Optimization Methods Exercise

Theory, Algorithms and Applications

作者：we

组织：Nanjing University of Aeronautics and Astronautics

时间：May. 2, 2021

版本：4.1

Bio：Information

ElegantLaTeX Program

Victory won't come to us unless we go to it.

第 1 章 第一次作业

本次作业包含三部分内容, 即基础题, 思考题和编程题. 作业提交形式为线上提交, 应使用 LaTeX 独立完成作业电子版.

Deadline: 2022 年 3 月 22 日 23:59.

1.1 基础题

说明：该部分的所有题目均为必做，所有题目及相关变种均可能出现在试卷上。

1.1.1 Sherman-Morrison 公式

已知 A 是一个 $n \times n$ 的可逆矩阵，并且 x 和 y 是两个 $n \times 1$ 向量，使得 $A + xy^{\mathrm{T}}$ 可逆，则有

$$(A + xy^{\mathrm{T}})^{-1} = A^{-1} - \frac{A^{-1}xy^{\mathrm{T}}A^{-1}}{1 + y^{\mathrm{T}}A^{-1}x}.$$

试验证该公式的正确性.

解 由于只需要进行验证，我们直接进行相乘即可. 首先我们证明是左逆：

$$\begin{aligned}\left(A^{-1} - \frac{A^{-1}xy^{\mathrm{T}}A^{-1}}{1 + y^{\mathrm{T}}A^{-1}x}\right)(A + xy^{\mathrm{T}}) =& I - \frac{A^{-1}xy^{\mathrm{T}}}{1 + y^{\mathrm{T}}A^{-1}x} + A^{-1}xy^{\mathrm{T}} - \frac{A^{-1}xy^{\mathrm{T}}A^{-1}xy^{\mathrm{T}}}{1 + y^{\mathrm{T}}A^{-1}x} \\ =& I - A^{-1}x\left(\frac{1}{1 + y^{\mathrm{T}}A^{-1}x} - 1 + \frac{y^{\mathrm{T}}A^{-1}x}{1 + y^{\mathrm{T}}A^{-1}x}\right)y^{\mathrm{T}} \\ =& I.\end{aligned}$$

同样的，我们容易证明是右逆. 因此，Sherman-Morrison 公式验证成立.

1.1.2 矩阵的子空间

1. 阅读 [1]，了解矩阵的四个基本子空间：列空间（Column Space）、行空间（Row Space）、零空间（Null Space）、左零空间（Left Null Space），写出对应的定义.
2. 阅读 Wikipedia 或是 Brilliant，基于 Gauss-Jordan 消元法证明秩-零化度定理 (Rank-Nullity Theorem).

解

1. 记考察的矩阵为 $A \in \mathbb{R}^{m \times n}$，我们有
   - 列空间 $C(A) = \{Ax | x \in \mathbb{R}^n\}$
   - 行空间 $C(A^{\mathrm{T}}) = \{A^{\mathrm{T}}x | x \in \mathbb{R}^m\}$
   - 零空间 $N(A) = \{x \in \mathbb{R}^n | Ax = \mathbf{0}\}$

- 左零空间 $N(A^{\mathrm{T}})=\{x\in\mathbb{R}^m|A^{\mathrm{T}}x=\mathbf{0}\}$

2. 注意到将矩阵转换为 Gauss-Jordan 形式不会改变列空间与零空间的秩. 又由于 Gauss-Jordan 形式对应列空间的秩为系数不全为零的行数，零空间的秩为系数全为零的行数. 由此秩-零化度定理得证（更详细的每一步可以参考Brilliant）.

1.1.3 矩阵求导

完成一下矩阵求导公式的推导:

1. $\frac{\partial}{\partial X}\mathrm{tr}(X^{\mathrm{T}}AX)=(A+A^{\mathrm{T}})X.$
2. $\frac{\partial}{\partial X}\mathrm{tr}(X^{\mathrm{T}}AY)=AY.$
3. $\frac{\partial}{\partial Y}\mathrm{tr}(X^{\mathrm{T}}AY)=A^{\mathrm{T}}X.$

解

1. 要求 $d\left(\mathrm{tr}\left(X^{\mathrm{T}}AX\right)\right)$，有

$$\begin{aligned}d\left(\mathrm{tr}\left(X^{\mathrm{T}}AX\right)\right)&=\mathrm{tr}\big(d(X^{\mathrm{T}}AX)\big)\\&=\mathrm{tr}\left(dX^{\mathrm{T}}AX+X^{\mathrm{T}}AdX\right)\\&=\mathrm{tr}\left(dX^{\mathrm{T}}AX\right)+\mathrm{tr}\left(X^{\mathrm{T}}AdX\right)\\&=\mathrm{tr}\left((dX)^{\mathrm{T}}AX\right)+\mathrm{tr}\left(X^{\mathrm{T}}AdX\right)\\&=\mathrm{tr}\left(X^{\mathrm{T}}A^{\mathrm{T}}dX\right)+\mathrm{tr}\left(X^{\mathrm{T}}AdX\right)\\&=\mathrm{tr}\left(X^{\mathrm{T}}(A^{\mathrm{T}}+A)dX\right).\end{aligned}$$

注意到 $df=\mathrm{tr}\left(\left(\frac{\partial f}{\partial X}\right)^{\mathrm{T}}dX\right)$，有

$$\frac{\partial\mathrm{tr}\left(X^{\mathrm{T}}AX\right)}{\partial X}=(A+A^{\mathrm{T}})X.$$

2. 要求 $d\left(\mathrm{tr}\left(X^{\mathrm{T}}AY\right)\right)$，有

$$\begin{aligned}d\left(\mathrm{tr}\left(X^{\mathrm{T}}AY\right)\right)&=\mathrm{tr}\left(d\left(X^{\mathrm{T}}AY\right)\right)\\&=\mathrm{tr}\left(dX^{\mathrm{T}}AY\right)\\&=\mathrm{tr}\left((dX)^{\mathrm{T}}AY\right)\\&=\mathrm{tr}\left(Y^{\mathrm{T}}A^{\mathrm{T}}dX\right).\end{aligned}$$

注意到 $df=\mathrm{tr}\left(\left(\frac{\partial f}{\partial X}\right)^{\mathrm{T}}dX\right)$，有

$$\frac{\partial\mathrm{tr}\left(X^{\mathrm{T}}AY\right)}{\partial X}=AY.$$

3. 要求 $d\left(\mathrm{tr}\left(X^{\mathrm{T}}AY\right)\right)$，有

$$\begin{aligned}d\left(\mathrm{tr}\left(X^{\mathrm{T}}AY\right)\right)&=\mathrm{tr}\left(d\left(X^{\mathrm{T}}AY\right)\right)\\&=\mathrm{tr}\left(X^{\mathrm{T}}AdY\right).\end{aligned}$$

注意到 $df = \mathrm{tr}\left(\left(\frac{\partial f}{\partial X}\right)^{\mathrm{T}} dX\right)$，有

$$\frac{\partial \mathrm{tr}\left(X^{\mathrm{T}} A Y\right)}{\partial Y} = A^{\mathrm{T}} X.$$

1.1.4 多元正态分布的等高面

1. 设 $X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$，其中 $\Sigma \succ 0$，X 的密度函数记为 $f(x;\mu,\Sigma)$. 任给 $a > 0$，试证明概率密度等高面

$$f(x;\mu,\Sigma) = a$$

是一个椭球面.

