Homework 1
CSCI4100

Han Hai
Rin:661534083
haih2@rpi.edu

September 7 2018

1. **[100 POINTS] Exercise 1.3:** The weight update rule in (1.3) has the nice interpretation that it moves in the direction of classifying x(t) correctly.

   (a) Show that $y(t)w^T(t)x(t) < 0$
   Solution: Since x(t) is misclassified, $w^(T)x(t)$ has a different sign from $y(t)$,thus the product of two will be negative.

   (b) Show that $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$
   Solution: According to the update rule $w(t+1) = w(t) + y(t)x(t)$
   so LHS becomes $y(t)(w^T(t) + x(t)y(t)) = (y(t)x(t))^2 + y(t)w^T(t)x(t)$
   Since $(y(t)x(t))^2 > 0$ assume both are not zeros, the LHS is larger than RHS

   (c) As far as classifying x(t) is concerned, argue that the move from w(t) to w(t+1) is a move "in the right direction".
   Solution: Since our goal is to classifies x(t) correctly, which means $y(t)w^T(t)x(t)$ should be positive, we want to increase $y(t)w^T(t)x(t)$. As we proved in (b), applying update rule will always increase the value. If it is originally a negative value, increasing it will move it towards positive. Thus, applying update rule is moving towards "the right direction"

2. **[100 points] Exercise 1.5:** Which of the following problems are more suited for the learning approach and which are more suited for the design approach?

   (a) Determining the age at which a particular medical test should be performed
   Learning approach: because we do not know how to derive the target function

   (b) Classifying numbers into primes and non-primes
   Design approach: because a target function can be derived mathematically without the need of data

   (c) Detecting potential fraud in credit card charges
   Learning approach: hard to find target function analytically, we can use data to learn

   (d) Determining the time it would take a falling object to hit the ground
   Design approach: physics equation can be derived analytically without the use of data

(e) Determining the optimal cycle for traffic lights in a busy intersection
Learning approach: hard to find target function, we can use data to learn

3. **[100 points] Exercise 1.6:** For each of the following tasks, identify which type of learning is involved (supervised, reinforcement, or unsupervised)and the training data to be used. If a task can fit more than one type, explain how and describe the training data for each type.

(a) Recommending a book to user in an online bookstore
The learning should be supervised and the input data should be user's attributes including user's previous purchase and the output data should be what books the use purchase the most recent time

(b) Playing tic-tac-toe
Reinforcement learning, the data should be previous tic-tac-toe moves and results, each move as input data should correspond to whether this move win the game or not as reward/grade

(c) Categorizing movies into different types
Supervised learning, because we can have human to classified the movie as output. Unsupervised learning might also work, we can just have movie as inputs and without classifying them by human, the machine can also find potential pattern

(d) learning to play music
Reinforcement learning, a reward should be playing the correct note

(e) Credit limit: Deciding the maximum allowed debt for each bank customer
Supervised learning: input data should be customer's attributes(age, past history etc.) and output data should be whether or not the customer is approved

4. **[100 points] Exercise 1.7** For each of the following learning scenarios in the above problem, evaluate the performance of g on the three points in X outside D. To measure the performance, compute how many of the 8 possible target functions agree with g on all three points, on two of them, on one of them and on none of them.

(a) H has only two hypotheses, one that always return bullet point and on that always return empty point. The learning algorithm picks the hypothesis that matches the data set the most.

Solution: I will assign three points if all three predictions matches, 2 with 2 matches, one with one match, and 0 for none.
In this case, g will return bullet for all three predictions,f_8 agrees with all of them, f_7, f_6, f_4 agrees with two of them, f_2, f_3, f_5 agrees with one of the prediction, while f_1 agrees with none.
$3 + 3 * 2 + 3 * 1 = 10$

(b) The same H, but the learning algorithm now picks the hypothesis that matches the data set the least.

