A BACKGROUNDS

Deep Deterministic Policy Gradient Incorporating a parameterized actor function $\mu_\theta(s)$, Deep Deterministic Policy Gradient uses the following actor and critic and loss to train the agent:

$$\begin{aligned} J_\pi(\theta) &= \hat{\mathbb{E}}_{s_t\sim\mathcal{D}}[-Q_\phi(s_t,a_t)|_{a_t=\mu_\theta(s_t)}], \\ J_Q(\phi) &= \hat{\mathbb{E}}_{s_t\sim\mathcal{D}}[(Q_\phi(s_t,a_t)-\hat{Q}(s_t,a_t))^2|_{a_t=\mu_\theta(s_t)}], \end{aligned} \tag{16}$$

where $\hat{Q}(s_t,a_t) = r_t + \gamma Q_{\bar{\phi}}(s_{t+1},\mu_{\bar{\theta}}(s_{t+1}))$, which is the target Q-value defined from a target network, θ and ϕ represents the parameters of the actor and the critic respectively, $\bar{\theta}$ and $\bar{\phi}$ represents the parameters of the target actor and the target critic respectively, and $\mathcal{D}$ represents the replay buffer. The weights of a target network are the exponentially moving average of the online network's weights.

Soft Actor-Critic Maximum entropy RL tackles an RL problem with an alternative objective function, which favors more random policies: $J = \hat{\mathbb{E}}_\pi[\sum_{t=0}^{\infty}\gamma^t r_t + \alpha H(\pi(\cdot \mid s_t))]$, where γ is the discount factor, α is a trainable coefficient of the entropy term and $H(\pi(\cdot \mid s_t))$ is the entropy of action distribution $\pi(\cdot \mid s_t)$. The Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018) optimizes it by training the actor π_θ and critic Q_ϕ with the following respective losses:

$$\begin{aligned} J_\pi(\theta) &= \hat{\mathbb{E}}_{s_t\sim\mathcal{D},a\sim\pi}[\alpha\log\pi_\theta(a \mid s_t) - Q_\phi(s_t,a)], \\ J_Q(\phi) &= \hat{\mathbb{E}}_{s_t,a_t\sim\mathcal{D}}[(Q_\phi(s_t,a_t)-\hat{Q}(s_t,a_t))^2], \end{aligned} \tag{17}$$

where $\hat{Q}(s_t,a_t) = r_t + \gamma Q_{\bar{\phi}}(s_{t+1},a_{t+1}) - \alpha\log\pi_\theta(a_{t+1}|s_{t+1})$, which is the target Q-value defined from a target network and $a_{t+1}\sim\pi_\theta(\cdot \mid s_{t+1})$, θ ,ϕ and $\bar{\phi}$ represents the parameters of the actor, the critic and the target critic respectively, and $\mathcal{D}$ represents the replay buffer. The weights of a target network are the exponentially moving average of the online network's weights.

Reinforcement Learning with Augmented Data Reinforcement Learning with Augmented Data (RAD) (Laskin et al., 2020) applies data augmentation in SAC by replacing the original observation with augmented observations in the training of the actor and critic. Given image transformation f_ν, the actor and critic losses are

$$\begin{aligned} J_\pi(\theta) &= \hat{\mathbb{E}}_{s_t\sim\mathcal{D},a\sim\pi}[\alpha\log\pi_\theta(a \mid f_\nu(s_t)) - Q_\phi(f_\nu(s_t),a)], \\ J_Q(\phi) &= \hat{\mathbb{E}}_{s_t,a_t\sim\mathcal{D}}[(Q_\phi(f_\nu(s_t),a_t)-\hat{Q}(f_\nu(s_t),a_t))^2], \end{aligned} \tag{18}$$

Data-Regularized Q Data-regularized Q (DrQ) (Kostrikov et al., 2020) extends RAD by using data augmentation in the training of the critic in two new ways. Given a type of image transformation f parameterized by ν, data augmentation is first applied in the calculation of the target Q-value for every transition (s,a,r,s'):

$$y = r + \gamma\frac{1}{K}\sum_{k=1}^{K} Q_{\bar{\phi}}(f_{\tau_k}(s'),a'), \text{ where } a'\sim\pi(\cdot \mid f_{\tau_k}(s')). \tag{19}$$

$Q_{\bar{\phi}}$ is the slowly updated target network. Then the critic is updated with different augmented s and this averaged target:

$$\ell_Q(\phi) = \frac{1}{NM}\sum_{i=1}^{N}\sum_{m=1}^{M}\Big(Q_\phi(f_{\tau_m}(s),a) - y\Big)^2. \tag{20}$$

Note that DrQ recovers RAD when $M=1$ and $K=1$.

