ML HW-3

(Q1) Monotonicity of loss & regularizer as the regularization parameter changes.

(A1) (a) By definition we have,

$$L(\tilde{\theta}^*) + \lambda r(\tilde{\theta}^*) \geq L(\theta^*) + \lambda r(\theta^*)$$

& similarly for $\tilde{\theta}$,

$$L(\theta^*) + \tilde{\lambda} r(\theta^*) \geq L(\tilde{\theta}^*) + \tilde{\lambda} r(\tilde{\theta}^*)$$

$\Rightarrow$ Subtracting the inequalities, we get the following results:

$$\Rightarrow L(\tilde{\tilde{\theta}}^*) + \lambda r(\tilde{\tilde{\theta}}^*) - L(\tilde{\theta}^*) - \tilde{\lambda} r(\tilde{\theta}^*)$$

$$\geq L(\theta^*) + \lambda r(\theta^*) - L(\theta^*) - \tilde{\lambda} r(\theta^*)$$

$$\Rightarrow (\lambda - \tilde{\lambda}) r(\tilde{\theta}^*) \geq (\lambda - \tilde{\lambda}) r(\theta^*),$$

&

moving $r(\theta^*)$ to the left hand side yields,

$$(\lambda - \tilde{\lambda})(r(\tilde{\theta}^*) - r(\theta^*)) \geq 0$$

But, since $\tilde{\lambda} \geq \lambda$, then

$$r(\tilde{\theta}^*) - r(\theta^*) \leq 0$$

Thus,

$$\boxed{r(\tilde{\theta}^*) \leq r(\theta^*)}$$

(b) We can use the same trick by multiplying both sides of the inequalities by $\frac{1}{\lambda}$, $\frac{1}{\tilde{\lambda}}$ respectively then,

$$\left(\frac{1}{\lambda}\right) L(\tilde{\theta}^*) + r(\tilde{\theta}^*) \geq \left(\frac{1}{\lambda}\right) L(\theta^*) + r(\theta^*)$$

&

$$\left(\frac{1}{\tilde{\lambda}}\right) L(\theta^*) + r(\theta^*) \geq \left(\frac{1}{\tilde{\lambda}}\right) L(\tilde{\theta}^*) + r(\tilde{\theta}^*)$$

Subtracting both inequalities,

$$\Rightarrow \left(\frac{1}{\lambda} - \frac{1}{\tilde{\lambda}}\right) L(\tilde{\theta}^*) \geq \left(\frac{1}{\lambda} - \frac{1}{\tilde{\lambda}}\right) L(\theta^*)$$

$$\Rightarrow \left(\frac{1}{\lambda} - \frac{1}{\tilde{\lambda}}\right)\left(L(\tilde{\theta}^*) - L(\theta^*)\right) \geq 0$$

Since $\lambda \leq \tilde{\lambda}$, we have $\left(\frac{1}{\lambda} - \frac{1}{\tilde{\lambda}}\right) \geq 0$, so

$$L(\tilde{\theta}^*) - L(\theta^*) \geq 0$$

Thus,

$$\boxed{L(\tilde{\theta}^*) \geq L(\theta^*)}$$

(Q2) MAP interpretation of regularized empirical loss minimization

(A2) @ $p(y|x;\theta) \sim N(\phi^T\theta, \sigma^2)$

&

$p(\theta) \sim N(0, \sigma_0^2 I)$

We deduce this:

$$P(y_i | x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\phi^T\theta - y_i)^2}{2\sigma^2}}$$

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2 I}} e^{\frac{-\theta^2}{2\sigma_0^2 I}}$$

$\Rightarrow$ Use above two values in MAP formulation we get the following:

$$\text{minimize} \sum_{i=1}^{n} -\log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\phi^T\theta - y_i)^2}{2\sigma^2}} - \log \frac{1}{\sqrt{2\pi\sigma_0^2 I}} e^{-\theta^2/2\sigma_0^2 I}$$

$\Rightarrow$ MAP depends on θ we remove the constant terms and we get:

$$\text{minimize} \sum_{i=1}^{n} (\phi^T\theta - y_i)^2 + \frac{\sigma^2}{\sigma_0^2 I} \|\theta\|^2$$

