最优化技术原理及应用

王冰川
bingcwang@csu.edu.cn
https://intleo.csu.edu.cn/index.html
中南大学自动化学院

无约束优化基础

主要参考中国科学院大学Xiao Wang老师PPT内容

主要内容

1. 数学基础
2. 最优解
3. 代表性算法

主要内容

1 数学基础

2 最优解

3 代表性算法

向量和矩阵

- 向量： $x \in \Re^n, x = (x_1, x_2, \dots, x_n)^{\mathrm{T}}$
- 内积： $x, y \in \Re^n, x^{\mathrm{T}} y = \sum_{i=1}^{n} x_i \, y_i$
- 矩阵： $A \in \Re^{m \times n}$
- 半正定： $A \in \Re^{n \times n}, \forall x \in \Re^n, x^{\mathrm{T}} A x \geq 0$
- 正定： $A \in \Re^{n \times n}, \forall x \in \Re^n, x^{\mathrm{T}} A x > 0$
- 正交： $Q \in \Re^{n \times n}, Q^{\mathrm{T}} Q = Q Q^{\mathrm{T}} = I$
- 特征值，特征向量： $Ax = \lambda x$

向量范数

- $l_1 - norm$： $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- $l_2 - norm$ ： $\|x\|_2 = (\sum_{i=1}^{n} x_i^2)^{1/2} = (x^{\mathrm{T}} x \quad)^{1/2}$
- $l_\infty - norm$ ： $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$
- $l_p - norm$ ： $\|x\|_2 = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$
- $\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty, \|x\|_\infty \le \|x\|_1 \le n\|x\|_\infty$
- Cauchy-Schwarz 不等式： $|x^{\mathrm{T}} y| \le \|x\|_2 \, \|y\|_2$

矩阵范数

- 定义：$A \in \Re^{m\times n}, \|A\| = \max_x \frac{\|Ax\|}{\|x\|}$
- $\|\mathrm{A}\|_1 = \max\limits_{j=1,\ldots,n} \sum_{i=1}^{m} |A_{ij}|$
- $\|\mathrm{A}\|_2 = (A^{\mathrm{T}} A)^{1/2}$最大的特征值
- $\|\mathrm{A}\|_\infty = \max\limits_{i=1,\ldots,m} \sum_{j=1}^{n} |A_{ij}|$
- *Frobenius*范数：$\|A\|_F = (\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2)^{1/2}$
- 条件数：$\kappa(A) = \|A\| \|A^{-1}\|$

- Frechet 可微：$f: \mathcal{R}^n \rightarrow \mathcal{R}$可微的充分条件：

$$\exists g \in \mathcal{R}^n, \lim_{y \to 0} \frac{f(x+y) - f(x) - g^T y}{\|y\|} = 0$$

- 梯度： $g(x) = \nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right) \in \mathcal{R}^n$

- Hessian: $H(x) = \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$

- 链式法则：$\alpha, \beta \in \mathcal{R}$

$$\frac{d\phi(\alpha(\beta))}{d\beta} = \phi'(\beta)\alpha'(\beta)$$

- 链式法则：$x, t \in \mathcal{R}; x = x(t); h(t) = f(x(t))$

$$\nabla h(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \nabla x_i(t) = \nabla x(t)^T \nabla f\big(x(t)\big)$$

- 直接求导: $D(f(x); p) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon p) - f(x)}{\epsilon} = \nabla f(x)^T p$

收敛速率（convergence rate）

$\{x_k\}$代表$\mathcal{R}^n$中收敛到x^*的一组序列，不同的收敛速率定义如下

- Q-linear：$\forall$充分大的$k, \exists r \in (0,1)$使得下式成立

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} \le r$$

- Q-superlinear：$\lim_{k\to\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = 0$

