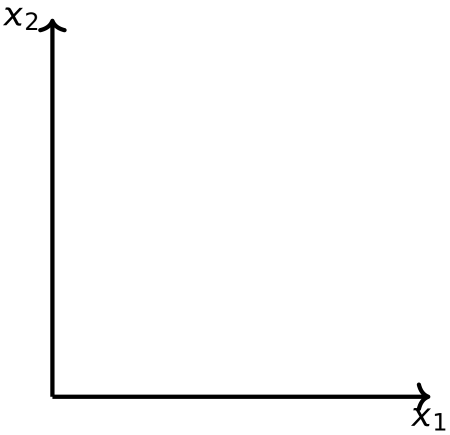
Deep Learning

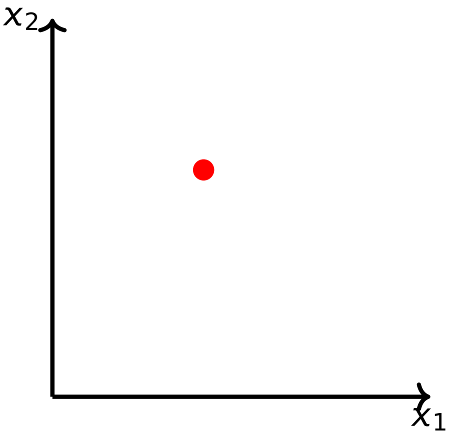
Lecture 6: Gradient Descent

Dr. Mehrdad Maleki

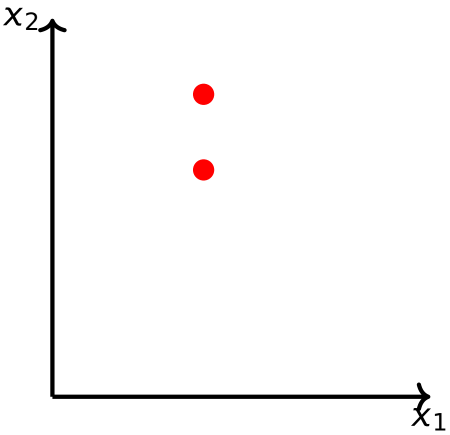
Motivation

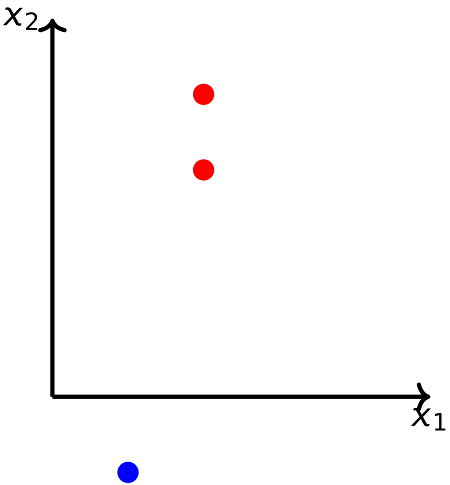
You have a NN that meant to compute desire function $g(\mathbf{X})$. But your NN actually compute $f(\mathbf{X}, \mathbf{W})$ and you need to find the value of $\mathbf{W}$ such that $||g(\mathbf{X}) - f(\mathbf{X}, \mathbf{W})||$ is minimum. But the value of $g(\mathbf{X})$ is not known for all $\mathbf{X}$. So we sample the function g. On input $\mathbf{X_i}$ if $\mathbf{y_i} = g(\mathbf{X_i})$ then for $i = 1, \ldots, N$, $\{(\mathbf{X_i}, \mathbf{y_i}) : 1 \leq i \leq N\}$ is the set of samples.

