凸优化和单调变分不等式的收缩算法

第十讲: 基于梯度投影的凸优化 收缩算法和下降算法

Projected gradient-based contraction method and descent method for convex optimization

南京大学数学系　　何炳生
hebma@nju.edu.cn

简单约束的凸优化问题

这一讲讨论简单约束可微凸优化问题

$$\min \{f(x) \mid x \in \Omega\}$$

的梯度算法, 其中 Ω 是$\Re^n$ 中的凸闭集, 并假设到 Ω 上的投影是容易实现的. 在第一讲中就已经提到, 简单约束可微凸优化问题等价于求变分不等式

$$\mathrm{VI}(\Omega, \nabla f) \qquad x^* \in \Omega, \qquad (x - x^*)^T \nabla f(x^*) \geq 0, \quad \forall x \in \Omega$$

的解. 这一讲的投影梯度方法, 分别是收缩算法和下降算法, 都不要用到函数值 $f(x)$, 只要对给定的 x, 能提供 $\nabla f(x)$. 收缩算法保证迭代点向解集靠近. 下降算法则隐含了目标函数值下降, 尽管目标函数值在计算过程中从不出现.

设 x^* 是变分不等式 $\mathrm{VI}(\Omega, \nabla f)$ 的解. 由于 $\tilde{x} = P_\Omega[x - \beta \nabla f(x)] \in \Omega$, 因此根据变分不等式的定义有第一个基本不等式

$$\text{(FI1)} \quad (\tilde{x} - x^*)^T \beta \nabla f(x^*) \geq 0.$$

由于 $\tilde{x}$ 是 $x - \beta \nabla f(x)$ 在 Ω 上的投影, $x^* \in \Omega$, 根据投影的基本性质, 有

$$\text{(FI2)} \quad (\tilde{x} - x^*)^T \left([x - \beta \nabla f(x)] - \tilde{x} \right) \geq 0.$$

1 凸优化的投影收缩算法

投影收缩算法由投影 $P_\Omega[x^k - \beta\nabla f(x^k)]$ 得到 $\tilde{x}^k$. $\quad x^k \in \Omega^* \Leftrightarrow x^k = \tilde{x}^k$.

误差度量函数 一个非负函数 $\varphi(x^k, \tilde{x}^k)$ 称作变分不等式 VI(Ω, F) 的误差度量函数, 如果有 $\delta > 0$, 使得

$$\varphi(x^k, \tilde{x}^k) \geq \delta\|x^k - \tilde{x}^k\|^2, \quad \text{并且} \quad \varphi(x^k, \tilde{x}^k) = 0 \Leftrightarrow x^k = \tilde{x}^k. \tag{1.1}$$

有利方向 设矩阵 G 正定, 凸优化的有利方向都是 $(x^k - \tilde{x}^k)$, 它满足

$$(x^k - x^*)^T G(x^k - \tilde{x}^k) \geq \varphi(x^k, \tilde{x}^k), \ \forall\, x^* \in \Omega^*. \tag{1.2}$$

初等收缩算法 考虑将 (1.1) 中的 $\varphi(x^k, \tilde{x}^k) \geq \delta\|x^k - \tilde{x}^k\|^2$ 改成条件

$$\varphi(x^k, \tilde{x}^k) \geq \frac{1}{2}\big(\|x^k - \tilde{x}^k\|_G^2 + \tau\|x^k - \tilde{x}^k\|^2\big), \quad (\tau > 0). \tag{1.3}$$

我们将条件 (1.2) 和 (1.3) 满足时, 用步长为 **1** 的迭代公式

$$x^{k+1} = x^k - (x^k - \tilde{x}^k) = \tilde{x}^k, \tag{1.4}$$

产生新迭代点的方法, 称为 **Primary Method** (初等方法).

1.1 凸二次优化的投影收缩算法

对凸二次优化, 我们将第一和第二个基本不等式

$$\begin{cases} (\tilde{x}-x^*)^T\beta(Hx^*+c)\geq 0 & (FI1)\\ (\tilde{x}-x^*)^T\big([x-\beta(Hx+c)]-\tilde{x}\big)\geq 0 & (FI2)\end{cases}$$

相加, 对所有的 $x\in\Re^n$, 都有

$$\{(x-x^*)-(x-\tilde{x})\}^T\{(x-\tilde{x})-\beta H(x-x^*)\}\geq 0$$

由上式和 H 半正定,

$$(x-x^*)^T(I+\beta H)(x-\tilde{x})\geq\|x-\tilde{x}\|^2. \qquad (1.1)$$

误差度量函数就是 $\varphi(x^k,\tilde{x}^k)=\|x^k-\tilde{x}^k\|^2$. 上式说明 $(\tilde{x}^k-x^k)$ 是未知距离函数 $\frac{1}{2}\|x-x^*\|^2_{(I+\beta H)}$ 在 x^k 处的下降方向. 记 $G=I+\beta H$, 就是(1.2) 的形式

$$(x^k-x^*)^TG(x^k-\tilde{x}^k)\geq\varphi(x^k,\tilde{x}^k),\ \forall\, x^*\in\Omega^*.$$

这里考虑的投影收缩算法, 要求它产生的迭代点使得 $\|x^k - x^*\|_G^2$ 严格单调下降. 以

$$x(\alpha) = x^k - \alpha(x^k - \tilde{x}^k) \tag{1.2}$$

产生依赖于步长 α 的新迭代点. 考察与 α 相关的距离平方缩短量

$$\vartheta(\alpha) = \|x^k - x^*\|_G^2 - \|x(\alpha) - x^*\|_G^2. \tag{1.3}$$

利用 (1.1) 就有

$$\begin{aligned} \vartheta(\alpha) &= \|x^k - x^*\|_G^2 - \|x^k - x^* - \alpha(x^k - \tilde{x}^k)\|_G^2 \\ &\geq 2\alpha\|x^k - \tilde{x}^k\|^2 - \alpha^2\|x^k - \tilde{x}^k\|_G^2. \end{aligned} \tag{1.4}$$

我们得到 $\vartheta(\alpha)$ 的一个下界、二次函数 $q(\alpha)$,

$$q(\alpha) = 2\alpha\|x^k - \tilde{x}^k\|^2 - \alpha^2\|x^k - \tilde{x}^k\|_G^2. \tag{1.5}$$

使 $q(\alpha)$ 达到极大的

$$\alpha_k^* = \|x^k - \tilde{x}^k\|^2 / \|x^k - \tilde{x}^k\|_G^2. \tag{1.6}$$

用迭代式

$$x^{k+1} = x^k - \gamma\alpha_k^*(x^k - \tilde{x}^k), \quad \gamma \in (0, 2) \tag{1.7}$$

产生新的迭代点 x^{k+1}. 这样的迭代序列 $\{x^k\}$ 满足

$$\|x^{k+1}-x^*\|^2_{(I+\beta H)} \le \|x^k-x^*\|^2_{(I+\beta H)} - \gamma(2-\gamma)\alpha_k^*\|x^k-\tilde{x}^k\|^2. \quad (1.8)$$

换句话说, 迭代序列 $\{x^k\}$ 在 $G=(I+\beta H)$-模下向解集收缩. 关于凸二次优化在$(I+\beta H)$-模下收缩的算法, 更详细的可参见文献 [5]. 一般说来, 如何选取适当的 β 对收敛速度影响很大.

