Algorithm 1 On-policy policy gradient with Monte Carlo estimator

Initialize θ_0
for *iterationk* $\in [0, \ldots, K]$ **do**
    sample trajectories $\{\tau_i\}$ by running $\pi_{\theta_k}(\mathbf{a}_t|\mathbf{s}_t)$ $\triangleright$ each τ_i consists of $\mathbf{s}_{i,0}, \mathbf{a}_{i,0}, \ldots, \mathbf{s}_{i,H}, \mathbf{a}_{i,H}$
    compute $\mathcal{R}_{i,t} = \sum_{t'=t}^{H} \gamma^{t'-t} r(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$
    fit $b(\mathbf{s}_t)$ to $\{\mathcal{R}i, t\}$ $\triangleright$ use constant $b_t = \frac{1}{N} \sum_i \mathcal{R}i, t$, or fit $b(\mathbf{s}_t)$ to $\{\mathcal{R}i, t\}$
    compute $\hat{A}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) = \mathcal{R}_{i,t} - b(\mathbf{s}_t)$
    estimate $\nabla_{\theta_k} J(\pi_{\theta_k}) \approx \sum_{i,t} \nabla_{\theta_k} \log \pi_{\theta_k}(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \hat{A}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$
    update parameters: $\theta_{k+1} \leftarrow \theta_k + \alpha \nabla_{\theta_k} J(\pi_{\theta_k})$
end for