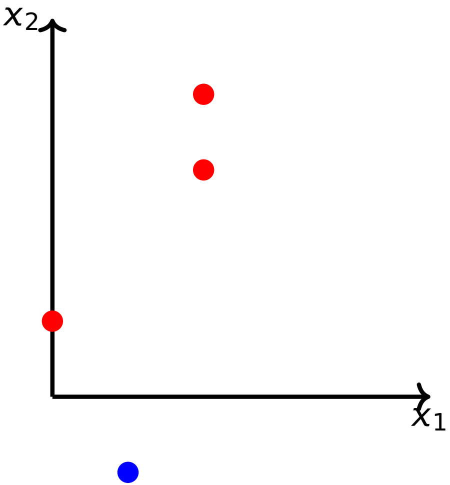

x_2
x_1



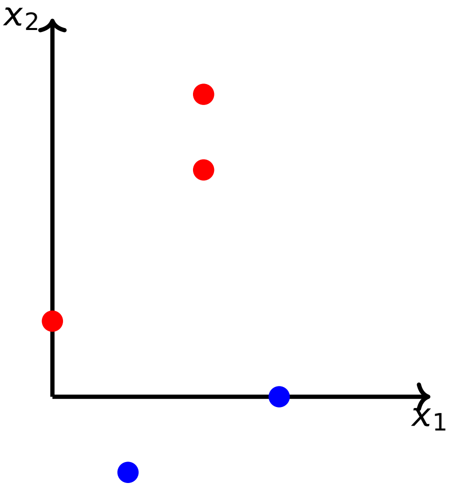
x_2
x_1

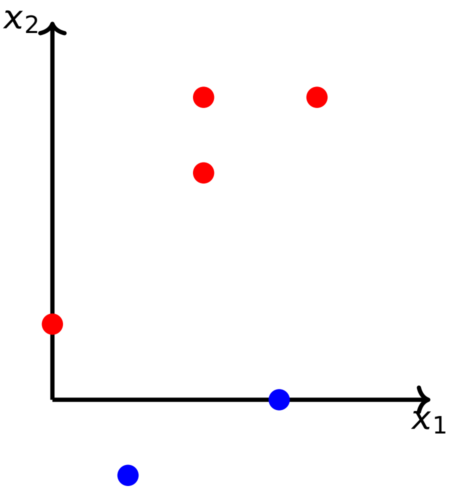

x_2
x_1

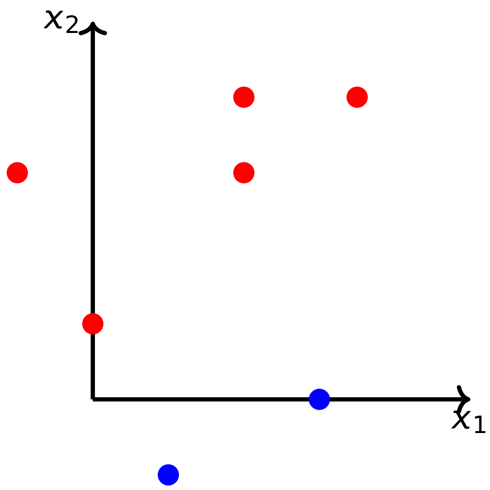

x_2
x_1

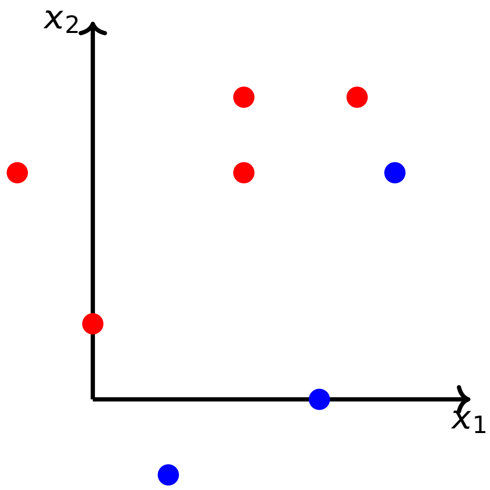


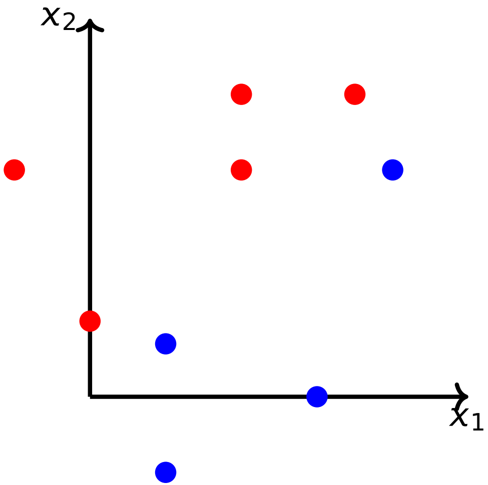
x_2
x_1

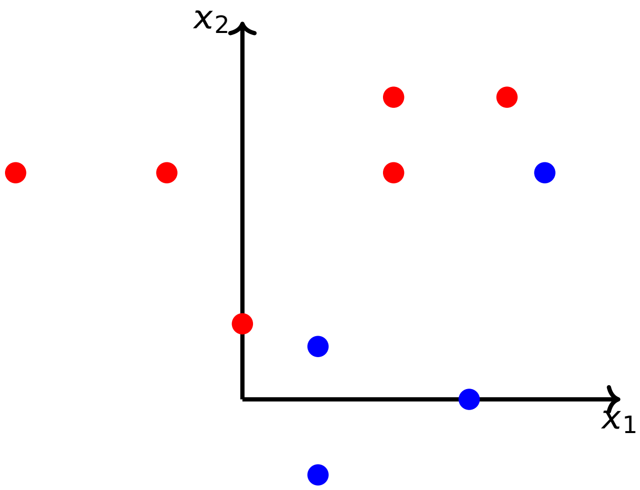

x_2
x_1

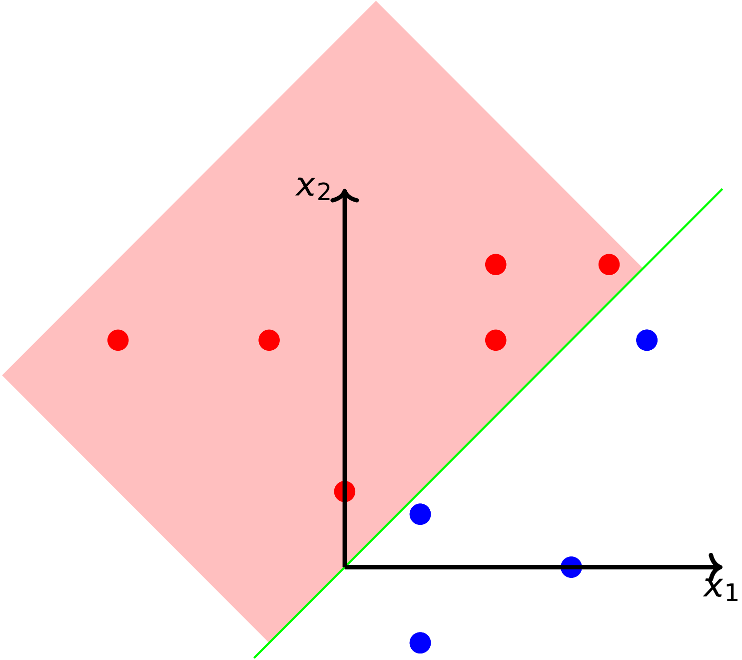

x_2
x_1

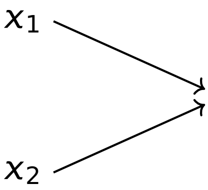

x_2
x_1

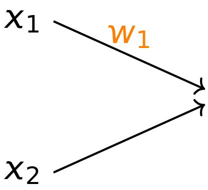

x_2
x_1

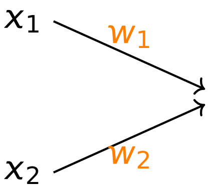

x_2
x_1

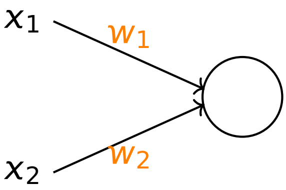

x_2
x_1


x_2
x_1

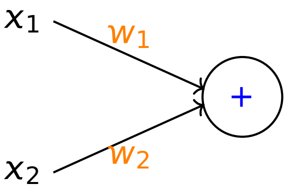

x_1
x_2

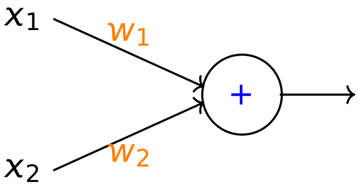

x_1
w_1
x_2

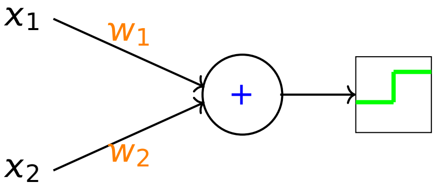

x_1
w_1
x_2
w_2

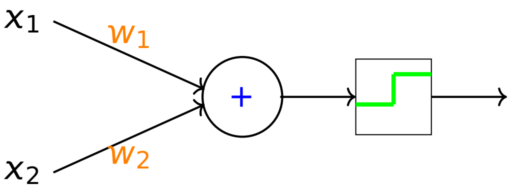


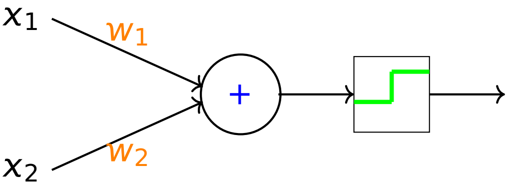
x_1
w_1
x_2
w_2


x_1
w_1
$+$
x_2
w_2

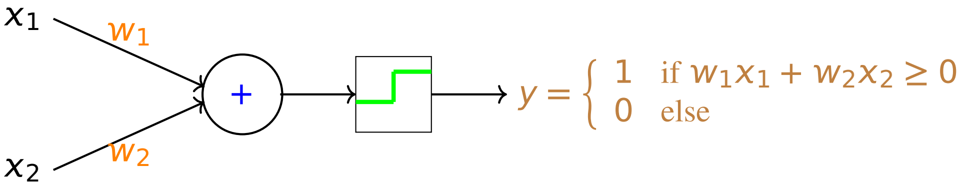

x_1
w_1
+
x_2
w_2


x_1
w_1
+
x_2
w_2


x_1
w_1
x_2
w_2
+



$$y = \begin{cases} 1 & \text{if } w_1 x_1 + w_2 x_2 \geq 0 \\ \end{cases}$$


