学海无涯苦作舟

优化篇

洗衣机

Last Update: 2025 年 5 月 4 日

目录

1 概念 & 定义

定义 1.1 (凸函数). 令 $\Omega \subset \mathbb{R}^n$ 为凸集，$\theta : \Omega \to \mathbb{R}$。若对任意 $u, v \in \Omega$ 和 $\lambda \in [0,1]$，都有

$$\theta\big(\lambda u + (1-\lambda)v\big) \le \lambda\,\theta(u) + (1-\lambda)\,\theta(v),$$

则称 θ 在 Ω 上为凸函数。

定义 1.2 (严格凸函数). 在同样记号下，若对任意 $u, v \in \Omega$ 且 $u \neq v$，以及 $\lambda \in (0,1)$，严格不等式

$$\theta\big(\lambda u + (1-\lambda)v\big) < \lambda\,\theta(u) + (1-\lambda)\,\theta(v)$$

成立，则称 θ 在 Ω 上为严格凸函数。

定义 1.3 (强凸函数). 若存在常数 $\sigma > 0$，使得对任意 $u, v \in \Omega$ 和 $\lambda \in (0,1)$，

$$\theta\big(\lambda u + (1-\lambda)v\big) \le \lambda\,\theta(u) + (1-\lambda)\,\theta(v) - \frac{\sigma}{2}\,\lambda(1-\lambda)\,\|u - v\|^2,$$

则称 θ 在 Ω 上为强凸函数。

备注 1.1. 由上述定义可知：

$$\text{强凸} \implies \text{严格凸} \implies \text{凸}.$$

定义 1.4 (可微凸函数的梯度不等式). 若 θ 在 Ω 上可微，则对任意 $u, v \in \Omega$ 有

$$\theta(u) \ge \theta(v) + \nabla\theta(v)^T(u - v)$$

当且仅当 θ 为凸函数。

定义 1.5 (次梯度与次微分). 令 $\theta : \Omega \to \mathbb{R}$，$\Omega$ 凸。若对于某向量 $s \in \mathbb{R}^n$，对任意 $u \in \Omega$，都有

$$\theta(u) \ge \theta(v) + s^T(u - v),$$

则称 s 为 θ 在 $v \in \Omega$ 处的次梯度（*Subgradient*）。所有次梯度构成的集合称为 θ 在 v 处的次微分（*Subdifferential*），记为 $\partial\theta(v)$。

定义 1.6 (Lipschitz 连续). 令 $F : \Omega \to \mathbb{R}^m$。若存在常数 $L > 0$，使得对所有 $u, v \in \Omega$，

$$\|F(u) - F(v)\| \le L\,\|u - v\|,$$

则称 F 在 Ω 上为 Lipschitz 连续（*Lipschitz continuous*）。

定义 1.7 (算子类型). 令 $\Omega \subset \mathbb{R}^n$ 为凸集，$F : \Omega \to \mathbb{R}^n$ 为映射；亦令 $(\mathcal{H}, \|\cdot\|)$ 为实 *Hilbert* 空间，$T : \mathcal{H} \to \mathcal{H}$。则：

1.（单调算子 *Monotone*）若对任意 $u, v \in \Omega$ 有

$$(u-v)^\top \big[F(u) - F(v)\big] \geq 0,$$

则称 F 在 Ω 上为单调算子；

2.（严格单调算子 *Strictly Monotone*）若对任意 $u, v \in \Omega,\ u \neq v$ 有

$$(u-v)^\top \big[F(u) - F(v)\big] > 0,$$

则称 F 为严格单调算子；

3.（强单调算子 *Strongly Monotone*）若存在 $\eta > 0$ 使得对任意 $u, v \in \Omega$，

$$(u-v)^\top \big[F(u) - F(v)\big] \geq \eta \|u - v\|^2,$$

则称 F 为强单调算子；

4.（伪单调算子 *Pseudo-monotone*）若对任意 $u, v \in \Omega$，

$$(u-v)^\top F(v) \geq 0 \implies (u-v)^\top F(u) \geq 0,$$

则称 F 为伪单调算子；

5.（余强制算子 *Co-coercive*）若存在 $\mu > 0$ 使得对任意 $u, v \in \Omega$，

$$(u-v)^\top \big[F(u) - F(v)\big] \geq \mu \|F(u) - F(v)\|^2,$$

则称 F 为余强制算子；

6.（非扩张算子 *Nonexpansive*）若对任意 $x, y \in \mathcal{H}$ 有

$$\|T(x) - T(y)\| \leq \|x - y\|,$$

则称 T 为非扩张算子；

7.（严格非扩张算子 *Strictly Nonexpansive*）若存在 $\rho \in [0, 1)$ 使得对任意 $x, y \in \mathcal{H}$，

