EECE7397 – Homework 1

1.

$$y(x, \mathbf{w}) \;=\; w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M \;\; = \sum_{j=0}^{M} w_j \big(x_n\big)^j$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big[y(x_n, \mathbf{w}) - t_n \Big]^2$$

$$\Rightarrow E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n \Big]^2$$

$$\frac{\partial E}{\partial w_i} = \frac{1}{2} \sum_{n=1}^{N} 2 \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] \frac{\partial}{\partial w_i} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big]$$

$$= \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i \left(\text{ As } \frac{\partial}{\partial w_i} w_j (x_n)^j = \delta_{ij} (x_n)^i \text{ if } j = i \right)$$

$$\frac{\partial E}{\partial w_i} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \Big[\sum_{j=0}^{M} w_j (x_n)^j - t_n \Big] (x_n)^i = 0$$

$$\Rightarrow \sum_{n=1}^{N} \sum_{j=0}^{M} w_j (x_n)^j (x_n)^i - \sum_{n=1}^{N} t_n (x_n)^i = 0$$

$$\Rightarrow \sum_{n=1}^{N} \sum_{j=0}^{M} w_j (x_n)^{j+i} - \sum_{n=1}^{N} t_n (x_n)^i = 0$$

$$\Rightarrow \sum_{j=0}^{M} w_j \sum_{n=1}^{N} (x_n)^{j+i} - \sum_{n=1}^{N} t_n (x_n)^i = 0$$

$$\Rightarrow \sum_{j=0}^{M} w_j \sum_{n=1}^{N} (x_n)^{j+i} = \sum_{n=1}^{N} t_n (x_n)^i$$

$$\Rightarrow \sum_{j=0}^{M} A_{ij} w_j = T_i \left(A_{ij} = \sum_{n=1}^{N} (x_n)^{j+i}, \quad T_i = \sum_{n=1}^{N} t_n (x_n)^i \right)$$

In matrix notation:

$$\mathbf{A}\,\mathbf{w} = \mathbf{T}$$

$$\Rightarrow \mathbf{w} = \mathbf{A}^{-1}\mathbf{T}$$

2.

$$y(x, \mathbf{w}) \;=\; w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M \;= \sum_{j=0}^{M} w_j \big(x_n\big)^j$$

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\Big[y(x_n, \mathbf{w}) - t_n\Big]^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$
$$\Rightarrow \tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n\Big]^2 + \frac{\lambda}{2}\sum_{j=0}^{M} w_j^2$$

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{2}\sum_{n=1}^{N}\Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n\Big]^2 + \frac{\partial}{\partial w_i}\frac{\lambda}{2}\sum_{j=0}^{M} w_j^2$$
$$= \frac{1}{2}\sum_{n=1}^{N} 2\Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n\Big]\frac{\partial}{\partial w_i}\Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n\Big] + \frac{\lambda}{2}\frac{\partial}{\partial w_i}\sum_{j=0}^{M} w_j^2$$
$$= \sum_{n=1}^{N}\Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n\Big](x_n)^i + \lambda w_i \left(\text{ As } \frac{\partial}{\partial w_i} w_j (x_n)^j = \delta_{ij}(x_n)^i \text{ if } j = i\right)$$

$$\frac{\partial \tilde{E}}{\partial w_i} = 0$$
$$\Rightarrow \sum_{n=1}^{N}\Big[\sum_{j=0}^{M} w_j \big(x_n\big)^j - t_n\Big](x_n)^i + \lambda w_i = 0$$
$$\Rightarrow \sum_{n=1}^{N}\sum_{j=0}^{M} w_j \big(x_n\big)^j (x_n)^i - \sum_{n=1}^{N}(x_n)^i t_n + \lambda w_i = 0$$
$$\Rightarrow \sum_{n=1}^{N}\sum_{j=0}^{M} w_j \big(x_n\big)^{j+i} - \sum_{n=1}^{N}(x_n)^i t_n + \lambda w_i = 0$$
$$\Rightarrow \sum_{j=0}^{M} w_j \sum_{n=1}^{N} \big(x_n\big)^{j+i} - \sum_{n=1}^{N}(x_n)^i t_n + \lambda w_i = 0$$
$$\Rightarrow \sum_{j=0}^{M} w_j \sum_{n=1}^{N} \big(x_n\big)^{j+i} + \lambda w_i = \sum_{n=1}^{N}(x_n)^i t_n$$
$$\Rightarrow \sum_{j=0}^{M} A_{ij} w_j + \lambda w_i = T_i \left(A_{ij} = \sum_{n=1}^{N}(x_n)^{j+i}, \quad T_i = \sum_{n=1}^{N} t_n (x_n)^i\right)$$

In matrix notation:

$$\mathbf{A}\,\mathbf{w} + \lambda\,\mathbf{w} = \mathbf{T}$$
$$\Rightarrow (\mathbf{A} + \lambda\,\mathbf{I})\,\mathbf{w} = \mathbf{T}$$
$$\Rightarrow \mathbf{w} = (\mathbf{A} + \lambda\,\mathbf{I})^{-1}\,\mathbf{T}$$

The regularization penalty term shrinks the parameter vector $\mathbf{w}$ toward 0 or smaller magnitudes, preventing large parameter values that can lead to overfitting. Another example of regularization is lasso, which adds a $\lambda\|\mathbf{w}\|$ penalty term. This is similar to the ridge regularization term $\frac{\lambda}{2}\|\mathbf{w}\|^2$ in that both push parameters toward zero; but, the lasso regularization penalty promotes sparsity, resulting in some coefficients resulting in 0.

