最优化方法

无约束问题的梯度下降法

张伯雷

南京邮电大学 计算机学院

bolei.zhang@njupt.edu.cn

http://bolei-zhang.github.io/course/opt.html

目录

上节课内容回顾

泰勒展开

$$f(x)=f\left(a\right)+\frac{f'\left(a\right)}{1!}\left(x-a\right)+\frac{f''\left(a\right)}{2!}\left(x-a\right)^2+\cdots+\frac{f^n\left(a\right)}{n!}\left(x-a\right)^n+o[\left(x-a\right)^n].$$

上节课内容回顾

泰勒展开

$$f(x)=f\left(a\right)+\frac{f'\left(a\right)}{1!}\left(x-a\right)+\frac{f''\left(a\right)}{2!}(x-a)^2+\cdots+\frac{f^n\left(a\right)}{n!}(x-a)^n+o[(x-a)^n].$$

凸函数的一阶条件

$$f(y)\geq f(x)+\nabla f(x)^T(y-x)$$

上节课内容回顾

泰勒展开

$$f(x)=f\left(a\right)+\frac{f'\left(a\right)}{1!}\left(x-a\right)+\frac{f''\left(a\right)}{2!}\left(x-a\right)^2+\cdots+\frac{f^n\left(a\right)}{n!}\left(x-a\right)^n+o[(x-a)^n].$$

凸函数的一阶条件

$$f(y)\geq f(x)+\nabla f(x)^T(y-x)$$

KKT条件

$$\begin{aligned}f_i\left(x^\star\right)&\leq 0,\ i=1,\ldots,m\\ h_i\left(x^\star\right)&=0,\ i=1,\ldots,p\\ \lambda_i^\star&\geq 0,\ i=1,\ldots,m\\ \lambda_i^\star f_i\left(x^\star\right)&=0,\ i=1,\ldots,m\end{aligned}$$

$$\nabla f_0\left(x^\star\right)+\sum_{i=1}^m\lambda_i^\star\nabla f_i\left(x^\star\right)+\sum_{i=1}^p\nu_i^\star\nabla h_i\left(x^\star\right)=0,$$

无约束优化问题

本章中，我们关注求解无约束的优化问题

$$\min \quad f(x)$$

- 其中，$f:\mathbb{R}^n \to \mathbb{R}$为二次连续可微的凸函数

无约束优化问题

本章中，我们关注求解无约束的优化问题

$$\min \quad f(x)$$

- 其中，$f:\mathbb{R}^n \to \mathbb{R}$为二次连续可微的凸函数

根据无约束凸优化问题的最优性条件(可以直接从KKT条件推出)，最优解x^*满足以下的条件

$$\nabla f(x^*) = 0$$

优化算法

例1:

$$\min \quad (1/2)x^T Px + q^T x + r$$

其中$P \in S_+^n$.

优化算法

例1:

$$\min \quad (1/2)x^T Px + q^T x + r$$

其中$P \in S_+^n$.

例2:

$$\min \quad f(x) = \log(\sum_{i=1}^{n} \exp(a_i^T x + b_i))$$

下降方法

迭代算法：从一个初始值$x^{(0)}$出发，计算出一个序列的x的值$x^{(0)}, x^{(1)}, \ldots \in \mathrm{dom} f$，使得当$k \to \infty$时，$f\left(x^{(k)}\right) \to p^*$

下降方法

迭代算法：从一个初始值$x^{(0)}$出发，计算出一个序列的x的值$x^{(0)}, x^{(1)}, \ldots \in \mathrm{dom} f$，使得当$k \to \infty$时，$f\left(x^{(k)}\right) \to p^*$

- 实际情况下，算法会在满足一定条件时终止，例如$\left|f\left(x^{(k)}\right) - p^*\right| \leq \epsilon$，$\epsilon > 0$

下降方法

迭代算法：从一个初始值$x^{(0)}$出发，计算出一个序列的x的值$x^{(0)}, x^{(1)}, \ldots \in \mathrm{dom} f$，使得当$k \to \infty$时，$f\left(x^{(k)}\right) \to p^*$

- 实际情况下，算法会在满足一定条件时终止，例如$\left|f\left(x^{(k)}\right) - p^*\right| \leq \epsilon$，$\epsilon > 0$

算法：通用下降方法

1. 输入：一个初始点$x^{(0)} \in \mathrm{dom} f$
2. 重复：
3. 找到一个下降的方向$\Delta x^{(k)}$
4. 选择一个步长$\alpha^{(k)}$
5. 更新：$x^{(k+1)} = x^{(k)} + \alpha^{(k)} \Delta x^{(k)}$
6. Until：满足停止条件

下降方向

为了找到合适的搜索方向，希望在每次迭代之后，都有

$$f_0(x^{(k+1)}) < f_0(x^{(k)})$$

下降方向

为了找到合适的搜索方向，希望在每次迭代之后，都有

$$f_0(x^{(k+1)}) < f_0(x^{(k)})$$

- 根据凸函数的一阶条件：$\nabla f_0(x^{(k)})^T(y - x^{(k)}) \geq 0$，可以推出

$$\nabla f_0(x^{(k)})^T \Delta x^{(k)} < 0$$

- 即下降方向需要和梯度方向成钝角

目录

步长

例：考虑一个二次函数的优化问题$\min f(x) = x^2$。取初始点为$x = 1$，此时的方向只有两种选择$\{-1, +1\}$。取$\Delta x^{(k)} = -sign(x^{(k)})$.

步长

例：考虑一个二次函数的优化问题$\min f(x) = x^2$。取初始点为$x = 1$，此时的方向只有两种选择$\{-1, +1\}$。取$\Delta x^{(k)} = -sign(x^{(k)})$.

考虑以下两种步长

$$\alpha_1^{(k)} = \frac{1}{3^{k+1}}, \alpha_2^{(k)} = 1 + \frac{2}{3^{k+1}}$$

步长

例：考虑一个二次函数的优化问题$\min f(x) = x^2$。取初始点为$x = 1$，此时的方向只有两种选择$\{-1, +1\}$。取$\Delta x^{(k)} = -sign(x^{(k)})$.