$$\|T(x) - T(y)\| \leq \rho\, \|x - y\|,$$

则称 T 为严格非扩张算子；

8.（α-平均算子 *α-Averaged*）若存在 $\alpha \in (0, 1)$ 与非扩张算子 N 使得

$$T = (1 - \alpha)\, I + \alpha\, N,$$

则称 T 为 α-平均算子。

2 分式规划 (Fractional Programming): Dinkelbach 算法

考虑如下问题

$$\min_{x\in\mathcal{S}} F(x) = \frac{f(x)}{g(x)}$$

其中 $f(x), g$ 是适当的闭凸函数, $g(x) > 0$.

Algorithm 1 Dinkelbach 算法

1: **Input:** 初始值 $\lambda^{(0)} \in \mathbb{R}$, 迭代次数 $k = 0$, 约束条件 $x \in \mathcal{S}$, 精度阈值 $\epsilon > 0$
2: **Output:** 最优解 x^* 和最优值 λ^*
3: **repeat**
4: **Step 1:** 求解子问题:

$$x^{(k)} = \arg\max_{x\in X}\{f(x) - \lambda^{(k)} g(x)\}$$

5: **Step 2:** 计算:

$$\phi(\lambda^{(k)}) = \max_{x\in X}\{f(x) - \lambda^{(k)} g(x)\}$$

6: **Step 3:** 更新:

$$\lambda^{(k+1)} = \frac{f(x^{(k)})}{g(x^{(k)})}$$

7: 更新迭代次数: $k \leftarrow k + 1$
8: **until** $\phi(\lambda^{(k)}) < \epsilon$

性质 2.1. *ϕ 关于 λ 单调递减: $\lambda_1 < \lambda_2 \Rightarrow \phi(\lambda_1) > \phi(\lambda_2)$.*

性质 2.2. *$\lambda = \lambda^* \Leftrightarrow \phi(\lambda) = 0$.*

证明. $(\Rightarrow)$: 令 $\lambda = \lambda^* = F(x^*) = \frac{f(x^*)}{g(x^*)}$. $\forall x \in \mathcal{S}, \lambda^* \le \frac{f(x)}{g(x)} \Rightarrow f(x) - \lambda^* g(x) \ge 0$, 因此 x^* 恰好取到 $\phi(\lambda^*)$ 的下界 0.

$(\Leftarrow)$: 假设存在 $\lambda' = F(x')$ 是更优解, 因此 $\lambda' = \frac{f(x')}{g(x')} < \lambda \Rightarrow f(x') - \lambda g(x') < 0$, i.e.$\phi(\lambda) < 0$, 矛盾.

□

3 酉不变性

酉矩阵保持内积不变

- 酉矩阵保持向量范数不变;
- 酉相似变换不改变矩阵的谱（特征值、奇异值）;

矩阵的 F 范数可以用奇异值刻画，因此也是酉不变的。

矩阵的诱导 2 范数等价于矩阵的谱半径，因此也是酉不变的。

4 变分不等式

要证明变分不等式问题 $\mathrm{VI}(F, \mathbb{R}_+^n)$ 与互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0$$

