Always provide justifications and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

You may find the following useful:

- Bernoulli distribution

$$p(x|\mu) = \mu^x (1-\mu)^{1-x}, \quad x \in \{0,1\}$$

- Binomial distribution

$$p(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Beta distribution

$$\mathrm{Beta}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- Gamma distribution

$$\mathrm{Gamma}(\tau|a,b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau)$$

- Gaussian distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)$$

- Wishart distribution

$$\mathcal{W}(\boldsymbol{\Sigma}|\boldsymbol{W},\nu) = B|\boldsymbol{\Sigma}|^{\frac{\nu-D-1}{2}} \exp\big(-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{W}^{-1}\boldsymbol{\Sigma})\big)$$

1 a Consider a binary random variable $y \in \{0,1\}$. If $y = 0$, it is assigned to a class $\mathcal{C}_0$, otherwise to a class $\mathcal{C}_1$ (there are only two classes). For an input $\boldsymbol{x} \in \mathbb{R}^D$, the distribution of the label y is given by

$$p(y=1|\boldsymbol{x}) = p(\mathcal{C}_1|\boldsymbol{x}), \quad p(y=0|\boldsymbol{x}) = 1 - p(y=1|\boldsymbol{x}).$$

We define the logistic sigmoid

$$\sigma(z) := \frac{1}{1+\exp(-z)}, \quad z := \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_0|\boldsymbol{x})} \tag{1}$$

i) Show that $\sigma(z) = p(\mathcal{C}_1|\boldsymbol{x})$

ii) Given a dataset $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ with $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0,1\}$, we use a Bernoulli likelihood

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \text{Ber}(y_n|\mu_n), \quad \mu_n := \sigma(z_n), \quad z_n := \boldsymbol{x}_n^\top \boldsymbol{\theta}$$

where $p(y_n = 1|\boldsymbol{x}_n) = \sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta}) = \mu_n$.

A) Assuming that the data is i.i.d., write down an expression for the likelihood $p(y_1, \ldots, y_N|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta})$.
Hint: Using the definition of μ_n will simplify the expression.

B) Compute the derivative of the negative log-likelihood with respect to the parameters $\boldsymbol{\theta}$.
Hint: The chain rule will be useful here.

C) Compared to linear regression, what problem do you encounter when you want to optimize $\boldsymbol{\theta}$? Suggest a solution to this problem, which you could use in practice.

b *In this part, you will receive +1 mark for a correct answer, 0 marks for no answer, and -1 mark for an incorrect answer. Your marks for this part are lower-bounded by 0.*

i) What is the dimension of the gradient $d\boldsymbol{y}/d\boldsymbol{X}$, where

$$\boldsymbol{y} = \exp(-|\boldsymbol{A}\boldsymbol{X}|)\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}^\top$$

with $\boldsymbol{A} \in \mathbb{R}^{D\times E}$ and $|\cdot|$ being the determinant?
Hint: You can determine the dimensions of $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{I}$, so that the functions are defined.

ii) True or False: MAP estimation in linear regression with a parameter prior $\mathcal{N}(0, \sigma^2)$ is equivalent to MLE for $\sigma \to 0$.

iii) True or False: In linear regression, maximum likelihood estimation always yields a unique solution.

iv) True or False: In linear regression withe a parameter prior $\mathcal{N}(\mathbf{0}, 10\boldsymbol{I})$, MAP estimation always yields a unique solution.

v) True or False: If a function has a local maximum at $\boldsymbol{x}_*$, the gradient at $\boldsymbol{x}_*$ must be zero and the Hessian at $\boldsymbol{x}_*$ must be positive definite.

vi) True or False: We can find different directed graphical models that give rise to the same joint distribution $p(a, b)$.

vii) True or False: It is possible that gradient descent does not find the global optimum, even if the function is convex, e.g., $f(x) = x^2$.

viii) Occam's razor states that we should always choose the model that explains the data best.

The two parts carry, respectively, 60% and 40% of the marks.

Always provide justifications and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

You may find the following useful:

- Bernoulli distribution

$$p(x|\mu) = \mu^x (1-\mu)^{1-x}, \quad x \in \{0, 1\}$$

- Binomial distribution

$$p(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Beta distribution

$$\mathrm{Beta}(\mu|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- Gamma distribution

$$\mathrm{Gamma}(\tau|a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau)$$

- Gaussian distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)$$

- Wishart distribution

$$\mathcal{W}(\boldsymbol{\Sigma}|\boldsymbol{W}, \nu) = B|\boldsymbol{\Sigma}|^{\frac{\nu-D-1}{2}} \exp\big(-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{W}^{-1}\boldsymbol{\Sigma})\big)$$

1 a Consider a binary random variable $y \in \{0, 1\}$. If $y = 0$, it is assigned to a class $\mathcal{C}_0$, otherwise to a class $\mathcal{C}_1$ (there are only two classes). For an input $\boldsymbol{x} \in \mathbb{R}^D$, the distribution of the label y is given by

$$p(y = 1|\boldsymbol{x}) = p(\mathcal{C}_1|\boldsymbol{x}), \quad p(y = 0|\boldsymbol{x}) = 1 - p(y = 1|\boldsymbol{x}).$$

We define the logistic sigmoid

$$\sigma(z) := \frac{1}{1 + \exp(-z)}, \quad z := \log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_0|\boldsymbol{x})} \tag{1}$$

i) Show that $\sigma(z) = p(\mathcal{C}_1|\boldsymbol{x})$

ii) Given a dataset $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ with $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0, 1\}$, we use a Bernoulli likelihood

$$p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \text{Ber}(y_n|\mu_n), \quad \mu_n := \sigma(z_n), \quad z_n := \boldsymbol{x}_n^\top \boldsymbol{\theta}$$

where $p(y_n = 1|\boldsymbol{x}_n) = \sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta}) = \mu_n$.

