Frank-Wolfe算法在目标函数为非凸，凸以及强凸情形下收敛性总结

2020.8.15

§1 Frank-Wolfe算法简介 [1]

Frank-Wolfe(FW)算法是众多非线性约束中的方法之一，近年来由于内存要求较小和无投影迭代的特点得到了广泛的应用。它可以解决如下的问题：

$$\underset{\boldsymbol{x}\in\mathcal{D}}{\text{minimize}}\, f(\boldsymbol{x}) \tag{1}$$

其中f可微并且是L – Lipschitz的，定义域$\mathcal{D}$是凸紧集。

FW算法首先设置一个初始点$\boldsymbol{x}_0$，并由此构造了一个迭代序列$\boldsymbol{x}_1, \boldsymbol{x}_2, ...$，这个序列可以收敛到问题的最优解。FW 算法的过程如下：

算法 1 FW Algorithm

1: **Input**:initial guess $\boldsymbol{x}_0$,tolerance $\delta > 0$
2: **for** $t = 0, 1, ...$ **do**
3: $\quad \boldsymbol{s}_t \in \underset{\mathbf{s}\in\mathcal{D}}{\text{argmax}} \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}\rangle$
4: $\quad \boldsymbol{d}_t = \boldsymbol{s}_t - \boldsymbol{x}_t$
5: $\quad g_t = \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{d}_t\rangle$
6: $\quad$ **if** $g_t < \delta$ **then**
7: $\quad\quad$ Return $\boldsymbol{x}_t$
8: $\quad$ **end if**
9: $\quad$ **Variant 1**:设置步长 $\gamma_t = \min\left\{\frac{\mathrm{g_t}}{\mathrm{L}\|\boldsymbol{d}_\mathrm{t}\|^2}, 1\right\}$
10: $\quad$ **Variant 2**:通过线搜索设置步长 $\gamma_t = \underset{\gamma\in[0,1]}{\text{argmin}}\, f(\boldsymbol{x}_t + \gamma \boldsymbol{d}_t)$
11: $\quad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$
12: **end for**
13: **Return** $\boldsymbol{x}_{t+1}$

与其他约束优化算法(如投影梯度下降法)相反，FW 算法不需要投影，因此有时称之为无投影算法，但需要求解**算法1**第3 步中的问题：

$$\boldsymbol{s}_t \in \underset{\boldsymbol{s}\in\mathcal{D}}{\text{argmin}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}\rangle \tag{2}$$

算法的其余部分主要是寻找合适的步长，以便$f(\boldsymbol{x})$ 朝减小的方向移动。在FW 算法所允许的许多不同的步长规则中，我们将其分为变形1 和变形2。第1 个变形很容易计算，只依赖于Lipschitz常数L 的知识。第2 个变形需要在每次迭代时解决一个一维问题。在某些情况下如当目标函数 f 为最小二乘问题且约束集$\mathcal{D}$ 是ℓ_1 范数球时，步长选择这个问题有一个显式的最优解，在这种情况下，应首选变型2 这种方法。但是，在一般情况下如果没有显式的最优解，在这种情况下，应首选变形1。事实证明，对于规模较大的问题的线性子问题既有闭式解，也有有效算法。与投影相比，使用线性最小化还有其他重要的结果。根据线性规划的性质，这种线性极小化问题的最优解总是约束集的一个顶点。还有一些其它设置步长大小的策略。例如，步长也可以设置为$\gamma_t = \frac{2}{t+2}$，它不依赖于优化产生的任何其它变量。因此，它在实践中不能与其它选择步长的策略竞争，尽管它确实达到了相同的理论收敛速度。

对于γ_t还有一种选择，是由Demyanov 和Rubinov 提出来的，与变形1比较相似：

$$\gamma_t = \min\Big\{\frac{\mathrm{g_t}}{\mathrm{Ldiam}(\mathcal{D})^2}, 1\Big\} \tag{3}$$

其中diam表示相对于欧几里德范数的直径。然而，由于根据直径的定义，我们总是有:

$$\|\boldsymbol{x}_t - \boldsymbol{s}_t\|^2 \leq \mathrm{diam}(\mathcal{D})^2 \tag{4}$$