Solution: It's the reverse of part a, the hypothesis will predict all three empty bullet.f_8 agrees with none, f_7, f_6, f_4 agrees with one of the prediction, f_2, f_3, f_5 agrees with two of the prediction, while f_1 agrees with all predictions
$3 + 3 * 2 + 3 * 1 = 10$

(c) $H = \{XOR\}$ where XOR returns black bullet if the number of 1's is odd, and empty bullet if the number of 1's is even

It will predict white, white, black. In this casef_2 match the prediction, f_1, f_4, f_6 match two of the prediction, f_3, f_5, f_8 matches one of the prediction, while f_7 matches none
$3 + 3 * 2 + 3 * 1 = 10$

(d) H contains all possible hypotheses, and the learning algorithm picks the hypothesis that agrees with all training examples, but otherwise disagrees the most with the XOR.
It's the reverse of part c, and the score is still the same 10
Conclusion: No matter what hypothesis to choose, performance is essentially the same

5. **[200 points] Problem 1.1** We have 2 opaque bags, each containing 2 balls. One bag has 2 black balls and the other has a black and white ball. You pick a bag at random and then pick one of the balls in that bag at random. When you look at the ball it is black. You now pick the second ball from the same bag. What is the probability that this ball is also black? [Hint: Use Baye's Theorem: $P[A \cap B] = P[A|B]P(B) = P[B|A]P[A]$]

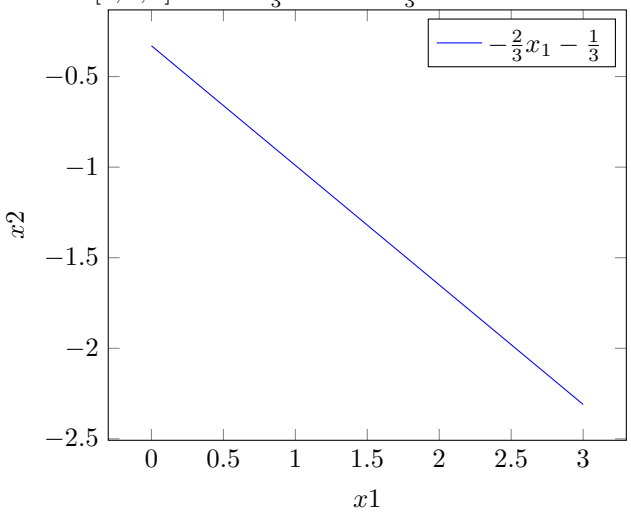
Solution: We are looking for the P[2nd ball is black given 1st ball is black], let it be P[2b|1b]
Derived from Bayes, it's equal to$P[1b \cap 2b]/P[1b] = \frac{1}{2} \div \frac{3}{4} = \frac{2}{3}$

6. **[200 points] Problem 1.2** Consider the perception in two dimensions: $h(x) = sign(w^T x)$ where $w = [w_0, w_1, w_2]^T$ and $x = [1, x_1, x_2]^T$ Technically, x has three coordinates, but we call this perceptron two-dimensional because the first coordinate is fixed at 1.

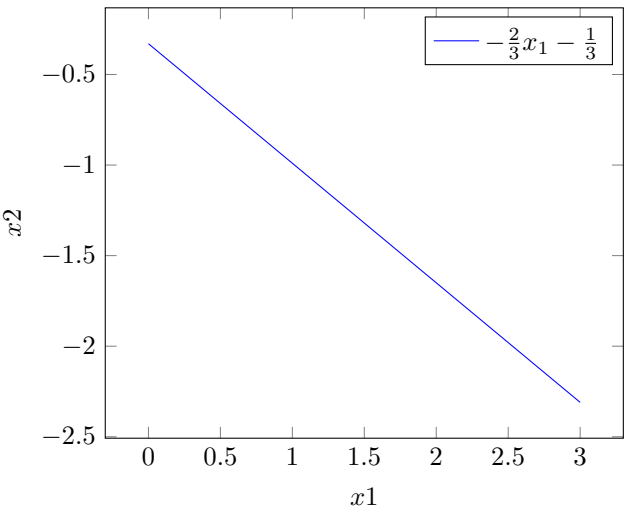
(a) Show that the regions on the plane where h(x)=+1 and h(x)=-1 are separated by a line. If we express this line by the equation $x_2 = ax_1 + b$, what are the slope a and intercept b in terms of w_0, w_1, w_2?
Since all the +1 points has $w^T x > 0$ and -1 has $w^T x < 0$, there should be a line between two plane and it satisfy $w^T x = 0$ which implies $w_0 + w_1 x_1 + w_2 x_2 = 0$
$x_2 = -\frac{w_o}{w_2} - \frac{w_1}{w_2} x_1$