考虑以下两种步长

$$\alpha_1^{(k)} = \frac{1}{3^{k+1}}, \alpha_2^{(k)} = 1 + \frac{2}{3^{k+1}}$$

可以得到

$$x_1^{(k)} = \frac{1}{2}(1 + \frac{1}{3^k}), x_2^{(k)} = \frac{(-1)^k}{2}(1 + \frac{1}{3^k})$$

精确线搜索算法 (Exact Line Search)

假设方向$\Delta x^{(k)}$ 给定，如何选择步长$\alpha^{(k)}$

$$\alpha^{(k)} = \arg \min_{\alpha^{(k)}>0} f_0(x^{(k)} + \alpha^{(k)} \Delta x^{(k)})$$

精确线搜索算法 (Exact Line Search)

假设方向$\Delta x^{(k)}$ 给定，如何选择步长$\alpha^{(k)}$

$$\alpha^{(k)} = \arg \min_{\alpha^{(k)}>0} f_0(x^{(k)} + \alpha^{(k)} \Delta x^{(k)})$$

- 一维、凸优化问题

精确线搜索算法 (Exact Line Search)

假设方向$\Delta x^{(k)}$ 给定，如何选择步长$\alpha^{(k)}$

$$\alpha^{(k)} = \arg\min_{\alpha^{(k)}>0} f_0(x^{(k)} + \alpha^{(k)}\Delta x^{(k)})$$

- 一维、凸优化问题
- （f 为凸函数当且仅当对所有的$x \in \mathrm{dom} f$ 与v, 满足$x + tv \in \mathrm{dom} f$ 的函数$g(t) = f(x + tv)$ 为凸函数.）

精确线搜索算法 (Exact Line Search)

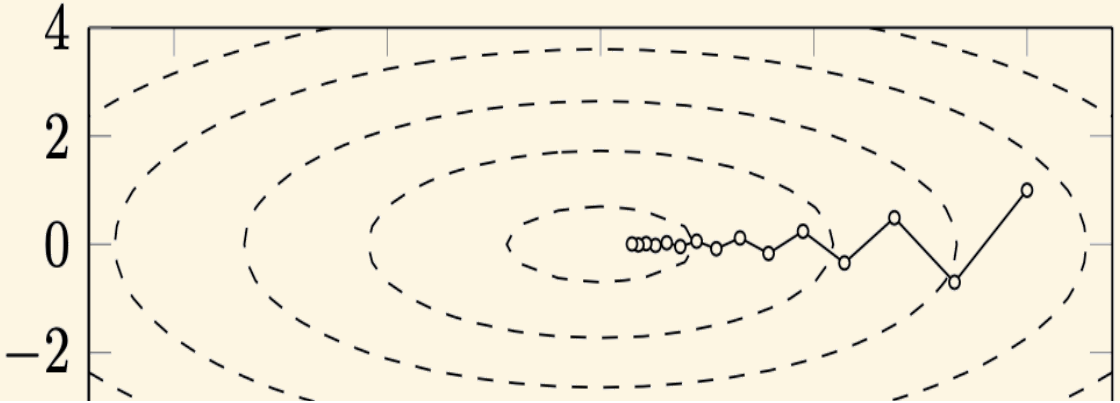
假设方向$\Delta x^{(k)}$ 给定，如何选择步长$\alpha^{(k)}$

$$\alpha^{(k)} = \arg\min_{\alpha^{(k)}>0} f_0(x^{(k)} + \alpha^{(k)}\Delta x^{(k)})$$

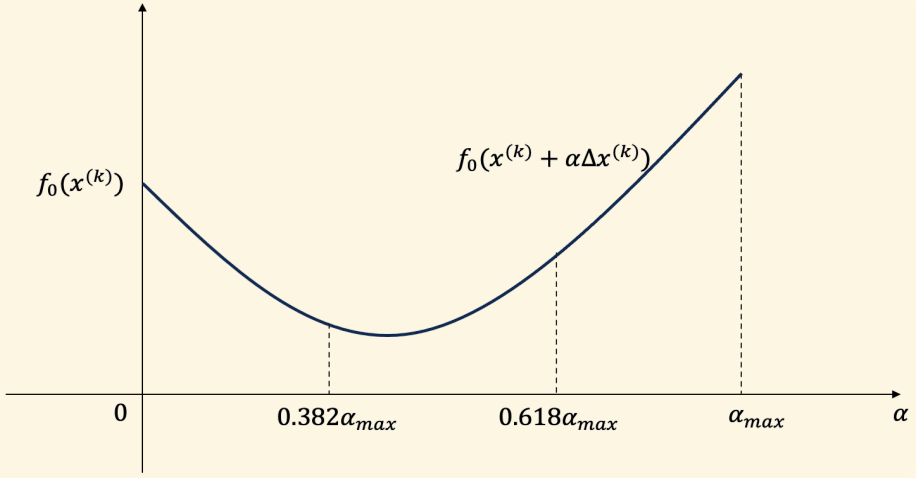
- 一维、凸优化问题
- （f 为凸函数当且仅当对所有的$x \in \mathrm{dom} f$ 与v, 满足$x + tv \in \mathrm{dom} f$ 的函数$g(t) = f(x + tv)$ 为凸函数.）
- 但是，由于需要在每一步的计算该步长，代价较高

精确步长搜索

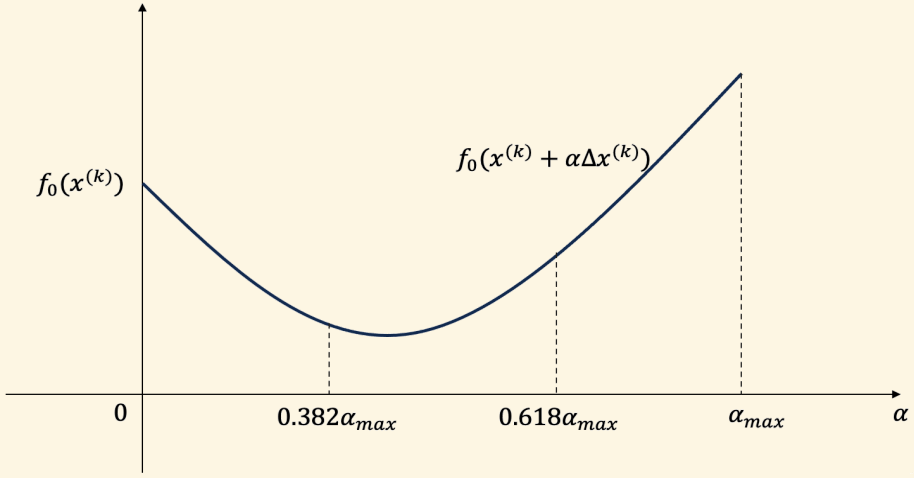
设二次函数$f(x,y)=x^2+10y^2$, 初始点$(x^{(0)},y^{(0)})$ 取为$(10,1)$, 假设搜索方向为$(-20,-20)$, 精确步长应该是多少?



