A. Starting from the "standard" form of each PDF/PMF, show that the following distributions are in an exponential family, and find the corresponding b, c, θ, and $a(\phi)$.

(i) $Y \sim \mathbf{N}(\mu, \sigma^2)$ **for known** σ^2

Let's begin by writing the PDF for the normal distribution:

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right\} \cdot \exp\left\{\log\left((2\pi\sigma^2)^{-1/2}\right)\right\} \\ &= \exp\left\{\frac{y^2}{-2\sigma^2} + y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)\right\} \end{aligned}$$

By letting $\theta = \mu, a(\phi) = \sigma^2, b(\theta) = \frac{1}{2}\mu^2$, and $c(y|\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$, we can write the normal distribution with fixed variance in the form of an exponential family:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y|\phi)\right\}$$

(ii) $Y = Z/N$**, where** $Z \sim \mathbf{Binom}(N, P)$ **for known** N

We can easily obtain the PMF of Y through a transformation of random variables:

$$\begin{aligned} P\left(Y = \frac{Z}{N}\right) &= P\left(Z = NY\right) \cdot \left|\frac{1}{N}\right| \\ &= \binom{N}{NY} P^{NY}(1-P)^{N-NY} \cdot \frac{1}{N} \\ &= \exp\left\{\log\left[\binom{N}{NY} P^{NY}(1-P)^{N-NY} \cdot \frac{1}{N}\right]\right\} \\ &= \exp\left\{\log\left[\binom{N}{NY}\right] + NY\log(P) + N\log(1-P) - NY\log(1-P) - \log(N)\right\} \\ &= \exp\left\{Y\left[N\log\left(\frac{P}{1-P}\right)\right] - N\log\left(\frac{1}{1-P}\right) + \log\left[\binom{N}{NY}\right] - \log(N)\right\} \end{aligned}$$

Let $\theta = N\log\left(\frac{P}{1-P}\right), b(\theta) = N\log\left(\frac{1}{1-P}\right), a(\phi) = 1$, and $c(y|\phi) = \log\left[\binom{N}{NY}\right] - \log(N)$ to get the form of an exponential family.

(iii) $Y \sim \textbf{Pois}(\lambda)$
You know the drill!

$$\begin{aligned} f(y|\lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \\ &= \exp\left\{\log\left[\frac{e^{-\lambda}\lambda^y}{y!}\right]\right\} \\ &= \exp\left\{-\lambda + y\log(\lambda) - \log(y!)\right\} \\ &= \exp\left\{y\log(\lambda) - \lambda + (-\log(y!))\right\} \end{aligned}$$

The above is in the desired exponential family form since we can let $\theta = \log(\lambda), a(\phi) = 1, b(\theta) = \lambda$, and $c(y|\phi) = -\log(y!)$. Therefore, we have shown that we can write the PDF of Y in the desired exponential family form.

B. We want to characterize the mean and variance of a distribution in the exponential family. To do this, we'll take an unfamiliar route, involving a preliminary lemma. Define the score $s(\theta)$ as the gradient of the log-likelihood with respect to θ:

$$s(\theta) = \frac{\partial}{\partial\theta} \log L(\theta).$$

While we think of the score as a function of θ, clearly the score also depends on the data. So a natural question is: what can we say about the distribution of the score over different random realizations of the data under the true data-generating process, i.e., at the true θ? It turns out we can say the following, sometimes referred to as the score equations:

$$\begin{aligned} E\left[s(\theta)\right] &= 0 \\ \mathcal{I}(\theta) \equiv \textbf{var}(s(\theta)) &= -E\left[H(\theta)\right], \end{aligned}$$

where the mean and variance are taken under the true θ. Prove these score equations.

First, we prove $E\left[s(\theta)\right] = 0$. Note that while the score is a function of θ, it's also dependent on the data y. Therefore, we can take the expected value of the score over the sample space $\mathcal{Y}$. Let's write out the form of $E\left[s(\theta)\right]$:

$$\begin{aligned} E\left[s(\theta)\right] &= \int_{\mathcal{Y}} s(\theta) f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta} \log L(\theta) \cdot f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial\theta} L(\theta)}{L(\theta)} \cdot f(y|\theta) dy \end{aligned}$$

Now, we use the statistical trick that we can rewrite the likelihood function as a PDF, since we integrate over $\mathcal{Y}$ with PDF $f(y|\theta)$:

$$\begin{aligned} E\left[s(\theta)\right] &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial\theta} f(y|\theta)}{f(y|\theta)} \cdot f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta} f(y|\theta) dy \\ &= \frac{\partial}{\partial\theta} \int_{\mathcal{Y}} f(y|\theta) dy \\ &= \frac{\partial}{\partial\theta} (1) = 0, \end{aligned}$$

where we assume that any necessary technical conditions are met to be able to switch the order of integration and differentiation.

