4 1 SGD:

$$X\hat{\omega} = t \Rightarrow X\hat{\omega} - t = 0 \quad \cdots (1)$$

In SGD, all x_i is contained in the span o X. And the SGD update ~~do~~ steps don't ever leave the span of X. Because, $\frac{d}{d\hat{\omega}_p}(x_i\hat{\omega}_p - t_i)^2 = 0$ will ~~giv~~ update $\hat{\omega}_p$ as ~~so~~ some combination of x_i, and t_i.

Thereby, we can assume the SGD solution is spanned by X : $\hat{\omega} = X^T S$, where S is ~~a~~ arbitrary matrix.

From (1), $X\hat{\omega} - t = 0$

$$\Rightarrow XX^T S - t = 0$$

$$\Rightarrow S = (XX^T)^{-1} t$$

So, $\hat{\omega} = X^T (XX^T)^{-1} t$

$$= \omega^*$$

[Showed]

(B)

(4) (2) Mini-batch SGD :

Yes, mini-batch SGD also obtains minimum norm solution on convergence.

Because the batch B is taken from the rows of X. So, the solution $\hat{\omega}$ is spanned by the rows of X.

$$\hat{\omega} = \underbrace{B}_{\text{Batch}} S = X\tilde{S}$$

So, $X\hat{\omega} - t = XX^T\tilde{S} - t = 0$

$$\Rightarrow \tilde{S} = (XX^T)^{-1} t$$

$$\therefore \hat{\omega} = X^T(XX^T)^{-1} t = \omega^*$$

④ ③ Adaptive Methods: Adagrad

$x_1 = [2, 1] \quad \omega_0 = [0, 0] \quad t = [2]$

Using minimum norm solution with GD,

we got $\omega^* = \begin{bmatrix} 0.8 \\ 0.4 \end{bmatrix}$ and

$$\nabla_{\omega^*} \mathcal{L}(\omega) = -2 x_1 t_1$$

Using Adagrad,

$$\omega_0 = [0, 0]$$

$$\omega_1 = \omega_0 - \frac{\eta}{\sqrt{G_{1}} + \epsilon} \nabla_{\hat{\omega}_0} \mathcal{L}(\omega)$$

$$G_1 = 0 + \left(\nabla_{\hat{\omega}_0} \mathcal{L}(\omega)\right)^2$$

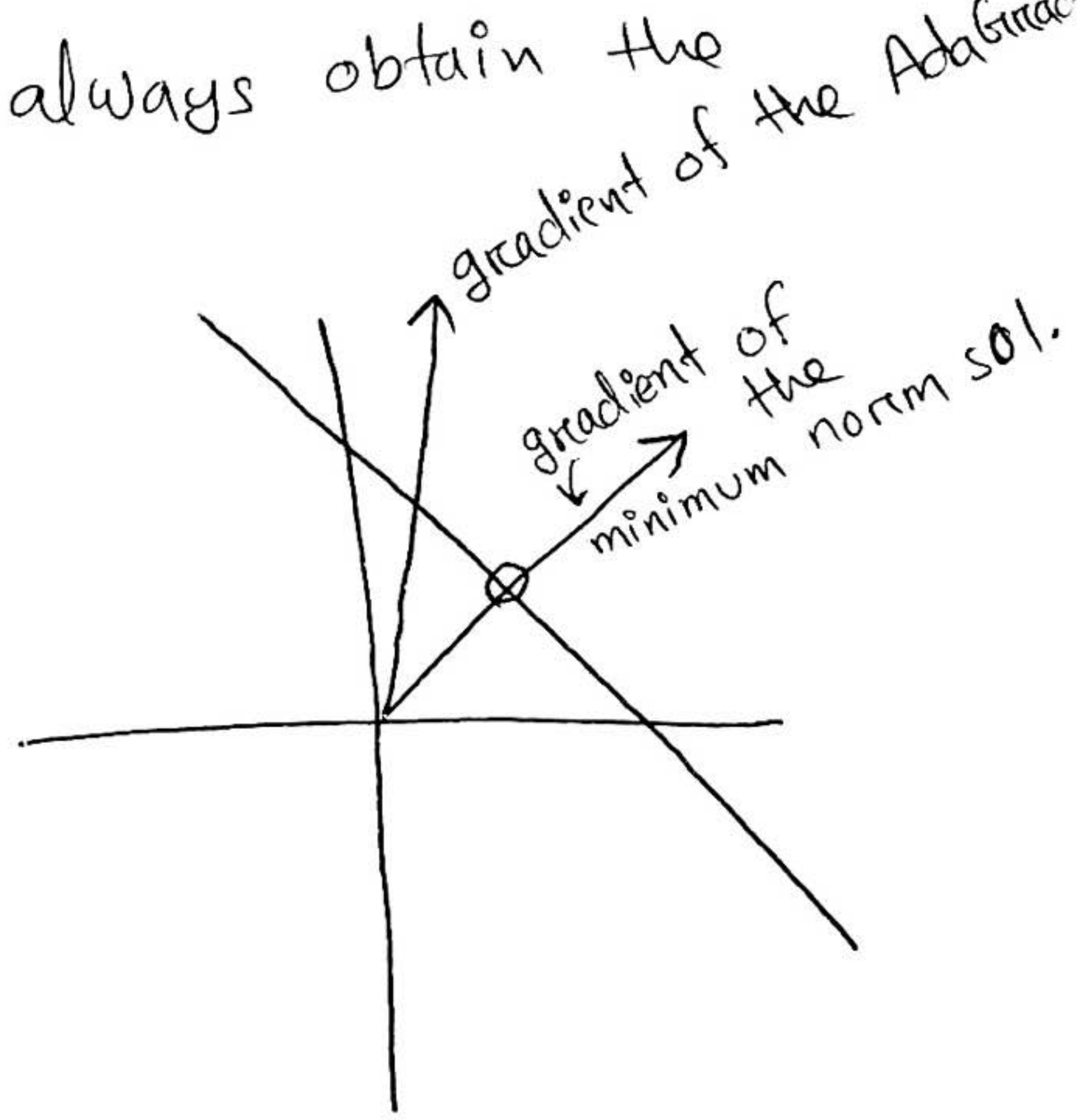
Let, assume, $\nabla_{\hat{\omega}_0} \mathcal{L}(\omega) = -2 x_1 t_1$ [similar to the GD]

then, $\omega_1 = \omega_0 - \dfrac{\eta}{(-2x_1 t_1) + \epsilon} \cdot (-2x_1 t_1)$

as ϵ is small, ω_1 looses x_1 term almost, ~~that means the direct~~ because numerator and denominator both contains $(-2x_1 t_1)$. So,

w_1 has little impact from x_1, which indicates the direction of the gradient is no longer along x_1 as much as the w^* case.
(GD with minimum norm sol.)

Thereby, AdaGrad doesn't always obtain the minimum norm solution.



This ~~Same~~ results holds true for other adaptive ~~models~~ methods (RMSProp, Adam) in general.

Because the scaling part in the weight update may divert solution gradient from the span of X and it may get outside of the span of X sometimes.