# 6.036 Fall 2018 Midterm Review
# **Solutions**

## 1 Spring 2013: Problem 1

1.1a) Here are plots of $\theta$ and the decision boundary $\theta \cdot x = 0$, obtained by simply running the perceptron algorithm on each point sequentially, with the initial value of $\theta$ as 0.

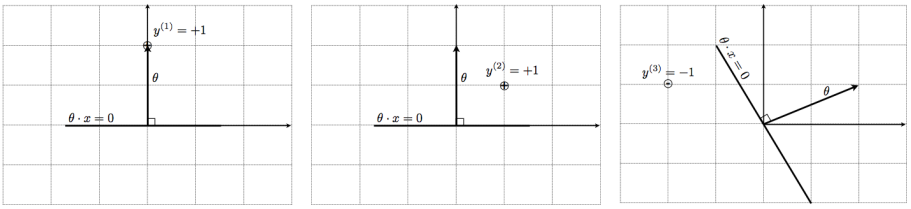



Figure 1: Problem 1.1a

1.1b) Here is one possible assignment of labels to the points such that the desired properties are satisfied.

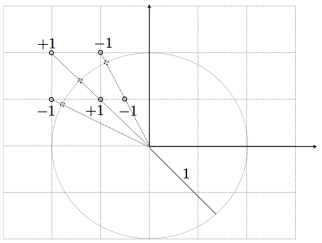



Figure 2: Problem 1.1b

**Explanation**: Observe that these original five points are not linearly separable, so perceptron wouldn't converge. Applying the feature mapping

$\phi(x) = x/\|x\|$ is just a fancy way of saying "project each point onto the unit circle." Observe that these projected points (marked on the circle) are indeed linearly separable, so perceptron would converge on the featurized data.

# 2 Spring 2013: Problem 3

3.1a) Statements 1, 3, and 4 should be marked (TRUE).

**Explanations**:
Statement 1 is TRUE because it is the optimality condition: we are just saying that the gradient of $J(\theta)$ at $\hat{\theta}$ is zero.

Statement 2 is FALSE because the greater than sign should be a less than sign: the optimal $\hat{\theta}$ minimizes $J(\theta)$, not maximizes it!

Statement 3 is TRUE because increasing $\lambda$ means we enforce more regularization.

Statement 4 is TRUE because we can always add frivolous features without increasing the training error for the optimal $\hat{\theta}$ (for example, we could set the coefficents of $\hat{\theta}$ corresponding to those features to 0). However, note that it may take longer for our learning algorithm to *find* this $\hat{\theta}$!

3.1b) A good classifier here would be $y = 1$ iff $\hat{\theta} \cdot \phi(x) \geq 0.5$.

**Explanation**: What you *don't* want to do is threshold at 0, i.e. $y = 1$ iff $\hat{\theta} \cdot \phi(x) \geq 0$. This is because the target ratings (training labels) are 0 or 1, so the regression function that we learn will tend to predict values that are between 0 and 1. This means we should use the midpoint value 0.5 as our threshold.

3.1c) The predictions will tend towards 0.

**Explanation**: If we increase $\lambda$, then $\|\hat{\theta}\|$ will decrease. As a result, the regression function values $\hat{\theta} \cdot \phi(x)$ will tend towards zero. Given the decision rule above, the predictions are going to become biased towards $y = 0$.

# 3 Spring 2014: Problem 3

3.1a) Yes. See the plot.

3.1b) No. See the plot.

3.1c)+3.1di) Here's a plot of the featurized data, $\hat{\theta}$, and separator.

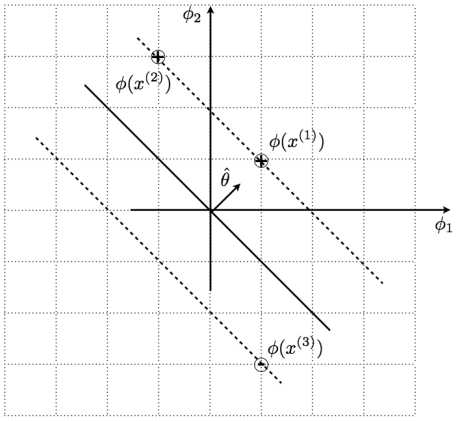



Figure 3: Problems: 3.1c and 3.1d(i)

3.1dii) The value of the margin is $\sqrt{2}$.

**Explanation**: the minimum distance from any data point to the separator, which you can see from the above plot.

3.1diii) $\|\hat{\theta}\| = \frac{1}{\sqrt{2}}$.