Now, we prove that $\text{var}(s(\theta)) = -E\left[H(\theta)\right]$. Using the provided hint, suppose we differentiate the first equation with respect to θ^T:

$$\begin{aligned}
\frac{\partial}{\partial\theta^T} E(s(\theta)) &= \frac{\partial}{\partial\theta^T}\int_{\mathcal{Y}} \frac{\partial}{\partial\theta}\log L(\theta) f(y|\theta)dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta^T}\left[\frac{\partial}{\partial\theta}\log L(\theta) f(y|\theta)\right] dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta}\log L(\theta)\cdot\frac{\partial}{\partial\theta^T} f(y|\theta) + f(y|\theta)\frac{\partial^2}{\partial\theta^T\theta}\log L(\theta)dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta}\log L(\theta)\cdot\frac{\partial}{\partial\theta^T} L(\theta)dy + \int_{\mathcal{Y}} f(y|\theta)\frac{\partial^2}{\partial\theta^T\theta}\log L(\theta)dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta}\log L(\theta)\cdot\frac{\partial}{\partial\theta^T}\log L(\theta)\cdot f(y|\theta)dy + E\left[\frac{\partial^2}{\partial\theta^T\theta}\log L(\theta)\right] \\
&= E\left[\frac{\partial}{\partial\theta}\log L(\theta)\cdot\frac{\partial}{\partial\theta^T}\log L(\theta)\right] + E\left[H(\theta)\right] \\
&= E\left[s(\theta)s(\theta)^T\right] + E\left[H(\theta)\right] \\
&\overset{set}{=} \frac{\partial}{\partial\theta^T}(0) = 0
\end{aligned}$$

Before we get to the big reveal, let's acknowledge the nice property that we used to obtain the fifth equality:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log L(\theta) &= \frac{\frac{\partial}{\partial\theta}L(\theta)}{f(y|\theta)} \\
\implies \frac{\partial}{\partial\theta}L(\theta) &= \frac{\partial}{\partial\theta}\log L(\theta)\cdot f(y|\theta)
\end{aligned}$$

Now, we see from the above that

$$\begin{aligned}
\text{var}\left[s(\theta)\right] &= E[s(\theta)s(\theta)^T] - (E[s(\theta)])^2 \\
&= E[s(\theta)s(\theta)^T] \\
&= -E[H(\theta)]
\end{aligned}$$

C. Use the score equations you just proved to show that, if $Y \sim f(y|\theta,\phi)$ is an exponential family, then

$$E(Y) = b'(\theta)$$
$$\textbf{var}(Y) = a(\phi)b''(\theta)$$

Thus, the variance of Y is a product of two terms: $b''(\theta)$ depends only on the canonical parameter θ, and hence on the mean, since we showed that $E(Y) = b'(\theta)$; $a(\phi)$ is independent of θ. Note that the most common form of a is $a(\phi) = \phi/w$ where ϕ is called a dispersion parameter and w is a known prior weight that can vary from one observation to another.

Recall that $E[s(\theta)] = E\left[\frac{\partial}{\partial\theta}\log L(\theta)\right] = 0$. Assume without loss of generality that there are n observations. For exponential families, we know that the log-likelihood is

$$\begin{aligned}\log L(\theta) &= \log\left[\prod_{i=1}^{n}\exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i|\phi)\right\}\right]\\ &= \sum_{i=1}^{n}\left[\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i|\phi)\right]\\ &= \frac{\theta}{a(\phi)}\sum_{i=1}^{n}y_i - \frac{nb(\theta)}{a(\phi)} + \sum_{i=1}^{n}c(y_i|\phi)\end{aligned}$$

By taking the expectation of the gradient of the log-likelihood with respect to θ, we obtain the following:

$$\begin{aligned}E[s(\theta)] &= \int_{\mathcal{Y}}\frac{\partial}{\partial\theta}\left[\frac{\theta}{a(\phi)}\sum_{i=1}^{n}y_i - \frac{nb(\theta)}{a(\phi)} + \sum_{i=1}^{n}c(y_i|\phi)\right]f(y|\theta)dy\\ &= \int_{\mathcal{Y}}\left[\frac{1}{a(\phi)}\sum_{i=1}^{n}y_i - \frac{nb'(\theta)}{a(\phi)}\right]f(y|\theta)dy\\ &= E\left[\frac{\sum_{i=1}^{n}y_i}{a(\phi)} - \frac{nb'(\theta)}{a(\phi)}\right]\\ &= \frac{1}{a(\phi)}\sum_{i=1}^{n}E(Y) - \frac{nb'(\theta)}{a(\phi)}\\ &\overset{set}{=} 0\end{aligned}$$

By manipulating the above equations, we get

$$E(Y) = b'(\theta)$$

Now, let's obtain the variance of Y:

$$\begin{aligned}
\text{var}(s(\theta)) &= \text{var}\left[\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right] \\
&= \frac{1}{a(\phi)^2}\sum_{i=1}^{n}\text{var}(Y) \\
&\overset{set}{=} -E\left[H(\theta)\right]
\end{aligned}$$

Note that

$$\begin{aligned}
-E[H(\theta)] &= -E\left[\frac{\partial}{\partial\theta^T}\left(\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right)\right] \\
&= -\int_{\mathcal{Y}}\frac{\partial}{\partial\theta^T}\left(\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right) f(y|\theta)dy \\
&= \int_{\mathcal{Y}}\frac{nb''(\theta)}{a(\phi)} f(y|\theta)dy \\
&= E\left[\frac{nb''(\theta)}{a(\phi)}\right] \\
&= \frac{nb''(\theta)}{a(\phi)}
\end{aligned}$$

By combining the two above derivations, we see that

$$\begin{aligned}
\frac{1}{a(\phi)^2}\sum_{i=1}^{n}\text{var}(Y) &= \frac{nb''(\theta)}{a(\phi)} \\
\implies \text{var}(Y) &= a(\phi)b''(\theta)
\end{aligned}$$

D. To convince yourself that your result in (C) is correct, use these results to compute the mean and variance of the N(μ, σ^2) distribution.

