优化算法复杂度分析（二）
光滑优化问题的复杂度分析

Hailiang Zhao
http://hliangzhao.me

2022 年 9 月 27 日

光滑优化问题

本 slide 中，我们将给出优化问题

$$f^* = \operatorname*{argmin}_{x \in X} f(x) \tag{1}$$

的复杂度分析，其中 $f(x)$ 是光滑的。

首先，我们将定义一些光滑函数的子集合，并列出它们的一些性质。

光滑函数子集合 $C_L^{k,p}(X)$

$\forall X \in \mathbb{R}^n$:

- 定义 $C_L^{k,p}(X)$ 为具有如下性质的函数集合:
  1. 任何函数 $f \in C_L^{k,p}(X)$ 在 X 上 k 次连续可微;
  2. 任何函数 $f \in C_L^{k,p}(X)$ 的 p 阶导数在 X 上 $Lipschitz$ 连续 (对于某常数 L):

$$\|f^{(p)}(x) - f^{(p)}(y)\| \leq L\|x - y\|, \forall x, y \in X. \tag{2}$$

- 定义 $\mathcal{F}_L^{k,p}(X)$ 为 $C_L^{k,p}(X)$ 和凸函数集合的交集。

光滑函数子集合的性质

性质 1：若 $f_1 \in C_{L_1}^{k,p}(X), f_2 \in C_{L_2}^{k,p}(X)$，且 $\alpha, \beta \in \mathbb{R}$，则对 $L_3 = |\alpha|L_1 + |\beta|L_2$ 有 $\alpha f_1 + \beta f_2 \in C_{L_3}^{k,p}(X)$.

性质 2：$f \in C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$ 当且仅当 $\forall x \in \mathbb{R}^n$，$\|\nabla^2 f(x)\| \leq L$。

性质 3：若 $f \in C_L^{1,1}(\mathbb{R}^n)$，则 $\forall x, y \in \mathbb{R}^n$，有

$$\left|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\right| \leq \frac{L}{2}\|y - x\|^2. \tag{3}$$

性质 4：若 $f \in C_L^{2,2}(\mathbb{R}^n)$，则 $\forall x, y \in \mathbb{R}^n$，$\exists M > 0$满足

$$\left\|\nabla f(y) - \nabla f(x) - \langle \nabla^2 f(x), y - x\rangle\right\| \leq \frac{M}{2}\|y - x\|^2, \tag{4}$$

$$\left|f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2}\langle \nabla^2 f(x)(y - x), y - x\rangle\right| \leq \frac{M}{6}\|y - x\|^3. \tag{5}$$

光滑函数子集合的性质

性质 5：令 $f \in C_L^{2,2}(\mathbb{R}^n)$ 且 $\|y-x\| = r$ 则$\exists M > 0$满足

$$\nabla^2 f(x) - MrI_n \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) + MrI_n. \quad (I_n\text{是单位阵}) \tag{6}$$

性质 6：若 $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$，则 $\forall x, y \in \mathbb{R}^n, \alpha \in [0,1]$ 有

$$0 \le f(y) - f(x) - \langle \nabla f(x), y-x \rangle \le \frac{L}{2}\|y-x\|^2, \tag{7}$$

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

$$\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \le \langle \nabla f(x) - \nabla f(y), x-y \rangle \le L\|x-y\|^2, \tag{9}$$

$$\begin{aligned} \alpha f(x) + (1-\alpha) f(y) &\le f(\alpha x + (1-\alpha)y) \\ &\quad + \frac{\alpha(1-\alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \end{aligned} \tag{10}$$

$$\begin{aligned} \alpha f(x) + (1-\alpha) f(y) &\le f(\alpha x + (1-\alpha)y) \\ &\quad + \alpha(1-\alpha)\frac{L}{2}\|x-y\|^2. \end{aligned} \tag{11}$$

