# 第 3 章：无约束优化

## (3) 牛顿法

授课教师：曹语

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

# 目录

# 背景和目标

第 3 章第 1 部分中，我们已经见过梯度下降法会沿着负梯度方向前进；第 2 部分中，我们又解释了它为什么会收敛，以及为什么条件数大时会变慢。

一个自然的问题是：

- 既然一阶信息只能告诉我们“ 朝哪里下降 ”，
- 那么如果再利用二阶信息，能不能更快地接近极小值点？

这节课讨论的就是一种典型的二阶方法：**牛顿法**。

# 学习目标

学完本节后，希望大家能够：

- 写出牛顿步 $\Delta \boldsymbol{x} = -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})$
- 用三种视角解释牛顿法为什么合理
- 说明牛顿步在什么条件下是下降方向
- 区分阻尼牛顿法与纯牛顿法
- 解释牛顿法局部二次收敛

# 回顾练习

- 若 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$，其中 $\boldsymbol{Q} \in \mathbf{S}^n_{++}$，怎样选一个正定矩阵 $\boldsymbol{P}$，使得最速下降方向

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

  更适合这个问题?

- 在第 2 部分里，为什么条件数大时梯度下降会变慢?

# 从最速下降到牛顿法

对于二次函数 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$，其中 $\boldsymbol{Q} \in \mathbf{S}^n_{++}$，若选择 $\boldsymbol{P} = \boldsymbol{Q}$，则最速下降方向变为

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

而对于一般函数，曲率矩阵会随位置变化，因此一个自然的局部选择是 $\boldsymbol{P} = \nabla^2 f(\boldsymbol{x})$。这和第 3.1 部分里“ 改造度量来改造下降方向 ”的思路一致。

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

这就是**牛顿步**。核心思想可以概括成一句话：

**用当前位置的曲率信息来“ 矫正 ”梯度方向。**

# 目录

# 视角 1：二阶近似的最优解

在点 $\boldsymbol{x}$ 附近，用二阶泰勒展开近似 $f$：

$$\hat{f}(\boldsymbol{x}+\boldsymbol{v}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{v}.$$

若 $\nabla^2 f(\boldsymbol{x}) \succ 0$，最小化这个二次模型可得

$$\nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})\boldsymbol{v} = 0,$$

即

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

直观理解：牛顿法每一步都在最小化当前位置附近的一个“ 局部二次模型 ”。

# 视角 2：Hessian 范数下的最速下降

对给定的正定矩阵 $\boldsymbol{P}$，定义范数

$$\|\boldsymbol{u}\|_{\boldsymbol{P}} = \left\|\boldsymbol{P}^{1/2}\boldsymbol{u}\right\|_2.$$

在这个范数下，最速下降方向与 $-\boldsymbol{P}^{-1}\nabla f(\boldsymbol{x})$ 同方向。

若在当前位置选 $\boldsymbol{P} = \nabla^2 f(\boldsymbol{x})$，就得到

$$\Delta\boldsymbol{x} \propto -\big(\nabla^2 f(\boldsymbol{x})\big)^{-1}\nabla f(\boldsymbol{x}).$$

直观理解：牛顿法是在当前曲率诱导的几何下看“ 最陡 ”。

# 视角 3：线性化最优性条件的解

极小值点满足一阶最优性条件

$$\nabla f(\boldsymbol{x}^\star) = \mathbf{0}.$$

若在当前点 $\boldsymbol{x}$ 附近对梯度做一阶展开，则下一步应该尽量让

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

更接近零。

于是我们把这个最优性条件线性化。

# 视角 3：线性化最优性条件的解（续）

在 $\boldsymbol{x}$ 附近做一阶展开：

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

令右边近似等于 $\mathbf{0}$，得到

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

直观理解：牛顿法是在每一步求解“ 线性化后的最优性方程 ”。

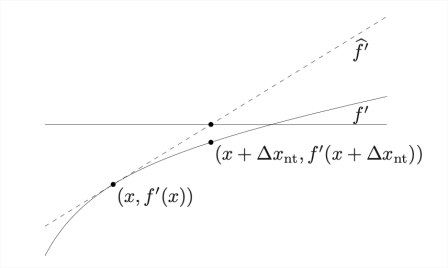
# 一维情形的直观图像

对于一维问题，若希望求解

$$f'(x) = 0,$$

则在点 $x$ 处对 $f'$ 做一阶近似：

$$f'(x+v) \approx f'(x) + f''(x)v.$$




令右边等于 $0$，得到

下降方向 $\quad v = -\dfrac{f'(x)}{f''(x)},$ $\quad \Longrightarrow \quad$ 纯牛顿法 $\quad x^+ = x - \dfrac{f'(x)}{f''(x)}.$

