，且对任意的$z^k \in N(z^*, \epsilon)$有

$$\begin{aligned}\|z^{k+1} - z^*\| &= \|z^k - J_k^{-1} F(z^k) - z^*\| \\ &\leq \|J_k^{-1}\| \cdot \|F(z^k) - F(z^*) - J_k(z^k - z^*)\| \\ &= o(\|z^k - z^*\|).\end{aligned}$$

故迭代(8) 是超线性收敛的. □

# 求解优化问题的半光滑牛顿算法

- 半光滑牛顿法也可用于求解优化问题

$$\min_x f(x), \tag{9}$$

其中$f(x)$是可微的，但不是二阶可微的．其最优性条件为$\nabla f(x)=0$．半光滑牛顿法迭代: $x^{k+1}=x^k-J_k^{-1}\nabla f(x^k)$.

- 若(9)为凸优化，则其广义海瑟矩阵是对称半正定的，因此可以选择任意的$J_k\in\partial_B(\nabla f(x^k))$，选取正则化参数$\mu_k>0$, 计算线性方程组

$$(J_k+\mu_k I)d^k=-\nabla f(x^k) \tag{10}$$

来得到半光滑牛方向$d^k$, 为优化问题(9)的下降方向.

- 可以利用Armijo 线搜索准则选取步长，即选取最小的非负常数$m_k$，满足

$$f(x^k+\rho^m d^k)\le f(x^k)+\sigma\rho^m\nabla f(x^k)^T d^k, \tag{11}$$

其中$\rho,\sigma\in(0,1)$ 是给定的常数．取下一步的迭代点为

$$x^{k+1}=x^k+\rho^{m_k}d^k. \tag{12}$$

- 二次收敛性： 设$x^*$是优化问题(9) 的最优解，$f(x)$ 是凸函数且具有利普希茨连续的梯度，其梯度$\nabla f(x)$ 是半光滑的和BD- 正则的． 如果$\sigma < \frac{1}{2}$，那么算法(10)-(11) 产生的序列$\{x^k\}$ 满足

  1. 存在整数$k_0$，使得对所有的$k \geq k_0$ 有$m_k = 0$.
  2. 整个序列$\{x^k\}$ 收敛到$x^*$ 且具有二次收敛性.

## ■应用举例

▶LASSO问题：基于近似点梯度法的半光滑牛顿算法

考虑LASSO问题

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

- 令$f(x) = \mu\|x\|_1$ 和$h(x) = \dfrac{1}{2}\|Ax - b\|^2$. 则利用近似点梯度法可将需要求解的问题(13)表示成求解下面的非线性方程组

$$F(x) = x - \mathrm{prox}_{tf}(x - t\nabla h(x)) = 0, \quad 其中\nabla h(x) = A^T(Ax - b).$$

- $F(x)$的一个广义雅克比矩阵可以表示为

$$J(x) = I - M(x)(I - tA^T A), \quad M(x) \in \partial \mathrm{prox}_{tf}(x - t\nabla h(x)).$$

- $f(x)$的邻近算子为收缩算子

$$\left(\mathrm{prox}_{tf}(x)\right)_i = \mathrm{sgn}(x_i) \max(|x_i| - \mu t, 0).$$

因此，$M(x)$是一个对角矩阵，其对角元素是

$$(M(x))_{ii} = \begin{cases} 1, & \text{if } |(x - t\nabla h(x))_i| > \mu t, \\ 0, & \text{otherwise.} \end{cases}$$

- 定义指标集合

$$\mathcal{I}(x) := \{i : |(x - t\nabla h(x))_i| > t\mu\} = \{i : (M(x))_{ii} = 1\},$$
$$\mathcal{O}(x) := \{i : |(x - t\nabla h(x))_i| \le t\mu\} = \{i : (M(x))_{ii} = 0\}.$$

