1. (a) From (3), we know that when we multiply a matrix $\mathbf{M} \in \mathbb{R}^{r\times c}$ by a vector $\mathbf{x} \in \mathbb{R}^c$, which is equivalent to a matrix in $\mathbb{R}^{c\times 1}$, the product is a column vector where each row is the dot product of the corresponding row of $\mathbf{M}$ with $\mathbf{x}$, i.e., the inner product of the two vectors. From (2), we know that the dot product of two vectors is a scalar. As a result, the product $\mathbf{Mx}$ is a column vector of r elements, which is equivalent to a matrix in $\mathbb{R}^{r\times 1}$. Therefore, $\mathbf{M} \in \mathbf{R}^{r\times c}$ implies that the function $f(\mathbf{x}) = \mathbf{Mx}$ can map $\mathbb{R}^{c\times 1} \to \mathbb{R}^{r\times 1}$.

   (b) From (4), we know that $\mathbf{y} = \mathbf{M_1x}$ is a column vector in $\mathbb{R}^d$. Applying the same rule to $\mathbf{M_2y}$, we can get $\mathbf{M_2y}$ is a column vector in $\mathbb{R}^r$. Therefore, the function $f(\mathbf{x}) = \mathbf{M_2M_1x}$ can map $\mathbb{R}^c \to \mathbb{R}^r$.

2. (a) $\frac{\partial}{\partial y}\ln(\frac{x^5}{y^2}) = \frac{\partial}{\partial y}(5\ln(x) - 2\ln(y)) = -2\frac{1}{y}$

   (b) $\frac{\partial}{\partial x_j}\ln\left(\sum_i x_iy_i\right) = \frac{\frac{\partial}{\partial x_j}\left(\sum_i x_iy_i\right)}{\sum_i x_iy_i} = \frac{y_j}{\sum_i x_iy_i}$

3. (a) i. $\mathbf{J}_g = \begin{bmatrix} \frac{\partial g_1}{\partial u} & \frac{\partial g_1}{\partial v} & \frac{\partial g_1}{\partial w} \\ \frac{\partial g_2}{\partial u} & \frac{\partial g_2}{\partial v} & \frac{\partial g_2}{\partial w} \\ \frac{\partial g_3}{\partial u} & \frac{\partial g_3}{\partial v} & \frac{\partial g_3}{\partial w} \end{bmatrix} = \begin{bmatrix} e^u & \frac{e^v}{1+e^v} & 0 \\ 2uw & 0 & u^2+1 \\ 0 & \frac{e^{-v}}{(1+e^{-v})^2} & 3w^2 \end{bmatrix}$

   ii. Substituting $u = 1, v = 0, w = 1$ into $\mathbf{J}_g$, we get $\mathbf{J}_g = \begin{bmatrix} e & \frac{1}{2} & 0 \\ 2 & 0 & 2 \\ 0 & \frac{1}{4} & 3 \end{bmatrix}$. The first column contains e while the other two columns contain only rational numbers.

   (b) i. This composition is invalid due to incompatible dimensions. The output of g is a vector in $\mathbb{R}^3$, while the input of p is a vector in $\mathbb{R}^2$.

   ii.
   - $\mathbf{g}(\mathbf{f}(s,t)) = \begin{bmatrix} e^{s^2t} + \ln(1+e^{s+e^t}) \\ \ln(1+s^2)\cdot((s^2t)^2+1) \\ \frac{1}{1+e^{-(s+e^t)}} + (\ln(1+s^2))^3 \end{bmatrix}$
   - $\frac{\partial}{\partial s}\mathbf{g}(\mathbf{f}(s,t)) = \begin{bmatrix} 2ste^{s^2t} + \frac{e^{s+e^t}}{1+e^{s+e^t}} \\ \frac{2s\cdot((s^2t)^2+1)}{1+s^2} + \ln(1+s^2)\cdot 4s^3t^2 \\ \frac{e^{-s-e^t}}{\left(1+e^{-s-e^t}\right)^2} + \frac{6s\ln^2\left(1+s^2\right)}{1+s^2} \end{bmatrix}$
   - $\mathbf{J_f}(s,t) = \begin{pmatrix} 2st & s^2 \\ 1 & e^t \\ \frac{2s}{1+s^2} & 0 \end{pmatrix}$

   To get $\mathbf{g}(\mathbf{f}(s,t))$ using the chain rule, we multiply $\mathbf{J_g}(\mathbf{f}(s,t))$ with the first column of $\mathbf{J_f}(s,t)$, that is, $\mathbf{J_g}(\mathbf{f}(s,t))\cdot\frac{\partial \mathbf{f}}{\partial s} = \begin{pmatrix} e^{s^2t} & \frac{e^{s+e^t}}{1+e^{s+e^t}} & 0 \\ 2s^2t\ln(1+s^2) & 0 & (s^2t)^2+1 \\ 0 & \frac{e^{-(s+e^t)}}{(1+e^{-(s+e^t)})^2} & 3(\ln(1+s^2))^2 \end{pmatrix}\begin{pmatrix} 2st \\ 1 \\ \frac{2s}{1+s^2} \end{pmatrix}$,

   which results in the same vector as $\frac{\partial}{\partial s}\mathbf{g}(\mathbf{f}(s,t)) = \begin{bmatrix} 2ste^{s^2t} + \frac{e^{s+e^t}}{1+e^{s+e^t}} \\ \frac{2s\cdot((s^2t)^2+1)}{1+s^2} + \ln(1+s^2)\cdot 4s^3t^2 \\ \frac{e^{-s-e^t}}{\left(1+e^{-s-e^t}\right)^2} + \frac{6s\ln^2\left(1+s^2\right)}{1+s^2} \end{bmatrix}$.