黄金分割法 (Inexact Line Search)



黄金分割法 (Inexact Line Search)



算法：黄金分割法（优选法，华罗庚等）

1. 输入：已知$f_0(x^{(k)})$与$f_0(x^{(k)}+\alpha_{\max}\Delta x^{(k)})$, 初始区间设为$[a,b]$, 这里$a=0, b=\alpha_{\max}$.
2. 重复：
3. 　分别计算$f_0(x_1)=f_0(x^{(k)}+0.382(b-a)\Delta x^{(k)})$与$f_0(x_2)=f_0(x^{(k)}+0.618(b-a)\Delta x^{(k)})$
4. 　如果$f_0(x_1)<f_0(x_2)$，说明最小值区间在$[a,x_2]$, 令$b=x_2$
5. 　如果$f_0(x_1)>f_0(x_2)$，说明最小值区间在$[x_1,b]$, 令$a=x_1$
6. Until：满足停止条件

目录

下降方向

为了找到合适的搜索方向，希望在每次迭代之后，都有

$$f_0(x^{(k+1)}) < f_0(x^{(k)})$$

下降方向

为了找到合适的搜索方向，希望在每次迭代之后，都有

$$f_0(x^{(k+1)}) < f_0(x^{(k)})$$

- 根据凸函数的一阶条件：$\nabla f_0(x^{(k)})^T(y - x^{(k)}) \geq 0$，可以推出

$$\nabla f_0(x^{(k)})^T \Delta x^{(k)} < 0$$

- 即下降方向需要和梯度方向成钝角

梯度下降法

方向选择为当前点的负梯度方向

$$\Delta x^{(k)} = -\nabla f_0(x^{(k)})$$

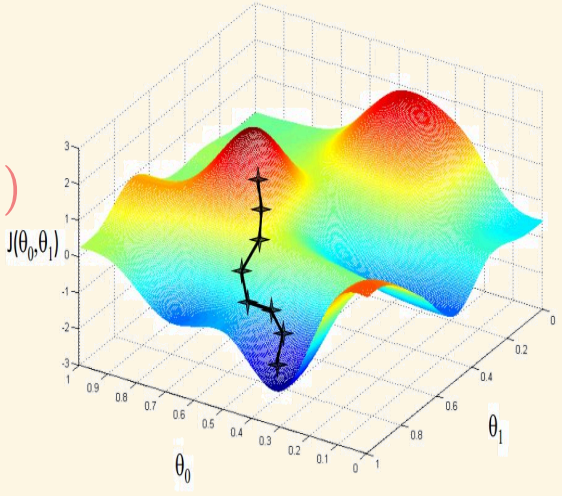
1. 输入：一个初始点$x^{(0)} \in \mathrm{dom} f$
2. 重复：
3. 　　选择一个步长 $\alpha^{(k)} = \arg\min_{\alpha^{(k)}>0} f_0(x^{(k)} + \alpha^{(k)}\Delta x^{(k)})$
4. 　　$\Delta x^{(k)} = -\nabla f_0(x^{(k)})$
5. 　　更新：$x^{(k+1)} = x^{(k)} + \alpha^{(k)}\Delta x^{(k)}$
6. Until：满足停止条件

梯度下降法

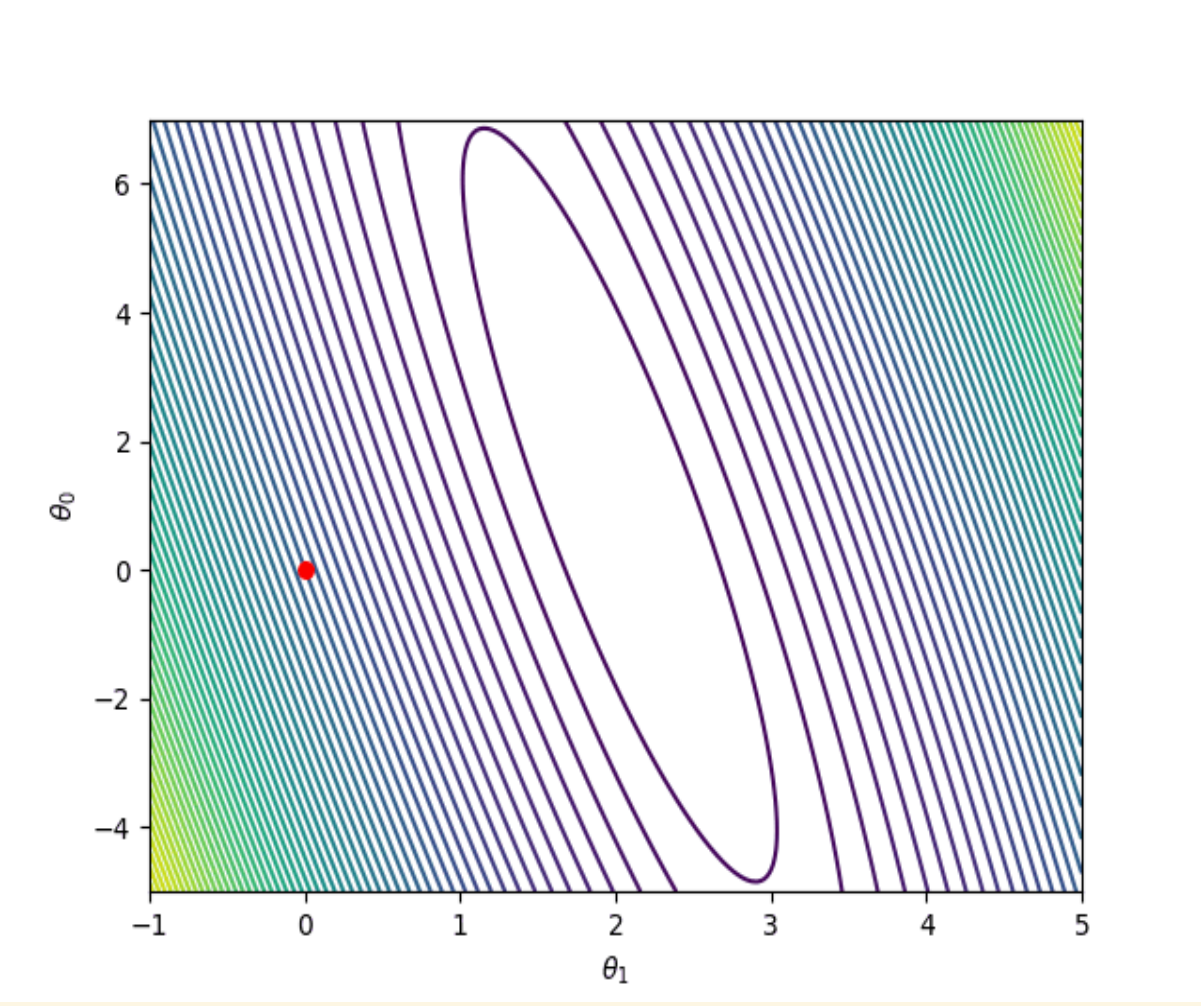
方向选择为当前点的负梯度方向

$$\Delta x^{(k)} = -\nabla f_0(x^{(k)})$$

1. 输入：一个初始点$x^{(0)} \in \mathrm{dom} f$
2. 重复：
3. 　　选择一个步长 $\alpha^{(k)} = \arg\min_{\alpha^{(k)}>0} f_0(x^{(k)} + \alpha^{(k)}\Delta x^{(k)})$
4. 　　$\Delta x^{(k)} = -\nabla f_0(x^{(k)})$
5. 　　更新：$x^{(k+1)} = x^{(k)} + \alpha^{(k)}\Delta x^{(k)}$
6. Until：满足停止条件

