李健宁
机器学习中的优化问题
2025 年 4 月 16 日

Homework 8

Question 1.

1. Backtracking 停止条件. backtracking line search 终止条件为:

$$f(\boldsymbol{x}+t\Delta\boldsymbol{x}) \leq f(\boldsymbol{x})+\alpha t\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x},$$

其中常数 $\alpha\in(0,\frac{1}{2})$。

我们利用函数在点 $\boldsymbol{x}$ 处的二阶泰勒展开:

$$f(\boldsymbol{x}+t\Delta\boldsymbol{x}) = f(\boldsymbol{x})+t\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}+\frac{1}{2}t^2\Delta\boldsymbol{x}^T\nabla^2 f(\boldsymbol{\xi})\Delta\boldsymbol{x},$$

其中 $\boldsymbol{\xi}$ 是位于 $\boldsymbol{x}$ 与 $\boldsymbol{x}+t\Delta\boldsymbol{x}$ 之间的某个点。

因为 $\nabla^2 f(\boldsymbol{\xi})\preceq M\boldsymbol{I}$, 因此

$$f(\boldsymbol{x}+t\Delta\boldsymbol{x}) \leq f(\boldsymbol{x})+t\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}+\frac{1}{2}Mt^2\|\Delta\boldsymbol{x}\|^2. \tag{1}$$

当 $t\leq\frac{-\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}}{M\|\Delta\boldsymbol{x}\|^2}$ 时, 因为 $\alpha\in(0,\frac{1}{2})$, 所以 $t\leq\frac{-2(1-\alpha)\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}}{M\|\Delta\boldsymbol{x}\|^2}$. 代入 (1), 有

$$\begin{aligned}
f(\boldsymbol{x}+t\Delta\boldsymbol{x}) \leq& f(\boldsymbol{x})+t(\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}+\frac{1}{2}Mt\|\Delta\boldsymbol{x}\|^2)\\
\leq& f(\boldsymbol{x})+t(\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}-(1-\alpha)\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x})\\
=& f(\boldsymbol{x})+\alpha t\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x},
\end{aligned}$$

2. Backtracking 迭代次数上界. 假设初始步长为 t_0, 每次回溯将当前步长乘以 $\beta\in(0,1)$, 第 k 次回溯后的步长为 $t_k=\beta^k t_0$.

迭代停止时有

$$t_k=\beta^k t_0\leq\frac{-\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}}{M\|\Delta\boldsymbol{x}\|^2}.$$

解不等式得

$$k\leq\left\lceil\log_{1/\beta}\left(\frac{Mt_0\|\Delta\boldsymbol{x}\|^2}{-\nabla f(\boldsymbol{x})^T\Delta\boldsymbol{x}}\right)\right\rceil.$$

Question 2.

定义函数 $\varphi(\alpha)=f(\mathbf{x}^{(k)}+\alpha\mathbf{d}^{(k)})$, 则 α_k 是 $\varphi(\alpha)$ 的极小点, 因此 $\varphi'(\alpha_k)=0$. 根据链式法则,

$$\varphi'(\alpha)=\nabla f(\mathbf{x}^{(k)}+\alpha\mathbf{d}^{(k)})^T\cdot\mathbf{d}^{(k)},$$

代入 α_k, 得到

$$\nabla f(\mathbf{x}^{(k+1)})^T\cdot\mathbf{d}^{(k)}=0.$$

由于 $\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}=\alpha_k\mathbf{d}^{(k)}$, 可得

$$\nabla f(\mathbf{x}^{(k+1)})^T\cdot(\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)})=\alpha_k\nabla f(\mathbf{x}^{(k+1)})^T\cdot\mathbf{d}^{(k)}=0.$$

所以

$$\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\perp\nabla f(\mathbf{x}^{(k+1)}).$$

Question 3.