Therefore a=$-\frac{w_1}{w_2}$ and b= $-\frac{w_0}{w_2}$

(b) for $w = [1, 2, 3]^T$ $a = -\frac{2}{3}$ and b=$-\frac{1}{3}$

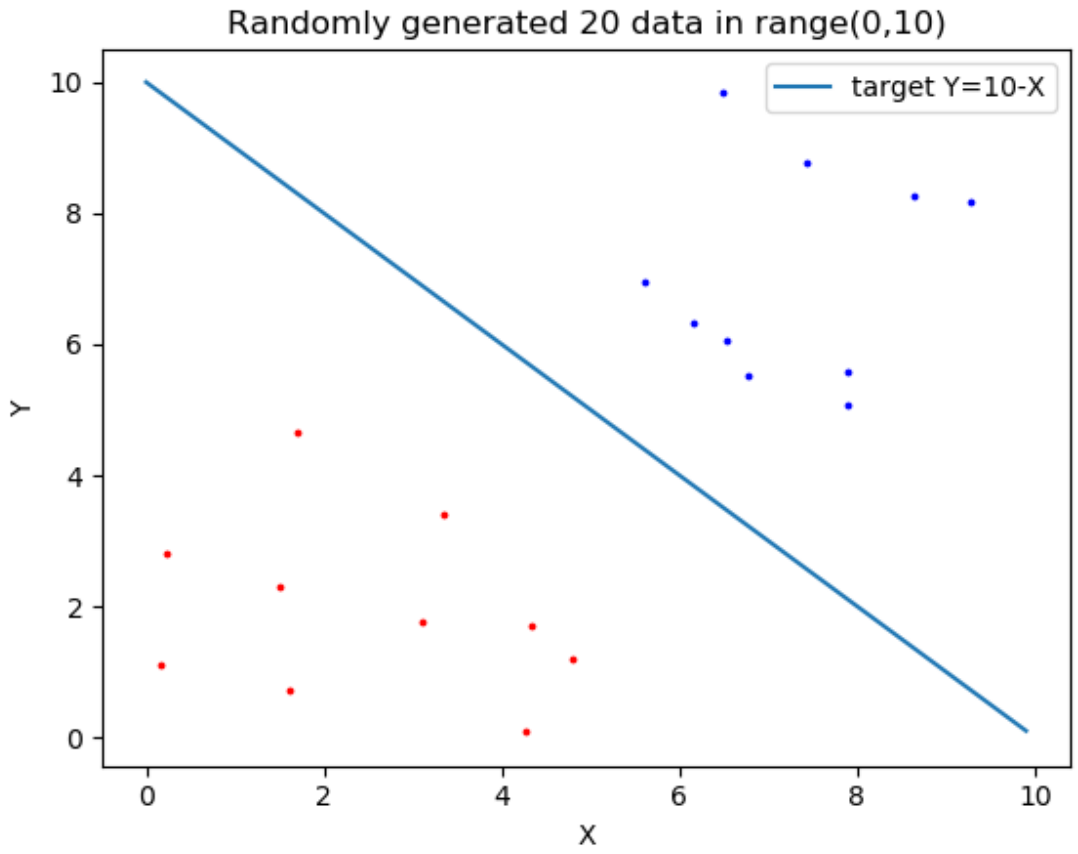


for $w = -[1, 2, 3]^T$ a =$-\frac{2}{3}$ and b= $-\frac{1}{3}$, so two lines are identical