Recall from (a) that $\theta = \mu, a(\phi) = \sigma^2$, and $b(\theta) = \frac{1}{2}\mu^2$. While (a) *did* assume that σ^2 was known, we see that the result still holds! From (c), we found that

$$\begin{aligned} E(Y) &= b'(\theta) \\ &= \frac{\partial}{\partial \mu}\left(\frac{1}{2}\mu^2\right) \\ &= \mu, \\ \mathrm{var}(Y) &= a(\phi)b''(\theta) \\ &= \sigma^2 \frac{\partial^2}{\partial \mu^2}\left(\frac{1}{2}\mu^2\right) \\ &= \sigma^2 \frac{\partial}{\partial \mu}(\mu) \\ &= \sigma^2, \end{aligned}$$

which are certainly the mean and variance of a N(μ, σ^2) distribution.

A. Deduce from your results above that, in a GLM,

$$\theta_i = (b')^{-1}\left(g^{-1}(x_i^T\beta)\right),$$
$$\mathbf{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i)$$

for some function V that you should specify in terms of the building blocks of the exponential family model. V is often referred to as the variance function, since it explicitly relates the mean and the variance in a GLM.

Let's start with proving the first equation. Recall that $E(Y_i) = b'(\theta_i)$ and, by definition of GLM, $E(Y_i) = \mu_i$. Additionally, by definition, $g(\mu_i) = x_i^T\beta$. Therefore, we can simply equate these equations in the following way:

$$\begin{aligned}
E(Y_i) &= b'(\theta_i) = \mu_i \\
\mu_i &= g^{-1}(x_i^T\beta), \\
\implies b'(\theta_i) &= g^{-1}(x_i^T\beta), \\
\implies \theta_i &= (b')^{-1}\left(g^{-1}(x_i^T\beta)\right),
\end{aligned}$$

which is the first equation.

Now, we want to prove the second equation, containing the variance of Y_i. Recall that $\text{var}(Y_i) = a(\phi)b''(\theta)$. In the formulation of the GLM, we see that $a(\phi) = \frac{\phi}{w_i}$. Therefore,

$$\begin{aligned}
\text{var}(Y_i) &= \frac{\phi}{w_i}b''(\theta_i) \\
&= \frac{\phi}{w_i}b''\left((b')^{-1}\left(g^{-1}(x_i^T\beta)\right)\right) \\
&= \frac{\phi}{w_i}b''\left((b')^{-1}(\mu_i)\right)
\end{aligned}$$

By letting $V(\mu_i) = b''\left((b')^{-1}(\mu_i)\right)$, we get the second equation. Notice how we wrote $V(\mu_i)$ as a function of μ_i using functions of the exponential family (i.e., function $b(\cdot)$).

B. Take two special cases. (1) Suppose that Y is a Poisson GLM, i.e., that the stochastic component of the model is a Poisson distribution. Show that $V(\mu) = \mu$. (2) Suppose that $Y = Z/N$ is a Binomial GLM, i.e., that the stochastic component of the model is a Binomial distribution $Z \sim$ Binom(N, P) and that Y is the fraction of yes outcomes. Show that $V(\mu) = \mu(1-\mu)$.

First special case:

We can always write the stochastic component of the model in the following form:

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i|\phi/w_i)\right\},$$

which resembles an exponential family where $a(\phi) = \frac{\phi}{w_i}$. Note that we can write the Poisson PMF as

$$\begin{aligned}
f(y_i|\lambda_i) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\
&= \exp\left\{\log\left[\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right]\right\} \\
&= \exp\left\{-\lambda_i + y_i\log(\lambda_i) - \log(y_i!)\right\},
\end{aligned}$$

which is an exponential family where $\theta_i = \log(\lambda_i)$, $b(\theta_i) = \lambda_i = e^{\theta_i}$, $a(\phi) = \frac{\phi}{w_i} = 1$, and $c(y_i|\phi/w_i) = -\log(y_i!)$. Recall that, for an exponential family,

$$\begin{aligned}
E(Y_i) &= b'(\theta_i), \\
\text{var}(Y_i) &= a(\phi)b''(\theta_i),
\end{aligned}$$

so we easily obtain

$$\begin{aligned}
E(Y_i) &= b'(\theta_i) \\
&= e^{\theta_i} \\
&= \lambda_i, \\
\text{var}(Y_i) &= a(\phi)b''(\theta) \\
&= \frac{\partial}{\partial\theta_i}(e^{\theta_i}) \\
&= \lambda_i,
\end{aligned}$$

By definition, $\mu_i = E(Y_i)$, so $\mu_i = \lambda_i$. Now, we set $V(\mu_i) = b''(\theta)$ to satisfy $\text{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i)$. Therefore, $V(\mu_i) = \mu_i$.