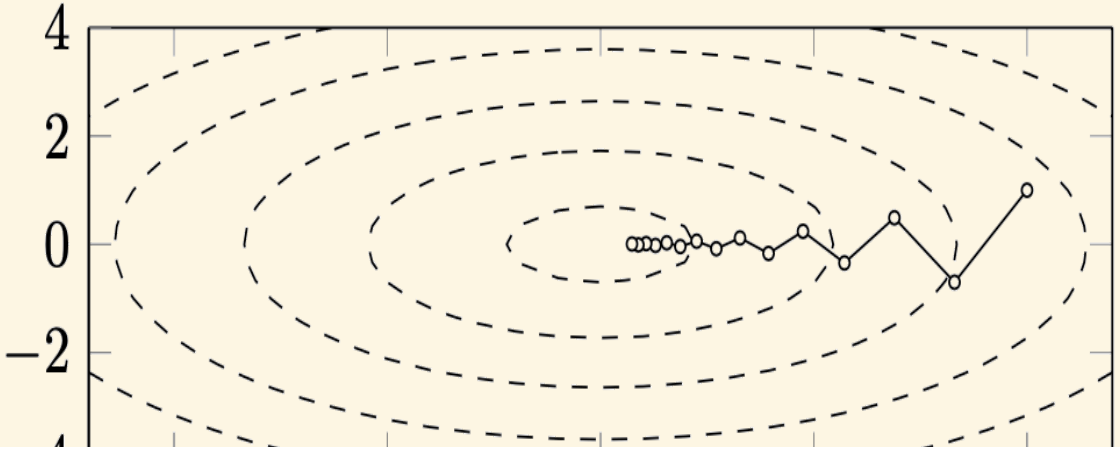


梯度下降法动画展示



二次函数的梯度法

设二次函数$f(x,y)=x^2+10y^2$, 初始点$(x^{(0)},y^{(0)})$ 取为$(10,1)$, 取固定步长$\alpha^{(k)}=0.085$. 我们使用梯度法$x^{(k+1)}=x^{(k)}-\alpha^{(k)}\nabla f(x^{(k)})$ 进行1次迭代，计算迭代一次之后的(x,y)的值



终止条件

- $|f_0(x^{(k)}) - p^*| \leq \epsilon$
- $|x^{(k)} - x^{(k+1)}| \leq \epsilon$
- $|f_0(x^{(k)}) - f_0(x^{(k+1)})| \leq \epsilon$
- $\nabla f_0(x) \leq \epsilon$

终止条件

- $|f_0(x^{(k)}) - p^*| \leq \epsilon$
- $|x^{(k)} - x^{(k+1)}| \leq \epsilon$
- $|f_0(x^{(k)}) - f_0(x^{(k+1)})| \leq \epsilon$
- $\nabla f_0(x) \leq \epsilon$

以上的条件都有一定的缺点

收敛性分析-函数的强凸性

一个函数f是强凸的，则$\exists m > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \succeq mI$

收敛性分析-函数的强凸性

一个函数f是强凸的，则$\exists m > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \succeq mI$

- 性质： $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} ||y - x||_2^2$

收敛性分析-函数的强凸性

一个函数f是强凸的，则$\exists m > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \succeq mI$

- 性质：$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}||y - x||_2^2$

还可以得到以下性质

- $f(x) - p^* \leq \frac{1}{2m}||\nabla f(x)||_2^2$
- $||x^* - x||_2 \leq \frac{2}{m}||\nabla f(x)||_2$

收敛性分析-函数的强凸性

一个函数f是强凸的，则$\exists m > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \succeq mI$

- 性质： $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} ||y - x||_2^2$

还可以得到以下性质

- $f(x) - p^* \leq \frac{1}{2m} ||\nabla f(x)||_2^2$
- $||x^* - x||_2 \leq \frac{2}{m} ||\nabla f(x)||_2$

结论: 对于强凸的函数$f_0(x)$, 当$\nabla f_0(x) \to 0$时， $f_0(x)$趋近于最优值, x逼近最优解

收敛性分析-函数的光滑性

一个函数f是光滑的， 则$\exists M > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \preceq MI$

收敛性分析-函数的光滑性

一个函数f是光滑的， 则$\exists M > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \preceq MI$

根据函数的光滑性，可以得到以下性质

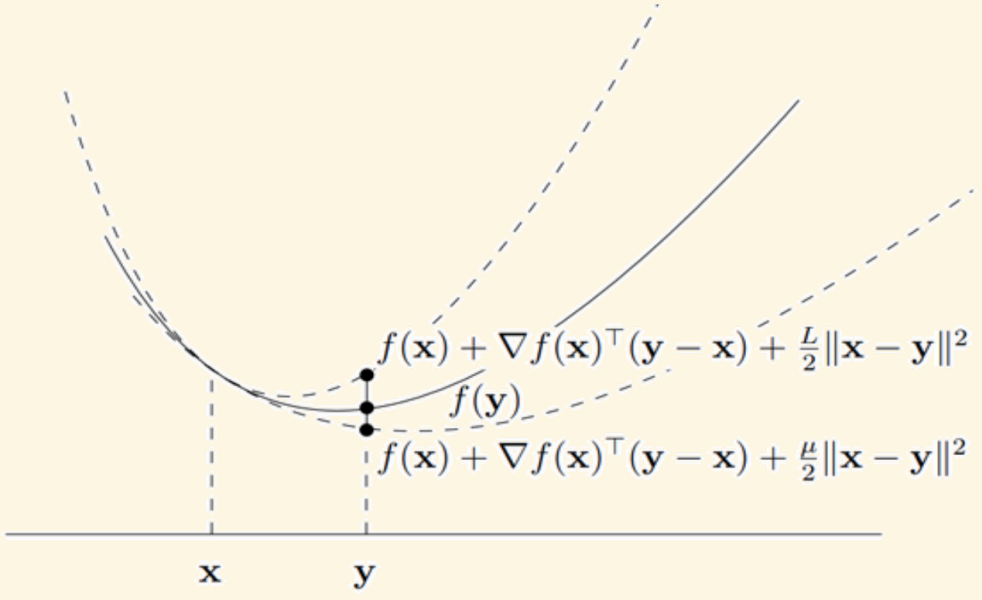
- $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} ||y - x||_2^2$
- $f(x) - p^* \geq \frac{1}{2M} ||\nabla f(x)||_2^2$

