1a) $P(x=x) = \left(\frac{1}{1+e^{-\lambda}}\right)^{[x=1]} \left(\frac{1}{1+e^{\lambda}}\right)^{[x=0]}$ $\quad x \in \{0,1\}$

$P(x=0) + P(x=1) = \frac{1}{1+e^{+\lambda}} + \frac{1}{1+e^{-\lambda}}$

$= \frac{1}{1+e^{\lambda}} + \frac{1}{1+1/e^{\lambda}} = \frac{1}{1+e^{\lambda}} + \frac{e^{\lambda}}{1+e^{\lambda}} = 1$

Thus this distribution is properly normalized.
Also note $P(x=0)$ is positive since $\frac{1}{1+e^{\lambda}} > 0 \quad \forall \lambda \in R$
and similarly $P(x=1)$ is positive since $\frac{1}{1+e^{-\lambda}} > 0 \quad \forall \lambda \in R$

1b) Given data set D containing a 1's and b 0's.
Maximum likelihood $= \prod_{n=1}^{N} P(z = z_n | \theta)$

Negative log likelihood $= -\sum_{n=1}^{N} P(z = z_n | \theta)$ which must be minimized.

∴ Minimizing negative log likelihood analytically.

$$L(D|\theta) = \left(\frac{1}{1+e^{-\lambda}}\right)^{a} \left(\frac{1}{1+e^{\lambda}}\right)^{b}$$

Negative log likelihood $\ell(D,\theta) = -\left(a \log\left(\frac{1}{1+e^{-\lambda}}\right) + b\left(\frac{1}{1+e^{\lambda}}\right)\right)$

$= -\left(-a \log(1+e^{-\lambda}) - b \log(1+e^{\lambda})\right)$

Equating gradient to zero. $\frac{d}{d\lambda} \ell(z,\theta) = 0$

$\Rightarrow \frac{a}{1+e^{-\lambda}} = \frac{b}{1+e^{\lambda}}$

$\Rightarrow \frac{a e^{\lambda}}{1+e^{\lambda}} = \frac{b}{1+e^{\lambda}} \Rightarrow e^{\lambda} = \frac{b}{a} \Rightarrow \lambda = \log\frac{b}{a}$

MLE

1c) Current distribution : $\left(\frac{e^{\lambda}}{1+e^{+\lambda}}\right)^{[x=1]} \left(\frac{1}{1+e^{\lambda}}\right)^{[x=0]}$

Standard parameterization : $\theta^{z}\ (1-\theta)^{1-z}$

If we replace θ with $\frac{e^{\lambda}}{1+e^{\lambda}}$ in the standard parameterization of the Bernoulli distribution, we get the current distribution $\left(\theta = \frac{e^{\lambda}}{1+e^{\lambda}}\right)$

2a) $\nabla \mathcal{L}(D,\theta) = \begin{bmatrix} \frac{d}{dw}\mathcal{L}(D,\theta) \\ \frac{d}{db}\mathcal{L}(D,\theta) \end{bmatrix}$

$$\frac{d}{dw}\mathcal{L}(D,\theta) = \sum_{n=1}^{N}\left(\frac{1}{1+e^{-y_n(wx_n^T+b)}}\right)\left(e^{-y_n(wx_n^T+b)}\right)(-y_n x_n^T) + 2\lambda|w-w_0|$$

$$\frac{d}{db}\mathcal{L}(D,\theta) = \sum_{n=1}^{N}\left(\frac{1}{1+e^{-y_n(wx_n^T+b)}}\right)\left(e^{-y_n(wx_n^T+b)}\right)(-y_n) + 2\lambda(b-b_0)$$

(Note: $\|w-w_0\|_2^2 = (w-w_0)^T(w-w_0) = w^Tw - w^Tw_0 - w_0^Tw + w_0^Tw_0$ whose derivative is $2|w-w_0|$ based on vector calculus)

3a) $\mathcal{L}(D,\theta) = \sum_{n=1}^{N} L_{k,\beta}(y_n, f(x_n,\theta))$ where $f(x_n,\theta) = wx_n^T + b$

$$\therefore \mathcal{L}(D,\theta) = \begin{cases} \sum_{n=1}^{N} \frac{1}{2k}(y_n - wx_n^T - b)^{2k} & \text{if } |y_n - wx_n^T - b| \le \beta \\ \sum_{n=1}^{N} \beta^{2k-1}\left(|y_n - wx_n^T - b| - \frac{2k-1}{2k}\beta\right) & \text{otherwise} \end{cases}$$

$$\nabla\mathcal{L}(D,\theta) = \begin{bmatrix} \frac{d}{dw}\mathcal{L}(D,\theta) \\ \frac{d}{db}\mathcal{L}(D,\theta) \end{bmatrix}$$