自调比投影收缩算法 = 故意缩短了步长的最速下降法

注意到在 $G=(I+\beta H)$ 时, "最优步长"

$$\alpha_k^* = \frac{\|x^k-\tilde{x}^k\|^2}{(x^k-\tilde{x}^k)^T(I+\beta H)(x^k-\tilde{x}^k)}. \quad (1.9)$$

对给定的 $\nu\in(0,1)$, 我们使用自调比法则选取 β 使得

$$(x^k-\tilde{x}^k)^T(\beta H)(x^k-\tilde{x}^k) \le \nu\|x^k-\tilde{x}^k\|^2, \quad (1.10)$$

代入 (1.9) 便有

$$\alpha_k^* \ge \frac{1}{1+\nu} > \frac{1}{2}.$$

这使得我们有可能在 (1.7) 中动态地取

$$\gamma_k = 1/\alpha_k^*, \qquad \text{因此} \qquad 1 < \gamma_k \le 1+\nu < 2.$$

迭代式 (1.7) 就成为

$$x^{k+1} = \tilde{x}^k = P_\Omega[x^k - \beta(Hx^k + c)]. \tag{1.11}$$

这样做使得有可能每步迭代只计算一次目标函数的梯度(矩阵与向量相乘).

在 (1.8) 中利用 $\gamma_k\alpha_k^* = 1,\ \gamma_k \le 1+\nu$ 以及 $\tilde{x}^k = x^{k+1}$ 就有

$$\|x^{k+1} - x^*\|^2_{(I+\beta H)} \le \|x^k - x^*\|^2_{(I+\beta H)} - (1-\nu)\|x^k - x^{k+1}\|^2. \tag{1.12}$$

♣ 求解 $\min\{\frac{1}{2}x^THx + c^Tx\}$ 的最速下降法是

$$x^{k+1} = x^k - \alpha_k^{SD}(Hx^k + c), \qquad \alpha_k^{SD} = \frac{\|Hx^k + c\|^2}{(Hx^k + c)^TH(Hx^k + c)}.$$

当 $\Omega = \Re^n$ 时, 根据 (1.11) 和 (1.10) 有

$$x^{k+1} = x^k - \beta(Hx^k + c) \qquad \text{和} \qquad \beta \le \nu\alpha_k^{SD}.$$

因此, [5] 中符合条件 (1.10) 的投影收缩算法(1.11) 相当于将最速下降法故意缩短了步长. 换句话说, 据此设计的求解无约束凸二次优化的算法, 恰是缩小了步长的最速下降法. 我们曾担心用这些方法求解无约束凸二次规划效果会比最速下降法还差. 事实上, 对最速下降法故意缩短步长, 收敛速率有令人难以置信的数量级提高. 读者容易用下面的例子来验证.

数值试验 试验问题中的 Hessian 矩阵是 Hilbert 矩阵, 即:

$$H = \{h_{ij}\}, \quad h_{ij} = \frac{1}{i+j-1}, \quad i = 1, \cdots, n; \quad j = 1, \cdots, n.$$

问题的规模(维数)分别从从 100 到 500. 因为 Hilbert 矩阵的条件很坏, 我们将最优解 x^* 设定为每个分量都是 1 的向量, 然后令 $c = -Hx^*$, 再用梯度类算法去求解. 试验中, 我们分别将零向量、 c 和 $-c$ 取作初始向量. 这里采用的停机准则是$\|Hx^k + c\|/\|Hx^0 + c\| \leq 10^{-7}$. 表 1–3 分别列出了在不同问题规模、不同缩扩(缩减或扩张)因子和不同初始向量下的迭代次数. 其中 n 表示问题规模, r 表示缩扩因子. $r = 1$ 时, 就是最速下降法.

初步的试验结果不但解除了我们的上述担心, 同时也印证了 Dai 和 Yuan [2] 提到的将最速下降法的步长乘上一个小于 1 的因子会加快收敛的结论.

表 1. 初始向量 $x^0 = 0$, 使用不同缩扩因子 r 时的迭代次数

n=	0.1	0.3	0.5	0.7	0.8	**0.9**	0.95	0.99	**1.00**	1.20
100	2863	1346	853	627	582	**437**	565	1201	**13169**	22695
200	3283	1398	923	804	541	**669**	898	1178	**14655**	21083
300	3497	1323	856	739	720	**568**	619	1545	**17467**	24027
500	3642	1351	1023	773	667	**578**	836	2024	**17757**	22750

初始向量 $x^0 = 0$, 迭代结束时平均相对误差 $\|x^k - x^*\|/\|x^0 - x^*\| = 3.0\text{e} - 3$

表 2. 初始向量 $x^0 = c$, 使用不同缩扩因子 r 时的迭代次数

n=	0.1	0.3	0.5	0.7	0.8	**0.9**	0.95	0.99	**1.0**	1.2
100	2129	1034	544	424	302	**438**	568	919	**5527**	9667
200	1880	808	568	482	372	**339**	446	713	**6625**	11023
300	1852	1002	741	531	610	**452**	450	917	**6631**	10235
500	2059	939	568	573	379	**547**	558	874	**7739**	11269

初始向量 $x^0 = c$, 迭代结束时平均相对误差 $\|x^k - x^*\|/\|x^0 - x^*\| = 1.8\text{e} - 3$

表 3. 初始向量 $x^0 = -c$, 使用不同缩扩因子 r 时的迭代次数

n=	0.1	0.3	0.5	0.7	0.8	**0.9**	0.95	0.99	**1.0**	1.2
100	2545	1221	666	591	498	**482**	638	1581	**14442**	20380
200	2826	990	874	470	526	**455**	578	841	**15222**	18892
300	2891	1299	918	738	549	**571**	608	2552	**18762**	21208
500	3158	1769	909	678	506	**512**	678	1240	**17512**	19790

初始向量 $x^0 = -c$, 迭代结束时平均相对误差 $\|x^k - x^*\|/\|x^0 - x^*\| = 3.8\text{e} - 3$

1.2 基于 FI2 的非线性凸优化的投影收缩算法

可微凸优化问题 $\min\{f(x)\,|\,x\in\Omega\}$ 与变分不等式 VI$(\Omega,\nabla f)$ 等价. 由

$$\tilde{x}=P_{\Omega}[x-\beta\nabla f(x)]$$

和投影的基本性质, 可以得到

$$(\tilde{x}-x^*)^T\big([x-\beta\nabla f(x)]-\tilde{x}\big)\geq 0.$$

上面的不等式其实就是第二讲中的第二个基本不等式. 据此可以推得

$$\begin{aligned}(\tilde{x}-x^*)^T(x-\tilde{x}) &\geq (\tilde{x}-x^*)^T\beta\nabla f(x)\\ &= \quad (x-x^*)^T\beta\nabla f(x)-(x-\tilde{x})^T\beta\nabla f(x).\end{aligned}\tag{1.13}$$