2. 特别地，令 $p = 2$ 且 $\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ $(\rho > 0)$，概率密度等高面就是平面上的一个椭圆. 试求该椭圆的方程、长轴和短轴.

解

1. 注意到等高面方程

$$(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu) = -2\ln\left((2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} a\right),$$

对正定对称矩阵 Σ^{-1} 对角化，并将矩阵 Σ 特征值降序排列得 Λ，并得到对应正交阵 U，得

$$(U(x-\mu))^{\mathrm{T}} \left(\frac{1}{-2\ln\left((2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} a\right)} \Lambda^{-1}\right) (U(x-\mu)) = 1.$$

即以 μ 为中心，U 第 i 行向量为对应第 i 长轴所在直线单位方向向量，且半轴长为

$$\left(-2\lambda_i \ln\left((2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} a\right)\right)^{\frac{1}{2}}$$

的超椭球面.

2. 考虑题中 $p = 2$ 且 $\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ $(\rho > 0)$ 时特定情况，可知椭圆方程如下：

$$\left(\frac{\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}(x-\mu)}{\sigma\sqrt{-2(1+\rho)\ln 2\pi a\sigma\sqrt{1-\rho^2}}}\right)^2 + \left(\frac{\begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}(x-\mu)}{\sigma\sqrt{-2(1-\rho)\ln 2\pi a\sigma\sqrt{1-\rho^2}}}\right)^2 = 1,$$

由此，容易观察得：

$$\text{长轴长} = 2\sigma\sqrt{-2(1+\rho)\ln 2\pi a\sigma\sqrt{1-\rho^2}},\ \text{短轴长} = 2\sigma\sqrt{-2(1-\rho)\ln 2\pi a\sigma\sqrt{1-\rho^2}}.$$

1.1.5 多元正态分布的简单性质

设 x 为 p 维随机向量 $x \sim \mathcal{N}(\mu, \Sigma)$，正态分布的概率密度函数为

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

完成以下问题:

1. 证明:
$$\mathbb{E}[x] = \int_{\mathbb{R}^n} xp(x|\mu, \Sigma)dx,$$
2. 计算: $\mathbb{E}[xx^{\mathrm{T}}] = \Sigma + \mu\mu^{\mathrm{T}}$ (提示, 考察与协方差矩阵的关系)
3. 设 A 为对称阵，计算二次型在正态分布下的期望:
$$\mathbb{E}[x^{\mathrm{T}}Ax] = \mathrm{tr}(\Sigma A) + \mu^{\mathrm{T}}A\mu$$
(提示 1. 多变量积分变量代换, 并使用对称矩阵的特征分解和正交矩阵的行列式为 ± 1. 提示 2. 另一种方法, 将二次型 $x^{\mathrm{T}}Ax$ 的期望写成矩阵迹的形式, 并利用矩阵的迹的性质进行变形计算.)
4. 当 $\mu = a\mathbf{1}_p$, $A = I_p - \frac{1}{p}\mathbf{1}_{p\times p}$, $\Sigma = \sigma^2 I_p$ 时，试利用 (1) 和 (2) 的结果证明 $\mathbb{E}(X^{\mathrm{T}}AX) = \sigma^2(p-1)$. 这里 $\mathbf{1}_p$ 表示所有分量均为 1 的列向量, $\mathbf{1}_{p\times p}$ 表示所有元素均为 1 的矩阵.

解

1. 通过计算我们容易知道
$$\mathbb{E}(X) = \mu.$$
2. 考虑方差 $\mathbb{E}\left((X-\mu)(X-\mu)^{\mathrm{T}}\right) = \Sigma$，展开得
$$\mathbb{E}(XX^{\mathrm{T}}) = \Sigma + \mu\mathbb{E}(X^{\mathrm{T}}) + \mathbb{E}(X)\mu^{\mathrm{T}} - \mu\mu^{\mathrm{T}} = \Sigma + \mu\mu^{\mathrm{T}}.$$
3. 引入不变量 tr 进行讨论，可知
$$\begin{aligned}
\mathbb{E}(X^{\mathrm{T}}AX) &= \mathrm{tr}\left(\mathbb{E}(X^{\mathrm{T}}AX)\right) \\
&= \mathbb{E}\left(\mathrm{tr}(X^{\mathrm{T}}AX)\right) \\
&= \mathbb{E}\left(\mathrm{tr}(AXX^{\mathrm{T}})\right) \\
&= \mathrm{tr}\left(\mathbb{E}(AXX^{\mathrm{T}})\right) \\
&= \mathrm{tr}\left(A\mathbb{E}(XX^{\mathrm{T}})\right) \\
&= \mathrm{tr}\left(A(\Sigma + \mu\mu^{\mathrm{T}})\right) \\
&= \mathrm{tr}(\Sigma A) + \mu^{\mathrm{T}}A\mu.
\end{aligned}$$
4. 直接带入得
$$\begin{aligned}
\mathbb{E}(X^{\mathrm{T}}AX) &= \mathrm{tr}(\Sigma A) + \mu^{\mathrm{T}}A\mu \\
&= \sigma^2 \mathrm{tr}\left(I_p - \frac{1}{p}\mathbf{1}_{p\times p}\right) + a^2\mathbf{1}_{\mathbf{p}}^{\mathrm{T}}\left(\mathbf{I_p} - \frac{\mathbf{1}}{\mathbf{p}}\mathbf{1_{p\times p}}\right)\mathbf{1_p} \\
&= \sigma^2(p-1) + a^2(p-p) \\
&= \sigma^2(p-1).
\end{aligned}$$

1.1.6 多元正态分布的极大似然估计

设 $X_{(i)}\ (i=1,2,\cdots,n)$ 为 p 元正态总体 $N_p(\mu,\Sigma)$ 的随机样本，记 $\overline{X}$ 为样本均值，A 为样本离差阵，有

$$\overline{X}=\frac{1}{n}\sum_{t=1}^{n}X_{(t)},\ A=\sum_{t=1}^{n}\left(X_{(t)}-\overline{X}\right)\left(X_{(t)}-\overline{X}\right)^{\mathrm{T}}.$$

完成以下问题

1. 推导 μ,Σ 的极大似然估计分别是 $\hat{\mu}=\overline{X},\hat{\Sigma}=\frac{1}{n}A.$
2. 证明 $\overline{X}\sim N\left(\mu,\frac{1}{n}\Sigma\right).$
3. 对于参数 θ 的一个估计 $\hat{\theta}$，我们称 $\hat{\theta}$ 是一个无偏估计，如果 $\mathbb{E}(\hat{\theta})=\theta$ 对于参数 θ 的任意取值都成立（通俗的说，就是假设真实分布中对应的参数为 θ，我们利用观测到的样本构造一个估计的方法 $\hat{\theta}$，这个样本满足 $\mathbb{E}(\hat{\theta})=\theta$）. 证明 $\overline{X}$ 是 μ 的无偏估计.

