A Experimental Details

The state value distribution $\hat{Y}_\psi$ is parameterized following Will et al. [17]: $\hat{Y}_\psi(s;\tau) = F_\psi\big(E_\psi(s) \odot T_\psi(\tau)\big)$ where F_ψ and E_ψ is a multi-layer perceptron (MLP) that maps an input into 64 and 1-dimensional output, and T_ψ is a cosine-based embedding function that maps a 64-dimensional cosine basis vector $[\cos(\pi i\tau)]_{i=0}^{63}$ into the same length feature vector with a single fully-connected layer with ReLU activation. For $\hat{Q}_\theta$, we used an MLP that maps a concatenation of state and action input to a single scalar value. We used a deterministic policy π_ϕ that maps a state into an action-dimensional vector. For every MLP, we used 2 fully-connected layers, and we trained five $\hat{Y}_\psi$s and $\hat{Q}_\theta$s with a single π_θ. For the action samples used in Eq. 8, we added an action noise following Fujimoto et al. [2]. For the efficiency in computing $\tilde{\pi}$, we reduce the search space by first finding the 100 nearest states in raw-state space and querying the actions of those states. We use an approximated nearest neighbor algorithm called Annoy [36]. Both in training $\hat{Y}_\psi$ and $\hat{Q}_\theta$, we adapted an n-step TD trick instead of using TD(0); we sampled $(s_t, a_t, \sum_{t'=t}^{t+9}\gamma^{t-1}r_t, s_{t+10})$ from a dataset instead of sampling (s, a, r, s'). Also, we used a slowly moving target network in calculating the bootstrapped distribution by keeping an exponential moving average of ψ [4] and using the averaged weight for bootstrapping.

We provide the hyperparameters used for the experiments in A.1. We use the provided hyperparameters unless mentioned otherwise for the ablation experiments. Code is also available [1].

Table A.1: Hyperparameters used in the experiments

	$\hat{Y}_\psi$	$\hat{Q}_\theta$	π_ϕ
γ		0.99	
n-steps		10	
# Ensembles	5	5	1
Batch Size		100	
# Training Iterations		1 million steps	
Learning Rate	1e-4	1e-3	3e-4
weight-decay	0	1e-8 w/ AdamW	0
$\lvert F\rvert$		64	
E_ψ		$\lvert S\rvert \to 256 + \text{ReLU} \to 256 + \text{ReLU} \to \lvert F\rvert$	
T_ψ		$[\cos(\pi i\tau)]_{i=0}^{63} \to 64 + \text{ReLU} \to \lvert F\rvert$	
F_ψ		$\lvert F\rvert \to 256 + \text{ReLU} \to 256 + \text{ReLU} \to 1$	
N, N'		16	
κ		1	
$\hat{Q}_\theta$		$\lvert S\rvert + \lvert A\rvert \to 256 + \text{swish} \to 256 + \text{swish} \to 1$	
λ		1.0 (`-medium-expert`), 0.1 (otherwise)	
# policy samples		10	
τ_1		0.9	
τ_2		0.1	
π_ϕ noise σ		0.3	
π_ϕ noise clip		0.5	
π_ϕ		$\lvert S\rvert \to 256 + \text{swish} \to 256 + \text{swish} \to \lvert A\rvert$	

[1] https://github.com/hiwonjoon/YOEO-public