Problem 1. ([Al] Ex. 3.2-3.3)

We discussed the discriminant functions $g_i(x), i \in [K]$ where K is the number of classes. When $K = 2$ we can also define a single discriminant

$$g(x) = g_1(x) - g_2(x)$$

and we choose C_1 if $g(x) > 0$ and C_2 if $g(x) < 0$.

1. In a two-class problem, the likelihood ratio is

$$\frac{p(x \mid C_1)}{p(x \mid C_2)}$$

Write a discriminant function in terms of the likelihood ratio.

Solution. We could introduce a discriminant function as

$$g(x) = \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \frac{P(x \mid C_1)}{P(x \mid C_2)} \frac{P(C_1)}{P(C_2)}$$

Then, we choose C_1 if $g(x) \geq 1$ and C_2 if $g(x) < 1$. ■

2. In a two-class problem, the log odds is defined as

$$\log \frac{P(C_1 \mid x)}{P(C_2 \mid x)}$$

Write a discriminant function in terms of the log odds.

Solution. We could introduce a similar discriminant function as

$$g(x) = \log \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \log \frac{P(x \mid C_1)}{P(x \mid C_2)} + \log \frac{P(C_1)}{P(C_2)}$$

Then, we choose C_1 if $g(x) \geq 0$ and C_2 if $g(x) < 0$. ■

Problem 2. ([AI] Ex. 3.4)

In a two-class, two-action problem, if the loss function is $\lambda_{11} = \lambda_{22} = 0, \lambda_{12} = 10$ and $\lambda_{21} = 5$, write the optimal decision rule. How does the rule change if we add a third action of reject with $\lambda = 1$? [Note: we don’t have 0/1 loss for this problem.]

Solution. The expected risks are

$$R(\alpha_1 \mid x) = \lambda_{11} P(C_1 \mid x) + \lambda_{12} P(C_2 \mid x) = 10 P(C_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21} P(C_1 \mid x) + \lambda_{22} P(C_2 \mid x) = 5 P(C_1 \mid x)$$

We choose C_1 if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$, or $10P(C_2 \mid x) < 5P(C_1 \mid x)$, $P(C_1 \mid x) > \frac{2}{3}$, choose C_2 if $P(C_1 \mid x) \leq \frac{2}{3}$. The risk of reject is

$$R(\alpha_3 \mid x) = \lambda P(C_1 \mid x) + \lambda P(C_2 \mid x) = \lambda = 1$$

Then, we choose C_1 if

$$\begin{cases} R(\alpha_1 \mid x) < R(\alpha_2 \mid x) \\ R(\alpha_1 \mid x) < R(\alpha_3 \mid x) \end{cases} \Rightarrow \begin{cases} 10P(C_2 \mid x) < 5P(C_1 \mid x) \\ 10P(C_2 \mid x) < 1 \end{cases} \Rightarrow \begin{cases} P(C_1 \mid x) > \frac{2}{3} \\ P(C_1 \mid x) > \frac{9}{10} \end{cases} \Rightarrow P(C_1 \mid x) > \frac{9}{10}$$

We choose C_2 if

$$\begin{cases} R(\alpha_2 \mid x) < R(\alpha_1 \mid x) \\ R(\alpha_2 \mid x) < R(\alpha_3 \mid x) \end{cases} \Rightarrow \begin{cases} 5P(C_1 \mid x) < 10P(C_2 \mid x) \\ 5P(C_1 \mid x) < 1 \end{cases} \Rightarrow \begin{cases} P(C_1 \mid x) < \frac{2}{3} \\ P(C_1 \mid x) < \frac{1}{5} \end{cases} \Rightarrow P(C_1 \mid x) < \frac{1}{5}$$

We reject if

$$\begin{cases} R(\alpha_3 \mid x) \leq R(\alpha_1 \mid x) \\ R(\alpha_3 \mid x) \leq R(\alpha_2 \mid x) \end{cases} \Rightarrow \begin{cases} 1 \leq 10P(C_2 \mid x) \\ 1 \leq 5P(C_1 \mid x) \end{cases} \Rightarrow \begin{cases} P(C_1 \mid x) \leq \frac{9}{10} \\ P(C_1 \mid x) \geq \frac{1}{5} \end{cases} \Rightarrow \frac{1}{5} \leq P(C_1 \mid x) \leq \frac{9}{10}$$

■

Problem 3. (Poisson MLE)

Let X be a random variable. $X \sim$ Poisson (λ) with the density

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

1. Find $\mathbb{E}[X]$ and $\text{Var}(X)$ if $X \sim$ Poisson (λ).

Solution.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \text{Img(X)}} x\mathbb{P}(X = x) \\ &= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^x}{x!} \\ &= e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda}\lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= e^{-\lambda}\lambda e^{\lambda} \\ &= \lambda \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \sum_{x \in \text{Img(X)}} x^2 \mathbb{P}(X = x) \\ &= \sum_{x=0}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} x^2 \frac{\lambda^x}{x!} \\ &= e^{-\lambda}\lambda \sum_{x=1}^{\infty} (x-1+1) \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda}\lambda \left(\sum_{x=1}^{\infty} (x-1) \frac{\lambda^{x-1}}{(x-1)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\ &= e^{-\lambda}\lambda \left(\lambda \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\ &= e^{-\lambda}\lambda \left(\lambda \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right) \\ &= e^{-\lambda}\lambda(\lambda e^{\lambda} + e^{\lambda}) \\ &= \lambda^2 + \lambda \\ \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

■

2. Consider the sample $\mathcal{X} = \{x_n\}_{n=1}^N$ where $x_n \sim^{i.i.d.}$ Poisson(λ). For the parameter λ above, write the likelihood $l(\lambda \mid \mathcal{X})$ and the log-likelihood $\mathcal{L}(\lambda \mid \mathcal{X})$.

Solution. Since they are i.i.d. samples,

$$l(\lambda \mid \mathcal{X}) = \prod_{n=1}^{N} \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}$$

By taking the logarithm of the likelihood,

$$\begin{aligned}\mathcal{L}(\lambda \mid \mathcal{X}) &= \log\left(\prod_{n=1}^{N} \frac{\lambda^{x_n} e^{-\lambda}}{x_n!}\right) \\ &= \sum_{n=1}^{N} \log\left(\frac{\lambda^{x_n} e^{-\lambda}}{x_n!}\right) \\ &= \sum_{n=1}^{N} \log(\lambda^{x_n}) + \log(e^{-\lambda}) - \log(x_n!) \\ &= -n\lambda + \log(\lambda)\sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\log(x_n!)\end{aligned}$$

■

3. Find the maximum likelihood estimator $\hat{\lambda}_{\text{MLE}}$.

Solution. To maximize $\mathcal{L}(\lambda \mid \mathcal{X})$, we need to solve

$$\hat{\lambda}_{\text{MLE}} = \underset{\lambda}{\text{argmax}} = \underbrace{-n\lambda + \log(\lambda)\sum_{n=1}^{N} x_n - \sum_{n=1}^{N}\log(x_n!)}_{f(\lambda)}$$

By the first order condition of the maximum,

$$\frac{\mathrm{d}f}{\mathrm{d}\lambda} = -n + \frac{1}{\lambda}\sum_{n=1}^{N} x_n = 0 \quad \Rightarrow \quad \hat{\lambda}_{\text{MLE}} = \frac{1}{n}\sum_{n=1}^{N} x_n$$

■

4. Is $\hat{\lambda}_{\text{MLE}}$ biased?

Solution. The bias of the estimator is

$$d_{\lambda}(\hat{\lambda}_{\text{MLE}}) = \mathbb{E}[\hat{\lambda}_{\text{MLE}}] - \lambda = \mathbb{E}\left[\frac{1}{n}\sum_{n=1}^{N} x_n\right] - \lambda = \frac{1}{n}\sum_{n=1}^{N}\mathbb{E}[x_n] - \lambda = \lambda - \lambda = 0$$

therefore it is unbiased. ■

Problem 4. (Uniform MLE) Let X be a random variable. $X \sim \text{Unif}(\theta)$ with the density

$$p(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \le \text{x} \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

1. Find $\mathbb{E}[X]$ and $\text{Var}(X)$ if $X \sim \text{Unif}(\theta)$.

Solution.

$$\begin{aligned}\mathbb{E}[X] &= \int_0^{\theta} x\frac{1}{\theta}\,\mathrm{d}x \\ &= \frac{1}{\theta}\left.\frac{x^2}{2}\right|_{x=0}^{\theta} \\ &= \frac{\theta}{2}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^{\theta} x^2\frac{1}{\theta}\,\mathrm{d}x \\ &= \frac{1}{\theta}\left.\frac{x^3}{3}\right|_{x=0}^{\theta} = \frac{\theta^2}{3} \\ \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}\end{aligned}$$

■

2. Consider the sample $\mathcal{X} = \{x_n\}_{n=1}^N$ where $x_n \sim^{i.i.d.} \text{Unif}(\theta)$. For the parameter θ above, write the likelihood $l(\theta \mid \mathcal{X})$ and the log-likelihood $\mathcal{L}(\theta \mid \mathcal{X})$.

Solution. Suppose $I(\cdot)$ is the indicator function. The likelihood function is,

$$l(\theta \mid \mathcal{X}) = \prod_{n=1}^N p(x_n \mid \theta) = \frac{1}{\theta^N} I\left(\{x_n\}_{n=1}^N \in [0, \theta]\right) = \frac{1}{\theta^N} I\left(\max \{x_n\}_{n=1}^N \leq \theta\right)$$

By taking the logarithm of the likelihood,

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log\left(\frac{1}{\theta^N} I\left(\max \{x_n\}_{n=1}^N \leq \theta\right)\right) = -N\log(\theta) + \log\left(I\left(\max \{x_n\}_{n=1}^N \leq \theta\right)\right)$$

■

3. Find the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$.

Solution. When $\theta < \max \{x_n\}_{n=1}^N$, $l(\theta \mid \mathcal{X}) = 0$. When $\theta \geq \max \{x_n\}_{n=1}^N$, $l(\theta \mid \mathcal{X}) = \frac{1}{\theta^N}$.

Since $\frac{1}{\theta^N}$ is monotonically decreasing, the maximum likelihood estimator is $\hat{\theta}_{\text{MLE}} = \max \{x_n\}_{n=1}^N$. ■

4. Is $\hat{\theta}_{\text{MLE}}$ biased?

Solution. In order to take the expectation of $\hat{\theta}_{\text{MLE}}$, we need to find its distribution. The CDF of the estimator is obvious,

$$P(\hat{\theta}_{\text{MLE}} \leq m) = P(\max \{x_n\}_{n=1}^N \leq m) = P(x_1 \leq m, x_2 \leq m, \ldots, x_N \leq m) = \underbrace{\left(\frac{m}{\theta}\right)^N}_{F(m)}$$

Then, we could get the PDF by,

$$f(m) = \frac{\mathrm{d}F(m)}{\mathrm{d}m} = \frac{1}{\theta} N \left(\frac{m}{\theta}\right)^{N-1}$$

The bias of the estimator is,

$$d_\theta(\hat{\theta}_{\text{MLE}}) = \mathbb{E}[\hat{\theta}_{\text{MLE}}] - \theta = \int_0^\theta m \frac{1}{\theta} N \left(\frac{m}{\theta}\right)^{N-1} \mathrm{d}m - \theta = \frac{N}{\theta^N} \left.\frac{m^{N+1}}{N+1}\right|_{m=0}^{\theta} - \theta = \frac{N}{N+1}\theta - \theta = -\frac{\theta}{N+1}$$

therefore it is biased. ■

Problem 5. (See [Al] Ch.16.2.2) Find $\hat{q}_{\text{MAP}}$ for the Bernoulli likelihood

$$p(\mathcal{X} \mid q) = \prod_{n=1}^N q^{x_n}(1-q)^{1-x_n}$$

with the beta prior

$$p(q) = beta(q \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1}$$

Solution.