解

1. 考虑似然函数，我们记为 $L(\mu,\Sigma)$，从而有

$$\begin{aligned}L(\mu,\Sigma)&=\prod_{i=1}^{n}\frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\exp\left(-\frac{1}{2}\left(x_{(i)}-\mu\right)^{\mathrm{T}}\Sigma^{-1}\left(x_{(i)}-\mu\right)\right)\\&=\frac{1}{\sqrt{(2\pi)^{np}|\Sigma|^n}}\exp\left(-\frac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{n}\left(x_{(i)}-\mu\right)^{\mathrm{T}}\Sigma^{-1}\left(x_{(i)}-\mu\right)\right)\right)\\&=\frac{1}{\sqrt{(2\pi)^{np}|\Sigma|^n}}\exp\left(\mathrm{tr}\left(-\frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}\left(x_{(i)}-\mu\right)^{\mathrm{T}}\left(x_{(i)}-\mu\right)\right)\right).\end{aligned}$$

其中利用了技巧“标量的迹是其本身”进行简化，我们注意到

$$\begin{aligned}&\sum_{i=1}^{n}\left(x_{(i)}-\mu\right)\left(x_{(i)}-\mu\right)^{\mathrm{T}}\\=&\sum_{i=1}^{n}\left(x_{(i)}-\overline{X}+\overline{X}-\mu\right)\left(x_{(i)}-\overline{X}+\overline{X}-\mu\right)^{\mathrm{T}}\\=&\sum_{i=1}^{n}\left(x_{(i)}-\overline{X}\right)\left(x_{(i)}-\overline{X}\right)^{\mathrm{T}}+n\left(\overline{X}-\mu\right)\left(\overline{X}-\mu\right)^{\mathrm{T}}.\end{aligned}$$

考虑对数似然函数

$$\ln L(\mu,\Sigma)=-\frac{np}{2}\ln 2\pi-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}A\right)-\frac{n}{2}\left(\overline{X}-\mu\right)^{\mathrm{T}}\Sigma^{-1}\left(\overline{X}-\mu\right).$$

对 μ,Σ 求导，我们可得

$$\begin{aligned}\frac{\partial}{\partial\mu}\ln L&=2\Sigma^{-1}(\mu-\overline{X}),\\\frac{\partial}{\partial\Sigma}\ln L&=-\frac{1}{2}\left(\Sigma^{-1}-\Sigma^{-1}\left(\sum_{i=1}^{n}\left(x_{(i)}-\mu\right)\left(x_{(i)}-\mu\right)^{\mathrm{T}}\right)\Sigma^{-1}\right),\end{aligned}$$

令之为零，立得 μ,Σ 的极大似然估计分别是 $\hat{\mu}=\overline{X},\hat{\Sigma}=\frac{1}{n}A.$

2. 由正态分布性质立得

$$\overline{X}=\frac{1}{n}\sum_{t=1}^{n}X_{(t)}\sim N\left(\mu,\frac{n}{n^2}\Sigma\right)=N\left(\mu,\frac{1}{n}\Sigma\right).$$

3. 注意到

$$\mathbb{E}(\overline{X})=\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_{i1})\\ \vdots\\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(x_{ip})\end{bmatrix}=\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}\mu_1\\ \vdots\\ \frac{1}{n}\sum_{i=1}^{n}\mu_p\end{bmatrix}=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_p\end{bmatrix}=\mu.$$

故 $\overline{X}$ 是 μ 的无偏估计.

1.1.7 范数的严格凸性

求证范数的严格凸性等价于下列条件:

$$\|x+y\|=\|x\|+\|y\|\ (\forall x\neq\theta,y\neq\theta)\implies x=cy\ (c>0).$$

解

- $\Longrightarrow$

  即要通过 $\|\lambda x+(1-\lambda)y\|<\lambda\|x\|+(1-\lambda)\|y\|\ (\lambda\in(0,1)\ )$ 与 $\|x+y\|=\|x\|+\|y\|\ (\forall x\neq\theta,y\neq\theta)$ 推出 $x=cy\ (c>0)$. 由反证法，我们假设 x 不能表示为 cy，此时，下式的表示是唯一的:

  $$\left\|\frac{1}{2}x+\frac{1}{2}y\right\|=\frac{1}{2}\|x\|+\frac{1}{2}\|y\|,$$

  这与严格凸性是矛盾的，因此假设不成立，原命题为真.

- $\Longleftarrow$

  我们将条件取为对应的逆否命题，即

  $$x\neq cy\implies\|x+y\|\neq\|x\|+\|y\|,$$

  注意到范数均是凸的，立刻有

  $$x\neq cy\implies\|x+y\|<\|x\|+\|y\|,$$

  从而

  $$\|\lambda x+(1-\lambda)y\|<\lambda\|x\|+(1-\lambda)\|y\|\ (\lambda\in(0,1)\ ).$$

1.1.8 凸函数极小值的性质

设 $\mathscr{X}$ 为线性赋范空间，函数 $\varphi:\mathscr{X}\to\mathbb{R}^1$ 称为凸的，如果不等式

$$\varphi\left(\lambda x+(1-\lambda)y\right)\leq\lambda\varphi(x)+(1-\lambda)\varphi(y)\ (\forall 0\leq\lambda\leq 1)$$

成立. 求证凸函数的局部极小值必然是全空间的极小值.

解 我们只要证明局部极小值的唯一性即可. 用反证法，我们假设有两个极小值点 $x_1, x_2 \in \mathscr{X}$. 不失一般性地，我们设 $\varphi(x_1) \geq \varphi(x_2)$. 注意到 x_1 为极小值，因此

$$\exists \varepsilon > 0, \forall x(\|x - x_1\| \leq \varepsilon), f(x_1) < f(x).$$

我们试图构造一种情况产生矛盾. 这里记 $y = \lambda x_1 + (1-\lambda)x_2$，此时：

$$\|y - x_1\| = (1-\lambda)\|x_1 - x_2\|.$$

注意到

$$\lim_{\lambda \to 1}(1-\lambda)\|x_1 - x_2\| = 0,$$

故

$$\exists \lambda^*, \|x(\lambda^*) - x_1\| \leq \varepsilon,$$

此时

$$f\left(x(\lambda^*)\right) > f(x_1)$$

又由凸函数性质

$$\begin{aligned} f\left(x(\lambda^*)\right) &= f(\lambda^* x_1 + (1-\lambda^*)x_2) \\ &< \lambda^* f(x_1) + (1-\lambda^*)f(x_2) \\ &\leq \lambda^* f(x_1) + (1-\lambda^*)f(x_1) \\ &= f(x_1). \end{aligned}$$

从而产生矛盾，因此假设不成立，原命题为真.

1.1.9 投影的最优性条件

给定数据集 $\{\boldsymbol{x}_i\}_{i=1}^N$, $\boldsymbol{x}_i \in \mathbb{R}^n$, 设其均值为 $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i$. 设以 $\boldsymbol{w}$ 为方向向量, 过点 $\bar{\boldsymbol{x}}$ 的直线为

$$\boldsymbol{L} = \{\boldsymbol{y} \in \mathbb{R}^n | \boldsymbol{y} = \bar{\boldsymbol{x}} + t\boldsymbol{w},\ t \in \mathbb{R}\}.$$

1. 对任何 $\boldsymbol{x} \in \mathbb{R}^n$, 定义 $\tilde{\boldsymbol{x}}_i = \arg\min_{\boldsymbol{y} \in \boldsymbol{L}} \|\boldsymbol{x}_i - \boldsymbol{y}\|$ 为 $\boldsymbol{x}$ 在 $\boldsymbol{L}$ 上的投影. 请计算数据 $\boldsymbol{x}_i$ 在 $\boldsymbol{L}$ 的投影所对应的点 $\tilde{\boldsymbol{x}}$.
2. 对于目标函数

   $$\min_{\boldsymbol{w}} J(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2$$

   如果 $\|\boldsymbol{w}\| = 1$, 求最优点所满足的最优性条件. (提示: 参考瑞利商)
3. 设 $t_i^* = \arg\min_t \|\boldsymbol{x}_i - (\bar{\boldsymbol{x}} + t\boldsymbol{w})\|$, 则 t_i^* 依赖于方向 $\boldsymbol{w}$. 对于目标函数

   $$\max_{\boldsymbol{w}} K(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^N (t_i^*)^2,$$

   如果 $\|\boldsymbol{w}\| = 1$, 求最优点所满足的最优性条件.