的等价性，我们需要分别证明两个方向的蕴含关系。

**1. 变分不等式 $\Rightarrow$ 互补问题 ** 假设 u 是 $\mathrm{VI}(F, \mathbb{R}_+^n)$ 的解，即

$$\forall v \in \mathbb{R}_+^n, \quad (v-u)^T F(u) \geq 0.$$

需要证明 u 满足互补问题的三个条件。

(1) 非负性 $u \geq 0$ 和 $F(u) \geq 0$ - 由于 $\mathbb{R}_+^n$ 是可行域，显然 $u \geq 0$。- 取 $v = u + e_i$（其中 e_i 是第 i 个单位向量），则 $v \in \mathbb{R}_+^n$，代入变分不等式得

$$(v-u)^T F(u) = e_i^T F(u) = F_i(u) \geq 0.$$

这对所有 i 成立，故 $F(u) \geq 0$。

(2) 正交性 $u^T F(u) = 0$ - 取 $v = 0$，代入变分不等式得

$$(0-u)^T F(u) = -u^T F(u) \geq 0 \quad \Rightarrow \quad u^T F(u) \leq 0.$$

- 取 $v = 2u$，代入得

$$(2u-u)^T F(u) = u^T F(u) \geq 0.$$

- 结合两者，得 $u^T F(u) = 0$。

综上，$\mathrm{VI}(F, \mathbb{R}_+^n)$ 的解 u 满足互补问题。

—

**2. 互补问题 $\Rightarrow$ 变分不等式 ** 假设 u 满足互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0,$$

需要证明 u 是 $\mathrm{VI}(F,\mathbb{R}^n_+)$ 的解。

对任意 $v \in \mathbb{R}^n_+$，有

$$(v-u)^T F(u) = v^T F(u) - u^T F(u) = v^T F(u).$$

由于 $v \geq 0$ 且 $F(u) \geq 0$，故 $v^T F(u) \geq 0$。因此

$$(v-u)^T F(u) \geq 0,$$

即 u 满足变分不等式。

—

** 结论 ** 变分不等式 $\mathrm{VI}(F,\mathbb{R}^n_+)$ 和互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0$$

在解集上完全等价。

5 凸的等价刻画

可微函数 $\theta(\cdot)$ 在非空凸集 Ω 上是凸的, 当且仅当其梯度 $\nabla\theta(\cdot)$ 是一个单调算子, 即对 $\nabla\theta(\cdot)$ 满足:

$$(u-v)^T(\nabla\theta(u)-\nabla\theta(v)) \geq 0, \quad \forall u,\ v \in \Omega.$$

证明关键在于将 $\theta(u+t(u-v))$ 视作一元函数

类似地, $\theta(\cdot)$ 是严格凸的等价于 $\nabla\theta(\cdot)$ 是严格单调的; $\theta(\cdot)$ 是强凸的等价于 $\nabla\theta$ 是强单调的. 但需要注意的是, 若一个映射 $F(\cdot)$ 是单调的, 我们并不能直接得到 $F(\cdot)$ 是某一凸函数 $\theta(\cdot)$ 的梯度; 其他情况也是类似的.

6 理论依据：梯度下降为何一定下降？

定义 6.1 (下降方向). 设 $f:\mathbb{R}^n\to\mathbb{R}$ 是定义在 $\mathbb{R}^n$ 上的连续可微函数。如果向量 $0\neq\mathbf{d}\in\mathbb{R}^n$ 在点 $\mathbf{x}$ 处的函数 f 的方向导数 $f'(\mathbf{x};\mathbf{d})$ 是负的，即

$$f'(\mathbf{x};\mathbf{d})=\nabla f(\mathbf{x})^T\mathbf{d}<0,$$

则称 $\mathbf{d}$ 是 f 在 $\mathbf{x}$ 处的下降方向。

下降方向是梯度下降的理论基础。

引理 6.1 (下降方向的下降性质). 设 f 是定义在 $\mathbb{R}^n$ 上的连续可微函数，且 $\mathbf{x}\in\mathbb{R}^n$。假设 $\mathbf{d}$ 是 f 在 $\mathbf{x}$ 处的下降方向。则存在 $\varepsilon>0$ 使得

$$f(\mathbf{x}+t\mathbf{d})<f(\mathbf{x})$$

对任意 $t\in(0,\varepsilon]$ 成立。

证明. 由于 $f'(\mathbf{x};\mathbf{d})<0$，根据方向导数的定义，有

$$\lim_{t\to0^+}\frac{f(\mathbf{x}+t\mathbf{d})-f(\mathbf{x})}{t}=f'(\mathbf{x};\mathbf{d})<0.$$

