第 3 章：无约束优化

(3)：牛顿法

曹语

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

目录

背景和目标

在最速下降法中，我们已经经验地看到对于坐标变换 $\bar{\boldsymbol{x}} = \boldsymbol{P}^{1/2}\boldsymbol{x}$，$\boldsymbol{P}$ 是正定矩阵；若

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

能够接近简单的二次函数 $|\bar{\boldsymbol{x}}|^2$，最速下降法的方向更有效。转换成原坐标轴得到：

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

的效率更好。

目标：如何通过这个思想构造加速算法

练习：如果 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}$ 本身具有二次形式，$\boldsymbol{Q} \in \mathbf{S}^n_{++}$，则最优的 $\boldsymbol{P}$ 是什么，能帮助我们实现在新坐标中类似 $\bar{\boldsymbol{x}}^\top\boldsymbol{I}_{n\times n}\bar{\boldsymbol{x}}$?

结论是我们会选择

$$\boldsymbol{P} = \boldsymbol{Q} = \nabla^2 f \qquad \rightarrow \qquad \Delta \boldsymbol{x} = -\boldsymbol{P}^{-1} \nabla f(\boldsymbol{x}) = -\left(\nabla^2 f\right)^{-1} \nabla f(\boldsymbol{x})$$

如果一般而言 Hessian 矩阵 $\nabla^2 f$ 取决于位置 $\boldsymbol{x}$，则我们可以考虑局部更新为

$$\Delta \boldsymbol{x} = -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})$$

该下降法被称为**牛顿法**

练习：上述的牛顿步径是下降方向吗?

目录

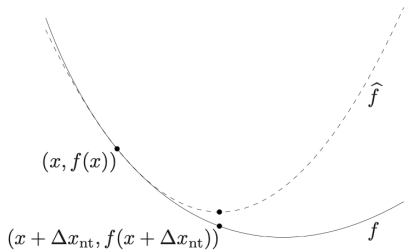
视角 1：二阶近似的最优解

若 $\boldsymbol{y} \approx \boldsymbol{x}$，则我们可知

$$
\begin{aligned}
f\big(\boldsymbol{y}\big) &\approx \hat{f}(\boldsymbol{y}) \\
\hat{f}(\boldsymbol{y}) &= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\top} \boldsymbol{v} \\
&\quad + \frac{1}{2} \boldsymbol{v}^{\top} \nabla^2 f(\boldsymbol{x}) \boldsymbol{v},
\end{aligned}
$$

其中，$\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{x}$。

若我们最小化右侧的二阶近似，则最优的 $\boldsymbol{v}$ 为牛顿步径。



视角 2：Hessian 范数的最速下降方向

对于范数 $\|\boldsymbol{u}\|_{\boldsymbol{P}} = \left\|\boldsymbol{P}^{1/2}\boldsymbol{u}\right\|_2$，其中 $\boldsymbol{P} = \nabla^2 f(\boldsymbol{x})$，则最速下降方向为

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

【该视角即坐标变换视角】

视角 3：线性化最优性条件的解

- 如果我们在 $\boldsymbol{x}$ 附近做 ∇f 的一阶展开，则

$$\nabla f(\boldsymbol{x}+\boldsymbol{v}) \approx \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})\boldsymbol{v}$$

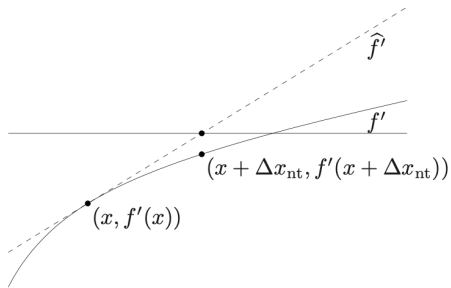
通过解，右侧 $= 0$，可知

$$\boldsymbol{v} = -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})$$

- 对于 1 维问题，为了求解 $f'(x) = 0$，上述意义下 “ 最好 ” 的方向是

$$v = -\frac{f'(x)}{f''(x)}$$

1D 的牛顿法示意图



目录

牛顿减量

确定方向后，我们需要确定最优的步长 t

$$g(t) = f\left(\boldsymbol{x} + t\Delta\boldsymbol{x}\right)$$

计算可知，

$$\left.\frac{\mathrm{d}g}{\mathrm{d}t}\right|_{t=0} = -\nabla f(\boldsymbol{x})^{\top}\left(\nabla^2 f\left(\boldsymbol{x}\right)\right)^{-1}\nabla f(\boldsymbol{x}) \overset{\text{标记为}}{=} -\left(\lambda(\boldsymbol{x})\right)^2$$