解

1. 我们知道

$$\arg\min_{\boldsymbol{y}\in\boldsymbol{L}} \|\boldsymbol{x}_i-\boldsymbol{y}\| = \arg\min_{\boldsymbol{y}\in\boldsymbol{L}} \|\boldsymbol{x}_i-\boldsymbol{y}\|^2.$$

对于给定方向 $\boldsymbol{w}$ 的直线 $\boldsymbol{L}$，我们相当于只要求 $\tilde{\boldsymbol{x}}=\bar{\boldsymbol{x}}+\tilde{t}\boldsymbol{w}$ 满足

$$\tilde{t}=\arg\min_{s} \|\boldsymbol{x}_i-\bar{\boldsymbol{x}}-s\boldsymbol{w}\|^2.$$

通过求导可知

$$\frac{\partial}{\partial s}\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}-s\boldsymbol{w}\|^2 = 2\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}-s\boldsymbol{w}) \implies \tilde{t}=\frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}.$$

通过二阶求导容易知道该点为极小值点，由凸性知道是全局极小值点. 从而得到投影点

$$\tilde{\boldsymbol{x}}=\bar{\boldsymbol{x}}+\frac{\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}\boldsymbol{w}.$$

2. 此时，我们首先进一步化简最小值点的形式

$$\tilde{\boldsymbol{x}}=\bar{\boldsymbol{x}}+\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}).$$

代入目标函数，我们有

$$\begin{aligned}
J(\boldsymbol{w}) &= \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\|^2\\
&= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}(I-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})^2(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\\
&= \operatorname{tr}\left(\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}(I-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})^2(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right)\\
&= \operatorname{tr}\left((I-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})^2\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right)\\
&= \operatorname{tr}\left((I-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})A\right)\\
&= \operatorname{tr}(A)-\boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w}.
\end{aligned}$$

值得说明的是，最后第二步利用 $(I-\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})$ 是幂等矩阵进行了降幂，并且引入了之前提及的样本离差阵进行简化. 容易看出最优性条件即 $\boldsymbol{w}^*$ 为 A 第一主成分所在方向的单位向量.

3. 我们首先对目标函数进行化简，有

$$\begin{aligned}
K(\boldsymbol{w}) &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right)^2\\
&= \frac{1}{N}\sum_{i=1}^{N}\operatorname{tr}\left((\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right)\\
&= \operatorname{tr}\left(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\mathrm{T}}\right)\\
&= \operatorname{tr}(\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}A)\\
&= \boldsymbol{w}^{\mathrm{T}}A\boldsymbol{w}.
\end{aligned}$$

与上一问类似，最优性条件即 $\boldsymbol{w}^*$ 为 A 第一主成分所在方向的单位向量.

1.1.10 KL 散度的简单介绍

对于空间 $\mathbb{R}^n$ 中的两个概率分布 p, q，我们定义对应的 KL 散度如下：

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = \int_{\mathbb{R}^n} p(x) \left(\log p(x) - \log q(x)\right) dx.$$

试完成以下问题

1. 验证 KL 散度是否满足对称性（即 $\forall p, q, \mathrm{D}_{\mathrm{KL}}(p\|q) = \mathrm{D}_{\mathrm{KL}}(q\|p)$）.
2. 利用 $\log x$ 的凸性与 Jensen 不等式验证 KL 散度是否满足非负性（即 $\forall p, q, \mathrm{D}_{\mathrm{KL}}(p\|q) \geq 0$）.

解

1. 通过直接代入，我们可以发现 KL 散度不具有对称性（因此它也不是范数）.
2. 我们有

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(p\|q) &= \int_{\mathbb{R}^n} p(x) \left(\log p(x) - \log q(x)\right) dx \\
&= \int_{\mathbb{R}^n} p(x) \left(\log \frac{p(x)}{q(x)}\right) dx \\
&= \mathbb{E}_{x\sim p} \left(\log \frac{p(x)}{q(x)}\right) \\
&= -\mathbb{E}_{x\sim p} \left(\log \frac{q(x)}{p(x)}\right) \\
&\geq -\log \mathbb{E}_{x\sim p} \left(\frac{q(x)}{p(x)}\right) \\
&= -\log \int_{\mathbb{R}^n} q(x) dx \\
&= 0.
\end{aligned}$$

从而 KL 散度满足非负性.

1.2 思考题

说明：该部分的所有题目均可以不做，给出精彩的解答可以弥补在基础题区域的失分，特别精彩的解答可能可以折算为平时成绩的额外加分（待定）。部分题目及相关变种均可能出现在试卷上。

1.2.1 Sherman-Morrison 公式（接1.1.1）

试推导 Sherman-Morrison 公式.

1.2.2 矩阵的子空间（接1.1.2）

1. 阅读 Wikipedia，了解正交补（Orthogonal Complement），并说明矩阵的四个基本子空间直接的正交关系.
2. 基于正交补关系与秩-零化度定理证明矩阵的行秩等于列秩.
3. 阅读 Wikipedia，了解正交基（Orthonormal Basis），并写出 Bessel 不等式与 Parseval 等式）.
4. 阅读 [2] 中关于等周问题相关章节，并与通过变分法给出的证明进行比较.

1.2.3 矩阵求导（接1.1.3）

完成以下矩阵求导公式的推导：

1. $$\frac{\partial \det(X)}{\partial X} = \det(X) X^{-\mathrm{T}}.$$
2. $$\frac{\partial Y^{-1}}{\partial x} = -Y^{-1} \frac{\partial Y}{\partial x} Y^{-1}.$$

1.2.4 多元正态分布的极大似然估计（接1.1.6）

1. 说明 A 可以写成 $\sum_{t=1}^{n-1} Z_t Z_t^{\mathrm{T}}$ 的形式，其中 $Z_i \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$ $(i = 1, 2, \cdots, n-1)$.
2. 说明 Σ 的极大似然估计 $\hat{\Sigma} = \frac{1}{n} A$ 不是无偏估计.
3. 验证样本协方差阵 $S = \frac{1}{n-1} A$ 是 Σ 的无偏估计.

1.3 编程题

说明：该部分注明为“思考题”的题目均可以不做，给出精彩的解答可以弥补在基础题区域的失分，特别精彩的解答可能可以折算为平时成绩的额外加分（待定）。

非“思考题”部分必做，主要考察思考过程，不会因为给出算法的效率差等扣分（不排除后期用 OJ 收代码，可能部分性能极差的代码会扣分）。

除特殊要求的题目外，所有题目的编程语言不限；除特殊申明，不限库函数的使用；担心提交代码无法编译的，可以备注开发环境。

代码可能会抽样查重，发现代码抄袭现象时（尤其是逐字符相同、最后修改日期不正常等），一切处理手段解释保留。

1.3.1 Gauss-Jordan 消元法

试用 C++ 实现，Gauss-Jordan 消元法，要求可以 AC洛谷 P3389，允许参考题解，但请自己写一遍（除源码外，请额外提交带帐号名的 AC 截图）.