# 目录

# 牛顿步是下降方向吗?

假设 $\nabla^2 f(\boldsymbol{x}) \succ 0$，且 $\nabla f(\boldsymbol{x}) \neq \mathbf{0}$。

令

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

则方向导数为

$$\nabla f(\boldsymbol{x})^\top \Delta \boldsymbol{x} = -\nabla f(\boldsymbol{x})^\top \big(\nabla^2 f(\boldsymbol{x})\big)^{-1} \nabla f(\boldsymbol{x}).$$

由于 $\big(\nabla^2 f(\boldsymbol{x})\big)^{-1} \succ 0$，故

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

因此，在 Hessian 正定时，牛顿步确实是下降方向。

# 牛顿减量

确定方向后，考虑沿牛顿方向的一维函数

$$g(t) = f\big(\boldsymbol{x} + t\Delta\boldsymbol{x}\big).$$

对牛顿步，有

$$g'(0) = \nabla f(\boldsymbol{x})^\top \Delta\boldsymbol{x} = -\nabla f(\boldsymbol{x})^\top \big(\nabla^2 f(\boldsymbol{x})\big)^{-1} \nabla f(\boldsymbol{x}).$$

也就是说，$g'(0)$ 的绝对值正好反映了牛顿方向的局部下降强度。

# 牛顿减量：定义

我们记

$$\lambda(\boldsymbol{x}) = \sqrt{\nabla f(\boldsymbol{x})^\top \left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})}.$$

于是

$$g'(0) = -\lambda(\boldsymbol{x})^2.$$

该量称为**牛顿减量**。

# 牛顿减量的含义

牛顿减量可以理解为：

- 在牛顿方向上，$t = 0$ 处的瞬时下降速度大小
- 当前点离最优解还有多远的一个局部指标
- 一个适合牛顿法的停止准则

在课本里常使用

$$\frac{\lambda(\boldsymbol{x})^2}{2} \leq \epsilon$$

作为停止条件。

直观理解：$\lambda(\boldsymbol{x})$ 越小说明局部二次模型已经暗示再往下也降不了多少了。

# 阻尼牛顿法 / 谨慎牛顿法

**算法流程**

1. 计算牛顿步

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

2. 计算牛顿减量 $\lambda(\boldsymbol{x}^{(k)})$
3. 若 $\left(\lambda(\boldsymbol{x}^{(k)})\right)^2/2 \leq \epsilon$，则停止
4. 用回溯直线搜索选择步长 $t^{(k)} > 0$
5. 更新

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + t^{(k)} \Delta \boldsymbol{x}^{(k)}$$

若最终接受的步长始终为 $t^{(k)} = 1$，则称该阶段为**纯牛顿阶段**。

# 为什么需要阻尼?

**如果直接取 $t = 1$**

如果当前点离最优解较远时，局部二次模型未必足够准确，因此可能会：

- 下降不稳定
- 走得过远
- 进入性质较差的区域

**更稳妥的做法**

先用牛顿步给出方向，再用回溯直线搜索控制步长。

总结： 阻尼让牛顿法在远离真解时也更稳。

# 例子 1：$f(x) = x - \ln x$

考虑定义域为 $(0, +\infty)$ 上的函数

$$f(x) = x - \ln x.$$

**基本性质**

$$f'(x) = 1 - \frac{1}{x}, \qquad f''(x) = \frac{1}{x^2} > 0.$$

因此 $f$ 是严格凸函数，唯一极小值点为

$$x^\star = 1.$$

思考：对这个例子，牛顿步和迭代格式能不能显式写出来?

