第 1 章 第三次作业

作业提交形式为线上提交, 应使用 LaTeX 独立完成作业电子版, 并提交相应程序.

Deadline: 2022 年 5 月 7 日 23:59.

1.1 基础题

1.1.1 凸函数

利用凸函数的二阶条件证明 $f(x) = \log \sum_{k=1}^{n} \exp(x_k)$ 是凸函数.

注: 本问题可以讨论, 查阅资料.

解 为了简便书写，我们记向量函数 $t = \Big[\exp(x_1), \exp(x_2), \cdots, \exp(x_n)\Big]^T$. 从而，$f(x)$ 写作如下形式

$$f(x) = \log \mathbf{1}^{\mathrm{T}} t.$$

从而我们得到对应的梯度和 Hessian 矩阵：

$$\begin{aligned}
\nabla f(x) &= \frac{1}{\mathbf{1}^{\mathrm{T}} t} \Big[\exp(x_i)\Big]_{1\times n} = s, \\
\nabla^2 f(x) &= \frac{1}{(\mathbf{1}^{\mathrm{T}} t)^2} \Big[\exp(x_i)\left(\delta_{ij}(\mathbf{1}^{\mathrm{T}} t) - \exp(x_j)\right)\Big]_{n\times n} = \operatorname{diag}(s) - s s^{\mathrm{T}}.
\end{aligned}$$

此处的 $s = t/(\mathbf{1}^{\mathrm{T}} t)$，可以近似理解为一种特殊的归一化（这里 $t \succ 0$，其实就是在 l^1 意义下的归一化）. 注意到有 Cauchy 不等式

$$\sum_i (s_i x_i^2) = \left(\sum_i s_i x_i^2\right)\left(\sum_i s_i \cdot 1^2\right) \geq \left(\sum_i 1 \cdot s_i x_i\right)^2.$$

从而 $\nabla^2 f(x) \succeq 0$，即凸性得证.

1.1.2 梯度下降法

设 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} + b^T \boldsymbol{x}$ 是正定二次函数, 其中 $Q = \operatorname{diag}(1, 2, 10)$, $b = (1, 0, 0)^{\mathrm{T}}$. 设初始点为 $\boldsymbol{x}_0 = (1, 1, 1)^{\mathrm{T}}$. 使用最速下降法, 用精确线搜索确定步长.

1. 算法的收敛速率是怎样的?
2. 要使 $f(x_T) - f(x^*) < 10^{-10}$, 大约需要多少次迭代?

解

1. 注意到采用精准线搜索确定步长，我们有:

$$\alpha_k = \frac{(g_k^{\mathrm{T}} g_k)^2}{g_k^{\mathrm{T}} Q g_k}.$$

对应的收敛速率为

$$\frac{f(x_{k+1})-f(x^*)}{f(x_k)-f(x^*)} \leq \left(1-\frac{2}{1+\kappa}\right)^2 = \frac{81}{121}.$$

2. 解方程

$$\left(\frac{81}{121}\right)^{\mathrm{T}} \cdot (f(x_0)-f(x^*)) \leq 10^{-10}. \implies T_{min} = 63.$$

大约需要 63 次迭代.

1.1.3 强凸与光滑问题判定

设 $\min_x f(x) = \frac{1}{2}\|Ax-y\|^2$, 其中 $A \in \mathbb{R}^{n\times n}$ 是非退化矩阵. 使用梯度下降法求解此问题

$$x_{k+1} = x_k - \alpha \nabla f(x_k),$$

其中 α 为步长.

1. 计算 $f(x)$ 相对于 x 的梯度.
2. 设 $A = \mathrm{diag}(-1, 5)$, 完成以下讨论:
   - 证明 $f(x)$ 是强凸函数, 计算参数 μ.
   - 证明 $f(x)$ 是光滑函数, 计算光滑函数的参数 L.
3. 设 $A = \mathrm{diag}(1, 2)$, 完成以下讨论:
   - 采用固定步长 $\alpha_k = 1/4$ 进行迭代, 求算法的单步下降幅度 $f(x_{k+1}) - f(x_k)$.
   - 采用固定步长进行迭代. 问如果要使迭代序列是下降序列, 步长应满足什么条件?
   - 如果 $f(x_0) - f(x^*) = 1$, 要使 $f(x_k) - f(x^*) < 10^{-3}$, 试估计需要多少次迭代.

