Ex. 1

See code and output at the end of this document.

Ex, 2

$$g_\theta = \theta^T X \qquad [\theta_0, \theta_1, \theta_2] \begin{bmatrix} x_{bmi} \\ x_{ht} \\ 1 \end{bmatrix}$$

$$\hat{\theta} = \arg\min \sum_{n=1}^{N} (y_n - g_\theta(x_n))^2$$

$$\Rightarrow \hat{\theta} = \arg\min \| y - X\theta \|^2$$

a) $J(\theta) = \| y - X\theta \|^2 = \sqrt{(y - X\theta)^2}^{\,2} = (y - X\theta)^2$

$\nabla_\theta J(\theta) = 0$

$$2X^T(y - X\hat{\theta}) = 0.$$

$$X^T y - X^T X \hat{\theta} = 0$$

$$\boxed{\hat{\theta} = (X^T X)^{-1} X^T y}$$

To have a unique minimizer, $\nabla^2_\theta J(\theta)$ needs to be positive definite.

If $X^T X$ is singular, we can use regularization & LASSO regression.

c) See code

d). Find $\nabla \mathcal{E}_{train}(\theta^k)$.

$$\theta^{k+1} = \theta^k - \alpha^k \nabla \mathcal{E}_{train}(\theta^k)$$

$$\nabla \mathcal{E}_{train}(\theta^k) = \nabla_\theta \| y_n - X\theta^{(k)} \|^2$$

$$= 2X^T(y_n - X\theta^{(k)})$$

$$= 2X^Ty - 2X^TX\theta^k$$

$$= 2(X^Ty - X^TX\theta^k) = \underline{d}$$

$$\theta^{k+1} = \theta^k - \alpha^k(b - A\theta^k)$$

$$\boxed{\theta^{k+1} = \theta^k - \alpha^k b + \alpha^k A\theta^k.}$$

$$\min_\alpha J(\underline{\theta} + \alpha\underline{d}) = \|\underline{y} - X(\underline{\theta} + \alpha d)\|^2.$$

$$\nabla_\alpha J(\underline{\theta} + \alpha d) = \frac{d}{d\alpha}\left[\|y\|^2 + \|X(\underline{\theta} + \alpha\underline{d})\|^2 - 2\underline{y}^T X(\theta + \alpha\underline{d})\right]$$

$$= \frac{d}{d\alpha}\left[\|y\|^2 + (\underline{\theta} + \alpha\underline{d})^T X^T X(\underline{\theta} + \alpha\underline{d}) - 2y^T X(\theta + \alpha d)\right]$$

$$= \frac{d}{d\alpha}\left[(\underline{\theta} + \alpha\underline{d})^T X^T X(\underline{\theta} + \alpha\underline{d}) - 2y^T X\alpha\underline{d}\right].$$

$$= \frac{d}{d\alpha}\left[\underline{\theta}^T X^T X(\underline{\theta} + \alpha\underline{d}) + \alpha\underline{d}^T X^T X(\underline{\theta} + \alpha\underline{d})\right] - 2y^T X\underline{d}$$

$$= \theta^T X^T X\underline{d} + \underline{d}^T X^T X\underline{\theta} + 2\alpha d^T X^T X\underline{d} - 2y^T X\underline{d}$$

$$= 2\theta^T X^T X\underline{d} + 2\alpha d^T X^T X\underline{d} - 2y^T X\underline{d}$$

$$= 2(\underline{\theta}^T X^T X\underline{d} + \alpha d^T X^T X\underline{d} - y^T X\underline{d})$$

$$\nabla_\alpha J(\underline{\theta} + \alpha \underline{d}) = 0.$$

$$\underline{\theta}^T X^T X \underline{d} + \alpha \| X \underline{d} \|^2 - y^T X \underline{d} = 0,$$

$$\alpha^k = - \frac{\underline{\theta}^T X^T X \underline{d}^k - \underline{y}^T X \underline{d}^k}{\| X \underline{d}^k \|^2}$$

$$= - \frac{(\theta^T X^T X - y^T X) \underline{d}^k}{\| X d^k \|^2}$$

$$= - \frac{\frac{1}{2} d^{kT} \cdot d^k}{d^{kT} X^T \cancel{X} d^k}$$

$$= - \frac{1}{2} \frac{d^{kT} \cdot d^k}{d^{kT} \cdot A \cdot d^k}$$

4) $$\hat{\theta}_\lambda = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|X\theta - y\|_2^2 + \alpha\|\theta\|_2^2 \quad (3)$$

$$\hat{\theta}_\alpha = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|X\theta - y\|_2^2 \quad \text{s.t.} \quad \|\theta\|_2^2 \le \alpha \quad (4)$$

$$\hat{\theta}_\epsilon = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|\theta\|_2^2 \quad \text{s.t.} \quad \|X\theta - y\|_2^2 \le \epsilon \quad (5)$$

a) Looking at (3).