针对 $\alpha = 0.1, \beta = 0.3$ 给出 $\log(f(x) - p^*)$ 和步长随迭代次数的变化。

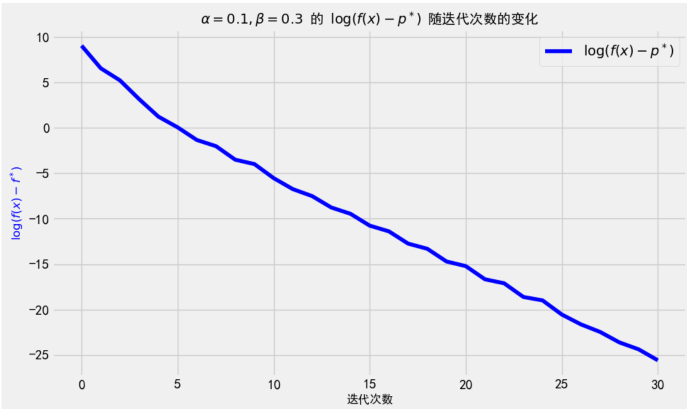


图 1. $\alpha = 0.1, \beta = 0.3$ 的 $\log(f(x) - p^*)$ 随迭代次数的变化

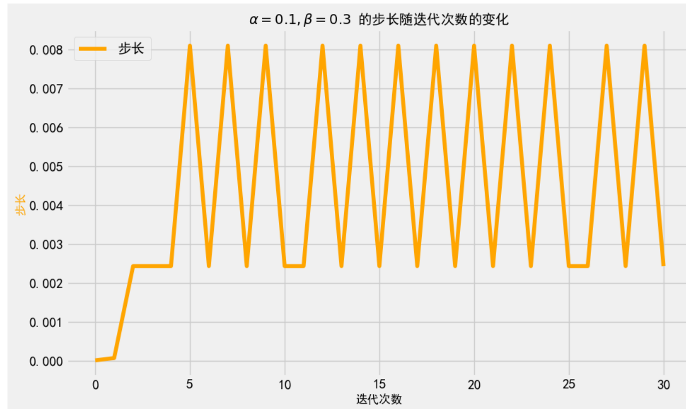


图 2. $\alpha = 0.1, \beta = 0.3$ 的步长随迭代次数的变化

我们选取了 $\alpha \in [0.01, 0.1, 0.3, 0.49], \beta \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ 绘制迭代次数和参数的关系。实验中，我们选取的停止标准是 $\nabla f(x) \leq \eta = 10^{-4}$，迭代次数上限是 1000 次。

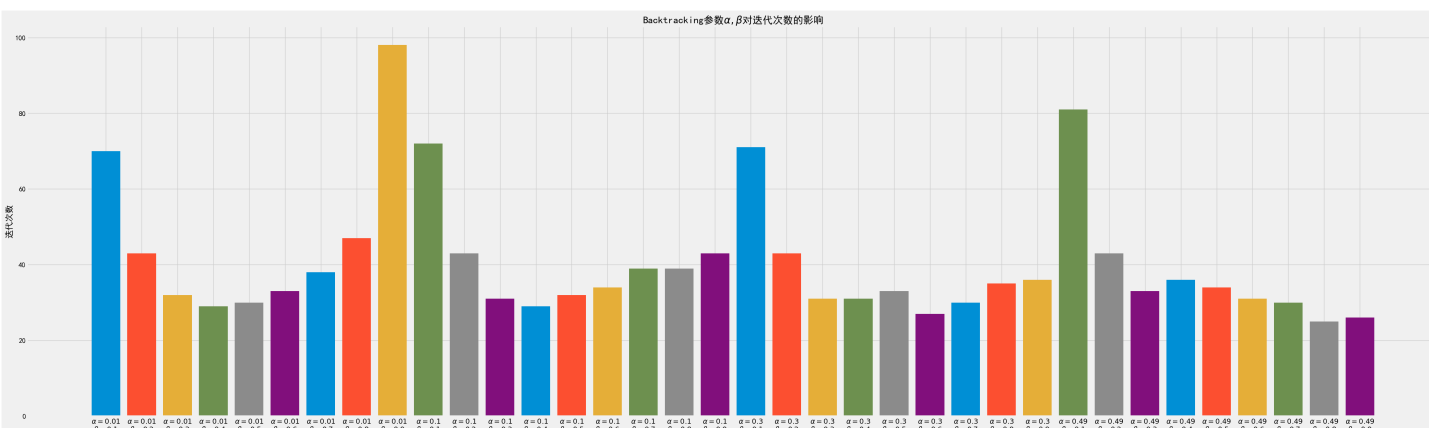


图 3. Backtracking 参数 α, β 对迭代次数的影响[1]

