第 3 章：无约束优化

(1)：下降法

曹语

课程主页：https://yucaoyc.github.io/math3806

背景和目标

我们介绍了一些（无约束的）凸优化问题的例子：

$$\text{minimize } f(\boldsymbol{x})$$

例如

- 无约束的最小二乘问题；该优化问题可被应用于拟合数据
- 无约束的几何规划

问题：为什么我们不考虑无约束的线性规划问题呢？

本章目标：介绍算法来近似计算上述优化问题的最优解

- 下降法（及其变式）$\rightarrow$ 这节课的内容
- 牛顿法

情景设定和算法的含义

目标是近似求解如下问题：

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}), \\ &\boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

假设

- f 是凸函数
- ∇f 和 $\nabla^2 f$ 可计算
- 最优解 $\boldsymbol{x}^\star$ 存在；将 $f(\boldsymbol{x}^\star)$ 标记为符号 $p^\star$

算法： 通过找到一个序列 $\{\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(k)}, \cdots\}$ 使得 $f(\boldsymbol{x}^{(i)})$ 随着 i 的增加而不断减小；基于误差阈值 ϵ，输出某个 $\boldsymbol{x}^{(k)}$，若 $f(\boldsymbol{x}^{(k)}) \leq p^\star + \epsilon$。

因此，事实上我们会考虑在如下的区间进行优化：

$$S = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leqslant f\left(\boldsymbol{x}^{(0)}\right) \right\}$$

为简化，我们直接假设 S 是闭集。

需要解决的算法问题：

Q1 基于 $\boldsymbol{x}^{(k)}$，如何更新得到 $\boldsymbol{x}^{(k+1)}$，使得 $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$?

Q2 如何选择终止的判断方式?

因此，事实上我们会考虑在如下的区间进行优化：

$$S = \left\{ \boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leqslant f\left(\boldsymbol{x}^{(0)}\right) \right\}$$

为简化，我们直接假设 S 是闭集。

需要解决的算法问题：

Q1 基于 $\boldsymbol{x}^{(k)}$，如何更新得到 $\boldsymbol{x}^{(k+1)}$，使得 $f(\boldsymbol{x}^{(k+1)}) < f(\boldsymbol{x}^{(k)})$?

Q2 如何选择终止的判断方式?

由于最优解满足 $\nabla f(\boldsymbol{x}) = 0$，因此我们可以选择 $\left\|\nabla f(\boldsymbol{x}^{(k)})\right\| \leq \eta$（其中 η 是某个提前选择的小的参数，例如 $\eta = 10^{-5}$）

目录

下降方法

我们根据某个方向 $\Delta \boldsymbol{x}^{(k)}$ 来更新

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t^{(k)} \Delta \boldsymbol{x}^{(k)}, \qquad t^{(k)} > 0$$

- $\Delta \boldsymbol{x}^{(k)}$ 被称为**搜索方向**
- $t^{(k)}$ 被称为称为第 k 次迭代的**步长**

考虑若需要沿着该方向的函数值（严格）变小；考虑 $g(t) = f(\boldsymbol{x}^{(k)} + t\Delta \boldsymbol{x}^{(k)})$

$$0 > g'(0) = \underline{\qquad ? \qquad}$$

因此，我们会需要

$$\nabla f(\boldsymbol{x}^{(k)})^\top \Delta \boldsymbol{x}^{(k)} < 0$$

我们称这样的方向为下降方向。

问题：有没有可能满足 $\nabla f(\boldsymbol{x}^{(k)})^{\top}\Delta\boldsymbol{x}^{(k)} < 0$ 的下降方向 $\Delta\boldsymbol{x}^{(k)}$ 无法找到呢?

可以讨论两种情况：$\nabla f(\boldsymbol{x}^{(k)}) \neq \boldsymbol{0}$ 和 $\nabla f(\boldsymbol{x}^{(k)}) = \boldsymbol{0}$。

通用下降方法

Algorithm 1: 通用下降方法

Input: $\boldsymbol{x}^{(0)}$

1 $k = 0$
2 **while** 终止条件不满足，且 $k < N$ **do**
3 　确定下降方向 $\Delta \boldsymbol{x}^{(k)}$
4 　直线搜索：选择步长 $t^{(k)} > 0$
5 　$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t^{(k)} \Delta \boldsymbol{x}^{(k)}$
6 　$k \leftarrow k + 1$
7 **end**