这个变形提供的步长总是小于变形1的步长，并且给出了更差的收敛性。对该步长的进一步改进包括用局部估计替换Lipschitz常数L，从而允许更大的步长。

可以看到Frank-Wolfe算法是一种通过求解一系列线性规划问题来解决潜在非线性问题的算法，这种方法的有效性与快速解决线性子问题的能力紧密相连。这与其他一阶方法(如Nesterov的加速方法)形成了对比，后者的前提是在每次迭代中能够解决由强凸函数定义的某些投影问题。在许多应用中，求解一个线性优化子问题要比求解相关的投影子问题简单得多。此外，在许多应用中，线性优化子问题的解通常是高度结构化的，并表现出特殊的稀疏性或低秩性，Frank-Wolfe算法可以利用这一点。Frank-Wolfe算法在每次迭代求解一个子问题，并产生一系列可行解，这些可行解都是之前所有子问题解的凸组合，对于适当选择的步长，可以推导出$\mathcal{O}(\frac{1}{k})$的收敛速度。由于子问题解的结构以及迭代是子问题解的凸组合，Frank-Wolfe方法返回的可行解通常也是高度结构化的。例如，当可行域是单位单形$\Delta_n = \{\lambda \in \mathbb{R}^n : e^T\lambda = 1, \lambda > 0\}$ 时, 线性规划的解总是返回一个极值点，则Frank-Wolfe算法在第k 次迭代时产生的解最多有k个非零项。可推广到矩阵优化设置:如果可行域是核范数诱导的球，则在迭代k时，该方法在第k步产生的矩阵的秩最大不超过k。在许多应用中，这样的结构特性是非常可取的，在这种情况下，Frank-Wolfe方法可能比一些加速方法更有吸引力，尽管Frank-Wolfe方法的收敛速度较慢。

§2　收敛性分析

Frank-Wolfe算法在一般情况下收敛。正如已经看到的，凸的目标函数是必要的，以获得收敛保证。如前所述，假设f 在L-Lipschitz 梯度下是可微的，$\mathcal{D}$ 是一个凸紧集。在这一部分，将给出两个主要的收敛结果：一个用于一般目标，一个用于凸目标。为简单起见，假设线性子问题被精确地求解，但是这些证明可以很容易地扩展到考虑近似线性最小化的问题上去。本节的其余部分结构如下：首先介绍了两个关键定义和技术引理，最后证明了收敛性结果。

定义1. 称$\boldsymbol{x}^* \in \mathcal{D}$为最优点，当且仅当：

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0, \forall \boldsymbol{x} \in \mathcal{D}. \tag{5}$$

这个定义的直观含义是，如果以$\boldsymbol{x}^*$起点的多面体中的每个方向都与梯度正相关，则$\boldsymbol{x}^*$是一个最优点。否则，如果以$\boldsymbol{x}^*$起点没有可行的下降方向，$\boldsymbol{x}^*$是一个最优点。

定义2. 将Frank-Wolfe间隙记成g_t，定义为：

$$g_t = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{s}_t \rangle . \tag{6}$$

下一个引理将连续两次迭代的目标函数值联系起来，对于凸和非凸目标，它在证明收敛结果的过程中起到了关键作用。

引理1. 令$\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots$是由FW 算法产生的迭代序列，对于任意的$\xi \in [0,1]$，有下面的不等式：

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \xi g_t + \frac{1}{2}\xi^2 L \mathrm{diam}(\mathcal{D})^2 \tag{7}$$

证明. 关于f的Lipschitz梯度假设的一个结果是，$y \in \mathcal{D}$ 的每一点有上界，即：

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \tag{8}$$

应用这个不等式，对定义域中的任何$\boldsymbol{x}, \boldsymbol{y}$ 都有效。对于$\gamma \in [0,1]$且$\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{y} = (1-\gamma)\boldsymbol{x}_t + \gamma \boldsymbol{s}_t$的情况，$\boldsymbol{y}$ 仍然在定义域中，我们有：

$$f((1-\gamma)\boldsymbol{x}_t + \gamma \boldsymbol{s}_t) \leq f(\boldsymbol{x})_t + \gamma \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{x}_t \rangle + \frac{L\gamma^2}{2} \|\boldsymbol{s}_t - \boldsymbol{x}_t\|^2 \tag{9}$$

