# A. Appendix

We present two variants of DDPG with the proposed smoothness-inducing regularizer. The first algorithm, DDPG-SR-A, directly learns a smooth policy with a regularizer that measures the non-smoothness in the actor network (policy). The second variant, DDPG-SR-C, learns a smooth Q-function with a regularizer that measure the non-smoothness in the critic network (Q-function). We present the details of DDPG-SR-A and DDPG-SR-C in Algorithm 2 and Algorithm 3, respectively.

**Algorithm 2** DDPG with smoothness-inducing regularization on the actor network (DDPG-SR-A).

**Input**: step size for target networks $\alpha \in (0,1)$, coefficient of regularizer $\lambda_{\mathrm{s}}$, perturbation strength $\epsilon$, number of iterations to solve inner optimization problem $D$, number of training steps $T$, number of training episodes $M$, step size for inner maximization $\eta_\delta$, step size for updating actor/critic network $\eta$.

**Initialize**: randomly initialize the critic network $Q_\phi(s,a)$ and the actor network $\mu_\theta(s)$, initialize target networks $Q_{\phi'}(s,a)$ and $\mu_{\theta'}(s)$ with $\phi' = \phi$ and $\theta' = \theta$, initialize replay buffer $\mathcal{R}$.

**for** episode = 1 …, M **do**
  Initialize a random process $\epsilon$ for action exploration.
  Observe initial state $s_1$.
  **for** $t = 1 \ldots T$ **do**
    Select action $a_t = \mu_\theta(s_t) + \epsilon_t$ where $\epsilon_t$ is the exploration noise.
    Take action $a_t$, receive reward $r_t$ and observe the new state $s_{t+1}$.
    Store transition $(s_t, a_t, r_t, s_{t+1})$ into the replay buffer $\mathcal{R}$.
    Sample mini-batch $B$ of transitions $\{(s_t^i, a_t^i, r_t^i, s_{t+1}^i)\}_{i\in B}$ from the replay buffer $\mathcal{R}$.
    Set $y_t^i = r_t^i + \gamma Q_{\phi'}(s_{t+1}^i, \mu_{\theta'}(s_{t+1}^i))$ for $i \in B$.
    Update the critic network: $\phi \leftarrow \operatorname{argmin}_{\widetilde{\phi}} \sum_{i\in B}(y_t^i - Q_{\widetilde{\phi}}(s_t^i, a_t^i))^2$.
    **for** $s_t^i \in B$ **do**
      Randomly initialize $\delta_i$.
      **for** $\ell = 1 \ldots D$ **do**
        $\delta_i \leftarrow \delta_i + \eta_\delta \nabla_\delta \big\|\mu_\theta(s_t^i) - \mu_\theta(s_t^i + \delta_i)\big\|_2^2$.
        $\delta_i \leftarrow \Pi_{\mathbb{B}_d(0,\epsilon)}(\delta_i)$.
      **end for**
      Set $\widehat{s}_t^i = s_t^i + \delta_i$.
    **end for**
    Update the actor network:

$$\theta \leftarrow \theta + \frac{\eta}{|B|}\sum_{i\in B}\left(\nabla_a Q_\phi(s,a)\big|_{s=s_t^i, a=u_\theta(s_t^i)} \nabla_\theta \mu_\theta(s)\big|_{s=s_t^i} - \lambda_s \nabla_\theta \big\|\mu_\theta(s_t^i) - \mu_\theta(\widehat{s}_t^i)\big\|_2^2\right).$$

    Update the target networks:

$$\theta' \leftarrow \alpha\theta + (1-\alpha)\theta',$$
$$\phi' \leftarrow \alpha\phi + (1-\alpha)\phi'.$$

  **end for**
**end for**

**Algorithm 3** DDPG with smoothness-inducing regularization on the critic network (DDPG-SR-C).

**Input**: step size for target networks $\alpha \in (0, 1)$, coefficient of regularizer $\lambda_{\mathrm{s}}$, perturbation strength $\epsilon$, number of iterations to solve inner optimization problem $D$, number of training steps $T$, number of training episodes $M$, step size for inner maximization $\eta_\delta$, step size for updating actor/critic network $\eta$.

**Initialize**: randomly initialize the critic network $Q_\phi(s, a)$ and the actor network $\mu_\theta(s)$, initialize target networks $Q_{\phi'}(s, a)$ and $\mu_{\theta'}(s)$ with $\phi' = \phi$ and $\theta' = \theta$, initialize replay buffer $\mathcal{R}$.

**for** episode = 1 …, M **do**
  Initialize a random process $\epsilon$ for action exploration.
  Observe initial state $s_1$.
  **for** $t = 1 \ldots T$ **do**
    Select action $a_t = \mu_\theta(s_t) + \epsilon_t$ where $\epsilon_t$ is the exploration noise.
    Take action $a_t$, receive reward $r_t$ and observe the new state $s_{t+1}$.
    Store transition $(s_t, a_t, r_t, s_{t+1})$ into replay buffer $\mathcal{R}$.
    Sample mini-batch $B$ of transitions $\{(s_t^i, a_t^i, r_t^i, s_{t+1}^i)\}_{i \in B}$ from the replay buffer $\mathcal{R}$.
    Set $y_t^i = r_t^i + \gamma Q_{\phi'}(s_{t+1}^i, \mu_{\theta'}(s_{t+1}^i))$ for $i \in B$.
    **for** $s_t^i \in B$ **do**
      Randomly initialize $\delta_i$.
      **for** $\ell = 1 \ldots D$ **do**
        $\delta_i \leftarrow \delta_i + \eta_\delta \nabla_\delta (Q_\phi(s_t^i, a_t^i) - Q_\phi(s_t^i + \delta, a_t^i))^2$.
        $\delta_i \leftarrow \Pi_{\mathbb{B}_d(0,\epsilon)}(\delta_i)$.
      **end for**
      Set $\widehat{s}_t^i = s_t^i + \delta_i$.
    **end for**
    Update the critic network:

$$\phi \leftarrow \operatorname*{argmin}_{\widetilde{\phi}} \sum_{i \in B} (y_t^i - Q_{\widetilde{\phi}}(s_t^i, a_t^i))^2 + \lambda_{\mathrm{s}} \sum_{i \in B} (Q_\phi(s_t^i, a_t^i) - Q_\phi(\widehat{s}_t^i, a_t^i))^2.$$

    Update the actor network:

$$\theta \leftarrow \theta + \frac{\eta}{|B|} \sum_{i \in B} \nabla_a Q_\phi(s, a)\big|_{s = s_t^i, a = u_\theta(s_t^i)} \nabla_\theta \mu_\theta(s)\big|_{s = s_t^i}.$$

    Update the target networks:

$$\theta' \leftarrow \alpha\theta + (1 - \alpha)\theta',$$
$$\phi' \leftarrow \alpha\phi + (1 - \alpha)\phi'.$$

  **end for**
**end for**