此处的通用方法中，有两处有待进一步确认：

任务 1 确定下降方向

任务 2 直线搜索

目录

方式 (1)：精确直线搜索/Exact line search

$$t = \operatorname{argmin}_{s \geqslant 0} f(\boldsymbol{x} + s\Delta\boldsymbol{x})$$

优劣对比：

- 好处：比较精确，且函数值下降快
- 坏处：计算成本一般而言比较高

使用场景： 如果计算该问题的成本远小于确定下降方向的成本时，我们可以考虑该方法（例如，计算函数值所需的时间远小于计算梯度的时间）

方式 (2)：回溯直线搜索/Backtracking line search

给定参数 $\alpha \in (0, 0.5), \beta \in (0, 1)$

Algorithm 2: 回溯直线搜索

Input: t_0 (一般选 $t_0 = 1$), 当前位置 $\boldsymbol{x}$, 下降方向 $\Delta \boldsymbol{x}$

Result: t

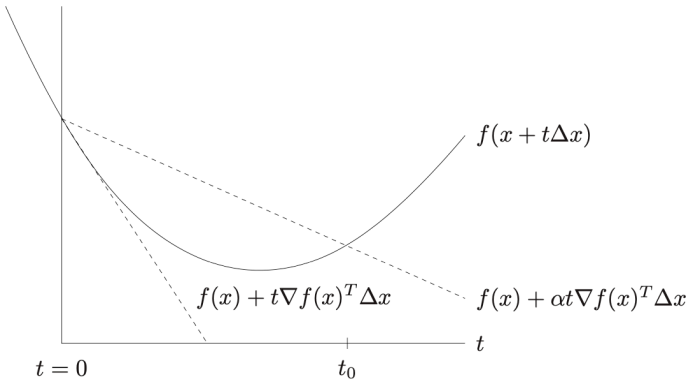
$t \leftarrow t_0$

while $f(\boldsymbol{x} + t\Delta \boldsymbol{x}) > f(\boldsymbol{x}) + \alpha t \nabla f(\boldsymbol{x})^\top \Delta \boldsymbol{x}$ **do**

  令 $t \leftarrow \beta t$

end

- 由于 $\Delta \boldsymbol{x}$ 是下降方向，只要 t 足够小，就一定可以满足上述终止条件
- 终止条件 $f(\boldsymbol{x} + t\Delta \boldsymbol{x}) \leq f(\boldsymbol{x}) + \alpha t \nabla f(\boldsymbol{x})^\top \Delta \boldsymbol{x}$ 也称为**Armijo's condition**



经验而言：正常取值 α 在 $0.01 \sim 0.3$ 之间；β 在 $0.1 \sim 0.8$ 之间

目录

梯度下降法

梯度下降法： 考虑

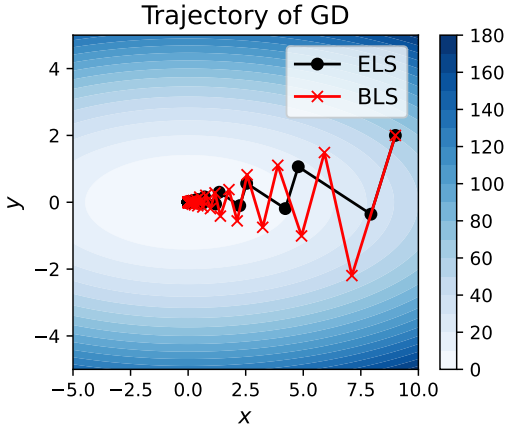
$$\Delta \boldsymbol{x} = -\nabla f(\boldsymbol{x})$$