收敛性分析-函数的光滑性

一个函数f是光滑的， 则$\exists M > 0$, 使得$\forall x \in \mathrm{dom} f, \nabla^2 f(x) \preceq MI$

根据函数的光滑性，可以得到以下性质

- $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} ||y - x||_2^2$
- $f(x) - p^* \geq \frac{1}{2M} ||\nabla f(x)||_2^2$



梯度下降法-收敛性分析

假设: f_0是二阶可微的，而且具有强凸性, $MI \succeq \nabla^2 f(x) \succeq mI$

线性收敛的定义:

$$\frac{\left\| f(x^{k+1}) - f(x^*) \right\|}{\left\| f(x^0) - f(x^*) \right\|} < c^k$$

收敛性证明 (精确线搜索)

1. 定义函数

$$\tilde{f}(\alpha) = f(x^{(k)} + \alpha \Delta x^{(k)}) = f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

收敛性证明 (精确线搜索)

1. 定义函数

$$\tilde{f}(\alpha) = f(x^{(k)} + \alpha \Delta x^{(k)}) = f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

2. 根据光滑性:

$$f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)}) + \nabla f(x^{(k)})^T(-\alpha \nabla f(x^{(k)})) + \frac{M}{2}|| - \alpha \nabla f$$

收敛性证明 (精确线搜索)

1. 定义函数

$$\tilde{f}(\alpha) = f(x^{(k)} + \alpha \Delta x^{(k)}) = f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

2. 根据光滑性:

$$f(x^{(k)} - \alpha \nabla f(x^{(k)})) \leq f(x^{(k)}) + \nabla f(x^{(k)})^T(-\alpha \nabla f(x^{(k)})) + \frac{M}{2}|| - \alpha \nabla f$$

3. 等价于:

$$\tilde{f}(\alpha) \leq f(x^{(k)}) - \alpha ||\nabla f(x^{(k)}))||_2^2 + \frac{M\alpha^2}{2}||\nabla f(x^{(k)}))||_2^2$$

收敛性证明 (精确线搜索)

4. 对以上不等式两边分别取极小(左边$\alpha = \alpha_{exact}$, 右边$\alpha = \frac{1}{M}$), 则

$$f(x^{(k+1)}) = \tilde{f}(\alpha_{exact}) \leq f(x^{(k)}) - \frac{1}{2M}||\nabla f(x^{(k)}))||_2^2$$

收敛性证明 (精确线搜索)

4. 对以上不等式两边分别取极小(左边$\alpha = \alpha_{exact}$, 右边$\alpha = \frac{1}{M}$), 则

$$f(x^{(k+1)}) = \tilde{f}(\alpha_{exact}) \leq f(x^{(k)}) - \frac{1}{2M}||\nabla f(x^{(k)}))||_2^2$$

5. 联合不等式$p^* \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2$, 可得

$$f(x^{(k+1)}) - p^* \leq (1 - \frac{m}{M})(f(x^{(k)}) - p^*)$$

收敛性证明 (精确线搜索)

4. 对以上不等式两边分别取极小(左边$\alpha = \alpha_{exact}$, 右边$\alpha = \frac{1}{M}$)， 则

$$f(x^{(k+1)}) = \tilde{f}(\alpha_{exact}) \leq f(x^{(k)}) - \frac{1}{2M}||\nabla f(x^{(k)}))||_2^2$$

5. 联合不等式$p^* \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2$，可得

$$f(x^{(k+1)}) - p^* \leq (1 - \frac{m}{M})(f(x^{(k)}) - p^*)$$

6. 令$c = 1 - m/M$，可以得到

$$f(x^{(k)}) - p^* \leq c^k(f(x^{(0)}) - p^*)$$

即随着$k \to \infty$, $f(x^{(k)})$趋近于p^*

梯度下降法的收敛性

结论：梯度下降法可以线性收敛，在最多 $\frac{\log((f(x^{(0)})-p^*)/\epsilon)}{\log(1/c)}$ 步之后收敛到 $f\left(x^{(k)}\right)-p^* \leq \epsilon$

梯度下降法的收敛性

结论：梯度下降法可以线性收敛，在最多 $\frac{\log((f(x^{(0)})-p^*)/\epsilon)}{\log(1/c)}$ 步之后收敛到 $f\left(x^{(k)}\right)-p^* \leq \epsilon$

- 根据分子可以看出，收敛速度取决于初始点的选择$x^{(0)}$，以及需求的精度 ϵ

梯度下降法的收敛性

结论：梯度下降法可以线性收敛，在最多 $\frac{\log((f(x^{(0)})-p^*)/\epsilon)}{\log(1/c)}$ 步之后收敛到 $f(x^{(k)}) - p^* \leq \epsilon$

- 根据分子可以看出，收敛速度取决于初始点的选择$x^{(0)}$，以及需求的精度 ϵ
- 分母中$c = 1 - m/M$，可以得出$\log(1/c) = -\log(1 - m/M) \approx m/M$，说明当这个值越小，收敛越慢

总结

- 对于无约束可微的凸优化问题，通过迭代法求解最优解
- 迭代法中有三个因素：步长、方向、停止条件
- 步长搜索为线搜索方法，分为精确线搜索和不精确线搜索
- 方向可以采用负梯度的方向
- 梯度下降法可以线性收敛

作业

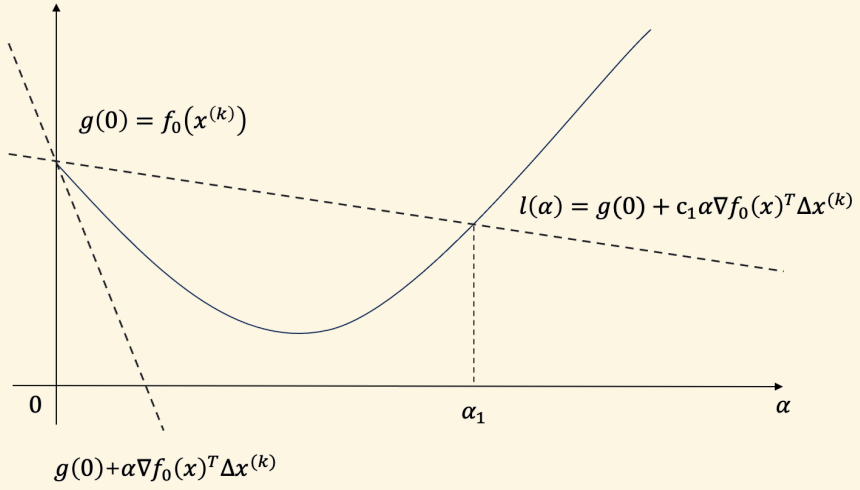
1. 对$f(x) = -\log x + x$，取步长为1，初始点为$x^{(0)} = 3$，经过一步的固定步长0.1的梯度下降法后的$x^{(1)}$取值为多少？
2. 对于函数$f(x) = x_1^2 + x_2$, 取初始点为$x^{(0)} = (2, 1)$，经过一步的精确步长的梯度下降法后的$x^{(1)}$取值为多少？
3. (选做) 证明强凸性的两个性质：
   - $f(x) - p^* \leq \frac{1}{2m}||\nabla f(x)||_2^2$
   - $||x^* - x||_2 \leq \frac{2}{m}||\nabla f(x)||_2$
4. (选做) 证明例2（Page 5）中的log-sum-sup为凸函数