$$\begin{aligned}
\hat{q}_{\mathrm{MAP}} &= \operatorname*{argmax}_q \ \mathbb{P}(q \mid \mathcal{X}) = \operatorname*{argmax}_q \ \log \mathbb{P}(q \mid \mathcal{X}) = \operatorname*{argmax}_q \ \log \frac{\mathbb{P}(\mathcal{X} \mid q)\mathbb{P}(q)}{\mathbb{P}(\mathcal{X})} = \operatorname*{argmax}_q \ \log \mathbb{P}(\mathcal{X} \mid q)\mathbb{P}(q) \\
&= \operatorname*{argmax}_q \ \log \prod_{n=1}^N \mathbb{P}(x_n \mid q)\mathbb{P}(q) = \operatorname*{argmax}_q \ \sum_{n=1}^N \log \mathbb{P}(x_n \mid q) + \log \mathbb{P}(q) \\
&= \operatorname*{argmax}_q \ \underbrace{\sum_{n=1}^N x_n \log q + (1-x_n)\log(1-q) + \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + (\alpha-1)\log q + (\beta-1)\log(1-q)}_{\mathcal{L}}
\end{aligned}$$

By the first order condition of the maximum,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial q} &= \sum_{n=1}^N \frac{\partial}{\partial q} x_n \log q + \frac{\partial}{\partial q}(1-x_n)\log(1-q) + \frac{\partial}{\partial q} \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + \frac{\partial}{\partial q}(\alpha-1)\log q + \frac{\partial}{\partial q}(\beta-1)\log(1-q) \\
&= \frac{1}{q}\sum_{n=1}^N x_n - \frac{1}{1-q}\sum_{n=1}^N (1-x_n) + 0 + \frac{\alpha-1}{q} - \frac{\beta-1}{1-q}
\end{aligned}$$

Let $\frac{\partial \mathcal{L}}{\partial q} = 0$ and we have

$$\begin{aligned}
\frac{1}{q}\sum_{n=1}^N x_n - \frac{1}{1-q}\sum_{n=1}^N (1-x_n) + \frac{\alpha-1}{q} - \frac{\beta-1}{1-q} &= 0 \\
q\left(\sum_{n=1}^N (1-x_n) + \beta - 1\right) &= (1-q)\left(\sum_{n=1}^N x_n + \alpha - 1\right) \\
q\left(\sum_{n=1}^N (1-x_n) + \sum_{n=1}^N x_n + \beta - 1 + \alpha - 1\right) &= \sum_{n=1}^N x_n + \alpha - 1 \\
q\left(N + \beta + \alpha - 2\right) &= \sum_{n=1}^N x_n + \alpha - 1 \\
q &= \frac{\sum_{n=1}^N x_n + \alpha - 1}{N + \beta + \alpha - 2}
\end{aligned}$$

We have

$$\hat{q}_{\mathrm{MAP}} = \frac{\sum_{n=1}^N x_n + \alpha - 1}{N + \beta + \alpha - 2}$$

■

Problem 6. (Exponential family) A probability distribution in the exponential family is given by

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right)$$

where $\boldsymbol{\eta}$ is the parameter vector.

1. Prove that $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$ with identity covariance (where $\boldsymbol{\mu}$ is the parameter) is in the exponential family.

   Solution. Suppose $\boldsymbol{x} \in \mathbb{R}^d$. For $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$, we have,

$$\begin{aligned}
p(\boldsymbol{x} \mid \boldsymbol{\mu}) &= (2\pi)^{\frac{-d}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top(\boldsymbol{x}-\boldsymbol{\mu})\right) \\
&= (2\pi)^{\frac{-d}{2}} \exp\left(\boldsymbol{\mu}^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\mu}\right) \\
&= (2\pi)^{\frac{-d}{2}} \exp\left(\langle \boldsymbol{\mu}, \boldsymbol{x} \rangle + \left\langle \operatorname{vec}\left(-\frac{1}{2}\boldsymbol{I}\right), \operatorname{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \right\rangle - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\mu}\right)
\end{aligned}$$

Then, we write $h(\boldsymbol{x}) = (2\pi)^{\frac{-d}{2}}$, $T(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \\ \operatorname{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \end{bmatrix}$, $\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\mu} \\ \operatorname{vec}\left(-\frac{1}{2}\boldsymbol{I}\right) \end{bmatrix}$, $A(\boldsymbol{\eta}) = \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\eta}^\top\boldsymbol{\eta} - \frac{d}{4})$ ■

2. Prove that
$$\nabla_{\boldsymbol{\eta}} A = \mathbb{E}_{\mathbf{x}\sim p(\boldsymbol{x}|\boldsymbol{\eta})}[T(\mathbf{x})].$$
Hint: Use the fact that $\int p(\boldsymbol{x} \mid \boldsymbol{\eta}) d\boldsymbol{x} = 1$ to get an expression of A first.

Solution. As the Hint shows,
$$\int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \mathrm{d}\boldsymbol{x} = 1$$
$$\exp(-A(\boldsymbol{\eta})) \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} = 1$$
$$A(\boldsymbol{\eta}) = \log \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}$$

Then, we take the derivative,
$$\begin{aligned}\nabla_{\boldsymbol{\eta}} A &= \frac{\partial}{\partial \boldsymbol{\eta}^\top}\left(\log \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}\right) \\ &= \frac{\int T(\boldsymbol{x}) h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}}{\int h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}} \\ &= \int T(\boldsymbol{x}) h(\boldsymbol{x}) \exp\left(\boldsymbol{\eta}^\top T(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \mathrm{d}\boldsymbol{x} \\ &= \mathbb{E}_{\mathbf{x}\sim p(\boldsymbol{x}|\boldsymbol{\eta})}[T(\mathbf{x})]\end{aligned}$$
■

3. Verify Part 2 using the example in Part 1.

Solution. We first consider

$$(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top])_{ij} = \mathbb{E}[(\boldsymbol{x}\boldsymbol{x}^\top)_{ij}] = \mathbb{E}[x_i x_j] = \operatorname{cov}(x_i, x_j) + \mathbb{E}[x_i]\mathbb{E}[x_j] = \Sigma_{ij} + \mu_i\mu_j$$

$$\mathbb{E}\left[\operatorname{vec}(\boldsymbol{x}\boldsymbol{x}^\top)\right] = \mathbb{E}\left[\begin{pmatrix} x_1^2 \\ \vdots \\ x_1 x_d \\ \vdots \\ x_d^2 \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}\left[x_1^2\right] \\ \vdots \\ \mathbb{E}\left[x_1 x_d\right] \\ \vdots \\ \mathbb{E}\left[x_d^2\right] \end{pmatrix} = \begin{pmatrix} \Sigma_{11} + \mu_1^2 \\ \vdots \\ \Sigma_{1d} + \mu_1\mu_d \\ \vdots \\ \Sigma_{dd} + \mu_d^2 \end{pmatrix} = \operatorname{vec}\left(\mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^\top\right]\right)$$

Therefore,
$$\mathbb{E}[T(\boldsymbol{x})] = \mathbb{E}\begin{bmatrix} \boldsymbol{x} \\ \operatorname{vec}\left(\boldsymbol{x}\boldsymbol{x}^\top\right) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \operatorname{vec}\left(\boldsymbol{I} + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right) \end{bmatrix}$$

We can verify that $\nabla_{\boldsymbol{\eta}} A = \mathbb{E}[T(\boldsymbol{x})]$ by direct differentiation. ■

Problem 1. KL and entropy

The Kullback-Leibler (KL) divergence of a distribution $p(\boldsymbol{x})$ from another distribution $q(\boldsymbol{x})$ is given by

$$D_{\mathrm{KL}}(p\|q) = -\int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}$$

Prove that $D_{\mathrm{KL}}(p\|q) \geq 0$.

Solution.

Theorem 1. Jensen's inequality *Suppose X is an integrable random variable, $f : \mathbb{R} \to \mathbb{R}$ is a concave function, such that $Y = f(X)$ is also integrable, then,*

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

Proof. Since f is concave, if $x, y \in \mathbb{R}$, we have

$$f(x) \leq f(y) + f'(y)(x - y)$$

Let $x = X$, $y = \mathbb{E}[X]$, then

$$f(X) \leq f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X])$$

This holds for all X. Thus, we could take the expectation on both side,

$$\mathbb{E}[f(X)] \leq \mathbb{E}[f(\mathbb{E}[X])] + \mathbb{E}[f'(\mathbb{E}[X])(X - \mathbb{E}[X])] = f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(\mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]]) = f(\mathbb{E}[X])$$

□

By Jensen's inequality, since log is concave, we have

$$\begin{aligned}
-D_{\mathrm{KL}}(p\|q) &= \int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right)\right] \\
&\leq \log\left(\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right]\right) \\
&= \log \int p(\boldsymbol{x}) \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\
&= 0
\end{aligned}$$

You can also solve it by Log Sum Inequality or Gibbs' Inequality. ■

* If you have much time, also think about the following problems. Even if you don't think about them, we will cover them later in this course. These will not be covered in the recitation.

1. The Kullback-Leibler (KL) divergence of a distribution $p(\boldsymbol{x})$ from another distribution $q(\boldsymbol{x})$ is given by

$$D_{\mathrm{KL}}(p\|q) = -\int p(\boldsymbol{x}) \log\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right) d\boldsymbol{x}$$

Let $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $q(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{S})$. Calculate $D_{\mathrm{KL}}(p\|q)$.

Solution. The density functions for $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ are

$$p(x) = \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \quad q(x) = \frac{1}{(2\pi)^{n/2} \det(\boldsymbol{S})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^\top \boldsymbol{S}^{-1}(\boldsymbol{x} - \boldsymbol{m})\right)$$

Then,

$$\begin{aligned}
D\left(p\|q\right) &= \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\log p(\boldsymbol{x}) - \log q(\boldsymbol{x})\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\log\det\boldsymbol{\Sigma} - \left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right) + \log\det\boldsymbol{S} + \left(\boldsymbol{x}-\boldsymbol{m}\right)^\top \boldsymbol{S}^{-1}\left(\boldsymbol{x}-\boldsymbol{m}\right)\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right) + \left(\boldsymbol{x}-\boldsymbol{m}\right)^\top \boldsymbol{S}^{-1}\left(\boldsymbol{x}-\boldsymbol{m}\right)\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top\right) + \operatorname{tr}\left(\boldsymbol{S}^{-1}\left(\boldsymbol{x}-\boldsymbol{m}\right)\left(\boldsymbol{x}-\boldsymbol{m}\right)^\top\right)\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} + \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{S}^{-1}\left(\boldsymbol{x}\boldsymbol{x}^\top - 2\boldsymbol{x}\boldsymbol{m}^\top + \boldsymbol{m}\boldsymbol{m}^\top\right)\right)\right] \\
&= \frac{1}{2}\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} - \frac{1}{2}n + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{S}^{-1}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top - 2\boldsymbol{m}\boldsymbol{\mu}^\top + \boldsymbol{m}\boldsymbol{m}^\top\right)\right) \\
&= \frac{1}{2}\left(\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} - n + \operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\boldsymbol{\mu}^\top\boldsymbol{S}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^\top\boldsymbol{S}^{-1}\boldsymbol{m} + \boldsymbol{m}^\top\boldsymbol{S}^{-1}\boldsymbol{m}\right)\right) \\
&= \frac{1}{2}\left(\log\frac{\det\boldsymbol{S}}{\det\boldsymbol{\Sigma}} - n + \operatorname{tr}\left(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}\right) + \left(\boldsymbol{m}-\boldsymbol{\mu}\right)^\top\boldsymbol{S}^{-1}\left(\boldsymbol{m}-\boldsymbol{\mu}\right)\right)
\end{aligned}$$

■

2. The entropy of a distribution $p(\boldsymbol{x})$ is given by

$$H(p) = -\int p(\boldsymbol{x})\log p(\boldsymbol{x}) d\boldsymbol{x}.$$

Calculate $H(p)$ where $p(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Solution. The density function for $p(\boldsymbol{x})$ is

$$p(x) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Then,

$$\begin{aligned}
H(x) &= -\int p(\boldsymbol{x})\log p(\boldsymbol{x}) d\boldsymbol{x} \\
&= -\mathbb{E}\left[\log\left(\frac{1}{(2\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right)\right] \\
&= -\mathbb{E}\left[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right)\right] \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\right]\right) \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{I}\right) \\
&= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}n
\end{aligned}$$

■

Problem 2. Ridge regression ([HaTF] Ex. 3.29)

Recall that in a ridge regression we minimize $\frac{1}{2}\|\boldsymbol{r}-\boldsymbol{X}\boldsymbol{w}\|^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2$. Suppose we run a ridge regression with parameter λ on a single variable x and get coefficient w (so the data matrix $\boldsymbol{X}$ is $N\times 1$, which can be denoted as a vector $\boldsymbol{x}\in\mathbb{R}^N$). We now include an exact copy $x^* = x$ and refit our ridge regression. Show that both coefficients are identical, and derive their value. Show in general that if m copies of a variable x_j are included in a ridge regression, their coefficients are all the same.