1.3.2 凸包的简单应用

文件 ‘2021JSCPC 热身赛题面.pdf’ 中有四道题，针对其中的 C 题，考虑基于凸包的算法（允许参考题解）。

1. 请用 algorithm2e 宏包给出对应的伪代码.
2. 给出 C++ 代码实现（请设置编译选项为 ‘-std=c++14 -O2’）.

1.3.3 BiCGSTAB（思考题）

参考稳定双共轭梯度法，利用 algorithm2e 宏包给出对应的伪代码. 感兴趣的同学可以利用 Matlab 或 Python 提供的函数进行测试，与其它算法的性能进行比较.

1.3.4 个人绩点分析（思考题）

1. 将教务处上的课程绩点，复制并保存为 csv 文件，自设一些可能的变量（如喜爱的程度、是否有预备基础等），并存入 MySQL 中.
2. 利用 Python 中 mysql.connector 等方法连接数据库，并保存为 DataFrame 文件（主要是练习数据库基础，也可直接导入 csv）.
3. 利用 statsmodels 提供的 ols 方法，提交 OLS Regression Results，并解释 R-squared, F-statistic, p-values 等变量的统计学含义（本小问仅作帮助理解假设检验用）.

Bibliography

[1] Gilbert Strang. The Four Fundamental Subspaces: 4 Lines. Website. https://web.mit.edu/18.06/www/Essays/newpaper_ver3.pdf (cit. on p. A).

[2] Elias M Stein and Rami Shakarchi. Fourier analysis: an introduction. Vol. 1. Princeton University Press, 2011 (cit. on p. J).

[3] Jianguo Huang and Jie-yong Zhou. "A direct proof and a generalization for a Kantorovich type inequality". In: Linear Algebra and its Applications 397 (2005), pp. 185–192 (cit. on p. O).

[4] David M. Blei. Exponential Families. Website. https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/exponential-families.pdf (cit. on p. S).

[5] Michael I. Jordan. The Exponential Family and Generalized Linear Models. Website. https://people.eecs.berkeley.edu/~jordan/courses/281A-fall04/lectures/lec-10-7.pdf (cit. on p. S).

第 2 章 第二次作业

本次作业包含三部分内容, 即基础题, 思考题和编程题. 作业提交形式为线上提交, 应使用 LaTeX 独立完成作业电子版.

Deadline: 2022 年 4 月 11 日 23:59.

2.1 基础题

说明：该部分的所有题目均为必做，所有题目及相关变种均可能出现在试卷上。

2.1.1 强凸函数性质

设 $f:\mathbb{R}^n\to\mathbb{R}$ 是连续可微函数, 并且 f 是强凸的, 即存在 $\mu>0$, 满足:

$$f(y)\geq f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{\mu}{2}\|y-x\|^2.$$

证明以下结论:

1. 函数 $h(x)=f(x)-\frac{\mu}{2}\|x\|^2$ 是凸函数.
2. 梯度的强单调性:
$$(\nabla f(x)-\nabla f(y))^{\mathrm{T}}(x-y)\geq\mu\|x-y\|^2.$$
3. 如果 $f(x)$ 是二阶连续可微的, 则 $\nabla^2 f(x)\succeq\mu I$.

解

1. 注意到
$$f(x)=h(x)+\frac{\mu}{2}\|x\|^2,$$
$$\nabla f(x)=\nabla h(x)+\mu x,$$
我们将之代入强凸的定义式，立得
$$h(y)\geq h(x)+\nabla h(x)^{\mathrm{T}}(y-x).$$
函数 $h(x)=f(x)-\frac{\mu}{2}\|x\|^2$ 是凸函数由此得证.
2. 考虑不等式
$$f(y)\geq f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{\mu}{2}\|y-x\|^2$$
$$f(x)\geq f(y)+\nabla f(y)^{\mathrm{T}}(x-y)+\frac{\mu}{2}\|x-y\|^2,$$
相加可得
$$(\nabla f(x)-\nabla f(y))^{\mathrm{T}}(x-y)\geq\mu\|x-y\|^2.$$
3. 考虑
$$(\nabla f(x+t)-\nabla f(x))^{\mathrm{T}}t\geq\mu\|t\|^2.$$

当 $t \to 0$ 时，有

$$t^{\mathrm{T}}\nabla^2 f(x)t \geq \mu\|t\|^2,$$

即 $\nabla^2 f(x) \succeq \mu I$.

2.1.2 光滑函数性质

设 $f : \mathbb{R}^n \to \mathbb{R}$ 是二阶连续可微函数, 并且其梯度 $\nabla f(x)$ 是 L-Lipschitz 连续的, 即

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y \in \mathbb{R}^n.$$

则有如下结论:

1. $\nabla^2 f(x) \preceq LI$.
2. 函数 $h(x) = \frac{L}{2}\|x\|^2 - f(x)$ 是凸函数.

解

1. 首先，我们注意到

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq \|\nabla f(x) - \nabla f(y)\|\|x - y\| \leq L\|x - y\|^2.$$

立刻有

$$(\nabla f(x + t) - \nabla f(x))^{\mathrm{T}}t \leq L\|t\|^2.$$

当 $t \to 0$ 时，有

$$t^{\mathrm{T}}\nabla^2 f(x)t \leq L\|t\|^2,$$

即 $\nabla^2 f(x) \preceq LI$.

2. 我们考虑这样的辅助函数

$$\begin{aligned}
&h(y) - h(x) - \nabla h(x)^{\mathrm{T}}(y - x)\\
=&\frac{L}{2}\|y - x\|^2 - \big(f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y - x)\big)\\
=&\frac{L}{2}\|y - x\|^2 - \left(\nabla f(\theta x + (1 - \theta)y) - \nabla f(x)\right)^{\mathrm{T}}(y - x)\\
\geq&\frac{L}{2}\|y - x\|^2 - \frac{L}{2}\|y - x\|^2\\
=&0
\end{aligned}$$

最后一步，如果没有二阶可微的条件，可以尝试考虑 L-Lipschitz 连续对应弱导数的存在性，并利用零测集对积分结果不影响的性质进行证明.

2.1.3 二次函数情形中梯度下降法的收敛性分析

给出二次函数

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}Q\boldsymbol{x} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} + c,$$

其中 Q 为严格正定的实对称矩阵，我们记

$$L \triangleq \lambda_{\max} = \|Q\|, \quad \mu \triangleq \lambda_{\min} = \|Q^{-1}\|^{-1}.$$

使用线搜索 $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{d}_t$ 求解上述问题. 完成以下讨论.

1. 参考Wikipedia和文献 [3] 等，给出康托罗维奇（Kantorovich）不等式的形式;
2. 如果使用负梯度方向为搜索方向, 精确线搜索确定步长, 计算每次迭代 α_t.
3. 请基于 Kantorovich 不等式证明采用精准线搜索步长梯度下降法的收敛性满足

$$\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2;$$

4. 如果采用固定步长 $\alpha = 1/L$，证明收敛性满足

$$\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \leq 1 - \frac{\mu}{L};$$

5. 根据条件数（Condition Number）κ（LaTeX 符号为\kappa）的定义，对上面的收敛性结论进行改写.