Armijo准则 (Inexact Line Search)

- 对于$c_1 \in (0,1)$，如果$f_0(x+\alpha\Delta x) \leq f_0(x) + c_1\alpha\nabla f(x)^T\Delta x$，则步长$\alpha$满足Armijo准则

Armijo准则 (Inexact Line Search)

- 对于$c_1 \in (0,1)$，如果$f_0(x+\alpha\Delta x) \leq f_0(x) + c_1\alpha\nabla f(x)^T\Delta x$，则步长$\alpha$满足Armijo准则
- 几何意义：点$(\alpha, \phi(\alpha))$必须在直线$l(\alpha)$的下方，则图中区间$(0, \alpha_1]$中的点都满足Armijo准则
- 由于是下降方向，直线$l(\alpha)$的斜率为负。$\nabla f(x)^T\Delta x < 0$, 一般取$c_1$为一个很小的正数



Armijo准则-选取α

回退法 (Backtracking)

1. 选取初始步长$\alpha_{\max}$, 参数$\gamma \in (0,1), c_1 \in (0,1)$
2. 重复:
3. 　　如果$f_0(x+\alpha\Delta x) > f_0(x) + c_1\alpha\nabla f(x)^T\Delta x$:
4. 　　　　$\alpha = \gamma\alpha$
5. 　　否则: 停止

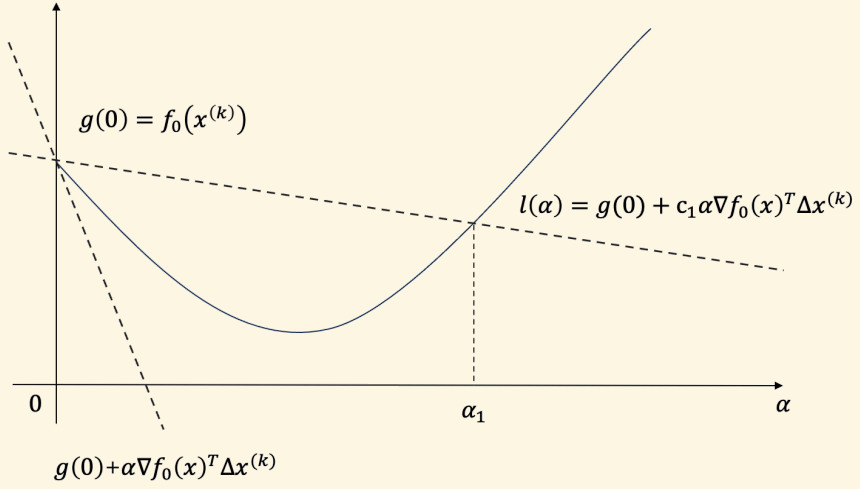
目录

Armijo准则 (Inexact Line Search)

- 对于$c_1 \in (0,1)$，如果$f_0(x+\alpha\Delta x) \leq f_0(x) + c_1\alpha\nabla f(x)^T\Delta x$，则步长$\alpha$满足Armijo准则

Armijo准则 (Inexact Line Search)

- 对于$c_1 \in (0,1)$，如果$f_0(x+\alpha\Delta x) \leq f_0(x) + c_1\alpha\nabla f(x)^T\Delta x$，则步长$\alpha$满足Armijo准则
- 几何意义：点$(\alpha, \phi(\alpha))$必须在直线$l(\alpha)$的下方，则图中区间$(0, \alpha_1]$中的点都满足Armijo准则
- 由于是下降方向，直线$l(\alpha)$的斜率为负。$\nabla f(x)^T\Delta x < 0$, 一般取$c_1$为一个很小的正数



Armijo准则-选取α

回退法 (Backtracking)

1. 选取初始步长$\alpha_{\max}$(一般为1), 参数$\gamma \in (0,1), c_1 \in (0,1)$
2. 重复:
3. 　　如果$f_0(x+\alpha\Delta x) > f_0(x) + c_1\alpha\nabla f(x)^T\Delta x$:
4. 　　　　$\alpha = \gamma\alpha$
5. 　　否则: 停止

收敛性证明 (Armijo Rule)

1. 定义函数

$$\tilde{f}(\alpha) = f(x^{(k)} + \alpha \Delta x^{(k)}) = f(x^{(k)} - \alpha \nabla f(x^{(k)})) = f(x^{(k+1)})$$

收敛性证明 (Armijo Rule)

1. 定义函数

$$\tilde{f}(\alpha) = f(x^{(k)} + \alpha \Delta x^{(k)}) = f(x^{(k)} - \alpha \nabla f(x^{(k)})) = f(x^{(k+1)})$$

2. 根据光滑性:

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \nabla f(x^{(k)})^T(-\alpha \nabla f(x^{(k)})) + \frac{M}{2}||-\alpha \nabla f(x^{(k)}))||_2^2$$

收敛性证明 (Armijo Rule)

1. 定义函数

$$\tilde{f}(\alpha) = f(x^{(k)} + \alpha \Delta x^{(k)}) = f(x^{(k)} - \alpha \nabla f(x^{(k)})) = f(x^{(k+1)})$$

2. 根据光滑性:

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \nabla f(x^{(k)})^T(-\alpha \nabla f(x^{(k)})) + \frac{M}{2}||-\alpha \nabla f(x^{(k)}))||_2^2$$

3. 等价于:

$$\tilde{f}(\alpha) \leq f(x^{(k)}) - \alpha||\nabla f(x^{(k)}))||_2^2 + \frac{M\alpha^2}{2}||\nabla f(x^{(k)}))||_2^2$$

收敛性证明 (Armijo Rule)

4. Armijo Rule

收敛性证明 (Armijo Rule)

4. Armijo Rule

$$\begin{aligned}\tilde{f}(\alpha) = f(x^{(k+1)}) &= f(x + \alpha \Delta x^{(k)}) \\ &\leq f(x^{(k)}) + c_1 \alpha \nabla f(x^{(k)})^T \Delta x^{(k)}\end{aligned}$$

收敛性证明 (Armijo Rule)

4. Armijo Rule

$$\begin{aligned}\tilde{f}(\alpha) = f(x^{(k+1)}) &= f(x + \alpha \Delta x^{(k)}) \\ &\leq f(x^{(k)}) + c_1 \alpha \nabla f(x^{(k)})^T \Delta x^{(k)}\end{aligned}$$

