Problem 1

(1) The main difference is that supervised learning uses labeled input and output data, while unsupervised learning does not.

(2) training data: used for learning to fit the parameters of the choose model.

Test data: used to assess the performance of the trained model.

Validation data: used to tune parameters. The specific reason that why we need validation data is that in the process of adjusting parameters, the final model may have the problem of overfitting, which may leads to bad generaliz-ation on new data. So we use part of training data to approximate out of sample error, that is validation, and we test the model and further, "validata" our hyper-parameter base on validation data.

(3) not necessarily. From the concertration inequality, we know $P(|r_i - \rho_i h_i| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$
so if $\sigma > 0.02$, the probility of right arm's length leaves the height by amount t is enlarged, which means it will concentration inequality for sub-Gussian couldn't hold.
If $\sigma^2 = 0.02$, the left arms and right arms are from the same distribution, it is better to use both of them.

(4). $X = V\Sigma U^T$, $V \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{d\times d}$, $\Sigma \in \mathbb{R}^{n\times d}$ and $V^TV = I$, $U^TU = I$

$X^TX = U\Sigma^T V^T V \Sigma U^T = U\Sigma^T\Sigma U^T$

$\because$ the X is full-lumn rank, $\Sigma = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$ and $\lambda_i \neq 0$

$\therefore$ so $\Sigma^T\Sigma > 0$, and $U\Sigma^T\Sigma U^T = \sum_{i=1}^{d} \lambda_i^2 U_i^2 > 0 \Rightarrow X^TX$ is positive definite

(5) set $f_\theta(X) = \|X\theta - y\|^2 + \lambda\|\theta\|^2$

$$= (\theta^TX^T - y^T)(X\theta - y) + \lambda\theta^T\theta$$

$$= \theta^TX^TX\theta - \theta^TX^Ty - y^TX\theta + y^Ty + \lambda\theta^T\theta$$

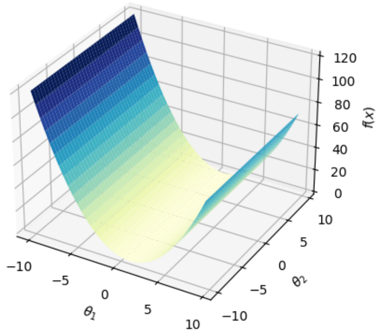
$$\frac{\partial f_\theta(X)}{\partial\theta} = 2X^TX\theta - 2X^Ty + 2\lambda\theta$$

$$= 0$$

$$\Rightarrow (X^TX + \lambda I)\theta^* = X^Ty$$

Problem 2

1. $\min\limits_{\theta} \| \theta_1 - 1 \|^2$, the 3D-figure for $f_\theta = (\theta_1 - 1)^2$ is:



(2) $\min\limits_{\theta} \| X\theta - y \|_2^2 = \min\limits_{z} \| Az - y \|_2^2$

$$\nabla f_z = 2 \| Az - y \| = 0 \Rightarrow A^T A \hat{z} = A^T y$$

$\because n < d \quad \therefore \text{rank}(A^T A) = n < d \Rightarrow \hat{z}$ is not unique

$\because z := U_1^T \theta$, $\quad U \in \mathbb{R}^{d \times d}$ are orthonormal $\therefore U^T = U^{-1}$

and $U_1^T \theta = \hat{z}$, $\quad \hat{\theta} = U^{-T} \hat{z} \Rightarrow \hat{\theta}$ is not unique.

Problem 3

(a) $\because P(\varepsilon_i) = \frac{1}{2b} e^{-\left| \frac{\varepsilon_i - 0}{b} \right|}$ and $\varepsilon_i \overset{iid}{\sim} L(0, b)$

$\therefore P(\varepsilon | \theta) = \pi P(\varepsilon_i | \theta) = \prod\limits_{i=1}^{n} \frac{1}{2b} e^{-\left| \frac{\varepsilon_i}{b} \right|} = \left(\frac{1}{2b}\right)^n e^{-\Sigma_i \left| \frac{\varepsilon_i}{b} \right|}$

and $\varepsilon = y - X\theta$, so further $P(\varepsilon | \theta) = \left(\frac{1}{2b}\right)^n e^{-\Sigma_i \left(\frac{y_i - x_i^T \theta}{b} \right)}$