x_1
w_1
x_2
w_2
$+$
$y = \begin{cases} 1 & \text{if } w_1x_1 + w_2x_2 \geq 0 \\ 0 & \text{else} \end{cases}$

We know that,

	Red		Blue
1	$g(2,3)=1$	7	$g(1,-1)=0$
2	$g(2,4)=1$	8	$g(3,0)=0$
3	$g(0,1)=1$	9	$g(4,3)=0$
4	$g(3.5,4)=1$	10	$g(1,0.7)=0$
5	$g(-1,3)=1$		
6	$g(-3,3)=1$		

we need to find the parameters w_1, w_2 such that if

$$f(x_1, x_2\,; w_1, w_2) = \begin{cases} 1 & \text{if } w_1 x_1 + w_2 x_2 \geq 0 \\ 0 & \text{else} \end{cases}$$

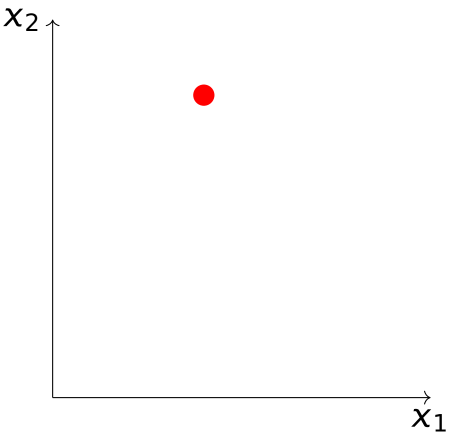
then,

$$\sum_{i=1}^{10} \| f(x_1^{(i)}, x_2^{(i)})\,; w_1, w_2) - g(x_1^{(i)}, x_2^{(i)}) \|$$

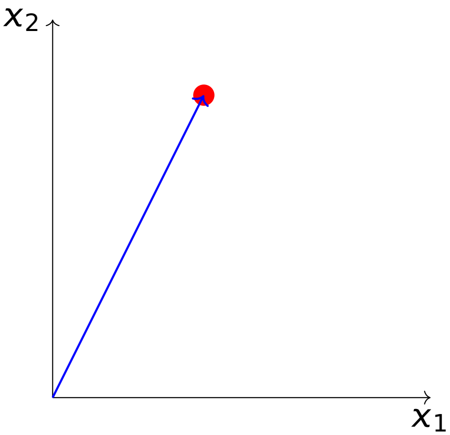
is minimum.

Suppose we only have one point with red label, i.e., 1. Then the optimal weight vector for this point is as follow,

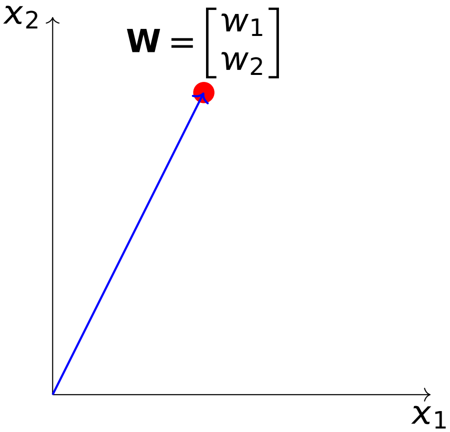
Suppose we only have one point with red label, i.e., 1. Then the optimal weight vector for this point is as follow,



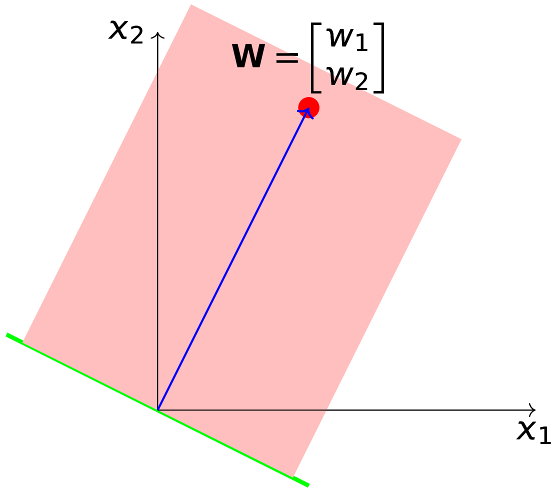
Suppose we only have one point with red label, i.e., 1. Then the optimal weight vector for this point is as follow,