解

1. 对于严格正定实对称矩阵 G，我们有 Kantorovich 不等式:

$$\frac{(x^{\mathrm{T}}Gx)(x^{\mathrm{T}}G^{-1}x)}{x^{\mathrm{T}}x} \leq \frac{(\lambda_{\max} + \lambda_{\min})^2}{4\lambda_{\max}\lambda_{\min}}.$$

2. 首先，我们求解最优的信息

$$\nabla f(\boldsymbol{x}^*) = Q\boldsymbol{x}^* + \boldsymbol{b} = \boldsymbol{0}.$$

对于上述方差进行求解，我们有

$$\boldsymbol{x}^* = -Q^{-1}\boldsymbol{b},$$

代入二次函数，从而有

$$f(\boldsymbol{x}^*) = -\frac{1}{2}\|\boldsymbol{b}\|_{Q^{-1}}^2 + c.$$

另一方面，我们注意到梯度下降法的更新方程 $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha \boldsymbol{g}_t$，我们考虑如何获取精准线搜索的步长:

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) &= \frac{1}{2}\|\boldsymbol{x}_t - \alpha\boldsymbol{g}_t\|_Q^2 - \frac{1}{2}\|\boldsymbol{x}_t\|_Q^2 - \alpha\langle \boldsymbol{b}, \boldsymbol{g}_t\rangle \\ &= \frac{1}{2}\alpha^2\|\boldsymbol{g}_t\|_Q^2 - \alpha\langle Q\boldsymbol{x}_t, \boldsymbol{g}_t\rangle - \alpha\langle \boldsymbol{b}, \boldsymbol{g}_t\rangle \\ &= \frac{1}{2}\alpha^2\|\boldsymbol{g}_t\|_Q^2 - \alpha\langle Q\boldsymbol{x}_t + \boldsymbol{b}, \boldsymbol{g}_t\rangle \\ &= \frac{1}{2}\alpha^2\|\boldsymbol{g}_t\|_Q^2 - \alpha\|\boldsymbol{g}_t\|^2. \end{aligned}$$

我们考虑以步长 α 为唯一自变量的二次函数 $\varphi(\alpha) = \frac{1}{2}\alpha^2\|\boldsymbol{g}_t\|_Q^2 - \alpha\|\boldsymbol{g}_t\|^2$. 注意到精准线搜索的定义，我们有:

$$\alpha = \frac{\|\boldsymbol{g}_t\|^2}{\|\boldsymbol{g}_t\|_Q^2}.$$

为了便于书写，我们这里的 α 表示精准线搜索的步长（事实上，写作 α^* 更好），立刻

得到

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*) = -\frac{1}{2} \cdot \frac{\|\boldsymbol{g}_t\|^4}{\|\boldsymbol{g}_t\|_Q^2}.$$

我们从而可以得到如下的结果

$$\begin{aligned}
\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} &= 1 + \frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \\
&= 1 + \frac{-\frac{1}{2} \cdot \frac{\|\boldsymbol{g}_t\|^4}{\|\boldsymbol{g}_t\|_Q^2}}{\frac{1}{2}\|\boldsymbol{x}_t\|_Q^2 + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_t + \frac{1}{2}\|\boldsymbol{b}\|_{Q^{-1}}^2} \\
&= 1 - \frac{\|\boldsymbol{g}_t\|^4}{\|\boldsymbol{g}_t\|_Q^2 \cdot \left\|Q^{\frac{1}{2}}\boldsymbol{x}_t + Q^{-\frac{1}{2}}\boldsymbol{b}\right\|^2} \\
&= 1 - \frac{\|\boldsymbol{g}_t\|^4}{\|\boldsymbol{g}_t\|_Q^2 \cdot \|Q\boldsymbol{x}_t + \boldsymbol{b}\|_{Q^{-1}}^2} \\
&= 1 - \frac{\|\boldsymbol{g}_t\|^4}{\|\boldsymbol{g}_t\|_Q^2 \|\boldsymbol{g}_t\|_{Q^{-1}}^2} \\
&\leq 1 - \frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2} \\
&= \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2.
\end{aligned}$$

3. 我们就更一般性的情况予以证明，对于二次连续可微函数 f，其 Hesse 矩阵满足 $O \prec \mu I \preceq \nabla^2 f(x) \preceq LI$，如果采用固定步长 $\alpha = 1/L$ 进行梯度下降，其收敛性满足

$$\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \leq 1 - \frac{\mu}{L};$$

我们首先有 L-光滑性，对应有

$$\begin{aligned}
f(\boldsymbol{x}_{t+1}) &= f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^{\mathrm{T}}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{1}{2}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)^{\mathrm{T}}\nabla^2 f(\boldsymbol{\xi})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) \\
&\leq f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^{\mathrm{T}}(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2.
\end{aligned}$$

代入步长 $\alpha = 1/L$，立刻有

$$\begin{aligned}
f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^{\mathrm{T}}\left(-\frac{1}{L}\nabla f(\boldsymbol{x}_t)\right) + \frac{L}{2}\left\|-\frac{1}{L}\nabla f(\boldsymbol{x}_t)\right\|^2 \\
&= f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2.
\end{aligned}$$

同样的，利用 μ-强凸性，对应有

$$\begin{aligned}
f(\boldsymbol{y}) &= f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}}\nabla^2 f(\boldsymbol{\xi})(\boldsymbol{y} - \boldsymbol{x}) \\
&\geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2
\end{aligned}$$

注意到，不等式两侧同时取极大值（极小值）不等式仍然成立，立刻有

$$\min_{\boldsymbol{y}} f(\boldsymbol{y}) \geq \min_{\boldsymbol{y}} \left(f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2\right)$$

容易知道，不等式右侧的极值点 $\boldsymbol{y}_{\mathrm{RHS}}^* = x - \nabla f(x)/\mu$. 注意到左侧的极值点即全局极小

值点 $\boldsymbol{y}^*_{\mathrm{LHS}} = x^*$. 我们取 $\boldsymbol{x} = \boldsymbol{x}_t$，立刻有

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}_t) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}_t)\|^2.$$

合并得

$$2\mu\left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)\right) \leq \|\nabla f(\boldsymbol{x}_t)\| \leq 2L\left(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)\right),$$

化简得

$$\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \leq 1 - \frac{\mu}{L}.$$

4. 利用到条件数 κ 的定义

$$\kappa \triangleq \mathrm{cond}\ (Q) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

我们可以得到：

- 采用精准线搜索步长梯度下降法的收敛性满足

$$\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2.$$

- 采用固定步长 $\alpha = 1/L$ 梯度下降法的收敛性满足

$$\frac{f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)} \leq 1 - \frac{1}{\kappa}.$$

2.1.4 Fisher 判别的简单公式推导

对于给定的 k 组 m 维数据 $\{X_i\}_{i=1}^N$，我们希望能将之投影到某一个（几个）方向上，使组与组之间尽可能地分开. 一种常见的做法是采用方差分析的思想来导出判别函数，我们不妨记投影方向为 w. 我们的目标是尽可能增大组间方差，减小组内方差，我们这里直接考虑组间平方和与组内平方和. 投影后的组间平方和是容易得到的：

$$\sum_{i=1}^{k} n_i w^{\mathrm{T}}\left(\overline{X^{(i)}} - \overline{X}\right)^2 = w^{\mathrm{T}}\left(\sum_{i=1}^{k} n_i\left(\overline{X^{(i)}} - \overline{X}\right)\left(\left(\overline{X^{(i)}} - \overline{X}\right)^{\mathrm{T}}\right)\right) w \triangleq w^{\mathrm{T}} S_{\mathrm{b}} w.$$

其中，$n_i, \overline{X^{(i)}}$ 分别是第 i 组数据的数量与组中心. 我们称 S_{b} 为组间离差阵（between）. 上式的含义其实就是把一个组看成一个整体，按组中数据量赋予权重进行求和，最后进行了一个写法上的简化.