另据凸函数的性质, 有

$$(x-x^*)^T\nabla f(x)\geq f(x)-f(x^*)\geq (x-\tilde{x})^T\nabla f(\tilde{x})+(f(\tilde{x})-f(x^*)).\tag{1.14}$$

以 (1.14) 代入 (1.13) 的右端, 则对所有的 $x\in\Re^n$, 都有

$$(\tilde{x}-x^*)^T(x-\tilde{x})\geq -\beta(x-\tilde{x})^T(\nabla f(x)-\nabla f(\tilde{x})).$$

由上式得

$$(x-x^*)^T(x-\tilde{x}) \geq \|x-\tilde{x}\|^2 - \beta(x-\tilde{x})^T(\nabla f(x)-\nabla f(\tilde{x})). \qquad (1.15)$$

我们定义

$$\varphi(x,\tilde{x}) = \|x-\tilde{x}\|^2 - \beta(x-\tilde{x})^T(\nabla f(x)-\nabla f(\tilde{x})). \qquad (1.16)$$

假设 ∇f 是 Lipschitz 连续的. 对于一个确定的 $\nu \in (0,1)$, 总可以采用 Armijo 技术对算子进行调比, 使得 $\beta\nabla f$ 的 Lipschitz 常数不大于 ν, 有

$$\beta\|\nabla f(x)-\nabla f(\tilde{x})\| \leq \nu\|x-\tilde{x}\|$$

式成立. 这样便有

$$(x-x^*)^T(x-\tilde{x}) \geq \varphi(x,\tilde{x}) \geq (1-\nu)\|x-\tilde{x}\|^2.$$

由 (1.16) 定义的 $\varphi(x,\tilde{x})$ 满足条件 (1.1), 其中的 $\delta = 1-\nu > 0$. 考虑欧氏模下的收缩算法, 由

$$x^{k+1} = x^k - \gamma\alpha_k^*(x^k-\tilde{x}^k),$$

产生新的迭代点 x^{k+1}, 其中

$$\alpha_k^* = \frac{\varphi(x^k, \tilde{x}^k)}{\|x^k - \tilde{x}^k\|^2}, \qquad \gamma \in (0, 2).$$

迭代序列 $\{x^k\}$ 就满足

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \gamma(2-\gamma)\alpha_k^* \varphi(x^k, \tilde{x}^k).$$

由 $\varphi(x, \tilde{x}) \geq (1-\nu)\|x - \tilde{x}\|^2$ 可以进一步推得

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \gamma(2-\gamma)(1-\nu)^2\|x^k - \tilde{x}^k\|^2.$$

有关这一节的相关论文和数值试验可见:

- B.S. He, L.Z. Liao, and X. Wang, Proximal-like contraction methods for monotone variational inequalities in a unified framework I: Effective quadruplet and primary methods, Comput. Optim. Appl., 51, 649-679, 2012
- B.S. He, L.Z. Liao, and X. Wang, Proximal-like contraction methods for monotone variational inequalities in a unified framework II: General methods and numerical experiments, Comput. Optim. Appl. 51, 681-708, 2012

2 Projected Gradient Descent (PDG) method for nonlinear convex optimization

The findings on projection and contraction method for solving the quadratic programming also contribute to solving the following differentiable convex optimization problem.

Let Ω be a convex closed set in R^n. The problem concerted in this subsection is to find $x^* \in \Omega$, such that

$$(x - x^*)^T g(x^*) \geq 0, \quad \forall\, x \in \Omega, \tag{2.1}$$

where $g(x)$ is a mapping from R^n into itself. We assume that $g(x)$ is the gradient of a certain convex function, say $f(x)$, however, $f(x)$ is not provided. Only for given x, $g(x)$ is observable (sometimes with costly expenses).

In other words, (2.1) is equivalent to the following convex optimization problem

$$\min \{f(x) \mid x \in \Omega\}. \tag{2.2}$$

We call (3.2) an **oracle** convex optimization problem, because only the gradient information $g(x)$ can be used for solving (3.2). For $x^* \in \Omega^*$, we assume $f(x^*) > -\infty$.

In addition, we also assume that $g(x)$ is Lipschitz continuous, i.e., there exists a constant $L > 0$ such that

$$\|g(x) - g(y)\| \leq L\|x - y\|, \quad \forall\, x,\, y \in \Re^n. \tag{2.3}$$

We require that $g(x)$ is Lipschitz continuous while it does not need to know the value of L in (2.3).

The methods presented in this section do not involve the value of $f(x)$, but they can guarantee that $f(x^k)$ is strict monotonically decreasing, hence they belong to the descending methods.

2.1 Steepest descent method for convex programming

Single step projected gradient method.

Step k. ($k \geq 0$) Set

$$x^{k+1} = P_\Omega[x^k - \beta_k g(x^k)], \tag{2.4a}$$

where the step size β_k satisfies the following condition:

$$(x^k - x^{k+1})^T (g(x^k) - g(x^{k+1})) \leq \frac{\nu}{\beta_k} \|x^k - x^{k+1}\|^2. \tag{2.4b}$$

Note that the condition (2.4b) automatically holds when $\beta_k \leq \nu/L$, where L is

the Lipschitz modulus of $g(x)$. The reason is

$$\begin{aligned}&(x^k - x^{k+1})^T \beta_k (g(x^k) - g(x^{k+1})) \\ &\quad \leq \quad \|x^k - x^{k+1}\| \cdot \beta_k L \|x^k - x^{k+1}\| \;\leq\; \nu \|x^k - x^{k+1}\|^2.\end{aligned}$$

2.2 Global convergence of the proposed method

In the following, we show an important lemma by using the basic properties of the projection and convex function.

Lemma 2.1 *For given x^k, let x^{k+1} be generated by* (2.4a)*. If the step-size β_k satisfies* (2.4b)*, then we have*

$$(x - x^{k+1})^T g(x^k) \geq \tfrac{1}{\beta_k} (x - x^{k+1})^T (x^k - x^{k+1}), \quad \forall x \in \Omega, \tag{2.5}$$

and

$$\begin{aligned}&\beta_k (f(x) - f(x^{k+1})) \\ &\quad \geq \; (x - x^{k+1})^T (x^k - x^{k+1}) - \nu \|x^k - x^{k+1}\|^2, \; \forall\, x \in \Omega.\end{aligned} \tag{2.6}$$

Proof. Note that x^{k+1} is the projection of $[x^k - \beta_k g(x^k)]$ on Ω. Using the projection's property, $(x - P_\Omega(z))^T(z - P_\Omega(z)) \leq 0$, $\forall x \in \Omega$, we have

$$(x - x^{k+1})^T\{[x^k - \beta_k g(x^k)] - x^{k+1}\} \leq 0, \quad \forall\, x \in \Omega.$$

