已知数据集$\mathcal{D}=\{(x_i,y_i)\}_{i=1}^n$，这里$x_i,y_i\in\mathbb{R}$，样本个数为$n$.

第一题

假设模型为$y_i=mx_i+c$, 这里$m,c\in\mathbb{R}$。模型损失定义如下：

$$\mathcal{L}(m,c)=\frac{1}{n}\sum_{i=1}^{n}(mx_i+c-y_i)^2$$

(1) 试求解m,c？
(2) 存在以下两个问题：(a) 易受噪声影响；(b) 无法检测多模型. 请给出解决方案.

A1:
(1) 分别关于m,c求导，并令其为0：

$$\begin{aligned}\frac{\partial\mathcal{L}(m,c)}{\partial m}&=\frac{2}{n}\sum_{i=1}^{n}x_i(mx_i+c-y_i)\doteq 0\\ \frac{\partial\mathcal{L}(m,c)}{\partial c}&=\frac{2}{n}\sum_{i=1}^{n}(mx_i+c-y_i)\doteq 0\\ \Longrightarrow&\begin{cases}m=\frac{\sum_{i=1}^n\sum_{j=1}^n x_iy_j-n\sum_{i=1}^n x_iy_i}{\left(\sum_{i=1}^n x_i\right)^2-n\sum_{i=1}^n x_i^2}\\ c=\frac{1}{n}\sum_{i=1}^n(y_i-mx_i)\end{cases}\end{aligned}$$

(2)

- 对于带有噪声的数据，可对参数进行正则化；

  $$\mathcal{L}(m,c)=\frac{1}{n}\sum_{i=1}^{n}(mx_i+c-y_i)^2+\lambda(m^2+c^2)$$

- 引入更高阶项，e.g., $x^2,x^3,\cdots$

第二题

假设模型为$y_i=w_0+w_1x_i+\cdots+w_dx_i^d$，这里数据的生成方式为

$$y_i=sin2\pi x_i+\epsilon,\epsilon\sim\mathcal{N}(0,0.01^2)$$

定义如下矩阵

$$X=\begin{pmatrix}1 & x_1 & x_1^2 & \cdots & x_1^d\\ 1 & x_2 & x_2^2 & \cdots & x_2^d\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^d\end{pmatrix}, w=\begin{pmatrix}w_0\\ w_1\\ \vdots\\ w_d\end{pmatrix}, \mathcal{Y}=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}$$

（1）当$n=4, d=3$时，模型有唯一解. 请利用Vandermonde给出解的形式.
（2）试给出下式w的解析解.

$$w^*=\underset{w}{argmin}\ ||Xw-\mathcal{Y}||_2^2$$

A2:
（1）参考这里可得Vandermonde矩阵的求逆公式，

$$X_{ij}^{-1}=(-1)^{i+1}\sum_{\substack{1\le p_1<\cdots<p_{n-i}\le n\\ p_1,\ldots,p_{n-i}\neq j}} x_{p_1}x_{p_2}\cdots x_{p_{n-i}} \Bigg/ \prod_{\substack{1\le k\le n\\ k\neq j}}(x_k-x_j)$$

因此，有

$$\begin{aligned}Xw&=\mathcal{Y}\\ w&=X^{-1}\mathcal{Y}\\ &=[X_{ij}^{-1}]_{4\times 4}\mathcal{Y}\end{aligned}$$

（2）
设，

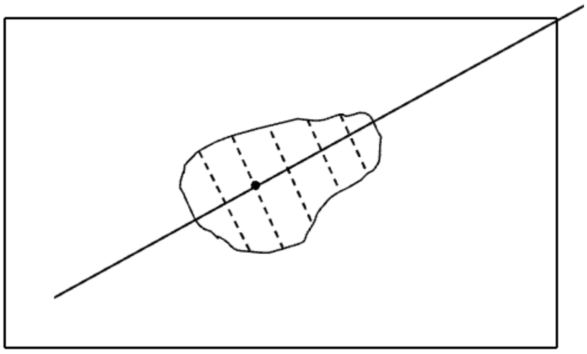
$$\mathcal{L}(w)=(Xw-\mathcal{Y})^\top(Xw-\mathcal{Y})$$