**Explanation**: Remember that in the SVM objective, we are secretly encoding the margin as $\frac{1}{\|\theta\|}$. So we have $\|\hat{\theta}\| = \frac{1}{\sqrt{2}}$, based on the previous answer.

3.1e) Please see the plot.

**Explanation**: We can solve for this analytically. We have $\hat{\theta} = [0.5, 0.5]^\top$ for the featurized data, which we found by looking at the above plot. That means the decision boundary is $\hat{\theta} \cdot \phi(x) = 0.5x_1 + 0.5x_2x_1 = 0$, which is solved by either $x_1 = 0$ or $x_2 = -1$. So, this is a non-linear decision boundary in the original feature space! And here's the plot.

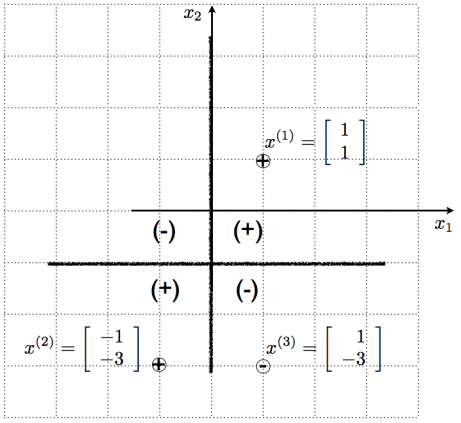



Figure 4: Problem: 3.1e

# 4 Spring 2016: Problem 1

1.1) No.

**Explanation**: For example, suppose that the first coordinate had the same value across all data points. Then, eliminating it would not change whether or not the data are linearly separable!

1.2) No.

**Explanation**: Perceptron does not do gradient descent on any particular loss function.

1.3) Objective Function $= 2 + \frac{1}{3}$.
Yes there is a solution which has lower loss and smaller $||\theta||$.

**Explanation**: Note that the margin is $\frac{1}{2}$ here, as shown by the dotted lines. Remember that in the SVM objective, we are secretly encoding the margin as $\frac{1}{\|\theta\|}$. So we have $\|\theta\| = 2$.

Now we need to find the loss. The top and bottom points are classified correctly and are outside the dotted lines, so they incur 0 loss. The + point that's on the decision boundary incurs loss 1 because $\theta \cdot x = 0$ for that point, and hinge loss is $1 - y\theta \cdot x$ if $y\theta \cdot x < 1$.

Putting it all together, we get $\frac{1}{n}(0 + 0 + 1) + \frac{\lambda}{2}\|\theta\|^2 = \frac{1}{3} + 2$.

A better separator would be one that is shifted down a bit (but maintains the same slope), so it passes equidistant between the lower + and the −.

1.4) The third option (labelled as (4) in the list) should be marked.

**Explanation**: Note that Statement 4 is wrong because we are doing stochastic gradient descent, so no need to divide by $n$ (the total number of data points).

# 5 Spring 2016: Problem 2

2.1) $\hat{\theta} = 0$.

**Explanation**: We are applying infinite regularization, so all we care about is that $\|\hat{\theta}\|$ is as small as possible. **Explanation**:

2.2) You only need to have the right shape of your plot.

**Explanation**: Training error is smallest when $\lambda = 0$ and increases with $\lambda$.

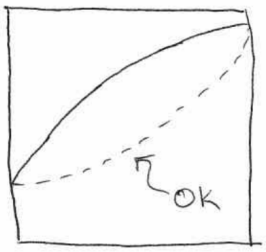
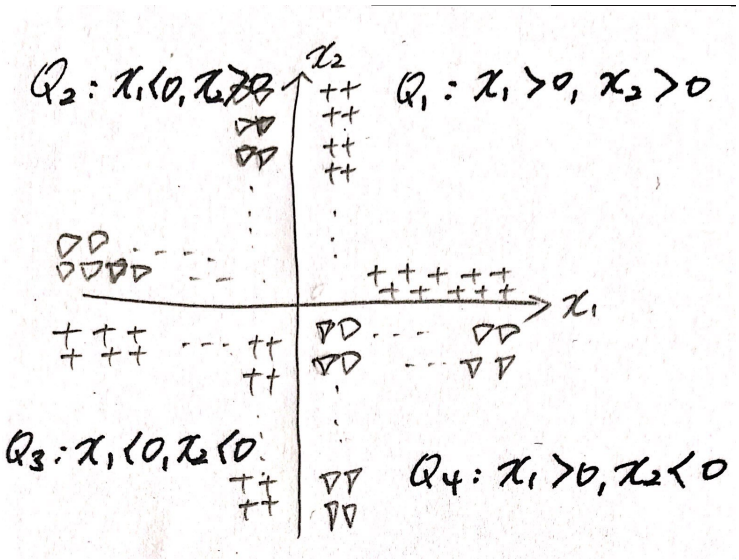



Figure 5: Problem 2.2

# 6 Spring 2016: Problem 4

4.1) (a) and (d) should be marked.