Second special case:

If the stochastic component is $Z \sim \text{Binom}(N, P)$, then we can write

$$f(z_i|\theta_i, \phi_i) = \exp\left\{\frac{z_i\theta_i - b(\theta_i)}{\phi/w_i} + c(z_i|\phi/w_i)\right\}$$

Note that since Z follows a binomial distribution, we can write its PMF as

$$\begin{aligned} f(z_i|\theta_i, \phi_i) &= \binom{N}{z_i} P^{z_i}(1-P)^{N-z_i} \\ &= \exp\left\{\log\left[\binom{N}{z_i} P^{z_i}(1-P)^{N-z_i}\right]\right\} \\ &= \exp\left\{z_i \log(P) - z_i \log(1-P) + N\log(1-P) + \log\left[\binom{N}{z_i}\right]\right\} \end{aligned}$$

Note that we can only write this PMF in the form of an exponential family for the GLM when N is known. So, let's assume N is known to show the PMF of z_i can be written in the desired exponential family form:

$$\begin{aligned} f(z_i|\theta_i, \phi_i) &= \exp\left\{z_i \log(P) - z_i \log(1-P) + N\log(1-P) + \log\left[\binom{N}{z_i}\right]\right\} \\ &= \exp\left\{z_i \log\left(\frac{P}{1-P}\right) - N\log\left(\frac{1}{1-P}\right) + \log\left[\binom{N}{z_i}\right]\right\}, \end{aligned}$$

where $a(\phi) = \frac{\phi}{w_i} = 1$, $\theta_i = \log\left(\frac{P}{1-P}\right)$, $b(\theta_i) = N\log\left(\frac{1}{1-P}\right)$, and $c(z_i|\phi/w_i) = \log\left[\binom{N}{z_i}\right]$. Once again, we find the form of $V(\mu_i)$ by setting it equal to $b''(\theta_i)$. Note that

$$\begin{aligned} \theta_i &= \log\left(\frac{P}{1-P}\right) \\ \implies e^{\theta_i} &= \frac{P}{1-P} \\ \implies e^{\theta_i} &= P + Pe^{\theta_i} \\ \implies P(1+e^{\theta_i}) &= e^{\theta_i} \\ \implies P &= \frac{e^{\theta_i}}{1+e^{\theta_i}} \end{aligned}$$

Knowing P, we can obtain the proper form for $b(\theta_i)$:

$$b(\theta_i) = N\log\left(1+e^{\theta_i}\right)$$

Now, we can obtain $b''(\theta_i)$:

$$\begin{aligned} b'(\theta_i) &= \frac{N}{1+e^{\theta_i}} \cdot e^{\theta_i} \\ b''(\theta_i) &= \frac{N}{1+e^{\theta_i}} \cdot e^{\theta_i} - \left(e^{\theta_i}\right)^2 \frac{N}{(1+e^{\theta_i})^2} = \frac{Ne^{\theta_i}}{(1+e^{\theta_i})^2} \end{aligned}$$

We can find μ_i with

$$\begin{aligned}
E(Z_i) &= b'(\theta_i) \\
&= \frac{N}{1+\frac{P}{1-P}} \cdot \frac{P}{1-P} \\
&= (1-P)N \cdot \frac{P}{1-P} \\
&= NP \\
&\overset{set}{=} \mu,
\end{aligned}$$

which we set to μ by definition of the GLM. Now, we can find $b''(\theta_i)$ in terms of $\mu = NP$:

$$\begin{aligned}
b''(\theta_i) &= NP - \left(\frac{P}{1-P}\right)^2 (1-P)^2 N \\
&= NP - NP^2 \\
&= NP(1-P)
\end{aligned}$$

Now, notice that all of the above work was in terms of $z_i \sim \text{Binom}(N, P)$; however, this question asks us to show that $V(\mu) = b''(\theta_i) = \mu(1-\mu)$ for $Y = \frac{Z}{N}$. We can simply divide $b''(\theta_i)$ by N to obtain the corresponding $b''(\theta_i)$ for Y. Therefore, $V(\mu) = P(1-P) = \mu(1-\mu)$.

C. To specify a GLM, we must choose the link function $g(\mu_i)$. Recall that g links the predictors with the mean of the response: $g(\mu_i) = x_i^T\beta$. Since you've shown that

$$\theta_i = (b')^{-1}\left\{g^{-1}(x_i^T\beta)\right\},$$

a "simple" choice of link function is one where $g^{-1} = b'$. This is known as the canonical link, in which case the canonical parameter simplifies to $\theta_i = x_i^T\beta$. So under the canonical link $g(\mu) = (b')^{-1}(\mu)$, we have the model

$$f(y_i|\beta,\phi) = \exp\left\{\frac{y_i x_i^T\beta - b(x_i^T\beta)}{\phi/w_i} + c(y_i|\phi/w_i)\right\}.$$

Now return to the two special cases from the previous problem and find the canonical link $g(\mu)$.