能够发现，$\nabla f(x) \leq \eta = 10^{-4}$ 的停止标准相对来说比较严苛。Backtracking 参数选取上，更加接近 0.5 的 β 和 α 收敛更快。在 β 较小时，α 的选择对收敛行为影响较大，其他情况则主要依赖于 β，但迭代次数变化的范围整体来看较小。

Question 4.

先给出伪代码。

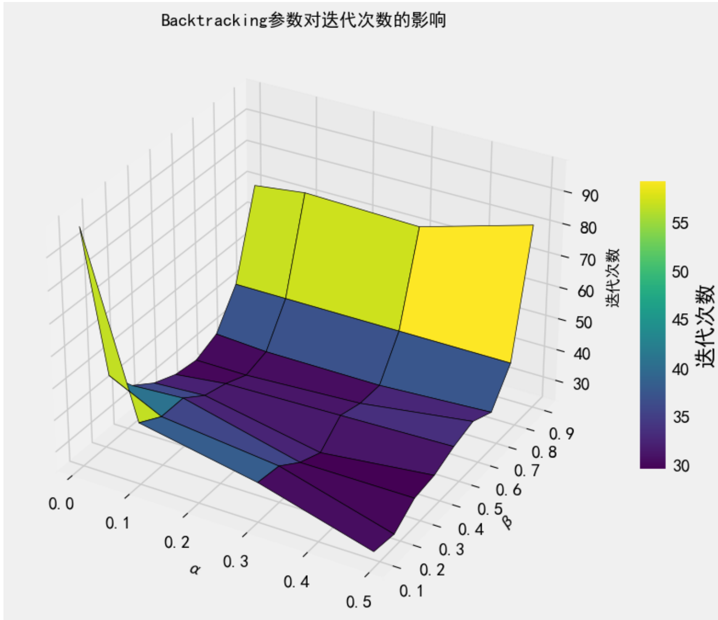


图 4. Backtracking 参数 α, β 对迭代次数的影响

Algorithm 1: Steepest Descent in ℓ_∞-norm

Input: Objective function $f(x)$, gradient $\nabla f(x)$, initial point x_0, tolerance η, backtracking parameters α, β

Output: Approximate minimizer x

$x \leftarrow x_0$;

repeat

  $g \leftarrow \nabla f(x)$;

  $d \leftarrow -\text{sign}(g)$;

  $t \leftarrow 1$;

  while $f(x + td) > f(x) + \alpha t g^T d$ **do**

    $t \leftarrow \beta t$;

  $x \leftarrow x + td$;

until $\|g\|_2 \leq \eta$;

