CSCI 378: HW2

One of the great joys of abandoning optimality in gradient descent is that you can also abandon other social mores of optimization, like the common assumption that the loss function is differentiable everywhere.

Suppose our loss function is $L(\theta) = |\theta - 3|$ for $\theta \in \mathbb{R}$.

(a) Where is $L(\theta)$ non-differentiable?

(b) Trace through the first 4 time steps of VANILLAGD, MOMENTUMGD, and ADAGRAD for $L(\theta)$, if $\theta^{(0)} = 3.75$

	VANILLA ($\alpha = 0.5$)	MOMENTUM ($\alpha = 0.5, \mu = 0.2$)	ADAGRAD ($\alpha = 0.5, \delta = 0$)
$\theta^{(0)}$	3.75	3.75	3.75
$\theta^{(1)}$			
$\theta^{(2)}$			
$\theta^{(3)}$			
$\theta^{(4)}$			

(c) Give a starting value for $\theta^{(0)}$ where VANILLA will break down.

(d) Why is such a situation unlikely to happen in theory?

(e) Why might it happen anyway in practice?