Solution. Let $\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2$, to find its minimum, we take the derivative,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = -\boldsymbol{X}^\top(\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}) + \lambda\boldsymbol{w}$$

By setting the derivative to 0, we have

$$\boldsymbol{w} = \left(\boldsymbol{X}^\mathrm{T}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\mathrm{T}\boldsymbol{r}$$

For $\boldsymbol{X} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{X}^\mathrm{T}\boldsymbol{X}$ is a scaler. We have

$$\hat{w} = \frac{\boldsymbol{X}^\mathrm{T}\boldsymbol{r}}{\boldsymbol{X}^\mathrm{T}\boldsymbol{X} + \lambda}$$

If we include a copy of $\boldsymbol{X}$, the target function turned to

$$\underset{w_1,w_2}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2\right\|^2 + \frac{1}{2}\lambda\left\|w_1\right\|^2 + \frac{1}{2}\lambda\left\|w_2\right\|^2$$

Let

$$\mathcal{L}(w_1, w_2) = \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2\right\|^2 + \frac{1}{2}\lambda\left\|w_1\right\|^2 + \frac{1}{2}\lambda\left\|w_2\right\|^2$$

Take derivative with respect to w_1 and w_2

$$\begin{aligned}\frac{\partial \mathcal{L}(w_1, w_2)}{\partial w_1} &= -\boldsymbol{X}^\mathrm{T}(\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2) + \lambda w_1\\ \frac{\partial \mathcal{L}(w_1, w_2)}{\partial w_2} &= -\boldsymbol{X}^\mathrm{T}(\boldsymbol{r} - \boldsymbol{X}w_1 - \boldsymbol{X}w_2) + \lambda w_2\end{aligned}$$

Set them to 0 and solve the system of equations, we have

$$\begin{bmatrix}\hat{w}_1\\ \hat{w}_2\end{bmatrix} = \frac{\boldsymbol{X}^\mathrm{T}\boldsymbol{r}}{2\boldsymbol{X}^\mathrm{T}\boldsymbol{X} + \lambda}\begin{bmatrix}1\\ 1\end{bmatrix}$$

They are the same.

If we have m copies of $\boldsymbol{X}$, the target function become

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}\sum_{j=1}^{m} w_j\right\|^2 + \frac{1}{2}\lambda\sum_{j=1}^{m}\|w_j\|^2$$

Let

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{X}\sum_{j=1}^{m} w_j\right\|^2 + \frac{1}{2}\lambda\sum_{j=1}^{m}\|w_j\|^2$$

Take derivative with respect to w_k for $1 \le k \le m$

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial w_k} = \boldsymbol{X}^\mathrm{T}\boldsymbol{X}\sum_{j=1}^{m} w_j - \boldsymbol{X}^\mathrm{T}\boldsymbol{r} + \lambda w_k$$

Set it to 0 and we have

$$\boldsymbol{X}^\mathrm{T}\boldsymbol{X}\sum_{j=1}^{m}\hat{w}_j + \lambda\hat{w}_k = \boldsymbol{X}^\mathrm{T}\boldsymbol{r}$$

We could see that if we change k to another k', the solution will be the same. Thus, all $\hat{w}_k$ are the same

$$\hat{w}_k = \frac{\boldsymbol{X}^\mathrm{T}\boldsymbol{r}}{m\boldsymbol{X}^\mathrm{T}\boldsymbol{X} + \lambda}, \quad 1 \le k \le m$$

■

Problem 3. Elastic net ([HaTF] Ex. 3.30)

Consider the elastic-net optimization problem:

$$\min_{\boldsymbol{w}} \|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\left[\alpha\|\boldsymbol{w}\|^2 + (1-\alpha)\|\boldsymbol{w}\|_1\right]$$

Show how one can turn this into a lasso problem using an augmented version of $\boldsymbol{X}$ and r :

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \gamma\boldsymbol{I} \end{bmatrix}, \text{ and } \quad \tilde{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{0} \end{bmatrix}$$

Solution. Suppose $\boldsymbol{r} \in \mathbb{R}^N$, $\boldsymbol{X} \in \mathbb{R}^{N\times(D+1)}$, and $\boldsymbol{w} \in \mathbb{R}^{(D+1)}$. Then,

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} \boldsymbol{X} \\ \gamma\boldsymbol{I}_{D+1} \end{bmatrix} \in \mathbb{R}^{(N+D+1)\times(D+1)}, \quad \text{and } \tilde{\boldsymbol{r}} = \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{0}_{D+1} \end{bmatrix} \in \mathbb{R}^{(N+D+1)}$$

Thus, the lasso problem of $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{r}}$ is,

$$\left\|\tilde{\boldsymbol{r}} - \tilde{\boldsymbol{X}}\boldsymbol{w}\right\|^2 + \tilde{\lambda}\|\boldsymbol{w}\|_1 = \left\| \begin{matrix} \boldsymbol{r} - \boldsymbol{X}\boldsymbol{w} \\ \gamma\boldsymbol{w} \end{matrix} \right\|^2 + \tilde{\lambda}\|\boldsymbol{w}\|_1 = \|\boldsymbol{r} - \boldsymbol{X}\boldsymbol{w}\|^2 + \gamma^2\|\boldsymbol{w}\|^2 + \tilde{\lambda}\|\boldsymbol{w}\|_1$$

To make it a lasso problem, we need $\gamma^2 = \alpha\lambda$, $\gamma = \sqrt{\lambda\alpha}$, and $\tilde{\lambda} = \lambda(1-\alpha)$.

Thus, to solve the elastic-net optimization problem, we first augment $\boldsymbol{X}$, $\boldsymbol{r}$ with $\gamma = \sqrt{\lambda\alpha}$, and solve the lasso problem of $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{r}}$ with $\tilde{\lambda} = \lambda(1-\alpha)$. ■

Problem 4. Kernel

1. Suppose $K(\boldsymbol{x}) \geq 0$ and $\int_{\mathbb{R}^d} K(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$. Show that the kernel estimator

$$\hat{p}(\boldsymbol{x}) = \frac{1}{Nh^d}\sum_{n=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right)$$

is a density.

Solution. Suppose we have $\mathcal{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, and we define,

$$\hat{f}(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} = \frac{1}{N}\mathbb{1}_{\mathcal{X}}(\boldsymbol{x}) = \begin{cases} \frac{1}{N} & \text{if } \boldsymbol{x} \in \mathcal{X} \\ 0 & \text{if } \boldsymbol{x} \notin \mathcal{X} \end{cases}, \quad \hat{K}(\boldsymbol{u}) = \frac{1}{h^d}K(\frac{\boldsymbol{u}}{h})$$

Then, the convolution of $\hat{f}$ and $\hat{K}$ is,

$$\begin{aligned}\left(\hat{f} * \hat{K}\right)(\boldsymbol{x}) &= \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{u})\hat{K}(\boldsymbol{x} - \boldsymbol{u})\ \mathrm{d}\boldsymbol{u} \\ &= \frac{1}{N}\sum_{n=1}^{N}\hat{K}(\boldsymbol{x} - \boldsymbol{x}_n) \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^d}K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right) \\ &= \frac{1}{Nh^d}\sum_{n=1}^{N}K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right) = \hat{p}(\boldsymbol{x})\end{aligned}$$

Then,

$$\begin{aligned}
\int_{\mathbb{R}^d} \hat{p}(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x} &= \int_{\mathbb{R}^d} \left(\hat{f} * \hat{K}\right)(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \hat{f}(\boldsymbol{u})\hat{K}(\boldsymbol{x}-\boldsymbol{u})\ \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d} \hat{K}(\boldsymbol{x}-\boldsymbol{u})\ \mathrm{d}\boldsymbol{x}\right)\hat{f}(\boldsymbol{u})\ \mathrm{d}\boldsymbol{u} \\
&= \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{u})\ \mathrm{d}\boldsymbol{u} \\
&= 1
\end{aligned}$$

■

2. Suppose $K(\boldsymbol{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\boldsymbol{x}\|^2}{2}\right]$. Show that each $K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$ can be written as a product of d univariate kernels.

Solution. Suppose $\boldsymbol{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{bmatrix}$, and $\boldsymbol{x}_n = \begin{bmatrix} x_n^{(1)} \\ x_n^{(2)} \\ \vdots \\ x_n^{(d)} \end{bmatrix}$.

The univariate Gaussian kernel is

$$K_0(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$$

Then,

$$\begin{aligned}
K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right) &= \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\|^2}{2}\right] \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\boldsymbol{x}-\boldsymbol{x}_n\|^2}{2h^2}\right] \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\left(x^{(1)}-x_n^{(1)}\right)^2+\left(x^{(2)}-x_n^{(2)}\right)^2+\ldots+\left(x^{(d)}-x_n^{(d)}\right)^2}{2h^2}\right] \\
&= \prod_{k=1}^d \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(x^{(k)}-x_n^{(k)}\right)^2}{2h^2}\right) \\
&= \prod_{k=1}^d K_0\left(\frac{x^{(k)}-x_n^{(k)}}{h}\right)
\end{aligned}$$

■

Problem 5. Smoother ([HaTF] Ex.6.8)

Suppose for continuous response Y and predictor X we model the joint density of X, Y using a multivariate Gaussian kernel estimator. This means that

$$\hat{p}(x,y) = \frac{1}{Nh^2}\sum_{n=1}^N K_h\left(x-x_n\right)K_h\left(y-y_n\right)$$

where $K_h(x) = K(x/h)$ and K is the Gaussian kernel. (cf. Problem 4 above.) Show that the conditional mean $\mathbb{E}[Y \mid X]$ derived from this estimate is a Nadaraya-Watson estimator.

Solution. By the definition of the conditional expectation,

$$\mathbb{E}[Y \mid X] = \int p(y \mid x)y\ \mathrm{d}y = \frac{\int p(x,y)y\ \mathrm{d}y}{p(x)}$$

The marginal distribution is,

$$\hat{p}(x) = \int \hat{p}(x, y) \, \mathrm{d}y = \frac{1}{Nh} \sum_{n=1}^{N} K_h(x - x_n)$$

Therefore,

$$\begin{aligned}
\mathbb{E}[Y \mid X] &= \int p(y \mid x) y \, \mathrm{d}y \\
&= \frac{\int p(x, y) y \, \mathrm{d}y}{p(x)} \\
&= \frac{\int \frac{1}{Nh^2} \sum_{n=1}^{N} K_h(x - x_n) K_h(y - y_n) y \, \mathrm{d}y}{\frac{1}{Nh} \sum_{n=1}^{N} K_h(x - x_n)} \\
&= \frac{\sum_{n=1}^{N} K_h(x - x_n)}{h \sum_{n=1}^{N} K_h(x - x_n)} \left(\int K_h(y - y_n) y \, \mathrm{d}y \right) \\
&= \frac{\sum_{n=1}^{N} K_h(x - x_n)}{h \sum_{n=1}^{N} K_h(x - x_n)} \left(\int K_h(y - y_n)(y - y_n) \, \mathrm{d}y + \int K_h(y - y_n) y_n \, \mathrm{d}y \right) \\
&= \frac{\sum_{n=1}^{N} K_h(x - x_n)}{h \sum_{n=1}^{N} K_h(x - x_n)} (0 + h y_n) \\
&= \frac{\sum_{n=1}^{N} K_h(x - x_n) y_i}{\sum_{n=1}^{N} K_h(x - x_n)}
\end{aligned}$$

Thus, it is a Nadaraya-Watson estimator. ■

Problem 6. EM (Midterm exam, Fall’21, Problem #-1)

Hilbert owns a PS5, an Xbox and a Switch (three different gaming systems). On each system there is a game. The outcome of each game is either win (“W”) or loss (“L”). Every day, he plays the Switch game. If he wins, he continues to play the PS5 game and records the outcome of the PS5 game; otherwise, he continues to play the Xbox game and records the outcome of the Xbox game. The outcomes he recorded for the last ten days of March are as follows:

W W L W W L W W L L

Suppose the event on each day is independent. Denote the (unknown) probabilities of ”W” for the PS5, the Xbox and the Switch games by p, q, π, respectively. Let y denote the random variable representing the final outcome, so that $y = 1$ if “W” is recorded, and $y = 0$ if “L” is recorded.

1. Suppose we use an expectation-maximization (EM) algorithm to find the maximum-likelihood solution of p, q, π. In plain language, describe what is the latent variable and the values it can take.