3.

Assuming for a simple function $f(x)$, change of variable using non-linear function $g(y)$ results in $h(y) = f\big(g(y)\big)$.

$$\frac{d}{dy}\Big[f\big(g(y)\big)\Big] \;=\; f'\big(g(y)\big)\,g'(y)$$

$h(y)$ has maximum at $\hat{y}$ (mode $\hat{y}$):

$$h'(\hat{y}) \;=\; f'\big(g(\hat{y})\big)\,g'(\hat{y}) \;=\; 0$$

Assuming $g'(\hat{y}) \neq 0$:

$$f'\big(g(\hat{y})\big) \;=\; 0$$

$f(x)$ has maximum at $\hat{x}$ (mode $\hat{x}$):

$$f'(\hat{x}) \;=\; 0$$

As $f'(\hat{x}) = 0$ and $f'\big(g(\hat{y})\big) = 0$:

$$\hat{x} \;=\; g(\hat{y})$$

The $\hat{y}$ that maximizes h in y maps directly to the $\hat{x}$ that maximizes f in x via $\hat{x} = g(\hat{y})$, which is the maximum of f in x.

For a probability density function $p_x(x)$ change of variable using non-linear function $g(y)$ results in $p_y(y) \;=\; p_x\big(g(y)\big)\,\big|g'(y)\big|$

$$\begin{aligned}
\frac{d}{dy}\,p_y(y) &= \frac{d}{dy}\Big[p_x\big(g(y)\big)\,\big|g'(y)\big|\Big] \\
&= \frac{d}{dy}\Big[p_x\big(g(y)\big)\Big]\cdot\big|g'(y)\big| \;+\; p_x\big(g(y)\big)\,\frac{d}{dy}\Big[\big|g'(y)\big|\Big] \\
&= p_x'\big(g(y)\big)\,g'(y)\,\cdot\big|g'(y)\big| \;+\; p_x\big(g(y)\big)\cdot\,\operatorname{sgn}\!\big[g'(y)\big]g''(y)
\end{aligned}$$

$p_x(x)$ has a maximum at $\hat{x}$ (mode $\hat{x}$):

$$p_x'(\hat{x}) = 0$$

$p_y(y)$ has a maximum at $\hat{y}$ (mode $\hat{y}$):

$$p_y'(\hat{y}) = 0$$

$$\Rightarrow p_x'\big(g(\hat{y})\big)\,g'(\hat{y})\,\cdot\big|g'(\hat{y})\big| \;+\; p_x\big(g(\hat{y})\big)\,\cdot\operatorname{sgn}\!\big[g'(\hat{y})\big]\,g''(\hat{y}) = 0$$

$$\Rightarrow p_x'\big(g(\hat{y})\big)\,g'(\hat{y})\cdot\,g'(\hat{y})\,\cdot\operatorname{sgn}\!\big[g'(\hat{y})\big] \;+\; p_x\big(g(\hat{y})\big)\operatorname{sgn}\!\big[g'(\hat{y})\big]\,g''(\hat{y}) = 0$$

$$\Rightarrow \operatorname{sgn}\!\big[g'(\hat{y})\big]\,\Big(p_x'\big(g(\hat{y})\big)\,g'(\hat{y})^2 \;+\; p_x\big(g(\hat{y})\big)\,g''(\hat{y})\Big) = 0$$

$$\Rightarrow p_x'\big(g(\hat{y})\big)\,g'(\hat{y})^2 \;+\; p_x\big(g(\hat{y})\big)\,g''(\hat{y}) = 0$$

$$\Rightarrow p_x'\big(g(\hat{y})\big)\,g'(\hat{y})^2 \;=\; -p_x\big(g(\hat{y})\big)g''(\hat{y})$$

Because $g''(\hat{y}) \neq 0$, the expression does not reduce to $p_x'\big(g(\hat{y})\big) = 0$. The extra term $p_x\big(g(\hat{y})\big)\,g''(\hat{y})$, arising from the derivative of the Jacobian factor $|g'(y)|$, shifts the maximum away from the point where $p_x'(x) = 0$. The mode of $p_y(y)$ does not necessarily map to the mode of $p_x(x)$ via $x = g(y)$.