SVEA In order to avoid non-deterministic Q-target and over-regularization, Hansen et al. (2021) propose using state without complex augmentation for calculating the target. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be a set of random shift and a set of random shift plus one of the data augmentation mentioned in the paper such as random convolution (Lee et al., 2019). The critic loss used for training is

$$L_Q(\phi) = \frac{1}{N}\sum_{i=1}^{N}\alpha_{\text{svea}}(Q_\phi(f_{\tau_{1,i}}(s),a)-y_i)^2 + \beta_{\text{svea}}(Q_\phi(f_{\tau_{2,i}}(s),a)-y_i)^2, \tag{21}$$

where α_{svea} and β_{svea} are constant coefficients for naively and complexly augmented data respectively, $\tau_{1,i} \in \mathcal{T}_1$ and $\tau_{2,i} \in \mathcal{T}_2$, $y_i = r + \gamma Q_{\bar{\phi}}(\tau_{1,i}(s'), a') + \alpha \log \pi(a' \mid f_{\tau_{1,i}}(s'))$, where $a' \sim \pi(\cdot \mid \tau_{1,i}(s'))$.

DrAC Instead of directly replacing the original samples with augmented samples in the training, Raileanu et al. (2021) use two regularization terms in the training of the actor and critic to explicitly enforce the invariance. When applying it in the PPO algorithm (Schulman et al., 2017) to learn a state-value estimator $V_\phi(s)$ and a policy $\pi_\theta(s)$, the regularization terms are

$$\begin{aligned} G_V &= (\hat{V}(s) - V_\phi(f_\nu(s)))^2, \\ G_\pi &= D_{KL}[\pi_\theta(a \mid s) \mid \pi_\theta(a \mid f_\nu(s)]. \end{aligned} \tag{22}$$

where $\hat{V}(s)$ is the sum of rewards collected by the agent after state s and ν is the random variable for parameterizing the image transformation.

Tangent Prop Regularization Tangent prop (Simard et al., 1991) is a regularization term used for learning invariance for a function $G(s)$ with respect to a small image transformation parameterized by α on s:

$$\sum \left\| \frac{\partial G(s, \alpha)}{\partial \alpha} \right\|^2 = 0 \tag{23}$$

B Expected loss under data augmentation

B.1 Actor loss

SAC as base algorithm For image-based control tasks, a data augmentation f parameterized by μ over $\mathcal{T}$ is applied on the observations. The actor loss with implicit regularization for the state s in a transition is

$$\begin{aligned} \ell_\theta^I(s, \mu) &= \hat{\mathbb{E}}_\mu \Big[\alpha \log \pi_\theta(\hat{a} \mid f_\mu(s)) - Q_\phi(f_\mu(s), \hat{a}) \mid_{\hat{a} \sim \pi_\theta(\cdot \mid f_\mu(s))} \Big] \\ &= \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| \exp(\frac{1}{\alpha} Q_\phi(f_\mu(s), \cdot) - \log Z(f_\mu(s))) \Big) \Big] \end{aligned} \tag{24}$$

Let $g(f_\mu(s), \cdot) = \exp(\frac{1}{\alpha} Q_\phi(f_\mu(s), \cdot) - \log Z(f_\mu(s)))$.