梯度法

接下来我们分析不同算法在不同光滑函数子集合中的收敛性和复杂度，并给出不同光滑函数子集合的复杂度上界。

我们首先考虑梯度法在求解

$$\min_{x\in\mathbb{R}^n} f(x)$$

时的复杂度（即 $X=\mathbb{R}^n$）。

在梯度法中，新的迭代点通过如下方式得到：

$$x_{k+1}=x_k-h_k\nabla f(x_k), \tag{12}$$

其中 h_k 为步长。

梯度法

在梯度法中，新的迭代点通过如下方式得到：

$$x_{k+1} = x_k - h_k \nabla f(x_k), \tag{13}$$

其中步长 h_k 有三种选择方式：

- **固定步长和变步长策略**：$h_k = h > 0$ 或 $h_k = \frac{h}{\sqrt{k+1}}$；
- **精确先搜索步长法**：$h_k = \operatorname{argmin}_{h \geq 0} f\big(x_k - h\nabla f(x_k)\big)$；
- Goldenstein-Armijo **准则**：令 $x_{k+1} = x_k - h_k \nabla f(x_k)$，寻找满足以下不等式的 h_k：

$$\begin{aligned} \alpha\langle \nabla f(x_k), x_k - x_{k+1}\rangle &\leq f(x_k) - f(x_{k+1}) \\ \beta\langle \nabla f(x_k), x_k - x_{k+1}\rangle &\geq f(x_k) - f(x_{k+1}), \end{aligned}$$

其中 $0 < \alpha < \beta < 1$。

梯度法求解问题集合 $C_L^{1,1}(\mathbb{R}^n)$

问题模型 1

考虑无约束非凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:

- 全局信息 Σ: 目标函数 $f \in C_L^{1,1}(\mathbb{R}^n)$，约束集合 $X \equiv \mathbb{R}^n$，且 $f(x)$ 不一定是凸函数; $f(x)$ 有下界 $(\exists M \in \mathbb{R}[\forall x \in X, f(x) \geq M])$;
- 局部信息 $\mathcal{O}$: $\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回 $f(x_0)$ 和 $\nabla f(x_0)$;
- 解的精度 $\mathcal{T}_\epsilon$: 求局部极小值的近似解 $\bar{x} \in \mathbb{R}^n$，使得 $\|\nabla f(\bar{x})\| \leq \epsilon$。

梯度法求解问题集合 $C_L^{1,1}(\mathbb{R}^n)$

定理 1

使用梯度法求解问题模型 1 时，取 $\bar{x}$ 满足

$$\|\nabla f(\bar{x})\| = \min_{0\le k\le N} \|\nabla f(x_k)\|, \tag{14}$$

则算法的收敛速度为

$$\|\nabla f(\bar{x})\| \le \frac{1}{\sqrt{N+1}} \left[\frac{L}{\omega}\big(f(x_0) - f^*\big)\right]^{\frac{1}{2}}. \tag{15}$$

分析复杂度上界为 $\mathcal{O}\big(\frac{1}{\epsilon^2}\big)$:

$$N(\epsilon) \le \frac{L(f(x_0) - f^*)}{\omega\epsilon^2}. \tag{16}$$

证明.

结合光滑函数子集合的性质 3，可以发现三种步长设定方式均可得到 $f(x_k) - f(x_{k+1}) \ge \frac{\omega}{L}\|\nabla f(x_k)\|^2$。 □

梯度法求解问题集合 $C_L^{2,2}(\mathbb{R}^n)$

问题模型 2

考虑无约束非凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:

- 全局信息 Σ:
  - 目标函数 $f \in C_L^{2,2}(\mathbb{R}^n)$，约束集合 $X \equiv \mathbb{R}^n$，且 $f(x)$ 不一定是凸函数;
  - $f(x)$ 存在局部极小值点 x^*，且 $\nabla^2 f(x^*)$ 正定;
  - $\exists 0 < m \leq M < \infty \left[mI_n \preceq \nabla^2 f(x^*) \preceq MI_n \right]$;
  - 初始点 x_0 距离 x^* 足够近;
- 局部信息 $\mathcal{O}$: $\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回 $f(x_0)$ 和 $\nabla f(x_0)$;
- 解的精度 $\mathcal{T}_\epsilon$: 求局部极小值的近似解 $\bar{x} \in \mathbb{R}^n$，使得 $\|\nabla f(\bar{x})\| \leq \epsilon$。