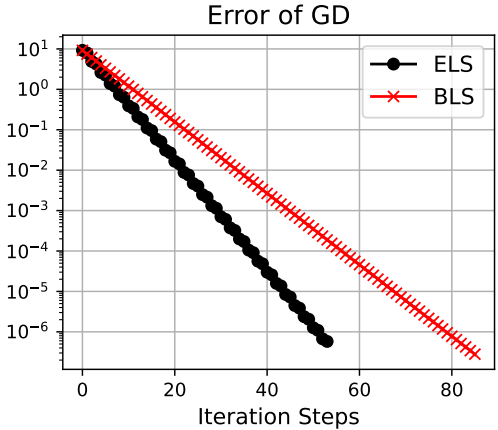
然后可以搭配精确直线搜索，或者回溯直线搜索来使用。

实验 1：不同直线搜索的对比

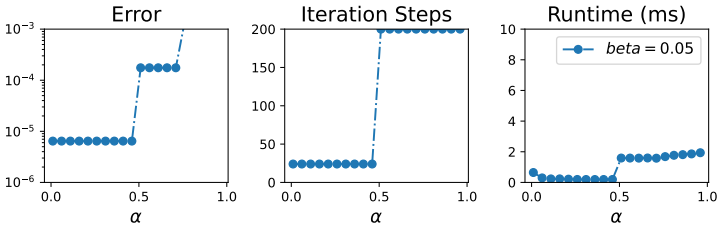
考虑如下的函数 $f(\boldsymbol{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$

某组实验的结果如下（具体见代码）：

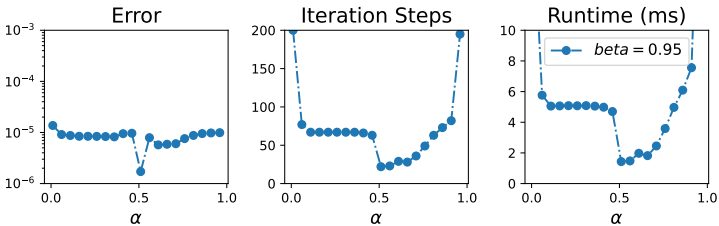




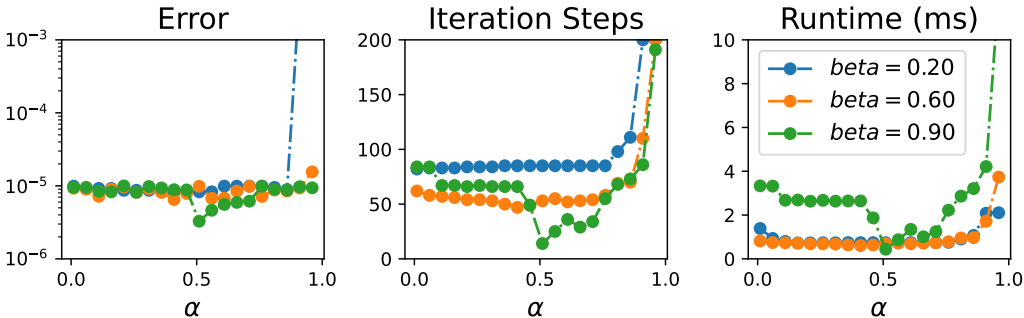
实验 2：回溯直线搜索的参数的影响



显然希望 $\alpha < 1/2$；若 $\alpha \in (0, 1/2)$ 之间，选择影响不大



我们不太希望选过大的 $\beta \approx 1$，因为这样运行时间太长



我们可得到如下观察：

- $\alpha < 1/2$，但也别太小，比如 α 在 $0.1 \sim 0.5$ 之间都还可以；
- $\beta \approx 1$ 不太好是因为搜索需要时间太长；$\beta \approx 0$ 不太好是因为搜索过于粗糙，使得结果没其他参数好；
- 在前面提到的经验的取值范围中，α 和 β 对于结果的影响并不显著。

实验 3：问题本身的影响

考虑 $f(\boldsymbol{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$

假设选择点 $\boldsymbol{x}^{(k)} = (a, b)$，对于梯度下降法 + 精确直线搜索，我们计算得知

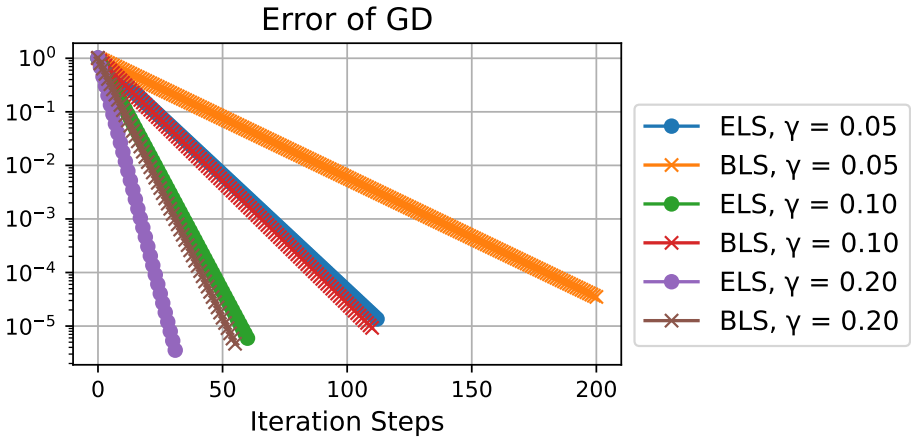
$$t = \frac{a^2 + b^2\gamma^2}{a^2 + b^2\gamma^3} \qquad \boldsymbol{x}^{(k+1)} = \left(\frac{ab^2(\gamma-1)\gamma^2}{a^2 + b^2\gamma^3}, -\frac{a^2 b(\gamma-1)}{a^2 + b^2\gamma^3}\right)$$