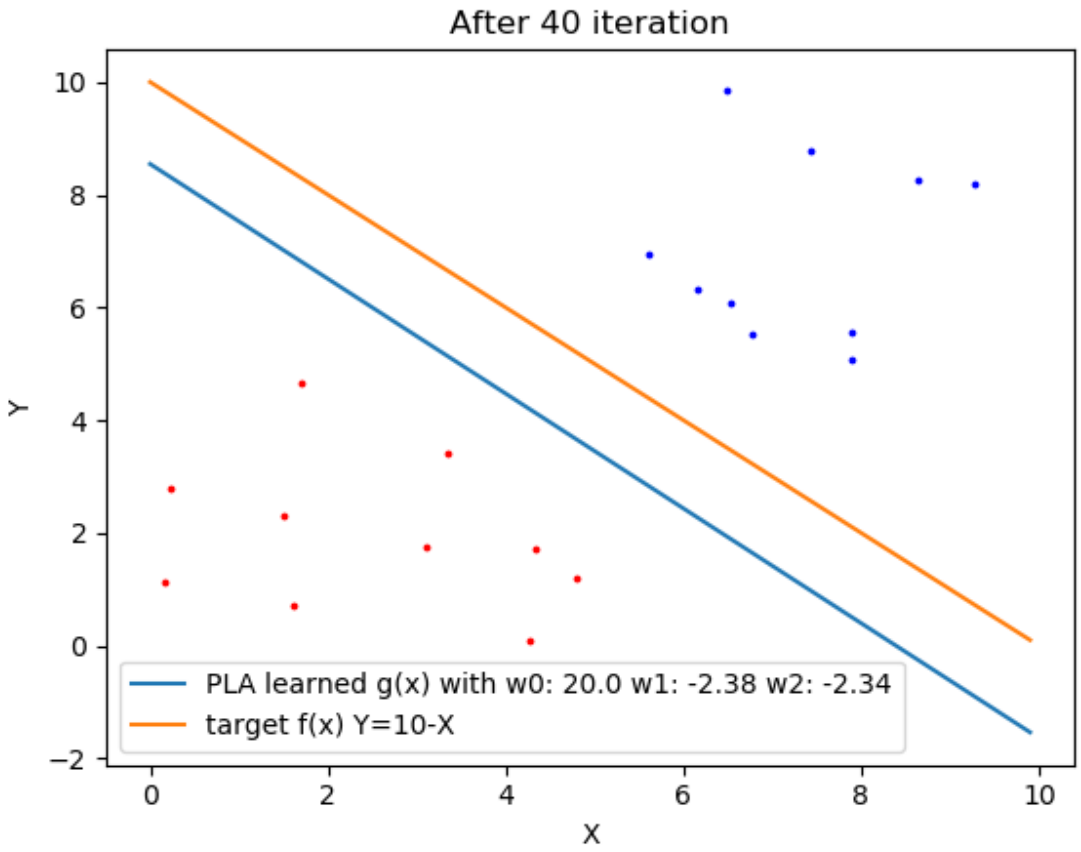

$-\frac{2}{3}x_1 - \frac{1}{3}$
−0.5
−1
−1.5
−2
−2.5
$x2$
0
0.5
1
1.5
2
2.5
3
$x1$

7. [**200 points**] **Problem 1.4** In Exercise 1.4, we use an artificial data set to study the perceptron learning algorithm. This problem leads you to explore the algorithm further with data sets of different sizes and dimensions.

   (a) Generate a linearly separable data set of size 20 as indicated in Exercise 1.4. Plot the examples (x_n, y_n) as well as the target function f on a plane. Be sure to mark the examples from different classes differently, and add labels to the axes of the plot.

