PROOF THAT GRADIENT DESCENT IS THE BEST WAY TO MINIMIZE COST FN.

Let the cost function be $C(\vec{X})$, where $\vec{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

After each iteration, $X_{n+1} = X_n + \Delta X$

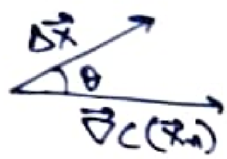
Taylor's Expansion in linear order:

$$C(X_{n+1}) = C(\vec{X}_n) + \Delta\vec{X} \cdot \vec{\nabla} C(\vec{X}_n)$$

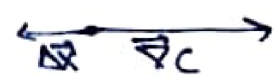
(neglect higher order terms as they will be very small)

then, in order to minimize C, we need to minimize

$$\Delta\vec{X} \cdot \vec{\nabla} C(\vec{X}_n) = |\Delta\vec{X}|\,|\vec{\nabla} C(\vec{X}_n)| \cos\theta$$



→ the value is minimum when $\cos\theta = -1$ $(\theta = \pi)$:



⇒ then, $\Delta\vec{X} = -\alpha \vec{\nabla} C$

where α is a positive constant.

$\Delta\vec{X} = \vec{X}_{n+1} - \vec{X}_n$,

$$\therefore \boxed{\vec{X}_{n+1} = \vec{X}_n - \alpha \vec{\nabla} C(\vec{X}_n)} \Rightarrow \boxed{x_i = x_i - \alpha \frac{\partial C}{\partial x_i}}$$

which is the formula for gradient descent.

And α (learning rate) is taken to be $\ll 1$, which is why the higher order terms (which will have $\alpha^2, \alpha^3, \ldots$) can be neglected.

→ This shows that gradient descent can be used to minimize the cost function.

$\vec{X}_{n+1} = \vec{X}_n - \alpha \vec{\nabla} C(\vec{X}_n)$

let the initial position be $\vec{X}_0$, initial cost C_0

$\vec{X}_1 = \vec{X}_0 - \alpha \vec{\nabla} C(\vec{X}_0)$

Linear order Taylor Expansion:

$C_1 = C_0 + (\vec{X}_1 - \vec{X}_0) \cdot \vec{\nabla} C_0$ (where C_n is $C(\vec{X}_n)$)

$= C_0 - \alpha \vec{\nabla} C_0 \cdot \vec{\nabla} C_0$

$= C_0 - \alpha |\vec{\nabla} C_0|^2$

$C_2 = C_1 - \alpha |\vec{\nabla} C_1|^2$

$\vdots$

$C_n = C_{n-1} - \alpha |\vec{\nabla} C_{n-1}|^2$

$C_n = C_0 - \alpha \left(\sum_{i=0}^{n-1} |\vec{\nabla} C_i|^2 \right)$

$\rightarrow$ if C has a lower bound, then C_n must be finite, which means that $\lim_{n\to\infty} \sum_{i=0}^{n-1} |\vec{\nabla} C_i|^2$ must converge.

in order for the sum to converge, $\lim_{n\to\infty} |\vec{\nabla} C_n|^2 = 0$

($|\vec{\nabla} C_n|^2$ has to approach 0 as $n \to \infty$)

$\therefore \lim_{n\to\infty} |\vec{\nabla} C_n| = 0$ $\rightarrow$ gradient approaches 0 after many iterations.

which means that $X_{n\,(n\to\infty)}$ must be a point of minima.

and $\lim_{n\to\infty}$, $\vec{X}_{n+1} = (\vec{X}_n - \alpha \cancel{|\vec{\nabla} C_n|^2}^{\,0}) = X_n$

so it ~~stays at the pt~~ converges to that pt.