我们可以直接验证：

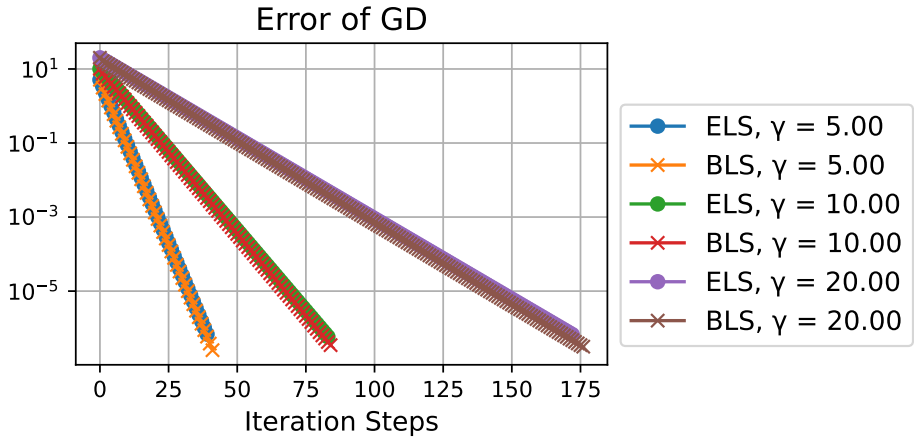
$$\frac{f\left(\boldsymbol{x}^{(k+1)}\right)}{f\left(\boldsymbol{x}^{(k)}\right)} = \frac{a^2b^2(\gamma-1)^2\gamma}{(a^2 + b^2\gamma)(a^2 + b^2\gamma^3)} = \frac{(\gamma-1)^2\gamma}{\left(1 + (b/a)^2\gamma\right)\left((a/b)^2 + \gamma^3\right)} \leq \left(\frac{\gamma-1}{\gamma+1}\right)^2$$

- 当 $\gamma \approx 0$ 或 $\gamma \approx \infty$ 时，收敛比较慢；当 $\gamma \approx 1$ 时，仅需几步即可收敛。
- 若初值 $\boldsymbol{x}^{(0)} = (\gamma, 1)$，上述的不等号成为等号，即 $\frac{f\left(\boldsymbol{x}^{(k+1)}\right)}{f\left(\boldsymbol{x}^{(k)}\right)} = \left(\frac{\gamma-1}{\gamma+1}\right)^2, \forall k$。