A) Assuming that the data is i.i.d., write down an expression for the likelihood $p(y_1, \ldots, y_N|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta})$.
Hint: Using the definition of μ_n will simplify the expression.

B) Compute the derivative of the negative log-likelihood with respect to the parameters $\boldsymbol{\theta}$.
Hint: The chain rule will be useful here.

C) Compared to linear regression, what problem do you encounter when you want to optimize $\boldsymbol{\theta}$? Suggest a solution to this problem, which you could use in practice.

---

i) ***[3 marks]***

$$\begin{aligned}
\sigma(z) &= \frac{1}{1 + \exp(-\log \frac{p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_0|\boldsymbol{x})})} \\
&= \frac{1}{1 + \exp(\log \frac{p(\mathcal{C}_0|\boldsymbol{x})}{p(\mathcal{C}_1|\boldsymbol{x})})} \\
&= \frac{1}{1 + \frac{p(\mathcal{C}_0|\boldsymbol{x})}{p(\mathcal{C}_1|\boldsymbol{x})}} \\
&= \frac{1}{\frac{p(\mathcal{C}_0|\boldsymbol{x}) + p(\mathcal{C}_1|\boldsymbol{x})}{p(\mathcal{C}_1|\boldsymbol{x})}} = p(\mathcal{C}_1|\boldsymbol{x})
\end{aligned}$$

since $p(\mathcal{C}_0|\boldsymbol{x}) + p(\mathcal{C}_1|\boldsymbol{x}) = 1.$

ii) A) ***[2 marks]****We use the definition of the Bernoulli distribution and arrive immediately at*

$$p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = \prod_{n=1}^{N} \mathrm{Ber}(y_n|\mu_n) = \prod_{n=1}^{N} \mu_n^{y_n}(1-\mu_n)^{1-y_n}$$

where $\mu_n = \sigma(\boldsymbol{x}_n^\top \boldsymbol{\theta})$

B) ***[5 marks]*** *The negative log-likelihood is given by*

$$\begin{aligned} NLL(\boldsymbol{\theta}) &= -\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \\ &= -\sum_{n=1}^{N} \left(y_n \log \mu_n + (1-y_n)\log(1-\mu_n)\right) \end{aligned}$$

We use the chain rule and get

$$\begin{aligned} \frac{dNLL}{d\boldsymbol{\theta}} &= \sum_n \frac{\partial NLL}{\partial \mu_n}\frac{\partial \mu_n}{\partial \boldsymbol{\theta}} \\ \frac{\partial NLL}{\partial \mu_n} &= y_n \frac{1}{\mu_n} - (1-y_n)\frac{1}{1-\mu_n} \\ \frac{\partial \mu_n}{\partial \boldsymbol{\theta}} &= \frac{\partial \sigma(z_n)}{\partial z_n}\frac{\partial z_n}{\partial \boldsymbol{\theta}} \end{aligned}$$

with $z_n = \boldsymbol{x}_n^\top \boldsymbol{\theta}$*. We get the partial derivatives*

$$\begin{aligned} \frac{\partial \sigma(z_n)}{\partial z_n} &= \frac{\exp(-z_n)}{(1+\exp(-z_n))^2} = \sigma(z_n)(1-\sigma(z_n)) \\ \frac{\partial z_n}{\partial \boldsymbol{\theta}} &= \boldsymbol{x}_n^\top \end{aligned}$$

Overall, we obtain

$$\frac{\partial NLL}{\partial \boldsymbol{\theta}} = \sum_n \left(\frac{y_n}{\mu_n} - \frac{(1-y_n)}{1-\mu_n}\right)\sigma(z_n)(1-\sigma(z_n))\boldsymbol{x}_n^\top$$

C) ***[2 marks]*** *A closed-form solution is not possible here. We can find the maximum likelihood estimator using gradient descent.*

Marks: **12**

b *In this part, you will receive +1 mark for a correct answer, 0 marks for no answer, and -1 mark for an incorrect answer. Your marks for this part are lower-bounded by 0.*

i) What is the dimension of the gradient $d\boldsymbol{y}/d\boldsymbol{X}$, where

$$\boldsymbol{y} = \exp(-|\boldsymbol{AX}|)\boldsymbol{I} + \boldsymbol{AA}^\top$$

with $\boldsymbol{A} \in \mathbb{R}^{D\times E}$ and $|\cdot|$ being the determinant?
Hint: You can determine the dimensions of $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{I}$, so that the functions are defined.

ii) True or False: MAP estimation in linear regression with a parameter prior $\mathcal{N}(0, \sigma^2)$ is equivalent to MLE for $\sigma \to 0$.

iii) True or False: In linear regression, maximum likelihood estimation always yields a unique solution.

iv) True or False: In linear regression withe a parameter prior $\mathcal{N}(\boldsymbol{0}, 10\boldsymbol{I})$, MAP estimation always yields a unique solution.

v) True or False: If a function has a local maximum at $\boldsymbol{x}_*$, the gradient at $\boldsymbol{x}_*$ must be zero and the Hessian at $\boldsymbol{x}_*$ must be positive definite.

vi) True or False: We can find different directed graphical models that give rise to the same joint distribution $p(a, b)$.

vii) True or False: It is possible that gradient descent does not find the global optimum, even if the function is convex, e.g., $f(x) = x^2$.

viii) Occam's razor states that we should always choose the model that explains the data best.

---

i) $(DxD)x(ExD)$ *or* $D^2 \times ED$

ii) *False*

iii) *False*

iv) *True*

v) *False*

vi) *True*

vii) *True*

viii) *False*

Marks: **8**

The two parts carry, respectively, 60% and 40% of the marks.