$$\begin{aligned} &\ell_\theta^I(s, \mu) \\ &= \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(f_\mu(s), \cdot) \Big) \Big] \\ &= \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(f_\mu(s), \cdot)) \Big) - D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(s, \cdot) \Big) \\ &+ D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(s, \cdot) \Big) \Big] \\ &= \hat{\mathbb{E}}_\mu \Big[\int_a \pi_\theta(a \mid f_\mu(s)) \log \frac{\pi_\theta(a \mid f_\mu(s))}{g(a \mid f_\mu(s))} - \int_a \pi_\theta(a \mid f_\mu(s)) \log \frac{\pi_\theta(a \mid f_\mu(s))}{g(a \mid s)} \Big] \\ &+ \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(s, \cdot) \Big) \Big] \\ &= \hat{\mathbb{E}}_\mu \Big[\int_a \pi_\theta(a \mid f_\mu(s)) \log \frac{g(a \mid s)}{g(a \mid f_\mu(s))} \Big] + \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(s, \cdot) \Big) \Big] \\ &= \hat{\mathbb{E}}_\mu \Big[\int_a \pi_\theta(a \mid f_\mu(s)) \log \frac{g(a \mid s)}{g(a \mid f_\mu(s))} \Big] + \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| g(s, \cdot) \Big) \\ &- D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| \pi_{\theta, sg}(\cdot \mid s) \Big) + D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| \pi_{\theta, sg}(\cdot \mid s) \Big) \Big] \\ &= \hat{\mathbb{E}}_\mu \Big[\int_a \pi_\theta(a \mid f_\mu(s)) \log \frac{g(a \mid s)}{g(a \mid f_\mu(s))} \Big] + \hat{\mathbb{E}}_\mu \Big[\int_a \pi_\theta(a \mid f_\mu(s)) \log \frac{\pi_{\theta, sg}(a \mid s)}{g(a \mid s)} \Big] \\ &+ \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot \mid f_\mu(s)) \| \pi_{\theta, sg}(\cdot \mid s) \Big) \Big]. \end{aligned} \tag{25}$$

If the invariance in the critic has been learned:

$$Q(f_\mu(s), a) = Q(s, a) \text{ for all } a \in \mathcal{A}, \tag{26}$$

the actor loss with implicit regularization becomes

$$\ell_\theta^I(s, \mu) = \hat{\mathbb{E}}_\mu \Big[\int_a \pi_\theta(a|f_\mu(s)) \log \frac{\pi_{\theta,sg}(a|s)}{g(a|s)} \Big] + \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot|f_\mu(s)) || \pi_{\theta,sg}(\cdot|s) \Big) \Big], \tag{27}$$

because

$$g(a|f_\mu(s)) = g(a|s) \text{ for all } a \in \mathcal{A}. \tag{28}$$

If the actor is well learned for state s, this actor loss become

$$\ell_\theta^I(s, \mu) = \hat{\mathbb{E}}_\mu \Big[D_{KL} \Big(\pi_\theta(\cdot|f_\mu(s)) || \pi_{\theta,sg}(\cdot|s) \Big) \Big], \tag{29}$$

because

$$g(a|s) = \pi_{\theta,sg}(a|s) \text{ for all } a \in \mathcal{A}. \tag{30}$$

DDPG as base algorithm For image-based control tasks, a data augmentation f parameterized by μ over $\mathcal{T}$ is applied on the observations. Considering that the Q-invariant transformation is also π^*-invariant, training all policies of the transformed states to get close to the same optimal policy is equivalent to training the policy of original state and enforce the invariance in the policy. Considering actor loss with implicit regularization, we can apply a Taylor expansion with respect to the optimal action $\pi^*(f_\mu(s)) = \pi^*(s) = \arg\max_a Q_\phi(s, a)$:

$$\begin{aligned}
&\ell_\theta^I(s, \mu) \\
&= \hat{\mathbb{E}}_\mu \Big[- Q_\phi(f_\mu(s), \hat{a}) \mid_{\hat{a}=\pi_\theta(f_\mu(s))} \Big] \\
&= -\hat{\mathbb{E}}_\mu \Big[Q_\phi(f_\mu(s), \pi^*(f_\mu(s)) + J(\hat{a} - \pi^*(f_\mu(s))) \\
&+ \frac{1}{2}(\hat{a} - \pi^*(f_\mu(s)))^T H(\hat{a} - \pi^*(f_\mu(s))) + o(\|\hat{a} - \pi^*(f_\mu(s))\|^2)|_{\hat{a}=\pi_\theta(f_\mu(s))} \Big] \\
&\approx -\frac{1}{2}\hat{\mathbb{E}}_\mu \Big[(\hat{a} - \pi^*(f_\mu(s)))^T H(\hat{a} - \pi^*(f_\mu(s)))|_{\hat{a}=\pi_\theta(f_\mu(s))} \Big] \\
&= -\frac{1}{2}\hat{\mathbb{E}}_\mu \Big[(\hat{a} - \pi_{\theta,sg}(s) + \pi_{\theta,sg}(s) - \pi^*(f_\mu(s)))^T H \\
&(\hat{a} - \pi_{\theta,sg}(s) + \pi_{\theta,sg}(s) - \pi^*(f_\mu(s)))|_{\hat{a}=\pi_\theta(f_\mu(s))} \Big] \\
&= -\frac{1}{2}\hat{\mathbb{E}}_\mu \Big[(\hat{a} - \pi_{\theta,sg}(s))^T H(\hat{a} - \pi_{\theta,sg}(s)) \\
&+ (\pi_{\theta,sg}(s) - \pi^*(f_\mu(s)))^T H(\pi_{\theta,sg}(s) - \pi^*(f_\mu(s))) \\
&+ 2(\hat{a} - \pi_{\theta,sg}(s))^T H(\pi_{\theta,sg}(s) - \pi^*(f_\mu(s)))|_{\hat{a}=\pi_\theta(f_\mu(s))} \Big].
\end{aligned} \tag{31}$$

The first term above is enforcing the invariance of the actor with respect to the transformation.

B.2 Critic loss

B.2.1 Linear Model

According to the analysis by Balestriero et al. (2022), the expected Mean Squared Error (MSE) under data augmentation for a linear regression model can be expressed by the expectation and variance of the transformed images. Now we want to derive a similar regularization term from the critic loss.

If we use linear model for the critic and actor:

$$Q(s, a) = \mathcal{W}_s * s + \mathcal{W}_a * a + b_0 \tag{32}$$

$$\bar{Q}(s, a) = \bar{\mathcal{W}}_s * s + \bar{\mathcal{W}}_a * a + \bar{b}_0 \tag{33}$$

$$\pi(s) = W_\tau * s + \epsilon W_\sigma * s + b_1, \epsilon \sim \mathcal{N}(0,1) \tag{34}$$

in which $\mathcal{W}_s \in \mathcal{R}^{1*|\mathcal{S}|}$, $\mathcal{W}_a \in \mathcal{R}^{1*|\mathcal{A}|}$, $\mathcal{W}_\tau \in \mathcal{R}^{1*|\mathcal{S}|}$, $\mathcal{W}_\sigma \in \mathcal{R}^{1*|\mathcal{S}|}$ and b_0, b_1 are parameters for the model. $\bar{Q}$ is the exponential moving average of Q.

The critic loss for a transition (s, a, r, s') under data augmentation $\nu \sim \mathcal{P}$ and $\mu \sim \mathcal{P}'$ for state s and next state s' is

$$\begin{aligned} \ell_\phi &= \mathbb{E}_\nu\Big[\Big(Q(f_\nu(s), a) - \hat{\mathbb{E}}_\mu[y]\Big)^2\Big] \\ &= \mathbb{E}_\nu\big[Q(f_\nu(s), a)^2\big] - 2\mathbb{E}_\nu\big[Q(f_\nu(s), a)\big]\hat{\mathbb{E}}_\mu[y] + \hat{\mathbb{E}}_\mu[y]^2) \end{aligned} \tag{35}$$

in which

$$y = r + \gamma\bar{Q}(f_\mu(s'), a') - \alpha \log \pi(a'|f_\mu(s'))|_{a' \sim \pi(\cdot|f_\mu(s'))}. \tag{36}$$

Considering the last term is not used to update Q, we only need to focus on the first two terms.