# 写出牛顿迭代

由 $f'(x) = 1 - \frac{1}{x}$ 和 $f''(x) = \frac{1}{x^2}$，可得牛顿步

$$\Delta x = -\frac{f'(x)}{f''(x)} = -\frac{1 - 1/x}{1/x^2} = x - x^2.$$

所以牛顿步就是 $x(1-x)$。

因此纯牛顿法为

$$x^{(k+1)} = x^{(k)} + \Delta x^{(k)} = x^{(k)} + x^{(k)} - \left(x^{(k)}\right)^2 = 2x^{(k)} - \left(x^{(k)}\right)^2.$$

阻尼牛顿法则写成

$$x^{(k+1)} = x^{(k)} + t^{(k)}\left(x^{(k)} - \left(x^{(k)}\right)^2\right), \qquad t^{(k)} \in (0, 1].$$

# 直观上：为什么 $t = 1$ 看起来安全?

设当前点 $x \in (0, 1)$，牛顿步为

$$\Delta x = x - x^2 = x(1 - x) > 0.$$

若取 $t = 1$，则新点为

$$x^+ = 2x - x^2 \in (x, 1).$$

这说明完整牛顿步不会把点推出定义域，而且会把点从 $x$ 推到 $(x, 1)$ 内，也就是更靠近最优解 $x^\star = 1$。

所以从直观上看，$t = 1$ 是一个很自然的候选步长。

但严格来说：回溯是否接受 $t = 1$，还要验证 Armijo 条件。

# 严格证明：为什么 Armijo 条件会接受 $t = 1$?

只需验证 Armijo 条件 $f(x + \Delta x) \le f(x) + \alpha f'(x)\Delta x$，其中 $\alpha \in (0, 1/2)$。

令 $s = 1 - x \in (0, 1)$，则 $x + \Delta x = 1 - s^2$，且 $f'(x)\Delta x = -s^2$。

$$\begin{aligned} f(x + \Delta x) - f(x) &= f(1 - s^2) - f(1 - s) \\ &= s - s^2 - \ln(1 + s) \\ &\le s - s^2 - \left( s - \frac{s^2}{2} \right) = -\frac{1}{2} s^2, \end{aligned}$$

这里用到了 $\ln(1 + s) \ge s - \frac{s^2}{2}$。

$$f(x + \Delta x) - f(x) \le -\frac{1}{2} s^2 \le -\alpha s^2 = \alpha f'(x)\Delta x$$

因为 $\alpha < 1/2$。

因此 $t = 1$ 满足 Armijo 条件，所以回溯直线搜索会接受 $t = 1$。

# 误差递推

令误差

$$e_k = \left| x^{(k)} - 1 \right|.$$

由纯牛顿迭代

$$x^{(k+1)} = 2x^{(k)} - \left(x^{(k)}\right)^2$$

可得

$$1 - x^{(k+1)} = 1 - 2x^{(k)} + \left(x^{(k)}\right)^2 = \left(1 - x^{(k)}\right)^2.$$

若 $x^{(0)} \in (0,1)$，则 $e_k = 1 - x^{(k)}$，从而

$$e_{k+1} = e_k^2.$$

这正是二次收敛最典型的形式。【实验部分见代码】

# 目录

# 收敛性：假设

为了得到一个比较完整的收敛结论，通常需要两类条件：

**条件 1：强凸且曲率有上下界**

$$m\boldsymbol{I}_{n\times n} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\boldsymbol{I}_{n\times n}$$

**条件 2：Hessian 变化不要太剧烈**

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

这两个条件分别控制：“ 曲率大小 ”与“ 曲率变化速度 ”。

# 收敛性：结论的两阶段图像

在上述假设下，阻尼牛顿法常被理解为有两个阶段：

1. 远离最优解时：依靠回溯直线搜索保证每一步都在稳定下降
2. 接近最优解后：回溯会接受 $t^{(k)} = 1$，算法进入纯牛顿阶段

进入纯牛顿阶段后，误差满足典型形式

$$e_{k+1} \le C e_k^2,$$

因此局部具有二次收敛。

这也是牛顿法比普通梯度下降快得多的根本原因。

# 局部二次收敛：一维直观推导

为简化起见，只看一维情形，并假设已经足够接近真解 $x^\star$。

纯牛顿法写为

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}.$$

若已经靠近真解，则把 $f'$ 在当前点附近再展开一层，就能看到主导误差来自二次项。

# 局部二次收敛：一维直观推导（续）

由于 $f'(x^\star)=0$，可得

$$x^{(k+1)}-x^\star = x^{(k)}-x^\star-\frac{f'(x^{(k)})}{f''(x^{(k)})}.$$

令函数 $g(x)=\frac{f'(x)}{f''(x)}$，对 $g$ 在 $x^\star$ 附近做二阶展开：

$$g(x^{(k)}) \approx \underbrace{g(x^\star)}_{=0}+\underbrace{g'(x^{(\star)})}_{=1}(x^{(k)}-x^\star)+\frac{g''(x^\star)}{2}(x^{(k)}-x^\star)^2$$