For a linear transformation, assume function $g(y) = a\,y + b$,

$$g'(y) = a, \quad g''(y) = 0.$$

For a probability density function $p_x(x)$ change of variable using linear function $g(y)$ results in $p_y(y) \;=\; p_x\big(a\,y+b\big)\,|a|$.

$$
\begin{aligned}
\frac{d}{dy}p_y(y) &= \frac{d}{dy}\Big[p_x(a\,y+b)\,|a|\Big] \\
&= |a|\,\frac{d}{dy}\Big[p_x(a\,y+b)\Big] \\
&= |a|\,\big[p_x'(a\,y+b)\,a\big] \\
&= a\,|a|\,p_x'(a\,y+b).
\end{aligned}
$$

$p_x(x)$ has a maximum at $\hat{x}$ (mode $\hat{x}$):

$$p_x'(\hat{x}) = 0$$

$p_y(y)$ has a maximum at $\hat{y}$ (mode $\hat{y}$):

$$
\begin{aligned}
p_y'(\hat{y}) &= 0 \\
\Rightarrow a\,|a|\,p_x'(a\,\hat{y}+b) &= 0 \\
\Rightarrow p_x'(a\,\hat{y}+b) &= 0
\end{aligned}
$$

As $p_x'(\hat{x}) = 0$ and $p_x'(a\,\hat{y}+b)) = 0$:

$$\hat{x} \;=\; a\,\hat{y}+b$$

For a general nonlinear $g(y)$, the transformed density includes the Jacobian factor $|g'(y)|$ which results in When you an extra term involving $g''(y)$ shifting the mode when finding the location of the maximum. Consequently, the value $\hat{y}$ that maximizes $p_y(y)$ does not, in general, map via $\hat{x} = g(\hat{y})$ to the value $\hat{x}$ that maximizes $p_x(x)$. This shift in the mode occurs because the Jacobian term $|g'(y)|$ affects where the maximum occurs.

If $g(y) = a\,y+b$ with $a \neq 0$, then $g'(y) = a$ and $g''(y) = 0$, so the Jacobian factor $|g'(y)| = |a|$ is a constant. Thus, maximizing $p_x\big(a\,y+b\big)\,|a|$ is equivalent to maximizing $p_x\big(a\,y+b\big)$.The mode $\hat{y}$ of $p_y(y)$ and the mode $\hat{x}$ of $p_x(x)$ map via $\hat{x} = a\,\hat{y}+b$.

4.

$$E\big[L(\mathbf{t},\mathbf{y}(\mathbf{x}))\big] = \iint \|\mathbf{y}(\mathbf{x})-\mathbf{t}\|^2\,p(\mathbf{x},\mathbf{t})\,d\mathbf{x}\,d\mathbf{t}$$

$$\|\mathbf{y}(\mathbf{x})-\mathbf{t}\|^2 \;=\; \big(\mathbf{y}(\mathbf{x})-\mathbf{t}\big)^{\top}\big(\mathbf{y}(\mathbf{x})-\mathbf{t}\big) \;=\; \sum_{i=1}^{M}\big[y_i(\mathbf{x})-t_i\big]^2$$

$$
\begin{aligned}
E[L] &= \iint \|\mathbf{y}(\mathbf{x})-\mathbf{t}\|^2\,p(\mathbf{x},\mathbf{t})\,d\mathbf{x}\,d\mathbf{t} \\
&= \iint \sum_{i=1}^{M}\big[y_i(\mathbf{x})-t_i\big]^2\,p(\mathbf{x},\mathbf{t})\,d\mathbf{x}\,d\mathbf{t}
\end{aligned}
$$

$$
\begin{aligned}
\delta E[L] &= \frac{\delta}{\delta y_j(\mathbf{x})}\iint \sum_{i=1}^{M}\big[y_i(\mathbf{x})-t_i\big]^2\,p(\mathbf{x},\mathbf{t})\,d\mathbf{x}\,d\mathbf{t} \\
&= \iint \frac{\delta}{\delta y_j(\mathbf{x})}\left[\sum_{i=1}^{M}\big(y_i(\mathbf{x})-t_i\big)^2\right]p(\mathbf{x},\mathbf{t})\,d\mathbf{x}\,d\mathbf{t} \\
&= \iint 2\,\big[y_j(\mathbf{x})-t_j\big]\,p(\mathbf{x},\mathbf{t})\,d\mathbf{x}\,d\mathbf{t} \\
&= \int 2\,\big[y_j(\mathbf{x})-t_j\big]\,p(\mathbf{x},\mathbf{t})\,d\mathbf{t}
\end{aligned}
$$