类似上一题，针对 $\alpha = 0.1, \beta = 0.1$ 给出 $\log(f(x) - p^*)$ 和步长随迭代次数的变化。

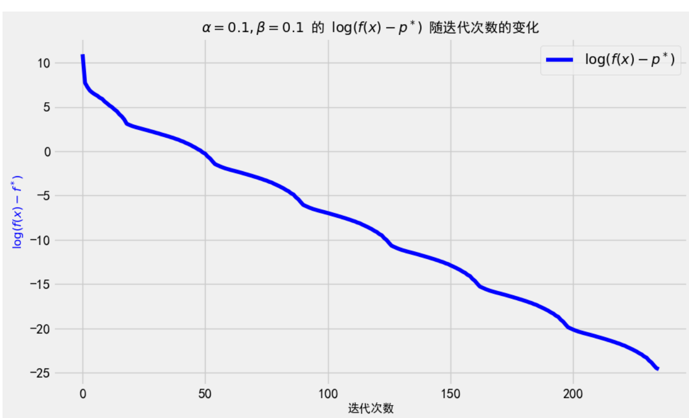


图 5. $\alpha = 0.1, \beta = 0.1$ 的 $\log(f(x) - p^*)$ 随迭代次数的变化

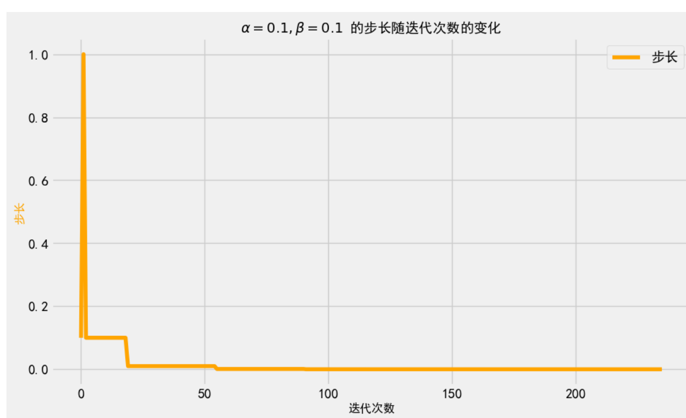


图 6. $\alpha = 0.1, \beta = 0.1$ 的步长随迭代次数的变化

我们选取了 $\alpha \in [0.01, 0.1, 0.3, 0.49], \beta \in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ 绘制迭代次数和参数的关系。实验中，我们选取的停止标准是 $\nabla f(x) \leq \eta = 10^{-4}$，迭代次数上限是 1000 次。

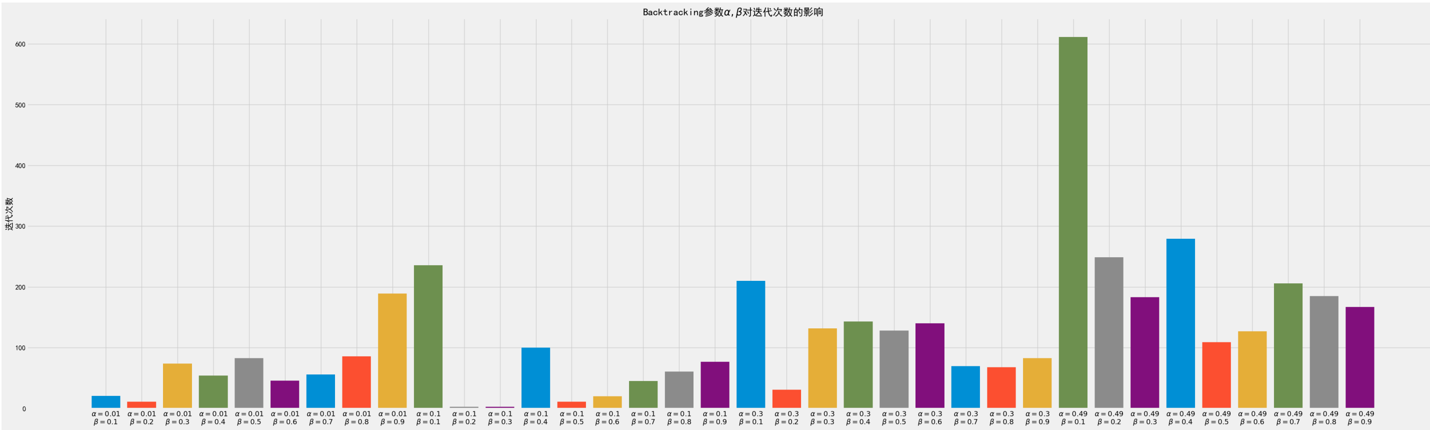


图 7. Backtracking 参数 α, β 对迭代次数的影响[2]