当$\gamma \in [0,1]$时，需要最小化上面公式的右边部分。这是一个最小化γ 的二次函数，将γ_t^*记为：

$$\gamma_t^* = \min\Big\{ \frac{\mathrm{g_t}}{\mathrm{L} \|\boldsymbol{x}_\mathrm{t} - \boldsymbol{s}_\mathrm{t}\|^2}, 1 \Big\}. \tag{10}$$

当$\gamma = \gamma_t^*$时，有：

$$\begin{aligned} f((1-\gamma_t^*)\boldsymbol{x}_t + \gamma_t^* \boldsymbol{s}_t) &\leq f(\boldsymbol{x}_t) - \gamma_t^* g_t + \frac{L(\gamma_t^*)^2}{2} \|\boldsymbol{s}_t - \boldsymbol{x}_t\|^2 \\ &= f(\boldsymbol{x}_t) + \min_{\xi \in [0,1]} \Big\{ -\xi \mathrm{g_t} + \frac{\mathrm{L}\xi^2}{2} \|\boldsymbol{s}_\mathrm{t} - \boldsymbol{x}_\mathrm{t}\|^2 \Big\} \\ &\leq f(\boldsymbol{x}_t) - \xi g_t + \frac{L\xi^2}{2} \|\boldsymbol{s}_t - \boldsymbol{x}_t\|^2 \quad (\forall \xi \in [0,1]) \\ &\leq f(\boldsymbol{x}_t) - \xi g_t + \frac{L\xi^2}{2} \mathrm{diam}(\mathcal{D})^2. \end{aligned} \tag{11}$$

上述不等式的右边已经包含引理中的结果，对于算法的变形1，我们有$f(\boldsymbol{x}_{t+1}) = f((1-\gamma_t^*)\boldsymbol{x}_t + \gamma_t^* \boldsymbol{s}_t)$，这种情况下$\gamma_t$ 和γ_t^* 相等。对于算法的变形2, 我们有$f(\boldsymbol{x}_{t+1}) \leq f((1-\gamma_t^*)\boldsymbol{x}_t + \gamma_t^* \boldsymbol{s}_t)$, 根据线搜索的定义，$f(\boldsymbol{x}_{t+1})$ 是最小化的目标函数值。因此，无论是哪种情况，我们都有：

$$f(\boldsymbol{x}_{t+1}) \leq f((1-\gamma_t^*)\boldsymbol{x}_t + \gamma_t^* \boldsymbol{s}_t). \tag{12}$$

将该式与前一个不等式联系起来，便得到要证明的不等式。 □

§2.1 目标函数非凸

下面是第一个收敛速度结果，对于梯度L-Lipschitz目标函数是有效的。

定理1. 如果f是可微的且具有L-Lipschitz 梯度，那么在最佳Frank-Wolfe间隙上有$\mathcal{O}(1/\sqrt{t})$的上界：

$$\min_{0\le i\le t} g_i \le \frac{\max\{2h_0, L*\mathrm{diam}(\mathcal{D})^2\}}{\sqrt{t+1}}, \tag{13}$$

其中$h_0 = f(\boldsymbol{x}_0) - \min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$是初始全局次优的。

证明. 根据引理1, 对于任意的$\xi \in [0,1]$有:

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\le f(\boldsymbol{x}_t) - \xi g_t + \frac{L\xi^2}{2}\mathrm{diam}(\mathcal{D})^2 \\ &\le f(\boldsymbol{x}_t) - \xi g_t + \frac{C\xi^2}{2} \end{aligned} \tag{14}$$

其中$C = L\mathrm{diam}(\mathcal{D})^2$. 考虑使右边公式最小化的$\xi$ 的值，可得$\xi^* = \min\{\mathrm{g_t/C}, 1\}$，现在根据$\xi^*$ 的值做出选择，如果$g_t \le C$,则$\xi^* = g_t/C$，利用上一个不等式中的值，得到：

$$f(\boldsymbol{x}_{t+1}) \le f(\boldsymbol{x}_t) - \frac{g_t^2}{2C} \tag{15}$$

如果$g_t > C$,则$\xi^* = 1$,有下面的不等式:

