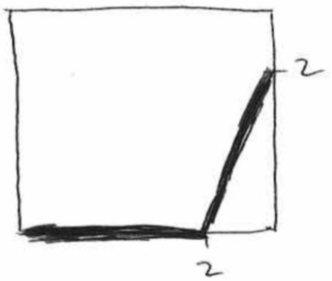
PROBLEM 12

(5.1) $x \leq -1$

(5.2)



(5.3) Y

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

(5.5) 0, except W_0

(5.6) A
C
B

Additional explanation:

12) a) The input to the first hidden unit is $z1 = xw_{11} + w_{01}$, but we're given that w_{11} is 1 and w_{01} is 1. Therefore, the ReLU activation turns inputs with $x \leq -1$ into 0.

b) It will look like a ReLU from 2 to 2.

c) Yes, from the graph, the training examples are linearly separable in the transformed coordinates.

d) We know that w_{02} will decrease when the derivative of loss with respect to w_{02} is positive, since the gradient descent update subtracts and the step size is positive. This occurs for $x \in (1, \infty)$.

e) All the parameters would stay as initialized, so everything (except the offsets) would stay 0.

f) A, C, B. Model A should allow a relatively low training error but a larger validation error (compare to B), as in the case of a simpler model without regularization. Model C should allow overfitting, either with lack of regularization or a high model complexity, to generalize poorly. Model B should use regularization or be simpler to have a higher training error but a lower validation error. This question is relative and requires thinking carefully about how A, B, and C compare to each other.