The main idea behind AAdam is to speed up the progress along dimensions in which gradient consistently points in the same direction as the current step, and to decrease progress along dimensions where the sign of the gradient continues to change. In addition to storing an exponentially decaying average of past squared gradients v_t and an exponentially decaying average of past gradients m_t as Adam does, AAdam also keeps an exponentially decaying average of past updates (which act like the momentum). Thus, the current update not only depends on the previous gradients, it also depends on the previous values of the update $\Delta\theta$. We keep track of past parameters updates with an exponential decay where (1 - β_1) (approximatively 0.9, the same β_1 in Adam) is the constant which is multiplied by the sign of the current gradient. The new update rule is summarized as follows:

$$\begin{aligned} \theta_{n+1} &= \theta_n - (\eta \frac{\beta_1}{\sqrt{\hat{v_n} + \epsilon}} \hat{m}_n + d) \\ d &= \Delta\theta_{n-1} * sign(\nabla_{\theta_n} J(\theta)) * (1 - \beta_1) \\ \text{where} \quad \hat{m_n}, \hat{v_n} &are \text{ Adam Parameters} \\ \text{and} \quad \Delta\theta_{n-1} &\text{is the Previous update step.} \end{aligned}$$

This very first idea (AAdam_Old) actually works well. However, the results are sometimes mitigate since the sign of the gradient may result to a very bad behavior even for simple convex problems such as logistic regression. To cope with this issue, we move out the sign and use its variation as an heuristic instead.

$$\begin{aligned} d &= -\Delta\theta_{n-1} * (1 - \beta_1) \quad \text{if sign}(\nabla_{\theta_n} J(\theta)) * sign(\nabla_{\theta_{n-1}} J(\theta)) < 0 \\ d &= \Delta\theta_{n-1} * (1 - \beta_1) \quad \text{if sign}(\nabla_{\theta_n} J(\theta)) * sign(\nabla_{\theta_{n-1}} J(\theta)) > 0. \end{aligned}$$

The value d is added to the update step size if the gradient did not change its direction, otherwise it is subtracted. Thus, if the size of past updates decreases, then the size of the current step will tend to decrease linearly, otherwise, it will increase. It is important to consider the absolute value of d since its current sign could change the expected behavior. We multiplied d by -1 if the direction changes and by $+1$ otherwise. In this way, we ensure that the value d

is actually added or subtracted from the current update step size of Adam.