1) Logistic Regression: Training stability
   a) Dataset A converges very soon, but dataset B doesn't seem to converge even after long time learning.

   b) The problem is with the fact that dataset B is strictly separable, which causes the loss to be unbounded below (need solid math derivation), which means theta is going to grow infinitely to try to get a better loss. This can be easily demonstrated by printing the loss function result: $\ell(\theta) = \Sigma_1^{\mathrm{m}}(y^{(i)} \log p^{(i)} + (1 - y^{(i)} \log(1 - p^{(i)}))$where p=h(x). There are many ways to fix problems, one of the fastest way is to simply randomly change one of the labels in the dataset (change from -1 to 1 or vice versa), and the training will converge very soon.

   c) Adding L2 regularization helps. Actually adding (theta / m) to 'grad' in function calc_grad, and then change the learning rate to 1 would make the convergence happen in less than 2000 iterations for both dataset A and dataset B.
   Adding noise to the dataset would bring misclassification to the problem, which means it is no longer strictly separable, which in turn means it also helps. (??)
   Others are not relevant to the separation issue.

   d) In SVM, θ is going to be normalized, strict separation issue would no longer be a problem. (Again, Solid Math derivation is needed)

   TBC

2) Model Calibration
   a) Prove calibration condition holds true for LR over range (a, b) = (0, 1)

$$\ell(\theta) = \sum_1^m y^{(i)} \log p^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - p^{(i)}\right)$$

$$p^{(i)} = g(z) \Rightarrow \nabla_\theta {p^{(i)}}' = g(z)\left(1 - g(z)\right){x^{(i)}}^T$$

$$\nabla_\theta \ell(\theta) = \sum_1^m (y^{(i)} p^{(i)} \, (1 - p^{(i)})/p^{(i)} \; + (1 - y^{(i)})(-p^{(i)})(1 - p^{(i)})/(1 - p^{(i)}))x^{(i)^T}$$

$$\nabla_\theta \ell(\theta) = \sum_1^m (y^{(i)} \, (1 - p^{(i)}) \; + (1 - y^{(i)})(-p^{(i)}))x^{(i)^T} = \vec{0} \Rightarrow$$

$$\sum_1^m y^{(i)} \, x^{(i)^T} = \sum_1^m p^{(i)} x^{(i)^T}$$

   Vectorize: $yX = pX$, with $y = \left[y^{(0)}, y^{(1)}, \ldots, y^{(m)}\right], p = [p^{(0)}, p^{(1)}, \ldots p^{(m)}]$ and $X \in \mathbb{R}^{m \times n}$

   Use the fact that we include a bias term: $x_0^{(i)} = 1$ we got all 1s for the first column of X:

$$\sum_{1}^{m} y^{(i)} = \sum_{1}^{m} p^{(i)} \Rightarrow \frac{\sum_{i \in I_{0,1}} P\left(y^{(i)} = 1 \middle| x^{(i)}; \theta\right)}{\left|i \in I_{0,1}\right|} = \frac{\sum_{i \in I_{0,1}} 1\{y^{(i)} = 1\}}{\left|i \in I_{0,1}\right|}$$

b) Perfect calibration doesn't mean perfect accuracy. If for any $(\mathrm{a}, \mathrm{b}) \subset [0,1]$ the property in the question holds, by switching two samples with different probabilities, the calibration is still the same. Conversely if the model achieves perfect accuracy, it is perfectly calibrated. This can be explained by clapping (a, b) to the probability of every single sample, the calibration equation holds.

c) What effect including L2 regularization in the logistic regression objective has on model calibration
L2 regularization filters high variance by penalizing on $\left|\left|\theta\right|\right|_2$. It introduces a degree of uncertainty in the vicinity of the decision boundary which is smoothed by the L2 regularization, which means the objective cost $\mathrm{h}(\theta)$ is changed (closer to 0.5). This will change the model calibration.

3) Bayesian Logistic Regression and weight decay

Prove that $\left|\left|\theta_{MAP}\right|\right|_2 \leq \left|\left|\theta_{ML}\right|\right|_2$given:

$$\theta_{ML} = \arg\max_{\theta_{ML}} \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta\right)$$
$$\theta_{MAP} = \arg\max_{\theta_{MAP}} p(\theta) \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta\right)$$
$$\theta \sim \mathcal{N}(0, \tau^2 I)$$

Prove: suppose $\left|\left|\theta_{MAP}\right|\right|_2 > \left|\left|\theta_{ML}\right|\right|_2$ then from definition of θ_{MAP}:

$$p(\theta_{MAP}) \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta_{MAP}\right) > p(\theta_{ML}) \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta_{ML}\right)$$

Yet from the definition of θ_{ML}, $\Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta_{ML}\right) \geq \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta_{MAP}\right)$, we get:

$$p(\theta_{MAP}) \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta_{MAP}\right) > p(\theta_{ML}) \Pi_{i=1}^{m} p\left(y^{(i)} | x^{(i)}; \theta_{MAP}\right) \Rightarrow$$

$$p(\theta_{MAP}) > p(\theta_{ML})$$

Which contradicts $\left|\left|\theta_{MAP}\right|\right|_2 \leq \left|\left|\theta_{ML}\right|\right|_2$ and $\theta \sim \mathcal{N}(0, \tau^2 I)$. Proved.