It follows that

$$(x - x^{k+1})^T \beta_k g(x^k) \geq (x - x^{k+1})^T(x^k - x^{k+1}), \quad \forall\, x \in \Omega, \tag{2.7}$$

and the first assertion (2.5) is proved. Using the convexity of f , we have

$$f(x) \geq f(x^k) + (x - x^k)^T g(x^k), \tag{2.8}$$

and

$$\begin{aligned} f(x^k) &\geq f(x^{k+1}) + (x^k - x^{k+1})^T g(x^{k+1}) \\ &= f(x^{k+1}) + (x^k - x^{k+1})^T g(x^k) \\ &\qquad -(x^k - x^{k+1})^T\big(g(x^k) - g(x^{k+1})\big) \\ &\geq f(x^{k+1}) + (x^k - x^{k+1})^T g(x^k) - \frac{\nu}{\beta_k}\|x^k - x^{k+1}\|^2. \end{aligned} \tag{2.9}$$

The last " $\geq$ " is due to (2.4b). From (2.8) and (2.9), we get

$$\begin{aligned} & f(x) - f(x^{k+1}) \\ & \quad \geq \quad f(x^k) + (x - x^k)^T g(x^k) \\ & \qquad -\Big\{ f(x^k) + (x^{k+1} - x^k)^T g(x^k) + \frac{\nu}{\beta_k} \|x^k - x^{k+1}\|^2 \Big\} \\ & \quad = \quad (x - x^{k+1})^T g(x^k) - \frac{\nu}{\beta_k} \|x^k - x^{k+1}\|^2. \end{aligned} \tag{2.10}$$

Substituting (2.5) in (2.10), we obtain

$$f(x) - f(x^{k+1}) \geq \frac{1}{\beta_k} (x - x^{k+1})^T (x^k - x^{k+1}) - \frac{\nu}{\beta_k} \|x^k - x^{k+1}\|^2,$$

and the second assertion of this lemma is proved. $\square$

The following theorem shows that the projected gradient method (2.4) is a descent method whose objective function value $\{f(x^k)\}$ is monotonically decreasing.

Theorem 2.1 *Let $\{x^k\}$ be the sequence generated by the single step projected gradient method* (2.4)*. Then, we have*

$$f(x^{k+1}) \leq f(x^k) - \frac{1-\nu}{\beta_k}\|x^k - x^{k+1}\|^2, \tag{2.11}$$

and

$$\begin{aligned} \|x^{k+1} - x^*\|^2 \quad \leq \quad & \|x^k - x^*\|^2 - (1-2\nu)\|x^k - x^{k+1}\|^2 \\ & -2\beta_k(f(x^{k+1}) - f(x^*)). \end{aligned} \tag{2.12}$$

Proof. Setting $x = x^k$ in (2.6) in Lemma 2.1, we obtain the assertion (2.11) immediately. Next, setting $x = x^*$ in (2.6), we have

$$\begin{aligned} & \beta_k(f(x^*) - f(x^{k+1})) \\ & \quad \geq \quad (x^* - x^{k+1})^T(x^k - x^{k+1}) - \nu\|x^k - x^{k+1}\|^2, \end{aligned}$$

and thus

$$\begin{aligned}
&(x^k - x^*)^T(x^k - x^{k+1}) \\
&\quad \geq \quad (1-\nu)\|x^k - x^{k+1}\|^2 + \beta_k(f(x^{k+1}) - f(x^*)).
\end{aligned}$$

Using the above inequality, we get

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 \\
&\quad = \quad \|(x^k - x^*) - (x^k - x^{k+1})\|^2 \\
&\quad = \quad \|x^k - x^*\|^2 - 2(x^k - x^*)^T(x^k - x^{k+1}) + \|x^k - x^{k+1}\|^2 \\
&\quad \leq \quad \|x^k - x^*\|^2 - 2(1-\nu)\|x^k - x^{k+1}\|^2 \\
&\qquad\qquad -2\beta_k(f(x^{k+1}) - f(x^*)) + \|x^k - x^{k+1}\|^2 \\
&\quad = \quad \|x^k - x^*\|^2 - (1-2\nu)\|x^k - x^{k+1}\|^2 \\
&\qquad\qquad -2\beta_k(f(x^{k+1}) - f(x^*)).
\end{aligned}$$

This completes the proof of the assertion (2.12). □

Directly from (2.12), it follows the following corollary**:**

Corollary 2.1 *Let $\{x^k\}$ be the sequence generated by the single step projected gradient method* (2.4)*. If $\nu \leq \frac{1}{2}$, we have $\|x^{k+1} - x^*\|^2 < \|x^k - x^*\|^2$, for any $x^* \in \Omega^*$. The generated sequence $\{x^k\}$ is in a compact set.*

2.3 Convergence rate of the proposed method

Below we show that the iteration-complexity of the single projected gradient method is $O(1/k)$. For the convenience, we assume $\beta_k \equiv \beta$.

Theorem 2.2 *Let $\{x^k\}$ be generated by the single step projected gradient method* (2.4)*. Then, we have*

$$\begin{aligned} & 2k\beta(f(x^k) - f(x^*)) \\ & \leq \|x^0 - x^*\|^2 - \sum_{l=0}^{k-1} \Big((1-2\nu) + 2l(1-\nu) \Big) \|x^l - x^{l+1}\|^2. \end{aligned} \tag{2.13}$$

Proof. First, it follows from (2.12) that, for any $x^* \in \Omega^*$ and all $l \geq 0$, we have

$$2\beta(f(x^*)-f(x^{l+1})) \geq \|x^{l+1}-x^*\|^2-\|x^l-x^*\|^2+(1-2\nu)\|x^l-x^{l+1}\|^2.$$

Summing the above inequality over $l = 0, \ldots, k-1$, we obtain

$$\begin{aligned} & 2\beta\Big(kf(x^*) - \sum_{l=0}^{k-1} f(x^{l+1})\Big) \\ \geq \quad & \|x^k - x^*\|^2 - \|x^0 - x^*\|^2 + \sum_{l=0}^{k-1}(1-2\nu)\|x^l - x^{l+1}\|^2. \end{aligned} \tag{2.14}$$

It follows from (2.11) that

$$2\beta l\big(f(x^l) - f(x^{l+1})\big) \geq 2l(1-\nu)\|x^l - x^{l+1}\|^2,$$

which can be rewritten as

$$2\beta\big(lf(x^l) - (l+1)f(x^{l+1}) + f(x^{l+1})\big) \geq 2l(1-\nu)\|x^l - x^{l+1}\|^2.$$

Summing the above inequality over $l = 0, \ldots, k-1$, it follows that

$$2\beta \sum_{l=0}^{k-1} \Big(l f(x^l) - (l+1) f(x^{l+1}) + f(x^{l+1}) \Big) \geq \sum_{l=0}^{k-1} 2l(1-\nu) \|x^l - x^{l+1}\|^2,$$

which simplifies to

$$2\beta \Big(-k f(x^k) + \sum_{l=0}^{k-1} f(x^{l+1}) \Big) \geq \sum_{l=0}^{k-1} 2l(1-\nu) \|x^l - x^{l+1}\|^2. \qquad (2.15)$$

Adding (2.14) and (2.15), we get

$$\begin{aligned} & 2k\beta \big(f(x^*) - f(x^k) \big) \\ & \quad \geq \quad -\|x^0 - x^*\|^2 + \sum_{l=0}^{k-1} \Big((1-2\nu) + 2l(1-\nu) \Big) \|x^l - x^{l+1}\|^2, \end{aligned}$$

which implies (2.13) and the theorem is proved. □

From (2.13) follows directly the following theorem.