- Q-quadratic:$\forall$充分大的$k, \exists M$使得下式成立

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \le M$$

主要内容

- 一个无约束优化问题定义如下

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

其中， $f\colon \mathcal{R}^n \to \mathcal{R}$是一个平滑函数

- 一般来说，有关f的信息并不容易获得，使用尽量少的信息获得最优值是值得考虑的。

- 全局最优：x^*是全局最优解，如果满足以下条件

$$f(x^*) \le f(x), \forall x \in \mathcal{R}^n$$

- 局部最优： x^*是局部最优解，如果在其领域$\aleph$满足以下条件

$$f(x^*) \le f(x), \forall x \in \aleph$$

- 严格局部最优： x^*是严格局部最优解

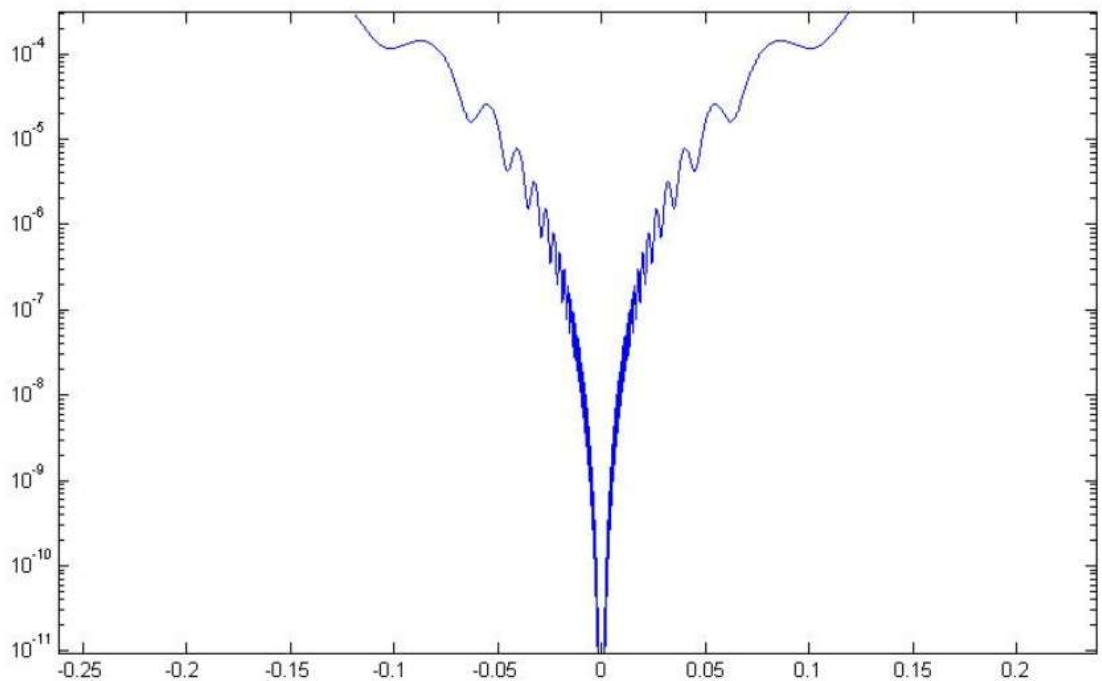
$$f(x^*) \le f(x), \forall x \in \aleph \text{ with } x \neq x^*$$

- 孤立（isolated）局部最优：x^*是孤立局部最优解

x^*是邻域内唯一的局部最优解

➢ 一个例子

$$f(x) = x^4 \cos\left(\frac{1}{x}\right) + 2x^4$$



局部最优判定

- 通过定义判定，如果目标函数f是平滑的，则可以根据根据其一阶导数（梯度向量）、二阶导数（Hessian矩阵）进行判定，主要是利用Taylor’s Theorem判定

- Taylor’s Theorem：假设$f\colon \mathcal{R}^n \rightarrow \mathcal{R}$连续可微，并且$p \in \mathcal{R}^n$,则有

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \quad , for\ some\ t \in (0,1)$$

如果f连续二阶可导，则有

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p d_t$$

- Taylor's Theorem：假设$f: \mathcal{R}^n \to \mathcal{R}$连续可微，并且$p \in \mathcal{R}^n$,则有

$$f(x+p) = f(x) + \nabla f(x+tp)^T p \quad , for\ some\ t \in (0,1)$$

如果f连续二阶可导，则有

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p d_t$$

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+tp) p \quad ,$$

$$for\ some\ t \in (0,1)$$

局部最优判定

- First-Order Necessary Conditions：假设x^*是f的局部最优， f在x^*的一个开邻域内（D）连续可微，则$\nabla f(x^*) = 0$

- 简要证明：反证法证明，假设$\nabla f(x^*) \neq 0$，定义$p = -\nabla f(x^*)$，因为D为开集合并且f连续可微，那么存在一个足够小的数t，使得$\forall \tau \in [0, t], x^*+\tau p \in D$并且$\nabla f(x^*+\tau p)^T p < 0$，根据Taylor's Theorem，存在$\theta \in (0, t)$满足

$$f(x^*+\tau p) < f(x^*) + \tau \nabla f(x^*+\theta p)^T p < f(x^*)$$

这与$f(x^*)$是局部最优矛盾

- Second-Order Necessary Conditions：假设x^*是f的局部最优， f在x^*的一个开邻域内（ D ）连续二阶可微，则$\nabla f(x^*) = 0$， $\nabla^2 f(x^*)$半正定

- 简要证明：反证法，假设$\nabla^2 f(x^*)$不是半正定，则存在向量p使得$p^T\nabla^2 f(x^*)p < 0$。存在一个足够小的数t，使得$\forall \tau \in [0, t], x^*+\tau p \in D$并且$p^T\nabla^2 f(x^*+\tau p)p < 0$，根据Taylor's Theorem， 存在$\theta \in (0, t)$满足

$$f(x^*+\tau p) < f(x^*) + \nabla f(x^*)\tau p + \tau^2 p^T \nabla^2 f(x^*+\theta p)p < f(x^*)$$