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) &\le f(\boldsymbol{x}_t) - g_t + \frac{C}{2} \\ &\le f(\boldsymbol{x}_t) - \frac{g_t}{2} \end{aligned} \tag{16}$$

结合两式得：

$$f(\boldsymbol{x}_{t+1}) \le f(\boldsymbol{x}_t) - \frac{g_t}{2}\min\Big\{\frac{g_t}{C}, 1\Big\} \tag{17}$$

其中$g_t^* = \min_{0\le i\le t} g_i$, 接下来对$g_t^*$ 进行选择，如果$g_t^* \le C$, 则$\min\{g_t^*/C, 1\} = g_t^*/C$解前面关于$g_t^*$的不等式得：

$$g_t^* \le \sqrt{\frac{2Ch_0}{t+1}} \le \frac{2h_0 + C}{2\sqrt{t+1}} \le \frac{\max\{2h_0, C\}}{\sqrt{t+1}} \tag{18}$$

在第二个不等式中利用Young不等式$ab \le \frac{a^2}{2} + \frac{b^2}{2}$, $a = \sqrt{2h_0}$, $b = \sqrt{C}$.

如果$g_t^* > C$，则$\min\{g_t^*/C, 1\} = 1$, 且有：

$$g_t^* \le \frac{2Ch_0}{t+1} \le \frac{2h_0}{\sqrt{t+1}} \le \frac{\max\{2h_0, C\}}{\sqrt{t+1}} \tag{19}$$

因此，在这两种情况下可得：

$$g_t^* \le \frac{\max\{2h_0, C\}}{\sqrt{t+1}} \tag{20}$$

□

§2.2 目标函数为凸

定理2. 如果f是可微凸函数，且梯度L-Lipschitz，则对于函数的次优性有如下的收敛速度:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{2L * \text{diam}(\mathcal{D})^2}{t+1} \tag{21}$$

证明. 由凸性可得:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \tag{22}$$

令$e_t = \boldsymbol{A}_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*))$, $\boldsymbol{A}_t$是正定的，$C = L\text{diam}(\mathcal{D})^2$，然后可以得到下面的不等式：

$$\begin{aligned} e_{t+1} - e_t &= \boldsymbol{A}_{t+1}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)) - \boldsymbol{A}_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \\ &\leq \boldsymbol{A}_{t+1}(f(\boldsymbol{x}_t) - \xi g_t + \frac{\xi^2 C}{2} - f(\boldsymbol{x}^*)) - \boldsymbol{A}_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \\ &\leq \boldsymbol{A}_{t+1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) - \xi(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) + \frac{\xi^2 C}{2}) - \boldsymbol{A}_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \\ &= ((1-\xi)\boldsymbol{A}_{t+1} - \boldsymbol{A}_t)(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) + \boldsymbol{A}_{t+1}\frac{\xi^2 C}{2} \end{aligned} \tag{23}$$

现在，选择 $\boldsymbol{A}_t = \frac{t(t+1)}{2}$, $\xi = 2/(t+2)$, 我们有：

$$(1-\xi)\boldsymbol{A}_{t+1} - \boldsymbol{A}_t = \frac{t(t+1)}{2} - \frac{t(t+1)}{2} = 0 \tag{24}$$

$$\boldsymbol{A}_{t+1}\frac{\xi^2}{2} = \frac{t+1}{t+2} \leq 1, \tag{25}$$

因此将式中的 $\boldsymbol{A}_t$ 和ξ替换, 则有：

$$e_{t+1} - e_t \leq C. \tag{26}$$

将不等式从0加到$t-1$，并使用$e_0 = 0$，可得：

$$e_t \leq tC \Longrightarrow f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{2C}{t+1}. \tag{27}$$

□

§2.3 目标函数为强凸

定义3. (光滑函数)设凸集$\mathcal{K} \subset \boldsymbol{E}$, 称函数$f : \boldsymbol{E} \to \mathcal{R}$ 是光滑函数, 当且仅当对于$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$有：