Suppose we only have one point with red label, i.e., 1. Then the optimal weight vector for this point is as follow,

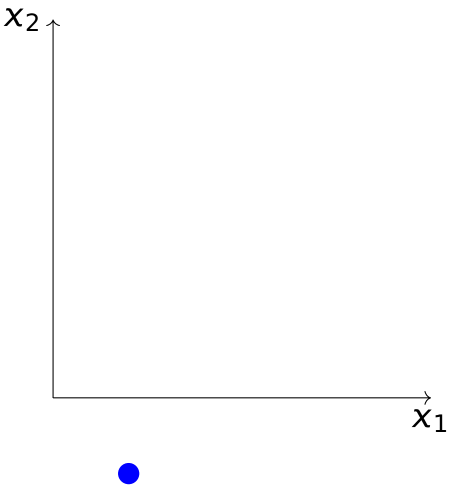


Suppose we only have one point with red label, i.e., 1. Then the optimal weight vector for this point is as follow,

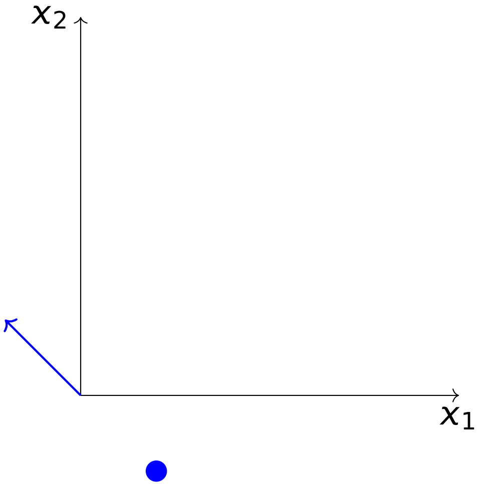


Suppose we only have one point with blue label, i.e., -1. Then the optimal weight vector for this point is as follow,

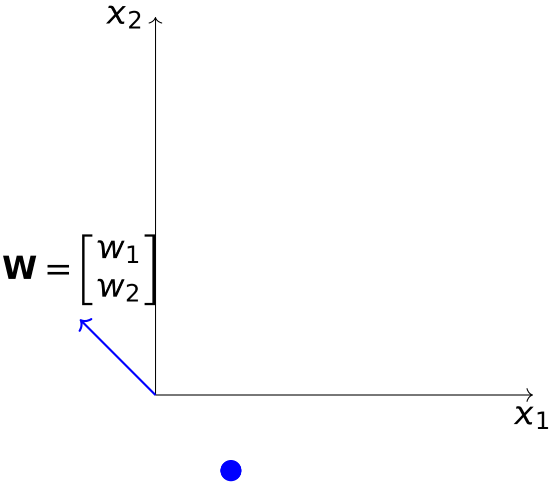
Suppose we only have one point with blue label, i.e., -1. Then the optimal weight vector for this point is as follow,



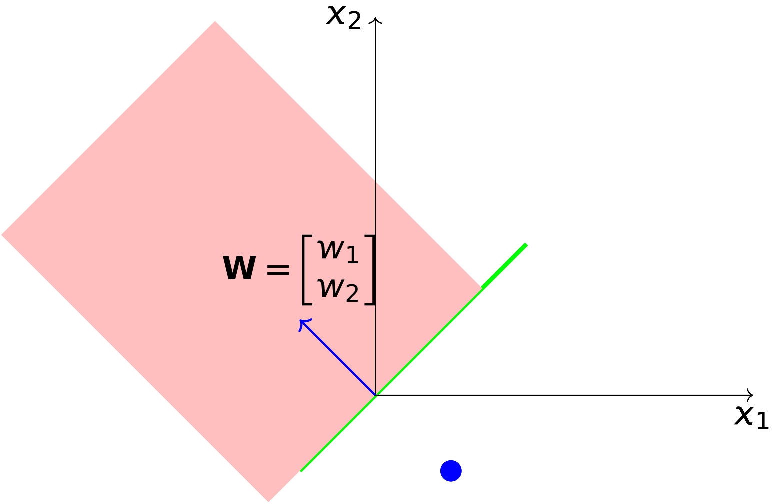
Suppose we only have one point with blue label, i.e., -1. Then the optimal weight vector for this point is as follow,

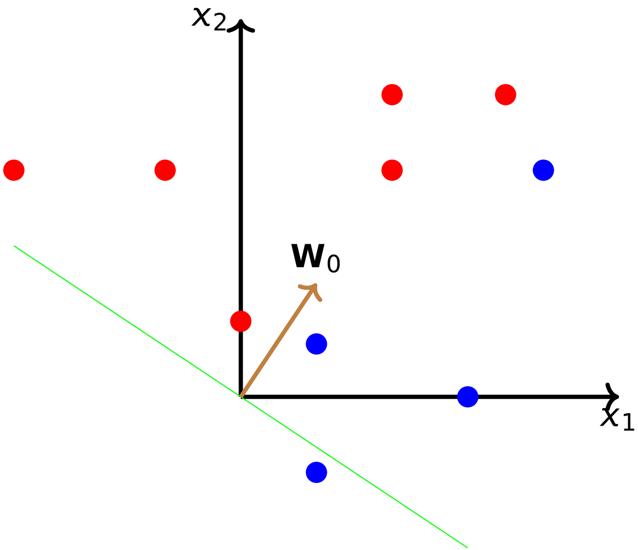


Suppose we only have one point with blue label, i.e., -1. Then the optimal weight vector for this point is as follow,

