Supplementary Material

Sheryl Paul, Jyotirmoy V. Deshmukh

1 Notes for Implementation

1.1 Default hyperparameters

Below are the default hyperparameters for each of the baselines. Unless specified otherwise in the subsequent sections, these will be employed.

1. **A2C:**

```
{
  policy: 'MlpPolicy', n_envs: 8, learning_rate: 0.0005, n_steps: 5, gamma: 0.99,
  gae_lambda: 1.0, ent_coef: 0.0, vf_coef: 0.5, max_grad_norm: 0.5, rms_prop_eps: 1e-5,
  use_rms_prop: True, use_sde: False, sde_sample_freq: -1, normalize_advantage: False,
  stats_window_size: 100, verbose: 1, seed: None, _init_setup_model: True
}
```

2. **PPO:**

```
{
  policy: 'MlpPolicy', n_envs: 8, learning_rate: 0.0003, n_steps: 2048,
  batch_size: 64, n_epochs: 10, gamma: 0.99, gae_lambda: 0.95, clip_range: 0.2,
  clip_range_vf: None, normalize_advantage: True, ent_coef: 0.0, vf_coef: 0.5,
  max_grad_norm: 0.5, use_sde: False, sde_sample_freq: -1,
  target_kl: None, stats_window_size: 100, tensorboard_log: None,
  policy_kwargs: None, verbose: 1, seed: None, _init_setup_model: True
}
```

3. **DQN:**

```
{
  policy: 'MlpPolicy', n_envs: 8, learning_rate: 0.0001, buffer_size: 1000000,
  learning_starts: 1000, batch_size: 32, tau: 1.0, gamma: 0.99, train_freq: 4,
  gradient_steps: 1, replay_buffer_class: None, replay_buffer_kwargs: None,
  optimize_memory_usage: False, target_update_interval: 10000, exploration_fraction: 0.1,
  exploration_initial_eps: 1.0,
  exploration_final_eps: 0.01, max_grad_norm: 10,
  stats_window_size: 100, tensorboard_log: None,
  policy_kwargs: None, verbose: 1, seed: None, _init_setup_model: True
}
```

4. **PPO-B:** Same as the PPO settings used for the base model of each environment.
5. **A2C-B:** Same as the A2C settings used for the base model of each environment.
6. **DQN-B:** Same as the DQN settings used for the base model of each environment.
7. **PPO-DR:** Same as the PPO settings used for the base model of each environment.

1.2 Specific Environment Settings

- ***CliffWalking:***
  1. PPO: { `n_steps: 512, learning_rate: 0.00025`}
  2. DQN: { `exploration_fraction: 0.8` }
  3. A2C: { `n_steps: 100`}

4. ERPO: { number_of_episodes_per_batch: 50, w: 0.3, epsilon: 0.01, alpha: 2 [1], nu: 0.05 }

- ***DistributionShift:***

1. PPO: { n_steps: 512, learning_rate: 0.00025}
2. DQN: { exploration_fraction: 0.8 }
3. A2C: { n_steps: 500}
4. ERPO: { number_of_episodes_per_batch: 50, w: 0.3, epsilon: 0.01, alpha: 2, nu: 0.01 }

- ***Walls&Lava:***

1. PPO: { n_steps: 512, learning_rate: 0.00025}
2. DQN: { exploration_fraction: 0.7 }
3. A2C: { n_steps: 500}
4. ERPO: { number_of_episodes_per_batch: 30, w: 0.5, epsilon: 0.01, alpha: 2, nu: 0.01 }

- ***Taxi:***

1. PPO: { n_steps: 4096, learning_rate: 0.00025}
2. DQN: { exploration_fraction: 0.8 }
3. ERPO: { number_of_episodes_per_batch: 20, w: 0.5, epsilon: 0.01, alpha: 3, nu: 0.01 }

- ***FrozenLake:***

1. PPO: { n_steps: 512, learning_rate: 0.00025}
2. DQN: { exploration_fraction: 0.8 }
3. A2C: { n_steps: 250, learning_rate: 0.0005}
4. ERPO: { number_of_episodes_per_batch: 30, w: 0.2, epsilon: 0.01, alpha: 3, nu: 0.05 }

[1] We introduce a scaling factor in the policy update