对于初值 $x_0 = (\gamma, 1)$，验证了 γ 很小的时候，问题找到精确解变难。

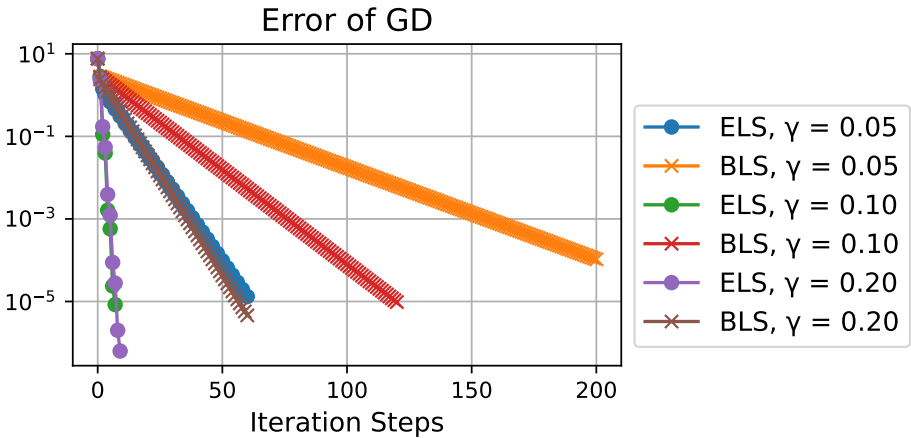


对于初值 $x_0 = (\gamma, 1)$，验证了 γ 很大的时候，问题找到精确解变难。

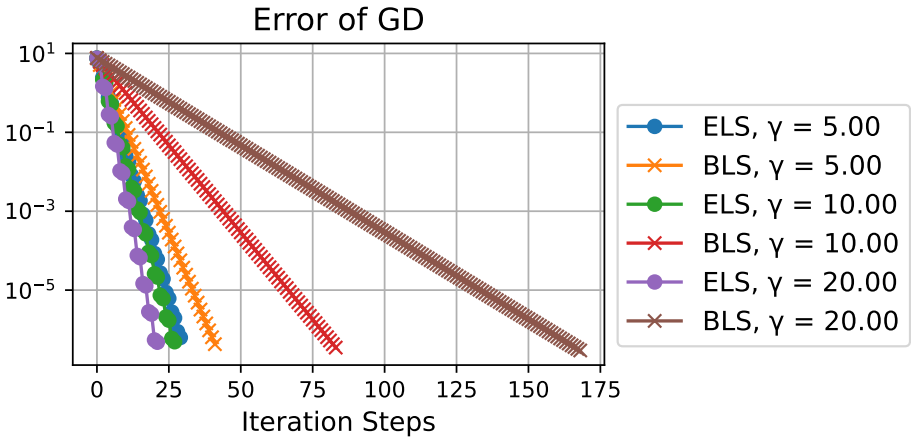


实验 4：初值对于问题的影响

问题的初值对于结果影响也很大：对此处的实验，我们取相同的 $\boldsymbol{x}^{(0)}$



问题的初值对于结果影响也很大：对此处的实验，我们取相同的 $\boldsymbol{x}^{(0)}$



之前描述的 $\gamma \to \infty$ 和 $\gamma \to 0$ 时，收敛变难是指针对所有点的最差情况，但是对于特殊的初始 $\boldsymbol{x}^{(0)}$，情况不一定完全吻合。

实验现象总结

- 梯度法具有线性收敛，即 $\log e_N \approx -aN + b$，或者说 $e_N \approx ce^{-aN}$ （$a, c > 0, b \in \mathbb{R}$）
- 回溯直线搜索的参数 α, β 对结果有影响，但不是特别本质；很多时候，回溯直线搜索的效果和精确直线搜索效果差不多，计算代价一般低很多
- 问题本身对于收敛速度的影响很大（在下一部分，我们会提到条件数的概念来量化这个；理论验证大多数时候无法考虑单个初值的情况，而是考虑最差的情况）

部分实验现象的理论解释将是本章第 (2) 部分需要尝试解决的问题。

目录

最速下降法

最速下降法： 考虑梯度下降法的规范化版本

$$\Delta \boldsymbol{x} = \arg\min \left\{ \nabla f(\boldsymbol{x})^\top \boldsymbol{v} \mid \|\boldsymbol{v}\| = 1 \right\}$$

- 若范数为 L^2，则 $\Delta \boldsymbol{x} = -\frac{\nabla f(\boldsymbol{x})}{\|\nabla f(\boldsymbol{x})\|_2}$
- 若范数为 L^1，则 $\Delta \boldsymbol{x} = -\text{sign}\left(\frac{\partial f(\boldsymbol{x})}{\partial x_i}\right) \boldsymbol{e}_i$，其中坐标 i 是数值 $\nabla f(\boldsymbol{x})_i$ 中绝对值意义下最大的值的坐标。

  例子： 若 $\nabla f(\boldsymbol{x}) = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$，则 $\Delta \boldsymbol{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