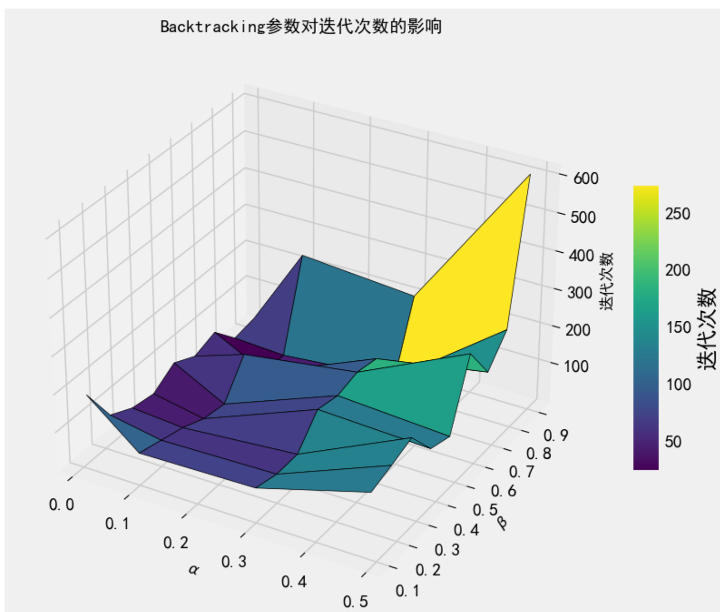


图 8. Backtracking 参数 α, β 对迭代次数的影响

能够发现，$\nabla f(x) \leq \eta = 10^{-4}$ 的停止标准相对来说依旧比较严苛。Backtracking 参数选取上，较小的 α 和 β 收敛更快。在 β 较大时，α 的选择对收敛行为影响较大，其他情况迭代次数的整体波动不大。

Question 5.

根据题意，我们分别采取 Damped Newton 法和 Gauss Newton 法求解优化问题，并展示求解结果。在 Damped Newton 法中，我们需要选择 Backtracking 的参数，这里选择的参数是 $\alpha = 0.3, \beta = 0.5$，根据课本中的典型参数值，同时希望能够较快收敛。（事实上，选取更小的 β 能够（在时间上）更快收敛。）根据结果（图9、10），Gauss Newton 法在解决非线性最小平方问题上确实表现上佳。

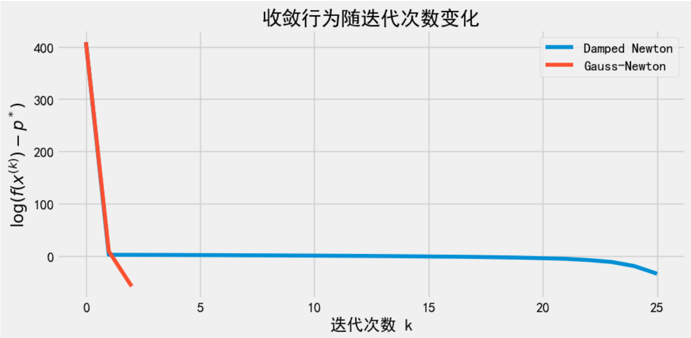


图 9. 收敛行为随迭代次数变化

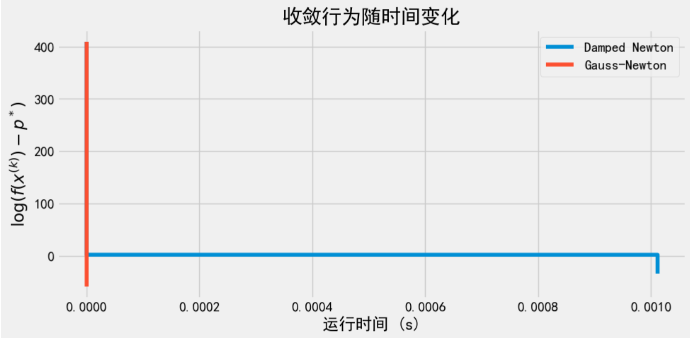


图 10. 收敛行为随时间变化

Question 6.