Expectation

$$\begin{aligned} \mathbb{E}_\nu\big[Q(f_\nu(s), a)\big] &= \mathbb{E}_\nu[\mathcal{W}_s f_\nu(s) + \mathcal{W}_a a + b_0] \\ &= \mathcal{W}_s\mathbb{E}_\nu[f_\nu(s)] + \mathcal{W}_a a + b_0 \\ &= Q(\mathbb{E}_\nu[f_\nu(s)], a) \end{aligned} \tag{37}$$

Variance

$$\begin{aligned} &\mathbb{E}_\nu\big[Q(f_\nu(s), a)^2\big] \\ &= \mathbb{E}_\nu\Big[(\mathcal{W}_s f_\nu(s) + \mathcal{W}_a a + b_0)^2\Big] \\ &= \mathbb{E}_\nu\Big[f^T(s,\nu)\mathcal{W}_s^T\mathcal{W}_s f_\nu(s) + (\mathcal{W}_a a + b_0)^2 + 2(\mathcal{W}_a a + b_0)\big(\mathcal{W}_s f_\nu(s)\big)\Big] \\ &= \mathbb{E}_\nu\Big[Tr\Big(\mathcal{W}_s^T\mathcal{W}_s f_\nu(s) f^T(s,\nu)\Big)\Big] + (\mathcal{W}_a a + b_0)^2 + 2(\mathcal{W}_a a + b_0)\big(\mathcal{W}_s\mathbb{E}_\nu[f_\nu(s)]\big) \\ &= Tr\Big(\mathcal{W}_s^T\mathcal{W}_s\mathbb{E}_\nu\big[f_\nu(s) f^T(s,\nu)\big]\Big) + (\mathcal{W}_a a + b_0)^2 + 2(\mathcal{W}_a a + b_0)\big(\mathcal{W}_s\mathbb{E}_\nu[f_\nu(s)]\big) \\ &= Tr\Big(\mathcal{W}_s^T\mathcal{W}_s\big(\mathbb{E}_\nu\big[f_\nu(s) f^T(s,\nu)\big] - \mathbb{E}_\nu\big[f_\nu(s)\big]\mathbb{E}_\nu\big[f^T(s,\nu)\big]\big)\Big) \\ &+ Tr\Big(\mathcal{W}_s^T\mathcal{W}_s\mathbb{E}_\nu\big[f_\nu(s)\big]\mathbb{E}_\nu\big[f^T(s,\nu)\big]\Big) + (\mathcal{W}_a a + b_0)^2 + 2(\mathcal{W}_a a + b_0)\big(\mathcal{W}_s\mathbb{E}_\nu[f_\nu(s)]\big) \\ &= Tr\Big(\mathcal{W}_s^T\mathcal{W}_s\mathbb{V}_\nu[f_\nu(s)]\Big) + Q(\mathbb{E}_\nu[f_\nu(s)], a)^2 \end{aligned} \tag{38}$$

Whole loss

$$\begin{aligned} \ell_\phi &= \sum_{i=1}^{N} \mathbb{E}_\nu\big[Q(f_\nu(s), a)^2\big] - 2\mathbb{E}_\nu\big[Q(f_\nu(s), a)\big]\hat{\mathbb{E}}_\mu[y] + \hat{\mathbb{E}}_\mu[y]^2 \\ &= \sum_{i=1}^{N} Tr\Big(\mathcal{W}_s^T\mathcal{W}_s\mathbb{V}_\nu[f_\nu(s)]\Big) + Q(\mathbb{E}_\nu[f_\nu(s)], a)^2 - 2\hat{\mathbb{E}}_\mu[y]Q(\mathbb{E}_\nu[T_\nu(s)], a) + \hat{\mathbb{E}}_\mu[y]^2 \\ &= \sum_{i=1}^{N} \Big(Q\big(\mathbb{E}_\nu[f_\nu(s)], a\big) - \hat{\mathbb{E}}_\mu[y]\Big)^2 + Tr\Big(\mathcal{W}_a^T\mathcal{W}_a\mathbb{V}_\nu[f_\nu(s)]\Big) \end{aligned} \tag{39}$$

B.2.2 NON-LINEAR MODEL

According to the analysis by Balestriero et al. (2022), the expected loss of transformed state has an upper bound related to the variance of the transformed state:

$$\mathbb{E}[(\ell \circ Q)(f(x))] \leq (\ell \circ Q)(\mathbb{E}[f(x)]) + \kappa(x)\|\mathcal{J}Q(\mathbb{E}[f(x)])H(x)\Lambda(x)^{\frac{1}{2}}\|_F^2, \tag{40}$$

in which variance of the transformed image can be decomposed into

$$\mathbb{V}[f(x)] = H(x)\Lambda(x)H(x)^T. \tag{41}$$

The second term in the RHS of Equation 40 recovers tangent prop regularization.

