2022秋季 学院路校区(凸优化与近似)
最优化方法　　第 4 次作业
提交日期：2022 年11 月10日周二课前

2022 年 10 月 26 日

1. 在针对光滑函数的Nesterov加速梯度法中，定义了标量序列

$$\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}.$$

证明序列$\{\lambda_t\}$是严格递增的，并且$\lambda_t \geq \frac{t}{2}$.

2. 考虑二元二次函数

$$f(x,y) = x^2 - 2xy + 10y^2 - 4x - 20y$$

的极小化.

   (a) 它是凸函数吗？求该函数的极小点.

   (b) 给出用坐标下降法得到的迭代序列，并画出初始点$(x_1, y_1) = (0, 0)$时迭代点的轨迹. 参考解答：$x_{t+1} = 2 + y_t, y_{t+1} = 1 + \frac{x_{t+1}}{10}$.

   (c) (b)中的序列收敛到该函数的极小点吗？

3. 假设$\|\cdot\|_a$和$\|\cdot\|_b$分别是$\mathbb{R}^m$和$\mathbb{R}^n$上的范数. 已知矩阵$X \in \mathbb{R}^{m\times n}$，可将其看作$\mathbb{R}^n$到$\mathbb{R}^m$的线性变换，从而可由范数$\|\cdot\|_a$和$\|\cdot\|_b$诱导出矩阵的算子范数

$$\|X\|_{a,b} = \sup\{\|Xu\|_a : \|u\|_b \leq 1\}.$$

当$\|\cdot\|_a$和$\|\cdot\|_b$均是欧氏范数时，诱导出的矩阵算子范数就是X的最大奇异值. 用$\|X\|_2$表示:

$$\|X\|_2 = \sigma_{\max}(X) = \sqrt{\lambda_{\max}(X^T X)}.$$

也称这个范数是矩阵的谱范数或者ℓ_2-范数. 请完成以下问题：

(a) 证明谱范数的如下变分表示

$$\|X\|_2 = \max_{u\in\mathbb{R}^n, v\in\mathbb{R}^m} v^TXu \text{ subject to } \|u\|_2 \le 1, \|v\|_2 \le 1.$$

并根此说明$\|X\|_2$关于矩阵X是凸函数.

(b) 基于(a)中的表述和块坐标下降法，给出求解矩阵X的最大奇异值的块坐标下降法.

4. 设$\|\cdot\|$是$\mathbb{R}^n$上的范数，对应的对偶范数，用$\|\cdot\|_*$表示，定义为

$$\|z\|_* = \sup\{z^Tx : \|x\| \le 1\}.$$

对偶范数可以解释为z^T的算子范数. 由$1\times n$矩阵在$\mathbb{R}^n$上的范数$\|\cdot\|$和$\mathbb{R}$上的绝对值能导出

$$\|z\|_* = \sup\{|z^Tx| : \|x\| \le 1\}.$$

所以，从对偶范数的定义，能得到对所有的$x, y\in\mathbb{R}^n$，不等式

$$|z^Tx| \le \|x\|\|z\|_*$$

成立. 该不等式在下述意义下是紧的：对任意x，存在z使得等式成立. 类似地，对任意z存在x使得等式成立.

完成以下问题：

(a) 证明欧氏(ℓ_2)范数的对偶范数是自己；

(b) 证明ℓ_1的对偶范数是ℓ_∞；

(c) 证明ℓ_∞的对偶范数是ℓ_1；

(d) 已知A是$n\times n$对称正定矩阵，证明椭球范数$\|x\|_A = \sqrt{x^TAx}$的对偶范数是$\|x\|_A^* = \sqrt{x^TA^{-1}x}$.

(e) 证明椭球范数$\|\cdot\|_A$度量下，可微函数f在x处的下降方向是$-A^{-1}\nabla f(x)$.