根据 Conjugate Direction 方法，有 $\boldsymbol{d}^{(k+1)} = -\boldsymbol{g}^{(k+1)} + \beta_k \boldsymbol{d}^{(k)}, k \geq 1$，所以根据共轭方向的性质 $\boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{d}^{(k-1)} = 0$ 能够得到

$$\boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{d}^{(k)} = -\boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{g}^{(k)} + \beta_k \boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{d}^{(k-1)} = -\boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{g}^{(k)}, k \geq 1.$$

当 $k = 0$ 时，由 $\boldsymbol{d}^{(0)} = -\boldsymbol{g}^{(0)}$，$\boldsymbol{d}^{(0)}\boldsymbol{Q}\boldsymbol{d}^{(0)} = -\boldsymbol{d}^{(0)}\boldsymbol{Q}\boldsymbol{g}^{(0)}$。因此对 $\forall k, \boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{d}^{(k)} = -\boldsymbol{d}^{(k)}\boldsymbol{Q}\boldsymbol{g}^{(k)}$。

Question 7.

根据 *Opt_book.pdf* 中引理 146，我们有

$$\mathbf{g}^{(k+1)^T}\mathbf{d}^{(j)} = 0, \quad j = 0, 1, \ldots, k. \tag{2}$$

根据算法，对 $1 \leq j \leq k-1$

$$\mathbf{d}^{(j)} = -\mathbf{g}^{(j)} + \beta_{j-1}\mathbf{d}^{(j-1)},$$

所以

$$0 = \mathbf{g}^{(k+1)^T}\mathbf{d}^{(j)} = -\mathbf{g}^{(k+1)^T}\mathbf{g}^{(j)} + \beta_{j-1}\mathbf{g}^{(k+1)^T}\mathbf{d}^{(j-1)}.$$

由于 $\mathbf{g}^{(k+1)^T}\mathbf{d}^{(j-1)} = 0$，有

$$\mathbf{g}^{(k+1)^T}\mathbf{g}^{(j)} = 0, 1 \leq j \leq k-1.$$

$j = 0$ 时，$\mathbf{d}^{(0)} = -\mathbf{g}^{(0)}$, 所以 $\mathbf{g}^{(k+1)^T}\mathbf{g}^{(0)} = -\mathbf{g}^{(k+1)^T}\mathbf{d}^{(0)} = 0$ 也成立。

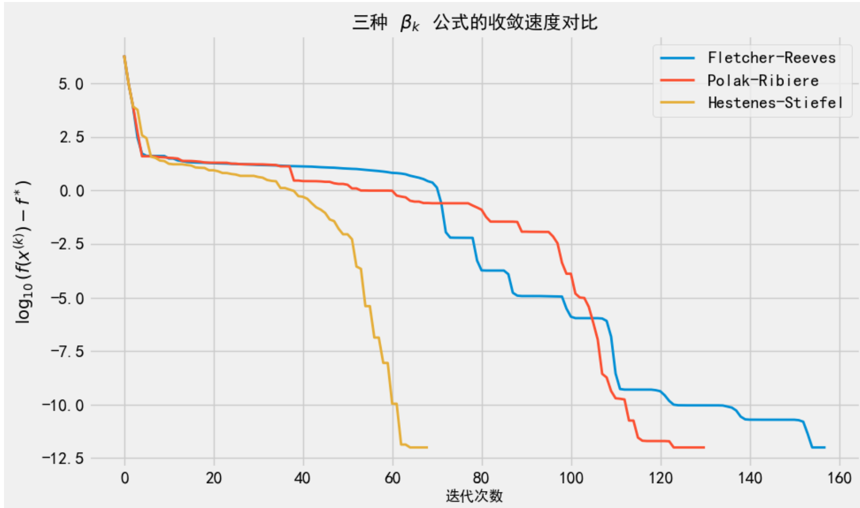
Question 8.

Backtracking 参数 $\alpha = 0.3, \beta = 0.5$，最多搜索 20 次；共轭梯度法的停止标准是 $\nabla f(x) \leq 10^{-6}$，最大迭代次数 20000 次。