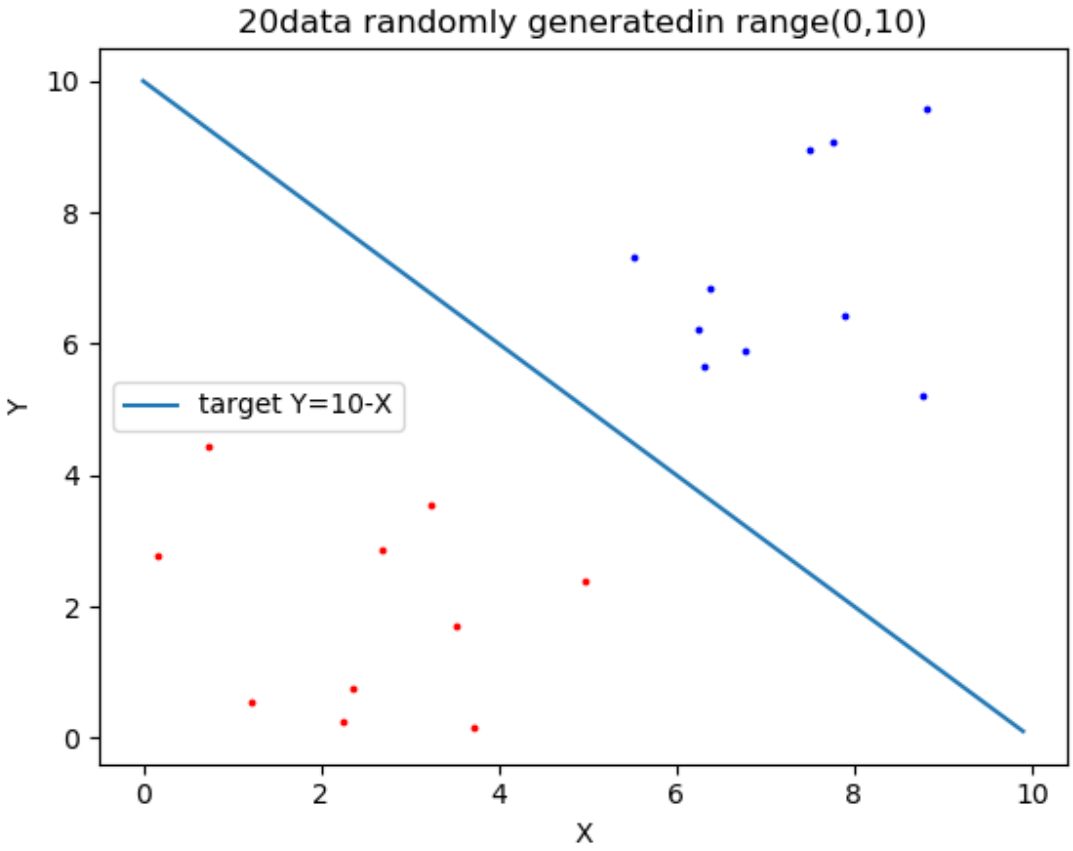


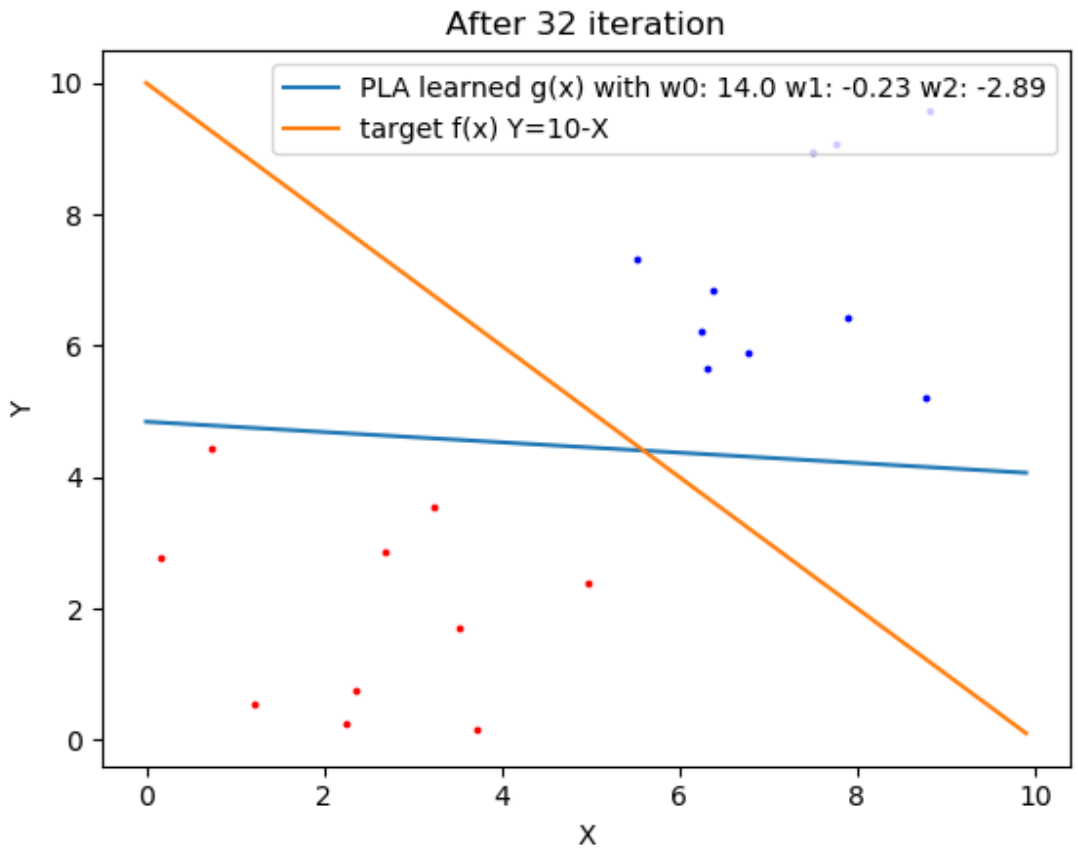
(b) Run the perceptron learning algorithm on the data set above. Report the number of updates that the algorithm takes before converging. Plot the examples, the target function f, and the final hypothesis g in the same figure. Comment on whether f is close to g.



