CS189 HW3

(a)

$$\theta^* = \arg\max_\theta \sum_{i=1}^{n} \log P(X_i|\theta) + n \log P(\theta)$$

$$= \arg\min_\theta \|X\theta - y\|_2^2 + \frac{\sigma^2}{\sigma_h^2}\|\theta - \mu_\theta\|_2^2$$

$$\nabla_\theta \left(\|X\theta - y\|_2^2 + \frac{\sigma^2}{\sigma_h^2}\|\theta - \mu_\theta\|_2^2\right)$$

$$= \nabla_\theta \left((X\theta - y)^T(X\theta - y) + \frac{\sigma^2}{\sigma_h^2}(\theta-\mu_\theta)^T(\theta-\mu_\theta)\right)$$

$$= \nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T y - y^T(X\theta) + y^Ty + \frac{\sigma^2}{\sigma_h^2}(\theta^T\theta - \theta^T\mu_\theta - \mu_\theta^T\theta + \mu_\theta^T\mu_\theta)\right)$$

$$= 2X^TX\theta - 2X^Ty + \frac{2\sigma^2}{\sigma_h^2}(\theta - \mu_\theta) = 0$$

$$X^TX\theta^* - X^Ty + \frac{\sigma^2}{\sigma_h^2}(\theta^* - \mu_\theta) = 0$$

$$\theta^* = \left(X^TX + \frac{\sigma^2}{\sigma_h^2}I\right)^{-1}\left(X^Ty + \frac{\sigma^2}{\sigma_h^2}\mu_\theta\right)$$

(b) The two figures for $\sigma_h = 1$ and $\sigma_h = 10$ are included in Appendix. The code is in problem1-starter.py from # Generating Data and Labels with Random Gaussian Noise to the end. After changing σ_h to 10, The range and variation of prior obviously increases. This is because σ_h is the standard deviation of prior distribution. σ_h increases means the variance of prior increases.

Code also includes the implement of MLE, MAP and prior.

(c) The two figures for $\mu_\theta = \begin{bmatrix}-5\\-5\end{bmatrix}$ and $\begin{bmatrix}3\\3\end{bmatrix}$ are included in Appendix. The code is in problem1-starter.py.

When change to $\begin{bmatrix}-5\\-5\end{bmatrix}$, the center of prior plot moves from (0, 0) to (-5, 5). When change to $\begin{bmatrix}3\\3\end{bmatrix}$, the center of prior plot moves from (-5, -5) to (3, 3). θ_{MLE} doesn't change according to the plots with respect to prior.

(d)

$$\theta_{MAP} = \arg\max_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} \log P(x_i|\theta) + \log P(\theta)$$

$$= \arg\max_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - X_i\theta_i)^2}{2\sigma^2}}\right) + \log\left(\frac{1}{2\sigma_h} e^{-\frac{|\theta-\mu|_1}{\sigma_h}}\right)$$

$$= \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} \frac{(y_i - X_i\theta_i)^2}{2\sigma^2} + \frac{|\theta-\mu|_1}{\sigma_h}$$

$$= \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} (y_i - X_i\theta_i)^2 + \frac{2\sigma^2}{\sigma_h}|\theta-\mu|_1$$

$$= \arg\min_{\theta\in\mathbb{R}^d} \|y - X\theta\|^2 + \frac{2\sigma^2}{\sigma_h}|\theta-\mu|_1$$

(e) The two plots for $p(\theta_i) \sim L(0,1)$ and $p(\theta_i) \sim L(0, 0.0001)$
The code is in problem1_starter.py. With implement of Laplacian_prior and MAP_contour.
After change σ_h from 1 to 0.0001 the weight of prior in θ_{MAP} increases a lot.
This is because: According to the result in (d)

$$\theta_{MAP} = \underset{\theta \in R^d}{\operatorname{argmin}} \|y - X\theta\|^2 + \underbrace{2\frac{\sigma^2}{\sigma_h^2} |\theta - \mu|_1}$$

This is the prior term in θ_{MAP}, its weight is $\frac{2\sigma^2}{\sigma_h^2}$. If I change σ_h from 1 to 0.0001, The weight changes to 10^8 times before

(f) If prior of θ is Unif (0, 1) then $P(\theta) = 1$

$\because y = X\theta + Z \quad Z \sim N(0,1) \therefore y \sim N(X\theta, 1)$

$$\therefore \theta_{MAP} = \underset{\theta \in R^d}{\operatorname{argmin}} \|y - X\theta\|^2 + 2\sigma^2 \times \log P(\theta)$$

$$= \underset{\theta \in R^d}{\operatorname{argmin}} \|y - X\theta\|^2 + 2\sigma^2(\log 1)$$

$$= \underset{\theta \in R^d}{\operatorname{argmin}} \|y - X\theta\|^2 = \theta_{MLE}$$

~~So Now~~

So, now θ_{MAP} and θ_{MLE} are equivalent.