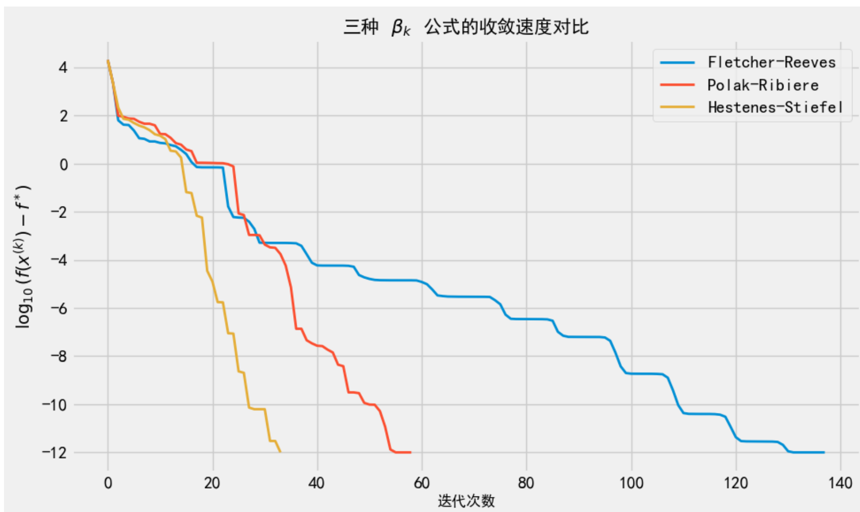
我们测试了函数中的 $\alpha \in [1, 100, 10000]$ 三种情况下三种 β 方法的表现。总体来讲 FR 方法初期下降最快，但接近目标值时速度放缓，HS 方法恰恰相反，初期下降较慢，但最终收敛最快，PR 中期较慢，其他时期介于两者之间。α 的变化影响目标函数的条件数，相对来讲 α 越大问题条件数越差，收敛速度也就越慢。

表 1. 不同 α 和 β 方法组合下的共轭梯度法结果对比

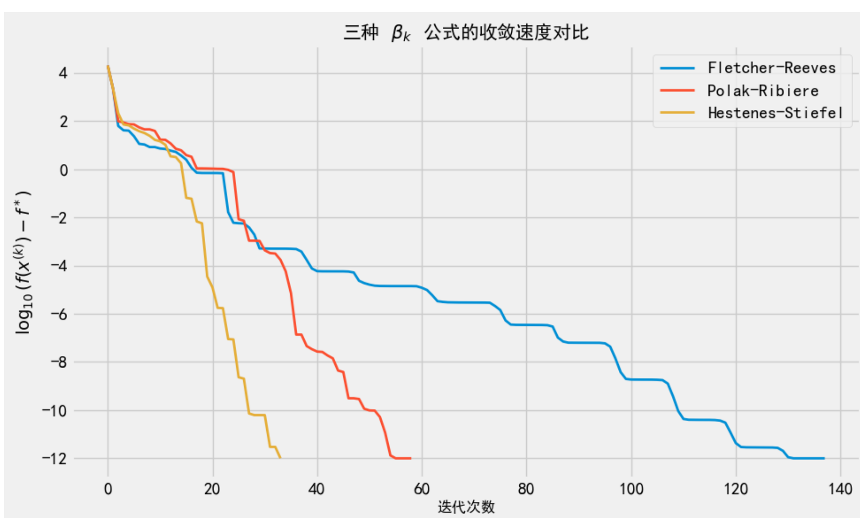
α	β 方法	迭代次数	最终梯度范数	耗时 (s)	最终函数值
	FR	157	5.03×10^{-7}	0.1632	2.54×10^{-13}
10000	PR	130	2.94×10^{-7}	0.1388	1.90×10^{-16}
	HS	68	6.67×10^{-7}	0.0711	8.09×10^{-17}
	FR	137	8.37×10^{-7}	0.1199	7.34×10^{-13}
100	PR	58	1.85×10^{-7}	0.0525	3.37×10^{-14}
	HS	33	7.78×10^{-8}	0.0256	2.12×10^{-16}
	FR	34	6.93×10^{-7}	0.0302	6.95×10^{-13}
1	PR	31	6.82×10^{-7}	0.0234	3.63×10^{-14}
	HS	30	4.10×10^{-7}	0.0231	7.85×10^{-14}



(A) $\alpha = 10000$ 时的收敛行为



(B) $\alpha = 100$ 时的收敛行为



(C) $\alpha = 1$ 时的收敛行为

图 11. 共轭梯度法在不同 α 下的收敛行为对比