分块坐标下降法

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：本教案由朱桢源协助准备

提纲

1. 分块坐标下降法
2. 应用举例
3. 收敛性分析

问题形式

考虑具有如下形式的问题：

$$\min_{x \in \mathcal{X}} \quad F(x_1, x_2, \cdots, x_s) = f(x_1, x_2, \cdots, x_s) + \sum_{i=1}^{s} r_i(x_i),$$

- $\mathcal{X}$是函数的可行域，自变量x拆分成s个变量块$x_1, x_2, \cdots, x_s$，每个变量块$x_i \in \mathbb{R}^{n_i}$.
- 函数f是关于x的可微函数，每个$r_i(x_i)$关于x_i是适当的闭凸函数，但不一定可微.
- 目标函数F的性质体现在f，每个r_i以及自变量的分块上. 通常情况下，f对于所有变量块x_i不可分，但单独考虑每一块自变量时，f有简单结构；r_i只和第i个自变量块有关，因此r_i在目标函数中是一个可分项.
- 求解该问题的难点在于如何利用分块结构处理不可分的函数f.

问题形式

- 分组LASSO 模型：参数$x=(x_1,x_2,\cdots,x_G)\in\mathbb{R}^p$ 可以分成G 组，且$\{x_i\}_{i=1}^G$中只有少数的非零向量.

$$\min_x \quad \frac{1}{2n}\|b-Ax\|_2^2+\lambda\sum_{i=1}^G\sqrt{p_i}\|x_i\|_2.$$

- K-均值聚类问题的等价形式：

$$\begin{aligned}\min_{\Phi,H} \quad & \|A-\Phi H\|_F^2,\\ \text{s.t.} \quad & \Phi\in\mathbb{R}^{n\times k}, \text{每一行只有一个元素为1，其余为0},\\ & H\in\mathbb{R}^{k\times p}.\end{aligned}$$

- 低秩矩阵恢复: 设$b\in\mathbb{R}^m$是已知的观测向量，$\mathcal{A}$是线性映射.

$$\min_{X,Y} \quad \frac{1}{2}\|\mathcal{A}(XY)-b\|_2^2+\alpha\|X\|_F^2+\beta\|Y\|_F^2,$$

其中$\alpha,\beta>0$为正则化参数.

问题形式

- 非负矩阵分解：设$\mathcal{M}$是已知张量，考虑求解如下极小化问题：

$$\min_{A_1,A_2,\cdots,A_N\geq 0} \quad \frac{1}{2}\|\mathcal{M}-A_1\circ A_2\circ\cdots\circ A_N\|_F^2+\sum_{i=1}^{N}\lambda_i r_i(A_i),$$

其中“$\circ$”表示张量的外积运算．

- 字典学习：设$A\in\mathbb{R}^{m\times n}$为$n$个观测，每个观测的信号维数是$m$，现在我们要从$A$中学习出一个字典$D\in\mathbb{R}^{m\times k}$和系数矩阵$X\in\mathbb{R}^{k\times n}$：

$$\begin{aligned}\min_{D,X} \quad & \frac{1}{2n}\|DX-A\|_F^2+\lambda\|X\|_1,\\ \text{s.t.} \quad & \|D\|_F\leq 1.\end{aligned}$$

在这里自变量有两块，分别为D和X，此外对D还存在球约束$\|D\|_F\leq 1$．

挑战和难点

- 函数f关于变量全体一般是非凸的，这使得问题求解具有挑战性
- 应用在非凸问题上的算法收敛性不易分析，很多针对凸问题设计的算法通常会失效
- 目标函数的整体结构十分复杂，变量的更新需要很大计算量
- **目标:** 发展一种更新方式简单且有全局收敛性（收敛到稳定点）的有效算法

变量划分

- **分块坐标下降法更新方式**：按照$x_1, x_2, \cdots, x_s$的次序依次固定其他$(s-1)$块变量极小化F，完成一块变量的极小化后，它的值便立即被更新到变量空间中，更新下一块变量时将使用每个变量最新的值.

- 变量划分

$$\mathcal{X}_i^k = \{x \in \mathbb{R}^{n_i} \mid (x_1^k, \cdots, x_{i-1}^k, x, x_{i+1}^{k-1}, \cdots, x_s^{k-1}) \in \mathcal{X}\}.$$

- 辅助函数

$$f_i^k(x_i) = f(x_1^k, \cdots, x_{i-1}^k, x_i, x_{i+1}^{k-1}, \cdots, x_s^{k-1}),$$

其中x_j^k表示在第k次迭代中第j块自变量的值，函数f_i^k 表示在第k次迭代更新第i块变量时所需要考虑的目标函数的光滑部分.

变量更新方式

在每一步更新中，通常使用以下三种更新格式之一：

$$x_i^k = \operatorname*{argmin}_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + r_i(x_i) \right\}, \tag{1}$$

$$x_i^k = \operatorname*{argmin}_{x_i \in \mathcal{X}_i^k} \left\{ f_i^k(x_i) + \frac{L_i^{k-1}}{2} \|x_i - x_i^{k-1}\|_2^2 + r_i(x_i) \right\}, \tag{2}$$

$$x_i^k = \operatorname*{argmin}_{x_i \in \mathcal{X}_i^k} \left\{ \langle \hat{g}_i^k, x_i - \hat{x}_i^{k-1} \rangle + \frac{L_i^{k-1}}{2} \|x_i - \hat{x}_i^{k-1}\|_2^2 + r_i(x_i) \right\}, \tag{3}$$

- $L_i^k > 0$为常数
- 在更新格式(3)中，$\hat{x}_i^{k-1}$采用外推定义：

$$\hat{x}_i^{k-1} = x_i^{k-1} + \omega_i^{k-1}(x_i^{k-1} - x_i^{k-2}), \tag{4}$$

其中$\omega_i^k \geq 0$为外推的**权重**，$\hat{g}_i^k \overset{\text{def}}{=\!=} \nabla f_i^k(\hat{x}_i^{k-1})$为外推点处的梯度．

Algorithm 1 分块坐标下降法

1: **初始化**：选择两组初始点$(x_1^{-1}, x_2^{-1}, \cdots, x_s^{-1}) = (x_1^0, x_2^0, \cdots, x_s^0)$.
2: **for** $k = 1, 2, \cdots$ **do**
3: **for** $i = 1, 2, \cdots$ **do**
4:   使用格式(1) 或(2) 或(3) 更新x_i^k.
5: **end for**
6: **if** 满足停机条件 **then**
7:   返回$(x_1^k, x_2^k, \cdots, x_s^k)$，算法终止.
8: **end if**
9: **end for**

- 三种格式都有其适用的问题，特别是子问题是否可写出显式解
- 在每一步更新中，三种迭代格式对不同自变量块可以混合使用，不必仅仅局限于一种.