$$
\begin{aligned}
\int 2\left[y_j(\mathbf{x}) - t_j\right] p(\mathbf{x},\mathbf{t})\, d\mathbf{t} &= 0 \\
\Rightarrow\ y_j(\mathbf{x}) \int p(\mathbf{x},\mathbf{t})\, d\mathbf{t} \ - \int t_j\, p(\mathbf{x},\mathbf{t})\, d\mathbf{t} &= 0 \\
\Rightarrow\ y_j(\mathbf{x}) \int p(\mathbf{x},\mathbf{t})\, d\mathbf{t} &= \int t_j\, p(\mathbf{x},\mathbf{t})\, d\mathbf{t} \\
\Rightarrow\ y_j(\mathbf{x}) &= \frac{\int t_j\, p(\mathbf{x},\mathbf{t})\, d\mathbf{t}}{\int p(\mathbf{x},\mathbf{t})\, d\mathbf{t}} \\
\Rightarrow\ y_j(\mathbf{x}) &= \frac{\int t_j\, p(\mathbf{x},\mathbf{t})\, d\mathbf{t}}{p(\mathbf{x})} \quad (\text{As } p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{t})\, d\mathbf{t}) \\
\Rightarrow\ y_j(\mathbf{x}) &= \frac{\int t_j\, p(\mathbf{t} \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{t}}{p(\mathbf{x})} \quad (\text{As } p(\mathbf{x},\mathbf{t}) = p(\mathbf{t} \mid \mathbf{x})\, p(\mathbf{x})) \\
\Rightarrow\ y_j(\mathbf{x}) &= \int t_j\, p(\mathbf{t} \mid \mathbf{x})\, d\mathbf{t} \\
\Rightarrow\ y_j(\mathbf{x}) &= \mathbf{E}\left[t_j \mid \mathbf{x}\right] \quad (\text{As } \mathbf{E}[x \mid y] = \int x\, p(x \mid y)\, dx)
\end{aligned}
$$

For multiple target variables $\mathbf{t} = (t_1, \dots, t_M)$:

$$\mathbf{y}(\mathbf{x}) = (y_1(\mathbf{x}), \dots, y_M(\mathbf{x})) = \mathbf{E}[\mathbf{t} \mid \mathbf{x}],$$

$$y_j(\mathbf{x}) = \int t_j\, p(\mathbf{t} \mid \mathbf{x})\, d\mathbf{t} = \mathbf{E}[t_j \mid \mathbf{x}]$$

For a single target variable $t \quad (j = 1 \text{ as } M = 1)$:

$$y(\mathbf{x}) = \int t\, p(t \mid \mathbf{x})\, dt = \mathbf{E}[t \mid \mathbf{x}]$$

5.

$$\binom{N}{m} = \frac{N!}{m!\,(N-m)!}$$

$$\binom{N}{m-1} = \frac{N!}{(m-1)!\,[N-(m-1)]!} = \frac{N!}{(m-1)!\,(N-m+1)!}$$

$$
\begin{aligned}
\binom{N}{m} + \binom{N}{m-1} &= \frac{N!}{m!(N-m)!} \ + \ \frac{N!}{(m-1)!(N-m+1)!} \\
&= \frac{N!}{m\,(m-1)!(N-m)!} \ + \ \frac{N!}{(m-1)!\,(N-m+1)\,(N-m)!} \\
&\quad (\text{As } m! = m\,(m-1)! \text{ and } (N-m+1)! = (N-m+1)\,(N-m)!) \\
&= \frac{N!}{(N-m)!(m-1)!}\,\frac{1}{m} \ + \ \frac{N!}{(N-m)!(m-1)!}\,\frac{1}{N-m+1} \\
&= \frac{N!}{(N-m)!(m-1)!}\left(\frac{1}{m} \ + \ \frac{1}{N-m+1}\right) \\
&= \frac{N!}{(N-m)!(m-1)!}\left(\frac{N-m+1+m}{m(N-m+1)}\right) \\
&= \frac{N!}{(N-m)!(m-1)!}\left(\frac{N+1}{m(N-m+1)}\right).
\end{aligned}
$$

$$\binom{N+1}{m} = \frac{(N+1)!}{m!\,\big[(N+1)-m\big]!} = \frac{(N+1)!}{m!\,(N+1-m)!}$$

$$= \frac{(N+1)\,N!}{m\,(m-1)!\,(N+1-m)\,(N-m)!}$$

$$= \frac{N!}{(N-m)!\,(m-1)!}\,\frac{N+1}{m\,(N+1-m)}.$$

As the expressions for $\binom{N}{m} + \binom{N}{m-1}$ and $\binom{N+1}{m}$ match, identity is proven:

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$$

Given the identity:

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$$

Proof of the Binomial Theorem by mathematical induction:

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m$$

Base Case N=0:

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m$$

$$\Rightarrow (1+x)^0 = \sum_{m=0}^{0} \binom{0}{m} x^m$$

$$\Rightarrow (1+x)^0 = \sum_{m=0}^{0} \binom{0}{0} x^0$$

$$\Rightarrow 1 = 1$$

The base case holds for N=0.

