Q1 [5x2=10 pts] Use machine learning terminologies to explain the following phenomenon.

Scenario	Fill in the blank with "overfitting" or "underfitting"
A student can solve every written assignment question after checking reference solutions, but the student fails to solve final exam questions.	This is overfitting.
A student can solve neither written assignment questions nor final exam questions.	This is underfitting.

Note: No explanation is needed.

Q2 [3x5=15 pts]. Consider $\mathbf{y} = \text{softmax}(\mathbf{s})$ for $\mathbf{y}, \mathbf{s} \in \mathbb{R}^K$, where $y_i = \frac{\exp\{s_i\}}{\sum_j \exp\{s_j\}}$.
(a) Is it true that $\text{softmax}(\boldsymbol{s}) = \text{softmax}(c\boldsymbol{s})$ for any positive constant c? (b) What happens if $c > 1$? (c) What happens if $0 < c < 1$? **Note**: A few words would suffice. No proof is needed.

a) No b) distribution peaker c) distribution smoother

Q3 [15 pts]. Directly adding a lower layer is common in modern neural architectures. Let $\mathbf{y} \in \mathbb{R}^d$ be computed by $\mathbf{y} = \mathbf{x} + \tilde{\mathbf{y}}$, where $\tilde{\mathbf{y}} = f(\mathbf{W}\mathbf{x} + \mathbf{b})$. $f(\cdot)$ is a scalar non-linear activation function. When fed with a vector, $f(\cdot)$ is applied to every element of the vector.

a) $\boldsymbol{y}$ is a d-dimensional vector. Infer the dimensions of $\tilde{\mathbf{y}}, \mathbf{W}, \mathbf{b}$, and $\mathbf{x}$.

b) Suppose $\frac{\partial J}{\partial y_i}$ is known for $i = 1, \cdots, d$, where y_i is an element of $\mathbf{y}$. Derive $\frac{\partial J}{\partial x_j}$ for some given j. **Note:** The derivative of f evaluated at z, denoted by $f'(z)$, can be directly used in the solution. Give key derivation steps.

a) $\tilde{y}, b, x \in \mathbb{R}^d$

$W \in \mathbb{R}^{d \times d}$

b) We have

$$\frac{\partial J}{\partial x_j} = \frac{\partial J}{\partial y_j} + \sum_i \frac{\partial J}{\partial \tilde{y}_i} \cdot \frac{\partial \tilde{y}_i}{\partial x_j}$$

$$= \frac{\partial J}{\partial y_j} + \sum_i \frac{\partial J}{\partial y_i} \cdot f'(w_i^T x + b_i)\, W_{ij}$$

w_j is the jth row of W as a column vector

Q4 [30 pts]. Consider a dataset $\mathcal{D} = \{x^{(m)}\}_{m=1}^{M}$, where each $x^{(m)} \in \{0, 1\}$ is iid sampled from $\mathrm{Bernoulli}(\pi)$, meaning that $\Pr[x^{(m)} = 1] = \pi$, $\Pr[x^{(m)} = 0] = 1 - \pi$.

a) Give a formula of the probability of ~~a sample $x^{(m)}$~~ $\mathcal{D}$, where $x^{(m)} = 1$ and $x^{(m)} = 0$ cases are unified.

b) Consider a Beta prior $\pi \sim \mathrm{Beta}(\alpha, \beta)$, i.e., $p(\pi; \alpha, \beta) = C(\alpha, \beta)\pi^{\alpha}(1 - \pi)^{\beta}$, where $\alpha, \beta > 0$. $C(\alpha, \beta)$ is a function depending on α and β, but is a constant with respect to π. $C(\alpha, \beta)$ serves as a normalizing factor to make $p(\pi; \alpha, \beta)$ a valid distribution over π.
Compute the posterior distribution $p(\pi|\mathcal{D})$, where the normalizing factor need not be expressed explicitly. Show that the posterior also follows a Beta distribution $\mathrm{Beta}(\alpha', \beta')$. What are α' and β'?

c) Give the max a posterior estimation of π under the above prior.