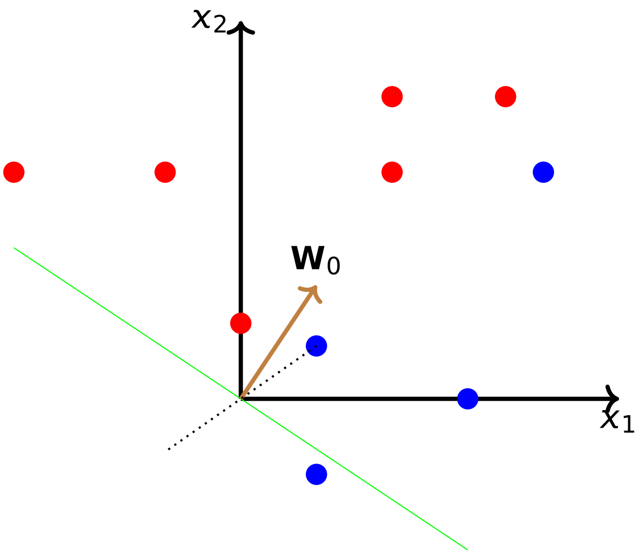


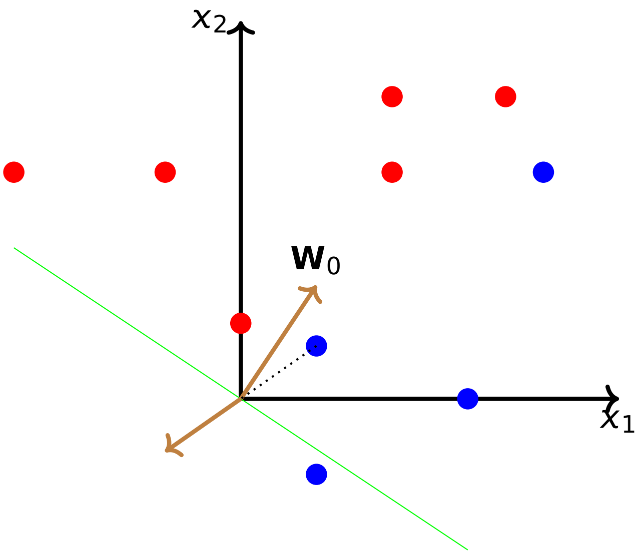
Suppose we only have one point with blue label, i.e., -1. Then the optimal weight vector for this point is as follow,



