

CMPUT 466 A2
Yuhan Ye 1463504

Q1.

$$(a) \theta_{MAP} = \underset{\theta}{\operatorname{argmax}} (p(D|\theta) p(\theta))$$

$$p(x_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma_0^2}}$$

$$p(D|\theta) = \prod_{i=1}^n p(x_i|\theta)$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma_0^n} \cdot e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2}}$$

$$p(D|\theta) p(\theta) = \frac{1}{(2\pi)^{\frac{n+1}{2}} \sigma_0^{n+1}} \cdot e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}}$$

$$LL(\theta) = \log(p(D|\theta) p(\theta)) = \log\left(\frac{1}{(2\pi)^{\frac{n+1}{2}} \sigma_0^{n+1}}\right) - \frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma_0^2} - \frac{(\theta-\mu)^2}{2\sigma^2}$$

$$\frac{\partial LL(\theta)}{\partial \theta} = \frac{\sum_{i=1}^n (x_i-\theta)}{\sigma_0^2} - \frac{\theta-\mu}{\sigma^2} = 0$$

$$\frac{\left(\sum_{i=1}^n x_i - n\theta\right) \cdot \sigma^2 - \sigma_0^2(\theta-\mu)}{\sigma_0^2 \sigma^2} = 0$$

$$\theta \cdot (-n\sigma^2 - \sigma_0^2) + \sum_{i=1}^n x_i \sigma^2 + \sigma_0^2 \mu = 0$$

$$\theta_{MAP} = \frac{\sum_{i=1}^n x_i \sigma^2 + \sigma_0^2 \mu}{n \sigma^2 + \sigma_0^2}$$

$$(b) p(D|\theta) = \frac{1}{(2\pi)^{\frac{n}{2}} 6_0^n} e^{-\sum_{i=1}^n (x_i - \theta)^2 / 2 6_0^2}$$

$$p(\theta) = \frac{1}{2b} e^{-|\theta|/b}$$

$$p(D|\theta)p(\theta) = \frac{1}{(2\pi)^{\frac{n}{2}} 6_0^n \cdot 2b} e^{-\sum_{i=1}^n (x_i - \theta)^2 / 2 6_0^2} - \frac{|\theta|}{b}$$

$$LL(\theta) = \log(p(D|\theta)p(\theta)) = \log\left(\frac{1}{(2\pi)^{\frac{n}{2}} 6_0^n \cdot 2b}\right) - \frac{\sum_{i=1}^n (x_i - \theta)^2}{2 6_0^2} - \frac{|\theta|}{b}$$

$$\frac{dLL(\theta)}{d\theta} = \frac{\sum_{i=1}^n (x_i - \theta)}{6_0^2} - \frac{|\theta|}{b\theta} = 0$$

$$\theta = \begin{cases} \frac{\sum_{i=1}^n x_i}{n} - \frac{6_0^2}{bn}, & \theta > 0 \\ \frac{\sum_{i=1}^n x_i}{n} + \frac{6_0^2}{bn}, & \theta < 0 \end{cases} \quad \text{or } \theta = \frac{\sum_{i=1}^n x_i}{n} - \frac{6_0^2 |\theta|}{bn}$$

We do not have a closed form for θ_{MAP}

Our objective is continuous but not smooth, it is not differentiable at $\theta=0$, making whole expression non-differentiable.

We can use proximal methods. Use gradient descent for the smooth component and for θ that are close to 0, set them to 0.

$$\text{With step size } \eta. \quad \theta = \theta_i + \frac{\sum_{i=1}^n x_i}{n}$$

$$\text{Proxy } (\theta_i) = \begin{cases} \theta_i - \eta \frac{6_0^2}{bn}, & \theta_i > \eta \frac{6_0^2}{bn} \\ 0 & |\theta_i| \leq \eta \frac{6_0^2}{bn} \\ \theta_i + \eta \frac{6_0^2}{bn}, & \theta_i < -\eta \frac{6_0^2}{bn} \end{cases}$$

$$(c) p(x_i|\theta) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^2} e^{-\frac{1}{2}(x_i - \theta)^T I \cdot (x_i - \theta)}$$

$$p(\theta) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^2} e^{-\frac{1}{2}\theta^T \Sigma^{-1} \theta} \quad \Sigma^{-1} = \sigma^{-2} I$$

$$p(D|\theta) = \frac{1}{(2\pi)^{\frac{nd}{2}} \sigma^2} e^{-\frac{1}{2} \sum_{i=1}^n [(x_i - \theta)^T I \cdot (x_i - \theta)]}$$

$$LL(\theta) = \log(p(D|\theta)p(\theta)) = \log\left(\frac{1}{(2\pi)^{\frac{d(n+1)}{2}} \sigma^2}\right) - \frac{1}{2} \sum_{i=1}^n [(x_i - \theta)^T I \cdot (x_i - \theta)] - \frac{1}{2} \theta^T \cdot \sigma^{-2} I \cdot \theta$$

$$\frac{\partial LL(\theta)}{\partial (\theta)} = \sum_{i=1}^n (x_i - \theta) - \sigma^{-2} \theta = 0$$

$$\sum_{i=1}^n x_i - n\theta - \sigma^{-2} \theta = 0$$

$$\theta_{MAP} = \frac{\sum_{i=1}^n x_i}{n + \sigma^{-2}}$$

(0-70)

2. (a) Increasing number of features to more than 69 raise numpy.linalg.LinAlgError: singular matrix
 It is because that X is at low-rank, $X^T X$ is not invertible. To solve this, we can use pseudo-inverse numpy.linalg.pinv

(c) Training error of RidgeRegression is larger than that of FSLRegression. While Test error of Ridge Regression is smaller. It indicates that FSLRegression is 'overfit', while ridge balances value of weights (penalize big values in w) and performs better on test set.

(f) SGD is faster to approach local minima, that weights is converged although it is 'noisy' at the beginning.
 SGD plot is smooth, since it decreases in a right direction each epoch, but takes longer time.

Hilroy

Bonus

- (c) Note that in the initial phase of iterations
Momentum has an offset to the initial value
($m_0 = 0$, ~~$v_0 = 0$~~ , so there are bias toward 0)
Thus Momentum without bias correction is
actually a biased estimate of the expected gradient