则$F(x)$的雅克比矩阵$J(x)$可表示成

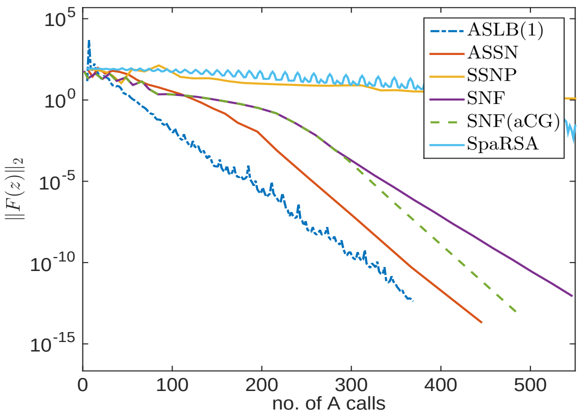
$$J(x) = \begin{pmatrix} t(A^TA)_{\mathcal{I}(x)\mathcal{I}(x)} & t(A^TA)_{\mathcal{I}(x)\mathcal{O}(x)} \\ 0 & I \end{pmatrix}.$$

- 利用$J(x)$的分块结构，可以降低求解线性方程组(10)的复杂度.
令$\mathcal{I} = \mathcal{I}(x^k)$和$\mathcal{O} = \mathcal{O}(x^k)$，则(10)可表示为

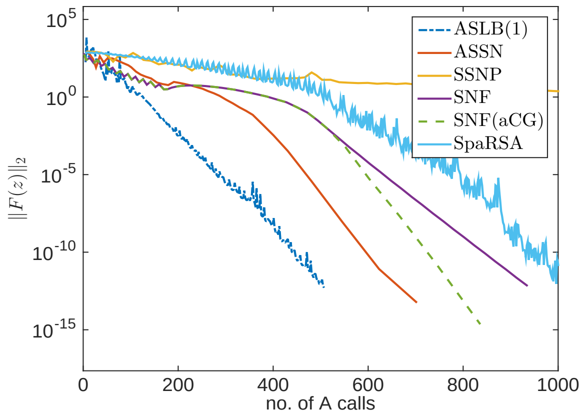
$$\begin{cases} (1+\mu_k)d_{\mathcal{O}}^k = -F_{\mathcal{O}}^k, \quad \Rightarrow \quad d_{\mathcal{O}}^k = -\dfrac{1}{1+\mu_k}F_{\mathcal{O}}^k, \\ (t(A^TA)_{\mathcal{I}\mathcal{I}} + \mu I)d_{\mathcal{I}}^k + t(A^TA)_{\mathcal{I}\mathcal{O}}d_{\mathcal{O}}^k = -F_{\mathcal{I}}^k, \end{cases} \Longrightarrow$$

$$(t(A^TA)_{\mathcal{I}\mathcal{I}} + \mu_k I)d_{\mathcal{I}}^k = -F_{\mathcal{I}}^k + \frac{t}{1+\mu_k}(A^TA)_{\mathcal{I}\mathcal{O}}F_{\mathcal{O}}^k. \tag{14}$$

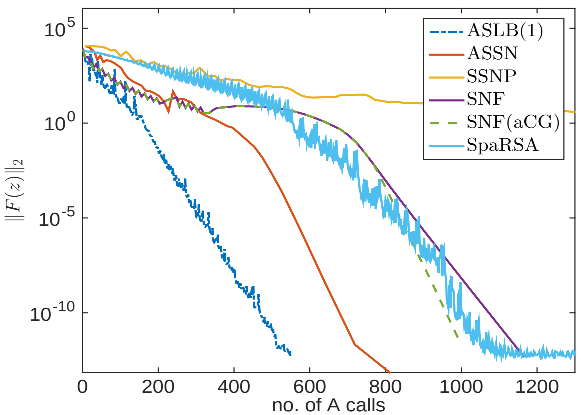
半光滑牛顿法求解LASSO (13)只需要求解规模为$|\mathcal{I}|$的线性方程组(14)即可. 由于$\ell_1$范数能够保证问题(13) 的解是稀疏的，而指标集$\mathcal{I}$恰好是解非零元的位置，因此在实际问题$|\mathcal{I}|$是非常小的. 这表明半光滑牛顿法很好地利用了问题的稀疏结构，求解牛顿方向的代价是比较小的.