**Explanation**: Note that the points labeled '+' live in the first and third quadrants (Q1, Q3), while the points labeled $\triangledown$ live in the second and fourth quadrants (Q2, Q4).

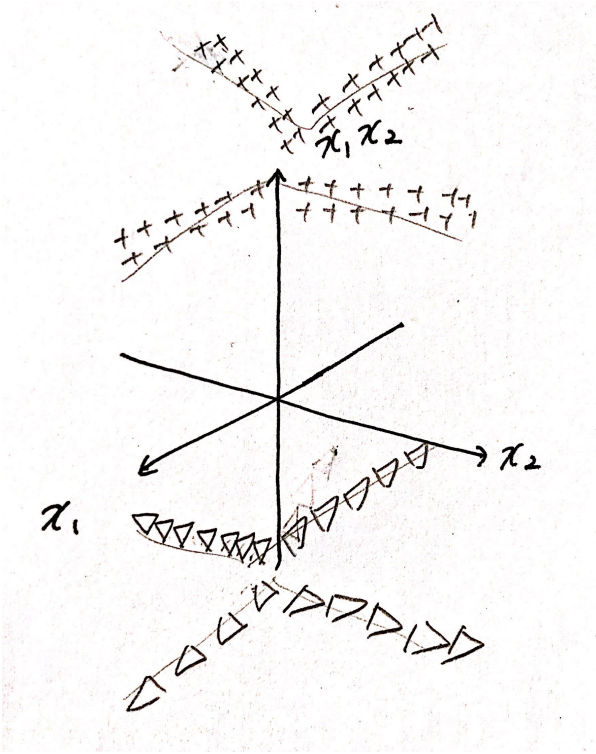


Also note that:

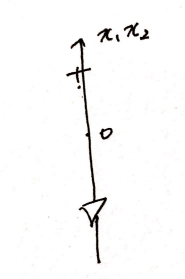
- In Q1: $x_1 > 0, x_2 > 0$ so $x_1 x_2 > 0$
- In Q2: $x_1 < 0, x_2 > 0$ so $x_1 x_2 < 0$
- In Q3: $x_1 < 0, x_2 < 0$ so $x_1 x_2 > 0$
- In Q4: $x_1 > 0, x_2 < 0$ so $x_1 x_2 < 0$

Thus, $x_1 x_2 > 0$ in Q1 and Q3 and $x_1 x_2 < 0$ in Q2 and Q4. Now let's go through each of the choices (a),(b),(c),(d) and see if data is linearly separable in each of the cases.

(a) We can see $[x_1, x_2, x_1 x_2]$ as the third dimension "height" added to the original picture above. Because $x_1 x_2 > 0$ in Q1 and Q3 and $x_1 x_2 < 0$ in Q2 and Q4, $[x_1, x_2, x_1 x_2]$ looks like the following:



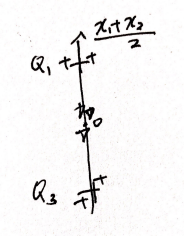

Or if we only look at the "height" dimension, like the following:




Thus, because of this "height" dimension created by $xy$, **there is a linear boundary (something like $xy = 0$ ) that linearly separates given data.**

(b) Note that $[x_1^2, x_2^2]$ maps all data to Q1. Because + and $\triangledown$ labeled points are (almost) symmetric with respect to the $x_1, x_2$ axis, + and $\triangledown$ labeled points may be mapped to (almost) same location in Q1. Thus, the first two dimensions do not help in any way for linear separability.

$\frac{x_1+x_2}{2}$ is always positive for points in Q1, always negative for points in Q3, can be either positive/negative/0 for points in Q2 and Q4.




As in the picture above, there is no linear boundary in the third dimension that linear separates data. Overall, data is **not linearly separable.**

(c) $tanh(x_1 + x_2)$ is always positive for points in Q1, always negative for points in Q3, can be either positive/negative/0 for points in Q2 and Q4. Similarly as in (b), data is **not linearly separable**.