Theorem 2.3 *Let $\{x^k\}$ be generated by the single step projected gradient method. If $\nu \leq \frac{1}{2}$, then we have*

$$f(x^k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2}{2k\beta}, \tag{2.16}$$

and thus the iteration-complexity of this method is $O(1/k)$.

What is about for any $\nu \in (0.5, 1)$? For such ν, we define

$$p(\nu) = \operatorname{argmin}\{l \,|\, l \geq 0 \text{ is a integer, } (1-2\nu) + 2l(1-\nu) \geq 0\}. \tag{2.17}$$

For any $\nu \in (0.5, 1)$, $p(\nu)$ is finite number. For example,we have

$\nu =$	0.9	0.8	0.7	$(0.5, 0.7)$
$p(\nu) =$	4	2	1	1

.

Since the term $\sum_{p(\nu)}^{k-1} \Big((1-2\nu) + 2l(1-\nu) \Big) \|x^l - x^{l+1}\|^2$ is positive, it

follows from Theorem 2.2 (see(2.13)) that

$$2k\beta(f(x^k)-f(x^*)) \leq \|x^0-x^*\|^2- \sum_{l=0}^{p(\nu)-1} \Big((1-2\nu)+2l(1-\nu)\Big)\|x^l-x^{l+1}\|^2.$$

The last inequality implies that $\lim_{k\to\infty}(f(x^k) - f(x^*)) = 0$.

The iteration-complexity of this method is $O(1/k)$ for any $\nu \in (0, 1)$.

Theorem 2.4 *Let $\{x^k\}$ be generated by the single step projected gradient method, then we have*

$$f(x^k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2 + D}{2k\beta}, \tag{2.18}$$

where

$$D = -\sum_{l=0}^{p(\nu)-1} \Big((1 - 2\nu) + 2l(1 - \nu)\Big)\|x^l - x^{l+1}\|^2.$$

and $p(\nu)$ is a finite integer defined in (2.17).

Self-adaptive projected gradient descent method

Self-adaptive projected gradient descent method.

Set $\beta_0 = 1$, $\mu = 0.5$, $\nu = 0.9$, $x^0 \in \Omega$ and $k = 0$. Provide $g(x^0)$.

For $k = 0, 1, \ldots$, if the stopping criterium is not satisfied, do

Step 1. $\tilde{x}^k = P_\Omega[x^k - \beta_k g(x^k)]$,

$r_k = \beta_k \|g(x^k) - g(\tilde{x}^k)\| / \|x^k - \tilde{x}^k\|$.

while $r_k > \nu$

$\beta_k := \beta_k * 0.8 / r_k$,

$\tilde{x}^k = P_\Omega[x^k - \beta_k g(x^k)]$,

$r_k = \beta_k \|g(x^k) - g(\tilde{x}^k)\| / \|x^k - \tilde{x}^k\|$.

end(while)

$x^{k+1} = \tilde{x}^k$,

$g(x^{k+1}) = g(\tilde{x}^k)$.

If $r_k \leq \mu$ then $\beta_k := \beta_k * 1.5$, **end(if)**

Step 2. $\beta_{k+1} = \beta_k$ and $k = k + 1$, go to Step 1.

Remark 2.1 *Instead of the condition* (2.4b)*, here we have*

$$\beta_k \|g(x^k) - g(x^{k+1})\| \leq \nu \|x^k - x^{k+1}\|.$$

Remark 2.2 *If $r_k \leq \nu$, we direct take $x^{k+1} = \tilde{x}^k$, and $g(x^{k+1}) = g(\tilde{x}^k)$ for the next iteration. We call the method Self-adaptive single step projected gradient method because it needs only once evaluation of the gradient $g(x^k)$ in each iteration when adjusting the parameter β_k is not necessary.*

Remark 2.3 *If $r_k > \nu$, we adjust the parameter β_k by $\beta_k := \beta_k * 0.8/r_k$. According to our limited numerical experiments, using the reduced β_k, the condition $r_k \leq \nu$ is satisfied.*

Remark 2.4 *Too small step size β_k will leads to slow convergence. If $r_k \leq \mu$, we will enlarge the trial step size β for the by $\beta_k := \beta_k * 1.5$.*

3 经济平衡问题上的一个应用

作为单步投影梯度法的应用, 我们讨论第一讲 §2 提到的保护资源、保障供给互补问题的求解方法. 读者可以从第一讲的 §2 中更好地了解问题的背景. 对所有的 $i = 1, \ldots, m,\ j = 1, \ldots, n$, 使用记号：

S_i: 该种商品的第 i 个资源地;

D_j: 该种商品的第 j 个需求地;

x_{ij}: 从 S_i 到 D_j 的交易量;

s_i: 经营者们在资源地 S_i 的总采购量, $s_i = \sum_{j=1}^{n} x_{ij}$;

d_j: 经营者在需求地 D_j 的总销售量, $d_j = \sum_{i=1}^{m} x_{ij}$;

h_i^s: 经营者在资源地 S_i 处的采购价;

h_j^d: 经营者在需求地 D_j 处的销售价;

t_{ij}: 从 S_i 到 D_j 的交易费用(包括运输费用);

y_i: 政府为避免资源过度开采而在资源地 S_i 向经营者征收的资源税;

z_j: 政府为保障供给而在需求地 D_j 给经营者的经营补贴.

3.1 经营者追求利益最大化之互补问题

显然, 为了获利, 如果 S_i 处的采购价、资源税及从 S_i 到 D_j 的交易费的和 $(h_i^s + y_i + t_{ij})$ 不小于需求地 D_j 处的销售价与政府补贴的和 $(h_j^d + z_j)$, 经营者是不会从 S_i 采购商品运到 D_j 销售的. 反之, 根据贪婪原理, 经营者会尽可能增大经营量, 直到 $(h_i^s + y_i + t_{ij})$ 和 $(h_j^d + z_j)$ 相等. 这种关系的数学表达式是下面的平衡问题:

$$h_i^s + y_i + t_{ij} \begin{cases} \geq h_j^d + z_j, & \text{if } x_{ij} = 0, \\ = h_j^d + z_j, & \text{if } x_{ij} > 0\,. \end{cases} \tag{3.1}$$

它可以写成互补问题的形式: 对任意的 $i = 1, \ldots, m$ 和 $j = 1, \ldots, n$, 都有

$$0 \leq x_{ij} \ \perp \ \big((h_i^s + y_i + t_{ij}) - (h_j^d + z_j)\big) \geq 0. \tag{3.2}$$

通常假设 S_i 处的采购价 f_i 仅与经营者们在该地的采购量 s_i 有关; D_j 处的销售价 g_j 仅与经营者们运到该地的销售量 d_j 有关; 从 S_i 到 D_j 的交易价仅与它们之间的交易量 x_{ij}有关. 对于 $y = 0$ 和 $z = 0$ 的互补问题 (3.2), 文献中 [12] 称之为空间价格平衡问题 (Spatial Price Equilibrium Problem). 已有的求解空

间价格平衡问题的方法[?, 8, 9] 都对函数 h^s, h^d 和 t 的表达式有一定的要求. 当 a_i, b_j, c_{ij}, ξ_i, η_j, ζ_{ij} 为常数,

$$h_i^s(s_i) = \xi_i + a_i s_i, \qquad a_i \geq 0; \tag{3.3a}$$

$$h_j^d(d_j) = \eta_j - b_j d_j, \qquad b_j \geq 0; \tag{3.3b}$$

$$t_{ij}(x_{ij}) = \zeta_{ij} + c_{ij} x_{ij}, \qquad c_{ij} \geq 0 \tag{3.3c}$$

时, (3.2) 就是一个单调对称线性互补问题, 可以用求解带非负约束凸二次规划的方法求解 [3]. 在实际生活中, 这个问题是由市场根据贪婪原理自行解决的.