Inductive Hypothesis:
Assume the theorem holds for some $N \geq 0$, then it has to proven in the inductive step that it holds for N+1

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m$$

Inductive Step:
To prove it holds for $N+1$:

$$(1+x)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} x^m$$

$$
\begin{aligned}
&(1+x)^{N+1} \\
&= (1+x)^N \cdot (1+x) \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m \cdot (1+x) \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=0}^{N} \binom{N}{m} x^m \cdot x \\
&= \sum_{m=0}^{N} \binom{N}{m} x^m + \sum_{m=0}^{N} \binom{N}{m} x^{m+1}
\end{aligned}
$$

Term x^k for $1 \leq k \leq N$:
In the first sum, the term $\binom{N}{m}x^m$ matches x^k when $m = k$, giving the coefficient $\binom{N}{k}$. In the second sum, $\binom{N}{m}x^{m+1}$ matches x^k when $m+1 = k$, giving the coefficient $\binom{N}{k-1}$. By the identity $\binom{N}{k} + \binom{N}{k-1} = \binom{N+1}{k}$ the total coefficient of x^k is $\binom{N+1}{k}$.
The constant term x^0 (when $m = 0$) comes from the first sum $\binom{N}{0} x^0 = 1$, and the x^{N+1} term comes from the second sum $\binom{N}{N} x^{N+1} = x^{N+1}$ when $m = N$. The total coefficient of the powers $x^1, x^2, \ldots, x^N$.Thus, the expression becomes:

$$
(1+x)^{N+1} = \sum_{m=0}^{N+1} \binom{N+1}{m} x^m
$$

$$
\begin{aligned}
&\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{-m} \\
&= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \left(\frac{\mu}{1-\mu}\right)^m
\end{aligned}
$$

Let x be $\frac{\mu}{1-\mu}$.

$$
(1+x)^N = \left(1 + \frac{\mu}{1-\mu}\right)^N = \left(\frac{1-\mu+\mu}{1-\mu}\right)^N = \left(\frac{1}{1-\mu}\right)^N = \left(1-\mu\right)^{-N}
$$

From the binomial theorem:

$$
(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m
$$

Thus, $\sum_{m=0}^{N} \binom{N}{m} x^m = (1-\mu)^{-N}$

$$
\begin{aligned}
&(1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \left(\frac{\mu}{1-\mu}\right)^m \\
&= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} x^m \\
&= \left(1-\mu\right)^N \left(1-\mu\right)^{-N} \\
&= 1
\end{aligned}
$$

Thus, $\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1$ is proven.

$$
\begin{aligned}
\mathbf{E}[\mathbf{X}] &= \sum_{m=0}^{N} m \cdot \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} m \cdot \frac{N!}{(N-m)!m!} \cdot \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} m \cdot \frac{N(N-1)!}{(N-m)! \cdot m(m-1)!} \cdot \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} \frac{N(N-1)!}{(N-m)!(m-1)!} \cdot \mu^m (1-\mu)^{N-m} \\
&= N \sum_{m=1}^{N} \frac{(N-1)!}{(N-m)!(m-1)!} \cdot \mu^m (1-\mu)^{N-m}
\end{aligned}
$$

Let $k = m-1$, then $m = k+1$. As m goes from 1 to N, k goes from 0 to $N-1$.

$$
\begin{aligned}
& N \sum_{m=1}^{N} \frac{(N-1)!}{(N-m)!(m-1)!} \cdot \mu^m (1-\mu)^{N-m} \\
&= N \sum_{k=0}^{N-1} \frac{(N-1)!}{(N-(k+1))!((k+1)-1)!} \cdot \mu^{k+1} (1-\mu)^{N-(k+1)} \\
&= N \sum_{k=0}^{N-1} \frac{(N-1)!}{(N-1-k)!k!} \cdot \mu^{k+1} (1-\mu)^{(N-1)-k} \\
&= N \sum_{k=0}^{N-1} \frac{(N-1)!}{(N-1-k)!k!} \cdot \mu^{k} \mu (1-\mu)^{(N-1)-k} \\
&= N\mu \sum_{k=0}^{N-1} \frac{(N-1)!}{(N-1-k)!k!} \cdot \mu^{k} (1-\mu)^{(N-1)-k} \\
&= N\mu \sum_{k=0}^{N-1} \binom{N-1}{k} \mu^{k} (1-\mu)^{(N-1)-k} \\
&= N\mu \\
& \left(As(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1 \right) \\
& \left(As(\mu+(1-\mu))^{N-1} = \sum_{k=0}^{N-1} \binom{N-1}{k} \mu^k (1-\mu)^{(N-1)-k} = 1 \right)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}[\mathbf{X^2}] &= \sum_{m=0}^{N} m^2 \cdot \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} (m(m-1)+m) \cdot \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \Big(\sum_{m=0}^{N} m(m-1) + \sum_{m=0}^{N} m \Big) \cdot \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} m(m-1) \binom{N}{m} \mu^m (1-\mu)^{N-m} + \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} m(m-1) \cdot \frac{N!}{(N-m)!m!} \mu^m (1-\mu)^{N-m} + \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} m(m-1) \cdot \frac{N(N-1)(N-2)!}{(N-m)!m(m-1)(m-2)!} \mu^m (1-\mu)^{N-m} + \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= \sum_{m=0}^{N} \frac{N(N-1)(N-2)!}{(N-m)!(m-2)!} \mu^m (1-\mu)^{N-m} + \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m} \\
&= N(N-1) \sum_{m=0}^{N} \frac{(N-2)!}{(N-m)!(m-2)!} \mu^m (1-\mu)^{N-m} + \sum_{m=0}^{N} m \binom{N}{m} \mu^m (1-\mu)^{N-m}
\end{aligned}
$$