   Solution. The latent variable is the result of the Switch Game. It takes value in $\{0, 1\}$. ■

2. For the final outcome y, consider its parametric likelihood $p(y \mid p, q, \pi)$. Is it true or false that $p(y \mid p, q, \pi) = \sum_z (p(z \mid p, q, \pi) + p(y \mid z, p, q, \pi))$, where the summation is over all possible values of the latent variable?

   Solution. False. It is multiply: $p(y \mid p, q, \pi) = \sum_z p(z \mid p, q, \pi) p(y \mid z, p, q, \pi)$ ■

3. Write $p(y \mid p, q, \pi)$ as a function of y, p, q, π.

   Solution. Since π is the probability of wining the Switch game, we have

$$p(z = 1 \mid p, q, \pi) = \pi, \quad p(z = 0 \mid p, q, \pi) = 1 - \pi$$

If he wins the Switch game, he will play the PS5 with p wining rate, thus

$$p(y \mid z=1,p,q,\pi) = p^y(1-p)^{1-y}$$

If he loses the Switch game, he will play the Xbox with q wining rate, thus

$$p(y \mid z=0,p,q,\pi) = q^y(1-q)^{1-y}$$

Therefore, we have

$$\begin{aligned} p(y \mid p,q,\pi) &= \sum_{z\in\{0,1\}} p(z \mid p,q,\pi)p(y \mid z,p,q,\pi) \\ &= p(z=0 \mid p,q,\pi)p(y \mid z=0,p,q,\pi) + p(z=1 \mid p,q,\pi)p(y \mid z=1,p,q,\pi) \\ &= (1-\pi)q^y(1-q)^{1-y} + \pi p^y(1-p)^{1-y} \end{aligned}$$

■

4. Using the data recorded on the last ten days of March, and the initial values

$$\left(p^{(0)}, q^{(0)}, \pi^{(0)}\right) = (0.5, 0.5, 0.5)$$

implement the EM algorithm for one E-step and one M-step. Calculate $\left(p^{(1)}, q^{(1)}, \pi^{(1)}\right)$.

Solution.

To help with understanding, we write this EM as the form of GMM model in the lecture. This is a mixture of univariate Bernoulli. Let $\tilde{\boldsymbol{z}}$ be a one hot vector, and $\tilde{z}_k = 1$ implies the choice of the k-th cluster ($K = 2$ in our case). The marginal distribution over $\tilde{\boldsymbol{z}}$ is given by,

$$p(\tilde{z}_k = 1) = \tilde{\pi}_k, \text{ where } 0 \le \tilde{\pi}_k \le 1, \sum_{k=1}^{K} \tilde{\pi}_k = 1$$

Then, we can write

$$p(\tilde{\boldsymbol{z}}) = \prod_{k=1}^{K} \tilde{\pi}_k^{\tilde{z}_k}, \quad p(y \mid \tilde{z}_k = 1) = \mathcal{B}(y \mid \mu_k), \quad p(y \mid \tilde{\boldsymbol{z}}) = \prod_{k=1}^{K} \mathcal{B}(y \mid \mu_k)^{\tilde{z}_k}$$

where $\mathcal{B}(y \mid \mu_k) = \mu_k^y(1-\mu_k)^{(1-y)}$ is the univariate Bernoulli.

Therefore, the likelihood is

$$p(y) = \sum_{\tilde{\boldsymbol{z}}} p(\tilde{\boldsymbol{z}})p(y \mid \tilde{\boldsymbol{z}}) = \sum_{k=1}^{K} \tilde{\pi}_k \mathcal{B}(y \mid \mu_k)$$

In order to derive the EM algorithm, we first write down the complete-data log likelihood function,

$$\log p(\boldsymbol{y}, \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \tilde{z}_{nk} \left\{\log \tilde{\pi}_k + y_n \log \mu_k + (1-y_n)\log(1-\mu_k)\right\}$$

Then, we take the expectation of the complete-data log likelihood with respect to the posterior distribution of the latent variables

$$\mathbb{E}_{\boldsymbol{z}}[\log p(\boldsymbol{y}, \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu})] = \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\tilde{z}_{nk})\left\{\log\tilde{\pi}_k + y_n\log\mu_k + (1-y_n)\log(1-\mu_k)\right\}}_{\mathcal{L}}$$

In the E-step, we evaluate the responsibility term by Bayes' theorem,

$$\gamma(\tilde{z}_{nk}) = \mathbb{E}[\tilde{z}_{nk}] = \frac{\sum_{\tilde{z}_{nk}}\tilde{z}_{nk}[\tilde{\pi}_k p(y_n \mid \mu_K)]^{\tilde{z}_{nk}}}{\sum_{\tilde{z}_{nj}}[\tilde{\pi}_j p(y_n \mid \mu_K)]^{\tilde{z}_{nj}}} = \frac{\tilde{\pi}_k p(y_n \mid \mu_k)}{\sum_{j=1}^{K}\tilde{\pi}_j p(y_n \mid \mu_j)}$$

In the M-step, we maximize $\mathcal{L}$ with respect to μ_k and $\tilde{\pi}_k$. For μ_k, we have

$$\frac{\partial\mathcal{L}}{\partial\mu_k} = \sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\left(\frac{y_n}{\mu_k} + \frac{1-y_n}{1-\mu_k}\right)$$

By the first order condition of the maximum, we have

$$\begin{aligned}
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\left(\frac{y_n}{\mu_k} - \frac{1-y_n}{1-\mu_k}\right) &= 0 \\
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\frac{y_n-\mu_k}{\mu_k(1-\mu_k)} &= 0 \\
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})y_n - \sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\mu_k &= 0 \\
\mu_k &= \frac{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})y_n}{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})}
\end{aligned}$$

For $\tilde{\pi}_k$, we need a Lagrange multiplier to fulfil the constraint,

$$\tilde{\mathcal{L}} = \mathbb{E}_{\boldsymbol{z}}[\log p(\boldsymbol{y}, \tilde{\boldsymbol{z}} \mid \tilde{\boldsymbol{\pi}}, \boldsymbol{\mu})] + \lambda\left(\sum_{k=1}^{K}\tilde{\pi}_k - 1\right)$$

and

$$\frac{\partial\tilde{\mathcal{L}}}{\partial\tilde{\pi}_k} = \sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\frac{1}{\tilde{\pi}_k} + \lambda$$

By the first order condition of the maximum, we have

$$\begin{aligned}
\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})\frac{1}{\tilde{\pi}_k} + \lambda &= 0 \\
\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\tilde{z}_{nk}) + \sum_{k=1}^{K}\tilde{\pi}_k\lambda &= 0 \\
\lambda = -\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(\tilde{z}_{nk}) &= -N \\
\tilde{\pi}_k = -\frac{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})}{\lambda} &= \frac{\sum_{n=1}^{N}\gamma(\tilde{z}_{nk})}{N}
\end{aligned}$$

For our problem, we have $K = 2$. If $k = 1$, it means he wins the Switch game, $\tilde{\pi}_1 = \pi$, $\mu_1 = p$. If $k = 2$, it means he loses the Switch game, $\tilde{\pi}_2 = 1 - \pi$, $\mu_2 = q$.

E-step: Since $p = q$, all $\gamma(\tilde{z}_k)$ are the same,

$$\gamma(\tilde{z}_k) = p(z = 1 \mid y = 1) = \frac{p(z=1)p(y=1 \mid z=1)}{p(z=0)p(y=1 \mid z=0) + p(z=1)p(y=1 \mid z=1)} = \frac{\pi^{(0)}p^{(0)}}{\pi^{(0)}q^{(0)} + \pi^{(0)}p^{(0)}} = 0.5$$

M-step:

$$\pi^{(1)} = \frac{1}{N}\sum_{n=1}^{N}\gamma(\tilde{z}_k) = \frac{1}{10}(10)(0.5) = 0.5$$

$$p^{(1)} = \frac{\sum_{n=1}^{N} y_n\gamma(\tilde{z}_k)}{\sum_{n=1}^{N}\gamma(\tilde{z}_k)} = \frac{6(0.5)}{10(0.5)} = 0.6$$

$$q^{(1)} = \frac{\sum_{n=1}^{N} y_n\gamma(\tilde{z}_k)}{\sum_{n=1}^{N}\gamma(\tilde{z}_k)} = \frac{6(0.5)}{10(0.5)} = 0.6$$

■

Problem 1. General view of GMM [Bi] Ex. 9.9

Recall that The expected value of the complete-data log likelihood function for GMM is given by

$$\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(z_{nk}\right) \left\{\ln \pi_k + \ln \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\}$$

With a fixed $\gamma\left(z_{nk}\right)$, find the maximizer $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ for $\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})]$.

Solution. To find the minimum, we take the derivative with respect to $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$,

$$\begin{aligned}
\frac{\partial \mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})]}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \ln \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-\frac{1}{2} \ln(\det(2\pi \boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right) \\
&= \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})]}{\partial \boldsymbol{\Sigma}_k} &= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \ln \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-\frac{1}{2} \ln(\det(2\pi \boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right) \\
&= \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}\right)
\end{aligned}$$

By the first order condition of the maximum, we set the derivatives to 0 and get,

$$\sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right) = 0 \Rightarrow \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \boldsymbol{x}_n - \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \boldsymbol{\mu}_k = 0$$

$$\tilde{\boldsymbol{\mu}}_k = \frac{\sum_{n=1}^{N} \gamma\left(z_{nk}\right) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma\left(z_{nk}\right)}$$

$$\sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-\frac{1}{2} \boldsymbol{\Sigma}_k^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}\right) = 0 \Rightarrow \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \left(-1 + (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}\right) = 0$$

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{\sum_{n=1}^{N} \gamma\left(z_{nk}\right) (\boldsymbol{x}_n - \tilde{\boldsymbol{\mu}}_k)(\boldsymbol{x}_n - \tilde{\boldsymbol{\mu}}_k)^\top}{\sum_{n=1}^{N} \gamma\left(z_{nk}\right)}$$

■

Problem 2. K-means as the limit of EM cf. [Bi] Ch.9.3.2

Consider the EM algorithm where the covariance matrices of the mixture components are all given by $\boldsymbol{\Sigma}_k = \epsilon \boldsymbol{I}, k = 1, \cdots, K$.

1. Write $p\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$.

   Solution.

   $$p\left(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \det(2\pi \boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right) = (2\pi\epsilon)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\epsilon} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2\right\}$$

   ■

2. Show that $\gamma\left(z_{nk}\right) \to r_{nk}$ as $\epsilon \to 0$, where $r_{nk}=1$ if $k=\operatorname{argmin}_j \left\|\boldsymbol{x}_n-\boldsymbol{\mu}_j\right\|^2$ and $r_{nk}=0$ otherwise.

Solution.

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}\|\boldsymbol{x}_n-\boldsymbol{\mu}_j\|^2\right\}}$$

When $\epsilon \to 0$, in the denominator the term for which $\|\boldsymbol{x}_n-\boldsymbol{\mu}_j\|^2$ is smallest will go to zero most slowly. Therefore, $\gamma(z_{nk})$ will go to zero except for term j, for which it will go to 1.

Note that this holds independent of the value of $\boldsymbol{\pi}$ (as long as it is not 0). Each data point is thereby assigned to the cluster having the closest mean. ■

3. Show that as $\epsilon \to 0$,

$$\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\right\|^2 + \text{ const.}$$

Solution.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] &= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left\{\ln\pi_k + \ln\mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\right\} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\{\ln\pi_k - \frac{1}{2}\ln(2\pi\epsilon) - \frac{1}{2\epsilon}\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2\right\} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(-\frac{1}{2\epsilon}\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2\right) + \text{ const.}\end{aligned}$$

■

Problem 3. Rayleigh quotient

The Rayleigh quotient for a real symmetric matrix $\boldsymbol{A}$ and a nonzero vector $\boldsymbol{v}$ is given by

$$\rho(\boldsymbol{v},\boldsymbol{A}) = \frac{\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{v}^\top\boldsymbol{v}}.$$

Prove that the $\rho(\boldsymbol{v},\boldsymbol{A}) \in [\lambda_{\min},\lambda_{\max}]$ where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\boldsymbol{A}$, respectively. For what $\boldsymbol{v}$ does $\rho(\boldsymbol{v},\boldsymbol{A})$ achieve the min and the max, respectively?