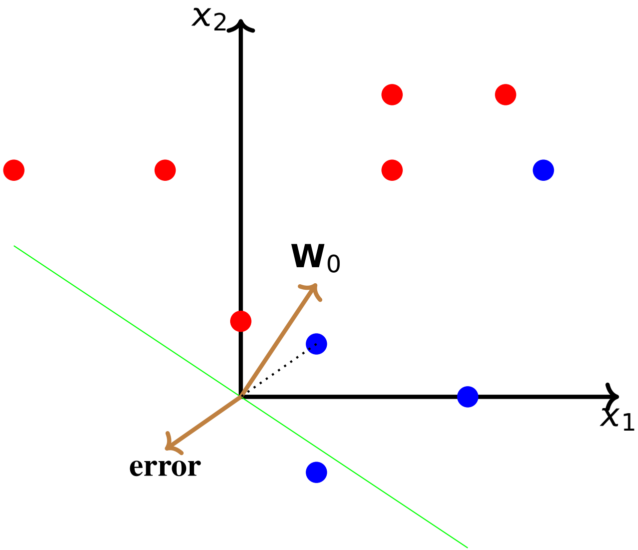


x_2
$\mathbf{w}_0$
x_1

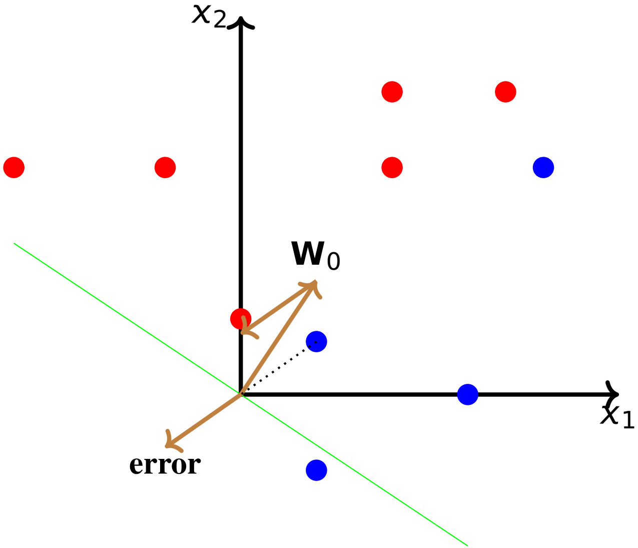

x_2
$\mathbf{w}_0$
x_1



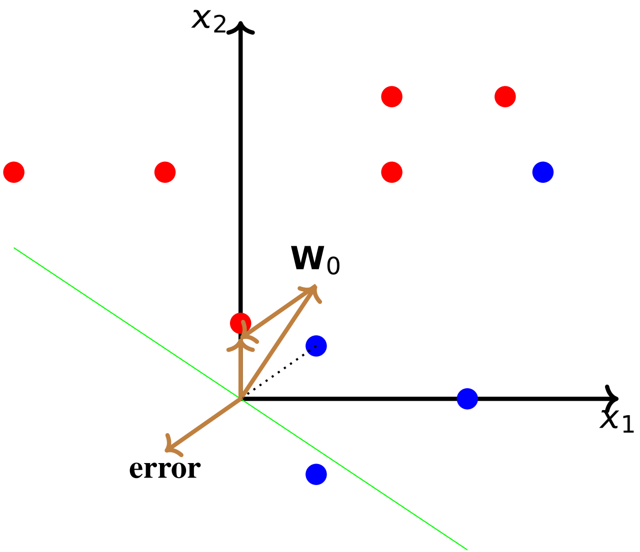
x_2
$\mathbf{w}_0$
x_1

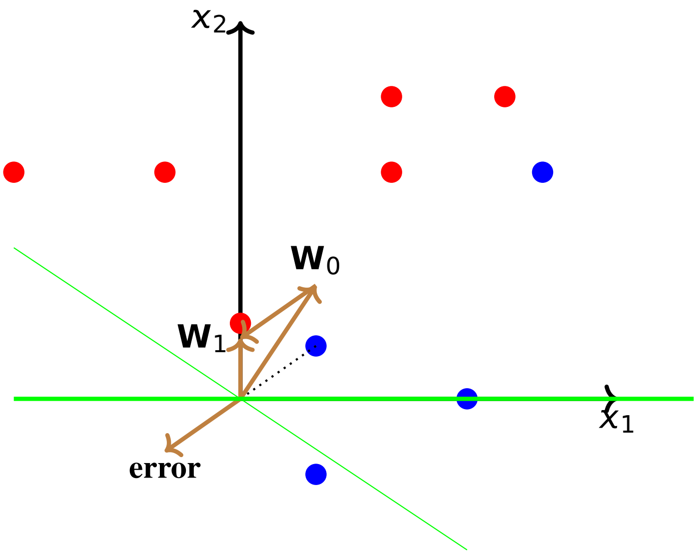


x_2
$\mathbf{w}_0$
x_1
error

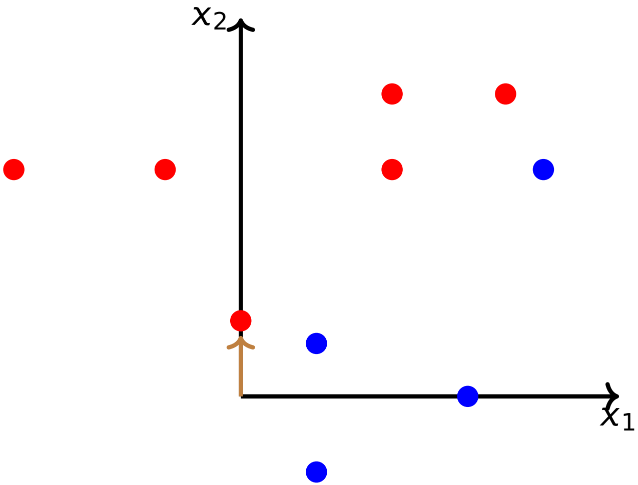


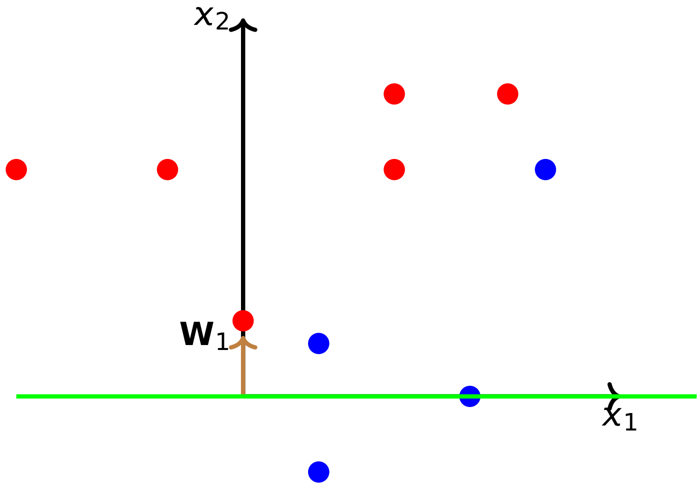
x_2
$\mathbf{w}_0$
x_1
error


x_2
$\mathbf{w}_0$
x_1
error

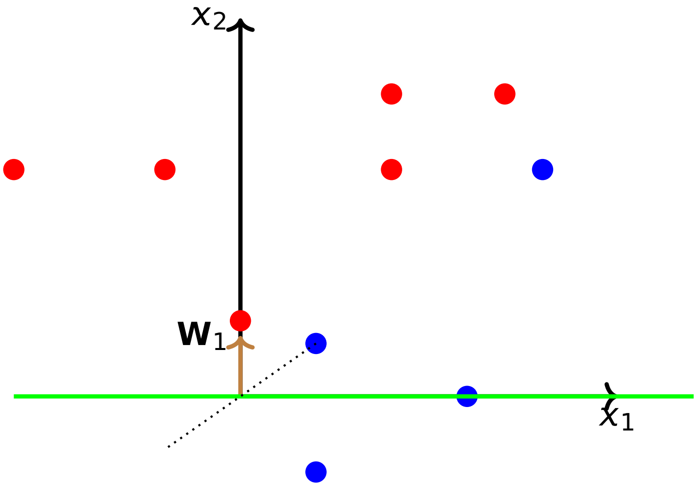


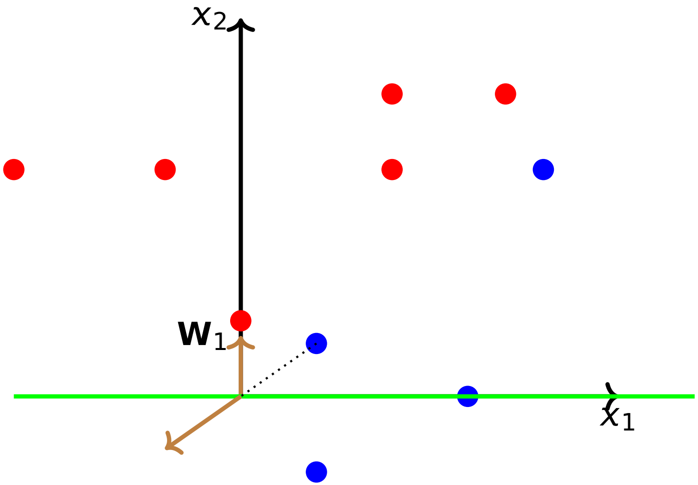
x_2
$\mathbf{w}_0$
$\mathbf{w}_1$
x_1
error