Let $k = m - 2$, then $m = k + 2$. As m goes from 2 to N, k goes from 0 to $N - 2$.

$$
\begin{aligned}
& N(N-1) \sum_{m=0}^{N} \frac{(N-2)!}{(N-m)!(m-2)!} \mu^m (1-\mu)^{N-m} \\
&= N(N-1) \sum_{k=0}^{N-2} \frac{(N-2)!}{(N-(k+2))!(k+2-2)!} \mu^m (1-\mu)^{N-m} \\
&= N(N-1) \sum_{k=0}^{N-2} \frac{(N-2)!}{(N-2-k)!k!} \mu^m (1-\mu)^{N-m} \\
&= N(N-1) \sum_{k=0}^{N-2} \binom{N-2}{k} \mu^{k+2} (1-\mu)^{N-(k+2)} \\
&= N(N-1) \sum_{k=0}^{N-2} \binom{N-2}{k} \mu^k \mu^2 (1-\mu)^{N-(k+2)} \\
&= N(N-1)\mu^2 \sum_{k=0}^{N-2} \binom{N-2}{k} \mu^k (1-\mu)^{(N-2)-k} \\
&= N(N-1)\mu^2 \\
& \Big(As (1+x)^N = \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1 \Big) \\
& \Big(As (\mu + (1-\mu))^{N-2} = \sum_{k=0}^{N-2} \binom{N-2}{k} \mu^k (1-\mu)^{(N-2)-k} = 1 \Big)
\end{aligned}
$$

Thus,

$$
\mathbf{E}[\mathbf{X^2}] = N(N-1)\mu^2 + N\mu
$$

$$
\begin{aligned}
\mathbf{var}(\mathbf{X}) &= \mathbf{E}[\mathbf{X^2}] - \mathbf{E}[\mathbf{X}]^2 \\
&= N(N-1)\mu^2 + N\mu - (N\mu)^2 \\
&= N^2\mu^2 - N\mu^2 + N\mu - N^2\mu^2 \\
&= N\mu - N\mu^2 \\
&= N\mu(1-\mu)
\end{aligned}
$$

6.

$$
\mathrm{Beta}(\mu \mid a, b) = p(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,\mu^{a-1}\,(1-\mu)^{b-1}
$$

$$
\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,d\mu = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}
$$

$$
\begin{aligned}
E[\mu] &= \int_0^1 \mu\,p(\mu)\,\mathrm{d}\mu \\
&= \int_0^1 \mu\,\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\,\mu^{a-1}\,(1-\mu)^{b-1}\,\mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\int_0^1 \mu\,\mu^{a-1}\,(1-\mu)^{b-1}\,\mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\int_0^1 \mu^{(a+1)-1}\,(1-\mu)^{b-1}\,\mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\cdot\frac{\Gamma(a+1)\,\Gamma(b)}{\Gamma\big((a+1)+b\big)} \quad \left(\text{As } \int_0^1 \mu^{a-1}(1-\mu)^{b-1}\,\mathrm{d}\mu = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}\right) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\cdot\frac{\Gamma(a+1)\,\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\cdot\frac{a\,\Gamma(a)\,\Gamma(b)}{(a+b)\,\Gamma(a+b)} \quad (\text{As } \Gamma(a+1) = a\,\Gamma(a),\ \Gamma(a+b+1) = (a+b)\,\Gamma(a+b)) \\
&= \frac{a}{a+b}
\end{aligned}
$$