2. Probabilistic Model of Linear regression

(a) $\because Y = X w_1 + w_0 + Z, \quad Z \sim N(0, 1)$

$\therefore Y|X=x \sim N(x w_1 + w_0, 1)$

$$P(Y|X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y-(Xw_1+w_0))^2}{2}}$$

(b) Log likelihood function

$$\underset{w}{\arg\max} \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i-(X_i w_1+w_0))^2}{2}}\right)$$

$$= \underset{w}{\arg\min} \sum_{i=1}^{n} (Y_i - (X_i w_1 + w_0))^2$$

$$= \underset{w}{\arg\min} \| Y - X w_1 - w_0 * \mathbf{1} \|^2 \qquad \mathbf{1} \in R^n \quad \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$= \underset{w}{\arg\min} (Y - X w_1 - w_0 \mathbf{1})^T (Y - X w_1 - w_0 \mathbf{1})$$

$$= \underset{w}{\arg\min} (Y^T - w_1 X^T - w_0 \mathbf{1}^T)(Y - X w_1 - w_0 \mathbf{1})$$

$$= \underset{w}{\arg\min} (Y^T Y - Y^T X w_1 - Y^T w_0 \mathbf{1} - w_1 X^T Y + w_1 X^T X w_1 + w_1 X^T w_0 \mathbf{1} - w_0 \mathbf{1}^T Y + w_0 \mathbf{1}^T X w_1 + w_0 \mathbf{1}^T w_0 \mathbf{1})$$

$$= \underset{w}{\arg\min} \sum_{i=1}^{n} Y_i^2 - 2 Y^T X w_1 - 2 w_0 \sum_{i=1}^{n} Y_i + w_1^2 X^T X + 2 w_0 w_1 \sum_{i=1}^{n} X_i + n w_0^2$$

$$\begin{cases} \nabla w_1 = -2 X^T Y + 2 X^T X w_1 + 2 w_0 \sum_{i=1}^{n} X_i \\ \nabla w_0 = -2 \sum_{i=1}^{n} Y_i + 2 w_1 \sum_{i=1}^{n} X_i + 2 n w_0 \end{cases}$$

$$\begin{cases} -2 X^T Y + 2 X^T X w_1 + 2 \sum_{i=1}^{n} X_i w_0 = 0 \\ -2 \sum_{i=1}^{n} Y_i + 2 w_1 \sum_{i=1}^{n} X_i + 2 n w_0 = 0 \end{cases}$$

$$\begin{cases} w_1 = \dfrac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i) + n \sum_{i=1}^{n} X_i Y_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \\ w_0 = \dfrac{(\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} X_i^2) + (\sum_{i=1}^{n} X_i Y_i)(\sum_{i=1}^{n} X_i)}{(\sum_{i=1}^{n} X_i)^2 - n \sum_{i=1}^{n} X_i^2} \end{cases}$$

(c) $\because Z \sim U[-0.5, 0.5] \quad Y = X w + Z$

$\therefore Y|X \sim U[-0.5 + Xw, 0.5 + Xw]$

(d) $Y_1 = X_1 W + Z_1$ log likelihood function

$\vdots$ $\log P(Y=Y_1 \mid X=X_1) \times \cdots \times P(Y=Y_n \mid X=X_n)$

$Y_n = X_n W + Z_n$ $= \sum_{i=1}^{n} \log P(Y=Y_i \mid X=X_i)$

$\because$ when $-0.5+X_iW \le Y_i \le 0.5+X_iW$, $P(Y=Y_i \mid X=X_i)=1$ otherwise it is 0

$\therefore$ To maximize log likelihood function, I need to make $-0.5+X_iW \le Y_i \le 0.5+X_iW$ hold for as many (X_i, Y_i) as possible.