x_2
x_1

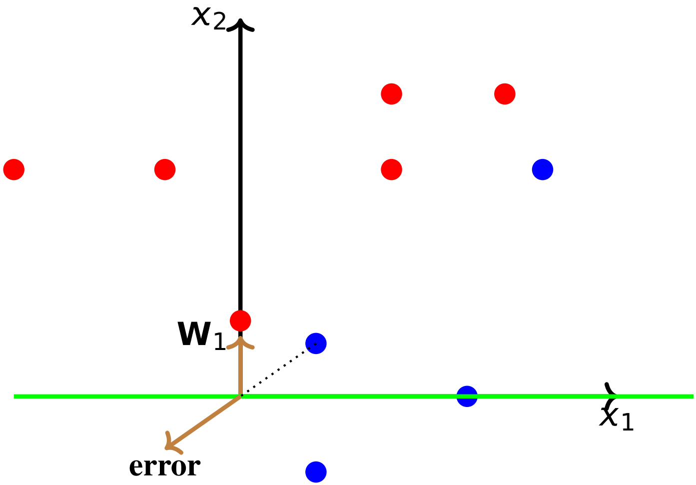


x_2
$\mathbf{w}_1$
x_1

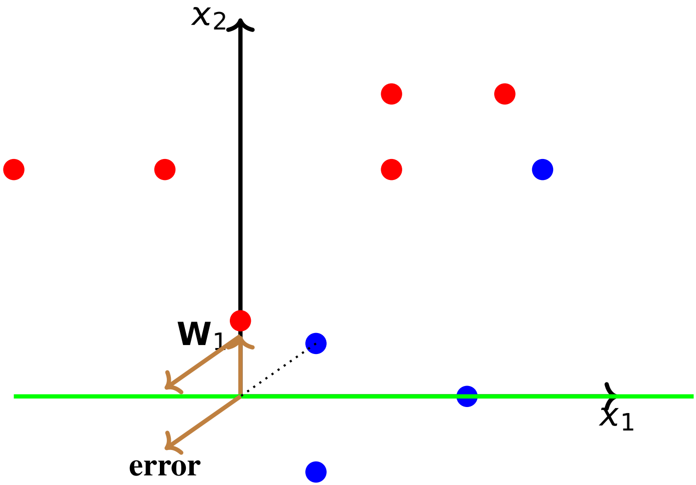

x_2
$\mathbf{w}_1$
x_1



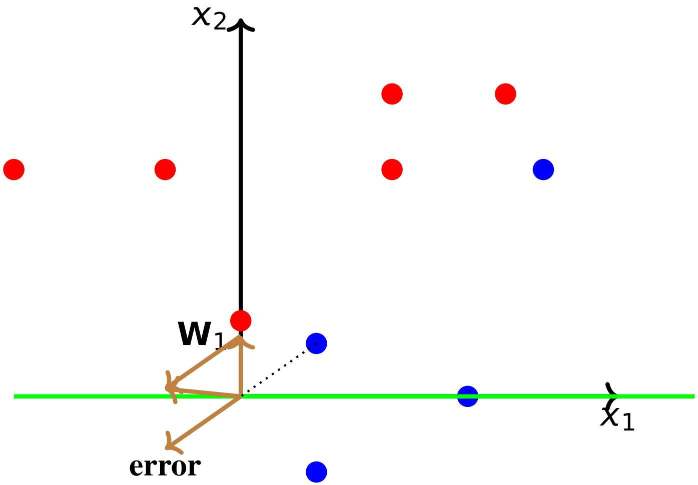
x_2
$\mathbf{w}_1$
x_1



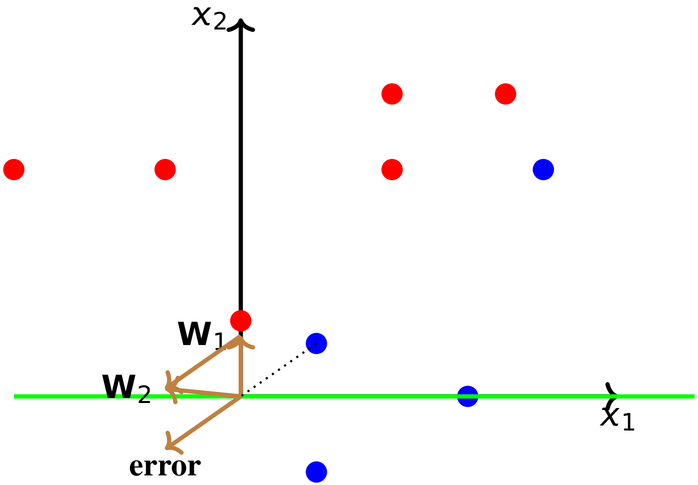
x_2
$\mathbf{w}_1$
x_1
error



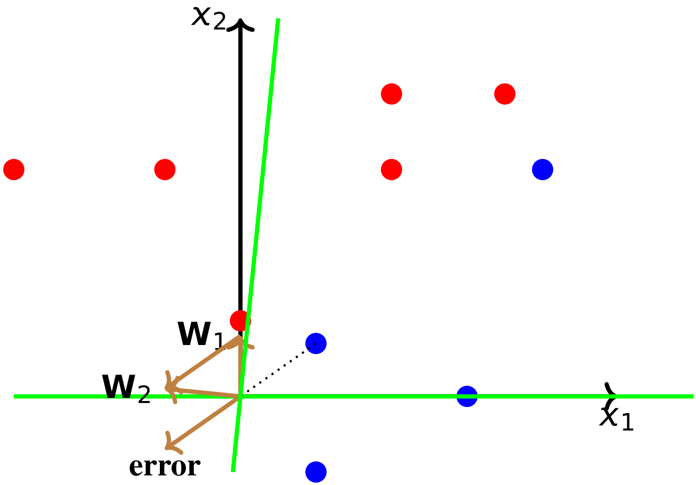
x_2
$\mathbf{w}_1$
x_1
error



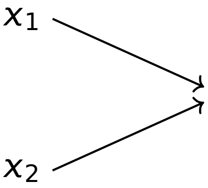
x_2
$\mathbf{w}_1$
x_1
error

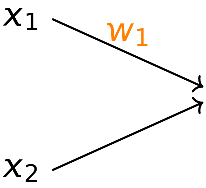


x_2
$\mathbf{w}_1$
$\mathbf{w}_2$
x_1
error

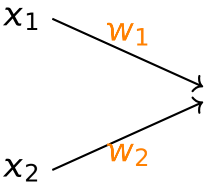


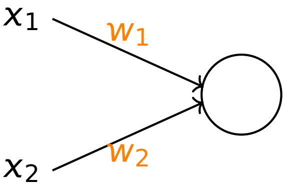
x_2
x_1
$\mathbf{w}_1$
$\mathbf{w}_2$
error