a). $$p(x^{(m)} \mid \pi) = \prod_{m=1}^{M} \pi^{x^{(m)}} (1-\pi)^{1-x^{(m)}}$$

b) $$p(\pi \mid \mathcal{D}; \alpha, \beta) \propto p(\pi) \cdot p(\mathcal{D} \mid \pi)$$
$$= C(\alpha, \beta)\pi^{\alpha}(1-\pi)^{\beta} \prod_{m=1}^{M} \pi^{x^{(m)}} (1-\pi)^{1-x^{(m)}}$$
$$= C(\alpha, \beta) \cdot \pi^{\alpha + \sum_{m=1}^{M} x^{(m)}} (1-\pi)^{\beta + M - \sum_{m=1}^{M} x^{(m)}}$$

c) Noticing the MLE is to $\max_{\pi} \pi^{\sum_{m=1}^{M} x^{(m)}} (1-\pi)^{M - \sum_{m=1}^{M} x^{(m)}}$

and yields $$\hat{\pi}_{MLE} = \frac{\sum_{m=1}^{M} x^{(m)}}{M},$$

we have MAP to $\max_{\pi} \pi^{\alpha + \sum_{m=1}^{M} x^{(m)}} (1-\pi)^{\beta + M - \sum_{m=1}^{M} x^{(m)}}$

and yield $$\hat{\pi}_{MAP} = \frac{\alpha + \sum_{m=1}^{M} x^{(m)}}{\alpha + \beta + M}$$

Q5 [30 pts]. Consider any continuous function $f : \mathbb{R}^d \to \mathbb{R}$. We would like to iteratively update the input $\mathbf{x}_k \in \mathbb{R}^d$ along a direction $\mathbf{p}_k \in \mathbb{R}^d$ with step size $\alpha_k > 0$, i.e., $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$ for the kth iteration step. For notational purposes, we denote the gradient vector by $\mathbf{g}_k = \nabla_{\mathbf{x}} f(\mathbf{x}_k)$ and the Hessian matrix by $\mathbf{H}_k = \nabla^2_{\mathbf{x}} f(\mathbf{x}_k)$.

a) In the lecture, we showed that, if $\mathbf{p}_k = -\mathbf{g}_k \neq \mathbf{0}$, then there exists some small α_k such that $f(\mathbf{x}_{k+1}) < f(\boldsymbol{x}_k)$. In other words, $-\mathbf{g}_k$ is a descending direction if not zero.
Prove that any direction $\mathbf{p}_k$ satisfying $\mathbf{p}_k^\top \mathbf{g}_k < 0$ is a descending direction.

b) Suppose f is a quadratic function with the form of $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{Q}$ is a symmetric and positive-definite matrix. Show that Newton's direction $-\mathbf{H}_k^{-1}\mathbf{g}_k$ with step size $\alpha_k = 1$ will move $\mathbf{x}_k$ to the global optimum of f in one step. **Hint:** In this case, $\mathbf{H}_k = \mathbf{Q}$ and $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$.

c) Now consider a convex function f (not necessarily quadratic). Prove that Newton's direction $-\mathbf{H}_k^{-1}\mathbf{g}_k$ is a descending direction, i.e., with a small enough α_k, we will have $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$.

Cheatsheet: Taylor's theorem suggests that, for a small enough α_k,

$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) = f(\mathbf{x}_k) + \alpha_k \mathbf{p}_k^\top \mathbf{g}_k + O(\alpha_k^2)$, and

$$f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) = f(\mathbf{x}_k) + \alpha_k \mathbf{p}_k^\top \mathbf{g}_k + \frac{1}{2}\alpha_k^2 \mathbf{p}_k^\top \mathbf{H}_k \mathbf{p}_k + O(\alpha_k^3),$$

where $O(\cdot)$ is a higher-order small term.

a) $f(x_{k+1}) = f(x_k) + \alpha_k \underbrace{p_k^T g_k}_{<0} + O(\alpha_k^2)$

Thus, $f(x_{k+1}) < f(x_k)$

b) Newton's direction $p_k^{Newton} = -H_k^{-1} g_k = -Q^{-1}(Qx_k - b) = -x_k + Q^{-1}b$

$x_{k+1} = x_k + \alpha_k^{Newton} \cdot p_k^{Newton} = x_k - x_k + Q^{-1}b = Q^{-1}b.$

The global optimum of $f(x)$ can be computed by

$\nabla_x f(x) = Qx - b \overset{set}{=} 0 \Rightarrow x^* = Q^{-1}b$

Thus $x_{k+1} = x^*$

c). $(-H_k^{-1} g_k)^T g_k = -g_k^T H_k^{-1} g_k < 0$ satisfying a)

Scrap paper. May be detached. Additional paper is available upon request as appropriate.
May be used as answer sheets if you

- Print your name and ID on every answer sheet (including additional sheets) submitted (1 bonus mark)
- Mark your solution and corresponding problem number clearly
- Submit the sheet by the end of the exam