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x}) + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{28}$$

定义4. (强凸函数)设凸集$\mathcal{K} \subset \boldsymbol{E}$, 称函数$f : \boldsymbol{E} \to \mathcal{R}$ 是α强凸函数, 当且仅当满足以下两个条件：

- 1.$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}$:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

- 2.$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}, \gamma \in [0,1]$:

$$f(\gamma\boldsymbol{x} + (1-\gamma)\boldsymbol{y}) \leq \gamma f(\boldsymbol{x}) + (1-\gamma)f(\boldsymbol{y}) - \frac{\alpha}{2}\gamma(1-\gamma)\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

若f是强凸的，$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmax}} f(\boldsymbol{x})$, 由一阶最优性条件可得，对任意$\boldsymbol{x} \in \mathcal{K}$：

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 \tag{29}$$

稍作变换，有：

$$\sqrt{\frac{2}{\alpha}(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))} \cdot \|\nabla f(\boldsymbol{x})\|_* \geq \|\boldsymbol{x} - \boldsymbol{x}^*\| \cdot \|\nabla f(\boldsymbol{x})\|_* \tag{30}$$

$$\geq (\boldsymbol{x} - \boldsymbol{x}^*) \cdot \nabla f(\boldsymbol{x}) \tag{31}$$

$$\geq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \tag{32}$$

其中第二个不等式利用Holder不等式，第三个不等式来自f的凸性。因此对任意$\boldsymbol{x} \in \mathcal{K}$ 有：

$$\|\nabla f(\boldsymbol{x})\|_* \geq \sqrt{\frac{\alpha}{2}} \cdot \sqrt{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)} \tag{33}$$

定义5. (强凸集)称凸集$\mathcal{K} \subset \boldsymbol{E}$是$\alpha$强凸集, 当且仅当$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{K}, \forall \gamma \in [0,1]$, 以及$\boldsymbol{z} \in \boldsymbol{E}$ 且$\|\boldsymbol{z}\| = 1$, 有：

$$\gamma\boldsymbol{x} + (1-\gamma)\boldsymbol{y} + \gamma(1-\gamma)\frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2\boldsymbol{z} \in \mathcal{K}. \tag{34}$$

也就是说$\mathcal{K}$包含以$\gamma\boldsymbol{x} + (1-\gamma)\boldsymbol{y}$ 为中心，半径为$\gamma(1-\gamma)\frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ 的球.

接下来考虑函数f, f是β_f光滑的以及α_f强凸，我们进一步假设可行集$\mathcal{K}$是$\alpha_{\mathcal{K}}$强凸的。在$\boldsymbol{x}_t$处的逼近误差记为h_t, $h_t = \underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmin}} f(\boldsymbol{x})$, 其中$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmin}} f(\boldsymbol{x})$.

引理2. 使用FW算法，在第t步迭代有：

$$h_{t+1} \leq h_t \cdot \max\{\frac{1}{2}, 1 - \frac{\alpha_{\mathcal{K}}\|\nabla f(\mathbf{x}_t)\|_*}{8\beta_f}\} \tag{35}$$

证明. 由$\boldsymbol{s}_t$的取法得：

$$(\boldsymbol{s}_t - \boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t) \leq (\boldsymbol{x}^* - \boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t) \leq f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) = -h_t \tag{36}$$

第二个不等式是由于f的凸性。记$\boldsymbol{c}_t = \frac{1}{2}(\boldsymbol{x}_t + \boldsymbol{s}_t), \boldsymbol{w}_t \in \underset{\boldsymbol{w}\in\mathbf{E},\|\boldsymbol{w}\|\leq 1}{\operatorname{argmin}} \boldsymbol{w} \cdot \nabla f(\boldsymbol{x}_t)$. 由Holder不等式可得$\boldsymbol{w}_t \cdot \nabla f(\boldsymbol{x}_t) = -\|\nabla f(\boldsymbol{x}_t)\|_*$, 由可行集的强凸性，点$\tilde{\boldsymbol{p}}_t = \boldsymbol{c}_t + \frac{\alpha_{\mathcal{K}}}{8}\|\boldsymbol{x}_t - \boldsymbol{p}_t\|^2\boldsymbol{z}$ 也在可行集$\mathcal{K}$中。再次利用$\boldsymbol{s}_t$的最优性可得：