$L(\theta | \varepsilon) = -n \log^{2b} - \Sigma_i (y_i - x_i^T \theta)$

learning problem: $\hat{\theta}_{ls} = \arg\min\limits_{\theta} \frac{1}{n} \sum\limits_{i=1}^{n} (y_i - x_i^T \theta)^2$

(b). $$H_u(x_i^T \theta - y_i) = \begin{cases} \| x_i^T \theta - y_i \|, & \| x_i^T \theta - y_i \| \geq u \\ \frac{\| x_i^T \theta - y_i \|^2}{2u} + \frac{u}{2}, & \| x_i^T \theta - y_i \| \leq u \end{cases}$$

$$\nabla H_u(x_i^T \theta - y_i) = \begin{cases} \| x_i \|, & \| x_i^T \theta - y_i \| \geq u \\ \frac{x_i (x_i^T \theta - y_i)}{u}, & \| x_i^T \theta - y_i \| \leq u \end{cases}$$

$$\therefore \nabla f(\theta) = \nabla \sum_i f_i(\theta) = \sum_{i=1}^{n} \nabla H_u(x_i^T \theta - y_i)$$

(c) (1) $\hat{\theta}_{LS} = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\theta^* + \varepsilon_1 + \varepsilon_2)$

$$= (X^T X)^{-1} (X^T X \theta^* + X\varepsilon_1 + X\varepsilon_2)$$

$$= \theta^* + (X^T X)^{-1} X (\varepsilon_1 + \varepsilon_2)$$

$$\| \hat{\theta}_{LS} - \theta^* \|_2 = \| (X^T X)^{-1} X (\varepsilon_1 + \varepsilon_2) \|_2$$

(2) The function of error with respect to iteration number is below:

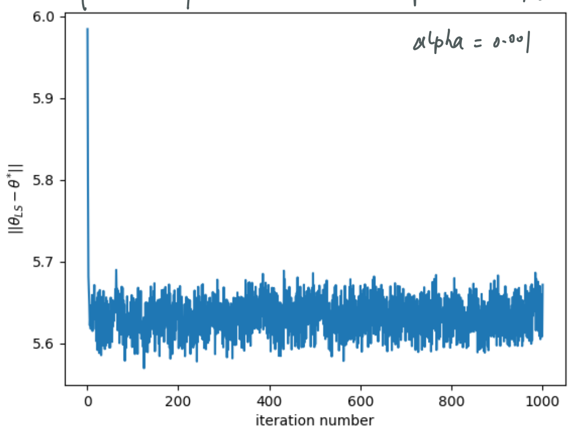
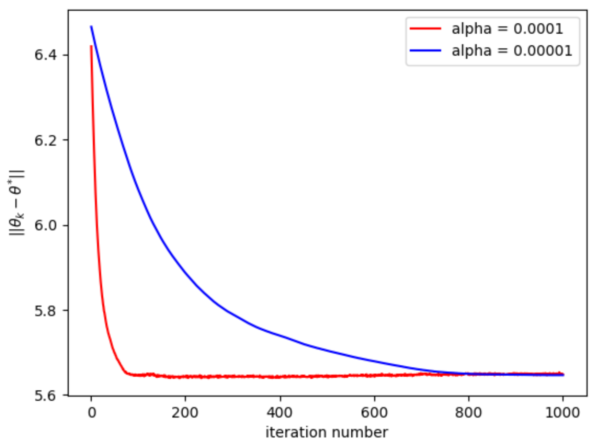


Figure 3-1

but we can see that the result doesn't converge. So I increase the iteration number to 100000, it still fails to converge. Therefore, I think it's the problem of our opitimization strategy. The step size α is fixed, which maybe too large to reach the optimal value in the end.

Further I decrease the alpha to 1e-4 and 1e-5 respectively, but result still doesn't change, just speed up the descent as follow:



In conclusion, this has to do nothing with our opitimization method. The key is that $\varepsilon_1 \in \mathbb{R}^n$ follows Gaussian distribution, and the figure 3-1 is a white Gaussian noise. We actually make it to converge, but maybe because the standard deviation of ε_1 is large, our convergence is not ideal.