$\Rightarrow$ This is of the form $L(\theta) + \lambda r(\theta)$ where

$$\lambda = \frac{\sigma^2}{\sigma_0^2 I}$$

(b) $P(y|x;\theta) \sim N(\phi^T\theta, \sigma^2)$

$$P(\theta) = \prod_{j=1}^{m} \frac{1}{2a} e^{-|\theta_j|/a}$$

$$-\log P(\theta) = -\log\left(\prod_{j=1}^{m} \frac{1}{2a} e^{-|\theta_j|/a}\right)$$

$$-\log P(\theta) = m \log 2a + \sum_{j=1}^{m} \frac{|\theta_j|}{a}$$

⇒ Taking from part (a) & substituting in MAP formulation we get:

⇒ minimize $\sum_{i=1}^{n} (\phi^T\theta - y_i)^2 + \frac{2\sigma^2 m}{a} \sum_{j=1}^{m} |\theta_j|$

⇒ This is of the form $L(\theta) + \lambda r(\theta)$ where

$$\lambda = \frac{2\sigma^2}{a}$$

(c) $P(y|x,\theta) = Pr[yu > 0]$ where $P(u|x,\theta) \sim N(\phi^T\theta, \sigma^2)$

We can do the following:

$$P(y|x,\theta) = Pr[yu \geq 0]$$

$$P(y|x,\theta) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{\left(\frac{-(u-y\phi^T\theta)^2}{2}\right)} du$$

$$P(y|x,\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y\phi^T\theta} e^{(-u^2/2)} du$$

$$P(y|x,\theta) = \Phi(y\phi^T\theta)$$

Using part a we get

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2 I}} e^{-\theta^2/2\sigma_0^2 I}$$

⇒ Substitution in MAP formula we get:

$$\text{minimize} \; -\log\left(\Phi(y\phi^T\theta)\right) + \frac{1}{2\sigma_0^2 I}\|\theta_j\|_2^2$$

⇒ This is of the form $L(\theta) + \lambda r(\theta)$ where

$$\lambda = \frac{1}{2\sigma_0^2 I}$$

(d) $P(y_i, x_i; \theta) = e^{\frac{-y_i\phi^T\theta}{1+e^{-y_i\phi^T\theta}}}$

& From (b) the value of $P(\theta)$ we can do following changes in our MAP formulation we get: $\text{minimize} \sum_{i=1} y_i\phi^T\theta + \|\theta\|_{1/2}$

⇒ This is of the form $L(\theta) + \lambda r(\theta)$ where

$$\lambda = 1/\alpha$$

(Q4) Hand-Written Digits classification

(A4) Given objective f:

$$\underset{\theta_1, \cdots, \theta_K}{\text{minimize}} \; \frac{1}{n} \sum_{i=1}^{n} \max_{c} \left(x_i^T \theta_c - x_i^T \theta_{y_i} + 1_{y_i \neq c} \right) + \lambda \sum_{j=1}^{m} \sqrt{\sum_{c=1}^{K} \theta_{jc}^2} \longrightarrow (1)$$

⇒ Randomly Select an x_i

⇒ Find subgradient at 'x_i' by differentiating eq (1)

⇒ θ corresponding to correct class.

$$\nabla_{\theta_{y_i}} L_i = -\left(\sum_{y \neq y_i}^{1} (\theta^T x_i - \theta_{y_i}^T x_i + \text{alpha}) x_i \right)$$

$$\Rightarrow \theta_{y_i}^{(t+1)} = \theta_{y_i}^{(t)} - r^{(t)} x_i$$

⇒ θ corresponding to incorrect class

$$\nabla_{\theta_j} L_i = 1\left(\theta_j^T x_j - \theta_{y_i}^T x_i + \text{alpha}\right) x_i$$

alpha here is $\underline{\underline{1}}$.

$$\Rightarrow \theta_{y_i}^{(t+1)} = \theta_{y_i}^{(t)} + r^{(t)} x_i$$

⇒ New θ(theta) is for proximal operation

$$\theta^{t+1} = \theta^t - \lambda(\text{Subgradient})$$

⇒ New weight that is theta is given as:

$$\text{Prox}(\theta^t - \gamma^{(t)} \lambda \nabla L)$$

⇒ Subgradient is calculated by block soft-thresholding on θ. The norm of each row of θ is calculated if it is less than $r^{(t)}\lambda$, the entire row is set to 0, otherwise magnitude of the row decreases by the amount of $r^{(t)}\lambda$.

$$\Rightarrow \theta_j^{(t+1)} = \begin{cases} 0 & : \text{if } v \le r^{(t)}\lambda \\ \left(1 - \frac{r^{(t)}\lambda}{v}\right)\theta_j^{(t+1)} & : \text{otherwise} \end{cases}$$

