```
Input: Set of arms $A$, exploration parameter $\epsilon$, action distribution $P(\cdot)$
Initialization: $R(a) \leftarrow 0$ for all $a \in A$;
Initialization: $N(a) \leftarrow 0$ for all $a \in A$;
for $t = 1$ to $T$ do
    if $random(0, 1) > \epsilon$ then
        Choose action $a_t = \arg\max_{s \in S} R(a)$;
    else
        Choose a random action $a_t \sim P(\cdot)$;
    end
    Perform action $a_t$ on the chosen arm;
    Observe reward $r_t$ obtained from the chosen action;
    Update action count: $N(a_t) \leftarrow N(a_t) + 1$;
    Update estimated reward value: $R(a_t) \leftarrow \hat{R}(a_t|r_t)$;
end
```