3.2 保护资源和保障供给的经济平衡问题

在现实生活中, 对(3.2) 中给定的 $y \geq 0$ 和 $z \geq 0$, 经济规律这只“无形的手”会让经营者们根据贪婪原理找到相应的 (x, s, d), 它是变分不等式(3.2) 的解. 换句话说, 原问题的解可由经营者们在经营活动中自行给出. 我们通常只知道函数 f, g 和 t 的一些性质而不知道它们的具体表达式, 能够观察到的是这个依赖于给定 y 和 z 的 s 与 d, 由于它是经营者根据贪婪原理给出的, 因此不可能顾及保护资源和保障供给. 假设从可持续发展的要求允许的最大资源消耗

量是$s^{\max}$, 为保障供给而必需的最小供应量是$d^{\min}$, 它们之间满足相容关系

$$\sum_{i=1}^{m} s_i^{\max} \geq \sum_{j=1}^{n} d_j^{\min}.$$

职能部门要采取经济手段来保证

$$s \leq s^{\max} \quad \text{和} \quad d \geq d^{\min}. \tag{3.4}$$

具体说来，在过度热销的资源地 S_i 通过对经营者征收资源税 y_i 保护资源，对供应不足的需求地 D_j 通过给经营者补贴 z_j 而吸引经营者增加供给. 我们的任务则是用数学方法帮助职能部门找到最优税收向量 $y^* \in \Re_+^m$ 和最优补贴向量 $z^* \in \Re_+^n$, 使得对我们给出的 (y^*, z^*) 以及经营者们由此产生的空间价格平衡问题 (3.2) 的解中的 s^* 和 d^* 满足

$$y^* \geq 0, \quad s^{\max} - s^* \geq 0, \quad {y^*}^T(s^{\max} - s^*) = 0, \tag{3.5}$$

和

$$z^* \geq 0, \quad d^* - d^{\min} \geq 0, \quad {z^*}^T(d^* - d^{\min}) = 0. \tag{3.6}$$

记

$$u = \begin{pmatrix} y \\ z \end{pmatrix} \quad \text{和} \quad F(u) = \begin{pmatrix} s^{\max} - s(u) \\ d(u) - d^{\min} \end{pmatrix}. \tag{3.7}$$

根据以上分析, 我们的任务可以归结为求解以下的隐式互补问题

$$u \geq 0, \quad F(u) \geq 0, \quad u^T F(u) = 0. \tag{3.8}$$

这里我们所说的'隐式' 是指我们不知道函数 F 的显式表达式而只能对给定的 $u \geq 0$ 观察到 $F(u)$ 的值. 本文的隐式互补问题是带有附加约束 (3.4) 的空间价格平衡问题 (3.2) 的对偶问题. 下面我们对函数 t_{ij}, h_i^s 和 h_j^d 作一定的假设后讨论隐式互补问题 (3.8) 的性质.

假设

A1. 对任意的 $i = 1, \ldots, m$ 和 $j = 1, \ldots, n$, 交易费用 $t_{ij}(x_{ij})$ 是交易量 x_{ij} 的非减函数.

A2. h_i^s 和 h_j^d 分别是 s_i 和 d_j 的一致严格增和一致严格减函数.

这样的假设应该说是合理的. 原因是由于道路拥挤, 单位交易费用(其中大部分为运输费用) 不会因交易量增加而减小, 资源地的采购价会因 "采购量"

的增大(货俏)而被生产者提高, 需求地的销售价会随“到货量”的增加而降低. 由以上的假设, 隐式互补问题 (3.8) 中的向量函数 $F(u)$ 是单调和 Lipschitz 连续的.

3.3 求解隐式互补问题的直接迭代方法

对映射 F 是单调和 Lipschitz 连续的隐式互补问题 (3.8), 文献中已有的在迭代过程中只用到 $F(u)$ 的值的方法, 主要是外梯度法 (Extra-gradient Method) 和投影收缩算法 (Projection and Contraction Method). 我们分别在第二讲和第三讲中已经做了介绍. 对于这一节讨论的隐式互补问题, 实际问题也只为我们提供了对应于自变量 $u \geq 0$ 的函数值 $F(u)$ 这样的信息.

由于每调用一次函数值就等同要进行一次税收和补贴政策的调整，实际问题要求我们在求解过程中尽可能减少调用函数值的次数.

♣ 注意到预测过程中至少调用二次函数值, 分别调用 $F(u^k)$ 和 $F(\tilde{u}^k)$ 以观察 $\tilde{u}^k$ 能否被接受为预测点. 在 β_k 选取适当, 进行了一次试探 $\tilde{u}^k$ 就被接受为预测点的情况时则预测过程恰好调用了二次函数值.

3.4 初步的数值试验情况

我们感兴趣的是用 §4 中介绍的单步投影梯度法, 求解保护资源和保障供给的经济平衡问题 (3.8). 为此, 对给定的 $y \in R_+^m$ 和 $z \in R_+^n$, 经营者们要解一个原问题 (3.2). 我们对原问题 (3.2), 按照 (3.3) 中的方式定义函数 h^s, h^d 和 t. 设 $m = 20$, $n = 50$. 对 $i = 1, \ldots, m$ 和 $j = 1, \ldots, n$, 取

$$a_i \in (1, 2), \qquad b_j \in (1, 2), \qquad c_{ij} \in (0.002, 0.005),$$
$$\xi_i \in (300, 400), \qquad \eta_j \in (600, 700), \qquad \zeta_{ij} \in (10, 20),$$

为一定范围中的随机数. 这样, 对给定的 y 和 z, 经营者们解的原问题 (3.2) 是一个单调对称的线性互补问题, 相当于变量个数为 1000 的带非负约束的凸二次规划. 这个本该由经营者们求解的问题, 也可用 §3 中介绍的方法去求解.

为保护资源和保障供给, 分别取 $s^{\max}$ 和 $d^{\min}$ 的每个分量均为 150 和 40. 我们要求解的隐式对偶问题的变量个数是 70. 计算以 $u^0 = 0$ 为初始点. 由于求解互补问题 (3.8) 时可以将 $\|\min\{u^k, F(u^k)\}\|_\infty$ 作为误差的一种度量, 我们在下表中给出对精度 $\|\min\{u^k, F(u^k)\}\|_\infty$ 不同要求时, 不同方法所需要的迭代次数和调用函数值 $F(u)$ 的次数.

