Algorithm: AdamE is the proposed algorithm for stochastic optimization. g_t^2 represents the elementwise multiplication of g_t by itself. The acceptable default learning rate $\alpha =$. The choice of M depends on a given problem and network architecture. Vector operations are element-wise.

Require

α: Learning rate (stepsize)
M: Averaging window
$f(\theta)$: Objective function with trainable parameters θ
θ_0: Initial values for all the parameters
$m_0 \leftarrow 0$ (First moment vector initialization)
$v_0 \leftarrow 0$ (Second moment vector initialization)
$g_M \leftarrow 0$ (Values of the gradient at $t - M + 1$ time step)
$t \leftarrow 0$ (Initial timestep)

while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Obtain gradients at timestep t)
if $t \leftarrow 1$ **do**
$g_M \leftarrow g_t$ (Fill in the buffer g_M with gradients required for the first averaging)
end if
if $t \ mod \ M + 1$ **do** (Averaging every M iterations)
$m_t \leftarrow m_{t-1} + \frac{1}{M}(g_t - g_{t-M+1})$ (Do averaging for first moment estimate)
$v_t \leftarrow v_{t-1} + \frac{1}{M}(g_t^2 - g_{t-M+1}^2)$ (Do averaging for the second moment estimate)
else do
$m_t \leftarrow m_{t-1} + \frac{1}{M} g_t$ (Biased first moment evaluation update)
$v_t \leftarrow v_{t-1} + \frac{1}{M} g_t^2$ (Biased first moment evaluation update)
end if

$\widehat{m}_t \leftarrow M \cdot m_t$ (Perform bias correction of the first moment estimate)
$\hat{v}_t \leftarrow M \cdot v_t$ (Perform bias correction of the second moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \frac{\widehat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}$ (Make an update to parameters)
if $t \ mod \ M + 1$ **do**
$m_t \leftarrow 0$
$v_t \leftarrow 0$
end if
end while
return θ_t (Final parameters)