该量 $\lambda(\boldsymbol{x})$ 被称为**牛顿减量**。

阻尼/谨慎牛顿法

Algorithm 1: 阻尼牛顿法/谨慎牛顿法

Input: $\boldsymbol{x}^{(0)}$，阈值 ϵ

while *True* **do**

- 计算牛顿步径 $\Delta \boldsymbol{x}^{(k)} = -\left(\nabla^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \nabla f(\boldsymbol{x}^{(k)})$
- 计算牛顿减量 $\lambda(\boldsymbol{x}^{(k)}) = \sqrt{\nabla f(\boldsymbol{x}^{(k)})^\top \left(\nabla^2 f\left(\boldsymbol{x}^{(k)}\right)\right)^{-1} \nabla f(\boldsymbol{x}^{(k)})}$
- 停止准则：如果 $\left(\lambda(\boldsymbol{x}^{(k)})\right)^2/2 \leq \epsilon$，则退出
- 通过回溯直线搜索选择步长 $t^{(k)} > 0$
- $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t^{(k)} \Delta \boldsymbol{x}^{(k)}$

end

若选择步长为 $t = 1$，则我们称之为纯牛顿法/纯牛顿阶段

例子 1：考虑凸函数 $f(x) = x - \ln(x)$，它具有唯一的极小值 $x^* = 1$。(a) 请写出阻尼牛顿法的格式，和纯牛顿法的格式；(b) 假设初值 $x^{(0)} \in (0, 1)$，验证对于阻尼牛顿法，使用回溯直线搜索，并 $\alpha = 1/2$ 时，步长一直为 $t = 1$，即实际一直处于纯牛顿阶段；(c) 假设 $x^{(0)} = 1/2$，请写出误差 $e_k = \left|x^{(k)} - x^*\right|$ 的具体表达式。

例子 2：2D 例子的数值实验见代码部分

目录

假设和结论

- 基本设定和下降法一样：即假设 $m\boldsymbol{I}_{n\times n} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}_{n\times n}$
- 额外假设 Hessian 矩阵具有 L-Lipschitz 条件：

$$\left\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\right\|_2 \leqslant L\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

（可以不严格地理解为三阶导数有界）

- 结论：迭代次数不会超过

$$\log_2 \log_2 \left(\frac{2m^3}{L^2\epsilon}\right) + \frac{M^2L^2/m^5}{\alpha\beta \min\left\{1, 9(1-2\alpha)^2\right\}} \left(f\left(\boldsymbol{x}^{(0)}\right) - p^\star\right)$$

“ 若 $\epsilon \to 0$, 条件数 $\kappa \gg \infty$，以及初始误差增加，则总体而言我们仍然可以预期所需迭代步数增加。”

直观解释：为何出现 $\log\log\left(1/\epsilon\right)$?

假设我们的某步已经比较接近真解 $\boldsymbol{x}^{(k)} \approx \boldsymbol{x}^*$，则

- 结论 1：回溯直线搜索得到 $t^{(k)} = 1$；因此，$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \Delta\boldsymbol{x}^{(k)}$，$\Delta\boldsymbol{x}^{(k)} = -\left(\nabla^2 f(\boldsymbol{x}^{(k)})\right)^{-1}\nabla f(\boldsymbol{x}^{(k)})$；
- 为简化，仅考虑一维情况，

$$
\begin{aligned}
f'(\boldsymbol{x}^{(k+1)}) &\approx f'(\boldsymbol{x}^{(k)}) + f''(\boldsymbol{x}^{(k)})\left(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right) + \frac{f'''(\boldsymbol{x}^{(k)})}{2}\left(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right)^2 \\
&= f'(\boldsymbol{x}^{(k)}) - f'(\boldsymbol{x}^{(k)}) + \frac{f'''(\boldsymbol{x}^{(k)})}{2}\left(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\right)^2 \\
&\approx \frac{f'''(\boldsymbol{x}^*)}{2\left(f''(\boldsymbol{x}^*)\right)^2}\left(f'(\boldsymbol{x}^{(k)})\right)^2
\end{aligned}
$$