同样的，有组内平方和，我们把对应的组内离差阵（within）记为 S_{w}，对应推导如下：

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(w^{\mathrm{T}}\left(X_j^{(i)} - \overline{X^{(i)}}\right)\right)^2 = w^{\mathrm{T}}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_j^{(i)} - \overline{X^{(i)}}\right)\left(X_j^{(i)} - \overline{X^{(i)}}\right)^{\mathrm{T}}\right) w \triangleq w^{\mathrm{T}} S_{\mathrm{w}} w.$$

因此，一维 Fisher 判别的优化目标为

$$\max_{w} \frac{w^{\mathrm{T}} S_{\mathrm{b}} w}{w^{\mathrm{T}} S_{\mathrm{w}} w}.$$

对于多维的 Fisher 判别，有多种的不同的目标，我们这里主要介绍一种通过 trace 构造的. 对于向多个方向投影，我们记最终选定的投影方向为 $\{w_i\}_{i=1}^p$. 我们把离差阵统一记为 S，那对应的平方和可以写作

$$\sum_{i=1}^{p} w_i^{\mathrm{T}} S w_i = \sum_{i=1}^{p}\sum_{j=1}^{p} \delta_{ij} w_i^{\mathrm{T}} S w_j = \operatorname{tr}\left(W^{\mathrm{T}} S W\right).$$

此处 δ_{ij} 为 Kronecker-δ，$W = \begin{bmatrix} w_1 & w_2 & \cdots & w_p \end{bmatrix}$. 一种令人困惑的思路是采用行列式进行构造，据称解释是因为行列式的值实际上是矩阵特征值的积，一个特征值可以表示在该特征向量上的发散程度。（如果你能想到一个好的解释，请务必指点一下愚蠢的 TA）.

结合上述介绍，与部分参考资料（机器学习幼儿园 · 线性模型、机器学习幼儿园 · 降维与度量学习、Fisher' s Linear Discriminant: Intuitively Explained、Linear discriminant analysis : a detailed tutorial）完成以下问题：

1. 对于一维 Fisher 判别的优化目标
$$\max_{w} \frac{w^{\mathrm{T}} S_{\mathrm{b}} w}{w^{\mathrm{T}} S_{\mathrm{w}} w}$$
进行求解.
2. 对于 p 维 Fisher 判别的优化目标
$$\max_{W} \frac{\operatorname{tr}\left(W^{\mathrm{T}} S_{\mathrm{b}} W\right)}{\operatorname{tr}\left(W^{\mathrm{T}} S_{\mathrm{w}} W\right)}$$
进行求解.
3. 总结 PCA 与 LDA 间的异同.

解

1. 这其实就是广义 Rayleigh 商，转化为优化形式，有
$$\max_{w^{\mathrm{T}} S_{\mathrm{w}} w = 1} w^{\mathrm{T}} S_{\mathrm{b}} w.$$
通过 Lagrange 乘子法得到最优点条件
$$S_{\mathrm{w}}^{-1} S_{\mathrm{b}} w = \lambda w$$
可知 λ 为 $S_{\mathrm{w}}^{-1} S_{\mathrm{b}}$ 的特征值，代入得最终目标有
$$w^{\mathrm{T}} S_{\mathrm{b}} w = w^{\mathrm{T}} \lambda S_{\mathrm{w}} w = \lambda.$$
因此，我们取 $\lambda = \|S_{\mathrm{w}}^{-1} S_{\mathrm{b}}\|$，$w$ 为对应的特征向量即可.
2. 解法是类似的
$$\max_{\operatorname{tr}(W^{\mathrm{T}} S_{\mathrm{w}} W = 1)} w^{\mathrm{T}} S_{\mathrm{b}} W.$$
通过 Lagrange 乘子法得到最优点条件
$$S_{\mathrm{w}}^{-1} S_{\mathrm{b}} W = \lambda W$$
值得说明的是，S_{w} 不可逆的时候，我们只需要取对应的 Moore-Penrose 逆 $S_{\mathrm{w}}^{\dagger}$ 即可.W 的结果即为 $S_{\mathrm{w}}^{\dagger} S_{\mathrm{b}}$ 对应前 p 大特征值的特征向量构成的矩阵.

3. 详见Comparison between PCA and LDA、线性判别分析 LDA 原理及推导过程等.

2.1.5 主成分的简单计算

设 p 元总体 X 的协方差阵为

$$\Sigma=\sigma^2\begin{bmatrix}1 & \rho & \cdots & \rho\\ \rho & 1 & \cdots & \rho\\ \vdots & \vdots & & \vdots\\ \rho & \rho & \cdots & 1\end{bmatrix}\quad(0<\rho\leq 1).$$

1. 试证明总体的第一主成分 $Z_1=\frac{1}{\sqrt{p}}(X_1+X_2+\cdots+X_p)$;
2. 试求第一主成分的贡献率.

解

1. 注意到

$$\Sigma=\sigma^2\left(\rho\mathbf{1}\mathbf{1}^{\mathrm{T}}+(1-\rho)I\right)$$

立刻有特征向量为 $\mathbf{1}$ 与所有与 $\mathbf{1}$ 正交的向量，特征值分别为 $\sigma\cdot(1+(p-1)\rho)$ 与 $\sigma\cdot(1-\rho)$，注意到 $1+(p-1)\rho>1-\rho$，只需对 $\mathbf{1}$ 归一化即可，从而得总体的第一主成分

$$Z_1=\frac{1}{\sqrt{p}}(X_1+X_2+\cdots+X_p).$$

2. 第一主成分的贡献率

$$\eta=\frac{\lambda_1}{\mathrm{tr}(\Sigma)}=\frac{1+(p-1)\rho}{p}$$

2.1.6 指数组函数与 PCA 的简单性质

指数族分布的形式如下

$$p(x|\eta)=h(x)\exp\left(\eta^{\mathrm{T}}T(x)-A(\eta)\right).$$

其中，η 为自然参数（natural parameter），$T(x)$ 为充分统计量（sufficient statistic），$h(x)$ 为测度（underlying measure），$A(\eta)$ 为对数归一化因子（log normalizer）. 相关内容可以参考 [4] 与 [5].

1. 请推导指数族分布的 moment matching 性质：

$$\nabla_\eta A(\hat{\eta})=\mathbb{E}_{\hat{\eta}}\left(T(x)\right).$$

2. 基于极大似然参数估计准则，进行进一步推导：

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{i=1}^{N} T\left(x^{(i)}\right).$$

3. 基于带约束问题的拉格朗日乘子优化方法，推导指数族分布为满足观测到的充分统计量的分布中，假设最少的模型，即最大熵模型.

$$\text{Constraints: } \mathbb{E}(f_k) = F_k,\ 1 \le k \le n$$

4. 从充分统计量的角度谈谈 PCA 的应用场景.

