Algorithm 1 Priority Experience Replay Constrained Stackelberg Q-Learning with MIP action selection (MIP-PCSQ)

Input: thresholds d_1, d_2, discount γ, soft-update rate ρ, exploration ε; PER hyper-parameters $\alpha, \beta_{\text{start}}, \varepsilon_p$; updates per step K; buffer capacity N
Init: critics ϕ_i (for Q_i), cost-critics ζ_i (for G_i), $i \in \{1, 2\}$; targets $\phi_i^{\text{targ}} \leftarrow \phi_i$, $\zeta_i^{\text{targ}} \leftarrow \zeta_i$; replay buffer D with PER; set $\beta \leftarrow \beta_{\text{start}}$
for $t = 1, 2, \ldots$ **do**
    Observe state s
    Build tables (online): $Q_1^{ij}(s), Q_2^{ij}(s), G_1^{ij}(s), G_2^{ij}(s)$; define safe sets $\mathcal{S}_i(s) = \{j \mid G_1^{ij}(s) \le d_1,\ G_2^{ij}(s) \le d_2\}$
    if rand() $< \varepsilon$ **then**
        pick (a_1, a_2) uniformly from $\bigcup_i \{(a_1^i, a_2^j) \mid j \in \mathcal{S}_i(s)\}$
    else
        Solve MIP (1) at s (use (13a)–(13h)) to get (a_1, a_2)
    end if
    Execute (a_1, a_2); observe r_i, c_i, s', d
    Push $(s, a_1, a_2, r_i, c_i, s', d)$ into D with priority $p_{\max}$
    for $k = 1$ **to** K **do** ▷ K critic updates per env-step
        $(\text{idx}, B, P) \leftarrow \text{PER.Sample}(D, |B|, \text{stratified} = \text{true})$
        $w \leftarrow \left(\frac{1/N}{P}\right)^{\beta}$; $\tilde{w} \leftarrow w / \max(w)$
        Build $Q_1^{ij}(s'), Q_2^{ij}(s'), G_1^{ij}(s'), G_2^{ij}(s')$ (online); define $\mathcal{S}_i(s')$
        Solve MIP (1) at s' to get (a_1', a_2')
        Targets (use target nets): $y_i = r_i + \gamma(1 - d) Q_i^{\text{targ}}(s', a_1', a_2')$, $\quad g_i = c_i + \gamma(1 - d) G_i^{\text{targ}}(s', a_1', a_2')$
        Residuals: $dQ_i \leftarrow Q_i(s, a_1, a_2) - y_i$, $\quad dG_i \leftarrow G_i(s, a_1, a_2) - g_i$
        Losses (weighted mean over B): $\mathcal{L}_{Q_i} = \frac{1}{|B|} \sum_{u \in B} \tilde{w}(u)\, [dQ_i(u)]^2$, $\quad \mathcal{L}_{G_i} = \frac{1}{|B|} \sum_{u \in B} \tilde{w}(u)\, [dG_i(u)]^2$
        GD steps: $\phi_i \leftarrow \phi_i - \eta_Q \nabla_{\phi_i} \mathcal{L}_{Q_i}$, $\quad \zeta_i \leftarrow \zeta_i - \eta_G \nabla_{\zeta_i} \mathcal{L}_{G_i}$
        Priority update (PER): $\Delta \leftarrow \lambda_{Q1}|dQ_1| + \lambda_{Q2}|dQ_2| + \lambda_{G1}|dG_1| + \lambda_{G2}|dG_2|$; $p_{\text{new}} \leftarrow (\Delta + \varepsilon_p)^{\alpha}$; PER.Update(idx, p_{new})
        Soft-update targets: $\phi_i^{\text{targ}} \leftarrow \rho\, \phi_i^{\text{targ}} + (1 - \rho)\phi_i$; $\zeta_i^{\text{targ}} \leftarrow \rho\, \zeta_i^{\text{targ}} + (1 - \rho)\zeta_i$
    end for
    if $d = 1$ **then** reset environment
    end if
end for

MIP (13a–13h) at state x ($x \in \{s, s'\}$):

$$\mathcal{S}_i(x) = \{j \mid G_1^{ij}(x) \le d_1,\ G_2^{ij}(x) \le d_2\};$$

$$
\begin{aligned}
\max_{x,y,v} \quad & \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} Q_1^{ij}(x)\, y_{ij} && \text{(13a)}\\
\text{s.t.} \quad & \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} y_{ij} = 1 && \text{(13b)}\\
& \sum_{j=1}^{m_2} y_{ij} = x_i,\ \forall i && \text{(13c)}\\
& x_i = 1 \Rightarrow v_i \geq Q_2^{i\ell}(x),\ \forall i,\ \forall \ell \in \mathcal{S}_i(x) && \text{(13d)}\\
& y_{ij} = 1 \Rightarrow v_i \geq Q_2^{ij}(x),\ \forall i, j && \text{(13e)}\\
& y_{ij} = 1 \Rightarrow v_i \leq Q_2^{ij}(x),\ \forall i, j && \text{(13f)}\\
& y_{ij} = 1 \Rightarrow G_1^{ij}(x) \leq d_1,\ \forall i, j && \text{(13g)}\\
& y_{ij} = 1 \Rightarrow G_2^{ij}(x) \leq d_2,\ \forall i, j && \text{(13h)}
\end{aligned}
$$