不同方法达到同样精度要求的迭代次数和调用 $F(u)$ 的次数

误差精度	外梯度方法		投影收缩算法D1		投影收缩算法D2		投影梯度法	
$\|e(u)\|_\infty$	迭代次数	调用梯度次数	迭代次数.	调用梯度次数	迭代次数	调用梯度次数	迭代次数	调用梯度次数
1	35	72	5	13	4	10	3	4
0.1	53	108	6	15	6	14	5	6
0.01	64	130	8	19	7	16	8	9
0.001	71	144	12	27	10	20	10	11

♣ 由于采用了精化的调比产生 β_k 的策略, 外梯度方法和投影收缩算法的每次迭代几乎只调用 2 次函数值. 就像在第二, 第三讲分析的, 投影收缩算法比外梯度方法收敛快. 外梯度方法的迭代次数和调用 $F(u)$ 的次数都比投影收缩算法的 6–7 倍.

♣ 在投影收缩算法中, 采用 d_2 方向的比采用 d_1 方向的快一些.

♣ 投影梯度法需要的迭代次数与投影收缩算法相当, 但调用 $F(u)$ 的次数比投影收缩算法少一半以上. 求解这类问题, 调用一次 $F(u)$, 相当于调整一次政策. 因此, 在上述算法中, 投影梯度法效率是最高的.

♣ 对变分不等式使用投影梯度法, 要求算子 F 是某个凸函数的梯度.

4 An accelerated two-steps P-G method

基于 Nesterov [10] 的思想, 采用 [1] 类似的做法, 可以构造一个只用梯度的快速算法. 这类算法在生成序列 $\{x^k\}$ 的同时, 还生成一个辅助序列 $\{y^k\}$.

A two-steps projected gradient method

Take $\beta > 0$, $x^1 \in R^n$. Set $y^1 = x^1$, $t_1 = 1$.

Step k. ($k \geq 1$) With given (x^k, y^k), let

$$x^{k+1} = P_\Omega[y^k - \beta_k g(y^k)], \tag{4.1a}$$

where the step size β_k is chosen to satisfy

$$(y^k - x^{k+1})^T (g(y^k) - g(x^{k+1})) \leq \frac{1}{2\beta_k} \|y^k - x^{k+1}\|^2. \tag{4.1b}$$

Set

$$y^{k+1} = x^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(x^{k+1} - x^k), \tag{4.1c}$$

where

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}. \tag{4.1d}$$

The method is called two-steps projected gradient method because each iteration consists of two steps. The k-th iteration begins with (x^k, y^k), the first step (4.1a) produces x^{k+1} and the second one (4.1c) updates y^{k+1}.

It is assumed that the positive sequence $\{\beta_k\}$ is non-increasing.

We show that the proposed two-steps projected gradient method is convergent with the iteration-complexity $O(1/k^2)$. The proof is similar as those in [1].

Lemma 4.1 *Let x^{k+1} be given by* (4.1a) *and the step size condition* (4.1b) *be satisfied. Then we have*

$$2\beta_k(f(x)-f(x^{k+1})) \geq \|y^k-x^{k+1}\|^2+2(x^{k+1}-y^k)^T(y^k-x),\ \forall\, x \in \Omega. \quad (4.2)$$

Proof. By setting $x^k = y^k$ and $\nu = \frac{1}{2}$ in (2.4a) and (2.4b), we get (4.1a) and (4.1b). Therefore, substituting $x^k = y^k$ and $\nu = \frac{1}{2}$ in (2.6), we get

$$\beta_k(f(x) - f(x^{k+1})) \geq (x - x^{k+1})^T(y^k - x^{k+1}) - \frac{1}{2}\|y^k - x^{k+1}\|^2,\ \forall\, x \in \Omega.$$

The above inequality can be rewritten as (4.2) and the lemma is proved. □

To derive the iteration-complexity of the two-steps projected gradient method, we need to

prove some properties of the corresponding sequence.

Lemma 4.2 *The sequences* $\{x^k\}$ *and* $\{y^k\}$ *generated by the proposed two-steps projected gradient method satisfy*

$$2\beta_k t_k^2 v_k - 2\beta_{k+1} t_{k+1}^2 v_{k+1} \geq \|u^{k+1}\|^2 - \|u^k\|^2, \ \forall k \geq 1, \tag{4.3}$$

where $v_k := f(x^{k+1}) - f(x^*)$ *and* $u^k := t_k x^{k+1} - (t_k - 1)x^k - x^*$.

Proof. By using Lemma 4.1 for $k+1$, $x = x^{k+1}$ and $x = x^*$ we get

$$2\beta_{k+1}\big(f(x^{k+1}) - f(x^{k+2})\big) \geq \|y^{k+1} - x^{k+2}\|^2 + 2(x^{k+2} - y^{k+1})^T(y^{k+1} - x^{k+1}),$$

and

$$2\beta_{k+1}\big(f(x^*) - f(x^{k+2})\big) \geq \|y^{k+1} - x^{k+2}\|^2 + 2(x^{k+2} - y^{k+1})^T(y^{k+1} - x^*).$$

Using the definition of v_k, we get

$$2\beta_{k+1}(v_k - v_{k+1}) \geq \|y^{k+1} - x^{k+2}\|^2 + 2(x^{k+1} - y^{k+1})^T(y^{k+1} - x^{k+2}), \tag{4.4}$$

and

$$-2\beta_{k+1} v_{k+1} \geq \|y^{k+1} - x^{k+2}\|^2 + 2(x^* - y^{k+1})^T(y^{k+1} - x^{k+2}). \tag{4.5}$$

To get a relation between v_k and v_{k+1}, we multiply (4.4) by $(t_{k+1}-1)$ and add it to (4.5):

$$2\beta_{k+1}\big((t_{k+1}-1)v_k - t_{k+1}v_{k+1}\big)$$
$$\geq\ t_{k+1}\|x^{k+2}- y^{k+1}\|^2 + 2(x^{k+2}- y^{k+1})^T\big(t_{k+1}y^{k+1}- (t_{k+1}-1)x^{k+1}- x^*\big).$$

Multiplying the last inequality by t_{k+1} and using

$$t_k^2 = t_{k+1}^2 - t_{k+1} \qquad \big(\text{and thus} \quad t_{k+1} = (1+\sqrt{1+4t_k^2})/2 \quad \text{as in} \quad (4.1\text{d})\big),$$

which yields

$$\begin{aligned} &2\beta_{k+1}\big(t_k^2 v_k - t_{k+1}^2 v_{k+1}\big) \\ &\quad\geq\quad \|t_{k+1}(x^{k+2}-y^{k+1})\|^2 \\ &\qquad\quad +2t_{k+1}(x^{k+2}-y^{k+1})^T\big(t_{k+1}y^{k+1}-(t_{k+1}-1)x^{k+1}-x^*\big). \end{aligned}$$

