CAP 6610 Homework 3

Siddhant Mittal (6061-8545)

1. (a) Given:

$$P(y/x;\theta) \sim N(\phi^T\theta, \sigma^2)$$
$$P(\theta) \sim N(0, \sigma_\circ^2 I)$$

From this we can deduce this:

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

Substituting the above two values in given explicit MAP formulation $minimize \sum_{i=1}^{n} - \log p(y_i/x_i, \theta) - \log p(\theta)$ we get the following:

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

Since MAP depends on θ we remove the constant terms and we get:

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

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

$$\lambda = \frac{\sigma^2}{\sigma_\circ^2 I}$$

(b) Given:

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

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

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

Taking from part (a) $P(y_i/x_i;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\phi^T\theta - y_i)^2}{2\sigma^2}}$ and substituting in MAP formulation we get

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

$$minimize \sum_{i=1}^{n}(\phi^T\theta - y_i)^2 + \frac{2\sigma^2}{a}||\theta_j||_1$$

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

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

(c) Given:
$p(y/x,\theta) = Pr[yu \geq 0]$ where $p(u/x,\theta) \sim N(\phi^T\theta, \sigma^2)$
We can do the following:

$$p(y/x,\theta) = Pr[yu \geq 0]$$

$$p(y/x,\theta) = \frac{1}{\sqrt{2\pi}}\int_0^{\infty} exp(-\frac{(u - y\phi^T\theta)^2}{2})du$$

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

$$p(y/x,\theta) = \Phi(y\phi^T\theta)$$

Using part a we get

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

Substituting in MAP formula we get

$$minimize - \log(\Phi(y\phi^T\theta)) + \frac{1}{2\sigma_\circ^2 I}||\theta_j||_2^2$$

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

$$\lambda = \frac{1}{2\sigma_\circ^2 I}$$

(d) Given:

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

and taking from part (b) the value of $p(\theta)$ we can do following changes in our MAP formulation and ignoring constant terms we get:

$$minimize \sum_{i=1}^{n} y_i \phi^T \theta + \frac{||\theta||_1}{a}$$

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

$$\lambda = \frac{1}{a}$$

2. (a) We know that

$$Prox_f(\theta) = argmin f(\theta) + \frac{1}{2}||\theta - \theta_1||^2$$

To find proximal mapping for θ we need to minimize the above equation that means we can write

$$\theta = Prox_f(\theta_1)$$

Differentiating RHS and using sum rule of sub differentiation we get

$$\frac{\partial}{\partial t}[f(\theta) + \frac{1}{2}||\theta - \theta_1||^2] \in 0$$

$$\partial f(\theta) + \frac{1}{2} 2(\theta - \theta_1)(1 - 0) \in 0$$

$$\theta_1 - \theta \in \partial f(\theta)$$

Since $\theta = Prox_f(\theta_1)$ we put that in

$$\theta_1 - Prox_f(\theta_1) \in \partial(Prox_f(\theta_1))$$

(b) Given $g_1 \in \partial f(\theta_1)$ and $g_2 \in \partial f(\theta_2)$ we have to prove $(g_1 - g_2)^T(\theta_1 - \theta_2) \geq 0$. By defination of subgradient of $f(\theta_1)$ and $f(\theta_1)$ we have following

$$f(\theta) \geq f(\theta_1) + g_1^T(\theta - \theta_1)\forall\theta$$
$$f(\theta) \geq f(\theta_2) + g_2^T(\theta - \theta_2)\forall\theta$$

Since it it for all θ this can also be written as

$$f(\theta_2) \geq f(\theta_1) + g_1^T(\theta_2 - \theta_1)\forall\theta$$
$$f(\theta_1) \geq f(\theta_2) + g_2^T(\theta_1 - \theta_2)\forall\theta$$