Solution. Note that the Rayleigh quotient is scaling invariant, i.e. $\rho(\boldsymbol{v},\boldsymbol{A}) = \rho(\alpha\boldsymbol{v},\boldsymbol{A})$. Without the loss of generality, we consider the following constrained problem:

$$\max_{\boldsymbol{v}\in\mathbb{R}^n:\|\boldsymbol{v}\|=1} \boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v}$$

Let $\boldsymbol{A}=\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ be the eigenvalue decomposition of $\boldsymbol{A}$, where $\boldsymbol{Q}=[\boldsymbol{q}_1,\ldots,\boldsymbol{q}_n]$ are orthogonal eigenvectors, $\boldsymbol{\Lambda}=\operatorname{diag}(\lambda_1,\ldots,\lambda_n)$ are eigenvalues. Then for any unit vector $\boldsymbol{v}$,

$$\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^\top(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top)\boldsymbol{v} = (\boldsymbol{v}^\top\boldsymbol{Q})\boldsymbol{\Lambda}(\boldsymbol{Q}^\top\boldsymbol{v}) = \boldsymbol{y}^\top\boldsymbol{\Lambda}\boldsymbol{y}$$

where $\boldsymbol{y}=\boldsymbol{Q}^\top\boldsymbol{v}$ is also a unit vector:

$$\|\boldsymbol{y}\|^2 = \boldsymbol{y}^T\boldsymbol{y} = \left(\boldsymbol{Q}^T\boldsymbol{v}\right)^T\left(\boldsymbol{Q}^T\boldsymbol{v}\right) = \boldsymbol{v}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{v} = 1$$

So the original optimization problem becomes the following one:

$$\max_{\boldsymbol{y}\in\mathbb{R}^n:\|\boldsymbol{y}\|=1} \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y}$$

To solve this new problem, write $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. It follows that

$$\boldsymbol{y}^T \boldsymbol{\Lambda} \boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2 \quad \text{(subject to } y_1^2 + y_2^2 + \cdots + y_n^2 = 1 \text{)}$$

Because $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, when $y_1^2 = 1, y_2^2 = \cdots = y_n^2 = 0$ (i.e., $\boldsymbol{y} = \pm \boldsymbol{e}_1$), the objective function attains its maximum value $\boldsymbol{y}^T \boldsymbol{\Lambda} \boldsymbol{y} = \lambda_1$.

In terms of the original variable $\boldsymbol{v}$, the maximizer is

$$\boldsymbol{v}_{\max} = \boldsymbol{Q} \boldsymbol{y}_{\max} = \boldsymbol{Q} (\pm \boldsymbol{e}_1) = \pm \boldsymbol{q}_1.$$

The minimum is the same procedure, resulting in $\boldsymbol{v}_{\min} = \pm \boldsymbol{q}_n$ ■

Problem 4. Graph Laplacian

1. Prove that all the eigenvalues of the graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ are non-negative.

Solution.

$$\begin{aligned}
\boldsymbol{z}^\top \boldsymbol{L} \boldsymbol{z} &= \boldsymbol{z}^\top (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{z} \\
&= \sum_{n=1}^{N} z_n d_n z_n - \sum_{n=1}^{N} \sum_{m=1}^{N} z_n W_{nm} z_m \\
&= \frac{1}{2} \sum_{n=1}^{N} z_n^2 \sum_{m=1}^{N} W_{nm} + \frac{1}{2} \sum_{m=1}^{N} z_m^2 \sum_{n=1}^{N} W_{nm} - \sum_{n=1}^{N} \sum_{m=1}^{N} z_n z_m W_{nm} \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (z_n^2 W_{nm} + z_m^2 W_{nm} - 2 z_n z_m W_{nm}) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} |z_n - z_m|^2 W_{nm} \geq 0
\end{aligned}$$

Therefore, $\boldsymbol{L}$ is positive semidefinite and therefore its eigenvalues are nonnegative. ■

2. Prove that all the eigenvalues of the normalized graph Laplacian $\boldsymbol{L}_{\text{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$ are in $[0, 2]$.

Solution. We could see that $\boldsymbol{L}_{\text{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2} (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2}$.

First, we show that 0 is an eigenvalue of $\boldsymbol{L}_{\text{sym}}$ using $\boldsymbol{x} = \boldsymbol{D}^{1/2} \boldsymbol{e}$,

$$\boldsymbol{L}_{\text{sym}} \boldsymbol{D}^{1/2} \boldsymbol{e} = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2} \boldsymbol{D}^{1/2} \boldsymbol{e} = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{e} = 0$$

since $\boldsymbol{D}\boldsymbol{e} - \boldsymbol{W}\boldsymbol{e} = 0$. Therefore, $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{L}_{\text{sym}}$ with eigenvalue 0. To show that it is the smallest, note that $\boldsymbol{L}_{\text{sym}}$ is also positive semidefinite,

$$\boldsymbol{z}^\top \boldsymbol{L}_{\text{sym}} \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2} \boldsymbol{z} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{|z_n - z_m|^2 W_{nm}}{\sqrt{d_n d_m}} \geq 0$$

Thus, the eigenvalues are non-negative and 0 is the smallest eigenvalue.

Similarly, we can show that $\boldsymbol{I} + \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$ is also positive semidefinite.

$$\boldsymbol{z}^\top (\boldsymbol{I} + \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}) \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{D}^{-1/2} (\boldsymbol{D} + \boldsymbol{W}) \boldsymbol{D}^{-1/2} \boldsymbol{z} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{|z_n + z_m|^2 W_{nm}}{\sqrt{d_n d_m}} \geq 0$$

Therefore, $\boldsymbol{z}^\top(\boldsymbol{I} + \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2})\boldsymbol{z} \geq 0$ and we have

$$-\boldsymbol{z}^\top \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{z} \leq \boldsymbol{z}^\top\boldsymbol{z} \Rightarrow \boldsymbol{z}^\top\boldsymbol{I}\boldsymbol{z} - \boldsymbol{z}^\top \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{z} \leq 2\boldsymbol{z}^\top\boldsymbol{z} \Rightarrow \frac{\boldsymbol{z}^\top \boldsymbol{L}_{\text{sym}}\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}} \leq 2$$

By Rayleigh quotient, $\lambda_{\max} \leq 2$. ■

Problem 5. One-class SVM

The optimization problem for one-class SVM is

$$\begin{aligned} \min \quad & R^2 + C\textstyle\sum_{n=1}^N \xi_n \\ \text{s.t.} \quad & \|\phi(\boldsymbol{x}_n) - \boldsymbol{a}\|^2 \leq R^2 + \xi_n \text{ for all } n \\ & \xi_n \geq 0 \text{ for all } n \end{aligned}$$

Write the Lagrangian and express it using only the Lagrange multipliers and the kernel $K(\boldsymbol{x}_n, \boldsymbol{x}_m) = \phi(\boldsymbol{x}_n)^\top \phi(\boldsymbol{x}_m)$.

Solution. The Lagrangian is,

$$L(R, a, \alpha_n, \xi_n) = R^2 + C\sum_{n=1}^N \xi_n - \sum_{n=1}^N \gamma_n\xi_n - \sum_{n=1}^N \alpha_n\left(R^2 + \xi_n - (\phi(\boldsymbol{x}_n) - \boldsymbol{a})^\top(\phi(\boldsymbol{x}_n) - \boldsymbol{a})\right)$$

with Lagrange multipliers $\alpha_i, \gamma_i \geq 0$. Then, we take the derivative with respect to the primal variables $\boldsymbol{a}$, ξ_i and R,

$$\begin{aligned} \frac{\partial L(R, a, \alpha_n, \xi_n)}{\partial \boldsymbol{a}} &= 2\sum_{n=1}^N \alpha_n(\boldsymbol{a} - \phi(\boldsymbol{x}_n)) \\ \frac{\partial L(R, a, \alpha_n, \xi_n)}{\partial \xi_n} &= C - \gamma_n - \alpha_n \\ \frac{\partial L(R, a, \alpha_n, \xi_n)}{\partial R} &= 2R - 2R\sum_{n=1}^N \alpha_n \end{aligned}$$

Set them to zero and we get $\boldsymbol{a} = \sum_{n=1}^N \alpha_n\phi(\boldsymbol{x}_n)$, $\gamma_n = C - \alpha_n$, $0 \leq \alpha_n \leq C$, and $\sum_{n=1}^N \alpha_n = 1$. Substituting them into the Lagrangian we obtain the following dual problem where we maximize with respect to α_i:

$$\begin{aligned} L(R, a, \alpha_n, \xi_n) &= R^2 + C\sum_{n=1}^N \xi_n - \sum_{n=1}^N \gamma_n\xi_n - \sum_{n=1}^N \alpha_n\left(R^2 + \xi_n - (\phi(\boldsymbol{x}_n) - \boldsymbol{a})^\top(\phi(\boldsymbol{x}_n) - \boldsymbol{a})\right) \\ &= R^2 + C\sum_{n=1}^N \xi_n - \sum_{n=1}^N (C - \alpha_n)\xi_n - R^2\sum_{n=1}^N \alpha_n - \sum_{n=1}^N \alpha_n\xi_n + \sum_{n=1}^N \alpha_n(\phi(\boldsymbol{x}_n) - \boldsymbol{a})^\top(\phi(\boldsymbol{x}_n) - \boldsymbol{a}) \\ &= \sum_{n=1}^N \alpha_n(\phi(\boldsymbol{x}_n) - \sum_{m=1}^N \alpha_m\phi(\boldsymbol{x}_m))^\top(\phi(\boldsymbol{x}_n) - \sum_{m=1}^N \alpha_m\phi(\boldsymbol{x}_m)) \\ &= \sum_{n=1}^N \alpha_n(\phi(\boldsymbol{x}_n)^\top\phi(\boldsymbol{x}_n)) - \sum_{n=1}^N\sum_{m=1}^N \alpha_n\alpha_m(\phi(\boldsymbol{x}_n)^\top\phi(\boldsymbol{x}_m)) \\ &= \sum_{n=1}^N \alpha_n K(\boldsymbol{x}_n, \boldsymbol{x}_n) - \sum_{n=1}^N\sum_{m=1}^N \alpha_n\alpha_m K(\boldsymbol{x}_n, \boldsymbol{x}_m) \end{aligned}$$

with constrains $0 \leq \alpha_n \leq C$, $\sum_{n=1}^N \alpha_n = 1$. ■

Problem 6. RKHS cf. [HaTF] Ex.5.16

Recall that $K(x, y) = \sum_{j=1}^\infty \gamma_j\phi_j(x)\phi_j(y)$ for which we can order $\gamma_1 \geq \gamma_2 \geq \cdots$ and $\{\phi_j\}_{j=1}^\infty$ is orthonormal:

$\langle \phi_i, \phi_j \rangle = \delta_{ij}$. Consider the ridge regression problem

$$\min_{\{c_j\}_{j=1}^{\infty}} \sum_{n=1}^{N} \left(y_n - \sum_{j=1}^{\infty} c_j \phi_j\left(x_n\right) \right)^2 + \lambda \sum_{j=1}^{\infty} \frac{c_j^2}{\gamma_j},$$

1. Explain why the problem is equivalent to

$$\min_{\boldsymbol{\alpha}} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha})^\top (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}.$$

Solution. In this setting, we have $f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$, $\|f\|_{\mathcal{H}_K}^2 = \sum_{i=1}^{\infty} c_i^2/\gamma_i$.

The solution have the form $f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$. By HW3, we have $\|f\|_{\mathcal{H}_K}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} K(x_i, x_j)\alpha_i\alpha_j$. Substitute them into the problem yield the results. ■

2. Assume $K(x, y) = \sum_{m=1}^{M} h_m(x)h_m(y)$ and $M \geq N$. Prove:

$$\boldsymbol{h}(x) = \boldsymbol{V}\boldsymbol{D}_\gamma^{1/2}\boldsymbol{\phi}(x)$$

where $\boldsymbol{h}(x) = [h_1(x), \cdots, h_M(x)]^\top$ and $\boldsymbol{\phi}(x) = [\phi_1(x), \cdots, \phi_M(x)]^\top$; $\boldsymbol{V}$ is an $M \times M$ orthogonal matrix and $\boldsymbol{D}_\gamma = \operatorname{diag}(\gamma_1, \cdots, \gamma_M)$. What are $\boldsymbol{V}$ and $\boldsymbol{D}_\gamma$? (Hint: $h_m = \sum_{j=1}^{M} \langle h_m, \phi_j \rangle \phi_j$).