因此，存在 $\varepsilon>0$ 使得

$$\frac{f(\mathbf{x}+t\mathbf{d})-f(\mathbf{x})}{t}<0$$

对任意 $t\in(0,\varepsilon]$ 成立，这直接意味着所需的结果。 □

7 优化算法收敛性证明的两类方法

7.1 不动点类

定理 7.1 (Banach 不动点定理). 设 (X,d) 是一个完备度量空间，$T: X \to X$ 是一个压缩映射，即存在常数 $0 \le c < 1$，使得对任意 $x, y \in X$，有

$$d(T(x), T(y)) \le c \cdot d(x, y).$$

则 T 在 X 上存在唯一的不动点 $x^* \in X$，即 $T(x^*) = x^*$。此外，对任意初始点 $x_0 \in X$，迭代序列 $x_{n+1} = T(x_n)$ 将以指数速度收敛到 x^*，即

$$d(x_n, x^*) \le \frac{c^n}{1-c} d(x_1, x_0).$$

Banach 不动点定理给了我们设计算法的思路。首先我们考虑单个函数的优化问题

$$\min f(x) \tag{1}$$

这里我们假设 $f(x)$ 是一个连续可微的函数。那么上述问题的解 x^* 就满足

$$\nabla f(x^*) = 0$$

如果我们想设计一个不动点迭代算法来求解这个问题，我们可以考虑构造一个映射 $T(x)$，使得

$$T(x^*) = x^*$$

那么这个 T 就应该是

$$T = Id() - \nabla f()$$

我们期待按照算法

$$x_{k+1} = T(x_k)$$

迭代的序列 x_k 会收敛到 x^*。因此，我们需要做的就是验证 T 是一个压缩映射。我们可以通过计算

$$\begin{aligned}\|T(x) - T(y)\|^2 =& \|x - y - \nabla f(x) + \nabla f(y)\|^2 \\ \le& \|x-y\|^2 + \|\nabla f(x) - \nabla f(y)\|^2 - 2\langle x - y, \nabla f(x) - \nabla f(y)\rangle\end{aligned} \tag{2}$$

对于一般的函数 $f(x)$ 并不会使得 T 满足压缩性质。因此，我们需要对 $f(x)$ 做一些假设。比如说存在 $\mu > 0, L > 0$ 使得

$$\|\nabla f(x) - \nabla f(y)\|^2 \le L\|x - y\|^2 \tag{3}$$

$$\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge \mu\|x - y\|^2 \tag{4}$$

于是(2)就变成了

$$d(T(x),T(y))^2 \le (1-2\mu+L)\|x-y\|^2 \tag{5}$$

那么我们就得到下面定理

定理 7.2. 设 $f(x)$ 满足(3)和(4)，并且 $2\mu-1 \le L \le 2\mu$，那么 $T(x)=x-\nabla f(x)$ 是一个压缩映射。并且对于任意的 x_0，迭代序列

$$x_{k+1}=T(x_k)$$

收敛到 x^*。

备注 7.1. 很容易看出(3)和(4)分别对应于 *Lipschitz* 连续性和单调性。也就是说我们可以通过对函数的 *Lipschitz* 连续性和强单调性来设计一个不动点迭代算法来求解优化问题。同时不动点迭代就是梯度下降法。类似的，牛顿法也可以理解为不动点迭代。

