Algorithm 1: Quicksort

Input: θ_0
for Learning iteration $i := 0 \dots \mathcal{I}$ **do**
  Policy weight update
  $\theta_{i+1} \leftarrow \theta_i - \alpha_i \hat{\mathbb{E}}_{\rho^{\pi^{\theta_i}}} \left[C^{\gamma}_{0:T} \nabla_\theta \log \pi^{\theta_i}(Z_{0:T}) \right]$
  $\alpha_i > 0$, learning rate
end for

return Near-optimal policy $\pi^{\theta_{\mathcal{I}}}$