First special case:
Recall that $b(\theta_i) = e^{\theta_i}$. We want to find $g(\mu)$ that satisfies $g(\mu) = (b')^{-1}(\mu)$. Note that $b'(\theta_i) = e^{\theta_i}$, so

$$\begin{aligned} g(\mu) &= (b')^{-1}(\mu) \\ \implies b'(g(\mu)) &= \mu \\ \implies e^{g(\mu)} &= \mu \\ \implies g(\mu) &= \log(\mu) \end{aligned}$$

Therefore, the canonical link is the log link, i.e., $g(\mu) = \log(\mu)$.
Second special case:
In this case, we found that

$$\begin{aligned} b(\theta_i) &= \log\left(1 + e^{\theta_i}\right), \\ b'(\theta_i) &= \frac{1}{1+e^{\theta_i}} \cdot e^{\theta_i}. \end{aligned}$$

We want to find $g(\mu)$ that satisfies $g(\mu) = (b')^{-1}(\mu)$:

$$\begin{aligned} b'(g(\mu)) &= \mu \\ \implies \frac{e^{g(\mu)}}{1+e^{g(\mu)}} &= \mu \\ \implies e^{g(\mu)} &= \mu + \mu e^{g(\mu)} \\ \implies e^{g(\mu)} &= \frac{\mu}{1-\mu} \end{aligned}$$

Therefore, our canonical link is $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

A. Using the chain rule

$$\frac{\partial}{\partial \beta} = \frac{\partial}{\partial \theta} \cdot \frac{\partial \theta}{\partial \mu} \cdot \frac{\partial \mu}{\partial \beta},$$

show that

$$s(\beta, \phi) \equiv \nabla_\beta \log L(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)},$$

where x_i is the vector of predictors for case i.

First, let's solve for $\frac{\partial \log(L)}{\partial \theta_i}$, $\frac{\partial \theta_i}{\partial \mu_i}$, and $\frac{\partial \mu_i}{\partial \beta}$:

$$\begin{aligned}
\frac{\partial \log(L)}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i}\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i|\phi/w_i)\right] \\
&= \frac{y_i - b'(\theta_i)}{\phi/w_i}, \\
\frac{\partial \theta_i}{\partial \mu_i} &= \frac{\partial}{\partial \mu_i}\left[(b')^{-1}(\mu_i)\right] \\
&= \left((b')^{-1}\right)'(\mu_i) \\
&= \frac{1}{b''\left\{(b')^{-1}(\mu_i)\right\}}, \\
\frac{\partial \mu_i}{\partial \beta} &= \frac{\partial}{\partial \beta}\left[g^{-1}(x_i^T\beta)\right] \\
&= (g^{-1})'(x_i^T\beta)x_i^T \\
&= \frac{x_i^T}{g'\left\{g^{-1}(x_i^T\beta)\right\}},
\end{aligned}$$

where we use the fact that $(g^{-1})'(x) = \frac{1}{g'\{g^{-1}(x)\}}$. Now, we multiply these three terms together to obtain $s_i(\beta, \phi) = \frac{\partial l}{\partial \beta}$:

$$\begin{aligned}
s_i(\beta, \phi) &= \left(\frac{y_i - b'(\theta_i)}{\phi/w_i}\right) \cdot \frac{1}{b''\left\{(b')^{-1}(\mu_i)\right\}} \cdot \frac{x_i^T}{g'\left\{g^{-1}(x_i^T\beta)\right\}} \\
&= \left(\frac{y_i - \mu_i}{\phi/w_i}\right) \cdot \frac{1}{b''\left\{\theta_i\right\}} \cdot \frac{x_i^T}{g'\left\{\mu_i\right\}} \\
&= \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)}
\end{aligned}$$

Therefore, we have shown that

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)}$$

B. Show that under the canonical link, $g'(\mu) = 1/V(\mu)$, the score function simplifies to

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi}.$$

Recall from (A) of this section that we were able to write the score as

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i) g'(\mu_i)}$$

By substituting $g'(\mu_i)$ with $\frac{1}{V(\mu_i)}$, we easily obtain the desired result:

$$\begin{aligned} s(\beta, \phi) &= \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi V(\mu_i)/V(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i}{\phi} \end{aligned}$$

I assume that this question was actually asking us to show $g'(\mu) = 1/V(\mu)$:

$$\begin{aligned} V(\mu_i) &= b''(\theta_i) \\ &= b''((b')^{-1}(\mu_i)) \\ &= (g^{-1})'(g(\mu_i)) \\ &= \frac{1}{g'(g^{-1}(g(\mu_i)))} \\ &= \frac{1}{g'(\mu_i)} \\ \implies g'(\mu) = 1/V(\mu) & \end{aligned}$$

C. Let's take the specific case of a GLM for a binomial outcome, where $Y_i \sim$ Binom(N_i, μ_i) for known sample size N_i, where $Y_i = Z_i/N_i$ is the observed success fraction, and where μ_i is related to the predictors $x_i \in \mathcal{R}^p$ via the canonical logistic link. This is called the logistic regression model. Write your own function that will fit a logistic regression model by gradient descent. Try to maintain some level of generality to your code, i.e., so that it could also work with different GLMs. Use the "wdbc.csv" dataset from the course website, and use the first 10 features for X.