$$
\begin{aligned}
E[\mu^2] &= \int_0^1 \mu^2\, p(\mu)\, \mathrm{d}\mu \\
&= \int_0^1 \mu^2\, \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, \mu^{a-1}\,(1-\mu)^{b-1}\, \mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_0^1 \mu^2\, \mu^{a-1}\,(1-\mu)^{b-1}\, \mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_0^1 \mu^{(a+2)-1}\,(1-\mu)^{b-1}\, \mathrm{d}\mu \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \cdot \frac{\Gamma(a+2)\,\Gamma(b)}{\Gamma((a+2)+b)} \quad \left(\text{As } \int_0^1 \mu^{a-1}(1-\mu)^{b-1}\, \mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\right) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \cdot \frac{\Gamma(a+2)\,\Gamma(b)}{\Gamma(a+b+2)} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \cdot \frac{(a+1)\,a\,\Gamma(a)\,\Gamma(b)}{(a+b+1)\,(a+b)\,\Gamma(a+b)} \\
&(\text{As } \Gamma((a+1)+1) = (a+1)\,\Gamma(a+1) = (a+1)a\Gamma(a)) \\
&(\text{As } \Gamma((a+b+1)+1) = (a+b+1)\,(a+b)\,\Gamma(a+b)) \\
&= \frac{a\,(a+1)}{(a+b)\,(a+b+1)}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[\mu] &= E[\mu^2] \;-\; \big(E[\mu]\big)^2 \\
&= \frac{a\,(a+1)}{(a+b)\,\big(a+b+1\big)} \;-\; \left(\frac{a}{a+b}\right)^2 \\
&= \frac{a\,(a+1)}{(a+b)\,(a+b+1)} \;-\; \frac{a^2}{(a+b)^2} \\
&= \frac{a\,(a+1)}{(a+b)\,(a+b+1)} \cdot \frac{(a+b)}{(a+b)} \;-\; \frac{a^2}{(a+b)^2} \cdot \frac{(a+b+1)}{(a+b+1)} \\
&= \frac{a\,(a+1)\,(a+b) - a^2\,(a+b+1)}{(a+b)^2\,(a+b+1)} \\
&= \frac{\big(a^3 + a^2 b + a^2 + ab\big) - \big(a^3 + a^2 b + a^2\big)}{(a+b)^2\,(a+b+1)} \\
&= \frac{ab}{(a+b)^2\,(a+b+1)}
\end{aligned}
$$

$$
p(\mu) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, \mu^{a-1}\,(1-\mu)^{b-1}
$$

$$
f(\mu) \;=\; \mu^{a-1}\,(1-\mu)^{b-1}
$$

As $\frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}$ is a constant (independent of μ), the value of μ that maximizes $p(\mu)$ also maximizes $f(\mu)$. To find the mode, the constant is ignored and $f(\mu)$ is maximized.

$$
\begin{aligned}
&\frac{d}{d\mu}\big[\mu^{a-1}(1-\mu)^{b-1}\big] \\
&= \mu^{a-1}\,\frac{d}{d\mu}\big[(1-\mu)^{b-1}\big] + (1-\mu)^{b-1}\,\frac{d}{d\mu}\big[\mu^{a-1}\big] \\
&= \mu^{a-1}\,\big[(b-1)(1-\mu)^{b-2}(-1)\big] + (1-\mu)^{b-1}\,\big[(a-1)\mu^{a-2}\big] \\
&= -\mu^{a-1}(b-1)(1-\mu)^{b-2} + \mu^{a-2}(1-\mu)^{b-1}(a-1) \\
&= \mu^{a-2}(1-\mu)^{b-2}\big[-\mu(b-1) + (1-\mu)(a-1)\big] \\
&= \mu^{a-2}(1-\mu)^{b-2}\big[(a-1) - (a+b-2)\mu\big]
\end{aligned}
$$

$$
\begin{aligned}
&\frac{df}{d\mu} = 0 \\
&\Rightarrow \mu^{a-2}(1-\mu)^{b-2}\big[(a-1) - (a+b-2)\mu\big] = 0 \\
&\Rightarrow \big[(a-1) - (a+b-2)\mu\big] = 0 \\
&\Rightarrow \mu = \frac{a-1}{a+b-2}
\end{aligned}
$$

$$
mode[\mu] = \frac{a-1}{a+b-2}
$$

7.