5. 当$0 \leq \alpha \leq \frac{1}{M}$时，有$(-\alpha + \frac{M\alpha^2}{2}) \leq -\frac{\alpha}{2}$

收敛性证明 (Armijo Rule)

4. Armijo Rule

$$\begin{aligned}\tilde{f}(\alpha) &= f(x^{(k+1)}) = f(x+\alpha\Delta x^{(k)}) \\ &\leq f(x^{(k)}) + c_1\alpha\nabla f(x^{(k)})^T\Delta x^{(k)}\end{aligned}$$

5. 当$0 \leq \alpha \leq \frac{1}{M}$时，有$(-\alpha + \frac{M\alpha^2}{2}) \leq -\frac{\alpha}{2}$
6. 代入，可以得出当$0 \leq \alpha \leq \frac{1}{M}$时， 此时$\alpha$满足Armijo Rule

$$\begin{aligned}\tilde{f}(\alpha) &\leq f(x^{(k)}) - \alpha||\nabla f(x^{(k)}))||_2^2 + \frac{M\alpha^2}{2}||\nabla f(x^{(k)}))||_2^2 \\ &\leq f(x^{(k)}) - \frac{\alpha}{2}||\nabla f(x^{(k)}))||_2^2 \\ &\leq f(x^{(k)}) - c_1\alpha||\nabla f(x^{(k)}))||_2^2 \quad (\text{当} c_1 \in (0, 0.5))\end{aligned}$$

收敛性证明 (Armijo Rule)

7. 因此，$\alpha = \alpha_{\max} = 1$或 $\alpha \geq \frac{\gamma}{M}$

收敛性证明 (Armijo Rule)

7. 因此，$\alpha = \alpha_{\max} = 1$或$\alpha \geq \frac{\gamma}{M}$
8. 根据光滑性, 右边$\alpha = \min < c_1, \frac{c_1\gamma}{M} >$， 则

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \min < c_1, \frac{c_1\gamma}{M} > ||\nabla f(x^{(k)}))||_2^2$$

收敛性证明 (Armijo Rule)

7. 因此，$\alpha = \alpha_{\max} = 1$或 $\alpha \geq \frac{\gamma}{M}$
8. 根据光滑性, 右边$\alpha = \min < c_1, \frac{c_1\gamma}{M} >$， 则

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \min < c_1, \frac{c_1\gamma}{M} > ||\nabla f(x^{(k)}))||_2^2$$

9. 联合不等式$p^* \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2$，可得

$$f(x^{(k+1)}) - p^* \leq (1 - \min < 2mc_1, \frac{2mc_1\gamma}{M} >)(f(x^{(k)}) - p^*)$$

收敛性证明 (Armijo Rule)

7. 因此，$\alpha = \alpha_{\max} = 1$或 $\alpha \geq \frac{\gamma}{M}$
8. 根据光滑性, 右边$\alpha = \min < c_1, \frac{c_1\gamma}{M} >$， 则

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \min < c_1, \frac{c_1\gamma}{M} > ||\nabla f(x^{(k)}))||_2^2$$

9. 联合不等式$p^* \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2$，可得

$$f(x^{(k+1)}) - p^* \leq (1 - \min < 2mc_1, \frac{2mc_1\gamma}{M} >)(f(x^{(k)}) - p^*)$$

6. 由于$\frac{2mc_1\gamma}{M} < 1$，因此线性收敛

例

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2), \ \ \gamma > 0$$

例

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2), \ \ \gamma > 0$$

- 矩阵的特征值为$1, \gamma$.

例

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2),\ \ \gamma > 0$$

- 矩阵的特征值为$1, \gamma$.
- 对于m和M，最紧（tightest）的选择为$m = \min < 1, \gamma >, M = \max < 1, \gamma >$

例

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2),\ \ \gamma > 0$$

- 矩阵的特征值为$1, \gamma$.
- 对于m和M，最紧（tightest）的选择为$m = \min < 1, \gamma >, M = \max < 1, \gamma >$
- 假设初始点为$x^{(0)} = (\gamma, 1)$, 根据梯度下降法可以推出

$$x_1^{(k)} = \gamma(\frac{\gamma - 1}{\gamma + 1})^k, x_2^{(k)} = \gamma(-\frac{\gamma - 1}{\gamma + 1})^k$$

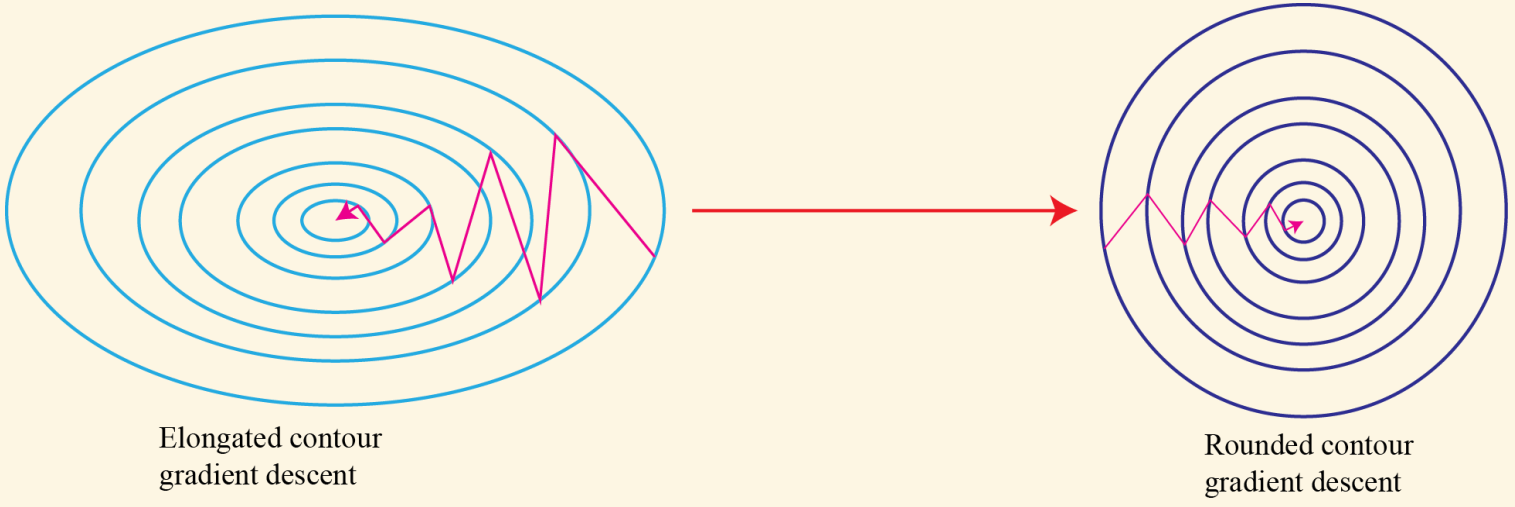
- 同时

$$f(x^{(k)}) = \frac{\gamma(\gamma + 1)}{2}(\frac{\gamma - 1}{\gamma + 1})^{2k} = (\frac{\gamma - 1}{\gamma + 1})^{2k} f(x^{(0)})$$