Adding above two equation we get

$$g_1^T(\theta_2 - \theta_1) + g_2^T(\theta_1 - \theta_2) \leq 0$$

Simplifying gives

$$(g_1 - g_2)^T(\theta_1 - \theta_2) \geq 0$$

(c)

From part (a) we have $\theta_1 - Prox_f(\theta_1) \in \partial(Prox_f(\theta_1))$ and from part (b) we have $(g_1 - g_2)^T(\theta_1 - \theta_2) \geq 0$. Putting values of θ_1 and θ_2 in form of part (a) in part (b) equation we get

Putting

$$g_1 = \theta_1 - Prox_f(\theta_1)$$
$$g_2 = \theta_2 - Prox_f(\theta_2)$$

$$(\theta_1 - Prox_f(\theta_1) - (\theta_2 - Prox_f(\theta_2))T(\theta_1 - \theta_2) \geq 0$$

By simplyfyinbg this we get

$$(Prox_f(\theta_1) - Prox_f(\theta_2))^T(\theta_1 - \theta_2) \geq ||Prox_f(\theta_1) - Prox_f(\theta_2||^2)$$

(d)

By applying Cauchy-Schwartz inequality property $a^T b \geq ||a||||b||$ is $b \geq ||a||$

We can modify the part equation to get nonexpansiveness property.

$$(\theta_1 - \theta_2) \geq ||Prox_f(\theta_1) - Prox_f(\theta_2)||$$

3. (a) Following algorithm is derived for solving this problem.

Given objective function.

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

→ Randomly ~~choo~~ select a x_i.

→ Find the subgradient at x_i by differentiatio equation (1)

→ ~~Eqn~~. θ corresponding to correct class.

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

θ corresponding to incorrect class.

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

delta here is one

→ new θ (theta) is ~~for~~ proximal operatio

$$\theta^{t+1} \leftarrow \theta^t - \lambda(\text{subgradient})$$

→ New weight that is theta is

$\cancel{\text{Prox}(\theta^t - \lambda\nabla L)}$. $\text{Prox}(\theta^t - y^{(t)}\lambda\nabla L$

→ These steps are repeated 10^6 iteration for 4 lamba value $\{10, 1, 0.1, 0.01\}$.

→ $y(t)$ is the step size which is diminishing step size at each iteration, step size is $\frac{1}{t+1}$ (0 iteration is the start)

Proximal operator calculation:

Prox of $L_{2,1}$ Mixed Norm

The problem is given by:

$$\arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \lambda\|X\|_{2,1}$$

Where $X, Y \in \mathbb{R}^{m \times n}$.

Again, this can be decomposed into working on each column of X separately:

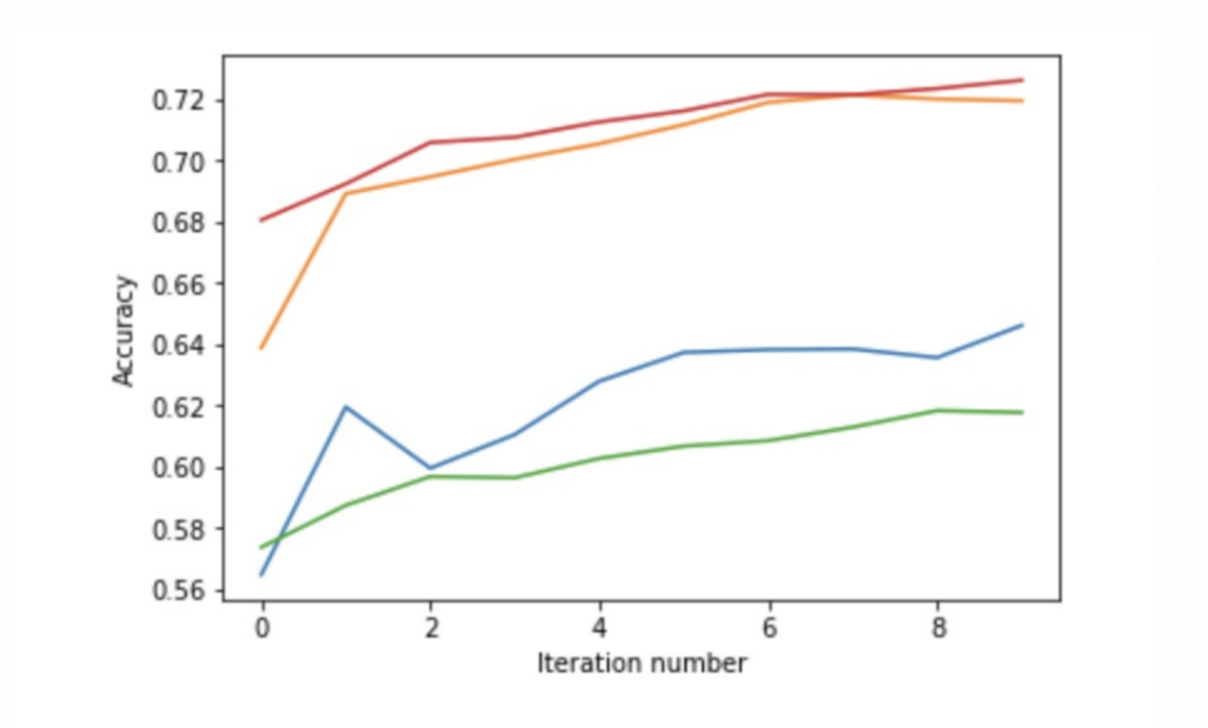
$$\begin{aligned}
\arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \lambda\|X\|_{2,1} &= \arg\min_X \sum_i \frac{1}{2}\|X_{:,i} - Y_{:,i}\|_2^2 + \sum_i \lambda\|X_{:,i}\|_2^2 \\
&= \arg\min_X \left(\frac{1}{2}\|X_{:,1} - Y_{:,1}\|_2^2 + \lambda\|X_{:,1}\|_2^2\right) \\
&+ \left(\frac{1}{2}\|X_{:,2} - Y_{:,2}\|_2^2 + \lambda\|X_{:,2}\|_2^2\right) \\
&+ \cdots \\
&+ \left(\frac{1}{2}\|X_{:,n} - Y_{:,n}\|_2^2 + \lambda\|X_{:,n}\|_2^2\right)
\end{aligned}$$

Each term in the brackets is independent Prox Function of L_2 Norm.
Hence the solution is given by:

$$\hat{X} = \arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \lambda\|X\|_{2,1}$$

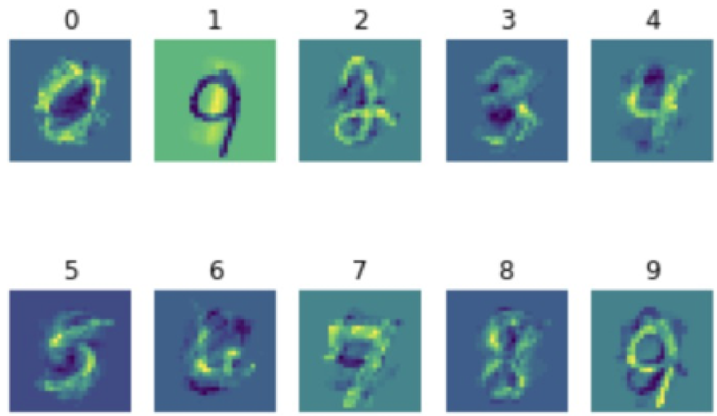
Where $\hat{X}_{:,i} = Y_{:,i}\left(1 - \frac{\lambda}{\max(\|Y_{:,i}\|_2, \lambda)}\right)$