(d) We saw in (a) that there exists a linear boundary in the direction of $x_1 x_2$ that linearly separates given data. Thus, data is **linearly separable**.

# 7 Spring 2016: Problem 5

5.1) $x \leq -1$.

**Explanation**: Because $f_1 = ReLU(z_1)$, $f_1 = 0$ iff $z_1 \leq 0$.

$$z_1 \leq 0 \Leftrightarrow xw_{11} + w_{01} \leq 0 \Leftrightarrow x + 1 \leq 0 \Leftrightarrow x \leq -1$$

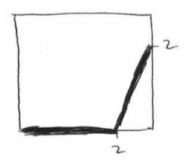



Figure 6: Problem 5.2

5.2) **Explanation**: From **5.1)**, we know that

$$f_1 = \begin{cases} x + 1, & \text{if } x \geq -1 \\ 0, & \text{otherwise} \end{cases}$$

Similarly,

$$f_2 = \begin{cases} 2x - 2, & \text{if } x \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

Thus, for $x \in [-2, 2]$, we get

$$(f_1, f_2) = \begin{cases} (0, 0), & \text{if } -2 \leq x < -1 \\ (x + 1, 0), & \text{if } -1 \leq x < 1 \\ (x + 1, 2x - 2), & \text{if } 1 \leq x < 2 \end{cases} \tag{1}$$

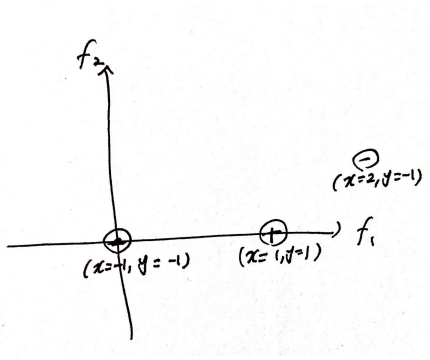
5.3) Yes

**Explanation**: From eq. (1),

we know that

- When $x = -1$, $(f_1, f_2) = (0, 0)$
- When $x = 1$, $(f_1, f_2) = (2, 0)$
- When $x = 2$, $(f_1, f_2) = (3, 2)$

Thus, in $(f_1, f_2)$ coordinates, given $(x, y)$ data pairs look like the following:




Thus the training examples are **linearly separable.**

5.4) $x \in (1, \infty)$

**Explanation**: After a gradient descent step, $w_{02}$ will decrease iff $\frac{\partial}{\partial w_{02}} Loss_h(yf) > 0$. Because it was given that $y = -1$,

$$\begin{aligned}
\frac{\partial}{\partial w_{02}} Loss_h(yf) &= \frac{\partial}{\partial w_{02}} 1 - yf \text{ if } f \geq -1 \\
&= \frac{\partial}{\partial w_{02}} f \text{ if } f \geq -1 \\
&= \frac{\partial}{\partial w_{02}} f_2 w_2 \text{ if } f \geq -1 \\
&= \frac{\partial}{\partial w_{02}} w_2(w_{12}x + w_{02}) \text{ if } f \geq -1 \text{and } x \geq 1
\end{aligned}$$

Otherwise, $\frac{\partial}{\partial w_{02}} Loss_h(yf) = 0$. Also, $\frac{\partial}{\partial w_{02}} Loss_h(yf) = \frac{\partial}{\partial w_{02}} w_2(w_{12}x + w_{02}) > 0$ iff $f > 0, x > 1$.

We can easily see that $f > 0$ automatically if $x > 1$.

Thus, gradient positive and $w_{02}$ decreases iff $x > 1$.

5.5) All weights stay the same except for $w_0$.

**Explanation**: Because of ReLU, $f_1 = ReLU(z_1) = ReLU(-1) = 0$, $f_2 = ReLU(z_2) = ReLU(-1) = 0$. So the derivative of $f_1, f_2$ with respect to $w_1 1, w_0 1, w_1 2, w_0 1$ will be 0, and all of $w_1 1, w_0 1, w_1 2, w_0 1$ will stay the same after a step of stochastic gradient descent update.

Because $f_1 = f_2 = 0$ always,

$$\frac{\partial}{\partial w_1} f = f_1 = 0, \frac{\partial}{\partial w_2} f = f_2 = 0$$

. Thus, both of $w_1, w_2$ will stay 0.

$$\frac{\partial}{\partial w_0} f = 1$$

, which is nonzero, so $w_0$ may be updated if $Loss(f)$ is a function that depends on $f$ (such as the non-zero case of hinge loss, $1 - yf$.)

