Adam

input : starting point $\vec{x}_0$, step factor α, decay rates γ_v, $\gamma_s \in [0, 1)$, small constant ϵ, tolerance E_s

output: $\vec{x}^*$, estimate of the minimizer, and $f(\vec{x}^*)$

// defaults from the original paper: $\alpha = 0.001$, $\gamma_v = 0.9$, $\gamma_s = 0.999$, $\epsilon = 10^{-8}$

// all vector products, square roots, and divisions below are elementwise

$\vec{v}_0 \leftarrow \vec{0}$, $\vec{s}_0 \leftarrow \vec{0}$, $k \leftarrow 0$

while $\|\nabla f(\vec{x}_k)\| > E_s$ **do**

$\vec{g}_k \leftarrow \nabla f(\vec{x}_k)$

$\vec{v}_{k+1} \leftarrow \gamma_v \vec{v}_k + (1 - \gamma_v) \vec{g}_k$ // biased decaying momentum

$\vec{s}_{k+1} \leftarrow \gamma_s \vec{s}_k + (1 - \gamma_s) (\vec{g}_k \odot \vec{g}_k)$ // biased decaying sq. gradient

$\hat{\vec{v}}_{k+1} \leftarrow \vec{v}_{k+1} / (1 - \gamma_v^{k+1})$ // corrected decaying momentum

$\hat{\vec{s}}_{k+1} \leftarrow \vec{s}_{k+1} / (1 - \gamma_s^{k+1})$ // corrected decaying sq. gradient

$$\vec{x}_{k+1} \leftarrow \vec{x}_k - \frac{\alpha}{\epsilon + \sqrt{\hat{\vec{s}}_{k+1}}} \odot \hat{\vec{v}}_{k+1}$$ // next iterate

$k \leftarrow k + 1$

end

return $\vec{x}^* = \vec{x}_k$, $f(\vec{x}^*)$