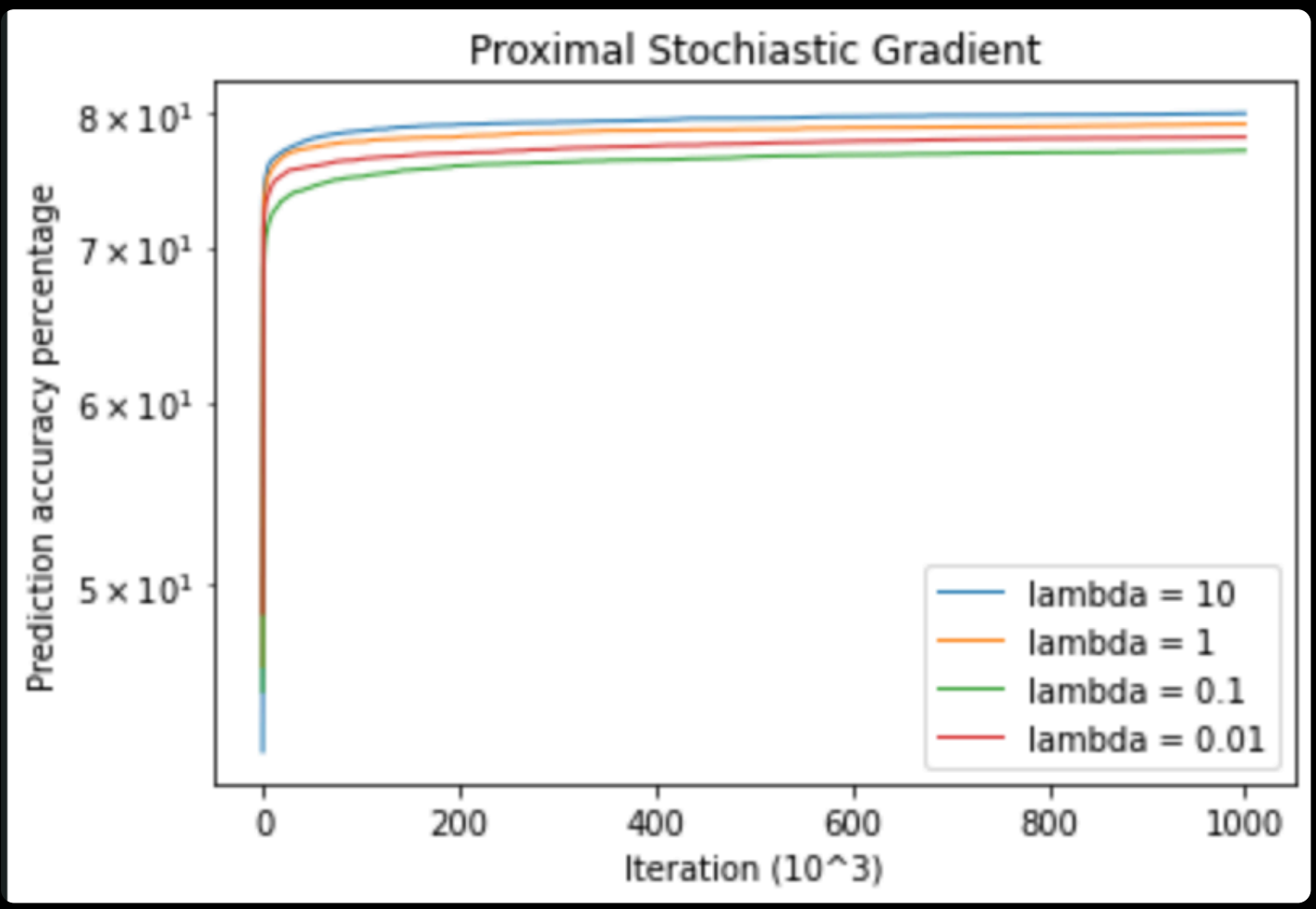
⇒ where $v = ||\theta_j^{(t+1)}||$

(b) Code attached in python file & ran on kaggle

https://www.kaggle.com/code/boltcoder/ml-hw3/notebook

(c) The testing error - rate decreases as the algorithm progresses as shown in the figure for all λ values.