算法格式

- BCD算法的子问题可采用三种不同的更新格式，这三种格式可能会产生不同的迭代序列，可能会收敛到不同的解，坐标下降算法的数值表现也不相同．

- 格式(1)是最直接的更新方式，它严格保证了整个迭代过程的目标函数值是下降的．然而由于f的形式复杂，子问题求解难度较大．在收敛性方面，格式(1)在强凸问题上可保证目标函数收敛到极小值，但在非凸问题上不一定收敛．

- 格式(2) (3) 则是对格式(1)的修正，不保证迭代过程目标函数的单调性，但可以改善收敛性结果．使用格式(2)可使得算法收敛性在函数F为非严格凸时有所改善．

- 格式(3)实质上为目标函数的一阶泰勒展开近似，在一些测试问题上有更好的表现，可能的原因是使用一阶近似可以避开一些局部极小值点．此外，格式(3)的计算量很小，比较容易实现．

例子：二元二次函数

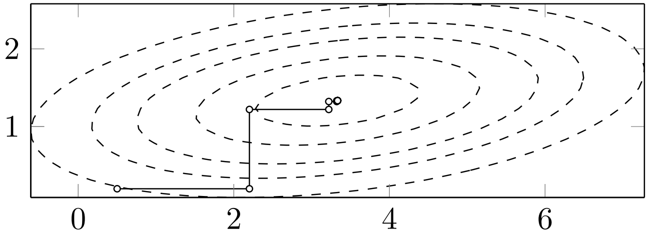
考虑二元二次函数的优化问题

$$\min \quad f(x,y) = x^2 - 2xy + 10y^2 - 4x - 20y.$$

故采用格式(1)的分块坐标下降法为

$$x^{k+1} = 2 + y^k, \quad y^{k+1} = 1 + \frac{x^{k+1}}{10}.$$

下图描绘了当初始点为 $(x,y) = (0.5, 0.2)$时的迭代点轨迹，可以看到在进行了7 次迭代后迭代点与最优解已充分接近.



不收敛反例

值得注意的是，对于非凸函数$f(x)$，分块坐标下降法可能失效．Powell 在1973 年就给出了一个使用格式(1)但不收敛的例子:

$$F(x_1,x_2,x_3) = -x_1x_2 - x_2x_3 - x_3x_1 + \sum_{i=1}^{3}[(x_i-1)_+^2 + (-x_i-1)_+^2],$$

其中$(x_i-1)_+^2$的含义为先对(x_i-1)取正部再平方．设$\varepsilon > 0$，初始点取为

$$x^0 = \left(-1-\varepsilon, 1+\frac{\varepsilon}{2}, -1-\frac{\varepsilon}{4}\right),$$

容易验证迭代序列满足

$$x^k = (-1)^k \cdot (-1,1,-1) + (-\frac{1}{8})^k \cdot \left(-\varepsilon, \frac{\varepsilon}{2}, -\frac{\varepsilon}{4}\right),$$

这个迭代序列有两个聚点$(-1,1,-1)$与$(1,-1,1)$，但这两个点都不是F的稳定点．

提纲

1 分块坐标下降法

2 应用举例

3 收敛性分析

LASSO 问题求解

下面介绍如何使用分块坐标下降法来求解LASSO 问题

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

将自变量x记为$x = \left[x_i, \bar{x}_i^\top\right]^\top$, 其中$\bar{x}_i$为$x$去掉第$i$个分量而形成的列向量. 而相应地，矩阵$A$在第$i$块的更新记为$A = \left[a_i \quad \bar{A}_i\right]$, 其中$\bar{A}_i$为矩阵$A$去掉第$i$列而形成的矩阵.
在第i块的更新中考虑格式(1)。做替换$c_i = b - \bar{A}_i\bar{x}_i$，原问题等价于

$$\min_{x_i} \quad f_i(x_i) \xlongequal{\text{def}} \mu|x_i| + \frac{1}{2}\|a_i\|^2 x_i^2 - a_i^{\mathrm{T}} c_i x_i.$$

可直接写出它的最小值点

$$x_i^k = \operatorname*{argmin}_{x_i} f_i(x_i) = \begin{cases} \frac{a_i^{\mathrm{T}} c_i - \mu_i}{\|a_i\|^2}, & a_i^{\mathrm{T}} c_i > \mu, \\ \frac{a_i^{\mathrm{T}} c_i + \mu_i}{\|a_i\|^2}, & a_i^{\mathrm{T}} c_i < -\mu, \\ 0, & \text{其他}. \end{cases}$$

- 当固定H时，设Φ的每一行为ϕ_i^{T}，那么根据矩阵分块乘法，

$$A - \Phi H = \begin{bmatrix} a_1^{\mathrm{T}} \\ a_2^{\mathrm{T}} \\ \vdots \\ a_n^{\mathrm{T}} \end{bmatrix} - \begin{bmatrix} \phi_1^{\mathrm{T}} \\ \phi_2^{\mathrm{T}} \\ \vdots \\ \phi_n^{\mathrm{T}} \end{bmatrix} H = \begin{bmatrix} a_1^{\mathrm{T}} - \phi_1^{\mathrm{T}} H \\ a_2^{\mathrm{T}} - \phi_2^{\mathrm{T}} H \\ \vdots \\ a_n^{\mathrm{T}} - \phi_n^{\mathrm{T}} H \end{bmatrix}.$$

注意到ϕ_i只有一个分量为1，其余分量为0，不妨设其第j个分量为1，此时$\phi_i^{\mathrm{T}} H$ 相当于将H 的第j行取出，因此$\|a_i^{\mathrm{T}} - \phi_i^{\mathrm{T}} H\|$为$a_i^{\mathrm{T}}$与$H$的第$j$个行向量的距离．我们的最终目的是极小化$\|A - \Phi H\|_F^2$，所以$j$应该选矩阵$H$中距离$a_i^{\mathrm{T}}$最近的那一行，即

$$\Phi_{ij} = \begin{cases} 1, & j = \underset{l}{\operatorname{argmin}} \|a_i - h_l\|, \\ 0, & \text{其他}. \end{cases}$$