解

1. 我们记考虑的空间为 $\mathscr{X}$，并将 $A(\hat{\eta})$ 展开为积分形式，从而：

$$\begin{aligned}\nabla_\eta A(\hat{\eta}) &= \nabla_\eta \log \int_{\mathscr{X}} h(x)\exp\left(\eta^{\mathrm{T}}T(x)\right)dx \\ &= \int_{\mathscr{X}} T(x)h(x)\exp\left(\eta^{\mathrm{T}}T(x) - A(\eta)\right)dx \\ &= \mathbb{E}_{\hat{\eta}}\left(T(x)\right).\end{aligned}$$

2. 注意到极大似然估计下，样本均值是对应的无偏估计，因此

$$\nabla_\eta A(\hat{\eta}) = \frac{1}{N}\sum_{i=1}^{N} T\left(x^{(i)}\right).$$

3. 不失一般性，我们令某一项约束表示概率函数的正则性（即 $\mathbb{E}(1) = 1$），注意结果对于约束的数量是不敏感的，我们考虑向量化表示，即 $\mathbb{E}(f) = F$, $\dim f \ge 1$。从而，我们得到如下优化问题

$$\begin{aligned}&\max \int_{\mathscr{X}} -p(x)\log p(x)dx, \\ &\text{s.t. } \mathbb{E}(f) = F.\end{aligned}$$

即

$$\begin{aligned}&\min \int_{\mathscr{X}} p(x)\log p(x)dx, \\ &\text{s.t. } \mathbb{E}(f) = F.\end{aligned}$$

我们构造 Lagrange 乘子项：

$$\mathcal{L} = \int_{\mathscr{X}} p(x)\log p(x)dx + \lambda^{\mathrm{T}}(\mathbb{E}(f) - F).$$

注意是对函数 p 优化，对应优化点应当满足 E-L 方程条件

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\partial}{\partial x}\frac{\partial \mathcal{L}}{\partial \dot{p}} = 0.$$

即

$$\int_{\mathscr{X}} \left(\log p(x) + 1 + \lambda^{\mathrm{T}} f\right)dx = 0.$$

注意到，积分值对积分空间应不敏感，即积分项几乎处处为零，即

$$p(x) = \exp(-\lambda^{\mathrm{T}} f - 1) \text{ a.e.}$$

在一般情况下，可以认为是指数族分布。

2.2 思考题

说明：该部分的所有题目均可以不做，给出精彩的解答可以弥补在基础题区域的失分，特别精彩的解答可能可以折算为平时成绩的额外加分（待定）。部分题目及相关变种均可能出现在试卷上。

本次思考题以调研为主，请做好文献引用。对于给出的两个方向，优质地完成一个即可加分（多选多加分）。每一个方向可以进入加分的最低篇幅为 3 页（不含参考文献列表）。

2.2.1 网络初始化的简单技巧

1. 从网络退化的角度解释为什么不能在初始化时将所有参数设置为相同数值;
2. 介绍 Xavier、MSRA，并选择至少一种初始化方法进行深入调研.

2.2.2 梯度消失与梯度爆炸的初步调研

1. 就梯度消失与梯度爆炸进行调研;
2. 就梯度消失与梯度爆炸的缓解方法进行调研.

2.3 编程题

说明：该部分注明为“思考题”的题目均可以不做，给出精彩的解答可以弥补在基础题区域的失分，特别精彩的解答可能可以折算为平时成绩的额外加分（待定）。

非“思考题”部分必做，主要考察思考过程，不会因为给出算法的效率差等扣分（不排除后期用 OJ 收代码，可能部分性能极差的代码会扣分）。

除特殊要求的题目外，所有题目的编程语言不限；除特殊申明，不限库函数的使用；担心提交代码无法编译的，可以备注开发环境。

代码可能会抽样查重，发现代码抄袭现象时（尤其是逐字符相同、最后修改日期不正常等），一切处理手段解释保留。

2.3.1 梯度法求解 LASSO 问题

阅读如下使用梯度法解 LASSO 问题的实例和对应的 MATLAB 代码:

- LASSO 问题的梯度法求解
- LASSO 问题的 Huber 光滑化梯度法
- LASSO 问题的连续化策略.

将上述实例中所有涉及的 MATLAB 程序转写成 Python 程序, 并实现上述实例. 注意, MATLAB 的 eig 函数可以由 numpy.linalg.eig 实现; MATLAB 的画图使用 matplotlib 实现.

2.3.2 多项式回归

使用多项式回归等方法, 拟合 data.m 中的数据, 作图并撰写报告. 要求实现以下模型: 最小二乘法, ℓ_2 范数的岭回归, ℓ_1 范数的 LASSO, 并使用梯度法寻找拟合参数, 不得直接调用 sklearn 的 fit 拟合 (但是可以阅读参考 fit 函数的代码).

2.3.3 二分法的简单应用

文件 ‘2020_icpc_shanghai_statement.pdf’ 中有十三道题，针对其中的 D 题，考虑基于二分法的算法，要求可以 AC第 45 届国际大学生程序设计竞赛（ICPC）亚洲区域赛（上海）D.Walker，具体如下：

1. 参考逆十字的答案，给出 C++ 代码实现.
2. 通常，我们认为 C++ 算法的需要进行 10^8 量级的运算可以在 1 秒中得出结果（性能高的平台可能可以达到 5×10^9）. 请结合二分法的收敛率，分析本题中实现 10^{-6} 次的精度迭代 100 次是一个合理的数值.
3. （思考题）参考五点共圆的答案，谈谈你对于闭式解于数值解的理解.

2.3.4 牛顿分形的简单介绍

参考【官方双语】（牛顿本人都不知道的）牛顿分形，学习牛顿分形的基本由来，并完成如下要求：

1. 在 3B1B 提供的Newton's Fractal交互平台上调出若干个自己觉得好看的分形，用 subfigure 环境提交.
2. 参考给大家看点好玩的，用 Python 画牛顿分形，用 Python 绘制不少于 4 个牛顿分形并提交分形与代码，要求如下：
   - 配色好看（可以参考 3B1B 的源码）；
   - （思考题）采用 CUDA 加速.

2.3.5 梯度下降法的简单变种

参考梯度下降法实现以下算法：

- Momentum 动量法
- Adagrad 法
- Adadelta 法
- RMSprop 法
- Adam 法

要求用 algorithm2e 包给出 Adam 法的伪代码，并自己实现各算法下降过程的可视化.

Bibliography

[1] Gilbert Strang. The Four Fundamental Subspaces: 4 Lines. Website. https://web.mit.edu/18.06/www/Essays/newpaper_ver3.pdf (cit. on p. A).

[2] Elias M Stein and Rami Shakarchi. Fourier analysis: an introduction. Vol. 1. Princeton University Press, 2011 (cit. on p. J).

[3] Jianguo Huang and Jie-yong Zhou. "A direct proof and a generalization for a Kantorovich type inequality". In: Linear Algebra and its Applications 397 (2005), pp. 185–192 (cit. on p. O).

[4] David M. Blei. Exponential Families. Website. https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/exponential-families.pdf (cit. on p. S).

[5] Michael I. Jordan. The Exponential Family and Generalized Linear Models. Website. https://people.eecs.berkeley.edu/~jordan/courses/281A-fall04/lectures/lec-10-7.pdf (cit. on p. S).