整理上述表达式可得，$x^{(k+1)}-x^{(k)}=-\frac{g''(x^\star)}{2}(x^{(k)}-x^\star)^2$，因此误差量会大致满足

$$e_{k+1}\approx Ce_k^2.$$

这里的重点是：下一步误差近似等于上一步误差的平方。

# 为什么会出现 $\log\log(1/\epsilon)$?

若误差满足

$$e_{k+1} \approx C e_k^2,$$

则取对数可得

$$\log e_{k+1} \approx \log C + 2\log e_k.$$

这说明：每做一步，“ 对数误差 ”大致又乘上一个 2。

当 $k$ 足够大、$e_k$ 足够小时，$\log e_k$ 会越来越快下降。

进一步看，$\log\big(-\log e_k\big) = \log\log(1/e_k)$ 与 $k$ 近似呈线性关系。

因此，为了使 $e_k \le \epsilon$，所需步数大致与 $\log\log(1/\epsilon)$ 同阶。

# 和梯度下降作比较

若普通梯度下降是线性收敛，则典型误差形式为

$$e_{k+1} \approx c e_k, \qquad 0 < c < 1.$$

因此梯度下降更像是“ 按固定比例缩小误差 ”。

这意味着 $\log(1/e_k)$ 关于 $k$ 近似是一次函数。

而牛顿法局部二次收敛时，$\log\log(1/e_k)$ 关于 $k$ 是一次函数。

一句话概括：

梯度下降是“ 稳步变小 ”，牛顿法在局部则是“ 越来越快地变小 ”。

# 目录

# $p$ 阶收敛的定义

设误差 $e_k = \left\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^\star \right\|$，并假设迭代收敛。

若对于某个 $p \geq 1$，有

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k^p} = C \neq 0,$$

则称该迭代过程具有$p$ 阶收敛。

也就是说，当误差足够小时，$e_{k+1} \approx C e_k^p$。

**特别地**

- $p = 1$：线性收敛
- $p > 1$：超线性收敛
- $p = 2$：二次收敛

# 收敛阶数的含义

**三种典型图像**

- 普通梯度下降法：$\log(1/e_k)$ 对 $k$ 近似线性
- 牛顿法：$\log\log(1/e_k)$ 对 $k$ 近似线性
- 一般 $p$ 阶方法：对应斜率与 $\log p$ 有关

**提示：**数值实验里即使我们未必能严格验证极限，但可通过作图观察趋势是否符合理论趋势。

# 目录

# 快速检查

问题：请快速回答下面三个问题：

**课堂小测**

1. 牛顿步的公式是什么?
2. 在什么条件下，牛顿步一定是下降方向?
3. 为什么牛顿法局部会比梯度下降快很多?

# 快速检查

问题：请快速回答下面三个问题：

**课堂小测**

1. 牛顿步的公式是什么?
2. 在什么条件下，牛顿步一定是下降方向?
3. 为什么牛顿法局部会比梯度下降快很多?

答案：1. $\Delta \boldsymbol{x} = -(\nabla^2 f(\boldsymbol{x}))^{-1} \nabla f(\boldsymbol{x})$。2. 当 $\nabla^2 f(\boldsymbol{x}) \succ 0$ 且 $\nabla f(\boldsymbol{x}) \neq 0$ 时，牛顿步是下降方向。3. 因为进入纯牛顿阶段后常有 $e_{k+1} \approx C e_k^2$，即局部二次收敛。

# 总结

本节要点：

- **牛顿步**： $\Delta \boldsymbol{x} = -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})$
- **三种视角**：二阶近似、Hessian 范数下的最速下降、线性化最优性条件
- **实际算法**：通常采用阻尼牛顿法，配合回溯直线搜索选步长
- **局部收敛**：进入纯牛顿阶段后，常表现出二次收敛

阅读作业 & 参考资料：

- 课本第 9.5 章
- 李庆扬等，《数值分析》第 5 版，第 6.3 章