$$J(\hat{\theta}) = \|X\hat{\theta} - y\|^2 + \alpha\|\hat{\theta}\|^2 = (X\hat{\theta} - y)^2 + \alpha\hat{\theta}^2$$

$$\nabla_\theta J(\hat{\theta}) = 0$$

$$2X^T(X\hat{\theta} - y) + 2\alpha\hat{\theta} = 0$$

$$2X^TX\hat{\theta} - 2X^Ty + 2\alpha\hat{\theta} = 0$$

$$(X^TX + \lambda I)\hat{\theta} = X^Ty$$

$$\hat{\theta}_\lambda = (X^TX + \alpha I)^{-1}X^Ty$$

4 b). $\hat{\theta}_\alpha = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|X\theta - y\|_2^2 \quad \text{s.t.} \quad \|\theta\|_2^2 \leq \alpha \qquad (4)$

$\hat{\theta}_\epsilon = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \|\theta\|_2^2 \quad \text{s.t.} \quad \|X\theta - y\|_2^2 \leq \epsilon \qquad (5)$

(i) For (3)

$$\mathcal{L}(\theta) = \|X\theta - y\|_2^2 + \lambda \|\theta\|_2^2$$

For (4) :

$$\mathcal{L}(\theta, \gamma_\alpha) = \|X\theta - y\|_2^2 - \gamma_\alpha (\alpha - \|\theta\|_2^2)$$

For (5)

$$\mathcal{L}(\theta, \gamma_\varepsilon) = \|\theta\|^2 - \gamma_\varepsilon (\varepsilon - \|X\theta - y\|^2)$$

4 b ii)

(3): stationarity: $\nabla_\theta \mathcal{L} = (X^T X + \lambda I)\hat{\theta} - X^T y = 0$

Primal feasibility: None because it has no constraint.

Dual feasibility: None . . . .

Complementary slackness: None . . . .

---

(4). stationarity: $\nabla_\theta \mathcal{L} = 0$.

$$\therefore \quad X^T(X\theta - y) - \gamma_\alpha \theta = 0$$

Primal feasibility: $(\alpha - \|\theta\|_2^2) \geq 0$

Dual feasibility: $\gamma_\alpha \geq 0$.

complementary slackness: $\gamma_\alpha \cdot (\alpha - \|\theta\|_2^2) = 0$.

---

(5) Stationarity: $\nabla_\theta \mathcal{L} = 0$

$$\theta - \gamma_\varepsilon X^T(X\theta - y) = 0.$$

Primal feasibility: $(\varepsilon - \|X\theta - y\|_2^2) \geq 0$

Dual feasibility: $\gamma_\varepsilon \geq 0$

Complementary slackness: $\gamma_\varepsilon (\varepsilon - \|X\theta - y\|_2^2) = 0$

4b) iii) $\hat{\theta}_\alpha = (X^T X + \alpha I)^{-1} X^T y$.

$X^T (X \hat{\theta}_\alpha - y) - \gamma_\alpha \hat{\theta}_\alpha = 0$ (Stationarity)

$X^T X \hat{\theta}_\lambda - X^T y - \gamma_\alpha \hat{\theta}_\alpha = 0$

$X^T X \hat{\theta}_\alpha - (X^T X + \lambda I) \hat{\theta}_\alpha - \gamma_\alpha \hat{\theta}_\alpha = 0.$

$(\cancel{X^T X} - \cancel{X^T X} - \lambda I - \gamma_\alpha I) \hat{\theta} = 0.$

$\boxed{\gamma_\alpha = \lambda} > 0$ (Primal feasibility).