接下来我们考虑一个更一般的优化问题

$$\min f(x)+g(x) \tag{6}$$

为了重点关注算法设计思路，我们假设 f,g 都是足够光滑的。因此最优解满足

$$\nabla f(x^*)+\nabla g(x^*)=0 \tag{7}$$

根据上式，我们要将 x^* 设计为函数的不动点。下面有两种设计方式。

- (Forward-Barckward Splitting) 从(7)出发，我们可以将其改写为

$$\nabla f(x^*)=-\nabla g(x^*),$$

  这就给了我们一个设计的思路。我们可以将其改写为

$$x^*+\nabla f(x^*)=x^*-\nabla g(x^*),$$

  进一步

$$x^*=(I+\nabla f)^{-1}(I-\nabla g)(x^*),$$

  于是我们定义

$$T(x)=(I+\nabla f)^{-1}(I-\nabla g)(x),$$

  这样我们就得到了一个不动点迭代算法

$$x_{k+1}=(I+\nabla f)^{-1}(I-\nabla g)(x_k).$$

  引入中间变量 $y_k=(I-\nabla g)(x_k)$，我们可以得到

$$y_{k+1}=(I-\nabla g)(x_k) \tag{8}$$

$$x_{k+1}=(I+\nabla f)^{-1}y_{k+1}. \tag{9}$$

  这就是经典的 FBS 算法. 可以看出(8)是对 g 作梯度下降，而(9)是找到沿着 f 作梯度上升后是 y_{k+1} 的点。我们可以将其理解为一个前向后向的迭代算法。

- (Peaceman -Rachford Splitting)

$$\begin{aligned}
0=\nabla f(x)+\nabla g(x) \Longleftrightarrow & 0=(I+\alpha\nabla f)x-(I-\alpha\nabla g)x \\
\Longleftrightarrow & 0=(I+\alpha\nabla g)x-R_{\alpha\nabla g}(I+\alpha\nabla g)x,\ R_{\alpha\nabla g}=2(I+\alpha\nabla g)^{-1}-I \\
\Longleftrightarrow & 0=(I+\alpha\nabla f)x-R_{\alpha\nabla g}z,\ z=(I+\alpha\nabla g)x \\
\Longleftrightarrow & R_{\alpha\nabla g}z=(I+\alpha\nabla f)J_{\alpha\nabla g}z,\ x=J_{\alpha\nabla g}z \\
\Longleftrightarrow & J_{\alpha\nabla f}R_{\alpha\nabla g}z=J_{\alpha\nabla g}z \\
\Longleftrightarrow & J_{\alpha\nabla f}R_{\alpha\nabla g}z=\frac{R_{\alpha\nabla g}+I}{2}z \\
\Longleftrightarrow & 2J_{\alpha\nabla f}R_{\alpha\nabla g}z-R_{\alpha\nabla g}z=z \\
\Longleftrightarrow & R_{\alpha\nabla f}R_{\alpha\nabla g}z=z
\end{aligned} \tag{10}$$

- (Douglas-Rachford Splitting)

$$\begin{aligned}
0=\nabla f(x)+\nabla g(x) &\Longleftrightarrow (\frac{R_{\alpha\nabla f}R_{\alpha\nabla g}}{2}+\frac{1}{2})z=z \\
&\Longleftrightarrow z=J_{\alpha\nabla f}(2J_{\alpha\nabla g}-I)z+(I-J_{\alpha\nabla g})z
\end{aligned} \tag{11}$$

7.2 能量函数类

首先还是考虑单个函数的优化问题

$$\min f(x) \tag{12}$$

假设 x^* 是最优解，为了证明算法迭代的收敛性，我们需要找到一个能量函数（一般称为 Lyapunov 函数），它刻画了当前迭代点和最优解之间的距离, 我们将能量函数记作 $E(x^*,x^k)$ 并且要求

- $E(x^*,x^k)\geq 0$;
- $E(x^*,x^{k+1})<E(x^*,x^k)$.

那么我们的算法就是在收敛的。

常见的能量函数有

- (1) $E(x^*,x^k)=\|x^k-x^*\|^2$;
- (2) $E(x^*,x^k)=f(x^k)-f(x^*)$;
- (3) $E(x^*,x^k)=f(x^*)-f(x^k)-\langle\nabla f(x^k),x^*-x^k\rangle$, Bregman 距离.