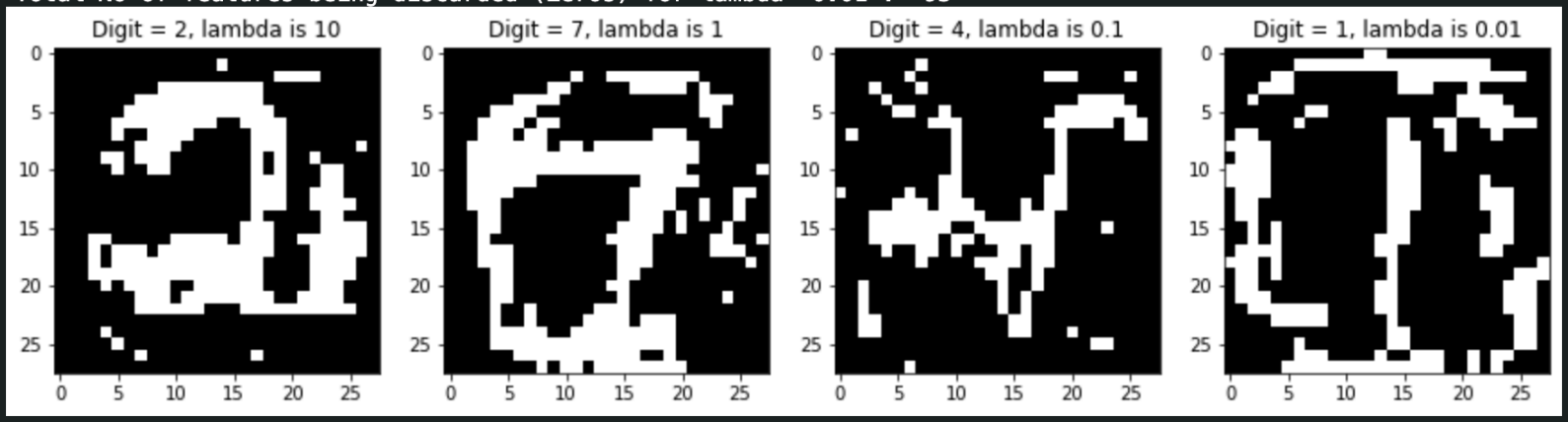
→ Larger value of λ has higher accuracy value



(d) For each λ we have a θ solution & each θ corresponds to 1 pixel.

Larger λ leads to fewer pixels being used for making predictions, where almost all pixels for $\lambda = 0.1$ & $\lambda = 0.01$ are used as shown in the figure.

```
Total No of features being discarded (Zeros) for lambda  10 :  257
Total No of features being discarded (Zeros) for lambda  1 :  170
Total No of features being discarded (Zeros) for lambda  0.1 :  114
Total No of features being discarded (Zeros) for lambda  0.01 :  93
```



(Q3) Proximal Operator for the group lasso regularizer

(A3) (a) $\|\theta_{g_i}\| = \sqrt{\sum_{c=1}^{k} \theta_{jc}^2}$

→ If $\theta_j \neq 0$, then partial differentiation is given as follows:

$$\frac{\partial \|\theta_i\|}{\partial \theta_{jc}} = \frac{\theta_{jc}}{\|\theta_i\|}$$

$$\Rightarrow \partial \|\theta_i\| = \frac{\theta_i}{\|\theta_i\|}$$

→ Else if $\theta_i = 0$,

subgradient $\partial \|\theta_i\|$

is an element of $\{z : \|z\| \leq 1\}$

→ This holds because, by definition, in order of $\partial \|\theta_i\|$ to be a subgradient of 'f' we must have

$\Rightarrow f(y) = \|y_2\| \geq f(x) + g^T(y-x) = g^T y$ $\forall y$

where $f(x) = \|\theta_j\|$,

$g = f \|\theta_j\|$ & $x = \theta_j$

$\Rightarrow \|y\| \geq g^T y$ holds true only if the condition for g is $\|g\| \leq 1$

$\Rightarrow \therefore$ The subdifferential is given as follows:

For $x \neq 0$, unique subgradient

$$g = \frac{\theta_j}{\|\theta_j\|}$$

&

For $x = 0$, subgradient θ_j is any element of $\{z : \|z\| \leq 1\}$

(6) minimize $\rho\|\theta_j\| + \frac{1}{2}\|\theta_j - \tilde{\theta}_j\|^2$

Differentiating equation w.r.t. θ_j we get:

$$\Rightarrow \rho v + \theta_j - \tilde{\theta}_j \rightarrow (1)$$

where $v = \begin{cases} \frac{\theta_j}{\|\theta_j\|} & : \theta_j \neq 0 \\ z : \|z\| \leq 1 & : \theta_j = 0 \end{cases}$

Now,

Equating (1) with 0 we get:

$$\rho v + \theta_j - \tilde{\theta}_j = 0$$

$$\Rightarrow \theta_j = \tilde{\theta}_j - \rho v$$