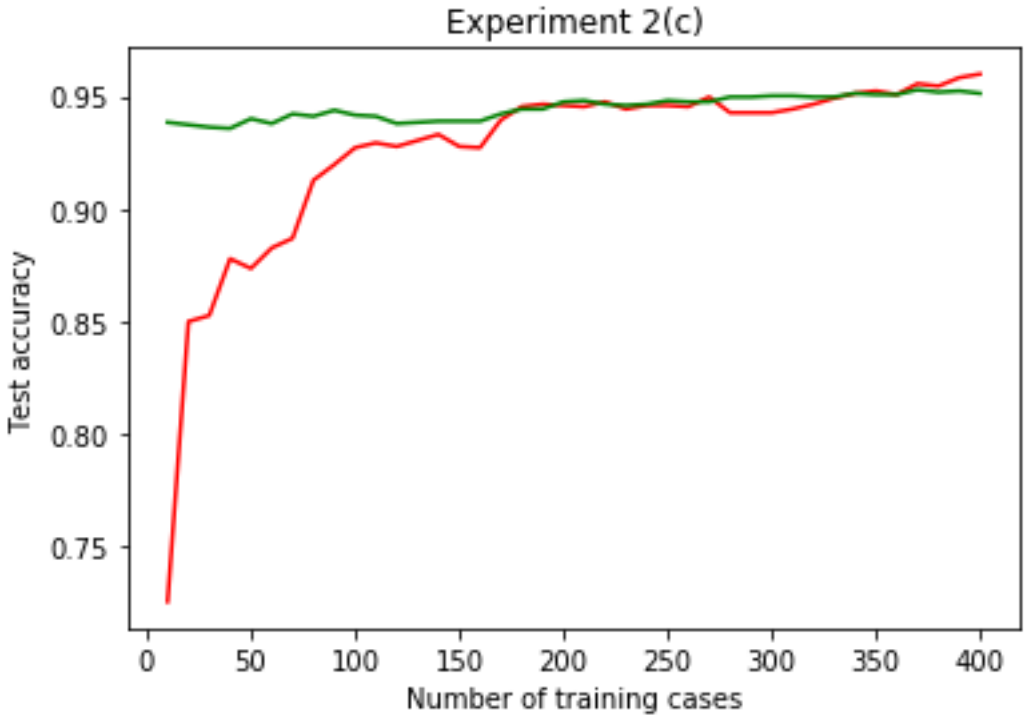
$$\frac{d}{dw}\mathcal{L}(D,\theta) = \begin{cases} \sum_{n=1}^{N} \frac{1}{2k}(2k)(y_n - wx_n^T - b)^{2k-1}(-x_n^T) & \text{if } |y_n - wx_n^T - b| \le \beta \\ \sum_{n=1}^{N} \beta^{2k-1}(-x_n^T) & \text{if } [y_n - wx_n^T - b] > 0 \\ \sum_{n=1}^{N} \beta^{2k-1}(x_n^T) & \text{if } [y_n - wx_n^T - b] < 0 \end{cases}$$

$$\frac{d}{db}\mathcal{L}(D,\theta) = \begin{cases} \sum_{n=1}^{N} \frac{1}{2k}(2k)(y_n - wx_n^T - b)^{2k-1}(-1) & \text{if } |y_n - wx_n^T - b| \le \beta \\ \sum_{n=1}^{N} \beta^{2k-1} & \text{if } [y_n - wx_n^T - b] > 0 \\ \sum_{n=1}^{N} -\beta^{2k-1} & \text{if } [y_n - wx_n^T - b] < 0 \end{cases}$$

(↑ Note: These are just brackets, not a step function)

2b) The given learning objective represents logistic regression with a non-zero mean spherical Gaussian prior. Since we are given a prior, MAP estimation must be used to learn the model parameters. The solution to MAP estimation is an optimization problem which we solved through numerical approximation using the fmin_l_bfgs_b method.

2c) The figure below represents two line plots showing test accuracy as a function of the number of training data cases for lambda = 0 and lamda = 10