Solution:I would say final hypothesis is close to target function, a little off, but good enough to separate data correctly

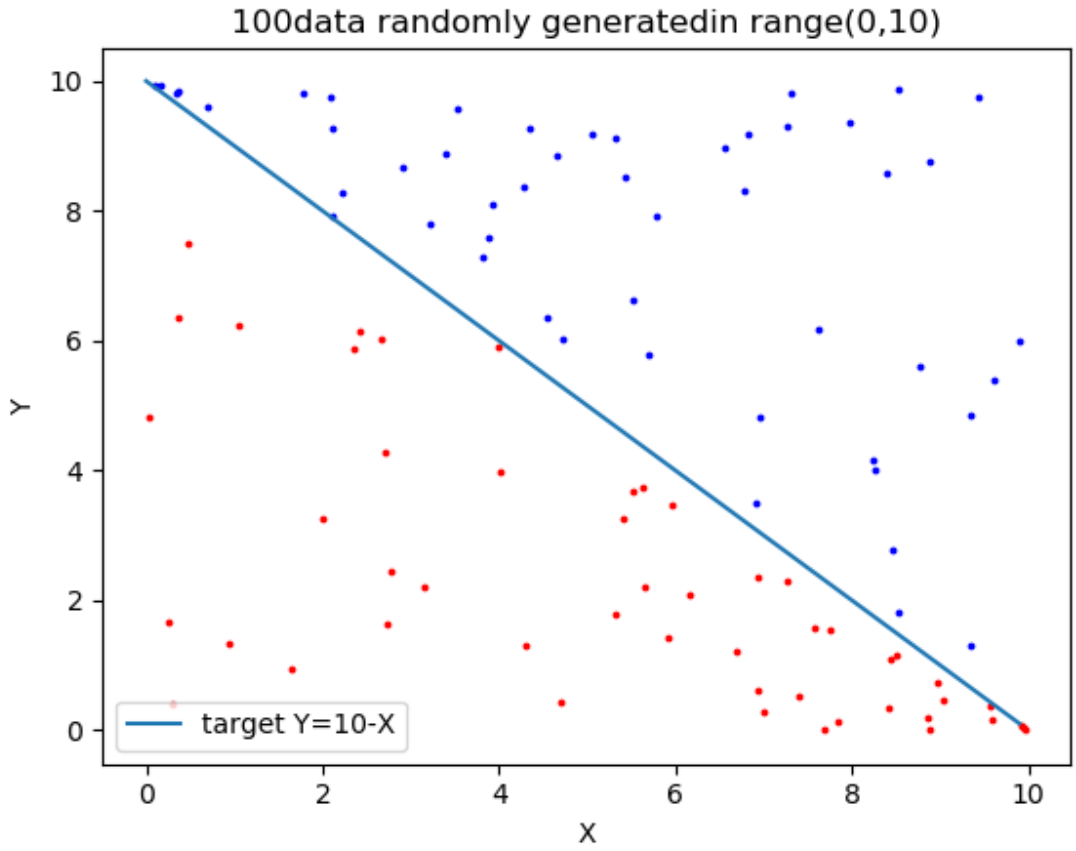
(c) Repeat everything in (b) with another randomly generated data set of size 20. Compare your results with(b).