其中h_l^{T}表示矩阵H的第l行．

K-均值聚类算法

- 当固定Φ时，此时考虑H的每一行h_j^{T}，根据目标函数的等价性有

$$\|A-\Phi H\|_F^2=\sum_{j=1}^{k}\sum_{a\in S_j}\|a-h_j\|^2,$$

因此只需要对每个h_j求最小即可．设$\bar{a}_j$是目前第j类所有点的均值，则

$$\begin{aligned}\sum_{a\in S_j}\|a-h_j\|^2&=\sum_{a\in S_j}\|a-\bar{a}_j+\bar{a}_j-h_j\|^2\\&=\sum_{a\in S_j}\left(\|a-\bar{a}_j\|^2+\|\bar{a}_j-h_j\|^2+2\left\langle a-\bar{a}_j,\bar{a}_j-h_j\right\rangle\right)\\&=\sum_{a\in S_j}\left(\|a-\bar{a}_j\|^2+\|\bar{a}_j-h_j\|^2\right),\end{aligned}$$

这里利用了交叉项$\sum_{a\in S_j}\left\langle a-\bar{a}_j,\bar{a}_j-h_j\right\rangle=0$的事实．因此容易看出，此时$h_j$直接取为$\bar{a}_j$即可达到最小值

非负矩阵分解

考虑最基本的非负矩阵分解问题

$$\min_{X,Y\geq 0} \quad f(X,Y)=\frac{1}{2}\|XY-M\|_F^2.$$

可以计算梯度

$$\frac{\partial f}{\partial X}=(XY-M)Y^{\mathrm{T}}, \quad \frac{\partial f}{\partial Y}=X^{\mathrm{T}}(XY-M).$$

注意到在格式(3)中，当$r_i(X)$为凸集示性函数时即是求解到该集合的投影，因此得到分块坐标下降法如下：

$$\begin{aligned} X^{k+1} &= \max\{X^k - t_k^x(X^kY^k-M)(Y^k)^{\mathrm{T}}, 0\}, \\ Y^{k+1} &= \max\{Y^k - t_k^y(X^k)^{\mathrm{T}}(X^kY^k-M), 0\}, \end{aligned}$$

其中t_k^x, t_k^y是步长，

$$\min_{D,X} \quad \frac{1}{2n}\|DX - A\|_F^2 + \lambda\|X\|_1 + \frac{\mu}{2}\|D\|_F^2.$$

- 当固定变量D时，考虑函数

$$f_D(X) = \frac{1}{2n}\|DX - A\|_F^2 + \lambda\|X\|_1.$$

使用格式(3)．通过直接计算可得$f_D(X)$中光滑部分的梯度为

$$G = \frac{1}{n}D^{\mathrm{T}}(DX - A),$$

因此格式(3)等价于

$$X^{k+1} = \mathrm{prox}_{t_k\lambda\|\cdot\|_1}\left(X^k - \frac{t_k}{n}(D^k)^{\mathrm{T}}(D^kX^k - A)\right),$$

其中t_k为步长．

$$\min_{D,X} \quad \frac{1}{2n}\|DX-A\|_F^2+\lambda\|X\|_1+\frac{\mu}{2}\|D\|_F^2.$$

- 当固定变量X时，考虑函数

$$f_X(D)=\frac{1}{2n}\|DX-A\|_F^2+\frac{\mu}{2}\|D\|_F^2.$$

使用格式(1)．计算关于D^{T}的梯度为

$$\nabla_{D^{\mathrm{T}}} f_X(D)=\frac{1}{n}X(X^{\mathrm{T}}D^{\mathrm{T}}-A^{\mathrm{T}})+\mu D^{\mathrm{T}},$$

令梯度为零向量，可得

$$D=AX^{\mathrm{T}}(XX^{\mathrm{T}}+n\mu I)^{-1}.$$

因为$X\in\mathbb{R}^{k\times n}$，其中$k\ll n$，所以$XX^{\mathrm{T}}$是一个比较小的矩阵，可以方便地求出它的逆．故格式(1)等价于

$$D^{k+1}=A(X^{k+1})^{\mathrm{T}}(X^{k+1}(X^{k+1})^{\mathrm{T}}+n\mu I)^{-1}.$$

最大割问题

$$
\begin{aligned}
(\text{半定松弛}) \quad \min \quad & \langle C, X\rangle, \\
\text{s.t.} \quad & X_{ii} = 1,\ i = 1, 2, \cdots, n, \\
& X \succeq 0. \\
(\text{非凸松弛}) \quad \min \quad & \left\langle C, V^{\mathrm{T}} V\right\rangle, \\
\text{s.t.} \quad & v_i \in \mathbb{R}^p,\ \|v_i\| = 1,\ i = 1, 2, \cdots, n, \\
& V = [v_1, v_2, \cdots, v_n].
\end{aligned}
$$

- 比较两种松弛方式可知，非凸松弛通过引入分解$X = V^{\mathrm{T}} V$并限制V的每一列的ℓ_2范数为1，将半定松弛中的X对角线元素为1以及X半正定的约束消去了.
- 这两个问题一般不等价，当p充分大时二者等价. 实际计算中通常选取一个较小的p.

最大割问题的非凸松弛

矩阵V是按列分成n块的，考虑格式(1)为例，取定i，固定其余v_j

$$\mathrm{Tr}\left(\begin{bmatrix} C_{11} & \cdots & C_{1i} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{i1} & \cdots & C_{ii} & \cdots & C_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{ni} & \cdots & C_{nn} \end{bmatrix}\begin{bmatrix} v_1^\mathrm{T}v_1 & \cdots & v_1^\mathrm{T}v_i & \cdots & v_1^\mathrm{T}v_n \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_i^\mathrm{T}v_1 & \cdots & v_i^\mathrm{T}v_i & \cdots & v_i^\mathrm{T}v_n \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_n^\mathrm{T}v_1 & \cdots & v_n^\mathrm{T}v_i & \cdots & v_n^\mathrm{T}v_n \end{bmatrix}\right),$$

根据以上矩阵分块示意图可知和v_i有关的部分为

$$C_{ii}v_i^\mathrm{T}v_i + \sum_{j\neq i}(C_{ij}+C_{ji})v_i^\mathrm{T}v_j.$$

由于约束$\|v_i\|=1$，上式中第一项是常数。最终在第i步子问题是：

$$\min f_i(v_i) = \left(\sum_{j\neq i} C_{ji}v_j^\mathrm{T}\right)v_i, \mathrm{s.t.} \|v_i\| = 1.$$

其解为：$v_i = -\left(\sum_{j\neq i} C_{ji}v_j\right) \Big/ \left\|\sum_{j\neq i} C_{ji}v_j\right\|.$

提纲

交替线性化方法

- 我们对格式(3) 在$s=2$且非凸的情况下进行收敛性分析. 定义:

$$\min \quad \Psi(x,y) \overset{\text{def}}{=} f(x)+g(y)+H(x,y), \quad (x,y)\in\mathbb{R}^n\times\mathbb{R}^m$$

  其中f 和g 为适当闭函数,H 为其定义域上的连续可微函数.