梯度法求解问题集合 $C_L^{2,2}(\mathbb{R}^n)$

定理 2

假设梯度法的初始点 x_0 距离极小值点 x^* 足够近，满足

$$r_0 = \|x_0 - x^*\| < \bar{r} = \frac{2m}{L}, \tag{17}$$

且步长取 $h_k = \frac{2}{m+M}$，则梯度法求解问题模型 2 的收敛速率为**局部线性收敛**：

$$\|x_k - x^*\| \leq \frac{\bar{r} r_0}{\bar{r} - r_0} \left(1 - \frac{2m}{M + 3m}\right)^k. \tag{18}$$

复杂度上界为

$$\frac{M + 3m}{2m} \left[\ln\left(\frac{\bar{r} r_0}{\bar{r} - r_0}\right) + \ln\frac{1}{\epsilon}\right]. \tag{19}$$

梯度法求解问题集合 $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

问题模型 3

考虑优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:

- 全局信息 Σ: $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$，或 $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$，或 $f \in \mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$，或 $f \in \mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$。分别是强凸函数集合、弱强凸函数集合和二阶增长类函数集合;
- $\mathcal{FO}$ 子程序（或 $\mathcal{PO}$ 子程序）;
- 解的精度 $\mathcal{T}_\epsilon$: 求**全局**极小值的近似解 $\bar{x} \in \mathbb{R}^n$，使得 $f(\bar{x}) - f^* \leq \epsilon$ 或 $\|\bar{x} - x^*\| \leq \epsilon$。

梯度法求解问题集合 $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

定理 3

若 $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$，令步长 $h_k \equiv h = \frac{2}{L}$，则梯度法求解问题模型 3 的收敛速率为**全局次线性收敛**：

$$f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|^2}{k+4}. \tag{20}$$

记 $D_0 = \|x_0 - x^*\|$，则复杂度上界为

$$\mathcal{O}\Big(\frac{LD_0}{\epsilon}\Big). \tag{21}$$

梯度法求解问题集合 $\mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$

定理 4

若 $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$，令步长 $h_k \equiv h = \frac{1}{L}$，则梯度法求解问题模型 3 的收敛速率为**全局线性收敛**：

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2, \tag{22}$$

$$\|x_k - x^*\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2. \tag{23}$$

其中 $Q = \frac{L}{\mu}$。目标函数值对应的复杂度上界为

$$\frac{\log\left(\frac{LD_0^2}{\epsilon}\right)}{\log\left(\frac{Q-1}{Q+1}\right)^2}. \tag{24}$$

梯度法求解问题集合 $\mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$

定理 5

若 $f \in \mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$，令步长 $h_k \equiv h = \frac{1}{L}$，则梯度法求解问题模型 3 的收敛速率为**全局线性收敛**：

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{k-1} \|x_0 - \bar{x}^k\|^2, \tag{25}$$

$$\|x_k - \bar{x}^k\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^{k} \|x_0 - \bar{x}^k\|^2. \tag{26}$$

其中$\bar{x}^k = \operatorname{argmin}_{k' \in \{1,\ldots,k\}} f(x_{k'})$。目标函数值对应的复杂度上界为

$$\frac{\log\left(\frac{LD_0^2}{\epsilon}\right)}{\log\left(\frac{Q-1}{Q+1}\right)}. \tag{27}$$

梯度法求解问题集合 $\mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$

定理 6

若 $f \in \mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$，令步长 $h_k \equiv h = \frac{1}{L}$，则梯度法求解问题模型 3 的收敛速率为**全局线性收敛**：

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q}{Q+1}\right)^{k-1} \|x_0 - \bar{x}^k\|^2, \tag{28}$$

$$\|x_k - \bar{x}^k\|^2 \leq \left(\frac{Q}{Q+1}\right)^{k} \|x_0 - \bar{x}^k\|^2. \tag{29}$$