$$(\boldsymbol{s}_t - \boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t) \leq (\tilde{\boldsymbol{p}}_t - \boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t) \tag{37}$$

$$= \frac{1}{2}(\boldsymbol{p}_t - \boldsymbol{x}_t)\nabla f(\boldsymbol{x}_t) + \frac{\alpha_{\mathcal{K}}\|\boldsymbol{x}_t - \boldsymbol{p}_t\|^2}{8}\boldsymbol{w}_t \cdot \nabla f(\boldsymbol{x}_t) \tag{38}$$

$$\leq -\frac{1}{2}h_t - \frac{\alpha_{\mathcal{K}}\|\boldsymbol{x}_t - \boldsymbol{p}_t\|^2}{8}\|\nabla f(\boldsymbol{x}_t)\|_* \tag{39}$$

上面最后一个不等式来自式子(36)。现在分析下降时的误差，由于f 是光滑函数得：

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \gamma_t(\boldsymbol{p}_t - \boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t) + \frac{\beta_f}{2}\gamma_t^2\|\boldsymbol{p}_t - \boldsymbol{x}_t\|^2. \tag{40}$$

上式两边同时减去$f(\boldsymbol{x}^*)$得：

$$h(\boldsymbol{x}_{t+1}) \leq h(\boldsymbol{x}_t) + \gamma_t(\boldsymbol{p}_t - \boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t) + \frac{\beta_f}{2}\gamma_t^2 \|\boldsymbol{p}_t - \boldsymbol{x}_t\|^2. \tag{41}$$

结合(37)式有：

$$h_{t+1} \leq h_t(1 - \frac{\gamma_t}{2}) - \gamma_t \frac{\alpha_{\mathcal{K}} \|\boldsymbol{x}_t - \boldsymbol{p}_t\|^2}{8} \|\nabla f(\boldsymbol{x}_t)\|_* + \frac{\beta_f}{2}\gamma_t^2 \|\boldsymbol{p}_t - \boldsymbol{x}_t\|^2 \tag{42}$$

$$= h_t(1 - \frac{\gamma_t}{2}) + \frac{\|\boldsymbol{p}_t - \boldsymbol{x}_t\|^2}{2}\big(\gamma_t^2\beta - \gamma_t \frac{\alpha_{\mathcal{K}} \|\nabla f(\boldsymbol{x}_t)\|_*}{4}\big). \tag{43}$$

当$\frac{\alpha_{\mathcal{K}} \|\nabla f(\boldsymbol{x}_t)\|_*}{4} \geq \beta_f$, 令$\gamma_t = 1$可得：

$$h_{t+1} \leq \frac{h_t}{2}. \tag{44}$$

其他情况下，令$\gamma_t = \frac{\alpha_{\mathcal{K}} \|\nabla f(\boldsymbol{x}_t)\|_*}{4}$可得：

$$h_{t+1} \leq h_t\big(1 - \frac{\alpha_{\mathcal{K}} \|\nabla f(\boldsymbol{x}_t)\|_*}{8}\big). \tag{45}$$

□

注意到**引理**2只依赖于可行集$\mathcal{K}$的强凸性，不需要利用f的凸性与函数光滑的条件，也不要求f是强凸的。

定理3. 令$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmin}} f(\boldsymbol{x}), M = \frac{\sqrt{\alpha_f}\alpha_{\mathcal{K}}}{8\sqrt{2}\beta_f}$, 记$D_{\mathcal{K}} = \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{K}} \|\boldsymbol{x} - \boldsymbol{y}\|$,由FW算法得到的第$t$ 个迭代点$\boldsymbol{x}_t$满足：

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{\max\{\frac{9}{2}\beta_f D_{\mathcal{K}}^2, 18M^{-2}\}}{(t+2)^2} = \mathcal{O}(\frac{1}{t^2}) \tag{46}$$

证明. 令$M = \frac{\sqrt{\alpha_f}\alpha_{\mathcal{K}}}{8\sqrt{2}\beta_f}, C = \max\{\frac{9}{2}\beta_f D_{\mathcal{K}}^2, 18M^{-2}\}$. 接下来通过数学归纳法证明$\forall t > 1$, $h_t \leq \frac{C}{(t+2)^2}$.