  该方法有时又称**坐标下降算法**

- **(二次 P-范数)** 若 $\|\boldsymbol{v}\|_{\boldsymbol{P}} := \sqrt{\boldsymbol{v}^\top \boldsymbol{P}\boldsymbol{v}} = \left\|\boldsymbol{P}^{1/2}\boldsymbol{v}\right\|_2$，其中 $\boldsymbol{P}$ 是正定矩阵，则

$$\Delta \boldsymbol{x} \propto -\boldsymbol{P}^{-1}\nabla f(\boldsymbol{x})$$

几何含义：考虑变换坐标 $\bar{\boldsymbol{x}} = \boldsymbol{P}^{1/2}\boldsymbol{x}$，优化问题等价于

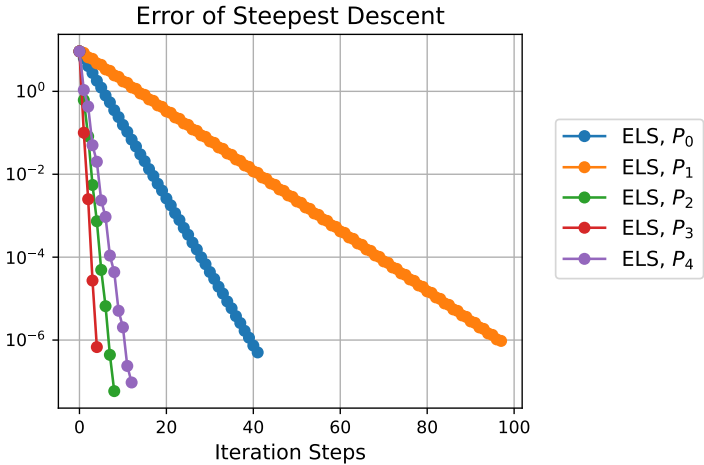
$$\min \bar{f}(\bar{\boldsymbol{x}}) = f\left(\boldsymbol{P}^{-1/2}\bar{\boldsymbol{x}}\right)$$

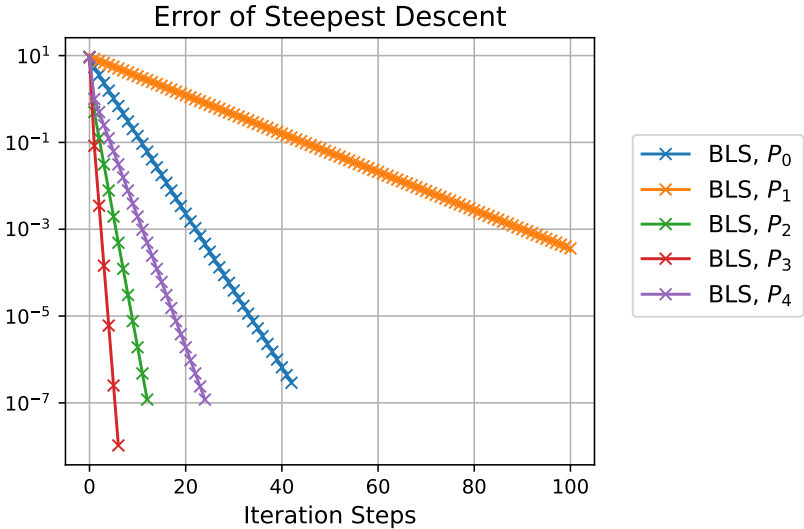
在 $\bar{\boldsymbol{x}}$ 的坐标中，Euclidean 距离意义下的最速下降方向为 $\Delta\bar{\boldsymbol{x}} \propto -\boldsymbol{P}^{-1/2}\nabla f(\boldsymbol{x})$，因此对应原坐标系

$$\Delta \boldsymbol{x} = \boldsymbol{P}^{-1/2}\Delta\bar{\boldsymbol{x}} \propto -\boldsymbol{P}^{-1}\nabla f(\boldsymbol{x})$$

不同最速下降法的选择

选择不同的 $\boldsymbol{P}_k = \begin{bmatrix} 1 & 0 \\ 0 & \theta_k \end{bmatrix}$; $\theta_0 = 1, \theta_1 = 4, \theta_2 = \frac{1}{4}, \theta_3 = \frac{1}{4.8}, \theta_4 = \frac{1}{10}, \gamma = 5$





实现坐标变换的矩阵 $\boldsymbol{P}$ 的选择很关键!

目录

总结

- 知道算法设计的核心是更新 $\boldsymbol{x}^{(k)}$ 和选择终止条件
- 了解使用 $\|\nabla f(\boldsymbol{x})\| \leq \eta$ 作为终止条件这一选择
- 能够知道更新 $\boldsymbol{x}^{(k)}$ 中需要的两个组件：直线搜索、确定下降方向
- 并且分别知道两个组件的两种选择
- 对于数值实验的结果有直观印象

阅读作业 & 参考资料：

- 课本第 9.1 - 9.4 章