- 对该问题,格式化为如下基本形式:

$$\begin{aligned} x^{k+1} &\in \operatorname{prox}_{c_k f}\left(x^k - c_k\nabla_x H\left(x^k, y^k\right)\right) \\ y^{k+1} &\in \operatorname{prox}_{d_k g}\left(y^k - d_k\nabla_y H\left(x^{k+1}, y^k\right)\right) \end{aligned}$$

  其中c_k, d_k 为步长参数. 由于f 和g 不是凸函数,相应地prox_f 和prox_g 是集合函数,在迭代过程中只要求x_{k+1} 和y_{k+1}是相应集合中的一个元素即可. 由于自变量只有两块,对光滑部分H 我们采用的是线性化处理,因此该格式又称为近似点交替线性化方法.

- 为了保证prox_f 和prox_g 是良定义的,还需要对f 和g 提出下界有限的假设.

(适当闭函数的邻近算子) 设 h 是适当闭函数(可以非凸)，且具有有限的下界，即满足$\inf_{x\in\mathbf{dom}h} h(x) > -\infty$, 定义 h 的邻近算子为

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

定理

设h 是适当闭函数且$\inf_{x\in\mathbf{dom}h} h(x) > -\infty$, 则$\forall x \in \mathbf{dom}\ h$, $\mathrm{prox}_h(x)$ 是$\mathbb{R}^n$ 上的非空紧集.

Proof.

定义 $g(u) = h(u) + \frac{1}{2}\|u-x\|^2$, 设 $\inf_{x\in\mathbf{dom}h} h(x) = l$.

取$u_0 \in \mathbf{dom}\ h$, 由于$\frac{1}{2}\|u-x\|^2$ 无上界，故$\exists R > 0$, 对$\forall$ 满足$\|u-x\| > R$ 的u, 成立$\frac{1}{2}\|u-x\|^2 > g(u_0) - l$，即$g(u) > g(u_0)$.

这说明下水平集$\{u \mid g(u) \leqslant g(u_0)\}$ 含于球$\|u-x\| \leqslant R$ 内, 即g 有一个非空有界下水平集. 显然$g(u)$ 是闭函数, 由Weierstrass 定理可知, $g(u)$ 的最小值点集合$\mathrm{prox}_h(x)$ 是非空紧集. $\square$

非光滑非凸问题函数的次微分

前面介绍了闭凸函数的邻近算子与次梯度的关系，而对于非凸函数有类似的结论。首先回顾一下非光滑非凸函数的次微分。

次微分

设$f:\mathbb{R}^n \to (-\infty,+\infty]$ 是适当下半连续函数.

- 对给定的$x \in \mathbf{dom}\, f$, 满足如下条件的所有向量$u \in \mathbb{R}^n$ 的集合定义为f 在点x 处的*Fréchet* 次微分:

$$\liminf_{y \to x, y \neq x} \frac{f(y) - f(x) - \langle u, y-x \rangle}{\|y-x\|} \geq 0,$$

  记为$\hat{\partial} f(x)$. 当$x \notin \mathbf{dom}\, f$ 时, 将$\hat{\partial} f(x)$ 定义为空集$\varnothing$.

- f 在点$x \in \mathbb{R}^n$ 处的极限次微分(或简称为次微分)定义为

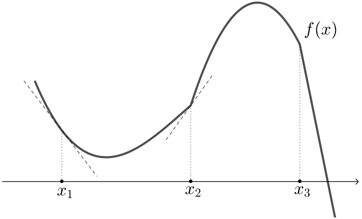
$$\partial f(x) = \{u \in \mathbb{R}^n \ : \ \exists\, x^k \to x, f(x^k) \to f(x), u^k \in \hat{\partial} f(x^k) \to u\}.$$

  极限次微分通过对x 附近的点处的*Fréchet* 次微分取极限得到.

- $\hat{\partial} f(x) \subseteq \partial f(x)$, 前者是闭凸集, 后者是闭集.并非在所有的$x \in \mathbf{dom}\, f$处都存在*Fréchet* 次微分.
- 凸函数的次梯度要求不等式

$$f(y) \geq f(x) + \langle g, y-x \rangle, \quad g \in \partial f(x)$$

在定义域内全局成立, 而非凸函数只要求在极限意义下成立.
- 当f 是可微函数时, *Fréchet* 次微分和次微分都退化成梯度.



如图，$f(x)$ 在 x_3 处不存在Fréchet 次微分, 但存在次微分

定理

设h 是适当闭函数(可非凸)且有下界, $u \in \mathrm{prox}_h(x)$, 则$x-u \in \partial h(u)$

假设条件

(1) $f : \mathbb{R}^n \to (-\infty, +\infty]$，$g : \mathbb{R}^m \to (-\infty, +\infty]$ 均为适当下半连续函数，$\inf_{\mathbb{R}^n \times \mathbb{R}^m} \Psi > -\infty$, $\inf_{\mathbb{R}^n} f > -\infty$，以及$\inf_{\mathbb{R}^m} g > -\infty$

(2) $H : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$是连续可微函数，且$\nabla H$在有界集上是联合利普希茨连续的. 即对于任意的$B_1 \times B_2 \subset \mathbb{R}^n \times \mathbb{R}^m$，存在$L > 0$使得对于任意的$(x_i, y_i) \in B_1 \times B_2, i = 1, 2$有

$$
\begin{aligned}
&\left\| \left(\nabla_x H(x_1, y_1) - \nabla_x H(x_2, y_2), \nabla_y H(x_1, y_1) - \nabla_y H(x_2, y_2) \right) \right\| \\
\leq &L \| (x_1 - x_2, y_1 - y_2) \|.
\end{aligned}
$$

假设条件

- 根据假设，在有界集上H 关于每个分量都是梯度L-利普希茨连续的，且参数与另一分量无关．即

$$\|\nabla_x H(x_1, y) - \nabla_x H(x_2, y)\| \leqslant L\|x_1 - x_2\|,$$
$$\|\nabla_y H(x, y_1) - \nabla_y H(x, y_2)\| \leqslant L\|y_1 - y_2\|.$$

- 可以直接写出$\Psi(x, y)$ 的次微分：

$$\partial\Psi(x, y) = (\nabla_x H(x, y) + \partial f(x), \nabla_y H(x, y) + \partial g(y))$$

其中“+”表示为集合间的加法.

证明梗概

- 充分下降：找到一个正常数ρ_1 使得

$$\rho_1\|z^{k+1}-z^k\|^2 \le \Psi(z^k)-\Psi(z^{k+1})$$

- 次梯度上界：假设算法产生的迭代序列有界，找到另一个常数ρ_2，使得次梯度有一个上界估计：

$$\|w^{k+1}\| \le \rho_2\|z^{k+1}-z^k\|, \quad w^k \in \partial\Psi(z^k)$$

- 利用KL 性质证明全序列收敛：假设Ψ 是一个KL函数，证明迭代序列$\{z^k\}_{k\in N}$是一个柯西列.