In a gradient descent algorithm, we can iteratively learn about our parameter of interest (β) by moving along the gradient of the log-likelihood. Suppose we denote the step-size of our gradient descent algorithm as γ. Then, we can update β with the following:

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} + \gamma \nabla_\beta \log L(\beta) \\ &= \beta^{(m)} + \gamma s(\beta, \phi),\end{aligned}$$

where we learned how to efficiently calculate $s(\beta, \phi)$ in (A). In this problem, we use a familiar binomial GLM, where $\phi/w_i = 1$ and the link function is the canonical logistic link. Therefore, our score simplifies to

$$\begin{aligned}s(\beta, \phi) &= \sum_{i=1}^{n} (Y_i - \mu_i) x_i \\ &= \sum_{i=1}^{n} (Y_i - g^{-1}(x_i^T \beta)) x_i\end{aligned}$$

The problem statement specifies that we use the canonical logistic link function, i.e. $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$. Therefore, the inverse of our link function is

$$g^{-1}(x_i^T \beta) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}},$$

which gives us

$$s(\beta, \phi) = \sum_{i=1}^{n} \left[Y_i - \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \right] x_i$$

Now, we can program a function in *R* that obtains the score.

To implement gradient descent, we also need a proper way to obtain γ. While we can specify some fixed value for γ, it may be best to iteratively obtain γ using a line search. Therefore, we can obtain the optimal γ by minimizing the

Algorithm 1 Gradient Descent Algorithm for Logistic Regression Model

```
Read in and scale data $X, y$
Initialize $\log L(\beta) = -100000$, $\beta = \text{rep}(0.1, \text{ncol}(X))$ and conv=FALSE
Set tolerance at $1e - 5$
while !conv do
    Set $g = 0$
    for i in 1:nrow(X) do
        Add $g = g + \left[ Y_i - \frac{e^{x_i^T \beta^{(i-1)}}}{1+e^{x_i^T \beta^{(i-1)}}} \right] x_i$
    end for
    Obtain $\gamma$ by minimizing the negative log-likelihood at $\beta^{(i-1)}$
    Calculate $\beta^{(i)} = \beta^{(i-1)} + \gamma g$
    Compute and store log-likelihood at $X\beta^{(i)}$
    if $|\text{log-likelihood}^{(i)} - \text{log-likelihood}^{(i-1)}| <$ tolerance then
        Set conv=TRUE
    end if
end while
```

negative log-likelihood given the current value of β, which can be done with the *optim()* function. The steps for this algorithm are reported in Algorithm 1.

After implementing the above algorithm, we obtain a vector of log-likelihood values; Figure 1 is a trace plot of these log-likelihood values. The red line in the plot is the log-likelihood reported by the *glm()* function in *R*. Using the *glm()* function with a binomial family as a means of comparison, we see that the above gradient descent algorithm works very well.

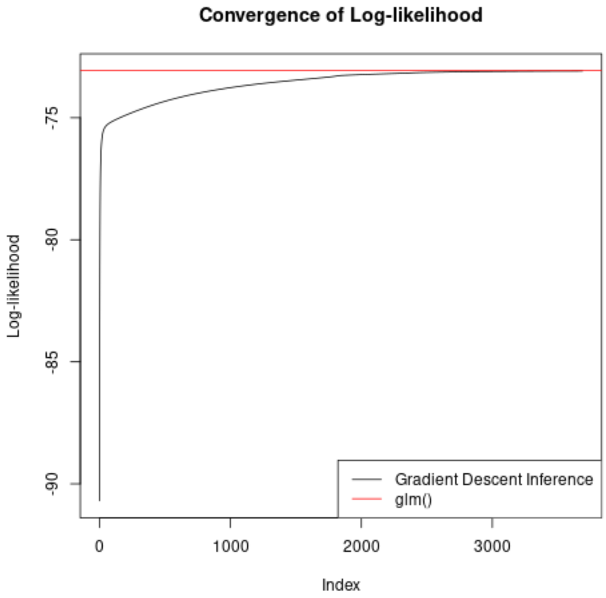

Figure 1: Log-likelihood values from Gradient Descent

D. Consider the Hessian matrix, i.e. the matrix of partial second derivatives of the log-likelihood

$$H(\beta, \phi) = \frac{\partial^2}{\partial\beta\partial\beta^T} \log L(\beta, \phi).$$

Give an expression for the Hessian matrix $H(\beta, \phi)$ of a GLM that is as simple as possible, ideally in matrix form. Assume the canonical link.