$\gamma_\alpha (\alpha - \|\hat{\theta}_\alpha\|^2) = 0$ (complementary slackness)

$\boxed{\alpha = \|(X^T X + \lambda I)^{-1} X^T y\|_2^2}$

4 b) iv).

$$\hat{\theta}_\lambda - \gamma_\varepsilon X^T(X\hat{\theta}_\lambda - y) = 0.$$

$$\hat{\theta}_\lambda - \gamma_\varepsilon X^TX\hat{\theta}_\lambda - \gamma_\varepsilon X^Ty = 0,$$

$$\hat{\theta}_\lambda - \gamma_\varepsilon X^TX\hat{\theta}_\lambda - \gamma_\varepsilon(X^TX + \lambda I)\hat{\theta}_\lambda = 0$$

$$\left[I - \gamma_\varepsilon(\cancel{X^TX} - \cancel{X^TX} + \lambda I)\right]\hat{\theta}_\lambda = 0$$

$$\left[(1 - \gamma_\varepsilon\lambda)I\right]\hat{\theta}_\lambda = 0$$

$$\boxed{\gamma_\varepsilon = \frac{1}{\lambda}} > 0 \quad \text{(stationarity)}$$

Complementary slackness: $\gamma_\varepsilon(\varepsilon - \|X\theta - y\|_2^2) = 0$

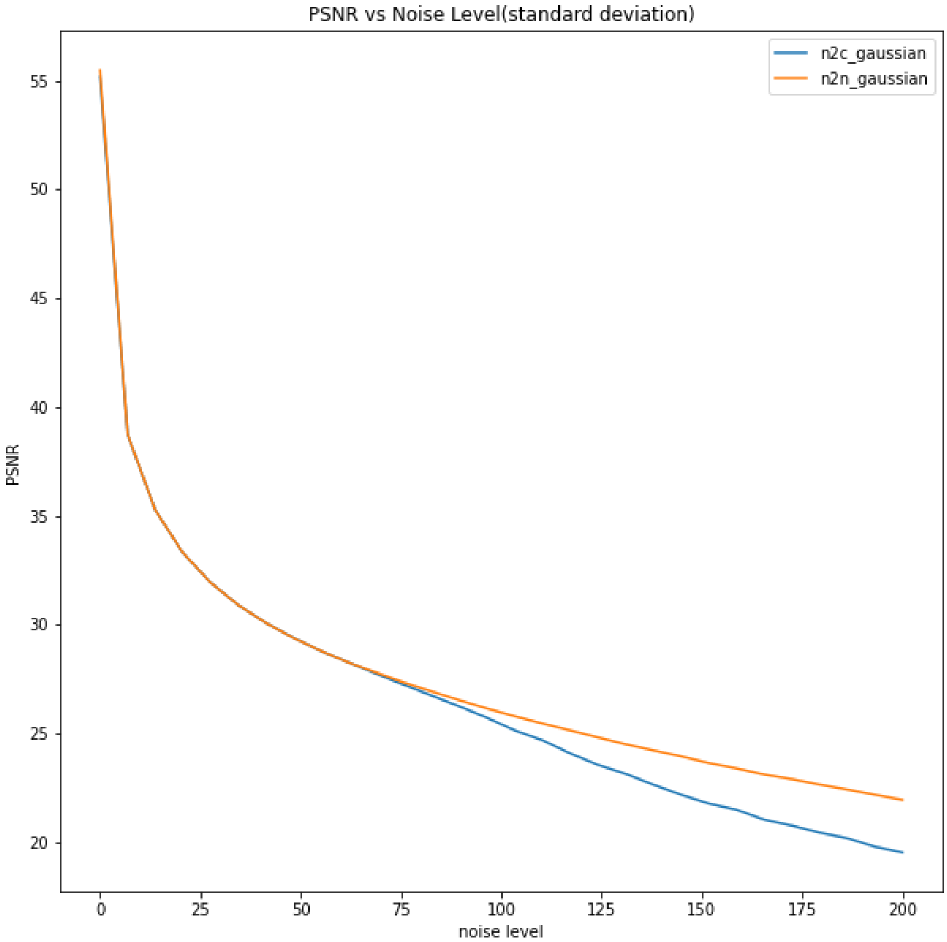
$$\varepsilon - \|X\hat{\theta}_\lambda - y\|_2^2 = 0$$

$$\varepsilon = \|X\hat{\theta}_\lambda - y\|_2^2$$

$$\boxed{\varepsilon = \|X(X^TX + \lambda I)^{-1}X^Ty - y\|_2^2}$$

4 b) v). KKT condition are necessary but not conclusive to guarantee a solution. However, since the problem we are optimizing is smooth, these conditions are sufficient for optimality, and we can claim that $\hat{\theta}_\lambda$ is the solution to (4).

5.



Noise levels are generate from 0 to 200 at normal distribution, then normalized by dividing by 255.

When the noise level is low, it's clear that both models perform similarly. However, as noise level get higher, the model trained with noise data (n2n) has a higher average PSNR score than the one trained with clean data.