$\because$ $-0.5+X_iW \le Y_i \le 0.5+X_iW \Rightarrow \frac{Y_i-0.5}{X_i} \le W \le \frac{Y_i+0.5}{X_i}$

$\therefore$ I think one appropriate estimate for W_{MLE} is : $\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i}$

(e) $Y_i = X_iW + Z_i$ $Z_i \sim N(0,1)$

$Y_i \sim N(X_iW, 1)$ $W \sim N(0, \sigma^2)$

$$\log P(W \mid Y, X) \propto \log P(Y \mid W, X) + \log P(W)$$

$$= \sum_{i=1}^{n} \log P(Y_i \mid X_i W) + \log P(W)$$

$$\propto -\sum_{i=1}^{n}\frac{(Y_i - X_iW)^2}{2} - \frac{W^2}{2\sigma^2}$$

$$\propto -\frac{\sigma^2\sum_{i=1}^{n}(Y_i - X_iW)^2 + W^2}{2\sigma^2}$$

$$\propto \frac{\left(\sum_{i=1}^{n}X_i^2\sigma^2+1\right)W^2 - \sum_{i=1}^{n}2\sigma^2X_iY_iW + \sum_{i=1}^{n}\sigma^2Y_i^2}{2\sigma^2}$$

$$\propto \frac{\left(W - \frac{\sum_{i=1}^{n}\sigma^2X_iY_i}{\sum_{i=1}^{n}\sigma^2X_i^2+1}\right)^2}{\frac{2\sigma^2}{\sum_{i=1}^{n}X_i^2\sigma^2+1}}$$

$\therefore$ $P(W \mid Y, X) \sim N(\mu', \sigma'^2)$ $\therefore$ posterior mean

where $\mu' = \frac{\sum_{i=1}^{n}\sigma^2X_iY_i}{\sum_{i=1}^{n}\sigma^2X_i^2+1}$ $\sigma'^2 = \frac{\sigma^2}{\sum_{i=1}^{n}\sigma^2X_i^2+1}$ is $\hat{\mu}' = \frac{\sum_{i=1}^{n}\sigma^2X_iY_i}{\sum_{i=1}^{n}\sigma^2X_i^2+1}$

(f) $Y_i = W^TX_i + Z_i$ $Z_i \sim N(0,1)$ $\therefore Y_i \sim N(W^TX_i, 1)$

log likelihood function

$$\sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i - W^T X_i)^2}{2}} \right]$$

$$W_{MLE} = \underset{W}{\arg\max} \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i - W^T X_i)^2}{2}} \right]$$

$$= \underset{W}{\arg\min} \sum_{i=1}^{n} \frac{(Y_i - W^T X_i)^2}{2} + \underbrace{\frac{1}{2} \log 2\pi}_{\text{has no } W}$$

$$= \underset{W}{\arg\min} \sum_{i=1}^{n} (Y_i - W^T X_i)^2$$

$$= \underset{W}{\arg\min} \| Y - W^T X \|^2 = W_{LS}$$

$\therefore$ In this setting, the maximum likelihood estimator for w is the solution to a lease square problem

(9) $Y_i = W^T X_i + Z_i \quad Z_i \sim N(0, 1) \quad W_i \sim N(0, \sigma^2)$

$p(w|x, Y) \propto P(X|W) P(W)$

$$\begin{pmatrix} \vdots \\ P(W_i | X, Y) \\ \vdots \end{pmatrix} \propto \begin{pmatrix} \vdots \\ e^{-\frac{\sum_{i=1}^{n} (Y_i - W^T X_i)^2}{2} - \frac{W_i^2}{2\sigma^2}} \\ \vdots \end{pmatrix}$$

I didn't manage to solve this problem.

3. Simple Bias - Variance Tradeoff

(a) ① $E\left(\frac{x_1+x_2+\dots+x_n}{n}-\mu\right) = \frac{1}{n}\times n\mu-\mu = 0$

② $E\left(\frac{x_1+x_2+\dots+x_n}{n+1}-\mu\right) = \frac{1}{n+1}\times n\mu-\mu = -\frac{\mu}{n+1}$

③ $E\left(\frac{x_1+x_2+\dots+x_n}{n+n_0}-\mu\right) = \frac{1}{n+n_0}\times n\mu-\mu = -\frac{n_0\mu}{n+n_0}$

④ $E(0-\mu) = -\mu$

(b) $X_1, X_2, \dots, X_n$ are i.i.d

① $Var(\hat{X}) = \frac{1}{n^2}\left[Var(X_1)+\dots+Var(X_n)\right] = \frac{1}{n^2}\times n\sigma^2 = \frac{\sigma^2}{n}$

② $Var(\hat{X}) = \frac{1}{(n+1)^2}\times n\sigma^2 = \frac{n\sigma^2}{(n+1)^2}$

③ $Var(\hat{X}) = \frac{1}{(n+n_0)^2}\times n\sigma^2 = \frac{n\sigma^2}{(n+n_0)^2}$