备注 7.2. 能量函数的选择考验直觉和经验。

8 Douglas-Rachford 分裂算法的等价形式

$$
\begin{aligned}
& 0 = \nabla f(x) + \nabla g(x) \\
\iff \quad & 0 = (I + \alpha \nabla f)x - (I - \alpha \nabla g)x \\
\iff \quad & 0 = (I + \alpha \nabla g)x - R_{\alpha \nabla g}(I + \alpha \nabla g)x, \quad R_{\alpha \nabla g} = 2(I + \alpha \nabla g)^{-1} - I \\
\iff \quad & 0 = (I + \alpha \nabla f)x - R_{\alpha \nabla g} z, \quad z = (I + \alpha \nabla g)x \\
\iff \quad & R_{\alpha \nabla g} z = (I + \alpha \nabla f) J_{\alpha \nabla g} z, \quad x = J_{\alpha \nabla g} z \text{可以称为解映射} \\
\iff \quad & J_{\alpha \nabla f} R_{\alpha \nabla g} z = J_{\alpha \nabla g} z \\
\iff \quad & J_{\alpha \nabla f} R_{\alpha \nabla g} z = \frac{R_{\alpha \nabla g} + I}{2} z \\
\iff \quad & 2 J_{\alpha \nabla f} R_{\alpha \nabla g} z - R_{\alpha \nabla g} z = z \\
\iff \quad & R_{\alpha \nabla f} R_{\alpha \nabla g} z = z \\
\iff \quad & \left(\frac{R_{\alpha \nabla f} R_{\alpha \nabla g}}{2} + \frac{1}{2} \right) z = z \\
\iff \quad & z = J_{\alpha \nabla f}(2 J_{\alpha \nabla g} - I) z + (I - J_{\alpha \nabla g}) z
\end{aligned}
\tag{13}
$$

1. $z_{k+1} = T(z_k)$, 其中 $T = \frac{R_{\alpha \nabla f} R_{\alpha \nabla g}}{2} + \frac{1}{2}$

2. $z_{k+1} = J_{\alpha \nabla f}(2 J_{\alpha \nabla g} - I) z + (I - J_{\alpha \nabla g}) z_k$

3. $z_{k+1} = z_k + \mathbf{prox}_{\alpha f}(2\mathbf{prox}_{\alpha g}(z_k) - z_k) - \mathbf{prox}_{\alpha g}(z_k)$

4. $z_{k+1} = (I + \alpha \partial f)^{-1}[(I - \alpha \partial f)(I + \alpha \partial g)^{-1} + \alpha \partial f] z_k$

5. $x_{k+1} = J_{\alpha \partial f}\left[J_{\alpha \partial g}(I - \alpha \partial f) + \alpha \partial f\right] x_k$

以上等价形式从不同角度揭示了 Douglas-Rachford 分裂方法的结构本质：既可以理解为两次反射的组合，也可以视为近端算子的协调操作，进而统一了最优化中的投影法、变分不等式法与不动点迭代法。

8.1 收敛性分析

DR 方法可以看成是一个不动点迭代，因此要证明收敛性，我们需要证明以下两个结论：

1. y_k 收敛到 $F(y)$ 的不动点 y^*

2. $x_{k+1} = \mathbf{prox}_f(z_k)$ 收敛到 $x^* = \mathbf{prox}_f(z^*)$

在证明收敛性之前，需要先定义两个映射：

$$F(z) = z + \mathbf{prox}_g\big(2\mathbf{prox}_f(z) - z\big) - \mathbf{prox}_f(z),$$
$$G(z) = z - F(z) = \mathbf{prox}_f(z) - \mathbf{prox}_g\big(2\mathbf{prox}_f(z) - z\big).$$

我们要用到这两个函数的 **firmly nonexpansive**（co-coercive with parameter 1）性质：

$$(F(z) - F(\hat{z}))^T(z - \hat{z}) \geq \|F(z) - F(\hat{z})\|_2^2, \quad \forall\, z, \hat{z},$$
$$(G(z) - G(\hat{z}))^T(z - \hat{z}) \geq \|G(z) - G(\hat{z})\|_2^2, \quad \forall\, z, \hat{z}.$$

证明. 令 $x = \mathbf{prox}_f(z)$, $\hat{x} = \mathbf{prox}_f(\hat{z})$,

$$\nu = \mathbf{prox}_g(2x - z), \quad \hat{\nu} = \mathbf{prox}_g(2\hat{x} - \hat{z}).$$