由于目标函数满足式子(33), 由**引理**2 得：

$$h_{t+1} \leq h_t \cdot \max\{\frac{1}{2}, 1 - \frac{\alpha_{\mathcal{K}}\sqrt{\alpha_f}}{8\sqrt{2}\beta_f}\sqrt{h_t}\} \tag{47}$$

$$= h_t \cdot \max\{\frac{1}{2}, 1 - Mh_t^{1/2}\} \tag{48}$$

当$t = 1$, 由于f是β_f光滑的可得:

$$f(\boldsymbol{x}_1) - f(\boldsymbol{x}^*) = f(\boldsymbol{x}_0 + \gamma_0(\boldsymbol{p}_0 - \boldsymbol{x}_0)) - f(\boldsymbol{x}^*) \tag{49}$$

$$\leq h_0 + \gamma_0(\boldsymbol{p}_0 - \boldsymbol{x}_0) \cdot \nabla f(\boldsymbol{x}_0) + \frac{\beta_f\gamma_0^2}{2} D_{\mathcal{K}}^2 \tag{50}$$

$$\leq h_0(1 - \gamma_0) + \frac{\beta_f\gamma_0^2}{2} D_{\mathcal{K}}^2. \tag{51}$$

令$\gamma_0 = 1$得$h_1 \leq \frac{\beta_f}{2} D_{\mathcal{K}}^2 \leq C/9$. 故当$t = 1$, $h_1 = f(\boldsymbol{x}_1) - f(\boldsymbol{x}^*) \leq C/4$ 成立。
假设对于任意$t \geq 1$有$h_t \leq \frac{C}{(t+2)^2}$. 由式子(47)得：

$$h_{t+1} \leq \frac{h_t}{2} \leq \frac{C}{2(t+2)^2} = \frac{C}{(t+3)^2} \cdot \frac{(t+3)^2}{2(t+2)^2} \leq \frac{C}{(t+3)^2} \tag{52}$$

如果$h_t \le \frac{C}{2(t+3)^2}$. 由式子(52)得，对任意$t \ge 1$ 有：

$$h_{t+1} \le h_t \le \frac{C}{2(t+2)^2} \le \frac{C}{(t+3)^2}. \tag{53}$$

若$h_t \ge \frac{C}{2(t+2)^2}$, 由式子(52)得：

$$\begin{aligned}
h_{t+1} &\le h_t(1 - Mh_t^{1/2}) && (54)\\
&< \frac{C}{(t+2)^2}\big(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\big) && (55)\\
&= \frac{C}{(t+3)^2} \cdot \frac{(t+3)^2}{(t+2)^2}\big(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\big) && (56)\\
&= \frac{C}{(t+3)^2} \cdot \frac{(t+2)^2 + 2t + 5}{(t+2)^2}\big(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\big) && (57)\\
&< \frac{C}{(t+3)^2}\big(1 + \frac{3(t+2)}{(t+3)^2}\big)\big(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\big) && (58)\\
&= \frac{C}{(t+3)^2}\big(1 + \frac{3}{t+3}\big)\big(1 - M\sqrt{\frac{C}{2}}\frac{1}{t+2}\big) && (59)\\
& && (60)
\end{aligned}$$

故对于$C \ge \frac{18}{M^2}$可得：

$$h_{t+1} \le \frac{C}{(t+2)^2}\big(1 + \frac{3}{t+2}\big)\big(1 - \frac{3}{t+2}\big) < \frac{C}{(t+3)^2} \tag{61}$$

□

参考文献

[1] Fabian Pedregosa.Notes on the Frank-Wolfe Algorithm, Part I，http://fa.bianp.net/blog/2018/notes-on-the-frank-wolfe-algorithm-part-i/

[2] Dan Garber,Elad Hazan,Faster Rates for the Frank-Wolfe Method over Strongly-Convex Sets