解

1. $f(x)$ 相对于 x 的梯度为

$$\nabla = A^{\mathrm{T}}(Ax-y).$$

2. 注意到 Hessian 矩阵

$$\nabla^2 f(x) = A^{\mathrm{T}}A = \mathrm{diag}(1, 25).$$

   - 注意到 $\nabla^2 f(x) \succeq I$, 故 $f(x)$ 强凸, 且强凸系数 $\mu = 1$.
   - 注意到 $\nabla^2 f(x) \preceq 25I$, 故 $f(x)$ 光滑, 且光滑系数 $L = 1$.

3. 注意到单步改进公式

$$\begin{aligned}
f(x_{t+1}) - f(x_t) &= \frac{1}{2}\|x_t - \alpha g_t\|_Q^2 - \frac{1}{2}\|x_t\|_Q^2 - \alpha\langle b, g_t\rangle \\
&= \frac{1}{2}\alpha^2\|g_t\|_Q^2 - \alpha\langle Qx_t, g_t\rangle - \alpha\langle b, g_t\rangle \\
&= \frac{1}{2}\alpha^2\|g_t\|_Q^2 - \alpha\langle Qx_t + b, g_t\rangle \\
&= \frac{1}{2}\alpha^2\|g_t\|_Q^2 - \alpha\|g_t\|^2.
\end{aligned}$$

- 代入 $\alpha = 1/4$，得到单步下降幅度
$$f(x_{k+1}) - f(x_k) = g_k^{\mathrm{T}} \left(\frac{1}{32} A^{\mathrm{T}} A - \frac{1}{4} I \right) g_k.$$
- 令 $f(x_{k+1}) - f(x_k) < 0$ 对于特定条件的 α 取值成立即可。考虑关于 α 的二次函数对于任意参数 g_t 均小于零，求解恒成立问题下的 α 取值范围：
$$\alpha \in \left[0, \min_{g_t} \frac{2\|g_t\|^2}{\|g_t\|_Q^2} \right] = \left[0, \frac{1}{2} \right].$$
- 采用固定步长需要迭代 25 次，采用精准线搜索需要迭代 7 次 [1].

1.1.4 共轭梯度法性质

设 $f : \mathbb{R}^3 \to \mathbb{R}$ 连续可微. 使用共轭梯度法优化 $f(x)$, 初始点为 $x^{(0)}$, 迭代序列为 $x^{(k+1)} = x^{(k)} + \alpha_k d_k$, 步长由精确线搜索确定. 如果第一次迭代方向为 $d_0 = (1, -1, 2)^{\mathrm{T}}$, 沿 d_0 精确线搜索得到 $x^{(1)}$ 并且在 $x^{(1)}$ 处

$$\left. \frac{\partial f(x)}{\partial x_1} \right|_{x^{(1)}} = \left. \frac{\partial f(x)}{\partial x_2} \right|_{x^{(1)}} = -2.$$

1. 求 $x^{(1)}$ 处的搜索方向
2. 如果在 $x^{(2)}$ 处梯度长度为 $\|\nabla f(x^{(2)})\| = 1$. 求 $\nabla f(x^{(2)})$.

解

1. 注意到正交性，有 $g_1^{\mathrm{T}} d_0 = 0$ 立得
$$g_1 = \begin{bmatrix} -2 & -2 & 0 \end{bmatrix}^{\mathrm{T}}.$$
由 Fletcher-Reeves 公式得：
$$d_1 = -g_k + \frac{g_1^T g_1}{g_0^T g_0} = \begin{bmatrix} \frac{10}{3} & \frac{2}{3} & \frac{8}{3} \end{bmatrix}^{\mathrm{T}}.$$
2. 考虑约束
$$\begin{bmatrix} d_0 & d_1 & g_2 \end{bmatrix}^{\mathrm{T}} g_2 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$$
解之得
$$g_2 = \pm \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}^{\mathrm{T}}.$$

1.1.5 共轭梯度法的等价性

设 $f(x) = \frac{1}{2} x^T Q x + b^T x + c$ 是二次函数, Q 是正定矩阵.