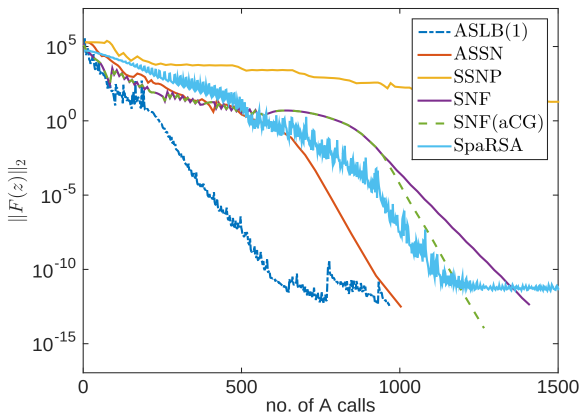

(a) 20dB



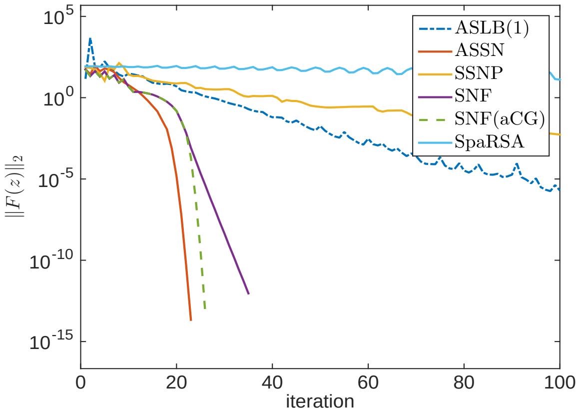

(b) 40dB



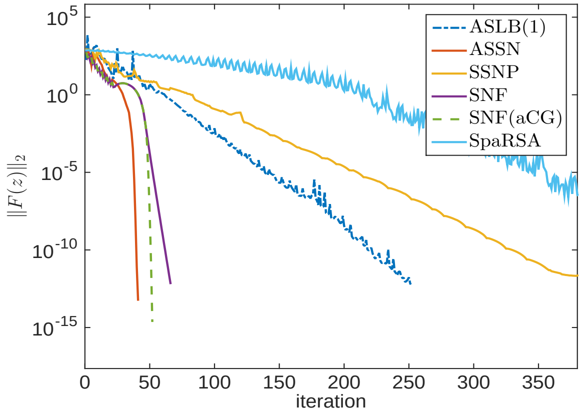

(c) 60dB



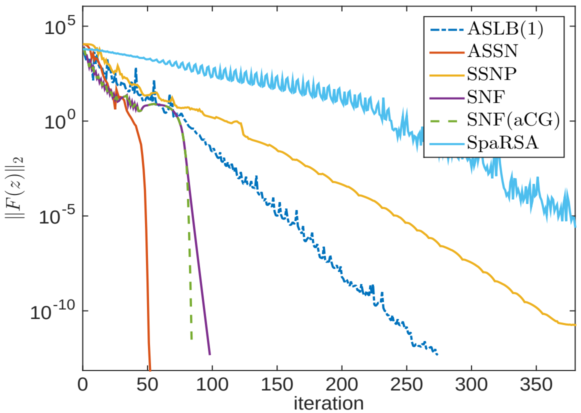

(d) 80dB



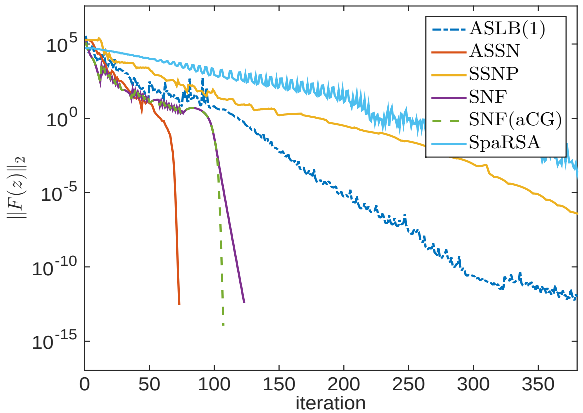

(e) 20dB




(f) 40dB




(g) 60dB




(h) 80dB

## ▶基追踪问题：基于DRS 算法的半光滑牛顿算法

考虑基追踪(BP) 问题

$$\min \ \|x\|_1 \quad \textbf{s.t.} \quad Ax = b, \tag{15}$$

其中$b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ 是行满秩的, 为简化假定$AA^\top = I$.

- 令$f(x) = 1_\Omega(Ax - b)$ 和$h(x) = \|x\|_1$，其中$\Omega = \{0\}$，则DRS 算法不动点映射对应的非线性方程组为

$$F(z) = \mathrm{prox}_{th}(z) - \mathrm{prox}_{tf}(2\mathrm{prox}_{th}(z) - z) = 0. \tag{16}$$

- $f(x)$ 的近似点算子可以写为

$$\mathrm{prox}_{tf}(z) = z - A^\top(Az - b).$$

- 函数$h(x)$ 的近似点算子是

$$(\mathrm{prox}_{th}(z))_i = \mathrm{sgn}(z_i)\max(|z_i| - t, 0).$$

