(4)

a) In the matrix form, OLS objective is given by,

$$f(\beta) = \min_{\beta} \|y - X\beta\|_2^2$$

$$\nabla f(\beta) = -2X^Ty + 2X^TX\beta$$
$$= 2X^T(X\beta - y)$$

Gradient descent formula is given by

$$\beta^t = \beta^{t-1} - \eta \nabla f(\beta^{t-1})$$
$$= \beta^{t-1} - \eta(2X^T(X\beta^{t-1} - y))$$
$$= \beta^{t-1} - 2\eta X^T(X\beta^{t-1} - y)$$

b) Incase of Newton Raphson algorithm, iterative steps, is given by

$$\beta^t = \beta^{t-1} - (\nabla^2 f(\beta^{t-1}))^{-1} \nabla f(\beta^{t-1})$$

$$\nabla^2 f(\beta) = \nabla(2X^TX\beta - 2X^Ty)$$
$$= 2\nabla(X^TX\beta)$$
$$= 2X^TX$$

In general,
$\nabla_\beta(A\beta) = A$

$$\beta^t = \beta^{t-1} - (2X^TX)^{-1}(2X^TX\beta^{t-1} - X^Ty)$$
$$= \beta^{t-1} - 1/2(X^TX)^{-1} * 2(X^TX\beta^{t-1} - X^Ty)$$
$$= \beta^{t-1} - (X^TX^{-1})(X^TX)\beta^{t-1} + (X^TX)^{-1}X^Ty$$

$$\beta^{t} = \beta^{t-1} - \beta^{t-1} + (X^T X)^{-1} X^T y$$

we know that $A^{-1}A = 1$

$$\boxed{\beta^{t} = (X^T X)^{-1} X^T y} = \beta_{OLS}$$

Observation → Optimal β is independent of initial β value. It seems like it converges to OLS.