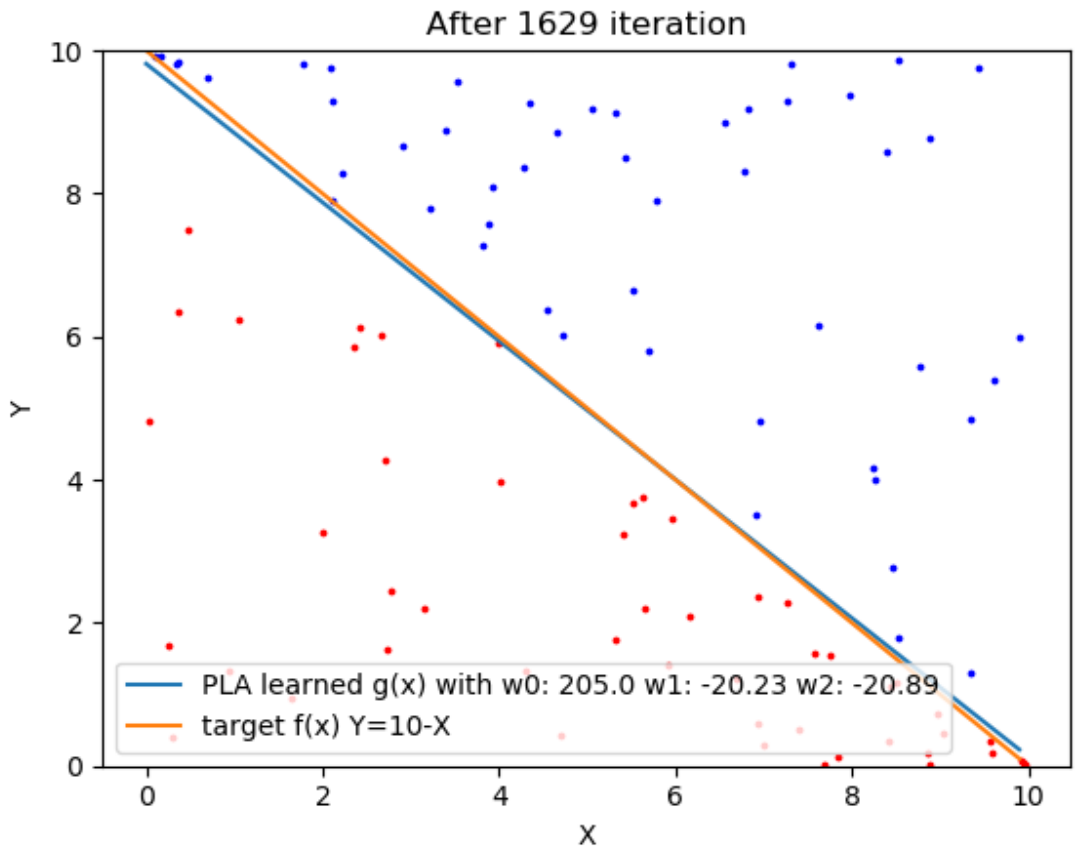




Compare to b, the number of updates to find a hypothesis is similar, though it's off by more from the target function.

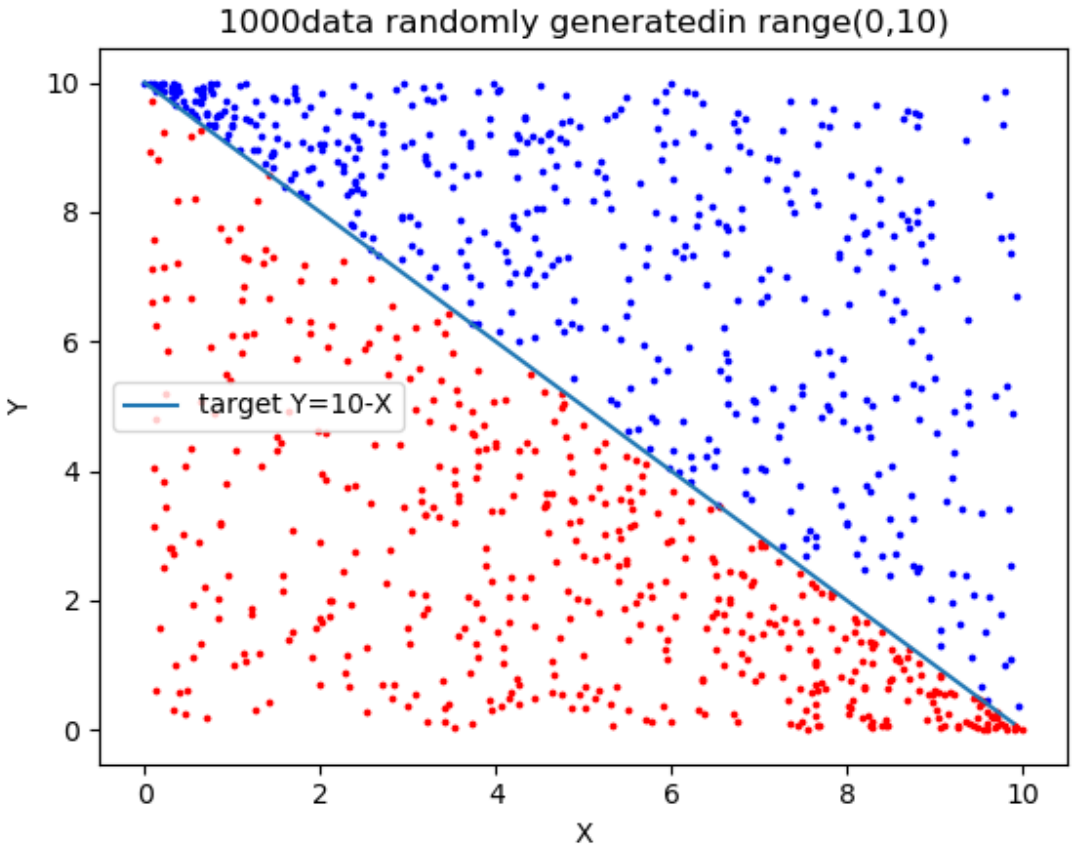
(d) Repeat everything in (b) with another randomly generated data set of size 100. Compare your results with (b)

