Supplementary Material: Autoregressive DRL with Learned Intrinsic Rewards for Portfolio Optimisation

Magdalene Hui Qi Lim
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
magdalenehq.lim@gmail.com

Nixie S Lesmana
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
Department of Physics, Faculty of Science, National University of Singapore, Singapore
nixiesap001@e.ntu.edu.sg

Chi Seng Pun
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
cspun@ntu.edu.sg

1 PSEUDOCODE FOR AUGMENTED AUTOA2C

To supplement our proposed algorithm as outlined in section 4 of the main paper, we provide pseudocodes for Augmented AutoA2C. We address the MMDP decomposition in Algorithm 1, and training and trading specifications in Algorithm 2.

Algorithm 1 Rollout for Autoregressive Model

Init: initialize $\mathcal{D} = \{\}$ and action vector $\mathbf{a_t} = (0, 0, \ldots, 0) \in \mathbb{R}^{30}$ corresponding to day t of MDP
for stock $d = 1$ to 30 **do**
  Input $u^{s_t}_{d-1} = \text{concat}(s_t, \mathbf{a_t})$
  Sample action $a^d_t \in \mathbb{R}$ using π_θ
  Update action vector $\mathbf{a_t}[\text{position } d] \leftarrow a^d_t$
  Observe intrinsic reward $r^{\text{in}}(u^{s_t}_{d-1}, a^d_t, u^{s_t}_d) \sim \pi_\eta$
  if $d < 30$ **then**
    Get reward $r^{\text{ex}}(u^{s_t}_{d-1}, a^d_t) = 0$
  else
    Execute action $\mathbf{a_t}$ in environment
    Observe next state s_{t+1}
    Observe inter-day reward $r^{\text{ex}}(u^{s_t}_{29}, a^{30}_t) = (b_{t+1} + \mathbf{p_{t+1}}^T \mathbf{h_{t+1}}) - (b_t + \mathbf{p_t}^T \mathbf{h_t})$
    Update $s_t \leftarrow s_{t+1}$
  end if
  Append $(u^{s_t}_{d-1}, a^d_t, u^{s_t}_d, r^{\text{ex}}, r^{\text{in}})$ to $\mathcal{D}$
end for
Return trajectory $\mathcal{D}$

For training, we maintain a reward buffer $\mathcal{R}$ for the purpose of transforming the inter-day rewards into some external reward function $U(\cdot)$ (see section 5.2) dependent on some history of inter-day rewards.

Algorithm 2 Augmented AutoA2C

Input: step size parameters α, β
for each moving window $\mathcal{W}$ **do**
  Initialize network parameters θ, η
  Initialize day $t = 0$
  Initialize reward buffer $\mathcal{R}$
  # Train
  repeat
    Sample trajectory $\mathcal{D} \sim \pi_\theta, \pi_\eta$ by Algorithm 1
    Store inter-day reward $r^{\text{ex}}(s_t, \mathbf{a_t})$ in $\mathcal{R}$
    Replace $r^{\text{ex}}(s_t, \mathbf{a_t})$ in $\mathcal{D}$ with $U(s_t, \mathbf{a_t}, s_{t+1}; \mathcal{R})$ or $U(s_{0:t+1}, \mathbf{a_{0:t}}; \mathcal{R})$
    Approximate $\nabla_\theta J^{\text{ex+in}}(\theta; \mathcal{D})$ by Equation 6
    Update policy actor $\theta' \leftarrow \theta + \alpha \nabla_\theta J^{\text{ex+in}}(\theta; \mathcal{D})$
    Update policy critic θ_v by Equation 7
    Approximate $\nabla_{\theta'} J^{\text{ex}}(\theta'; \mathcal{D})$ by Equation 11
    Approximate $\nabla_\eta \theta'$ by Equation 12
    Compute $\nabla_\eta J^{\text{ex}} = \nabla_{\theta'} J^{\text{ex}}(\theta'; \mathcal{D}) \nabla_\eta \theta'$
    Update intrinsic reward actor $\eta' \leftarrow \eta + \beta \nabla_\eta J^{\text{ex}}$
    Update intrinsic reward critic η_v by Equation 8
    $t \leftarrow t + 1$
  until done
  # Trade
  for trade day $t = 1$ to 63 **do**
    Sample d subactions $a^1, \ldots, a^d$ corresponding to s_t according to π_θ
    Execute action $\mathbf{a_t} = [a^1, \ldots, a^d]$ in environment
    Record portfolio value $b_t + \mathbf{p_t}^T \mathbf{h_t})$
  end for
end for