Applying the relation

$$\|a-b\|^2 + 2(a-b)^T(b-c) = \|a-c\|^2 - \|b-c\|^2$$

to the right-hand side of the last inequality with

$$a := t_{k+1}x^{k+2}, \qquad b := t_{k+1}y^{k+1}, \qquad c := (t_{k+1}-1)x^{k+1} + x^*,$$

and using the fact $2\beta_k t_k^2 v_k \geq 2\beta_{k+1} t_k^2 v_k$ (since $\{\beta_k\}$ is non-increasing), we get

$$\begin{aligned} & 2\beta_k t_k^2 v_k - 2\beta_{k+1} t_{k+1}^2 v_{k+1} \\ \geq \quad & \|t_{k+1} x^{k+2} - (t_{k+1} - 1) x^{k+1} - x^*\|^2 \\ & - \|t_{k+1} y^{k+1} - (t_{k+1} - 1) x^{k+1} - x^*\|^2. \end{aligned}$$

In order to write the above inequality in the form (4.3) with

$$u^k = t_k x^{k+1} - (t_k - 1) x^k - x^*,$$

we need only to set

$$t_{k+1} y^{k+1} - (t_{k+1} - 1) x^{k+1} - x^* = t_k x^{k+1} - (t_k - 1) x^k - x^*.$$

From the last equality we obtain

$$y^{k+1} = x^{k+1} + \Big(\frac{t_k - 1}{t_{k+1}}\Big)(x^{k+1} - x^k).$$

This is just the form (4.1c) in the accelerated two-steps version of the projected gradient method. $\square$

To proceed the proof of the main theorem, we need the following Lemma 4.3 and Lemma

4.4, which have also been considered in [1]. We omit their proofs as they are trivial.

Lemma 4.3 *Let $\{a_k\}$ and $\{b_k\}$ be positive sequences of reals satisfying*

$$a_k - a_{k+1} \geq b_{k+1} - b_k \quad \forall\, k \geq 1.$$

Then, $a_k \leq a_1 + b_1$ for every $k \geq 1$.

Lemma 4.4 *The positive sequence $\{t_k\}$ generated by*

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}, \qquad \textit{with} \qquad t_1 = 1$$

satisfies

$$t_k \geq \frac{k+1}{2}, \quad \forall\, k \geq 1.$$

Now, we are ready to show that the proposed two-steps projected gradient method is convergent with the rate $O(1/k^2)$.

Theorem 4.1 *Let $\{x^k\}$ and $\{y^k\}$ be generated by the proposed two-steps projected gradient method. Then, for any $k \geq 1$, we have*

$$f(x^k) - f(x^*) \leq \frac{2\|x^1 - x^*\|^2}{\beta_k k^2}, \quad \forall\, x^* \in \Omega^*. \tag{4.6}$$

Proof. Let us define the quantities

$$a_k := 2\beta_k t_k^2 v_k, \qquad b_k := \|u^k\|^2.$$

By using Lemma 4.2 and Lemma 4.3, we obtain

$$2\beta_k t_k^2 v_k \leq a_1 + b_1,$$

which combined with the definition v_k and $t_k \geq (k+1)/2$ (by Lemma 4.4) yields

$$f(x^{k+1}) - f(x^*) = v_k \leq \frac{2(a_1+b_1)}{\beta_k (k+1)^2} \leq \frac{2(a_1+b_1)}{\beta_{k+1}(k+1)^2}. \tag{4.7}$$

Since $t_1 = 1$, and using the definition of u_k given in Lemma 4.2, we have

$$a_1 = 2\beta_1 t_1^2 v_1 = 2\beta_1 v_1 = 2\beta_1 \big(f(x^2) - f(x^*)\big), \qquad b_1 = \|u^1\|^2 = \|x^2 - x^*\|^2.$$

Setting $x = x^*$ and $k = 1$ in (4.2), we have

$$\begin{aligned} 2\beta_1 (f(x^2) - f(x^*)) &\leq 2(y^1 - x^*)^T (y^1 - x^2) - \|y^1 - x^2\|^2 \\ &= \|y^1 - x^*\|^2 - \|x^2 - x^*\|^2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} a_1 + b_1 &= 2\beta_1(f(x^2) - f(x^*)) + \|x^2 - x^*\|^2 \\ &\leq \|y^1 - x^*\|^2 - \|x^2 - x^*\|^2 + \|x^2 - x^*\|^2 \\ &= \|x^1 - x^*\|^2. \end{aligned}$$

Substituting it in (4.7), the assertion is proved. □

Based on Theorem 2.2, for obtaining an ε-optimal solution (denoted by $\tilde{x}$) in the sense that $f(\tilde{x}) - f(x^*) \leq \varepsilon$, the number of iterations required by the proposed two-steps projected gradient method is at most $\lceil C/\sqrt{\varepsilon} - 1 \rceil$ where $C = 2\|x^1 - x^*\|^2/\beta$.

需要说明的是, 对这一讲 §5 中的问题, 我们并不提倡用这个附录中的快速方法, 原因是在 (4.1) 的 k-次迭代中, 需要至少用到两次梯度的信息, $g(y^k)$ 和 $g(x^{k+1})$. 这里的 $g(\cdot)$ 相当于 §5 中的 $F(\cdot)$. 在实际问题中, $F(\cdot)$ 的获取往往是代价不菲的. 此外, (4.1b) 中要求

$$(y^k - x^{k+1})^T(g(y^k) - g(x^{k+1})) \leq \tfrac{1}{2\beta_k}\|y^k - x^{k+1}\|^2,$$

并要求 $\{\beta_k\}$ 单调不增. 这些条件远不如单步投影梯度法中相应的条件 (2.4b) 宽松. 单步投影梯度法计算实践说明限制 $\{\beta_k\}$ 单调不增会使收敛变慢许多.

References

[1] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Imaging Science, 2 (2009), pp. 183-202.

[2] Y.H. Dai and Y. Yuan, Alternate minimization gradient method. *IMA J. Numerical Analysis* **23** (2003) 377–393.

[3] R. Fletcher, *Practical Methods of Optimization*, Second Edition, John Wiley $\&$ Sons, 1987.

[4] B.S. He, *A new method for a class of linear variational inequalities*, Math. Progr., **66**, pp. 137-144, 1994.

[5] B.S. He, *Solving a class of linear projection equations*, Numerische Mathematik, **68**, pp. 71-80, 1994.

[6] B.S He and L-Z Liao, *Improvements of some projection methods for monotone nonlinear variational inequalities*, Journal of Optimization Theory and Applications, **112**, pp. 111-128, 2002

[7] G. M. Korpelevich, The Extragradient Method for Finding Saddle Points and Other Problems, *Ekonomika i Matematchskie Metody* **12** (1976) 747-756.

[8] P. Marcotte, G. Margquis and L. Zubieta, *A Newton-SOR method for spatial price equilibrium*, Transportation Science **26** (1992) 36-47.

[9] A. Nagurney, *An Algorithm for the single commodity spatial price equilibrium problem*, Reg. Sci. and Urban Econ. **16** (1986) 573-588.

[10] Y. E. Nesterov, A method for solving the convex programming problem with convergence rate $O(1/k^2)$, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543-547.

[11] Y. E. Nesterov, On an approach to the construction of optimal methods ofminimization ofsmooth convex functions, Ekonom. Mat. Metody, 24 (1988), pp. 509-517.

[12] P. A. Samuelson, *Spatial price equilibrium and linear programming*, Amer. Ecom. Rev. **42** (1952) 283-303.