注：前两个步骤是证明多数算法的基本步骤，当这两个性质成立时，对任意的算法产生的迭代序列的聚点集合都为非空连通紧集，且这些聚点都是Ψ 的临界点.

近似点交替线性化方法下降量

设$h: \mathbb{R}^d \to \mathbb{R}$是连续可微函数，梯度$\nabla h$是利普希茨连续的，相应的常数为$L_h$，$\sigma : \mathbb{R}^d \to (-\infty, +\infty]$是适当下半连续函数且$\inf_{\mathbb{R}^d} \sigma > -\infty$. 固定$t < \frac{1}{L_h}$，则对任意的$u \in \mathbf{dom}\, \sigma$和$\tilde{u} \in \mathrm{prox}_{t\sigma}(u - t\nabla h(u))$，有

$$h(\tilde{u}) + \sigma(\tilde{u}) \le h(u) + \sigma(u) - \frac{1}{2}\left(\frac{1}{t} - L_h\right)\|\tilde{u} - u\|^2.$$

证明：首先根据σ的假设，$\tilde{u}$是良定义的. 根据$\tilde{u}$的最优性，有

$$\langle \tilde{u} - u, \nabla h(u)\rangle + \frac{1}{2t}\|\tilde{u} - u\|^2 + \sigma(\tilde{u}) \le \sigma(u).$$

再结合二次上界，有

$$\begin{aligned} h(\tilde{u}) + \sigma(\tilde{u}) &\le h(u) + \langle \tilde{u} - u, \nabla h(u)\rangle + \frac{L_h}{2}\|\tilde{u} - u\|^2 + \sigma(\tilde{u}) \\ &\le h(u) + \frac{L_h}{2}\|\tilde{u} - u\|^2 + \sigma(u) - \frac{1}{2t}\|\tilde{u} - u\|^2 \\ &= h(u) + \sigma(u) - \frac{1}{2}\left(\frac{1}{t} - L_h\right)\|\tilde{u} - u\|^2. \end{aligned}$$

充分下降定理

在假设条件下，设$\{z^k\} = \{(x^k, y^k)\}$为迭代格式产生的迭代序列，且假设z^k有界．取步长$c_k = d_k = \frac{1}{\gamma L}$，其中$\gamma > 1$是常数，$L$为$\nabla H$的利普希茨系数，则以下结论成立：

(1) 迭代点处的函数值序列$\{\Psi(z^k)\}$是单调下降的，且

$$\frac{\rho_1}{2}\|z^{k+1} - z^k\|^2 \le \Psi(z^k) - \Psi(z^{k+1}), \quad \forall\, k \ge 0,$$

其中$\rho_1 = (\gamma - 1)L$；

(2) 序列$\{\|z^{k+1} - z^k\|\}_{k=1}^{\infty}$平方可和，即

$$\sum_{k=1}^{\infty} (\|x^{k+1} - x^k\|^2 + \|y^{k+1} - y^k\|^2) = \sum_{k=1}^{\infty} \|z^{k+1} - z^k\|^2 < +\infty,$$

并由此推出$\lim_{k\to\infty} \|z^{k+1} - z^k\| = 0$．

证明

(1) 根据假设条件的(2)，$H(x,y)$关于每个分量都是利普希茨连续的，由第30页的结论可得到每一步关于x^k和y^k的下降量估计：

$$
\begin{aligned}
& H(x^{k+1}, y^k) + f(x^{k+1}) \\
\le\; & H(x^k, y^k) + f(x^k) - \frac{1}{2}\left(\frac{1}{c_k} - L\right)\|x^{k+1} - x^k\|^2 \\
=\; & H(x^k, y^k) + f(x^k) - \frac{1}{2}(\gamma - 1)L\|x^{k+1} - x^k\|^2,
\end{aligned}
$$

以及

$$
\begin{aligned}
& H(x^{k+1}, y^{k+1}) + g(y^{k+1}) \\
\le\; & H(x^{k+1}, y^k) + g(y^k) - \frac{1}{2}\left(\frac{1}{d_k} - L\right)\|y^{k+1} - y^k\|^2 \\
=\; & H(x^{k+1}, y^k) + g(y^k) - \frac{1}{2}(\gamma - 1)L\|y^{k+1} - y^k\|^2.
\end{aligned}
$$

证明

将上述两个不等式相加，消去$H(x^{k+1},y^k)$，得到

$$
\begin{aligned}
&\Psi(z^k)-\Psi(z^{k+1})\\
=&H(x^k,y^k)+f(x^k)+g(y^k)-H(x^{k+1},y^{k+1})-f(x^{k+1})-g(y^{k+1})\\
\geq&\frac{1}{2}(\gamma-1)L\left(\|x^{k+1}-x^k\|^2+\|y^{k+1}-y^k\|^2\right).
\end{aligned}
$$

由此立即可得

$$\frac{\rho_1}{2}\|z^{k+1}-z^k\|^2\leq\Psi(z^k)-\Psi(z^{k+1}). \tag{5}$$

此外，容易得知迭代点处的函数值$\{\Psi(z^k)\}$关于k是单调递减的．根据假设$\inf\Psi>-\infty$ 可知$\Psi(z^k)$ 单调下降收敛到一个有限的数Ψ^*．

(2) 设N为任意的整数，在(5) 式中对k求和，得

$$\sum_{k=0}^{N-1}\|z^{k+1}-z^k\|^2 \leq \frac{2}{\rho_1}(\Psi(z^0)-\Psi(z^N)) \leq \frac{2}{\rho_1}(\Psi(z^0)-\Psi^*).$$

令$N\to\infty$即可得$\sum_{k=0}^{\infty}\|z^{k+1}-z^k\|^2<+\infty$，从而

$$\lim_{k\to\infty}\|z^{k+1}-z^k\|=0$$

注：定理表明进行一轮近似点交替线性化迭代后，函数值下降量的下界可被相邻迭代点之间的距离控制．几乎所有下降类的算法在一定条件下都满足这个性质．到此我们完成了收敛性分析的第一个步骤．

次梯度上界

- 在上一步中我们证明了迭代点处的函数值Ψ^k最终会收敛到某个值
- 但是这个值和局部最优解的关系还没有明确说明
- 序列$\{z^k\}$的收敛性质在上面的定理中也没有体现
- 在这一部分我们将讨论序列$\{z^k\}$是否会趋于某个临界点，这是收敛性框架中的第二个步骤

次梯度上界

在假设条件下，设$\{z^k\}$是迭代格式产生的有界序列，对任意的整数k，定义

$$A_x^k = \frac{1}{c_{k-1}}(x^{k-1} - x^k) + \nabla_x H(x^k, y^k) - \nabla_x H(x^{k-1}, y^{k-1}),$$

以及

$$A_y^k = \frac{1}{d_{k-1}}(y^{k-1} - y^k) + \nabla_y H(x^k, y^k) - \nabla_y H(x^k, y^{k-1}).$$

则有$(A_x^k, A_y^k) \in \partial \Psi(x^k, y^k)$ 且

$$\|(A_x^k, A_y^k)\| \leq \|A_x^k\| + \|A_y^k\| \leq \rho_2 \|z^k - z^{k-1}\|,$$

其中$\rho_2 = (2\gamma + 3)L$.