Solution. From the definition of the kernel, we have

$$K(x, y) = \sum_{m=1}^{M} h_m(x)h_m(y) = \sum_{j=1}^{\infty} \gamma_j \phi_j(x)\phi_j(y)$$

Multiply both side by $\phi_k(x)$ yields,

$$\sum_{m=1}^{M} \langle h_m(x), \phi_k(x) \rangle h_m(y) = \sum_{j=1}^{\infty} \gamma_j \langle \phi_j(x), \phi_k(x) \rangle \phi_j(y) = \sum_{j=1}^{\infty} \gamma_j \delta_{jk} \phi_j(y) = \gamma_k \phi_j(y)$$

Let $g_{km} = \langle h_m(x), \phi_k(x) \rangle$ and multiply both side by $\phi_l(y)$ yields,

$$\sum_{m=1}^{M} g_{km} \langle h_m(y), \phi_l(y) \rangle = \gamma_k \langle \phi_j(y), \phi_l(y) \rangle \quad \Rightarrow \quad \sum_{m=1}^{M} g_{km} g_{lm} = \gamma_k \delta_{kl}$$

Let $\boldsymbol{G}_M = \{g_{nm}\} \in \mathbb{R}^{M \times N}$, we have

$$\boldsymbol{G}_M \boldsymbol{G}_M^\top = \operatorname{diag}\{\gamma_1, \gamma_2, \ldots, \gamma_M\} = \boldsymbol{D}_\gamma$$

Let $\boldsymbol{V}^\top = \boldsymbol{D}_\gamma^{-\frac{1}{2}} \boldsymbol{G}_M$, then

$$\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{G}_M \boldsymbol{D}_\gamma^{-1} \boldsymbol{G}_M^\top = \boldsymbol{I}$$

Then, we have

$$\sum_{m=1}^{M} g_{km} h_m(y) = \gamma_k \phi_j(y) \Rightarrow \boldsymbol{G}_M h(x) = \boldsymbol{D}_\gamma \boldsymbol{\phi}(x) \Rightarrow \boldsymbol{V}\boldsymbol{D}_\gamma^{-\frac{1}{2}} \boldsymbol{G}_M h(x) = \boldsymbol{V}\boldsymbol{D}_\gamma^{-\frac{1}{2}} \boldsymbol{D}_\gamma \boldsymbol{\phi}(x) \Rightarrow h(x) = \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}} \boldsymbol{\phi}(x)$$

■

Problem 1. Sample from Cauchy [Bi] Ex.11.3 Given a random variable z uniformly distributed over $(0,1)$, find a transformation $y = f(\mathrm{z})$ such that y has Cauchy distribution

$$p_{\mathrm{y}}(y) = \frac{1}{\pi}\frac{1}{1+y^2}.$$

Solution. Note that

$$\int \frac{1}{a^2+u^2}\,\mathrm{d}u = \frac{1}{a}\tan^{-1}\left(\frac{u}{a}\right) + C$$

We need

$$z = h(y) = \int_{-\infty}^{y} p_{\mathrm{y}}(y)\,\mathrm{d}y = \frac{1}{\pi}\tan^{-1}(y) + \frac{1}{2}$$

Therefore

$$y = h^{-1}(z) = \tan\left(\pi\left(z - \frac{1}{2}\right)\right)$$

■

Problem 2. Box-Muller [Bi] Ex.11.4 Suppose z_1 and z_2 are uniformly distributed over the unit circle (disk). Show that

$$y_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}, \quad y_2 = z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$$

where $r = z_1^2 + z_2^2$, has the joint density

$$p_{(y_1,y_2)}(y_1,y_2) = \left[\frac{1}{\sqrt{2\pi}}\exp\left(-y_1^2/2\right)\right]\left[\frac{1}{\sqrt{2\pi}}\exp\left(-y_2^2/2\right)\right]$$

Solution. We know that

$$p(y_1,y_2) = p(z_1,z_2)\left|\frac{\partial(z_1,z_2)}{\partial(y_1,y_2)}\right|$$

To find the Jacobian, we use the polar coordinate as intermediate and apply chain role. We define polar coordinate as

$$\begin{aligned}\theta &= \tan^{-1}\frac{z_2}{z_1} \\ r^2 &= z_1^2 + z_2^2\end{aligned} \quad \text{and,} \quad \begin{aligned} z_1 &= r\cos\theta \\ z_2 &= r\sin\theta\end{aligned}$$

Using the polar coordinate, we have

$$\frac{\partial(z_1,z_2)}{\partial(r,\theta)} = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix} \quad \left|\frac{\partial(z_1,z_2)}{\partial(r,\theta)}\right| = r\left(\cos^2\theta + \sin^2\theta\right) = r.$$

We can represent y as

$$y_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\cos\theta \tag{1}$$

$$y_2 = z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2} = \left(-2\ln r^2\right)^{1/2}\sin\theta \tag{2}$$

and thus

$$\frac{\partial(y_1,y_2)}{\partial(r,\theta)} = \begin{pmatrix} -2\cos\theta\left(-2\ln r^2\right)^{-1/2}r^{-1} & -2\sin\theta\left(-2\ln r^2\right)^{-1/2}r^{-1} \\ -\sin\theta\left(-2\ln r^2\right)^{1/2} & \cos\theta\left(-2\ln r^2\right)^{1/2}\end{pmatrix}$$

$$\left|\frac{\partial(r,\theta)}{\partial(y_1,y_2)}\right| = \left|\frac{\partial(y_1,y_2)}{\partial(r,\theta)}\right|^{-1} = \left(-2r^{-1}\left(\cos^2\theta + \sin^2\theta\right)\right)^{-1} = -\frac{r}{2}.$$

Applying the chain role, we have

$$\left|\frac{\partial\left(z_1, z_2\right)}{\partial\left(y_1, y_2\right)}\right|=\left|\frac{\partial\left(z_1, z_2\right)}{\partial(r, \theta)} \frac{\partial(r, \theta)}{\partial\left(y_1, y_2\right)}\right|=\left|\frac{\partial\left(z_1, z_2\right)}{\partial(r, \theta)}\right|\left|\frac{\partial(r, \theta)}{\partial\left(y_1, y_2\right)}\right|=-\frac{r^2}{2}$$

We will only use the absolute value of this.

By squaring both side of (1) and (2) and adding them together, we have

$$y_1^2+y_2^2=-2 \ln r^2 \quad \Rightarrow \quad r^2=\exp \left(-\frac{y_1^2+y_2^2}{2}\right)$$

Since $\left(z_1, z_2\right)$ is uniform, we have $p\left(z_1, z_2\right)=\frac{1}{\pi}$. Finally,

$$p(y_1, y_2)=p(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right|=\frac{1}{\pi} \frac{r^2}{2}=\frac{1}{2 \pi} \exp \left(-\frac{y_1^2+y_2^2}{2}\right)$$

■

Problem 3. Gibbs sampling Consider the Gibbs sampler for a vector of parameters $\boldsymbol{x}=\left(x_1, \cdots, x_M\right)^{\top}$. Suppose at the s-th step $\boldsymbol{x}^{(s)}$ is sampled from the target distribution $p(\boldsymbol{x})$ and then $\boldsymbol{x}^{(s+1)}$ is generated using the Gibbs sampler. Show that the marginal probability $P\left(\boldsymbol{x}^{(s+1)} \in \mathbb{A}\right)$ equals the target distribution $\int_{\mathbb{A}} p(\boldsymbol{x}) d \boldsymbol{x}$.

Solution. We can write $\boldsymbol{x}^{(s+1)}$ as,

$$p(\boldsymbol{x}^{(s+1)})=p(x_i^{(s+1)} \mid \boldsymbol{x}_{-i}^{(s+1)}) p(\boldsymbol{x}_{-i}^{(s+1)})=p(x_i^{(s+1)} \mid \boldsymbol{x}_{-i}^{(s)}) p(\boldsymbol{x}_{-i}^{(s)})=\frac{p(x_i^{(s+1)} \mid \boldsymbol{x}_{-i}^{(s)})}{p(x_i^{(s)} \mid \boldsymbol{x}_{-i}^{(s)})} p(\boldsymbol{x}^{(s)})$$

Therefore, the marginal probability converges to the target distribution. ■

Problem 4. Entropy Recall that the entropy of a discrete random variable X is defined to be

$$H(X)=-\sum_{x \in \mathbb{X}} p(x) \log_2 p(x)$$

where $\mathbb{X}$ is the set of all possible values of X.

1. A fair coin is flipped until the first head occurs. Let X denote the number of flips required. Find the entropy $H(X)$ in bits.

   Solution. Since $X=n$ means that first $n-1$ flips are tail and last flip is head. Suppose the probability of a head is p. Then, we have

   $$P(X=n)=(1-p)^{n-1}(p)^{-1}$$

   Thus, the entropy is

   $$\begin{aligned} H(X) & =-\sum_{n=1}^{\infty}(1-p)^{n-1} p \log \left((1-p)^{n-1} p\right) \\ & =-\left[\sum_{n=1}^{\infty}(1-p)^{n-1} p \log p+\sum_{n=1}^{\infty}(1-p)^{n-1} p \log (1-p)^{n-1}\right] \\ & =-\left[\sum_{m=0}^{\infty}(1-p)^m p \log p+\sum_{m=0}^{\infty} m(1-p)^m p \log (1-p)\right] \\ & =\frac{-p \log 0}{1-(1-p)}-\frac{p(1-p) \log (1-p)}{p^2} \\ & =\frac{-p \log p-(1-p) \log (1-p)}{p} \end{aligned}$$

Here we have a fair coin, $p = 1/2$, and the entropy is $H(X) = 2$ bits. ■

2. What is the relationship of $H(X)$ and $H(Y)$ if $Y = 2^X$?

Solution. Suppose $y = f(x)$, then

$$p(y) = \sum_{x:y=f(x)} p(x)$$
$$p(x) \leq \sum_{x:y=f(x)} p(x) = p(y)$$

Thus,

$$\sum_{x:y=f(x)} p(x)\log p(x) \leq p(y)\log p(y)$$

Then, we have

$$\begin{aligned} H(X) &= -\sum_{x\in X} p(x)\log p(x) \\ &= -\sum_{y\in Y}\sum_{x:y=f(x)} p(x)\log p(x) \\ &\geq -\sum_{y\in Y} p(y)\log p(y) = H(Y) \end{aligned}$$

It is equal if and only if $f(x)$ is one to one.

Since $Y = 2^X$ is one to one, $H(Y) = H(X)$ ■

Problem 5. Differential entropy Calculate the (differential) entropy of the following.

1. The exponential density $p(x) = \lambda e^{-\lambda x}, x \geq 0$.

Solution.

$$\begin{aligned} h(x) &= -\lambda\int_0^\infty e^{-\lambda x}\ln\left(\lambda e^{-\lambda x}\right)\mathrm{d}x \\ &= \lambda\int_0^\infty e^{-\lambda x}\ln\left(\frac{1}{\lambda}e^{\lambda x}\right)\mathrm{d}x \\ &= -\ln(\lambda)\lambda\int_0^\infty e^{-\lambda x}\mathrm{d}x + \lambda^2\int_0^\infty xe^{-\lambda x}\mathrm{d}x \\ &= -\ln(\lambda)\lambda\left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty + \lambda^2\left(\left[-\frac{1}{\lambda}xe^{-\lambda x}\right]_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\mathrm{d}x\right) \\ &= -\ln(\lambda) + \lambda\left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty \\ &= 1 - \ln(\lambda) \end{aligned}$$

■

2. The sum of x_1 and x_2 where x_1 is independent from x_2 and $p_{x_i}(x) = \mathcal{N}\left(x \mid \mu_i, \sigma_i^2\right)$ for $i = 1, 2$.