收敛速度

根据精确步长的梯度下降法，当$m \to M$时，收敛速度最快

根据Armijo Rule的梯度下降法，需要步长选取的较小，但是太小的补偿导致收敛速度更慢



目录

光滑

光滑

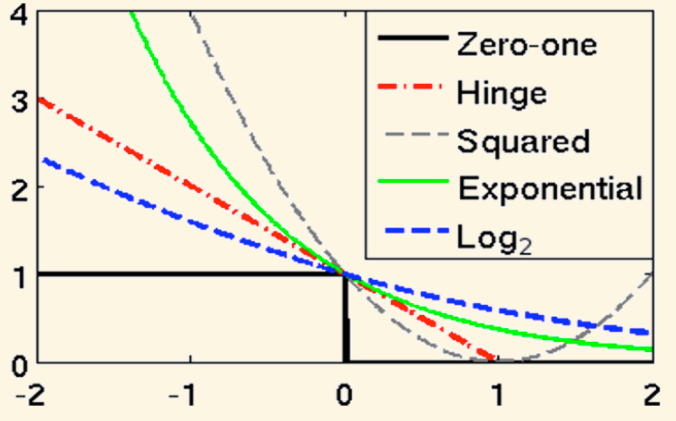
- $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$
- $f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2}||y - x||_2^2$

非光滑

- f不一定是可微的
- f的梯度，即使存在，不一定满足光滑性

常见的损失函数

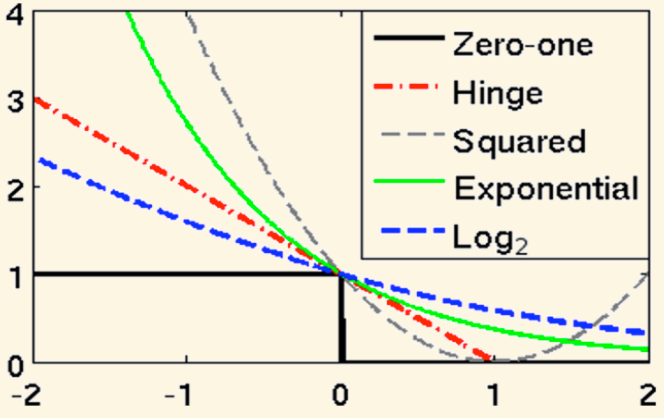
- s = $y \cdot y'$



$s = y \cdot y'$

常见的损失函数

- s = $y \cdot y'$
- 0-1 loss: 当$s < 0, f(s) = 1$，否则$f(s) = 0$
- Hinge loss: $f(s) = \max(0, 1 - s)$
- Squared loss: $f(s) = (s - 1)^2$
- Exponential Loss: $f(s) = e^{-s}$
- Logistic Loss: $f(s) = \log(1 + e^{-s})$

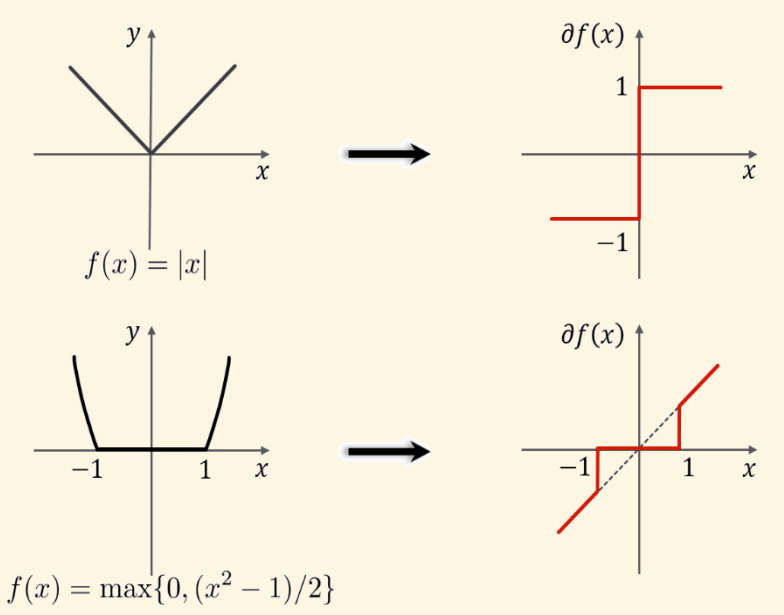


$s = y \cdot y'$

机器学习

机器学习中常见的非光滑应用

- 损失函数： Hinge loss, perception loss, l1-loss, etc.
- 正则化: l1-norm, total variation, elastic net, etc.
- 激活函数: ReLU, Leaky ReLU, etc.
- 似然函数: Laplacian noise.



例： LASSO

Least Absolute Shrinkage and Selection Operator (LASSO)

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|_1$$

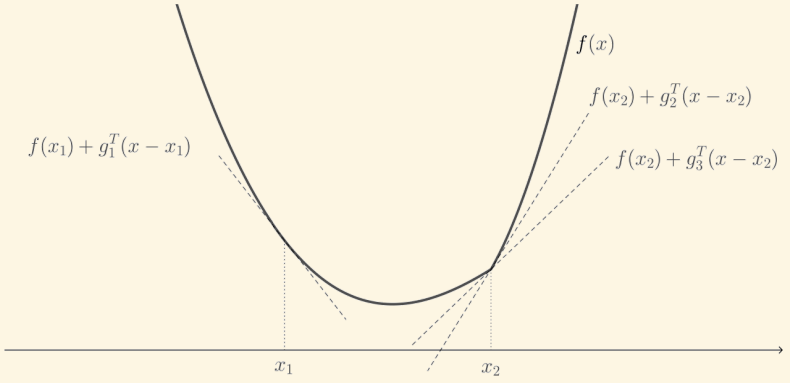
其中，$\|w\|_1 = \sum_{j=1}^{d} |w_j|$为L1正则项

次梯度

$f_0(x)$ 为连续凸函数,但不可微

次梯度

$$f(y) \geq f(x) + g^T(y - x)$$



次梯度法

$$x^{k+1} = x^k - \alpha^k g^k$$

次梯度法在凸优化问题中是可以收敛的，但收敛速度可能较慢，且对于非光滑问题，次梯度法能够比传统的梯度下降法更加有效。为了确保收敛性，步长的选择尤为关键，递减的步长通常能保证收敛。

谢 谢！！