证明

由迭代格式中更新x^k的一阶最优性条件可知

$$\nabla_x H(x^{k-1}, y^{k-1}) + \frac{1}{c_{k-1}}(x^k - x^{k-1}) + u^k = 0,$$

其中$u^k \in \partial f(x^k)$为f的一个次梯度．因此我们有

$$\nabla_x H(x^{k-1}, y^{k-1}) + u^k = \frac{1}{c_{k-1}}(x^{k-1} - x^k).$$

同理，由迭代格式中关于y^k的更新可知

$$\nabla_y H(x^k, y^{k-1}) + v^k = \frac{1}{d_{k-1}}(y^{k-1} - y^k),$$

其中$v^k \in \partial g(y^k)$为g的一个次梯度．由A_x^k, A_y^k的定义和$\partial\Psi$的表达式可得

$$A_x^k = \nabla_x H(x^k, y^k) + u^k \in \partial_x \Psi(x^k, y^k),$$
$$A_y^k = \nabla_y H(x^k, y^k) + v^k \in \partial_y \Psi(x^k, y^k).$$

即有$(A_x^k, A_y^k) \in \partial\Psi(x^k, y^k)$，我们需要证明的第一个结论因此成立．

证明

下面估计A_x^k和A_y^k的模长．这里需要借助假设的(2)，即∇H在有界集上关于(x, y)是联合利普希茨连续的．因此对$\|A_x^k\|$我们有

$$
\begin{aligned}
\|A_x^k\| &\leq \frac{1}{c_{k-1}}\|x^{k-1} - x^k\| \ + \|\nabla_x H(x^k, y^k) - \nabla_x H(x^{k-1}, y^{k-1})\| \\
&\leq \frac{1}{c_{k-1}}\|x^{k-1} - x^k\| + L\left(\|x^{k-1} - x^k\| + \|y^{k-1} - y^k\|\right) \\
&= \left(L + \frac{1}{c_{k-1}}\right)\|x^{k-1} - x^k\| + L\|y^{k-1} - y^k\| \\
&= (\gamma + 1)L\|x^{k-1} - x^k\| + L\|y^{k-1} - y^k\| \\
&\leq (\gamma + 2)L\|z^{k-1} - z^k\|.
\end{aligned}
$$

其中，第二个不等式是根据∇H的利普希茨连续性，最后一个不等式是将$\|x^{k-1} - x^k\|$和$\|y^{k-1} - y^k\|$ 统一放大为$\|z^{k-1} - z^k\|$．

证明

另一方面，对$\|A_y^k\|$的估计只需要用到∇H关于y的利普希茨连续性：

$$\begin{aligned}
\|A_y^k\| &\le \frac{1}{d_{k-1}}\|y^k - y^{k-1}\| + \|\nabla_y H(x^k, y^k) - \nabla_y H(x^k, y^{k-1})\| \\
&\le \frac{1}{d_{k-1}}\|y^k - y^{k-1}\| + L\|y^k - y^{k-1}\| \\
&= \left(\frac{1}{d_{k-1}} + L\right)\|y^k - y^{k-1}\| \\
&\le (\gamma + 1)L\|z^k - z^{k-1}\|.
\end{aligned}$$

结合这两个估计我们最终得到

$$\|(A_x^k, A_y^k)\| \le \|A_x^k\| + \|A_y^k\| \le (2\gamma + 3)L\|z^k - z^{k-1}\| = \rho_2\|z^k - z^{k-1}\|.$$

子列收敛性

- 上面的分析表明，$\partial\Psi(z^k)$将会包含一个模长不断趋于0的向量，这暗示着某种收敛性．由于有界序列$\{z^k\}$一定有收敛的子列，因此猜想$\{z^k\}$的极限点应该和Ψ的临界点有一定的关系．我们有：
- 定义$\omega(z^0)$为近似点交替线性化方法从点z^0出发产生迭代序列的所有极限点集，且$\{z^k\}$是有界序列，则以下结论成立：
  (1) $\emptyset \neq \omega(z^0) \subset \mathrm{crit}\ \Psi$，其中$\mathrm{crit}\ \Psi$定义为$\Psi$所有的临界点；
  (2) z^k与集合$\omega(z^0)$的距离趋于0，即
  $$\lim_{k\to\infty} \mathrm{dist}(z^k, \omega(z^0)) = 0;$$
  (3) $\omega(z^0)$ 是非空的连通紧集；
  (4) Ψ在$\omega(z^0)$上是一个有限的常数．

子列收敛性

- 上面的结论表明从点z^0出发产生的点列$\{z^k\}$的极限点都是Ψ的临界点（次梯度集含有零向量）.
- 至此我们已经得到了迭代序列$\{z^k\}$的子列收敛性，这至少保证了算法在迭代过程中与临界点越来越接近.
- 一个自然的问题就是：$\{z^k\}$全序列在何种条件下收敛?
- 这就要进入理论分析的第三个步骤：利用函数的KL 性质.

KL 性质

- 定义Φ_η是凹连续函数$\varphi:[0,\eta)\to\mathbb{R}_+$的集合且满足如下条件：(i) $\varphi(0)=0$；(ii) φ 在$(0,\eta)$ 内连续可微，在点0处连续；(iii) 对任意的$s\in(0,\eta)$，都有$\varphi'(s)>0$.
- 设$\sigma:\mathbb{R}^d\to(-\infty,+\infty]$是适当下半连续函数.

  (1) 称函数σ在给定点$\bar{u}\in\textbf{dom}\,\partial\sigma\overset{\text{def}}{=\!=}\{u\mid\partial\sigma(u)\neq\emptyset\}$处具有KL 性质，若存在$\eta\in(0,+\infty]$和$\bar{u}$的一个邻域$U$以及函数$\varphi\in\Phi_\eta$，使得

  $$\forall\, u\in U\cap[\sigma(\bar{u})<\sigma<\sigma(\bar{u})+\eta],$$

  以下不等式成立：

  $$\varphi'(\sigma(u)-\sigma(\bar{u}))\cdot\operatorname{dist}(0,\partial\sigma(u))\geq 1,$$

  其中$\operatorname{dist}(x,S)$表示点x到集合S的距离.

