Algorithm 1 DRL-TSBC 的学习过程

Require: 超参数: 经验池大小 M, 批次大小 B, 折扣因子 γ, 随机动作概率 ϵ, 最大模拟次数 E, 学习频率 P, 模型更新频率 O, 首班车时间 t_s, 末班车时间 t_e

Ensure: 训练好的 DQN 参数 θ

初始化经验池 D，主网络参数 θ，和目标网络参数 $\theta^- = \theta$
使用 θ 初始化主网络 $Q(s, a; \theta)$，并且使用 θ^- 初始化目标网络 $Q(s, a; \theta^-)$
for $episode = 1$ 到 E **do**
  初始化公交仿真环境
  for 所有 $i = t_s$ 转到 t_e 进行 **do**
    if 在 $[0, 1]$ 区间内随机取出的一个实数小于 ϵ 则 **then**
      随机选择动作 a
    else
      $a = \arg\max_a Q(s, a; \theta)$
    end if
    通过 a 计算 I^{up} and I^{down}
    if $I^{up} < T_{min}$ 则 **then**
      $a_{up} \leftarrow 0$
    end if
    if $I^{up} > T_{max}$ 则 **then**
      $a_{up} \leftarrow 1$
    end if
    if $I^{down} < T_{min}$ 则 **then**
      $a_{down} \leftarrow 0$
    end if
    if $I^{down} > T_{max}$ 则 **then**
      $a_{down} \leftarrow 1$
    end if
    $a \leftarrow (a_{up}, a_{down})$
    在环境中应用动作 a 并得到奖励 r 和下一个状态 s'
    将四元组 (s, a, r, s') 添加到经验池 D 中，同时如果 $|D| > M$，删除经验池中最旧的经验
    $s = s'$, $i = i + 1$
    if $|D| > M$ 且 $i \bmod P = 0$ 则 **then**
      从经验池 D 中随机取出 B 个四元组 (s, a, r, s')
      通过式 (2.18) 计算损失函数 L
      使用 Adam 反向传播通过 L 更新 θ
      每进行 O 次学习，更新目标网络参数 $\theta^- = \theta$
    end if
  end for
end for

Algorithm 2 DRL-TSBC 的推理过程

Require: 训练好的 DQN 参数 θ, 最大发车间隔 $T_{\max}$, 最小发车间隔 $T_{\min}$
Ensure: 最终公交时刻表（发车时间点列表）

初始化公交仿真环境
for 所有 $i = t_s$ 转到 t_e 进行 **do**
  $a = \arg\max_a Q(s, a; \theta)$
  通过 a 计算 I^{up} and I^{down}
  if $I^{up} < T_{min}$ 则 **then**
    $a_{up} \leftarrow 0$
  end if
  if $I^{up} > T_{max}$ 则 **then**
    $a_{up} \leftarrow 1$
  end if
  if $I^{down} < T_{min}$ 则 **then**
    $a_{down} \leftarrow 0$
  end if
  if $I^{down} > T_{max}$ 则 **then**
    $a_{down} \leftarrow 1$
  end if
  $a \leftarrow (a_{up}, a_{down})$
  在环境中应用动作 a 并得到下一个状态 s'
  $s = s'$, $i = i + 1$
end for
选择发车次数更多的方向
从时刻表中删除该方向的倒数第二次发车
$k \leftarrow$ 此时该方向的总发车次数
while 时刻表中的第 k 次发车和第 $k-1$ 次发车的间隔大于 T_{max} 进行 **do**
  将第 $k-1$ 次发车的时间推迟直到其与下一次发车的间隔为 T_{max}
  $k = k - 1$
end while