这与$f(x^*)$是局部最优矛盾

局部最优判定

- 必要条件是假设x^*为局部最优的前提下，得出$\nabla f(x^*)$和$\nabla^2 f(x^*)$的相关性质
- 如果$\nabla f(x^*) = 0$，则x^*被称为稳定点
- 稳定点不一定是局部最优点 $f(x) = x^3$

- Second-Order Sufficient Conditions: 假设$f: D \to \mathcal{R}$二阶可导，$x^* \in D, \nabla f(x^*) = 0, \nabla^2 f(x^*)$正定，则$x^*$是一个严格的局部最优解

- 简要证明：以x^*为中心，选择一个半径（r）充分小的邻域，对于该邻域内的任意向量$\|p\| \le r$，$\nabla^2 f(x^* + tp) > 0$，根据Taylor's Theorem，存在$t \in (0,1)$，使得满足

$$f(x^* + p) = f(x^*) + \nabla f(x^*) p + p^T \nabla^2 f(x^* + tp) \\ > f(x^*)$$

注意：严格的局部最优解不一定满足该充分条件！

局部最优判定

x^*是凸优化问题的局部最优，则对于可行域内的任意一个解y,有$f(y) \geq f(x^*)$:

- 由于局部最优性， $z \in \Omega \cap \aleph, f(z) \geq f(x^*)$;
- 由于可行域的凸性，x和y的连线也在可行域内；
- 在$\Omega \cap \aleph$中的连线上选择一个不等于x^*的解： $z = \alpha y + (1-\alpha)x^*$
- 凸函数性质： $f(z) \leq \alpha f(y) + (1-\alpha) f(x^*) = f(x^*) + \alpha(f(y) - f(x^*))$
- 因为$f(z) \geq f(x^*)$，所以$\alpha(f(y) - f(x^*)) \geq 0$，即$f(y) - f(x^*)$

主要内容

1 数学基础

2 最优解

3 代表性算法

- 选择初始点x_0，可根据先验知识选择，亦可任意选择
- 以此为起点，优化算法产生一个序列

$$x_1, x_2, \ldots$$

- 当达到某一个终止条件时，算法停止

- 如何产生解序列是一个优化算法的重点
- 终止条件包括解没有改进，或者迭代次数达到给定值等等

算法使用有关 x_k 处函数 f 的信息，也可能使用来自早期迭代的信息，决定如何从x_k 移动到x_{k+1}

- **单调算法**：找到函数值低于 x_k的新迭代 x_{k+1}
- **非单调算法**：不坚持每一步都减小f，而是要求f在规定的m次迭代后减小，即$f(x_k) < f(x_{k+m})$

从当前点 x_k移动到新的迭代 x_{k+1}有两种代表策略：**线搜索**和**信赖域方法**

- 选择方向向量：p_k
- 从当前迭代 x_k 开始沿此方向搜索具有较低函数值的新迭代

$$x_{k+1} \leftarrow x_k + \alpha_k\, p_k,\ \ f(x_k) < f(x_{k+1})$$

- 在新迭代点，计算新的方向（ p_{k+1} ）和步长（ α_{k+1} ），并重复以上过程达到终止条件

- 如何确定搜索步长α_k？
- 当搜索方向p_k确定以后，可以通过最小化以下问题来确定α_k

$$\alpha_k = \min_{\alpha>0} f(x_k + \alpha\, p_k)$$

- 一般来说，有两种方法设定α_k
  - 直接求解以上问题，但是优化过程将添加额外的开销，而且没有必要
  - 线搜索算法可以生成有限数量的试验步长，直到找到一个近似最小步长的α_k

信赖域方法关键步骤如下：

- 利用有关 f 的信息构建模型函数 （model function） m_k，使其在x_k附近与f相近
- 但是当x远离 x_k时，模型m_k可能不能很好地对 f 近似，因此需要将基于 m_k的 搜索限制在 x_k 周围的信赖域 （trust region）内

信赖域方法通过求解以下问题得到搜索方向：

$$p_k = \min_p m_k(x_k + p)\,, x_k + p\text{在信赖域内}$$

如果f不能充分下降，则对信赖域进行收缩

- 通常信赖域定义为半径大小为σ的球体：$\|p\| \le \sigma$；也可以使用椭球体和超立方体等定义信赖域
- 模型函数m_k通常定义为二次函数

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p$$

- m_k和f在x_k处的一阶导数相同， B_k一般为Hessian矩阵或者其近似

搜索策略比较

线搜索策略和信赖域策略的区别在于如何确定**搜索方向**和**搜索步长**

- 线搜索策略先确定搜索方向p_k，在此基础上确定搜索步长α_k
- 信赖域策略先确定信赖域大小（对搜索步长进行了限制），在此基础上确定搜索方向
- 后续章节会对两种搜索策略进行详细介绍

Thanks for the attentions!

Q&A