1. a) $\min_{\theta} \sum_{i=1}^{n} -\log p(y_i|x_i,\theta) - \log p(\theta)$

$p(y|x,\theta) \sim \mathcal{N}(\phi^T\theta, \sigma^2)$ $\quad$ $p(\theta) \sim \mathcal{N}(0, \sigma_0^2 I)$

applying distribution to the function,

$$\min_{\theta} \sum_{i=1}^{n} -\log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\phi^T\theta)^2}{2\sigma^2}\right)\right] - \log\left[\frac{1}{\sqrt{(2\pi\sigma_0^2 I)}} \exp\left(-\tfrac{1}{2}(\theta-0)^T (\sigma_0^2 I)^{-1}(\theta-0)\right)\right]$$

$$\downarrow\downarrow$$

$$\frac{1}{(2\pi\sigma_0^2)^{d/2}} \exp\left(-\frac{1}{2\sigma_0^2}\|\theta\|^2\right)$$

$$\min_{\theta} \sum_{i=1}^{n}\left(\underbrace{\tfrac{1}{2}\log(2\pi\sigma^2)}_{\text{const}} + \frac{(y_i-\phi^T\theta)^2}{2\sigma^2} + \underbrace{\frac{d}{2}\log(2\pi\sigma^2)}_{\text{const}} + \frac{1}{2\sigma_0^2}\|\theta\|^2\right)$$

$$\min_{\theta} \sum_{i=1}^{n} \frac{(y_i-\phi^T\theta)^2}{2\sigma^2} + \frac{n}{2\sigma_0^2}\|\theta\|^2 = \min_{\theta} \underbrace{\sum_{i=1}^{n}(y_i-\phi^T\theta)^2}_{\text{quadratic loss}} + \underbrace{\frac{n\sigma^2}{\sigma_0^2}\|\theta\|^2}_{\text{ridge}}$$

So, $\boxed{\lambda = \frac{n\sigma^2}{\sigma_0^2}}$

b) $p(y|x,\theta) \sim \mathcal{N}(\phi^T\theta, \sigma^2)$ $\quad$ $p(\theta) = \prod_{j=1}^{m} \frac{1}{2a}\exp\left(-\frac{|\theta_j|}{a}\right)$

$$\min_{\theta} \sum_{i=1}^{n}\left(\tfrac{1}{2}\log(2\pi\sigma^2) + \frac{(y_i-\phi^T\theta)^2}{2\sigma^2} - \left(\sum_{j=1}^{m}\left(\underbrace{-\log(2a)}_{\text{const}} - \frac{|\theta_j|}{a}\right)\right)\right)$$

$$\min_{\theta} \sum_{i=1}^{n}\left(\frac{y_i-\phi^T\theta}{2\sigma^2}\right)^2 + \frac{n}{a}\|\theta_j\|_1 \quad \left[\sum_{j=1}^{m}|\theta_j| \rightarrow l_1 \text{ norm}\right]$$

So, $\min_{\theta} \underbrace{\sum_{i=1}^{n}(y_i-\phi^T\theta)^2}_{\text{quadratic loss}} + \underbrace{\frac{2n\sigma^2}{a}\|\theta\|_1}_{\text{lasso}}$

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

c) $p(y|x,\theta) = P_\theta[yu \geq 0]$ where $y = \pm 1$, $p(u|x,\theta) \sim \mathcal{N}(\phi^T\theta, \sigma^2)$

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

$$p(y|x,\theta) = P_\theta[yu \geq 0] = \frac{1}{2\pi\sigma^2}\int_0^\infty \exp\left(-\frac{1}{2}\frac{(u-\phi^T\theta)^2}{\sigma^2}\right) du$$

$$= \frac{1}{2\pi\sigma^2}\int_{-\infty}^{y\phi^T\theta} \exp\left(-\frac{u^2}{2}\right) du = \Phi(y\phi^T\theta)$$

So, minimizing function $\min_\theta \sum_{i=1}^n -\log \Phi(y_i \phi_i^T\theta) - \frac{n}{2\sigma_0^2}\|\theta\|^2$

$-\log \Phi(y_i \phi_i^T\theta)$ → probit loss; λ → from derivation in a.