5. (AdaBoost的逐坐标下降解释) 提升技术的基本思想是利用弱学习算法构造一个强学习者. 为此，提升技术均使用集成方法. AdaBoost 是著名的提升算法，伪码描述见算法1. 这里给了m个已标号的训练样本$(x_1, y_1), \cdots, (x_m, y_m)$，其中$x_i$ 属于某定义域$\mathcal{X}$，标号$y_i \in \{-1, +1\}$. 在每一轮$t = 1, \cdots, T$，如算法所描述，计算m个样本的一个分布D_t，然后应用所给的弱学习算法找到弱假设$h_t : \mathcal{X} \mapsto \{-1, +1\}$，弱学习者的目的是找到关于分布$D_t$的加权误差$\epsilon_t$最低的弱假设. 计算加权弱假设

$$f_\alpha(x) = \sum_{t=1}^T \alpha_t h_t(x)$$

Algorithm 1 The boosting algorithm AdaBoost

Given: $(x_1, y_1), \cdots, (x_m, y_m)$ where $x_i \in \mathcal{X}, y_i \in \{-1, +1\}$.

Initialize $D_1(i) = 1/m$ for $i = 1, \cdots, m$.

for $t = 1$ to M **do**

Train weak learner using distribution D_t.

Get weak hypothesis $h_t : \mathcal{X} \mapsto \{-1, 1\}$.

Aim: select h_t with the lowest weighted error $\epsilon_t = \mathbb{P}_{i \sim D_t}(h_t(x_i) \neq y_i)$.

Choose $\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$

Update, for $i = 1, \cdots, m$:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

end for

Output the final hypothesis $h(x) = \text{sign} \left(\sum_{m=1}^{M} \alpha_m h_m(x) \right)$.

的符号作为最终或者组合假设h. 这里假设与分类器是通用的.

可将AdaBoost理解成贪心极小化指数损失对应的经验风险：

$$\min_{\alpha \in \Delta_M} F(\alpha) := \frac{1}{m} \sum_{i=1}^{m} \exp(-y_i f_\alpha(x_i))$$

的程序，其中$f_\alpha(x) = \sum_{j=1}^{M} \alpha_j h_j(x)$，

$$\Delta_M = \{\alpha \in \mathbb{R}^M : \alpha_j \geq 0, \forall j, \sum_{j=1}^{M} \alpha_j = 0\}. \quad (1)$$

这样的理解是很有帮助的. 首先，这种理解明晰了算法的目的，对证明收敛性有帮助. 其次，将算法与它的目标分离，有可能对于相同的目标得到更好或者更快的算法，或者也有可能推广AdaBoost 解决新的挑战性问题. 为了理解可将AbaBoost看作一种特定的坐标下降法(其中每步贪心地沿着一个坐标方向在前进)，请完成以下问题：

(a) 将算法中的归一化因子Z_t表示成加权误差ϵ_t的函数.

(b) 设e_t表示$\mathbb{R}^M$中第t个单位向量，即第t个分量为1，其它分量为0，设$\alpha^{(t-1)} \in \mathbb{R}^M$ 表示由前$t-1$个系数得到的向量[1]：$t > 1$时，

$$\alpha^{(t-1)} = (\alpha_1, \cdots, \alpha_{t-1}, 0, \cdots, 0) \in \mathbb{R}^M.$$

置$\alpha^{(0)} = (0, \cdots, 0) \in \mathbb{R}^M$. 请完成以下问题：

[1]为了避免与向量的分量混淆，这里用上标$(t-1)$ 表示迭代指标

(i) (2分)计算函数$F(\alpha)$ 在$\alpha^{(t-1)}$处的梯度.

(ii) (5分)给出函数$F(\alpha)$ 在$\alpha^{(t-1)}$沿坐标方向$e_1, \cdots, e_M$ 的方向导数，并确定沿哪个坐标方向的方向导数最小.

(iii) (5分) 确定使得一元函数$F(\alpha^{(t-1)} + \eta e_t)$ 取到最小值的η.

6. [**选做题：**Powell, 1973年构造的CD法不收敛的例子]考虑三元分片二次函数

$$f(x_1, x_2, x_3) = -x_1x_2 - x_2x_3 - x_3x_1 + \sum_{i=1}^{3}[(x_i - 1)_+^2 + (-x_i - 1)_+^2]$$

其中$(a)_+ = \max\{a, 0\}$.

(a) 它是凸函数吗？有极小点吗？有驻点吗？

(b) 已知初始点$x^0 = (-1-\epsilon, 1+\frac{\epsilon}{2}, -1-\frac{\epsilon}{4})$，其中$\epsilon > 0$. 给出坐标下降法极小化该函数的求解过程和得到的迭代点列.

(c) (b)中的序列有聚点吗？收敛到该函数的极小点吗？