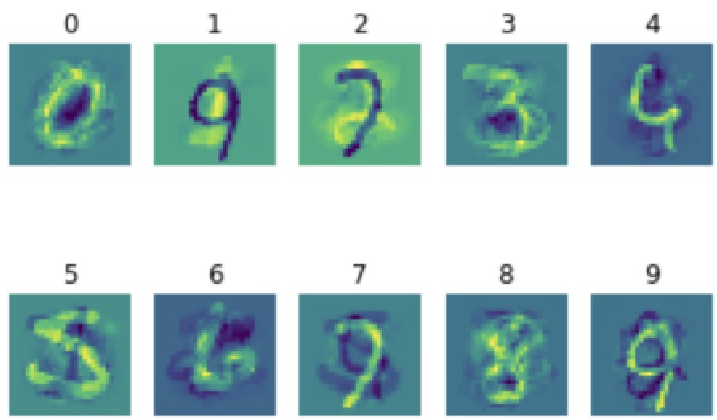
(c) Red line is lamba = 0.01, Green line is lamba = 0.1, Orange line is lamba = 1, Blue line is lamba = 10

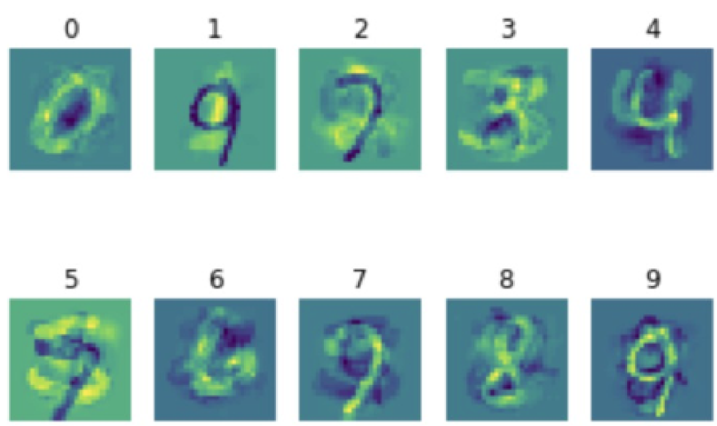


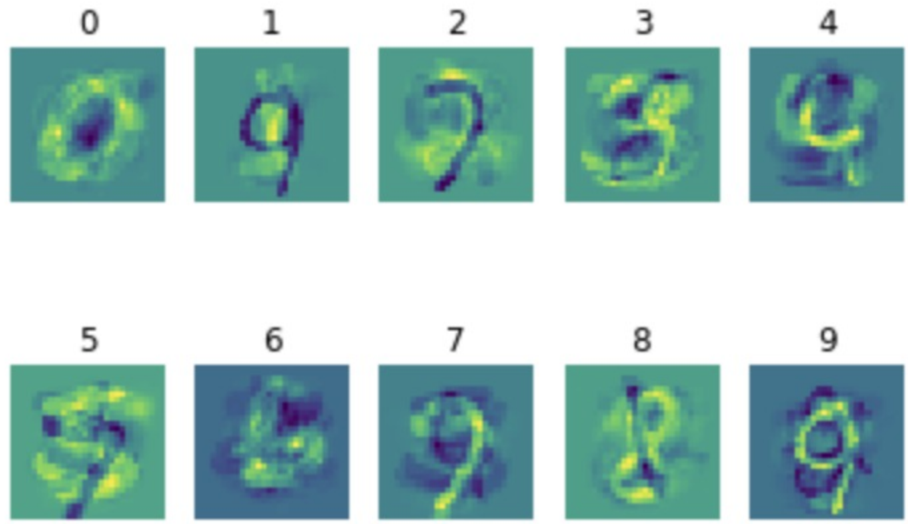
(d) Yes it true that a large leads to a more sparse solution. Below shows exactly the same.

Weights are the order of lamba 10,1,0,1,0.01 respectively


0 1 2 3 4
5 6 7 8 9


0 1 2 3 4
5 6 7 8 9


0 1 2 3 4
5 6 7 8 9


0
1
2
3
4
5
6
7
8
9