1. 给定下降方向 d, 计算精确线搜索确定的步长.

[1] 这里似乎没有明说，但看上文应该是讨论的固定步长的情况吧，考试的时候应该会明确指明（如果不幸忘了写，按上下文分析是采用哪一种）.

2. 证明, 在上述问题上, 使用精确线搜索确定步长时, 三种共轭梯度法所确定的点列 $\{x_i\}$ 是相同的.

解

1. 考虑如下情况
$$\frac{\partial}{\partial \alpha} f(x+\alpha d) = \nabla f(x+\alpha d)^{\mathrm{T}} d = (Q(x+\alpha d)+b)^{d} = 0.$$
注意到 $Qx+b=\nabla f(x)$，因此
$$\alpha = \frac{\nabla f(x)^{\mathrm{T}} d}{d^{\mathrm{T}} Q d}.$$
2. 只需要证明两组等价关系即可
   - 注意到 $g_k - g_{k-1} = \alpha G d_{k-1}$，故 Hestenes-Stiefel 公式与 Crowder-Wolfe 公式等价.
   - 注意到 $g_k^{\mathrm{T}} d_{k-1} = 0, g_k^{\mathrm{T}} g_{k-1} = 0$ 故 Crowder-Wolfe 公式与 Fletcher-Reeves 公式等价.

   从而，在上述问题上，使用精确线搜索确定步长时，三种共轭梯度法所确定的点列 $\{x_i\}$ 是相同的.

1.1.6 共轭梯度法的子空间性质与收敛性

设 $f(x) = \frac{1}{2}\boldsymbol{x}^T Q \boldsymbol{x} + b^T \boldsymbol{x} + c$ 是二次函数, Q 是正定矩阵. 使用共轭梯度法, 证明:

1. 每次迭代的搜索方向 $\boldsymbol{d}_i$ 都可以写成下面向量组的线性组合
$$\{\boldsymbol{d}, Q\boldsymbol{d}, Q^2\boldsymbol{d}, \cdots, Q^i\boldsymbol{d}\}.$$
2. 对于第 $i+1$ 次迭代的序列 $\boldsymbol{x}_{i+1}$, 证明, 对于任意由上述向量线性组合的向量 $\boldsymbol{x}'$, 都有 $f(\boldsymbol{x}') \geq f(\boldsymbol{x}_{i+1})$.
3. 使用上述结论证明, 共轭梯度法在二次函数 $f(\boldsymbol{x})$ 上迭代 K 次即可收敛到最优点, 其中 K 是矩阵 Q 的不同的特征值的数目 (number of distinct eigenvalues). (说明, 如果 Q 的特征值为 $1, 3, 3, 5, 5, 7$, 则分离特征值数目为 4 个.)

解

1. 考虑归纳法进行证明
   - $k=0$ 时，$d_0 = d \in \operatorname{span}\{d\}$ 成立.
   - 假设 $k \leq n$ 时，$d_k \in \operatorname{span}\{d, Qd, \cdots, Q^k d\}$ 成立.
   - 当 $k = n+1$ 时，注意到
   $$d_{n+1} = -g_{n+1} + \beta d_n = -\left(g_n + \alpha Q d_n\right) + \beta d_n \in \operatorname{span}\{d, Qd, \cdots, Q^n d\}$$
   成立.