目标函数值对应的复杂度上界为

$$\frac{\log\left(\frac{LD_0^2}{\epsilon}\right)}{\log\left(\frac{Q}{Q+1}\right)}. \tag{30}$$

可以发现，$\mathcal{F}_{L,\mu}^{1,1,}(\mathbb{R}^n)$，$\mathcal{W}_{L,\mu}^{1,1,}(\mathbb{R}^n)$，$\mathcal{S}_{L,\mu}^{1,1,}(\mathbb{R}^n)$ 的复杂度上界依次递减。

牛顿法

在牛顿法中，新的迭代点通过如下方式得到：

$$x_{k+1} \leftarrow x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k). \tag{31}$$

问题模型 4

考虑无约束非凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$：

- 全局信息 Σ：
  - 目标函数 $f \in C_L^{2,2}(\mathbb{R}^n)$，约束集合 $X \equiv \mathbb{R}^n$，且 $f(x)$ 不一定是凸函数；
  - $f(x)$ 存在局部极小值点 x^*，且 $\nabla^2 f(x^*)$ 正定；
  - $\exists m > 0 \left[m I_n \preceq \nabla^2 f(x^*) \right]$；
  - 初始点 x_0 距离 x^* 足够近；
- 局部信息 $\mathcal{O}$：$2\text{nd}\mathcal{O}$ 子程序，对于任意给定的 x_0 返回 $f(x_0)$、$\nabla f(x_0)$ 和 $\nabla^2 f(x_0)$；
- 解的精度 $\mathcal{T}_\epsilon$：求局部极小值的近似解 $\bar{x} \in \mathbb{R}^n$，使得 $\|\bar{x} - x^*\| \leq \epsilon$。

牛顿法求解问题集合 $C_L^{2,2}(\mathbb{R}^n)$

定理 7

假设牛顿法的初始点 x_0 距离极小值点 x^* 足够近，满足

$$\|x_0 - x^*\| < \bar{r} = \frac{3m}{L}, \tag{32}$$

则对任意的 k 满足

$$\|x_k - x^*\| \leq \bar{r}. \tag{33}$$

牛顿法求解问题模型 4 的收敛速率为**局部二阶收敛**：

$$\|x_{k+1} - x^*\| \leq \frac{L\|x_k - x^*\|^2}{2(m - L\|x_k - x^*\|)}. \tag{34}$$

复杂度上界为 $c \ln \ln \frac{\gamma}{\epsilon}$，其中 c 和 γ 为常数。

光滑凸优化问题的复杂度下界

定理 8

对任意 $x_0 \in \mathbb{R}^n$ 和 $1 \leq k \leq \frac{1}{2}(n-1)$，存在 $f \in C_L^{\infty,1}(\mathbb{R}^n)$，使得对任意的 $x_k \in x_0 + \text{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}$（即一阶算法，只利用梯度信息的算法）都有

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}. \tag{35}$$

令 $D_0 = \|x_0 - x^*\|$，令上式右端等于 ϵ，可以得到一阶算法求解光滑凸优化问题的复杂度下界为

$$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}D_0\right). \tag{36}$$

定理 3 告诉我们，梯度法求解非光滑凸优化问题的复杂度上界为 $\mathcal{O}(\frac{LD_0}{\epsilon})$，相比之下可以发现梯度法并非最优算法。求解某一类问题集合的最优算法应当让其复杂度上界和下界尽可能接近。

光滑强凸优化问题的复杂度下界

定理 9

对任意 $x_0 \in \mathbb{R}^\infty, \mu, L > 0$ 和 $Q = \frac{L}{\mu} > 1$，存在 $f \in C^{\infty,1}_{L,\mu}(\mathbb{R}^\infty)$，使得对任意 $x_k \in x_0 + \text{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}$ 都有

$$f(x_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|^2. \tag{37}$$

令上式右端等于 ϵ，可以得到一阶算法求解光滑强凸优化问题的复杂度下界为

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \max\left(\log \frac{\mu D_0}{\epsilon}, 1\right)\right). \tag{38}$$

