• This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.
• Start writing when instructed. Stop writing when your time is up.
• Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

(1) (7 Points) Consider the following depictions of bivariate Gaussian probability density functions (pdfs), each graphed via level sets / contours. For each set of bivariate Gaussian parameters below, identify which (if any) of the pdfs could correspond under some setting of the variables $a$, $b$, $c$, and $d$. (Select all that apply.)

(a) $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$

**Solution:** (iv). The variances of both variables are equal, with no correlation, yielding a spherical distribution.

(b) $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma =$ Identity Matrix

**Solution:** (iv). The variances of both variables are equal, with no correlation, yielding a spherical distribution.

(c) $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$

**Solution:** (ii), (iii), and (iv). The only restriction here is that the mean, $\mu$, is the origin. The covariance matrix is of the general form.

(d) $\mu = \begin{bmatrix} a \\ b \end{bmatrix}$ $\Sigma = \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix}$

**Solution:** (ii), (iv). The only restriction here is that there must be zero correlation.
(2) (2 points) Your friend needs help selecting between two linear regression models, desiring the one that generalizes best. Which model should they use? (Circle your answer and provide a short justification.)

**Model 1**
Training MSE: 100, Validation MSE: 105

**Model 2**
Training MSE: 50, Validation MSE: 110

**Solution:** Model 1 since even though it has worse training error, it yields slightly better validation error. Since validation error is a better predictor of future performance, this model is more likely to generalize best.

(3) (2 points) Consider the two datasets shown below: D1 and D2. When fitting a linear regression model to predict the y-axis value as output, given the x-axis value as input, which of these two training sets will have the lowest training mean squared error (MSE)? (Circle your answer. No need to provide justification.)

**Solution:** D1 since the line of best fit lies closer to the training points in this dataset – thus, the squared error of the best predictor will be smaller.

(4) (3 points) Recall, the original training loss used for standard least squares regression:

$$\text{Loss}(w) = \sum_{i=1}^{n} \left(y^{(i)} - w^T x^{(i)}\right)^2$$

The individual datapoints are indexed by $i$, so that $x^{(i)}$ is the input and $y^{(i)}$ is the output for the $i$th datapoint in your training set of $n$ datapoints. Suppose $x^{(i)} \in \mathbb{R}^d$ and $y \in \mathbb{R}$.

Suppose that your friend has proposed a new training loss to use for learning a regression model:

$$\text{NewLoss}(w) = \log \sum_{i=1}^{n} \left(y^{(i)} - w^T x^{(i)}\right)^2$$

Suppose the dataset is fixed – i.e. you are going to use the same training data with both losses – and that you are able to minimize both losses exactly during training. Will these two objectives always yield the same regression model? Why or why not? (Briefly justify your answer.)
Solution: Yes, both objectives, when applied to the same dataset, will always have the same \textit{argmin} and thus yield the same $w$. Notice that:

$$\text{NewLoss}(w) = \log(\text{Loss}(w))$$

Since \text{log} is a strict monotonic increasing function, the $w$ that minimizes \text{Loss} also minimizes \text{NewLoss}, and vice versa.