4. (a) i. $\mathbf{r}'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ \text{undefined} & x = 0 \end{cases}$

   ii. $\mathbf{J_r} = \begin{bmatrix} \mathbf{r}'(x_1) & 0 & 0 \\ 0 & \mathbf{r}'(x_2) & 0 \\ 0 & 0 & \mathbf{r}'(x_3) \end{bmatrix}$

(b) $\mathbf{J_A} = \begin{bmatrix} 2x_1 & 0 \\ 0 & 2x_2 \end{bmatrix}, \mathbf{J_B} = \begin{bmatrix} 2x_1 & 1 \\ 1 & 2x_2 \end{bmatrix}$. Since the non-diagonal elements of $\mathbf{J_B}$ are non-zero, $\mathbf{J_B}$ is not a diagonal matrix while $\mathbf{J_A}$ is a diagonal matrix. This is because that function B is not a elements-wise function, and thus each output element is not only related to one input element.

5. (a) i. Since the total probability must sum to 1, $\mathbb{P}[\hat{Y} = 0] = \mathbb{P}[\hat{Y} = 1] = 0.5$.
   ii. No, Their assumption is not correct. Their assumption implies that the classifier would have a fifty percent chance of classifying an image as a cat or a dog, regardless of the input image. However, if the input image is a clear image of a cat, the classifier should have a much higher chance of classifying it as a cat than as a dog, and thus the probability should not be 0.5.

6. (a) i. A. $\mathbb{E}[Z] = -2 \times 0.3 + 0 \times 0.4 + 2 \times 0.3 = 0, \mathbb{V}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$. Since $\mathbb{E}[Z] = 0$, $\mathbb{V}[Z] = \mathbb{E}[Z^2] = (-2)^2 \times 0.3 + 0^2 \times 0.4 + 2^2 \times 0.3 = 2.4$.
   B. $\mathbb{E}[Z^2] = (-2)^2 \times 0.3 + 0^2 \times 0.4 + 2^2 \times 0.3 = 4 \times 0.3 + 0 + 4 \times 0.3 = 2.4 = \mathbb{V}[Z]$
   ii. The mean and the variance of the distribution of $X+Y$ are $2-1 = 1$ and $9+4 = 13$ respectively. Since X and Y are all Normal distributions, $X + Y$ is also a Normal distribution. Therefore, $X + Y \sim \mathcal{N}(1, 13)$.
   iii. The mean and the variance of the distribution of $3X - 2Y + 5$ are $3 \times 2 - 2 \times (-1) + 5 = 13$ and $3^2 \times 9 + (-2)^2 \times 4 = 81 + 16 = 97$ respectively. Since X and Y are all Normal distributions, $3X - 2Y + 5$ is also a Normal distribution. Therefore, $3X - 2Y + 5 \sim \mathcal{N}(13, 97)$.

   (b) i. Since the expected value for each entry A_{ij} in the matrix $\mathbf{A}$ is 0, the expected value of the matrix itself is a zero matrix. Therefore, $\mathbb{E}[\mathbf{A}] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. When we multiply it by the vector $\mathbf{v}$, the expected value of the product is a zero vector.
   ii. $\mathbb{V}[(\mathbf{Av})_1] = \mathbb{V}[A_{11}] + \mathbb{V}[A_{12}] = 1 + 1 = 2$.
   iii. $\mathbb{V}[(\mathbf{Av})_i] = \mathbb{V}[A_{i1}] + \mathbb{V}[A_{i2}] = v1^2 + v2^2$. Since $v1^2 + v2^2 = c$, $\mathbb{V}[(\mathbf{Av})_i] = c$.

   (c) i. Since y is a linear function of the normally distributed variable ϵ_1, y is also normally distributed. The mean and variance of y at $x = 1$ are $2 + 3 + \mathbb{E}[\epsilon_1] = 5$ and $\mathbb{V}[\epsilon_1] = 1$ respectively. Therefore, $y \sim \mathcal{N}(5, 1)$. Given this information, the probability density function of y at $x = 1$ is $f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(4.8-5)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{0.04}{2}} = \frac{1}{\sqrt{2\pi}} e^{-0.02}$.

   (d) i. $\mathbb{E}[\bar{M}] = \mu$.
   ii. $\mathbb{V}[\bar{M}] = \frac{\sigma^2}{16}$.
   iii. To make the variance 4 times smaller than the variance we get from the sampling distribution of sample size $n = 16$, the sample size needs to be 4 times larger, which is $16 \times 4 = 64$.