Solution. We know that the sum of two independent Gaussian distributions is still a Gaussian distribution (wiki). Let $y = x_1 + x_2$, we have

$$p_{x_1}(x) = \mathcal{N}\left(x \mid \mu_1, \sigma_1^2\right), \quad p_{x_2}(x) = \mathcal{N}\left(x \mid \mu_2, \sigma_2^2\right), \quad p_y(y) = \mathcal{N}\left(y \mid \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$$

Then, for the differential entropy of Gaussian distribution, we have

$$\begin{aligned}
h(X) &= -\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)\mathrm{d}x \\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\ln\left(\sigma\sqrt{2\pi}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\right)\mathrm{d}x \\
&= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-t^2\right)\ln\left(\sigma\sqrt{2\pi}\exp\left(t^2\right)\right)\mathrm{d}t \qquad \left(\text{substituting } t=\frac{x-\mu}{\sqrt{2}\sigma}\right) \\
&= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}\left(\ln(\sigma\sqrt{2\pi})+\ln\left(\exp\left(t^2\right)\right)\right)\exp\left(-t^2\right)\mathrm{d}t \\
&= \frac{\ln(\sigma\sqrt{2\pi})}{\sqrt{\pi}}\int_{-\infty}^{\infty}\exp\left(-t^2\right)\mathrm{d}t+\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty}t^2\exp\left(-t^2\right)\mathrm{d}t \\
&= \frac{\sqrt{\pi}\ln(\sigma\sqrt{2\pi})}{\sqrt{\pi}}+\frac{1}{\sqrt{\pi}}\left(\left[-\frac{t}{2}\exp\left(-t^2\right)\right]_{-\infty}^{\infty}+\frac{1}{2}\int_{-\infty}^{\infty}\exp\left(-t^2\right)\mathrm{d}t\right) \\
&= \ln(\sigma\sqrt{2\pi})+\frac{1}{2\sqrt{\pi}}\int_{-\infty}^{\infty}\exp\left(-t^2\right)\mathrm{d}t \\
&= \ln(\sigma\sqrt{2\pi})+\frac{\sqrt{\pi}}{2\sqrt{\pi}} \\
&= \ln(\sigma\sqrt{2\pi})+\frac{1}{2}
\end{aligned}$$

Therefore, $h(Y)=\ln\left(\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}\right)+\frac{1}{2}$ ■

Problem 6. Change of variable

Recall that $H(\boldsymbol{x})=-\int p_{\boldsymbol{x}}(\boldsymbol{x})\ln p_{\boldsymbol{x}}(\boldsymbol{x})\ \mathrm{d}\boldsymbol{x}$. Prove:

$$H(\boldsymbol{A}\boldsymbol{x})=\ln|\det(\boldsymbol{A})|+H(\boldsymbol{x}).$$

Solution. Recall that when we make a change of variables, the probability density is transformed by the Jacobian of the change of variables.

$$p(\boldsymbol{x})=p(\boldsymbol{y})\left|\frac{\partial y_i}{\partial x_j}\right|=p(\boldsymbol{y})\det\boldsymbol{A}$$

Then the entropy of $\boldsymbol{y}$ is

$$H(\boldsymbol{y})=-\int p(\boldsymbol{y})\ln p(\boldsymbol{y})\ \mathrm{d}\boldsymbol{y}=-\int p(\boldsymbol{x})\ln\left(p(\boldsymbol{x})\det(\boldsymbol{A})^{-1}\right)\mathrm{d}\boldsymbol{x}=H(\boldsymbol{x})+\ln|\det(\boldsymbol{A})|$$

■

Problem 1. (90pt)

1. Suppose we want to estimate the mean μ and the variance σ^2 of a Gaussian density $\mathcal{N}\left(\mu, \sigma^2\right)$.

   (a) (20pt) Is the maximum-likelihood estimator for μ biased or unbiased? You don't need to provide your reason.

   Solution. Unbiased. ■

   (b) (20 pt) Is the maximum-likelihood estimator for σ^2 biased or unbiased? You don't need to provide your reason.

   Solution. Biased. ■

2. (20pt) Suppose f and g are two different estimators for the same parameter $\boldsymbol{\theta}$ in some density $p(\boldsymbol{x} \mid \boldsymbol{\theta})$. Is it possible that the bias of f is smaller than the bias of g, while at the same time the variance of f is also smaller than the variance of g ? If yes, give an example of such f and g; otherwise, briefly state the reason.

Solution. Yes. Suppose we are estimating μ in 1D Gaussian $\mathcal{N}(\mu, 1)$. Given a sample $\mathcal{X} = \{x_n\}_{n=1}^N$. Let f be the sample mean, $\sum_{n=1}^N x_n/N$, and $g = 2(f+1)$. ■

3. Let X be a random variable. $X \sim \text{Unif}(\theta)$ with the density

$$p(x) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

   (a) (20pt) Consider the sample $\mathcal{X} = \{x_n\}_{n=1}^N$ where $x_n \sim^{i.i.d}$. Unif (θ). For the parameter θ above, write the likelihood $l(\theta \mid \mathcal{X})$ and the log-likelihood $\mathcal{L}(\theta \mid \mathcal{X})$.

   Solution. Suppose $I(\cdot)$ is the indicator function. The likelihood function is,

$$l(\theta \mid \mathcal{X}) = \prod_{n=1}^N p(x_n \mid \theta) = \frac{1}{\theta^N} I\left(\{x_n\}_{n=1}^N \in [0, \theta]\right) = \frac{1}{\theta^N} I\left(\max\{x_n\}_{n=1}^N \leq \theta\right)$$

   By taking the logarithm of the likelihood,

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log\left(\frac{1}{\theta^N} I\left(\max\{x_n\}_{n=1}^N \leq \theta\right)\right) = -N\log(\theta) + \log\left(I\left(\max\{x_n\}_{n=1}^N \leq \theta\right)\right)$$

   ■

   (b) (10 pt) Find the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$.

   Solution. When $\theta < \max\{x_n\}_{n=1}^N$, $l(\theta \mid \mathcal{X}) = 0$. When $\theta \geq \max\{x_n\}_{n=1}^N$, $l(\theta \mid \mathcal{X}) = \frac{1}{\theta^N}$.
   Since $\frac{1}{\theta^N}$ is monotonically decreasing, the maximum likelihood estimator is $\hat{\theta}_{\text{MLE}} = \max\{x_n\}_{n=1}^N$. ■

Problem 2. (110pt)

1. (30pt) Consider applying K-means with $K = 2$ clusters to the four data points

$$\boldsymbol{x}_1 = (0,0)^{\mathrm{T}}, \boldsymbol{x}_2 = (2,0)^{\mathrm{T}}, \boldsymbol{x}_3 = (0,3)^{\mathrm{T}}, \boldsymbol{x}_4 = (2,3)^{\mathrm{T}}$$

in $\mathbb{R}^2$. Suppose the initial centers are set to be $\boldsymbol{\mu}_1 = (0,0)^{\mathrm{T}}$ and $\boldsymbol{\mu}_2 = (4,3)^{\mathrm{T}}$. Write the E-step and the M-step for the first iteration. You need to clearly state the locations of the centers and the labels of the data points.

Solution.

E-step: Distances are

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$
$\boldsymbol{\mu}_1$	**0**	**2**	**3**	$\sqrt{13}$
$\boldsymbol{\mu}_2$	5	$\sqrt{13}$	4	**2**

We assign $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$ to $\boldsymbol{\mu}_1$, and $\boldsymbol{x}_4$ to $\boldsymbol{\mu}_2$.

M-step: $\boldsymbol{\mu}_1 = \frac{\boldsymbol{x}_1+\boldsymbol{x}_2+\boldsymbol{x}_3}{3} = (\frac{2}{3}, 1)^{\mathrm{T}}$, $\boldsymbol{\mu}_2 = \boldsymbol{x}_4 = (2,3)^{\mathrm{T}}$ ■

2. Consider same data points as in Part 1. Denote $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{4}$. Suppose we fit a Gaussian Mixture Model (GMM) to the data set using the general EM algorithm. Answer the following questions.

(a) (20pt) According to the general EM algorithm, is $\boldsymbol{X}$ the complete dataset or not? If not, what would make it complete?

Solution. No, it is not the complete. It needs latent variable z. ■

(b) (20pt) In the context of the general EM algorithm, briefly explain the relationship between "expectation" and "responsibility".

Solution. The relationship should be $\mathbb{E}[z_{nk}] = \gamma(z_{nk})$. ■

(c) (40pt) Suppose in the current iteration, the two Gaussians are given by $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{I})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{I})$, where $\mu_1 = (0,0)^{\mathrm{T}}$ and $\mu_2 = (4,3)^{\mathrm{T}}$. Moreover, $\pi_1 = \pi_2 = 0.5$. What are the values of the "responsibilities" for $\boldsymbol{x}_1$, the first data point? Your answer may contain exponential functions.

Solution. Let γ_1, γ_2 be the "responsibilities" two components, respectively.

$$\pi_1\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{I}) = 0.5\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{\mu}_1\|^2}{2}\right) = \frac{1}{2\sqrt{2\pi}}$$

$$\pi_2\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_2, \boldsymbol{I}) = 0.5\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{\mu}_2\|^2}{2}\right) = \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{25}{2}\right)$$

$$\gamma_1 = \frac{\pi_1\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{I})}{\pi_1\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{I}) + \pi_2\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_2, \boldsymbol{I})} = \frac{1}{1+\exp(-12.5)}$$

$$\gamma_2 = \frac{\pi_2\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_2, \boldsymbol{I})}{\pi_1\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{I}) + \pi_2\mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_2, \boldsymbol{I})} = \frac{\exp(-12.5)}{1+\exp(-12.5)}$$

■

Problem 3. (80pt) Suppose we are given a dataset $\mathcal{X} = \{\boldsymbol{x}_n, y_n\}_{n=1}^4$ where the inputs (independent variables) are

$$\boldsymbol{x}_1 = (0,0)^{\mathrm{T}}, \boldsymbol{x}_2 = (2,0)^{\mathrm{T}}, \boldsymbol{x}_3 = (0,2)^{\mathrm{T}}, \boldsymbol{x}_4 = (1,1)^{\mathrm{T}}$$

and the target values (dependent variables) are

$$y_1 = y_2 = y_3 = 0, y_4 = 1.$$

We want to perform a kernel ridge regression to this dataset.

1. (20pt) Explain briefly why ridge regression can be regarded as a Bayesian approach.

   Solution. Ridge regression is a MAP estimator with a special normal prior. ■

2. (60 pt) Suppose we choose the polynomial kernel K defined by

   $$K(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} + 1\right)^2.$$

   Perform the kernel ridge regression to predict the target value for the input $\boldsymbol{x} = (2,2)^{\mathrm{T}}$ with the regularization parameter $\lambda = 1$. Your answer may contain vector and matrix operations and you don't need to do the computation.

   Solution. The equation is $y = k(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{t}$.

   $$k(\boldsymbol{x}) = \begin{bmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}) \\ \vdots \\ K(\boldsymbol{x}_4, \boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} 1 \\ 25 \\ 25 \\ 25 \end{bmatrix} \quad \boldsymbol{K} = [K(\boldsymbol{x}_n, \boldsymbol{x}_m)]_{nm} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 25 & 1 & 9 \\ 1 & 1 & 25 & 9 \\ 1 & 9 & 9 & 9 \end{bmatrix} \quad \boldsymbol{t} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

   Thus, we have

   $$y = \begin{bmatrix} 1 & 25 & 25 & 25 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 26 & 1 & 9 \\ 1 & 1 & 26 & 9 \\ 1 & 9 & 9 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = 2.1512$$

   ■

Problem 4. (20pt) Recall, that given a log-likelihood function $\mathcal{L}(\boldsymbol{\theta} \mid \mathcal{X}) = \sum_{n=1}^N \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})$, the score function is defined to be

$$\mathcal{S}(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{\partial \mathcal{L}(\boldsymbol{\theta} \mid \mathcal{X})}{\partial \boldsymbol{\theta}}$$

Prove: $\mathbb{E}[\mathcal{S}(\boldsymbol{\theta} \mid \mathcal{X})] = 0$

Solution. Since $\boldsymbol{x}_n$ are independent, we have

$$\mathbb{E}[\mathcal{S}(\boldsymbol{\theta} \mid \mathcal{X})] = \mathbb{E}\left[\frac{\partial \mathcal{L}(\boldsymbol{\theta} \mid \mathcal{X})}{\partial \boldsymbol{\theta}}\right] = \mathbb{E}\left[\sum_{n=1}^N \frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \sum_{n=1}^N \mathbb{E}\left[\frac{\partial \log p(\boldsymbol{x}_n \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial \log p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right] &= \int p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right) \frac{\partial \log p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \, \mathrm{d}\boldsymbol{x} \\
&= \int p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right) \frac{1}{p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)} \frac{\partial p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \, \mathrm{d}\boldsymbol{x} \\
&= \int \frac{\partial p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \, \mathrm{d}\boldsymbol{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int p\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right) \, \mathrm{d}\boldsymbol{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} 1 = 0
\end{aligned}$$

Therefore, $\mathbb{E}[\mathcal{S}(\boldsymbol{\theta} \mid \mathcal{X})] = \sum_{n=1}^{N} 0 = 0$ ■

Problem 1. [S-S] Problem 3.1

Monotonicity of Sample Complexity: Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot, \cdot)$. Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in each of its parameters. That is, show that given $\delta \in (0,1)$, and given $0 < \epsilon_1 \leq \epsilon_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$. Similarly, show that given $\epsilon \in (0,1)$, and given $0 < \delta_1 \leq \delta_2 < 1$, we have that $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.

