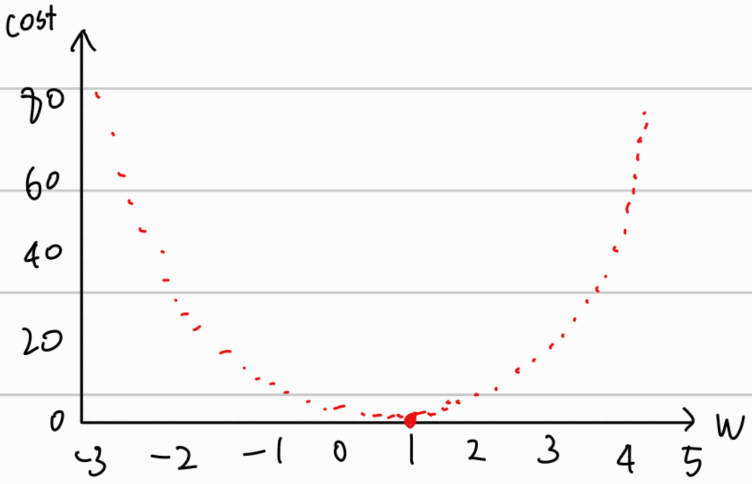
Recall the shape of the cost function:

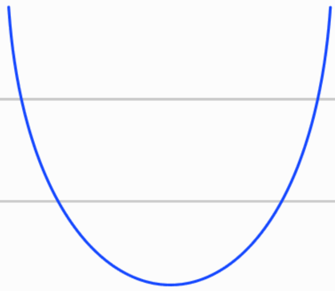
$$cost(W, b) = \frac{1}{m} \sum_{i=1}^{m} (H(x_i) - y_i)^2 \text{ when } H(x) = Wx + b$$



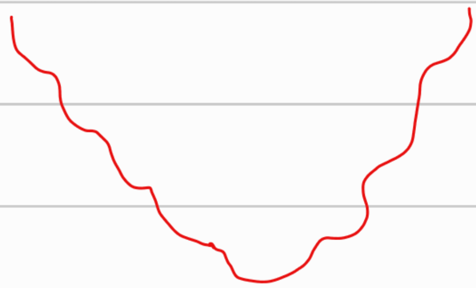
Let us examine the shape of cost w/ the logistic hypothesis:

$$cost(W, b) = \frac{1}{m} \sum_{i=1}^{m} (H(x_i) - y_i)^2$$

$H(x) = Wx + b$



$H(X) = \frac{1}{1 + e^{-W^T X}}$



* since $H(X)$ returns values $0 \sim 1$, the weird shape is the result.

Important observation:

- Gradient descent won't work on the sigmoid hypothesis due to there being multiple local minimums. We can't reliably reach the global minimum.

Solution : New cost function

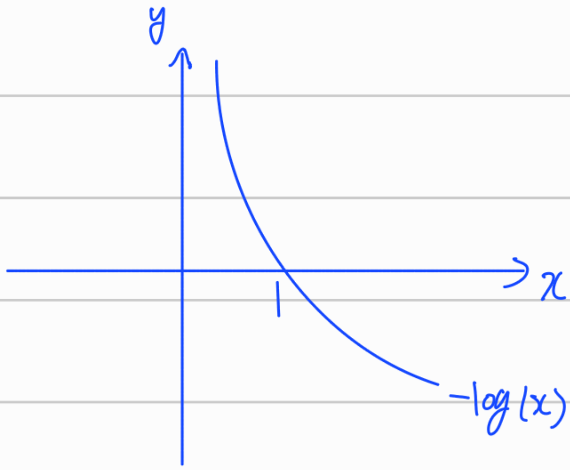
$$cost(W) = \frac{1}{m} \sum c(H(x), y)$$

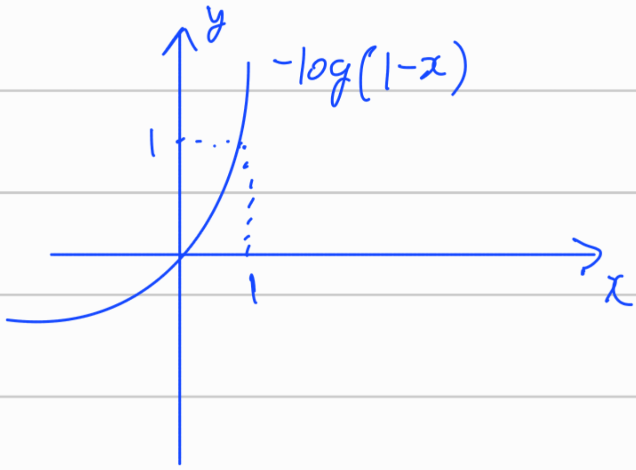
why not ln?

$$c(H(x), y) = \begin{cases} -\log(H(x)) & : y = 1 \\ -\log(1 - H(x)) & : y = 0 \end{cases}$$

Why is this cost function useful?

- Recall $H(x)$ has e in the denominator, w/ x acting acting as the exponent. The log function counteracts this to make a "smooth" continuous cost function
- Examine the behaviour of $-\log(x)$ and $-\log(1-x)$





Note that since our hypothesis cannot be larger than one, we can ignore when $x > 1$.

Consider what the cost function represents.

Cost is minimized when our hypothesis $(H(x))$ is close to our true data (y). We want to maximize this when $H(x)$ is not close to y.

So each value of y has 2 cases each:

$y=1$:

$H(x)=0$:

$$-\log(H(x)) \approx \infty$$

$H(x)=1$:

$$-\log(H(x)) = 0$$

$y=0$:

$H(x)=0$:

$$-\log(1-H(x)) = 0$$

$H(x)=1$:

$$-\log(1-H(x)) = \infty$$

$\therefore$ we can see that if $H(x)$ matches y, the cost is 0

and when $H(x)$ doesn't match y, the cost is ∞.

The cost function can be expressed in one line to cover both cases of y:

$$C(H(x), y) = -y\log(H(x)) - (1-y)\log(1-H(x))$$

Classification Gradient Descent:

- $cost(W) = -\frac{1}{m}\sum y\log(H(x)) + (1-y)\log(1-H(x))$
- $W := W - \alpha\frac{\partial}{\partial W}cost(W)$