**Algorithm 7.4:** Adam optimization

```
Input: Training set of data points indexed by $n \in \{1, \dots, N\}$
       Batch size $B$
       Error function per mini-batch $E_{n:n+B-1}(\mathbf{w})$
       Learning rate parameter $\eta$
       Decay parameters $\beta_1$ and $\beta_2$
       Stabilization parameter $\delta$
Output: Final weight vector $\mathbf{w}$
$n \leftarrow 1$
$\mathbf{s} \leftarrow \mathbf{0}$
$\mathbf{r} \leftarrow \mathbf{0}$
repeat
    Choose a mini-batch at random from $\mathcal{D}$
    $\mathbf{g} = -\nabla E_{n:n+B-1}(\mathbf{w})$ // evaluate gradient vector
    $\mathbf{s} \leftarrow \beta_1 \mathbf{s} + (1 - \beta_1)\mathbf{g}$
    $\mathbf{r} \leftarrow \beta_2 \mathbf{r} + (1 - \beta_2)\mathbf{g} \odot \mathbf{g}$ // element-wise multiply
    $\widehat{\mathbf{s}} \leftarrow \mathbf{s}/(1 - \beta_1^{\tau})$ // bias correction
    $\widehat{\mathbf{r}} \leftarrow \mathbf{r}/(1 - \beta_2^{\tau})$ // bias correction
    $\Delta\mathbf{w} \leftarrow -\eta \dfrac{\widehat{\mathbf{s}}}{\sqrt{\widehat{\mathbf{r}}} + \delta}$ // element-wise operations
    $\mathbf{w} \leftarrow \mathbf{w} + \Delta\mathbf{w}$ // weight vector update
    $n \leftarrow n + B$
    if $n + B > N$ then
        shuffle data
        $n \leftarrow 1$
    end if
until convergence
return $\mathbf{w}$
```