则根据

$$F(z) = z + \nu - x, \quad F(\hat{z}) = \hat{z} + \hat{\nu} - \hat{x}$$

有

$$\begin{aligned}
(F(z) - F(\hat{z}))^\top(z - \hat{z}) &\leq (z + \nu - x - \hat{z} - \hat{\nu} + \hat{x})^\top(z - \hat{z}) - (x - \hat{x})^\top(z - \hat{z}) + \|x - \hat{x}\|^2 \\
&= (\nu - \hat{\nu})^T(z - \hat{z}) + \|z - x - (\hat{z} - \hat{x})\|_2^2 \\
&= (v - \hat{v})^\top(2x - z - 2\hat{x} + \hat{z}) - \|v - \hat{v}\|^2 + \|F(z) - F(\hat{z})\|^2 \\
&\geq \|F(z) - F(\hat{z})\|_2^2,
\end{aligned}$$

其中最后一步用到了 $\mathbf{prox}_f, \mathbf{prox}_g$ 算子的 firm nonexpansiveness 性质：

$$(x - \hat{x})^T(z - \hat{z}) \geq \|x - \hat{x}\|_2^2, \qquad (2x - z - 2\hat{x} + \hat{z})^T(\nu - \hat{\nu}) \geq \|\nu - \hat{\nu}\|_2^2.$$

同理可证 G 的 firm nonexpansiveness 性质。证毕。 □

然后我们可以根据以下的不动点迭代公式证明前面提到的收敛性：

$$z_{k+1} = (1 - \rho_k)\, z_k + \rho_k\, F(z_k) = z_k - \rho_k\, G(z_k),$$

其中需假设 F 的不动点存在，且满足 $0 \in \partial f(x^*) + \partial g(x^*)$，以及松弛参数

$$\rho_k \in [\rho_{\min}, \rho_{\max}], \quad 0 < \rho_{\min} < \rho_{\max} < 2.$$

证明. 设 z^* 为 $F(z)$ 的不动点（也即 $G(z)$ 的零点），考虑 $\{z_k\}$ 步进化：

$$\|z_{k+1} - z^*\|_2^2 - \|z_k - z^*\|_2^2 = 2(z_{k+1} - z_k)^T(z_k - z^*) + \|z_{k+1} - z_k\|_2^2.$$

带入 $z_{k+1} = z_k - \rho_k G(z_k)$，并利用 G 的 firm nonexpansiveness，可得

$$\|z_{k+1} - z^*\|_2^2 \le -\rho_k(2-\rho_k)\|G(z_k)\|_2^2 \le -M\|G(z_k)\|_2^2,$$

其中 $M = \rho_{\min}(2-\rho_{\max}) > 0$。上述不等式说明

$$M\sum_{k=0}^{\infty}\|G(z_k)\|_2^2 \le \|z_0 - z^*\|_2^2, \quad \|G(z_k)\| \to 0.$$

还可以得到 $\|z_k - z^*\|_2$ 是单调不增的，因此有界。再由 $\|z_k - z^*\|_2$ 单调不增、故极限 $\lim_{k\to\infty}\|z_k - z^*\|_2$ 存在，又由于有界, 故存在收敛子序列。

记 $\bar{z}$ 为一个收敛子序列收敛到的极限点，根据 G 的连续性有

$$0 = \lim_{k\to\infty} G(z_{k_j}) = G(\bar{z}),$$

即 $\bar{z}$ 是 G 的零点，且极限 $\lim_{k\to\infty}\|z_{k_j} - z^*\|_2$ 存在。

接着需要证明唯一性。假设 $\bar{z}, \hat{z}$ 是两个不同的极限点，收敛极限

$$\lim_{k\to\infty}\|z_k - \bar{z}\|_2, \quad \lim_{k\to\infty}\|z_k - \hat{z}\|_2$$

都存在，因此

$$\|\bar{z} - \hat{z}\|_2 = \lim_{k\to\infty}\|z_k - \hat{z}\|_2 = \lim_{k\to\infty}\|z_k - \bar{z}\|_2 = 0.$$

从而 $\bar{z} = \hat{z}$，即极限唯一。 □

- Fejér 单调性蕴含 $\{x_k\}$ 有界，且对于任意 $p \in C$，距离序列 $\{\|x_k - p\|\}$ 收敛；
- 结合 Bolzano-Weierstrass 引理（有界序列存在收敛子列）与极限点唯一性，可推出全序列收敛，且其极限落在 C 中的某一点。