By (A) and (B),

$$\frac{\partial}{\partial\beta} \log L(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi}$$

From here, we can take the partial derivative of the above with respect to β^T:

$$\begin{aligned}
\frac{\partial}{\partial\beta^T} \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi} &= \frac{\partial}{\partial\beta^T} \sum_{i=1}^{n} \frac{w_i(Y_i - g^{-1}(x_i^T\beta))x_i}{\phi} \\
&= -\frac{1}{\phi} \sum_{i=1}^{n} w_i \frac{\partial}{\partial\beta^T} \left[g^{-1}(x_i^T\beta)\right] x_i \\
&= -\frac{1}{\phi} \sum_{i=1}^{n} w_i (g^{-1})'(x_i^T\beta) x_i^T x_i \\
&= -\frac{1}{\phi} \sum_{i=1}^{n} w_i b''(x_i^T\beta) x_i^T x_i
\end{aligned}$$

Suppose we want to write the above expression in matrix notation. First, we consider the following lemma.
Lemma: $A_{n\times p}B_{p\times k} = \sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i^T$, where $\mathbf{a}_i$ and $\mathbf{b}_i$ denote the columns of A, B.

From the above lemma, we learned that we can write the product of matrices as the sum of column i in A times row i in B; this is known as the *outer product*. Now, we can apply this iteratively to obtain our answer. Note that x_i denotes the ith row of matrix X. Switching to our notation in the problem,

$$AB = \sum_{i=1}^{n} a_i^T b_i,$$

since a_i^T denotes the column of A and b_i denotes the row of B. Suppose we denote $W = \text{diag}\left(\frac{w_i}{\phi} b''(x_i^T\beta)\right)$. Consider $X^T W X$. Using the above lemma, we obtain

$$\begin{aligned} X^T W X &= \left(\sum_{i=1}^{n} x_i \frac{w_i}{\phi} b''(x_i^T \beta) \right) X \\ &= \sum_{i=1}^{n} \left(x_i \frac{w_i}{\phi} b''(x_i^T \beta) \right)^T x_i \\ &= \sum_{i=1}^{n} \frac{w_i}{\phi} b''(x_i^T \beta) x_i^T x_i \end{aligned}$$

Therefore,

$$H(\beta, \phi) = -X^T W X$$

E. Now, consider a point $\beta_0 \in \mathbb{R}^p$, which serves as an intermediate guess for our vector of regression coefficients. Show that, for any GLM, the second-order Taylor approximation of $\mathcal{L}(\beta) = \log L(\beta, \phi)$ around the point β_0 can be expressed in the form

$$q(\beta; \beta_0) = -\frac{1}{2}(\tilde{y} - X\beta)^T W(\tilde{y} - X\beta) + c,$$

where $\tilde{y}$ is a vector of "working responses" and W is a diagonal matrix of "working weights". Give explicit expressions for the diagonal elements W_{ii} and for $\tilde{y}$, which will necessarily involve the point β_0 around which you're doing the expansion. Again, we're assuming the canonical link to make the algebra simpler.

Our second-order Taylor approximation is

$$\begin{aligned} q(\beta; \beta_0) &= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T|_{\beta=\beta_0}(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^T H(\beta_0)(\beta - \beta_0) \\ &= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T|_{\beta=\beta_0}(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T W X(\beta - \beta_0), \end{aligned}$$

where we use the Hessian from (D). Now, let's find $\nabla_\beta \mathcal{L}(\beta)^T$:

$$\begin{aligned} \nabla_\beta \mathcal{L}(\beta)^T &= \sum_{i=1}^n \frac{w_i(y_i - \mu_i)x_i^T}{\phi} \\ &= \sum_{i=1}^n \frac{w_i b''(x_i^T\beta)}{\phi} \cdot \frac{(y_i - \mu_i)x_i^T}{b''(x_i^T\beta)} \\ &= \tilde{z}^T W X, \end{aligned}$$

where $\tilde{z} = \left[\dots \frac{y_i - \mu_i}{b''(x_i^T\beta)} \dots \right]$. Now, we revisit the second-order Taylor approximation:

$$\begin{aligned} q(\beta; \beta_0) &= \mathcal{L}(\beta_0) + \nabla_\beta \mathcal{L}(\beta)^T|_{\beta=\beta_0}(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T W X(\beta - \beta_0) \\ &= \mathcal{L}(\beta_0) + \tilde{z}_{\beta_0}^T W X(\beta - \beta_0) - \frac{1}{2}(\beta - \beta_0)^T X^T W X(\beta - \beta_0) \\ &= \tilde{z}_{\beta_0}^T W X\beta - \frac{1}{2}\beta^T X^T W X\beta + \beta_0^T X^T W X\beta + c^* \\ &= -\frac{1}{2}\beta^T X^T W X\beta + \tilde{y}^T W X\beta + c^* \\ &= -\frac{1}{2}(\tilde{y} - X\beta)^T W(\tilde{y} - X\beta) + c, \end{aligned}$$

where $\tilde{y}^T = \beta_0^T X^T + \tilde{z}_{\beta_0}^T = \beta_0^T X^T + \left[\dots \frac{y_i - \mu_i|_{\beta=\beta_0}}{b''(x_i^T\beta_0)} \dots \right]$ and $W_{ii} = \frac{w_i}{\phi} b''(x_i^T\beta_0)$.

F. Read up on Newton's method for optimizing smooth functions (c.f. Nocedal and Wright, Chapter 2). Implement it for the logistic regression model and test it out on the same data set you just used to test out the gradient descent. Note: while you could do line search, there is a natural step size of 1 in Newton's method. Verify that your solution replicates the β estimates you get when using a program solver.