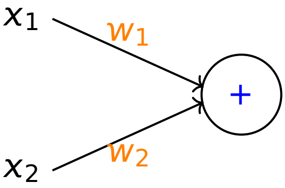

x_1
x_2

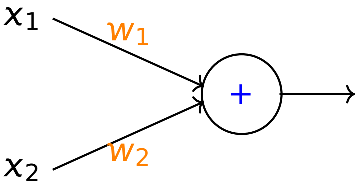


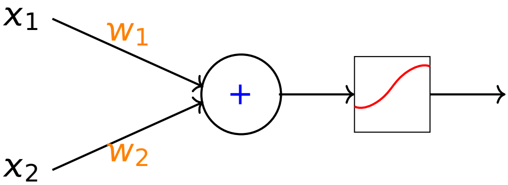
x_1
w_1
x_2


x_1
w_1
x_2
w_2

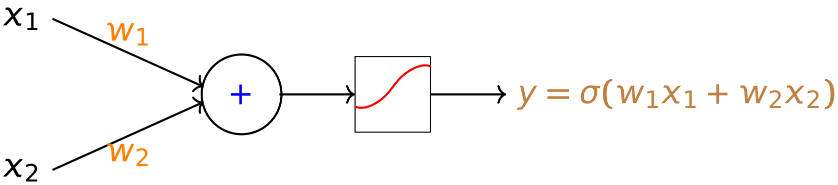

x_1
w_1
x_2
w_2


x_1
w_1
+
x_2
w_2


x_1
w_1
$+$
x_2
w_2



x_1
w_1
+
x_2
w_2


x_1
w_1
x_2
w_2
$+$
$y = \sigma(w_1 x_1 + w_2 x_2)$

y is the probability that the output labe is equal to 1 if x_1 and x_2 are given,

$$\sigma(w_1 x_1^{(i)} + w_2 x_2^{(i)}) = P(label = 1 | x_1^{(i)}, x_2^{(i)})$$

because,

$$w_1 x_1^{(i)} + w_2 x_2^{(i)} \geq 0 \Rightarrow \sigma(w_1 x_1^{(i)} + w_2 x_2^{(i)}) \geq \frac{1}{2}$$

$$w_1 x_1^{(i)} + w_2 x_2^{(i)} < 0 \Rightarrow \sigma(w_1 x_1^{(i)} + w_2 x_2^{(i)}) < \frac{1}{2}$$

So the pmf of the output $\hat{y}^{(i)}$ where $\hat{y}^{(i)} \in \{0, 1\}$ is the Bernoulli distribution, i.e.,

$$P(output = \hat{y}^{(i)} | x_1^{(i)}, x_2^{(i)}) = \sigma(z^{(i)})^{\hat{y}^{(i)}} (1 - \sigma(z^{(i)})^{1-\hat{y}^{(i)}}$$

where $z^{(i)} = w_1 x_1^{(i)} + w_2 x_2^{(i)}$. So we need to solve the optimization problem,

$$\max_{w_1, w_2} P(output = \hat{y}^{(i)} | x_1^{(i)}, x_2^{(i)})$$

but instead we could solve the following minimization problem,

$$\min_{w_1, w_2} -\hat{y}^{(i)} \log(\sigma(z^{(i)})) - (1 - \hat{y}^{(i)}) \log(1 - \sigma(z^{(i)}))$$

If we get average over all sample points we have definition of the **loss function**, i.e. ,

$$\mathcal{L}(w_1, w_2) = \frac{1}{N}\sum_{i=1}^{N} -\hat{y}^{(i)} \log(\sigma(z^{(i)})) - (1 - \hat{y}^{(i)}) \log(1 - \sigma(z^{(i)}))$$

So the goal of the learing is to solve the following minimization problem,

$$\min_{w_1, w_2} \mathcal{L}(w_1, w_2)$$

Gradient Descent

To solve unconstraint optimization problem like,

$$\min_{x_1, x_2} f(x_1, x_2)$$

we make a guss (x_1^0, x_2^0) that minimize the function f and we start improving this guess by moving in a direction that have a smaller value than $f(x_1^0, x_2^0)$. But this direction is in the opposite of the gradient at (x_1^0, x_2^0). So the next step is,

$$(x_1^1, x_2^1) = (x_1^0, x_2^0) - \alpha \mathbf{J}_f(x_1^0, x_2^0)$$

where α is the **learning rate**.

Gradient Descent

1. Make a guess: (x_1^0, x_2^0)
2. $n = 0$
3. Update: $(x_1^{n+1}, x_2^{n+1}) = (x_1^n, x_2^n) - \alpha \mathbf{J}_f(x_1^n, x_2^n)$
4. $n = n + 1$
5. If $\|(x_1^{n+1}, x_2^{n+1}) - (x_1^n, x_2^n)\| < \epsilon$ stop otherwise go to 3.

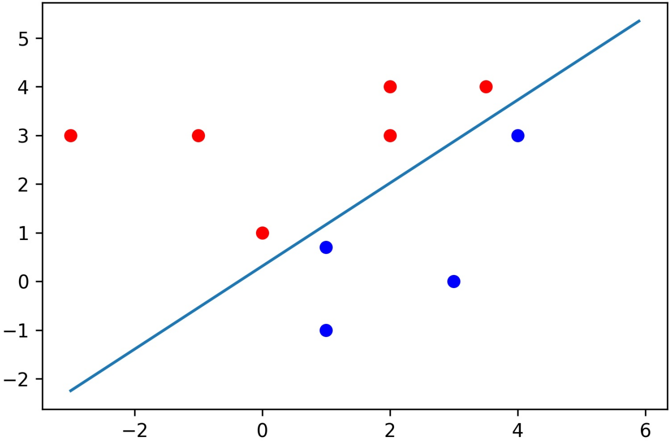


Figure: Gradient descent for Red/Blue example with 50000 iteration and learning rate $\alpha = 0.001$

Thank You