关于w求导，并令其为0，

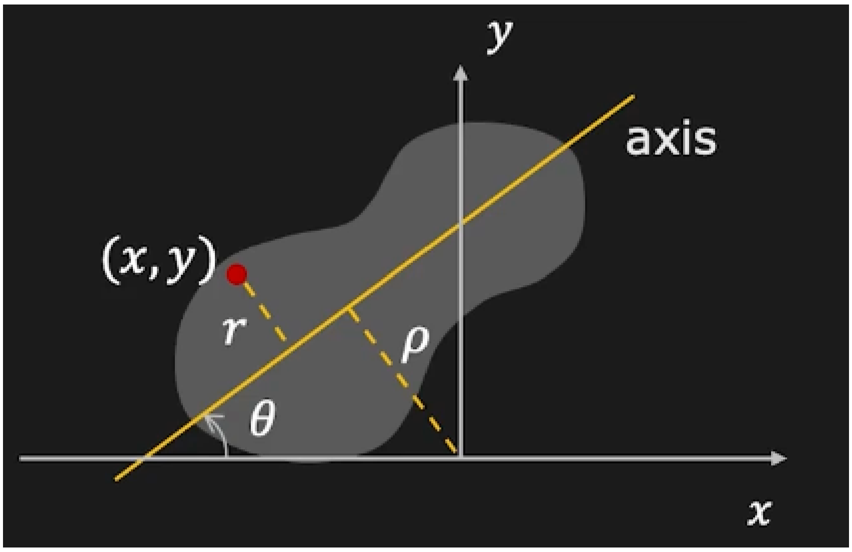
$$\begin{aligned}&\frac{\partial\mathcal{L}(w)}{\partial w}=2X^\top(Xw-\mathcal{Y})\doteq 0\\ &\Longrightarrow w=(X^\top X)^{-1}X^\top\mathcal{Y}\end{aligned}$$

附加题

试在二值图中寻找一个主方向，使得轮廓上的每个点到其距离之和最小.（提示：直线用极坐标表示）



Ans:



建立如图所示坐标系，做如下定义

- 直线方程：$x\sin\theta - y\cos\theta + \rho = 0$
- 对于点$(x,y) \in I$

  $$b(x,y) = \begin{cases} 1, 如果点(x,y)位于区域内部 \\ 0, 如果点(x,y)位于区域外部 \end{cases}$$

  易得，
- 区域面积: $A = \iint\limits_I b(x,y)dxdy$
- 区域的x-轴中心: $\bar{x} = \frac{1}{A}\iint\limits_I x\, b(x,y)dxdy$
- 区域的y-轴中心: $\bar{y} = \frac{1}{A}\iint\limits_I y\, b(x,y)dxdy$
- 点(x,y)到直线距离：$r = |\frac{x\sin\theta - y\cos\theta + \rho}{\sqrt{\sin^2\theta + \cos^2\theta}}| = |x\sin\theta - y\cos\theta + \rho|$

  所以，这里我们的优化目标为

$$E=\iint_I(x\sin\theta-y\cos\theta+\rho)^2\,b(x,y)dxdy$$

对E关于ρ求导可得，

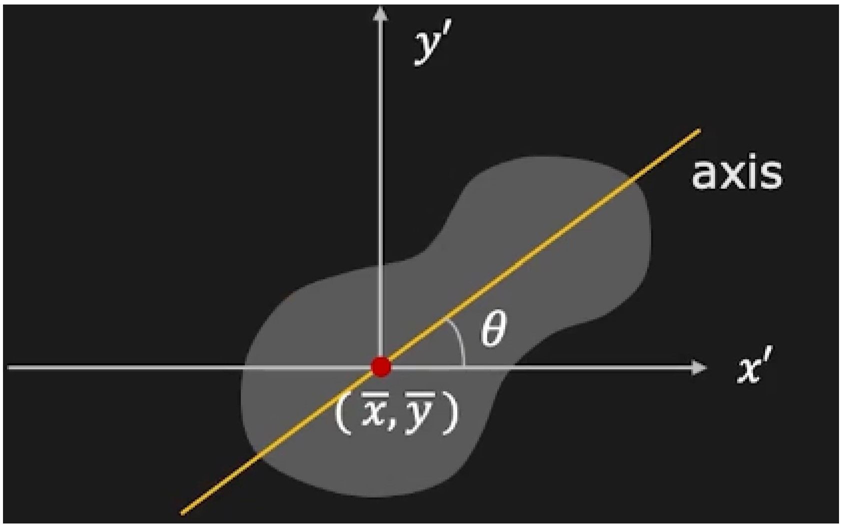
$$\frac{\partial E}{\partial \rho}=0\Longrightarrow A(\bar{x}\sin\theta-\bar{y}\cos\theta+\rho)=0$$

这里$A\neq 0$, 所以只能有 $\bar{x}\sin\theta-\bar{y}\cos\theta+\rho=0$ ，即直线一定穿过中心点$(\bar{x},\bar{y})$，因此我们将坐标系进行变换，即

$$\begin{cases}x'=x-\bar{x}\\ y'=y-\bar{y}\end{cases}$$

故而，直线变换为：

$$x\sin\theta-y\cos\theta+\rho=x'\sin\theta-y'\cos\theta$$



进一步，优化目标E可化简为：

$$E=a\sin^2\theta-b\sin\theta\cos\theta+c\cos^2\theta$$

这里a,b,c 的计算公式如下

$$a=\iint_{I'}\left(x'\right)^2 b(x,y)dx'dy'$$

$$b=2\iint_{I'}\left(x'y'\right)b(x,y)dx'dy'$$

$$c=\iint_{I'}\left(y'\right)^2 b(x,y)dx'dy'$$

令E关于θ的偏导为0，

$$\frac{dE}{d\theta} = (a-c)\sin 2\theta - b\cos 2\theta = 0$$

我们可以得到，

$$\tan 2\theta = \frac{b}{a-c}$$

$$\theta = \frac{1}{2}\arctan\left(\frac{b}{a-c}\right)$$