其广义雅可比矩阵$M(z) \in \partial\mathrm{prox}_{th}(z)$为对角矩阵，且对角元为

$$M(z)_{ii} = \begin{cases} 1, & \text{if } |z_i| > t, \\ 0, & \text{otherwise.} \end{cases}$$

- 映射$F(z)$ (16)的一个广义雅可比矩阵可表示为

$$J(z) = M(z) + (I - A^\top A)(I - 2M(z)). \tag{17}$$

- 令$W = I - 2M(z)$ 和$H = W + M(z) + \mu I$，则$W$ 和$H$为对角阵，其对角元：

$$W(z)_{ii} = \begin{cases} -1, & \text{if } |z_i| > t, \\ 1, & \text{otherwise}, \end{cases} \qquad H(z)_{ii} = \begin{cases} \mu, & \text{if } |z_i| > t, \\ 1+\mu, & \text{otherwise}. \end{cases}$$

因此, $WH^{-1} = \dfrac{1}{1+\mu} I - S$, 其中$S$ 也是对角矩阵，其对角元为

$$S_{ii}(z) = \begin{cases} \dfrac{1}{\mu} + \dfrac{1}{1+\mu}, & \textbf{if } |z_i| > t, \\ 0, & \textbf{otherwise}. \end{cases} \tag{18}$$

根据SMW公式，可得

$$\begin{aligned} (J(z) + \mu I)^{-1} &= (H - A^\top A W)^{-1} \\ &= H^{-1} + H^{-1} A^\top (I - AWH^{-1}A^\top)^{-1} AWH^{-1}. \end{aligned} \tag{19}$$