Nesterov 加速梯度算法框架

定理 8 告诉我们，函数集合 $C_L^{\infty,1}(\mathbb{R}^n)$ 相对于解算法集合 $x_k \in x_0 + \text{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}$ 的复杂度下界为 $\mathcal{O}(1/\sqrt{\epsilon})$，复杂度上界为 $\mathcal{O}(1/\epsilon)$。这意味着梯度法相对于该解算法集合而言并不是最优的。接下来我们介绍 Nesterov 加速梯度法。

Algorithm 1: Nesterov 加速梯度算法框架

Input: 令 $x_0 = y_0$，选取序列 $\{\gamma_k\}$ 和 $\{\beta_k\}$ 满足 $L\gamma_k \leq \beta_k$，且 $\gamma_1 = 1$

1 **for** $k = 1, 2, ..., N$ **do**

2 $\quad z_k \leftarrow (1 - \gamma_k)y_{k-1} + \gamma_k x_{k-1}$

3 $\quad x_k \leftarrow \text{argmin}_{x \in X}\{\langle \nabla f(z_k), x \rangle + \frac{\beta_k}{2}\|x - x_{k-1}\|_2^2\}$

4 $\quad y_k \leftarrow (1 - \gamma_k)y_{k-1} + \gamma_k x_k$

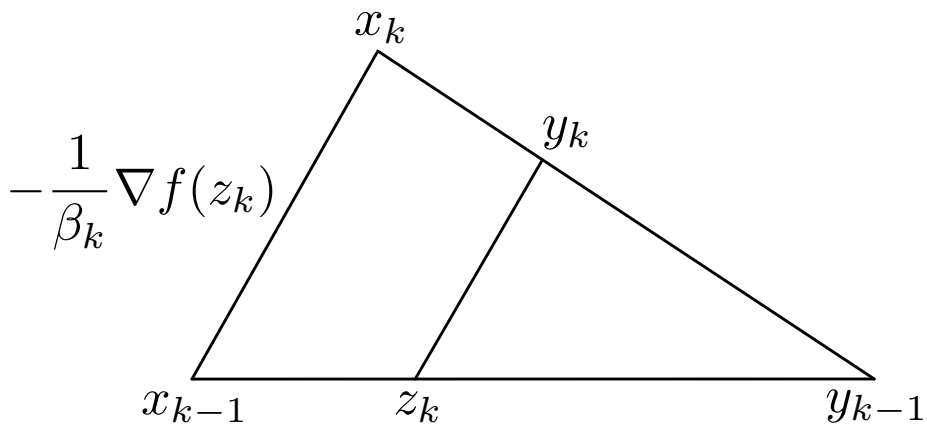
5 **end for**

Output: y_N

该算法中，收敛点列为 $\{y_k\}$。若 $X \equiv \mathbb{R}^n$，则第二步等价于 $x_k = x_{k-1} - \frac{1}{\beta_k}\nabla f(z_k)$（直接代入求出最优的 x）。

Nesterov 加速梯度算法框架

该算法中，收敛点列为 $\{y_k\}$。若 $X \equiv \mathbb{R}^n$，则第二步等价于 $x_k = x_{k-1} - \frac{1}{\beta_k}\nabla f(z_k)$（直接代入求出最优的 x）。此时，x_k、y_k 和 z_k 的迭代关系如下图所示：



注意到，这两个三角形是相似的，且相似比为 γ_k。

加速梯度算法的收敛速度

选取不同的参数 γ_k 和 β_k，可以得到不同的收敛速度。接下来，我们通过构造序列 $\{\Gamma_k\}$ 来分析收敛性。

引理 1

令 $\gamma_t \in (0,1], t = 1, 2, \ldots,$ 构造序列

$$\Gamma_t = \begin{cases} 1 & t = 1 \\ (1-\gamma_t)\Gamma_{t-1} & t \geq 2. \end{cases} \tag{39}$$