Legend: Red Line: Lambda = 0, Green Line: Lambda = 10

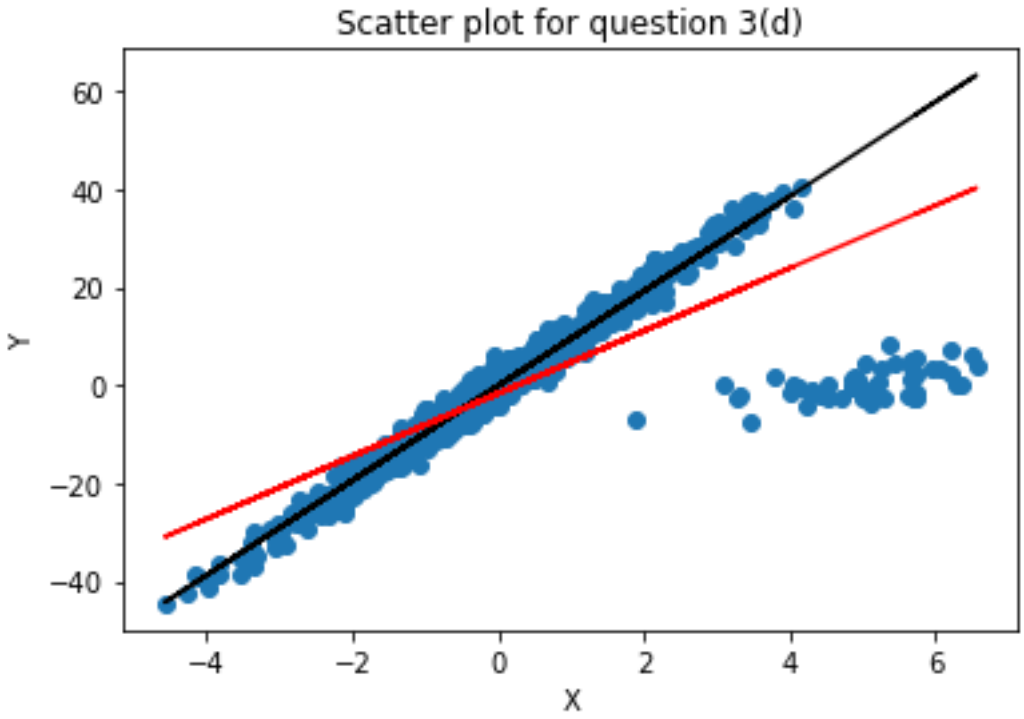
2d) It is known that MAP estimation performs far better than MLE when N (number of samples) is small. Setting lamda to zero is equivalent to computing MLE as we remove influence of prior on the objective function. When lambda is set to 10, the objective function represents MAP estimation and therefore we see better accuracy for small N. However, as N increases, MAP and MLE converge so we do not see the influence of lambda or equivalently the prior for large N.

3b) The given learning objective represents a loss minimization problem which falls under the empirical risk minimization class of problems. The parameters for the model are learnt by minimizing the loss between expected and observed outputs. We optimized the parameters through fmin_l_bfgs_b numerical approximation method which approximates the hessian to derive the optimal values.

3c) The MSE values are:

- Robust Regression MSE = 115.834685804
- Sklearn Linear Regression MSE = 76.0431153827

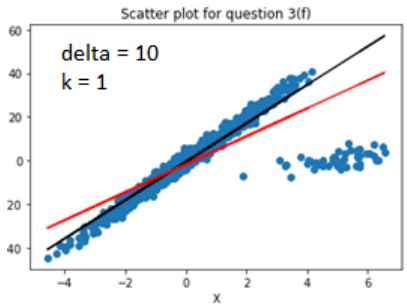
3d)

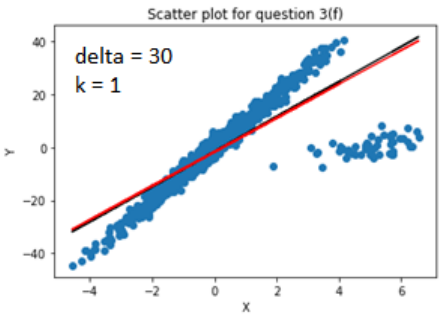


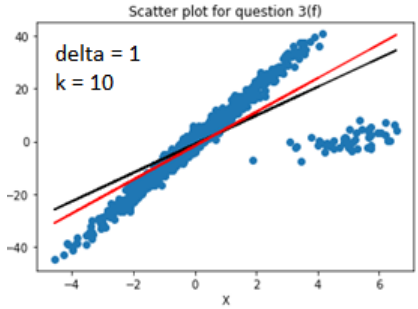
Legend: Blue Dots: data points, Black Line: Robust regression, Red Line: Linear regression

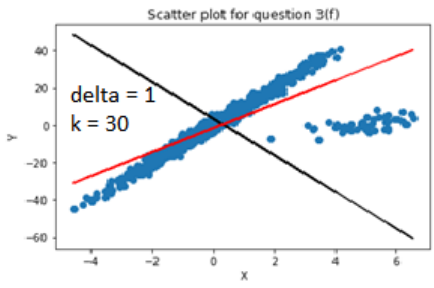
3e) Based on the plot, it is clear that the robust regression model would have better generalization performance on the future test data. As can be seen above, the robust regression model is less influenced by outliers. Although the MSE for robust regression is higher on the training data, it tends to offer better generalized prediction as the robust regression line passes through the bulk of the mass in the scatter plot.

3f) A few scatter plots obtained by varying the values of delta and k are given below.









As k or delta increases, the robust regression line appears to move closer towards the linear regression line. The impact of k is much more significant than that of delta, owing to a part of the objective loss function being represented as a polynomial to the degree 2k.

The ideal values for robust linear regression seem to be delta < 1 and k = 1 or 2.