$$\boxed{\lambda = \frac{n}{2\sigma_0^2}}$$

d) $p(y|x,\theta) = \frac{1}{1+\exp(-y\phi^T\theta)}$ where $y = \pm 1$, $p(\theta) = \prod_{j=1}^m \frac{1}{2a}\exp\left(-\frac{|\theta_j|}{a}\right)$

So, minimizing function, $\min_\theta \sum_{i=1}^n \left(-\log\left(1/1+\exp(-y_i\phi^T\theta)\right) - \sum_{j=1}^m -\log 2a - \frac{|\theta_j|}{a}\right)$

$\sum_{j=1}^m$ → const

$$\min_\theta \sum_{i=1}^n \log\left(1+\exp(-y_i\phi^T\theta)\right) + \sum_{i=1}^n \|\theta\|_1$$

$$\Rightarrow \min_\theta \sum_{i=1}^n \log\left(1+\exp(-y_i\phi^T\theta)\right) + n\|\theta\|_1$$

$\log(1+\exp(-y_i\phi^T\theta))$ → logistic loss; n → λ

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

By definition, $Prox_f(\theta_1)$ minimizes the problem. From optimality,

$$0 \in \partial f(Prox_f(\theta_1)) + Prox_f(\theta_1) - \theta_1$$

a)

$$\Rightarrow \theta_1 - Prox_f(\theta_1) \in \partial f(Prox_f(\theta_1)) \quad \text{— Proved.} \quad (1)$$

b) $g_1 \in \partial f(\theta_1)$ $\qquad$ $g_2 \in \partial f(\theta_2)$

g_1 is a subgradient of f at θ_1, g_2 is a subgradient of f at θ_2.

Now, subgradient is a global underestimator.

$$f(\theta_2) \geq f(\theta_1) + g_1^T(\theta_2 - \theta_1) \quad -(i)$$

$$f(\theta_1) \geq f(\theta_2) + g_2^T(\theta_1 - \theta_2) \quad -(ii)$$

adding (i) and (ii), $0 \geq g_1^T(\theta_2 - \theta_1) + g_2^T(\theta_1 - \theta_2)$

So, $(g_1 - g_2)^T(\theta_1 - \theta_2) \geq 0$ — Proved. (ii)

c) From a and b,

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

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

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

— Proved

d) $\|Prox_f(\theta_1) - Prox_f(\theta_2)\|^2 \leq \|Prox_f(\theta_1) - Prox_f(\theta_2)\| \cdot \|\theta_1 - \theta_2\|$

(Applying Cauchy Schwarz)

$$\therefore \|Prox_f(\theta_1) - Prox_f(\theta_2)\| \leq \|\theta_1 - \theta_2\|$$

3. a)

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

At iteration t,

For each class $c = 1 \ldots K$

(1) Pick a sample (i) from class c train data

(2) Pick gradient based on data sample i,

gradient based on $\nabla \theta_{y_i} L_i = -\left(\sum_{c \neq y_i} 1\left[x_i^T\theta_c - x_i^T\theta_{y_i} + 1 > 0\right)\right) x_i$

$$\nabla_{\theta_j} L_i = 1\left(x_i^T\theta_c - x_i^T\theta_{y_i} + 1 > 0\right) x_i$$

Update θ accordingly.

$[\delta^{(t)} = 1/t]$

$$\theta_{y_i} = \theta - 1/t \; \nabla \theta_{y_i} L_i$$

$$\theta_j = \theta_j - 1/t \; \nabla \theta_j L_i$$

(3) Apply Proximal operator,

For $j = 1$ to m,

$\theta_j^{(t+1)} = \text{Prox}_{\delta\lambda}\, \theta_j^{(t)}$, where $\text{Prox}_{\delta\lambda}(\theta) = \begin{cases} \left(1 - \frac{\delta\lambda}{\|\theta\|}\right)\theta & \text{—(i)} \\ 0 & \text{—(ii)} \end{cases}$

(i) $\|\theta_j\| > \delta\lambda$

(ii) $\|\theta_j\| \leq \delta\lambda$

d) Upon the implementation of our algo, the feature vector 784, we see we are getting different count of 0 vectors depending on regularization parameter.

(i) $\lambda = 10$, 249 features are discarded.

(ii) $\lambda = 1$, 179 features are discarded.

(iii) $\lambda = .1$, 116 features are discarded.

(iv) $\lambda = .01$, 91 features are discarded.