  (2) 若σ在$\textbf{dom}\,\partial\sigma$上处处满足KL 性质，则称$\sigma$是一个KL 函数.

KL性质的解释

- 一大类函数都具有KL 性质，该性质刻画了函数本身在给定点$\bar{u}$处的某种行为．
- 如果点$\bar{u}$不是函数σ的临界点，那么KL 性质在点$\bar{u}$处自然成立．因此KL 性质成立的不平凡情形是$\bar{u}$是σ的临界点，即$0 \in \partial\sigma(\bar{u})$．
- 这种情况下KL 性质保证了"函数σ可被锐化"．直观上来说，令

$$\tilde{\varphi}(u) = \varphi(\sigma(u) - \sigma(\bar{u})),$$

KL 性质在某种条件下可以改写成

$$\mathrm{dist}(0, \partial\tilde{\varphi}(u)) \geq 1,$$

其中u的取法需要保证$\sigma(u) > \sigma(\bar{u})$．

- 以上性质表明，无论u多么接近临界点$\bar{u}$，$\tilde{\varphi}(u)$ 的次梯度的模长均大于1．所以KL 性质也被称为是函数σ 在**重参数化子**φ下的一个**锐化**，这种几何性质在分析一阶算法的收敛性时起到关键作用．

半代数, 次分析以及对数指数函数是KL函数

- $\mathbb{R}^d$ 的子集S 是一个半代数集，如果存在有限个实多项式函数$g_{ij}, h_{ij} : \mathbb{R}^d \to \mathbb{R}$ 使得

$$S = \cup_{j=1}^{p} \cap_{i=1}^{q} \{u \in \mathbb{R}^d : g_{ij}(u) = 0, h_{ij}(u) < 0\}$$

- 函数$h : \mathbb{R}^d \to (-\infty, +\infty]$ 称为半代数的，如果它的图$\{(u, t) \in \mathbb{R}^{d+1} : h(u) = t\}$ 是$\mathbb{R}^{d+1}$ 的半代数子集
- 设$\sigma(u) : \mathbb{R}^d \to (-\infty, +\infty)$ 是下半连续的恰当函数. 若σ 是半代数的，则它在$\mathbf{dom}\ \sigma$中任一点处满足KL性质.

例子：

- 实多项式函数.
- 半代数集的指示函数.
- 半代数函数的有限和与有限乘积.
- 半代数函数的复合.
- 上极限/下极限类函数. 例如，当g是半代数函数并且C 是半代数集时，$\sup\{g(u,v) : v \in C\}$ 是半代数的.
- 半正定矩阵锥，Stiefel流形以及恒秩矩阵都是半代数集.
- S 是$\mathbb{R}^d$ 中的非空半代数子集，则函数$x \to \mathrm{dist}(x,S)^2$ 是半代数的.
- $\|\cdot\|_0, \|\cdot\|_p$是半代数函数，其中p 是有理数.

一致KL性质

由于非凸问题有多个临界点，有时单个点$\bar{u}$处的KL 性质是不够的，我们需要引入一致KL 性质：

- 设Ω是紧集，$\sigma : \mathbb{R}^d \to (-\infty, +\infty]$是适当下半连续函数，在$\Omega$上为常数且在$\Omega$的每个点处都满足KL 性质，则存在$\varepsilon > 0, \eta > 0, \varphi \in \Phi_\eta$使得对任意$\bar{u} \in \Omega$和所有满足以下条件的$u$：

$$\{u \in \mathbb{R}^d \ : \ \mathrm{dist}(u, \Omega) < \varepsilon\} \cap [\sigma(\bar{u}) < \sigma < \sigma(\bar{u}) + \eta],$$

有

$$\varphi'(\sigma(u) - \sigma(\bar{u}))\mathrm{dist}(0, \partial\sigma(u)) \geq 1.$$

证明

- 因为$\mathbb{R}^d$上的紧集可以由有限多个开集覆盖，因此该问题可在有限个点上进行讨论．设μ是σ在Ω上的取值．由于Ω是紧集，根据有限覆盖定理，存在有限多个开球$B(u_i, \varepsilon_i)$（其中$u_i \in \Omega, i = 1, 2, \cdots, p$）使得$\Omega \subset \bigcup_{i=1}^{p} B(u_i, \varepsilon_i)$．
- 现在考虑这些点u_i．在点u_i 上KL 性质成立，设$\varphi_i : [0, \eta_i) \to \mathbb{R}_+$是对应的重参数化子，则对任意$u \in B(u_i, \varepsilon_i) \cap [\mu < \sigma < \mu + \eta_i]$，有逐点的KL 性质：

$$\varphi_i'(\sigma(u) - \mu)\mathrm{dist}(0, \partial\sigma(u)) \geq 1.$$

取充分小的$\varepsilon > 0$使得

$$U_\varepsilon \stackrel{\text{def}}{=\!=} \{u \in \mathbb{R}^d \mid \mathrm{dist}(u, \Omega) \leq \varepsilon\} \subset \bigcup_{i=1}^{p} B(u_i, \varepsilon_i).$$

证明

- 取$\eta = \min_i \eta_i$，以及

$$\varphi(s) = \int_0^s \max_i \varphi_i'(t)\mathrm{d}t, \quad s \in [0, \eta).$$

容易验证$\varphi \in \varPhi_\eta$.

- 对任意的$u \in U_\varepsilon \cap [\mu < \sigma < \mu + \eta]$，$u$必定落在某个球$B(u_{i_0}, \varepsilon_{i_0})$中，我们有

$$\begin{aligned}\varphi'(\sigma(u) - \mu)\mathrm{dist}(0, \partial\sigma(u)) &= \max_i \varphi_i'(\sigma(u) - \mu)\mathrm{dist}(0, \partial\sigma(u)) \\ &\geq \varphi_{i_0}'(\sigma(u) - \mu)\mathrm{dist}(0, \partial\sigma(u)) \geq 1.\end{aligned}$$

即一致KL 性质成立.

有限长度性质

设Ψ是KL 函数，且满足假设条件，则以下结论成立：

1. 序列$\{z^k\}$的长度有限，即

$$\sum_{k=1}^{\infty} \|z^{k+1} - z^k\| < +\infty.$$

2. 序列$\{z^k\}$收敛到Ψ的一个临界点$z^* = (x^*, y^*)$.