9 Scaling 缩放技巧

10 Kurdyka-Łojasiewicz 条件

KL 条件是用于分析非凸优化问题中算法收敛性的强有力工具，尤其在凸但不可微或非凸优化问题中起着核心作用。KL 条件本质上是函数在临界点附近的一种“渐进良性行为”的刻画。

定义 10.1 (KL 函数). 设 $\phi : \mathbb{R}^n \to (-\infty, +\infty]$ 是下半连续的，并且在某点 x^* 附近是有限值。若存在：

- 一个邻域 U 使得 $x^* \in U$；
- 一个常数 $\eta \in (0, +\infty]$；
- 一个函数 $\varphi \in \Phi_\eta$，其中

$$\Phi_\eta := \left\{ \varphi \in C^0([0, \eta)) \cap C^1((0, \eta)) \mid \varphi(0) = 0,\ \varphi' > 0 \right\};$$

使得对于所有 $x \in U$ 满足 $x \neq x^*$ 且 $\phi(x^*) < \phi(x) < \phi(x^*) + \eta$，都有

$$\varphi'(\phi(x) - \phi(x^*)) \cdot \|\partial \phi(x)\| \geq 1,$$

其中 $\partial \phi(x)$ 表示 *Clarke* 次微分或 *Fréchet* 次微分。

则称 ϕ 在 x^* 附近满足 ***KL*** **条件**。

备注 10.1. *KL* 条件并不要求目标函数是凸的或光滑的，只要满足一定的“正则性”。很多常见的非凸函数，如半代数函数（*polynomial, piecewise linear,* ℓ_1 范数等）都满足 *KL* 条件。

示例 10.1 (KL 函数的形式). 常用的 *KL* 函数 $\varphi(s)$ 形式如下：

$$\varphi(s) = c s^{1-\theta}, \quad \theta \in [0, 1),\ c > 0.$$

这种形式能反映不同类型的收敛速度：

- $\theta = 0$ 时：有限步收敛；
- $\theta \in (0, \frac{1}{2}]$：线性收敛；
- $\theta \in (\frac{1}{2}, 1)$：亚线性收敛。

备注 10.2. 如果一个优化问题的目标函数满足 *KL* 条件，并且算法具有适当的下降性质，则可以几乎自动推导出全局收敛性和速率。

在基于 KL 性质证明迭代算法收敛性的过程中，通常需要借助以下三个关键不等式：

1. **KL 不等式（梯度模长 $\geq$ 目标函数值下降量）：**

$$\|\nabla f(x_k)\| \geq \varphi'\left(f(x_k) - f^*\right),$$

其中 φ 是 KL 函数，f^* 表示目标函数的极小值。该不等式表明，当函数值靠近极小值时，梯度趋于零。

2. **充分下降：**

$$|f(x_k) - f(x_{k+1})| \geq \alpha\|x_k - x_{k+1}\|^2,$$

其中 $\alpha > 0$ 是某个固定常数。此不等式说明函数值在每次迭代中具有充分下降性，是算法收敛的重要依据。

3. **梯度模长估计：**

$$\|x_k - x_{k+1}\| \geq \beta\|\nabla f(x_k)\|,$$

其中 $\beta > 0$ 是某个固定常数。该不等式用于连接变量变化幅度与梯度大小，从而与 KL 不等式结合，形成收敛性的闭环链式估计。

11 Farkas 引理

Farkas 引理是凸分析、线性规划和优化理论中的基础结果，描述了线性系统可行性的一种二择一关系。

引理 11.1 (Farkas 引理，标准版). *设 $A \in \mathbb{R}^{m \times n}$，$b \in \mathbb{R}^m$。恰好有且只有以下两种情况之一成立：*

1. *存在 $x \in \mathbb{R}^n$，使得 $Ax = b$ 且 $x \geq 0$；*
2. *存在 $y \in \mathbb{R}^m$，使得 $A^T y \geq 0$ 且 $b^T y < 0$。*

直观理解：

- 要么系统 $Ax = b, x \geq 0$ 有解；
- 要么存在一个向量 y，作为"证伪者"，证明无解。

本质上反映了凸集的"要么有交集，要么可以被超平面严格分开"的原理。