若序列 $\{\Delta_t\}_{t\geq 0}$ 满足

$$\Delta_t \leq (1-\gamma_t)\Delta_{t-1} + B_t, \quad t = 1, 2, \ldots, \tag{40}$$

则 $\forall k$ 有

$$\Delta_k \leq \Gamma_k(1-\gamma_1)\Delta_0 + \Gamma_k \sum_{t=1}^{k} \frac{B_t}{\Gamma_t}. \tag{41}$$

加速梯度算法的收敛速度

一般来说，我们会令 $\Delta_k = f(x_k) - f(x^*)$ 或 $\Delta_k = \|x_k - x^*\|_2^2$，由此，(40)变为

$$f(x_k) - f(x^*) \leq (1 - \gamma_k)\Big(f(x_{k-1}) - f(x^*)\Big) + B_k \tag{42}$$

或者

$$\|x_k - x^*\|_2^2 \leq (1 - \gamma_k)\|x_k - x^*\|_2^2 + B_k. \tag{43}$$

通常我们会构造序列 $\{\gamma_k\}$ 使其满足 $\gamma_1 = 1$，由此(41)变成

$$f(x_k) - f(x^*) \leq \Gamma_k \sum_{t=1}^{k} \frac{B_t}{\Gamma_t} \tag{44}$$

或

$$\|x_k - x^*\|_2^2 \leq \Gamma_k \sum_{t=1}^{k} \frac{B_t}{\Gamma_t}. \tag{45}$$

加速梯度算法在凸函数问题集合上的收敛速度

可以发现，加速梯度法的收敛速度与 B_k 和 Γ_k 有关。对于 B_k 我们通过放缩估计其上界，对于 γ_k 我们可以采用不同的构造方式。 例如，

$$\begin{aligned} \gamma_k &= \frac{1}{k} \quad &\Rightarrow \quad \Gamma_k &= \frac{1}{k} \\ \gamma_k &= \frac{1}{k+1} \quad &\Rightarrow \quad \Gamma_k &= \frac{2}{k(k+1)} \\ \gamma_k &= \frac{3}{k+2} \quad &\Rightarrow \quad \Gamma_k &= \frac{6}{k(k+1)(k+2)} \end{aligned} \tag{46}$$

令 $D_X = \sup_{x,y\in X} \|x-y\|$，下面给出加速梯度算法在求解问题模型 3 时的收敛速率。

定理 10

若 $f \in C_L^{1,1}(\mathbb{R}^n)$ 且是凸函数，则 Nesterov 加速梯度算法求解问题模型 3 的速率为

- 取 $\beta_k = L, \gamma_k = \frac{1}{k}$，则 $\Gamma_k = \frac{1}{k}, \frac{\beta_k\gamma_k}{\Gamma_k} = L$，我们有

$$f(y_k) - f(x^*) \le \frac{L}{2k} D_X^2, \quad f(y_k) - f(x^*) \le \frac{L}{2k}\|x_0 - x^*\|^2. \tag{47}$$

加速梯度算法在凸函数问题集合上的收敛速度

定理 10（续）

若 $f \in C_L^{1,1}(\mathbb{R}^n)$ 且是凸函数，则 Nesterov 加速梯度算法求解问题模型 3 的速率为

- 取 $\beta_k = \frac{2L}{k}, \gamma_k = \frac{2}{k+1}$，则 $\Gamma_k = \frac{2}{k(k+1)}, \frac{\beta_k \gamma_k}{\Gamma_k} = 2L$，我们有

$$f(y_k) - f(x^*) \leq \frac{2L}{k(k+1)} D_X^2, \tag{48}$$

$$f(y_k) - f(x^*) \leq \frac{4L}{k(k+1)} \|x_0 - x^*\|^2. \tag{49}$$

- 取 $\beta_k = \frac{3L}{k+1}, \gamma_k = \frac{3}{k+2}$，则 $\Gamma_k = \frac{6}{k(k+1)(k+2)}, \frac{\beta_k \gamma_k}{\Gamma_k} = \frac{3LK}{2} \geq \frac{\beta_{k-1} \gamma_{k-1}}{\Gamma_{k-1}}$，我们有

$$f(y_k) - f(x^*) \leq \frac{9L}{2(k+1)(k+2)} D_X^2. \tag{50}$$

加速梯度算法在凸函数问题集合上的收敛速度

证明.

