Environment

$\vec{o}_t, r_t, d_t$

$\vec{o}_t$

$(\vec{a}_t, \vec{o}_t, r_t, d_t)$
$\vec{a}_t \in \mathcal{A}$
$\vec{o}_t \in \Omega$
$r_t \in \mathbb{R}$
$d_t \in \{\text{TRUE}, \text{FALSE}\}$
ER Buffer

The paper uses a model that takes as inputs $(\vec{o}_t, r_t, d_t, \vec{h}_t)$. I.e., the same architecture as the DQN network. However, they only input a single observation $\vec{o}_t$, no other variables.

Outside of training, the algorithm keeps a rolling average for the $\mu_{\vec{z}}$ and $\Sigma_{\mathbb{Z}}$. Not sure why, these statistics are only used to compute the loss.

Encoder
(A.K.A $\Phi\colon \Omega \to \mathbb{R}^{32}$)

Conv: 32 filters, size 8×8, stride 4
ReLU
Conv: 2 filters, size 4×4, stride 2
(64 channels out!)
ReLU
Conv: 1 filter, size 3×3, stride 1
ReLU
FC ($64 \cdot 7 \cdot 7$ inputs, 32 out for Atari)

$\vec{z}_t \in \mathbb{R}^{32} = \mathcal{Z}$

Batch of N pairs $(\vec{z}_{t_i}, \vec{z}_{t_i+k})$ for (non-chronological) timesteps $t_1, t_2, \ldots t_N$.

W-MSE Loss for Encoder

Pass through encoder Φ

$\mathbf{Z} \leftarrow [\Phi(\vec{o}_{t_1})\Phi(\vec{o}_{t_1+k})\Phi(\vec{o}_{t_2})\Phi(\vec{o}_{t_2+k}) \ldots \Phi(\vec{o}_{t_N})\Phi(\vec{o}_{t_N+k})]$
Compute sample mean:
$\vec{\mu}_{\mathbf{Z}} \leftarrow \frac{1}{2N}\sum_{i=1}^{2N} \mathbf{Z}_i$
Compute sample covariance matrix:
$\vec{\Sigma}_{\mathbf{Z}} \leftarrow \frac{1}{2N-1}\sum_{i=1}^{2N}(\mathbf{Z}_i - \vec{\mu}_{\mathbf{Z}})(\mathbf{Z}_i - \vec{\mu}_{\mathbf{Z}})^\top$
Apply *Cholesky decomposition:*
$\mathbf{W}_{\mathbf{Z}}^\top \mathbf{W}_{\mathbf{Z}} \leftarrow \vec{\Sigma}_{\mathbf{Z}}^{-1}$
"Whiten" the latent encodings for each $i \in \{t_1, t_1+k, \ldots, t_n, t_N+k\}$:
$\vec{v}_i \leftarrow \mathbf{W}_{\mathbf{Z}}(\vec{Z}_i - \vec{\mu}_{\mathbf{Z}})$
Compute loss for encoder:
$L_\Phi = L_{\text{W-MSE}} \leftarrow \frac{1}{N}\sum_{(t_i, t_i+k), 1 \le i \le N} ||\vec{v}_{t_i} - \vec{v}_{t_i+k}||_2^2$
(Note the *pairwise* Euclidean distances)

This "whitening" achieves that:
- The mean of $\mathbf{V} = [\vec{v}_{t_1}, \vec{v}_{t_1+k}, \ldots, \vec{v}_{t_N+k}]$ is $\vec{0}$.
- The covariance of each $\vec{v}_i$ is I_{32} (i.e., its entries are independent).

$\hat{\vec{z}}_{t+1}$

Latent World Model

FC $\mathbb{R}^{256} \to \mathbb{R}^{32} = \mathcal{Z}$
FC $\mathbb{R}^{256} \to \mathbb{R}^{256}$
GRU
$\mathbb{R}^{128} \times \mathbb{R}^{256} \to \mathbb{R}^{256} \times \mathbb{R}^{256}$
FC $\mathbb{R}^{32+|\mathcal{A}|} \to \mathbb{R}^{128}$

$(\vec{h}_0 = \vec{0})$ $\vec{h}_t$ $\vec{h}_{t+1}$

$\begin{bmatrix}\vec{z}_t \\ \vec{a}_t\end{bmatrix}$

Agent & LWM loss

$\vec{r}_{i,t+1} \leftarrow ||\hat{\vec{z}}_{t+1} - \vec{z}_{t+1}||_2^2$
$\mu_{t+1} \leftarrow m \cdot \mu_t + (1-m) \cdot mean(\vec{r}_i)$
$\sigma_{t+1} \leftarrow m \cdot \sigma_t + (1-m) \cdot std(\vec{r}_i)$
$r_{\text{DQB},t+1} \leftarrow r_t + \beta \times \min(10, \max(-10, \frac{r_{i,t} \cdot \mu_{t+1}}{\sigma_{t+1}}))$

$r_{\text{DQN},t+1}$ is loss for the agent

Why is $\vec{r}_i$ a vector? It is in the code, unclear in supplementary.

$r_{i,t+1}$ (pred. err.) is loss for LWM

Legend:

- $\mathcal{Z}$: latent state space.
- d_t: Boolean Episode ended y/n.
- $\hat{\vec{z}}_t$: *predicted* latent state.
- $\vec{o}_t$: obsevation, 1 game frame
- k: timesteps between temporally close observations. ($k = 2$ in paper)

Variable.
Neural Network.

Note: tensors are processed in batches