Algorithm 1 GPOMDP

Input:

- Initial policy parameters $\theta_0 = (\theta_0^1, \ldots, \theta_0^{D_\theta})^T$
- Learning rate $\{\alpha_k\}$
- Number of trajectories M

Output: Approximation of the optimal policy $\pi_{\theta^*} \approx \pi_*$

1: Initialize $k = 0$
2: **repeat**
3: Sample M trajectories $h^{(m)} = \{(s_t^{(m)}, a_t^{(m)}, r_{t+1}^{(m)}\}_{t=0}^{T^{(m)}}$ of the MDP under policy π_{θ_k}
4: Compute the optimal baseline

$$\widehat{b}_k^n = \frac{\sum_{m=1}^{M} \left[\sum_{i=0}^{T^{(m)}} \partial_{\theta_k} \log \pi_\theta \left(s_i^{(m)}, a_i^{(m)}\right)\right]^2 \sum_{j=0}^{T^{(m)}} \gamma^j r_{j+1}^{(m)}}{\sum_{m=1}^{M} \left[\sum_{i=0}^{T^{(m)}} \partial_{\theta_k} \log \pi_\theta \left(s_i^{(m)}, a_i^{(m)}\right)\right]^2}$$

5: Approximate policy gradient

$$\frac{\partial}{\partial \theta^n} J_{\text{start}}(\theta_k) \approx \widehat{g}_k^n = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=0}^{T^{(m)}} \frac{\partial}{\partial \theta^n} \log \pi_{\theta_k} \left(s_i^{(m)}, a_i^{(m)}\right) \left(\sum_{j=i}^{T^{(m)}} \gamma^j r_{j+1}^{(m)} - \widehat{b}_k^n\right)$$

6: Update actor parameters $\theta_{k+1} = \theta_k + \alpha_k \widehat{g}_k$.
7: $k \leftarrow k + 1$
8: **until** converged