5.6) A,C,B

**Explanation**: With regularization: you increase the training error but lead to better generalization/test error compared to the case with no regularization. When you increase the number of hidden units, you are potentially overfitting so you get smaller training error but high test error.

# 8 Spring 2017: Problem 1

1.1) $\theta_0 = -7$

**Explanation**: Let $\alpha_i$ be the number of mistakes that the perceptron makes on the point $x^{(i)}$ with label $y^{(i)}$. The resulting offset parameter is: $\theta_0 = \sum_{i=1}^{8} \alpha_i y^{(i)} = -7$.

1.2) No

**Explanation**: When perceptron is initialized to all zeros, the first point considered is always a mistake. Since no mistakes were made on the point (4,4) labeled +1, it could not have been the first point considered.

1.3) $\theta = [1 \quad 1], \theta_0 = -5$

**Explanation**: The margin of a separator is the minimal distance between the separator and any point in the dataset. The equation of the line that maximizes the margin on the given points is $x_1 + x_2 - 5 = 0$. The parameters corresponding to the maximum margin separator are:
$\theta^T = [1 \quad 1], \theta_0 = -5$.

1.4) Margin = $\frac{1}{\sqrt{2}}$

**Explanation**: The support vectors (points closest to the max-margin separator) are (2,2), (2, 4) and (5,1). The distance between any one of these points and the separator is $\frac{\sqrt{2}}{2}$ Alternatively, we know the margin is $\frac{1}{||\theta||} = \frac{1}{\sqrt{2}}$.

1.5) 0

**Explanation**: Since the points are perfectly separated, the hinge loss is 0. Alternatively, the sum of the hinge losses can be calculated by:
$\sum_{i=1}^{8} \max\{0, 1 - y^{(i)}(\theta \cdot x^{(i)} + \theta_0)\}$.

1.6) 1.5

**Explanation**: The sum of the hinge losses for the new parameters is:

$$\sum_{i=1}^{8} \max\{0, 1 - \frac{1}{2} y^{(i)}(\theta \cdot x^{(i)} + \theta_0)\} = 0 + 0 + 0 +$$
$$(1 - \frac{1}{2}(-1)([1 \quad 1] \cdot [2 \quad 2] - 5)) +$$
$$(1 - \frac{1}{2}(1)([1 \quad 1] \cdot [5 \quad 1] - 5)) + 0 +$$
$$(1 - \frac{1}{2}(1)([1 \quad 1] \cdot [2 \quad 4] - 5)) + 0$$

# 9 Spring 2017: Problem 2

1.1) $\theta \leftarrow (1 - \eta\lambda)\theta + \eta \begin{cases} y^{(i)}x^{(i)}, & \text{if } y^{(i)}\theta \cdot x^{(i)} \leq 1 \\ 0, & \text{otherwise} \end{cases}$

**Explanation**:

$$\begin{aligned} \theta &\leftarrow \theta - \eta\Delta_\theta[\text{Loss}_h(y^{(i)}\theta \cdot x^{(i)}) + \frac{\lambda}{2}||\theta||^2] \\ &= \theta - \eta\Delta_\theta[\text{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] - \eta\Delta_\theta[\frac{\lambda}{2}||\theta||^2] \\ &= \theta - \eta\Delta_\theta[\text{Loss}_h(y^{(i)}\theta \cdot x^{(i)})] - \eta\lambda\theta \\ &= (1 - \eta\lambda)\theta + \eta \begin{cases} y^{(i)}x^{(i)}, & \text{if } y^{(i)}\theta \cdot x^{(i)} \leq 1 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

1.2) ((B)) small $\lambda$, small $\eta$;
((A)) small $\lambda$, large $\eta$;
((C)) large $\lambda$, large $\eta$.

**Explanation**: With small $\eta$, we should see very little change to the classifier; thus, this corresponds to figure B. With large $\lambda$ and large $\eta$,we should see both an increase in the margin and large update to the classifier. This matches figure C, leaving figure A as small $\lambda$ and large $\eta$. Also note that in

figure A, the new $\theta$ can be visually obtained by adding a fraction of vector $x$ (the point) to the previous $\theta$. As a result, $\lambda$ has to be small.

# 10 Spring 2017: Problem 4

4.1) See the plot below.