Solution. For fixed $\delta \in (0,1)$, suppose $0 < \epsilon_1 \leq \epsilon_2 \leq 1$. We need $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$. Given a training sequence of size $m \geq m_{\mathcal{H}}(\epsilon_1, \delta)$, we have that with probability at least $1-\delta$, we could learn a hypothesis h such that $L_{\mathcal{D},f}(h) \leq \epsilon_1 \leq \epsilon_2$. By the minimality of $m_{\mathcal{H}}(\epsilon_2, \delta)$, we get that $m_{\mathcal{H}}(\epsilon_2, \delta) \leq m_{\mathcal{H}}(\epsilon_1, \delta)$.

For fixed $\epsilon \in (0,1)$, suppose $0 < \delta_1 \leq \delta_2 < 1$. We need $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$. Given a training sequence of size $m \geq m_{\mathcal{H}}(\epsilon, \delta_1)$, we have that with probability at least $1 - \delta_1$, we could learn a hypothesis h such that $L_{\mathcal{D},f}(h) \leq \epsilon$. This also holds with $1 - \delta_2$. By the minimality of $m_{\mathcal{H}}(\epsilon_2, \delta)$, we get that $m_{\mathcal{H}}(\epsilon_2, \delta) \leq m_{\mathcal{H}}(\epsilon_1, \delta)$. ■

Problem 2. [S-S] Problem 3.2

Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{\text{Singleton}} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, where for each $z \in \mathcal{X}, h_z$ is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely, $\forall x \in X, h^-(x) = 0$. The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

1. Describe an algorithm that implements the ERM rule for learning $\mathcal{H}_{\text{Singleton}}$ in the realizable setup.

   Solution. Given a training set $\{(z_1, y_1), \ldots, (z_m, y_m)\}$, return h_{z_i} for the first (z_i, y_i) such that $y_i = 1$. If there is no such sample, return h^-. ■

2. Show that $\mathcal{H}_{\text{Singleton}}$ is PAC learnable. Provide an upper bound on the sample complexity.

   Solution. If $f = h^-$, our algorithm will clearly identify the correct hypothesis and have zero error. Hence, we may assume without loss of generality that $f = h_{z_0}$ for some $z_0 \in \mathcal{X}$. In this case, our algorithm will have nonzero error if and only if the training set S does not contain z_0, which will occur with probability

$$\mathcal{D}^m(\{S \mid z_0 \notin S\}) = (1 - \mathcal{D}(\{z_0\}))^m$$

in which case our algorithm will erroneously return h^- with error

$$L_{(\mathcal{D},f)}(h^-) = \mathcal{D}(\{z \mid h^-(z) \neq f(z)\}) = \mathcal{D}(\{z_0\})$$

Thus, assuming that $L_{(\mathcal{D},f)}(h^-) > \epsilon$ implies that $\mathcal{D}(\{z_0\}) > \epsilon$, and hence

$$(1 - \mathcal{D}(\{z_0\}))^m < (1 - \epsilon)^m \leq e^{-m\epsilon}$$

Thus, $\mathcal{H}_{\text{Singleton}}$ is PAC learnable, and to satisfy thebounds it suffices to find an m such that $e^{-m\epsilon} \leq \delta$, i.e.

$$m_{\mathcal{H}_{\text{Singleton}}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$$

■

Problem 3. [S-S] Problem 3.5

Let $\mathcal{X}$ be a domain and let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_m$ be a sequence of distributions over $\mathcal{X}$. Let $\mathcal{H}$ be a finite class of binary classifiers over $\mathcal{X}$ and let $f \in \mathcal{H}$. Suppose we are getting a sample S of m examples, such that the instances are independent but are not identically distributed; the i th instance is sampled from $\mathcal{D}_i$ and then y_i is set to be $f\left(\mathbf{x}_i\right)$. Let $\bar{D}_m$ denote the average, that is, $\overline{\mathcal{D}}_m = \left(\mathcal{D}_1 + \cdots + \mathcal{D}_m\right)/m$.

Fix an accuracy parameter $\epsilon \in (0,1)$. Show that

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } L_{\left(\overline{\mathcal{D}}_m, f\right)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right] \leq |\mathcal{H}| e^{-\epsilon m}.$$

Hint: Use the geometric-arithmetic mean inequality.

Solution. Fix some $h \in \mathcal{H}$ with $L_{\left(\overline{\mathcal{D}}_m, f\right)}(h) > \epsilon$. By definition,

$$\frac{\mathbb{P}_{X\sim\mathcal{D}_1}[h(X) = f(X)] + \ldots + \mathbb{P}_{X\sim\mathcal{D}_m}[h(X) = f(X))]}{m} < 1 - \epsilon.$$

We now bound the probability that h is consistent with S (i.e., that $L_S(h) = 0$) as follows:

$$\begin{aligned}
\underset{S\sim\prod_{i=1}^m \mathcal{D}_i}{\mathbb{P}}\left[L_S(h) = 0\right] &= \prod_{i=1}^m \underset{X\sim\mathcal{D}_i}{\mathbb{P}}[h(X) = f(X)] \\
&= \left(\left(\prod_{i=1}^m \underset{X\sim\mathcal{D}_i}{\mathbb{P}}[h(X) = f(X)]\right)^{\frac{1}{m}}\right)^m \\
&\leq \left(\frac{\sum_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]}{m}\right)^m \\
&< (1-\epsilon)^m \\
&\leq e^{-\epsilon m}
\end{aligned}$$

The first inequality is the geometric-arithmetic mean inequality. Applying the union bound, we conclude that the probability that there exists some $h \in \mathcal{H}$ with $L_{\left(\overline{\mathcal{D}}_m, f\right)}(h) > \epsilon$, which is consistent with S is at most $|\mathcal{H}| \exp(-\epsilon m)$. ■

Problem 4. [S-S] Problem 3.6

Let $\mathcal{H}$ be a hypothesis class of binary classifiers. Show that if $\mathcal{H}$ is agnostic PAC learnable, then $\mathcal{H}$ is PAC learnable as well. Furthermore, if A is a successful agnostic PAC learner for $\mathcal{H}$, then A is also a successful PAC learner for $\mathcal{H}$.

Solution. Suppose that $\mathcal{H}$ is agnostic PAC learnable, and let A be a learning algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. We show that $\mathcal{H}$ is PAC learnable using A.

Let $\mathcal{D}, f$ be an (unknown) distribution over $\mathcal{X}$, and the target function respectively. We may assume w.l.o.g. that $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \{0,1\}$, where the conditional probability of y given x is determined deterministically by f. Since we assume realizability, we have $\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$. Let $\epsilon, \delta \in (0,1)$. Then, for every positive integer $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, if we equip A with a training set S consisting of m i.i.d. instances which are labeled by f, then with probability at least $1-\delta$ (over the choice of $\left.S\right|_x$), it returns a hypothesis h with

$$\begin{aligned}
L_{\mathcal{D}}(h) &\leq \inf_{h'\in\mathcal{H}} L_{\mathcal{D}}\left(h'\right) + \epsilon \\
&= 0 + \epsilon \\
&= \epsilon.
\end{aligned}$$

■

Problem 5. [S-S] Problem 6.1

Show the following monotonicity property of VC-dimension: For every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$ then $\operatorname{VCdim}(\mathcal{H}') \leq \operatorname{VCdim}(\mathcal{H})$.

Solution. Let $\mathcal{H}' \subseteq \mathcal{H}$ be two hypothesis classes for binary classification. Since $\mathcal{H}' \subseteq \mathcal{H}$, then for every $C = \{c_1, \ldots, c_m\} \subseteq \mathcal{X}$, we have $\mathcal{H}'_C \subseteq \mathcal{H}_C$. In particular, if C is shattered by $\mathcal{H}'$, then C is shattered by $\mathcal{H}$ as well. Thus, VCdim $(\mathcal{H}') \leq \operatorname{VCdim}(\mathcal{H})$. ■

Problem 6. [S-S] Problem 6.4

We proved Sauer's lemma by proving that for every class $\mathcal{H}$ of finite VCdimension d, and every subset A of the domain,

$$|\mathcal{H}_A| \leq | \{B \subseteq A : \mathcal{H} \text{ shatters } B\} | \leq \sum_{i=0}^{d} \left(\begin{array}{c} |A| \\ i \end{array} \right)$$

Show that there are cases in which the previous two inequalities are strict (namely, the $\leq$ can be replaced by $<$) and cases in which they can be replaced by equalities. Demonstrate all four combinations of $=$ and $<$.

Solution. Let $\mathcal{X} = \mathbb{R}^d$. We will demonstrate all the 4 combinations using hypothesis classes defined over $\mathcal{X} \times \{0,1\}$. Remember that the empty set is always considered to be shattered.

- $(<,=)$: Let $d \geq 2$ and consider the class $\mathcal{H} = \left\{\mathbb{1}_{[\|x\|_2 \leq r]} : r \geq 0\right\}$ of concentric balls. The VC-dimension of this class is 1 . To see this, we first observe that if $\mathbf{x} \neq (0, \ldots, 0)$, then $\{\mathbf{x}\}$ is shattered. Second, if $\|\mathbf{x}_1\|_2 \leq \|\mathbf{x}_2\|_2$, then the labeling $y_1 = 0, y_2 = 1$ is not obtained by any hypothesis in $\mathcal{H}$. Let $A = \{\mathbf{e}_1, \mathbf{e}_2\}$, where $\mathbf{e}_1, \mathbf{e}_2$ are the first two elements of the standard basis of $\mathbb{R}^d$. Then, $\mathcal{H}_A = \{(0,0),(1,1)\}, \{B \subseteq A : \mathcal{H} \text{ shatters } B\} = \{\emptyset, \{\mathbf{e}_1\}, \{\mathbf{e}_2\}\}$, and $\sum_{i=0}^{d} \left(\begin{array}{c} |A| \\ i \end{array} \right) = 3$.

- $(=,<)$: Let $\mathcal{H}$ be the class of axis-aligned rectangles in $\mathbb{R}^2$. We have seen that the VC-dimension of $\mathcal{H}$ is 4 . Let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = (0,0), \mathbf{x}_2 = (1,0), \mathbf{x}_3 = (2,0)$. All the labelings except $(1,0,1)$ are obtained. Thus, $|\mathcal{H}_A| = 7, | \{B \subseteq A : \mathcal{H} \text{ shatters } B\} | = 7$, and $\sum_{i=0}^{d} \left(\begin{array}{c} |A| \\ i \end{array} \right) = 8$.

- $(<,<)$: Let $d \geq 3$ and consider the class $\mathcal{H} = \{\operatorname{sign}\langle w, x\rangle : w \in \mathbb{R}^d\}^2$ of homogenous halfspaces (see Chapter 9). We will prove in Theorem 9.2 that the VC-dimension of this class is d. However, here we will only rely on the fact that $\operatorname{VCdim}(\mathcal{H}) \geq 3$. This fact follows by observing that the set $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is shattered. Let $A = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, where $\mathbf{x}_1 = \mathbf{e}_1, \mathbf{x}_2 = \mathbf{e}_2$, and $\mathbf{x}_3 = (1,1,0,\ldots,0)$. Note that all the labelings except $(1,1,-1)$ and $(-1,-1,1)$ are obtained. It follows that $|\mathcal{H}_A| = 6, | \{B \subseteq A : \mathcal{H} \text{ shatters } B\} | = 7$, and $\sum_{i=0}^{d} \left(\begin{array}{c} |A| \\ i \end{array} \right) = 8$.

- $(=,=)$: Let $d = 1$, and consider the class $\mathcal{H} = \left\{\mathbb{1}_{[x \geq t]} : t \in \mathbb{R}\right\}$ of thresholds on the line. We have seen that every singleton is shattered by $\mathcal{H}$, and that every set of size at least 2 is not shattered by $\mathcal{H}$. Choose any finite set $A \subseteq \mathbb{R}$. Then each of the three terms in "Sauer's inequality" equals $|A| + 1$. ■