C EXPLICIT VS IMPLICIT REGULARIZATION

C.1 CRITIC LOSS IN EXPLICIT REGULARIZATION

For $\tau = (s, a, s', r)$ sampled from the replay buffer $\mathcal{D}$, given current estimation Q_ϕ and true estimation Q^* without error, the bias of the target $\mathbb{E}_{a' \sim \pi(s')} y(s', a')$ is smaller than the target $Q_{\phi,sg}(s, a)$:

$$
\begin{aligned}
&\mathbb{E}_{\tau\sim\mathcal{D}}\Big[(\mathbb{E}_{a'\sim\pi(\cdot|s')}[y(s',a')] - Q^*(s,a))^2\Big] \\
=&\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\Big(\mathbb{E}_{a'\sim\pi(\cdot|s')}[r + \gamma Q_{\bar{\phi}}(s',a') - \alpha\log\pi(a'|s')] \\
&- (r + \gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q^*(s',a') - \alpha\log\pi(a'|s')]\Big)^2\Big] \\
=&\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\Big(\gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q_{\bar{\phi}}(s',a') - Q^*(s',a')]\Big)^2\Big] \\
\approx&\gamma^2\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\Big(\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q_{\phi,sg}(s',a') - Q^*(s',a')]\Big)\Big] \\
<&\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\Big(\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q_{\phi,sg}(s',a') - Q^*(s',a')]\Big)^2\Big] \\
<&\mathbb{E}_{\tau\sim\mathcal{D},a'\sim\pi(\cdot|s')}\Big[\Big(Q_{\phi,sg}(s',a') - Q^*(s',a')\Big)^2\Big] \\
=&\mathbb{E}_{\tau\sim\mathcal{D}}\Big[\Big(Q_{\phi,sg}(s,a) - Q^*(s,a)\Big)^2\Big]
\end{aligned}
\tag{42}
$$

The bias of using a target $\bar{y}$ in the explicit regularization is

$$
\begin{aligned}
\epsilon(\bar{y}) =&\mathbb{E}_\tau\Big[\Big(\Big(Q_\phi(f_\nu(s),a) - \bar{y}\Big)^2 - \Big(Q_\phi(f_\nu(s),a) - Q^*(s,a)\Big)^2\Big)^2\Big] \\
=&\mathbb{E}_\tau\Big[\Big(2Q_\phi(f_\nu(s),a)(Q^*(s,a) - \bar{y}) + \bar{y}^2 - Q^*(s,a)^2\Big)^2\Big].
\end{aligned}
\tag{43}
$$

We only need to consider the first term $2Q_\phi(f_\nu(s), a)(Q^* - \bar{y})$, considering that other terms is constant with respect to ϕ. So the bias of using different targets in the regularization term is decided by the bias of the target compared to the true estimation. The bias of using $\mathbb{E}_{a'\sim\pi(s')}[y(s', a')]$ in the explicit regularization term is smaller than using $Q_{\phi,sg}(s, a)$ according to the equations above:

$$
\epsilon(\bar{y} = \mathbb{E}_{a'\sim\pi(\cdot|s')}[y(s',a')]) < \epsilon(\bar{y} = Q_{\phi,sg}(s,a)),
\tag{44}
$$

In practice, we use the sampled value $y(s', a')$ as the target, which leads to a smaller bias and relatively larger variance.