   因此，每次迭代的搜索方向 $\boldsymbol{d}_i$ 都可以写成下面向量组的线性组合
$$\{\boldsymbol{d}, Q\boldsymbol{d}, Q^2\boldsymbol{d}, \cdots, Q^i\boldsymbol{d}\}.$$

2. 考虑张成空间中扰动

$$\sum_{k=0}^{i} \alpha_k d_k \triangleq D\alpha.$$

考虑极值点的导数

$$\frac{\partial}{\partial \alpha} f(x_0 + D\alpha) = D^{\mathrm{T}} \nabla f(x) = \mathbf{0}.$$

即共轭梯度法对应的正交性

$$\forall k = 0, 1, \cdots, i \quad d_k^{\mathrm{T}} \nabla f(x) = 0.$$

故 $f(\boldsymbol{x}') \geq f(\boldsymbol{x}_{i+1})$ 成立

3. 注意到相同特征值对应的向量构成特征子空间，在同一个特征值对应的特征子空间上迭代一次即可达到最优点，因此迭代 K 次即可收敛到最优点.

1.1.7 约束优化问题的 KKT 条件

考虑由 PCA 的约束优化问题: 设 $\boldsymbol{x}_i$ 是一组数据, $\boldsymbol{x}_i$ 的均值为零, $S = \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}_i^T$ 是协方差矩阵, 一般 PCA 问题可以写成:

$$\begin{aligned} &\arg\max_{V} \operatorname{tr}(V^{\mathrm{T}} S V) \\ &\text{s.t. } V^{\mathrm{T}} V = I_k \end{aligned}$$

其中 $V \in \mathbb{R}^{n \times k}$, 完成以下问题.

1. 写出其 KKT 条件, 并分析 KKT 点.
2. 如果使用投影梯度法进行求解上述问题, 写出投影公式.

解

1. 构造 Lagrange 函数

$$\mathcal{L} = \operatorname{tr}(V^{\mathrm{T}} S V) + \lambda(V^{\mathrm{T}} V - I_k).$$

对 V, λ 求导，并令之为 0，得 V 为 S 的前 k 大的特征值对应的单位特征向量构成的矩阵.

2. 略去，不在考纲内，有兴趣的同学可以课后联系李老师或助教.

1.2 思考题

思考以下问题, 了解共轭方向与特征向量的关系

1.2.1 共轭方向与特征向量

思考以下问题, 了解共轭方向与特征向量的关系

1.2.2 共轭方向与特征向量

设 $f(x)=\frac{1}{2}\boldsymbol{x}^TQ\boldsymbol{x}+b^T\boldsymbol{x}$ 是正定二次函数, Q 的特征分解为 $Q=U\Lambda U^T$, 特征值满足 $\lambda_1>\lambda_2>\cdots>\lambda_n>0$, 其对应的特征向量分别为 $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n$. 令 $\boldsymbol{x}^*$ 表示 f 的最小值点. 取初始点 $\boldsymbol{x}_0$, 并且 $\boldsymbol{x}_0=\boldsymbol{x}^*+\sum_{i=1}^n\beta_i\boldsymbol{u}_i$ 满足 $\beta_i\neq 0,i=1,\dots,n$. 构造迭代算法

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\beta_k\boldsymbol{u}_k$$

1. 证明: 方向组 $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n$ 是关于 Q 相互共轭的.
2. 对于任意 $k=1,\dots,n-1$, 求单步改进幅度 $f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}_k)$.
3. 如果对于 $k=1,\dots,n-1$, 每一次迭代的改进幅度 $\delta_k=|f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}_k)|$ 都相等, 则步长 β_k 满足什么条件?

1.3 编程题

阅读以下页面内容, 并实现相应算法.

1.3.1 使用 Newton-CG 与 L-BFGS 法求解 LASSO 问题

注: 点击题目会跳转相应网页.

1. Newton-CG 法

   阅读理解上述 MATLAB 程序, 实例程序转写成 Python 程序, 并实现 Newton-CG 算法.
2. 应用牛顿-共轭梯度法解逻辑回归问题.

   阅读上述程序, 并用 Python 实现上述实例.
3. L-BFGS 算法的 MATLAB 实现.

   阅读理解上述 MATLAB 程序, 实例程序转写成 Python 程序, 并实现 L-BFGS 算法.
4. 利用 L-BFGS 方法求解逻辑回归问题.

   阅读上述程序, 并用 Python 实现上述实例.