**Explanation**: From the arrows indicated on the left plot of (4.1), note that $[w_{11}, w_{12}] = [1, 0]$ and $[w_{21}, w_{22}] = [0, 1]$. Thus,

$$f(z_1) = max\{0, x_1\}, \ f(z_2) = max\{0, x_2\}$$

Thus each of the points $a, b, c, d, e, f$ in the $(x_1, x_2)$ space is mapped to $(f(z_1), f(z_2))$ space like the following:

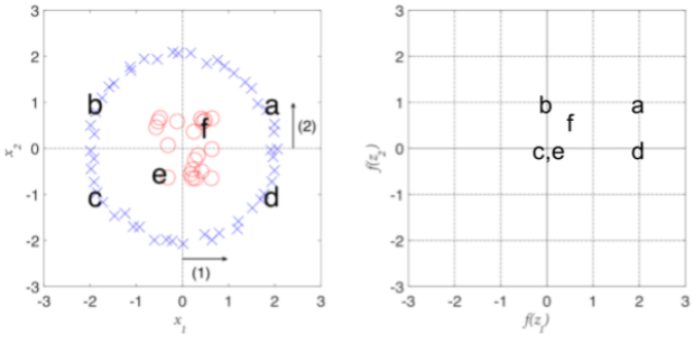



Figure 7: Problem 4.1

4.2) No

**Explanation**:Since points c and e are mapped to the origin but have opposite label, it is impossible to classify them both correctly.

4.3) Yes

**Explanation**: Instead of mapping points $a, b, c, d, e, f$ to the 2-dimensional $(f(z_1), f(z_2))$, if we map the points to a higher dimensional space, we can linearly separate data.

4.4) True, True, True

**Explanation**:

- Note that $tanh(x)$ behaves like a linear function when $x$ has a small absolute value near 0.
- Note that as $x$ gets large, $tanh(x)$ behaves like $tanh(x) = +1$ if $x > 0$, $tanh(x) = -1$ if $x < 0$.
- Consider the unit at the last layer of the given network. What does it mean that $tanh(v_1 f(z_1) + v_2 f(z_2))$ behaves like a sign function? That means

$$tanh(v_1 f(z_1) + v_2 f(z_2)) = \begin{cases} +1, & \text{if } v_1 f(z_1) + v_2 f(z_2) > 0 \\ -1, & \text{if } v_1 f(z_1) + v_2 f(z_2) < 0 \\ 0 & \text{otherwise} \end{cases}$$

and will have gradient 0 with respect to $v_1, v_2$. Thus, a network with sign units cannot be effectively trained with stochastic gradient descent.

# 11 Fall 2017: Problem 1

1. (a) The second option should be checked (Linearly separable but not through the origin).

   **Explanation**: From the plot, we see that any separators through the origin that classify both '+' labeled points correctly will also apply a '+' label to the negative points. So, the points are not separable through the origin, but are linearly separable if we allow an offset, for example, using the line that passes through $(0, 4)$ and $(4, 0)$.

   (b) There are many examples of linear separators that properly classify these points. One possible answer:
   $(-1)x_1 + (-1)(x_2) + (0)x_1^2 + (0)x_2^2 + (4) = 0$

   **Explanation**: We will consider the separator that passes through $(0, 4)$ and $(4, 0)$. The normal vector should have slope 1 and point towards the origin (to classify the '+' points positively), so we can take $\theta = [-1, -1]^T$. Then, computing the offset by plugging in $(0, 4)$ yields $\theta_0 = 4$. Thus, our separator is defined by

   $$\begin{aligned} \theta^T x + \theta_0 &= 0 \\ -x_1 - x_2 + 4 &= 0 \end{aligned}$$

   (c) Check the third option (False for all data sets).

   **Explanation**: We have simply performed a linear transformation of our data, so the separation problem has not gotten any harder. To see this mathematically, assume that we have found a linear separator $\theta', \theta'_0$ of the transformed data. We will expand the equation of the separator:

   $$\begin{aligned} \theta'^T \phi(x_1, x_2) + \theta'_0 &= 0 \\ \theta'_1 \cdot 2x_1 + \theta'_2 \cdot (x_1 + x_2) + \theta'_0 &= 0 \\ (2\theta'_1 + \theta'_2)x_1 + \theta'_2 x_2 + \theta'_0 &= 0 \end{aligned}$$

   Thus, we see that $\theta = [2\theta'_1 + \theta'_2, \theta'_2]^T, \theta_0 = \theta'_0$ separates our original dataset. So, if our original dataset is not separable, then neither is the transformed dataset.

