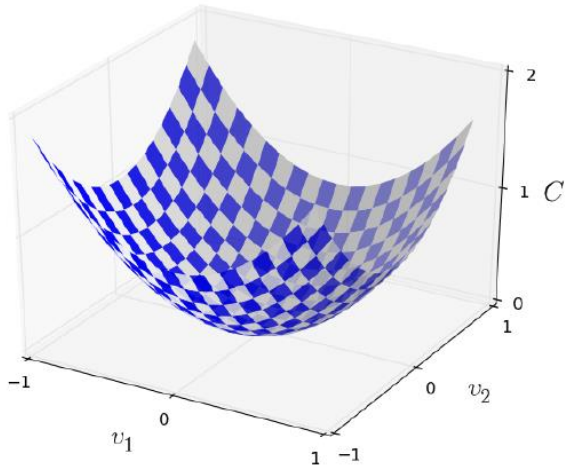
Learning with gradient descent

Assuming C is a function of two variables v_1 and v_2, as illustrated in the figure below, and we want to find where C achieve its global minimum.



let's think about what happens when we move a small amount Δv_1 in the direction v_1, and a small amount Δv_{21} in the direction v_2. Calculus tells us that C changes as follows:

$$\Delta C \approx \frac{\partial C}{\partial v_1}\Delta v_1 + \frac{\partial C}{\partial v_2}\Delta v_2.$$

If we can find a way of choosing Δv_1 and Δv_2 so as to make ΔC negative, we should be able to find where C achieve its global minimum.

Define Δv to be the vector of changes. In 2D, it is:

$$\Delta v \equiv (\Delta v_1, \Delta v_2)^T$$

Where T is the transpose operation.

Also, define the gradient of C to be the vector of partial derivatives:

$$\nabla C \equiv \left(\frac{\partial C}{\partial v_1}, \frac{\partial C}{\partial v_2}\right)^T.$$

Then, ΔC can be rewritten in terms of Δv and ∇C:

$$\Delta C \approx \nabla C \cdot \Delta v$$

If we choose:

$$\Delta v = -\eta \nabla C \qquad \text{(a)}$$

Where η is a small, positive parameter (known as the ***learning rate***)

Then:

$$\Delta C \approx -\eta \nabla C \cdot \nabla C = -\eta \|\nabla C\|^2$$

Because $\|\nabla C\|^2 \geq 0$, this guarantees that $\Delta C \leq 0$, i.e., ***C* will always decrease**, never increase, if we change v according to the prescription above. So we will use equation (a) to compute a value for Δv, then the new value of v is:

$$v \to v' = v - \eta \nabla C$$

To make gradient descent work correctly, we need to choose the learning rate η to be small enough. To use gradient descent to find the weights w_k and biases b_l which minimize the Quadratic cost, we have:

$$w_k \to w'_k = w_k - \eta \frac{\partial C}{\partial w_k}$$

$$b_l \to b'_l = b_l - \eta \frac{\partial C}{\partial b_l}$$

Then, the question is: How to compute the gradient of the cost function?