When using Newton's method, we update β with the following:

$$\beta^* = \beta + H^{-1}_{\mathcal{L}(\hat{\beta})} \nabla \mathcal{L}(\beta),$$

where H is the Hessian matrix we found in (D) (i.e. $H = -X^TWX$) and $\nabla\mathcal{L}(\beta) = s(\beta, \phi)$ is the score function. Our template for Newton's method can be found in Algorithm 2. Figure 2 displays the relatively quick convergence of the log-likelihood to that found using *glm()* in *R*.

Algorithm 2 Newton's Method for Logistic Regression Model

```
Read in and scale data $X, y$
Initialize $\log L(\beta) = -100000$, $\beta = \text{rep}(0.1, \text{ncol}(X))$ and conv=FALSE
Set tolerance at $1e-5$
while !conv do
  Set $g = 0$
  for i in 1:nrow(X) do
    Add $g = g + \left[Y_i - \frac{e^{x_i^T\beta^{(i-1)}}}{1+e^{x_i^T\beta^{(i-1)}}}\right] x_i$
  end for
  Calculate $H = -X^TWX$
  Calculate $\beta^{(i)} = \beta^{(i-1)} + H^{-1}g$
  Compute and store log-likelihood $\mathcal{L}(\beta^{(i)})$
  if $|\mathcal{L}(\beta^{(i)}) - \mathcal{L}(\beta^{(i-1)})|/|\mathcal{L}(\beta^{(i-1)})| <$ tol then
    Set conv=TRUE
  end if
end while
```

Below is a table comparing the results from my implementation of Newton's method and the results from *glm()*. With these minor discrepancies in so few iterations, it is natural to ask, "Why would I use gradient descent over Newton's method?" Let's take a look at the differences between these two methods:

1. Gradient descent is parametric according to the learning rate γ. If we hope to learn about some learning rate, then clearly gradient descent is the right selection. (While a parametric version of Newton's method exists, it is only applicable when we operate with a polynomial function with multiple roots.)

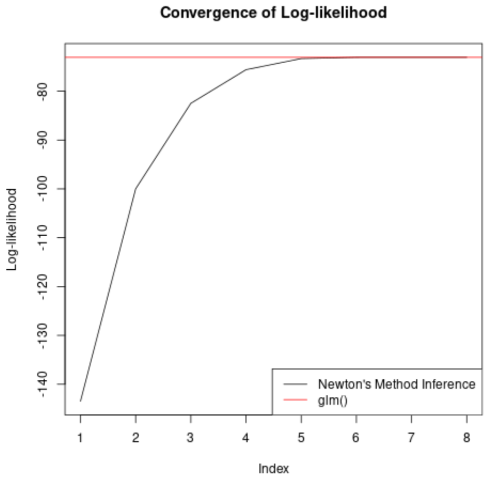

Figure 2: Convergence of Newton’s Method

2. Newton’s method requires second derivatives, whereas gradient descent can be applied using only the first derivative. Thus, Newton’s method has stronger constraints in terms of the differentiability of the function.

3. If gradient descent reaches a stationary point, it continues to run; however, parameters won’t update. Conversely, if Newton’s method hits a stationary point, then the algorithm will terminate due to division by zero.

Newton’s beta	**glm()’s beta**
0.48701671	0.48701675
-7.22184989	-7.22185053
1.65475612	1.65475615
-1.73763049	-1.73763027
14.00484503	14.00484560
1.07495327	1.07495329
-0.07723455	-0.07723455
0.67512312	0.67512313
2.59287422	2.59287426
0.44625630	0.44625631
-0.48248419	-0.48248420

G. Standard asymptotic theory implies that the maximum likelihood estimator is consistent and asymptotically normal around the true value β_0:

$$\hat{\beta} \sim N(\beta_0, \mathcal{I}(\beta_0, \phi)^{-1})),$$

where $\mathcal{I}(\beta_0, \phi)$ is called the *Fisher information matrix* and is defined as the variance of the score equations:

$$\mathcal{I}(\beta_0, \phi) \equiv \mathbf{var}(s(\beta_0, \phi)) = -E[H(\beta_0, \phi)]].$$

The fact that Fisher information is the negative of the expected Hessian motivates the following idea: use the inverse of the negative Hessian matrix at the MLE to approximate the inverse Fisher information, i.e. the covariance matrix of the estimator. Happily, you get this Hessian matrix for free when fitting by Newton's method.
For your logistic regression on the WDBC data fit via Newton's method, compute the square root of each diagonal element of the inverse Hessian matrix, evaluated at the MLE. Compare these to the standard errors you get when using a package solver.

The following table contains the standard errors arising from (1) the logistic regression model fit via Newton's method and (2) fit via *glm()* in *R*. The difference in standard errors is minimal.

Newton's standard errors	**glm()'s standard errors**
0.5642741	0.5643200
13.0939632	13.0949439
0.2775461	0.2775752
12.2742249	12.2749905
5.8903905	5.8909033
0.4493906	0.4494181
1.0742854	1.0743433
0.6473007	0.6473276
1.1069398	1.1070102
0.2914167	0.2914299
0.6040244	0.6040610