# 12 Fall 2017: Problem 4

4. (a) i. Yes

**Explanation**: $\theta$ matches the normal to the separator, and we can easily verify the offset by plugging in any point on the separator.

ii. $(1, 1) : 0$,
$(1, 3) : 1$,
$(3, 2) : 2$,
$(3, 4) : 0$
**Explanation**: These are direct computation, using $y^{(i)}$ according to the labels given in the diagram, and the hinge loss given in the problem statement.

iii. $J = 1.75$
**Explanation**: This is a direct computation using the objective function provided in the problem.

(b) (1): N/A or -1,
(2): 1,
(3): 1
**Explanation**: We can compute the margin of each point with respect to the separator using the equation

$$\frac{y^{(i)}(\theta^T x^{(i)} + \theta_0)}{\|\theta\|}$$

The margin of each separator with respect to the entire training set is then the minimum of the margins for each point.

Alternatively, we can plot the separators on the diagram and take the minimum (signed) distance of any point to the separator. We see that for $\theta^{(1)}$ this corresponds to a distance of -1 to the point at $(3, 2)$, while the other two separators have a distance of 1 to all four points.

(c) Select the third separator (3)
**Explanation**: Computing the the loss for each separator,

$$J(\theta^{(1)}, \theta_0^{(1)}) = \frac{1}{4}(0 + 1 + 2 + 0) = 0.75$$

$$J(\theta^{(2)}, \theta_0^{(2)}) = \frac{1}{4}(0.9 + 0.9 + 0.9 + 0.9) = 0.9$$

$$J(\theta^{(3)}, \theta_0^{(3)}) = \frac{1}{4}(0 + 0 + 0 + 0) = 0$$

So, we prefer the third separator.

(d) Yes
**Explanation**: Yes, the large $\lambda$ causes us to favor separators with small magnitude:

$$J(\theta^{(1)}, \theta_0^{(1)}) = 0.75 + 100 \cdot 1 = 100.75$$

$$J(\theta^{(2)}, \theta_0^{(2)}) = 0.9 + 100 \cdot 0.1 = 10.9$$
$$J(\theta^{(3)}, \theta_0^{(3)}) = 0 + 100 \cdot 100 = 10000$$

So, we would choose the second separator.

# 13 Fall 2017: Problem 5

5. a) 3

**Explanation**: $\max(1 \cdot (-2) + 3, 0) \cdot 3 + 0 = 3$.

b) $\frac{\partial L(f_2, y)}{\partial z_2} = w_2 \max(0, w_1 x + b_1) + b_2 - y$

**Explanation**:

$$\begin{aligned}\frac{\partial L(f_2, y)}{\partial z_2} &= (f_2 - y)\frac{\partial}{\partial z_2} f_2 \\ &= (f_2 - y) \\ &= w_2 \max(0, w_1 x + b_1) + b_2 - y\end{aligned}$$

c) $\frac{\partial L(f_2, y)}{\partial z_1} = \begin{cases} 0, & \text{if } w_1 x + b_1 < 0 \\ w_2(w_2 \max(0, w_1 x + b_1) + b_2 - y), & \text{o.w..} \end{cases}$

**Explanation**:

$$\begin{aligned}\frac{\partial L(f_2, y)}{\partial z_1} &= \frac{\partial f_1}{\partial z_1}\frac{\partial z_2}{\partial f_1}\frac{\partial}{\partial z_2} L(f_2, y) \\ &= w_2(f_2 - y)\begin{cases} 0, & \text{if } w_1 x + b_1 < 0 \\ 1, & \text{o.w..} \end{cases} \\ &= \begin{cases} 0, & \text{if } w_1 x + b_1 < 0 \\ w_2(w_2 \max(0, w_1 x + b_1) + b_2 - y), & \text{o.w..} \end{cases}\end{aligned}$$

d) $w_1 = -5$

**Explanation**:

$$\begin{aligned}\frac{\partial L(f_2, y)}{\partial w_1} &= \frac{\partial z_1}{\partial w_1}\frac{\partial L(f_2, y)}{\partial z_1} \\ &= w_2(f_2 - y)x\begin{cases} 0, & \text{if } w_1 x + b_1 < 0 \\ 1, & \text{o.w..} \end{cases}\end{aligned}$$

Given the values provided,
$w_1 \leftarrow w_1 - 0.5\frac{\partial}{\partial w_1} L(f_2, y) = -2 - 0.5(1 \cdot 1 \cdot 3 \cdot (3 - 1)) = -5$.