注:上述定理的(1) 有别于第一个步骤中充分下降定理的(2)：后者只得到了$\|z^{k+1} - z^k\|$平方可和的结论，而前者则说明从z^0出发，迭代序列的轨迹长度是有限的．这个结论显然比充分下降定理要强，也是推导全序列收敛的关键．

证明

- 由于$\{z^k\}$是有界序列，存在收敛子列$\{z^{k_q}\} \to \bar{z}, q \to \infty$. 和之前的推导类似，不管全序列$\{z^k\}$收敛性如何，对应的函数值列$\{\Psi(z^k)\}$总是收敛的，且

$$\lim_{k\to\infty} \Psi(z^k) = \Psi(\bar{z}). \tag{6}$$

以下不妨设$\Psi(\bar{z}) < \Psi(z^k)$. 这是因为若存在$\bar{k}$使得$\Psi(z^{\bar{k}}) = \Psi(\bar{z})$，由充分下降性可知$z^{\bar{k}+1} = z^{\bar{k}}$，进而有$z^k = z^{\bar{k}}, \forall\, k > \bar{k}$. 结论自然成立.

- 由极限(6) 和极限点集$\omega(z^0)$的性质$\lim_{k\to\infty} \mathrm{dist}(z^k, \omega(z^0)) = 0$ 可知对任意的$\varepsilon, \eta > 0$，存在充分大的正整数l，使得对任意的$k > l$，

$$\Psi(z^k) < \Psi(\bar{z}) + \eta, \quad \mathrm{dist}(z^k, \omega(z^0)) < \varepsilon.$$

- 以上的分析说明当k充分大时，迭代点序列最终会满足一致 KL 性质的前提. 下面就在这个结论下分别证明定理的两个结论.

证明

(1) 根据临界点的性质，$\omega(z^0)$是非空紧集，且Ψ在$\omega(z^0)$上是常数. 在一致KL性质中令$\Omega=\omega(z^0)$，对任意的$k>l$，

$$\varphi'(\Psi(z^k)-\Psi(\bar{z}))\mathrm{dist}(0,\partial\Psi(z^k))\geq 1.$$

根据第二个步骤中次梯度上界的引理可知

$$\mathrm{dist}(0,\partial\Psi(z^k))\leq\|(A_x^k,A_y^k)\|\leq\rho_2\|z^k-z^{k-1}\|.$$

代入KL 性质有

$$\varphi'(\Psi(z^k)-\Psi(\bar{z}))\geq\frac{1}{\rho_2}\|z^k-z^{k-1}\|^{-1}. \tag{7}$$

另外，由φ的凹性，有

$$\begin{aligned}&\varphi(\Psi(z^k)-\Psi(\bar{z}))-\varphi(\Psi(z^{k+1})-\Psi(\bar{z}))\\ \geq&\varphi'(\Psi(z^k)-\Psi(\bar{z}))(\Psi(z^k)-\Psi(z^{k+1})).\end{aligned} \tag{8}$$

证明

为了表示方便，定义

$$\Delta_{p,q} = \varphi(\Psi(z^p) - \Psi(\bar{z})) - \varphi(\Psi(z^q) - \Psi(\bar{z})),$$

其中p, q为任意正整数. 定义常数

$$C = \frac{2\rho_2}{\rho_1} > 0.$$

根据不等式(8)，使用(7) 式和第一个步骤中的充分下降定理分别估计不等号右边的两项，有

$$\begin{aligned}\Delta_{k,k+1} &\geq \varphi'(\Psi(z^k) - \Psi(\bar{z}))(\Psi(z^k) - \Psi(z^{k+1})) \\ &\geq \frac{1}{\rho_2}\|z^k - z^{k-1}\|^{-1} \cdot \frac{\rho_1}{2}\|z^{k+1} - z^k\|^2 \\ &= \frac{\|z^{k+1} - z^k\|^2}{C\|z^k - z^{k-1}\|},\end{aligned}$$

等价于

$$\|z^{k+1} - z^k\| \leq \sqrt{C\Delta_{k,k+1}\|z^k - z^{k-1}\|}.$$

证明

根据基本不等式$2\sqrt{ab} \leq a+b,\ \forall\, a,b>0$，我们取$a=\|z^k-z^{k-1}\|$, $b=C\Delta_{k,k+1}$，则

$$2\|z^{k+1}-z^k\| \leq \|z^k-z^{k-1}\|+C\Delta_{k,k+1}.$$

对任意的$k>l$，在上式中把k替换成i并对$i=l+1,l+2,\cdots,k$求和，得

$$\begin{aligned} 2\sum_{i=l+1}^{k}\|z^{i+1}-z^i\| &\leq \sum_{i=l+1}^{k}\|z^i-z^{i-1}\|+C\sum_{i=l+1}^{k}\Delta_{i,i+1} \\ &\leq \sum_{i=l+1}^{k}\|z^{i+1}-z^i\|+\|z^{l+1}-z^l\|+C\Delta_{l+1,k+1}. \end{aligned}$$

最后一个不等式是因为$\Delta_{p,q}+\Delta_{q,r}=\Delta_{p,r}$.

证明

注意到上式不等号右边刚好可以和左边部分抵消，我们有

$$\begin{aligned}&\sum_{i=l+1}^{k}\|z^{i+1}-z^{i}\|\\ \leq&\|z^{l+1}-z^{l}\|+C\Big(\varphi(\Psi(z^{l+1})-\Psi(\bar{z}))-\varphi(\Psi(z^{k+1})-\Psi(\bar{z}))\Big)\\ \leq&\|z^{l+1}-z^{l}\|+C\varphi(\Psi(z^{l+1})-\Psi(\bar{z})).\end{aligned}$$

不等式右边是有界的数且与k无关，由级数收敛的定义立即可得

$$\sum_{k=1}^{\infty}\|z^{k+1}-z^{k}\|<+\infty.$$

证明

(2) 在$\sum_{k=1}^{\infty}\|z^{k+1}-z^k\|<+\infty$的前提下$\{z^k\}$全序列收敛是显然的．这等价于证明$\{z^k\}$是柯西列．对任意$q>p>l$，

$$z^q-z^p=\sum_{k=p}^{q-1}(z^{k+1}-z^k),$$

根据三角不等式，

$$\|z^q-z^p\|=\left\|\sum_{k=p}^{q-1}(z^{k+1}-z^k)\right\|\le\sum_{k=p}^{q-1}\|z^{k+1}-z^k\|,$$

而$\|z^{k+1}-z^k\|$的可和性意味着$\sum_{k=l+1}^{\infty}\|z^{k+1}-z^k\|$趋于0．因此$\{z^k\}$是一个柯西列，算法产生的迭代序列有全序列收敛性.