2. For the gradient boosting algorithm, we want to unpack step 2(c) of algorithm 10.3 (ESL) to derive the optimal value of the weights (γ_{jm}) for each leaf j at boosting step m. (The derivations in Chen and Guestrin (2016) or Bujokas (2022) may be clearer.)

a. Derive γ_{jm} for both the MSE (L_2 norm) and binomial deviance loss functions.

b. Do the same for *Newton boosting* (Chen and Guestrin 2016), where we use a second-order rather than a first-order approximation to the loss function.

a)

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^N L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm},\ j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.

$$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

For MSE:

$$f(\gamma) = \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma) \qquad \frac{1}{K}$$

$$= \sum_{x_i \in R_{jm}} [y_i - (f_{m-1}(x_i) + \gamma_{jm})]^2$$

Take derivative, we have

$$\frac{\partial f(\gamma)}{\partial \gamma_{jm}} = \frac{\partial}{\partial \gamma_{jm}} \sum_{x_i} \cancel{y_i^2} - 2y_i(\cancel{f_{m-1}(x_i)} + \gamma_{jm}) + (f_{m-1}(x_i) + \gamma_{jm})^2$$

$$= \sum_{x_i} \frac{\partial}{\partial \gamma_{jm}} (-2y_i)\gamma_{jm} + \cancel{f_{m-1}^2(x_i)} + 2f_{m-1}(x_i)\gamma_{jm} + \gamma_{jm}^2$$

$$= \sum_{x_i} -2y_i + 2f_{m-1}(x_i) + 2\gamma_{jm}$$

$$\frac{\partial f(\hat{\gamma})}{\partial \hat{\gamma}_{jm}} = 0 \Rightarrow 0 = \sum_{x_i} (-2\hat{y}_i + 2f_{m-1}(x_i) + 2\gamma_{jm})$$

$$\Rightarrow \sum_{x_i} \hat{y}_i = \sum_{x_i} f_{m-1}(x_i) + \sum_{x_i} \gamma_{jm}$$

$$\Rightarrow n\hat{y}_i = \sum_{x_i} (y_i - f_{m-1}(x_i))$$

$$\Rightarrow \hat{y}_i = \frac{1}{n}\sum_{x_i}(y_i - f_{m-1}(x_i))$$

Binomial deviance loss :

$$f(\gamma) = \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_j$$

$$= \sum_{x_i} -\log(1+e^{-2y_i(f_{m-1}(x_i)+\gamma})$$

$$\frac{\partial f(\gamma)}{\partial \gamma_{jm}} = \sum_{x_i} -2y_i \cdot \frac{-e^{2y_i(f_{m-1}(x_i)+\gamma_{jm})}}{1+e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})}}$$

$$\frac{\partial f(\hat{\gamma})}{\partial \hat{\gamma}_{jm}} = 0 \Rightarrow \sum_{x_i} 2y_i \cdot \frac{e^{2y_i(f_{m-1}(x_i)+\gamma_{jm})}}{1+e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})}} = 0$$

b) For Newton Boosting, we have $g_i = \frac{\partial f(\gamma)}{\partial(\gamma)} L(y_i, f_{m-1}(x_i)$

$h_i = \frac{\partial^2 f(\gamma)}{\partial \gamma^2} L(y_i, f_{m-1}(x_i))$

$$f(\gamma) = \sum_{x_i}(L(y_i, f_{m-1}(x_i)) + \frac{\partial L}{\partial f_{m-1}} L(y_i, f_{m-1}(x_i))\gamma_{jm} + \frac{1}{2}\frac{\partial^2}{\partial f_{m-1}^2} L(y_i, f_{m-1}(x_i))\gamma_{jm}^2$$

For MSE

Let $f(\gamma) = 0$, we have

$$g_i = \frac{\partial L}{\partial f_{m-1}} L(y_i, f_{m-1}(x_i))\gamma_{jm} = -2(y_i - f_{m-1}(x_i))$$

and

$$\frac{1}{2}h_i = \frac{1}{2}\frac{\partial^2}{\partial f_{m-1}^2} L(y_i, f_{m-1}(x_i))\gamma_{jm}^2 = 1$$

Thus: $$\hat{\gamma}_{jm} = \frac{\sum_{x_i} \cancel{2}(y_i - f_{m-1}(x_i))}{\sum_{x_i} \cancel{2}} = \frac{1}{n}\sum_{x_i} y_i - f_{m-1}(x_i)$$

For binomial deviance

$$g_i = 2y_i \cdot \frac{e^{2y_i(f_{m-1}(x_i)+\gamma_{jm})}}{1+e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})}}$$

$$\text{and } h_i = \frac{4y_i^2 e^{2y_i(f_{m-1}(x_i)+\gamma_{jm})} + 4y_i}{(1+e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})})^2}$$

By chen ...,

$$\hat{\gamma}_{jm} = -\frac{\sum_{x_i} g_i}{\sum_{x_i} h_i + \lambda} = \frac{\sum_{x_i} 2y_i \cdot \frac{e^{2y_i(f_{m-1}(x_i)+\gamma_{jm})}}{1+e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})}}}{\sum_{x_i} \frac{4y_i^2 e^{2y_i(f_{m-1}(x_i)+\gamma_{jm})} + 4y_i}{(1+e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})})^2} + \lambda}$$