e) If $x = 0$, or if $w_2 = 0$, or if $w_1 x + b_1 < 0$, or if $f_2 = y$, which corresponds to the condition $\max(w_1 x + b_1, 0) w_2 + b_2 = y$

**Explanation**: $w_1$ is unchanged during backpropagation if the gradient evaluates to 0. We use our expression from (d) for the gradient (or partial derivative, since we are in 1-d) of the loss with respect to $w_1$:

$$\frac{\partial}{\partial w_1} L(f_2, y) = x \cdot \left\{ \begin{array}{ll} 0 & \text{if } w_1 x + b_1 < 0 \\ 1 & \text{otherwise} \end{array} \right\} w_2 (f_2 - y)$$

This is 0 if $x = 0$, if $w_2 = 0$, if $w_1 x + b_1 < 0$, or if $f_2 = y$, which corresponds to the condition

$$\max(w_1 x + b_1, 0) w_2 + b_2 = y$$

# 14 Fall 2017: Problem 6

6. a) $c_1 = 1, c_2 = 1$ OR $c_1 = \frac{1}{2}, c_2 = \frac{1}{2}$. (The rest of the answers below assume $c_1 = c_2 = 1$.)

**Explanation**: For squared error, we don't penalize differently for over or underestimating, so $c_1 = c_2 = c$. Minimizing this loss is equivalent for whatever positive constant $c$ we choose, though you will most often see $c$ set to 1 or 1/2 (for cleanliness when differentiating). In lecture notes, we take the average squared loss so in that case $c_1 = c_2 = 1/2$.

c) Assuming $c_1 = c_2 = 1$.

$$\theta = \theta - 2\eta x(g - y) \begin{cases} c_1, & \text{if } g > y \\ c_2 & \text{o.w.} \end{cases}$$

$$\theta_0 = \theta_0 - 2\eta(g - y) \begin{cases} c_1, & \text{if } g > y \\ c_2 & \text{o.w.} \end{cases}$$

# 15 Fall 2017: Problem 7

7. a) i. One-hot encoding

**Explanation**: We have categorical data that has no numerical interpretation, so it makes the most sense to use a one-hot encoding.

ii. Divide by 50

**Explanation**: We would like our features to have approximately the same magnitude, so we divide by 50.

iii. Divide by 1 billion

**Explanation**: We would like our features to have approximately the same magnitude, so we divide by 1,000,000.

iv. Omit
**Explanation**:The company name is unlikely to be indicative of how the stock will perform in the future, so we omit it.

b) i. $\alpha$) 1
$\beta$) sigmoid
$\gamma$) NLL
Also okay: 1 unit, linear, hinge or 2 units, softmax, NLL
**Explanation**: We can try to predict the probability that the company will have an IPO (and use a threshold probability of 0.5 to decide which classification to make), which can be done with a single sigmoid unit and NLL loss.
Other valid solutions are to use a single linearly activated unit with hinge loss (which is effectively SVM to perform the classification into the two classes) or two units with softmax activation and NLL loss (and choosing the larger of the two probabilities for our classification).

ii. $\alpha$) 1
$\beta$) linear
$\gamma$) squared-error
**Explanation**: We would like to predict a single numerical value that spans the real numbers, so we will use a linear activation and squared error.

iii. $\alpha$) 100
$\beta$) sigmoid
$\gamma$) NLL (individually)
**Explanation**: Here, we are asked to perform 100 independent 2-class classification problems. So, we can have 100 separate sigmoid activated units each with their own NLL loss (which we sum to get the total loss). Each unit is responsible for performing the prediction for one specific client.
We can also adapt the other solutions from (a), just using 100 copies of whatever approach we choose.

# 16 Fall 2017: Problem 9

9. (a) 0.01: Weight trajectory: (d) , Objective: (g)
**Explanation**: For the smallest step size, we see the slowest convergence to the optimum, which corresponds to slowest moving trajectory (d) and the slowly decreasing objective plot (g).

(b) 0.05: Weight trajectory: (a) , Objective: (h)
**Explanation**: As the step size increases, speed of convergence increases, as seen in trajectory (a) and objective plot (h).

(c) 0.50: Weight trajectory: (c) , Objective: (f)

**Explanation**: Convergence is even faster in trajectory (c), though we start seeing some oscillation. The objective plot is the most sharply decreasing plot (f).

(d) 1.00: Weight trajectory: (b) , Objective: (e)

**Explanation**: The step size has increased too much, and our weights and objective diverge. This is seen in (b) and (e).