- 由(18)知，

$$I - AWH^{-1}A^\top = \left(1 - \frac{1}{1+\mu}\right) I + ASA^\top.$$

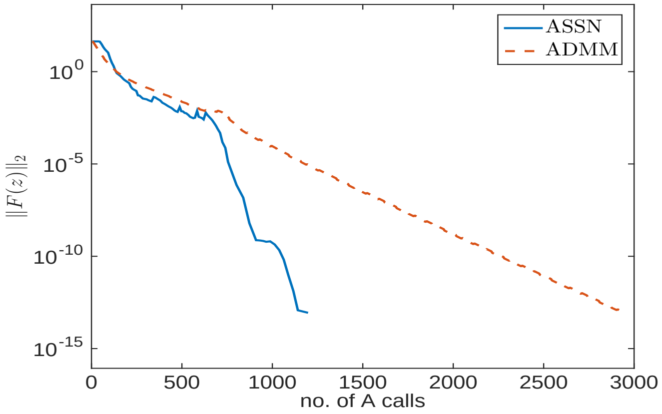
定义指标集合

$$\mathcal{I}(x) := \{i : |(z)_i| > t\} = \{i : M_{ii}(x) = 1\},$$
$$\mathcal{O}(x) := \{i : |(z)_i| \le t\} = \{i : M_{ii}(x) = 0\},$$

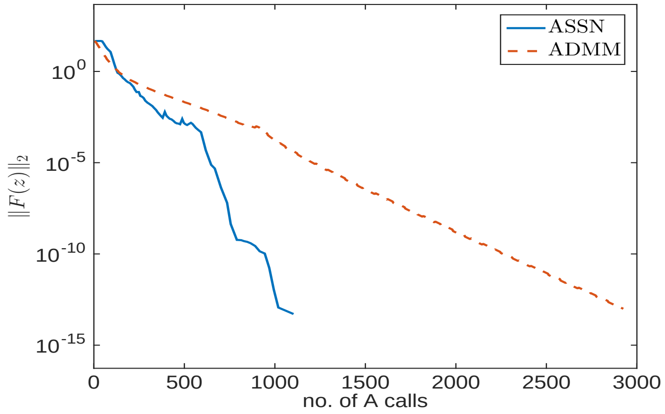
则有

$$ASA^\top = (\frac{1}{\mu} + \frac{1}{1+\mu}) A_{\mathcal{I}(x)} A_{\mathcal{I}(x)}^\top. \tag{20}$$

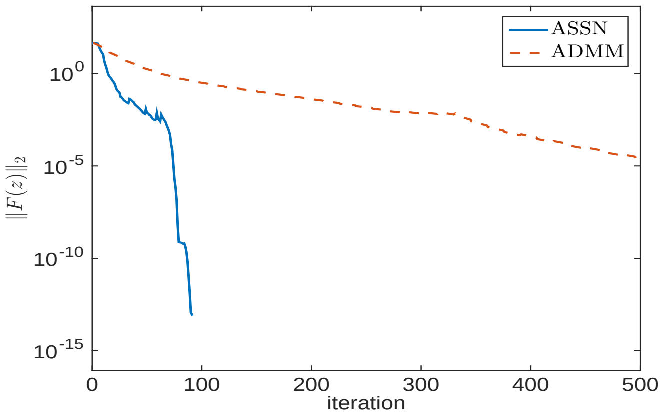
- 由(20)知，$I - AWH^{-1}A^\top$ 是正定的．如果子矩阵$A_{\mathcal{I}(x)}$ 是容易获得的，可以避免在求解牛顿方向时使用更大的矩阵$A$，从而降低求解牛顿方向的计算复杂度.



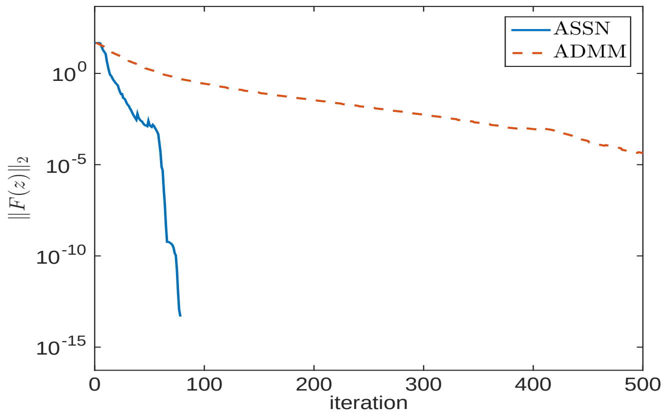

(i) 60dB




(j) 80dB




(k) 60dB




(l) 80dB