④ $Var(\hat{X}) = 0$

(c) $E[(\hat{X}-X')^2] = Var(\hat{X}-X') + [E(\hat{X}-X')]^2$

$= Var(\hat{X}) + Var(X') + \{E[(\hat{X}-\mu)+(\mu-X')]\}^2$

$= Var(\hat{X}) + \sigma^2 + [E(\hat{X}-\mu)+E(\mu-X')]^2$

$= Var(\hat{X}) + \sigma^2 + [E(\hat{X}-\mu)]^2$

$= Var(\hat{X}) + \sigma^2 + [bias(\hat{X})]^2$

$E[(\hat{X}-\mu)^2] = Var(\hat{X}-\mu) + [E(\hat{X}-\mu)]^2$

$= Var(\hat{X}) + [bias(\hat{X})]^2$

Compare them, there is a σ^2 in $E[(\hat{X}-X')^2]$ but not in $E[(\hat{X}-\mu)^2]$

$\hat{X}$ is calculated based on training samples

so $\hat{X}$ will perform worse on fresh sample X', therefore the error will be larger.

(d) ① $= Var(\hat{X}) + bias^2(\hat{X}) = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}$

② $= \frac{n}{(n+1)^2}\sigma^2 + \left(\frac{\mu}{n+1}\right)^2 = \frac{n\sigma^2+\mu^2}{(n+1)^2}$

③ $= \frac{n\sigma^2}{(n+n_0)^2} + \frac{n_0^2\mu^2}{(n+n_0)^2} = \frac{n\sigma^2+n_0^2\mu^2}{(n+n_0)^2}$

④ $= 0 + \mu^2 = \mu^2$

(e) ① is when $n_0 = 0$ ② is when $n_0 = 1$ ④ is when $n_0 = \infty$

(f) As n_0 increases, $|bias|$ increases (since bias ≤ 0 here, actually bias is decreasing), variance decreases.

(g) $n_0 = \alpha n$, $Var(\hat{X}) = \frac{n\sigma^2}{(n+\alpha n)^2}$ $bias(\hat{X}) = -\frac{\alpha n\mu}{n+\alpha n}$

$$\text{expected total error} = \frac{n\sigma^2}{(n+\alpha n)^2} + \frac{(\alpha n \mu)^2}{(n+\alpha n)^2}$$

$$= \frac{n\sigma^2 + \alpha^2 n^2 \mu^2}{(n+\alpha n)^2}$$

$$\frac{\partial}{\partial \alpha} \text{expected total error} = \frac{n^2\mu^2 \times 2\alpha(n+\alpha n)^2 - (n\sigma^2 + \alpha^2 n^2 \mu^2) \times 2(n+\alpha n) \times n}{(n+\alpha n)^4}$$

$$= 0$$

$$2n^2\mu^2\alpha(n+\alpha n)^2 = 2n(n+\alpha n)(n\sigma^2 + \alpha^2 n^2 \mu^2)$$

$$n\mu^2(n+\alpha n)\alpha = n\sigma^2 + \alpha^2 n^2 \mu^2$$

$$n^2\mu^2\alpha + n^2\mu^2\alpha^2 = n^2\mu^2\alpha^2 + n\sigma^2$$

$$n^2\mu^2\alpha = n\sigma^2$$

$$\alpha = \frac{n\sigma^2}{n^2\mu^2} = \frac{\sigma^2}{n\mu^2}$$

(h) When σ is large and μ is small (close to 0)

$$\alpha = \frac{\sigma^2}{n\mu^2} \rightarrow \infty \text{ (very large)}$$

(i) $X' = X - \mu_0$

$E(X') = E(X - \mu_0) = \mu - \mu_0$

$Var(X') = Var(X - \mu_0) = Var(X) = \sigma^2$

(j) In ridge regression, as λ increases, model bias $\uparrow$, model variance $\downarrow$.

This is very similar to α's influence of $\hat{X}$'s bias and variance.

So in ridge regression, we can use cross validation to select the λ value with the smallest validation error.

4 Robotic Learning of Controls from Demonstrations and Images.

(a) The 0th 10th and 20th images in the training set are plotted in the appendix

Their corresponding control vectors are shown in appendix.

The code is in robotic_ridge_code_starter.py under section # 4(a).

(b) The code to do this is under section # 4(b)

When I attempt to do this, it will raise Singular matrix error

This is because $\det(X^TX) = 0$, X^TX is not invertible

(c) For each λ in [0.1, 1.0, 10.0, 100.0, 1000.0] the result is shown in appendix

The code is under section # 4(c)

(d) The result is shown in appendix

The code is under section # 4(d)

(e) The result is shown in appendix

The code is under section # 4(e)

By increasing λ value, the bias will increase because $\lambda \uparrow$ means the model complexity will decrease, therefore bias will increase

By increasing λ value, the varience will decrease because $\lambda \uparrow$, model complexity $\downarrow$, therefore varience $\downarrow$.

(f) The result is shown in appendix

The code is under section # 4(f)