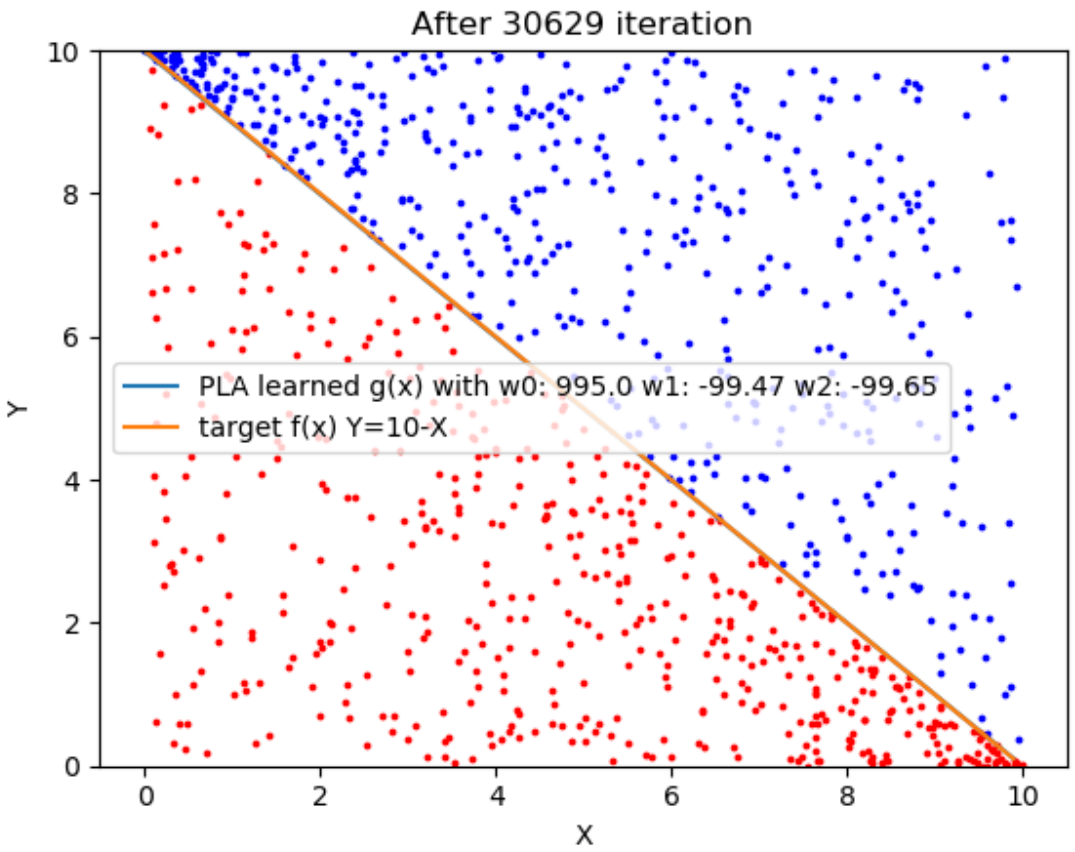




data set 100 takes much more updates to find a hypothesis function, and it's much closer to target function due to higher requirement of accuracy

(e) Repeat everything in (b) with another randomly generated data set of size 1000. Compare your results with (b)





It takes even more updates/iteration to reach a hypothesis function, and g(x) is almost identical to the target function f(x)

Conclusion: The larger the data set is, the more updates it takes to find g(x). Also, larger data set requires a more accurate g(x) that is closer to f(x).