充分利用 f 的凸性和 Nesterov 的步骤可以得到：对任意序列 $\{\beta_k\},\{\gamma_k\},\{\Gamma_k\}$ 满足 $L\gamma_k \leq \beta_k$ 时有

$$f(y_k) - f(x) \leq \frac{\beta_k \gamma_k}{2} D_X^2. \tag{51}$$

若还有

$$\frac{\beta)k\gamma_k}{\Gamma_k} \geq \frac{\beta_{k-1}\gamma_{k-1}}{\Gamma_{k-1}}, \forall k \geq 2, \tag{52}$$

则

$$f(y_k) - f(x) \leq \Gamma_k \frac{\beta_1 \gamma_1}{2} \|x_0 - x\|^2. \tag{53}$$

依次带入不同的序列设定即可。 □

Nesterov 加速梯度算法框架（面向非凸函数）

对非凸函数问题集合直接利用 Nesterov 加速梯度算法框架不一定会有收敛性结果。我们需要对其做一些修改，从而使得收敛性存在。

我们首先给出非凸函数问题模型：

问题模型 5

考虑无约束的非凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$：

- 全局信息 Σ：目标函数 $f \in C_L^{1,1}(\mathbb{R}^n)$，可能是非凸函数，约束集合 $X \equiv \mathbb{R}^n$；
- 局部信息 $\mathcal{O}$：$\mathcal{FO}$ 子程序；
- 解的精度 $\mathcal{T}_\epsilon$：求局部极小值点 $\bar{x} \in \mathbb{R}^n$ 使得 $\|\nabla f(\bar{x})\| \leq \epsilon$。

Nesterov 加速梯度算法框架（面向非凸函数）

非凸函数的加速梯度算法的步骤如下：

Algorithm 2: 非凸函数的加速梯度算法

Input: 令 $x_0 = y_0, \gamma_k \in (0, 1]$

for $k = 1, 2, ..., N$ **do**

$z_k \leftarrow (1 - \gamma_k) y_{k-1} + \gamma_k x_{k-1}$

$x_k \leftarrow x_{k-1} - \frac{1}{\beta_k} \nabla f(z_k)$

$y_k \leftarrow z_k - \frac{1}{\gamma_k} \nabla f(z_k)$

end for

Output: z_N

注意，该算法中，收敛点列为 $\{z_k\}$。

加速梯度算法在非凸函数问题集合上的收敛速度

定理 11

取 $\gamma_k = \frac{2}{k+1}, \gamma_k = 2L$，则上述算法求解问题模型 5 时的收敛速率如下：

- 若取 $\beta_k \in \left[\frac{4k+4}{2k+3}L, 2L\right]$，则

$$\min_{0\leq k\leq N} \|\nabla f(z_k)\|^2 \leq \frac{6L(f(x_0)-f^*)}{N}. \tag{54}$$

- 若问题模型 5 中的 f 是凸函数，取 $\beta_k = \frac{4L}{k}$，则

$$\min_{0\leq k\leq N} \|\nabla f(z_k)\|^2 \leq \frac{96L^2\|x_0-x^*\|^2}{N(N+1)(N+2)}. \tag{55}$$

可以发现，在凸函数情况下，该算法的收敛速度为 $\mathcal{O}(\frac{1}{N^{3/2}})$，而梯度法的收敛速度为 $\mathcal{O}(\frac{1}{N^{1/2}})$。这意味着，加速梯度法的确是有效果的。然而，当 f 是非凸函数时，该算法和梯度法的收敛速度相同（没有起到加速效果，但至少保证了收敛性的存在）。