因此，误差 $e_k := \left|f'(\boldsymbol{x}^{(k)})\right|$ 满足如下关系

$$
e_{k+1} = Ce_k^2
$$

问题 1：$e_{k+1} \approx Ce_k^2$ 的实际含义是什么呢?

$$\underbrace{\log e_{k+1}}_{\text{标记为} y_{k+1}} \approx \log(C) + 2\underbrace{\log e_k}_{\text{标记为} y_k}$$

$$\rightarrow \qquad y_{k+2} \approx \log(C) + 2y_{k+1} \approx \log(C) + 2\Big(\log(C) + 2y_k\Big)$$

对于一般的 $N > k$，

$$\begin{aligned} y_N &\approx (1 + 2 + \cdots + 2^{N-k-1})\log(C) + 2^{N-k}y_k \\ &= \frac{2^{N-k} - 1}{(2-1)}\log(C) + 2^{N-k}y_k \end{aligned}$$

若 $\log(C) + y_k < 0$，则当 $N \to \infty$，$y_N \to -\infty$，即 $e_N \to 0$。

问题 2：为什么该条件或者假设可以满足?

由于 $y_N \approx 2^{N-k}\Big(\log(C) + y_k\Big) < 0$，我们考虑

$$-y_N \approx 2^{N-k} \underbrace{D}_{\text{某个正数}}$$

$$\log(-y_N) \approx (N-k)\log(2) + \log(D)$$

大致意思就是

$$\underbrace{\log\Big(-\log(e_N)\Big)}_{=\log\log(1/e_N)} \approx (N-k)\log(2) + \log(D)$$

对于牛顿法，$\log\log(1/e_N)$ 是关于 N 的一次函数。

而对于普通梯度法里面，$\log(1/e_N)$ 是关于 N 的一次函数。

【这里我们解释了问题 1 的答案】

$\log(C) + y_k = \log(C) + \log(e_k)$ 为什么能保证是负数?

答案：我们先假设通过一些迭代 e_k 足够小（无论是由于初值选择很巧，或者由于一段迭代时间后产生这个效果），$\log(e_k)$ 是一个绝对值很大的负数，而 $\log(C)$ 是某个常数，只要 e_k 足够小，就能实现 $\log(C) + y_k < 0$。

因而我们可以得到如下重要结论：

纯牛顿法在真解附近一定收敛：

(1D) 若 $f''(x^*) \neq 0$，$f'''(x^*) \neq 0$，且初值 x_0 足够接近 x^*，则纯牛顿法一定会收敛，且收敛速度是超指数收敛。由于误差

$$\log\log(1/e_N) = \log(2)N + c$$

因此，为了满足误差小于 ϵ，我们可以选择 N 满足[1]

$$N = \log_2\log_2(1/\epsilon) - c$$

[1]为简化，取 log 的基底为 2

目录

定义 (基于文献 2)

对于任意的 $p \geq 1$，假设迭代过程收敛，若迭代误差 $e_k = \left\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^* \right\|$ 在 $k \to \infty$ 时满足

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k^p} = C \neq 0 \qquad \left(e_k \text{ 很小时，} e_{k+1} \approx C e_k^p\right)$$

则称该迭代过程为 p 阶收敛。

- 当 $p = 1$，线性收敛
- 当 $p > 1$，超线性收敛
- 当 $p = 2$，平方收敛

收敛阶数的含义

普通梯度下降法	一阶收敛	$\log\left(\frac{1}{e_N}\right)$ 为关于 N 的一次函数
Newton 法	2 阶收敛/平方收敛	$\log\left(\log\left(\frac{1}{e_N}\right)\right) \approx N\log(2) + \theta_2$
一般情况的算法	p 阶收敛	$\log\left(\log\left(\frac{1}{e_N}\right)\right) \approx N\log(p) + \theta_p$

其中 θ_2 和 θ_p 是一些和 N 无关的常数；$\log$ 的基底只要统一即可

目录

总结

主要需要掌握的知识：

- 能够复述牛顿法的形式
- 了解三种不同的推导出牛顿法的视角
- 能够复述写出阻尼牛顿法的伪代码
- 能够描述牛顿法和普通梯度法的误差在数值实验中的差异
- 能知道牛顿法是二阶格式，而普通梯度法是一阶格式；能够描述 p-阶格式的定义和含义

阅读作业 & 参考资料：

1. 课本第 9.5 章
2. 李庆扬、王能超、易大义，数值分析（第 5 版），华中科技大学出版社，2018，第 6.3 章