$$
\begin{aligned}
\mathbf{E}_{x,y}[x+ay] &= \iint (x+ay)\,p(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint (x+ay)\,p(x)\,p(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint x\,p(x)\,p(y)\,\mathrm{d}x\,\mathrm{d}y \;+\; \iint a\,y\,p(x)\,p(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \Big(\int x\,p(x)\,\mathrm{d}x\Big)\Big(\int p(y)\,\mathrm{d}y\Big) \;+\; a\,\Big(\int p(x)\,\mathrm{d}x\Big)\Big(\int y\,p(y)\,\mathrm{d}y\Big) \\
&= \int x\,p(x)\,\mathrm{d}x \;+\; a\int y\,p(y)\,\mathrm{d}y \quad (\text{As} \int p(x)\,\mathrm{d}x = 1 \text{ and } \int p(y)\,\mathrm{d}y = 1) \\
&= \mathbf{E}_x[x] \;+\; a\,\mathbf{E}_y[y] \quad (\text{As } \mathbf{E}_x[x] = \int x\,p(x)\,\mathrm{d}x,\; \mathbf{E}_y[y] = \int y\,p(y)\,\mathrm{d}y)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}_{x,y}[x+ay] &= \mathbf{E}_{x,y}[(x+ay)^2] \;-\; \mathbf{E}_{x,y}[x+ay]^{\,2} \\
&= \iint (x+ay)^2\, p(x,y)\,\mathrm{d}x\,\mathrm{d}y \;-\; \Big(\mathbf{E}_x[x] + a\,\mathbf{E}_y[y]\Big)^2 \\
&= \iint \big(x^2 + 2axy + a^2y^2\big)\, p(x)p(y)\,\mathrm{d}x\,\mathrm{d}y \;-\; \Big(\mathbf{E}_x[x] + a\,\mathbf{E}_y[y]\Big)^2 \\
&= \Big(\int x^2\, p(x)\,\mathrm{d}x\Big)\Big(\int p(y)\,\mathrm{d}y\Big) \;+\; 2a\,\Big(\int x\,p(x)\,\mathrm{d}x\Big)\Big(\int y\,p(y)\,\mathrm{d}y\Big) \\
&\quad\; +\; a^2\,\Big(\int p(x)\,\mathrm{d}x\Big)\Big(\int y^2\,p(y)\,\mathrm{d}y\Big) \;-\; \Big(\mathbf{E}_x[x] + a\,\mathbf{E}_y[y]\Big)^2 \\
&= \mathbf{E}_x[x^2] \;+\; 2a\,\mathbf{E}_x[x]\,\mathbf{E}_y[y] \;+\; a^2\,\mathbf{E}_y[y^2] \;-\; \Big(\mathbf{E}_x[x] + a\,\mathbf{E}_y[y]\Big)^2 \\
&\quad \Big(\text{As } \int p(x)\,\mathrm{d}x = 1,\; \int p(y)\,\mathrm{d}y = 1,\; \int x\,p(x)\,\mathrm{d}x = \mathbf{E}_x[x],\; \int y\,p(y)\,\mathrm{d}y = \mathbf{E}_y[y]\Big) \\
&= \mathbf{E}_x[x^2] + 2a\,\mathbf{E}_x[x]\mathbf{E}_y[y] + a^2\,\mathbf{E}_y[y^2] \;-\; \Big(\mathbf{E}_x[x]^2 + 2a\,\mathbf{E}_x[x]\mathbf{E}_y[y] + a^2\,\mathbf{E}_y[y]^2\Big) \\
&= \big(\mathbf{E}_x[x^2] - \mathbf{E}_x[x]^2\big) \;+\; a^2\,\big(\mathbf{E}_y[y^2] - \mathbf{E}_y[y]^2\big) \\
&= \mathrm{var}_x[x] \;+\; a^2\,\mathrm{var}_y[y] \\
&\quad (\text{As } \mathrm{var}_x[x] = \mathbf{E}_x[x^2] - (\mathbf{E}_x[x])^2)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}[x] &= \iint x\,p(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint x\,p(x\mid y)\,p(y)\,\mathrm{d}x\,\mathrm{d}y \quad (\text{As } p(x,y) = p(x\mid y)\,p(y)) \\
&= \int\Big(\int x\,p(x\mid y)\,\mathrm{d}x\Big)\,p(y)\,\mathrm{d}y \\
&= \int \mathbf{E}_x[\,x\mid y\,]\,p(y)\,\mathrm{d}y \quad \Big(\text{As } \mathbf{E}_x[x] = \int x\,p(x)\,\mathrm{d}x\Big) \\
&= \mathbf{E}_y\big[\mathbf{E}_x[\,x\mid y\,]\big] \quad \Big(\text{As } \mathbf{E}_y[y] = \int y\,p(y)\,\mathrm{d}x\Big)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}_x[x] &= \mathbf{E}_x[x^2] - (\mathbf{E}_x[x])^2 \\
&= \mathbf{E}_y\big[\mathbf{E}_x[\,x^2\mid y\,]\big] - \Big(\mathbf{E}_y\big[\mathbf{E}_x[\,x\mid y\,]\big]\Big)^2 \quad \big(\text{As } \mathbf{E}[x] = \mathbf{E}_y\big[\mathbf{E}_x[\,x\mid y\,]\big]\big) \\
&= \mathbf{E}_y\big[\mathrm{var}_x[x\mid y] + \mathbf{E}_x[\,x\mid y\,]^2\big] - \Big(\mathbf{E}_y\big[\mathbf{E}_x[\,x\mid y\,]\big]\Big)^2 \quad (\text{As } \mathbf{E}_x[x^2] = \mathrm{var}_x[x] + (\mathbf{E}_x[x])^2) \\
&= \mathbf{E}_y\big[\mathrm{var}_x[x\mid y]\big] + \mathbf{E}_y\big[\mathbf{E}_x[\,x\mid y\,]^2\big] - \Big(\mathbf{E}_y\big[\mathbf{E}_x[\,x\mid y\,]\big]\Big)^2 \\
&= \mathbf{E}_y\big[\mathrm{var}_x[x\mid y]\big] + \mathrm{var}_y[\mathbf{E}_x[x\mid y]] \quad (\text{As } \mathrm{var}_y[y] = \mathbf{E}_y[y^2] - (\mathbf{E}_y[y])^2)
\end{aligned}
$$