**Input:** A confidence parameter $\delta \in (0,1)$, $\mathcal{S}$ and $\mathcal{A}$.

**Initialization:** Set $t := 1$, and observe the initial state $s_1$.

**For** episodes $k = 1, 2, \ldots$ **do**

**Initialize episode** $k$:

1. Set the start time of episode $k$, $t_k := t$.
2. For all $(s,a)$ in $\mathcal{S} \times \mathcal{A}$ initialize the state-action counts for episode $k$, $v_k(s,a) := 0$. Further, set the state-action counts prior to episode $k$,
$$N_k(s,a) := \#\{\tau < t_k : s_\tau = s, a_\tau = a\}.$$
3. For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ set the observed accumulated rewards and the transition counts prior to episode $k$,
$$R_k(s,a) := \sum_{\tau=1}^{t_k - 1} r_\tau \mathbb{1}_{s_\tau = s, a_\tau = a},$$
$$P_k(s,a,s') := \#\left\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\right\}.$$
Compute estimates $\hat{r}_k(s,a) := \frac{R_k(s,a)}{\max\{1, N_k(s,a)\}}$, $\hat{p}_k(s'|s,a) := \frac{P_k(s,a,s')}{\max\{1, N_k(s,a)\}}$.

**Compute policy** $\tilde{\pi}_k$:

4. Let $\mathcal{M}_k$ be the set of all MDPs with states and actions as in $M$, and with transition probabilities $\tilde{p}(\cdot|s,a)$ close to $\hat{p}_k(\cdot|s,a)$, and rewards $\tilde{r}(s,a) \in [0,1]$ close to $\hat{r}_k(s,a)$, that is,
$$\left|\tilde{r}(s,a) - \hat{r}_k(s,a)\right| \leq \sqrt{\tfrac{7\log(2SAt_k/\delta)}{2\max\{1, N_k(s,a)\}}} \quad \text{and} \tag{3}$$
$$\left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 \leq \sqrt{\tfrac{14S\log(2At_k/\delta)}{\max\{1, N_k(s,a)\}}}. \tag{4}$$
5. Use extended value iteration (see Section 3.1) to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that
$$\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} \rho(M', \pi, s') - \frac{1}{\sqrt{t_k}}.$$

**Execute policy** $\tilde{\pi}_k$:

6. **While** $v_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_k(s_t, \tilde{\pi}_k(s_t))\}$ **do**
   (a) Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward $r_t$, and observe next state $s_{t+1}$.
   (b) Update $v_k(s_t, a_t) := v_k(s_t, a_t) + 1$.
   (c) Set $t := t + 1$.

Figure 1: The UCRL2 algorithm.