C.2 CRITIC LOSS CONNECTION

Assume given $\ell^E_\phi(s, a, r, s', \nu)$, by appropriately setting the random variables in Equations 5, it recovers the critic loss in explicit regularization (Equation 7), as shown below. If the distributions of $\hat{\nu}$ and $\hat{\mu}$ are defined as follows:

$$
\mathbb{P}(\hat{\nu} = \tau) = \begin{cases} \frac{\mathbb{P}(\nu=\tau)\alpha_Q+1}{\alpha_Q+1}, & \text{if } \tau = \tau_0 \\ \frac{\mathbb{P}(\nu=\tau)\alpha_Q}{\alpha_Q+1}, & \text{if } \tau \neq \tau_0 \end{cases} \quad \mathbb{P}(\hat{\mu}) = \begin{cases} 1, & \text{if } \hat{\mu} = \tau_0 \\ 0, & \text{if } \hat{\mu} \neq \tau_0 \end{cases},
\tag{45}
$$

then we have for any sample (s, a, r, s'):

$$
(\alpha_Q + 1)\ell^I_\phi(s, a, r, s', \hat{\nu}, \hat{\mu}) = \ell^E_\phi(s, a, r, s', \nu)
$$

C.3 ACTOR LOSS

Considering that the policy is parameterized as normal distribution in SAC, we first define:

$$
\pi_{\theta,sg}(\cdot \mid f_\eta(s)) = \mathcal{N}(\lambda_\eta, \sigma^2_\eta), \pi_\theta(\cdot \mid f_\mu(s)) = \mathcal{N}(\lambda_\mu, \sigma^2_\mu)
\tag{46}
$$

For simplicity, we consider μ and η are defined over discrete set $\mathcal{T}$ with probability $P(\mu = \tau_i) = P(\eta = \tau_i) = P_i, \tau_i \in \mathcal{T}$. The derivation can be easily extended to using a continuous set.

$$\pi_{avg}(\cdot \mid s) = \hat{\mathbb{E}}_\eta[\pi_{\theta,sg}(\cdot \mid f_\eta(s))] = \mathcal{N}(\lambda_{avg}, \sigma^2_{avg}) = \mathcal{N}(\sum_i P_i \lambda_{\tau_i}, \sum_i P_i^2 \sigma^2_{\tau_i}) \tag{47}$$

$$\begin{aligned} &\hat{\mathbb{E}}_\eta\big[D_{KL}(\pi_{\theta,sg}(\cdot \,|\, f_\eta(s)) \,\|\, \pi_\theta(\cdot \,|\, f_\mu(s))\big] \\ =&\hat{\mathbb{E}}_\eta\big[\log \frac{\sigma_\mu}{\sigma_\eta} + \frac{\sigma^2_\eta}{2\sigma^2_\mu} + \frac{(\lambda_\eta - \lambda_\mu)^2}{2\sigma^2_\mu} - \frac{1}{2}\big] \\ =&\log \sigma_\mu - \sum_i P_i \log \sigma_{\tau_i} + \frac{\sum_i P_i \sigma^2_{\tau_i}}{2\sigma^2_\mu} + \frac{\sum_i P_i (\lambda_{\tau_i} - \lambda_\mu)^2}{2\sigma^2_\mu} - \frac{1}{2} \end{aligned} \tag{48}$$

$$\begin{aligned} &D_{KL}(\pi_{avg}(\cdot \,|\, s) \,\|\, \pi_\theta(\cdot \,|\, f_\mu(s)) \\ =&\log \frac{\sigma_\mu}{\sigma_{avg}} + \frac{\sigma^2_{avg}}{2\sigma^2_\mu} + \frac{(\lambda_{avg} - \lambda_\mu)^2}{2\sigma^2_\mu} - \frac{1}{2} \\ =&\log \sigma_\mu - \frac{1}{2} \log \sum_i P_i^2 \sigma^2_{\tau_i} + \frac{\sum_i P_i^2 \sigma^2_{\tau_i}}{2\sigma^2_\mu} + \frac{(\sum_i P_i \lambda_{\tau_i} - \lambda_\mu)^2}{2\sigma^2_\mu} - \frac{1}{2} \end{aligned} \tag{49}$$

Comparing the two equations above, the first term and the last term are the same, and the second term is a constant with respect to the parameter θ of the actor. For the third term, it is obvious that $\frac{\sum_i P_i \sigma^2_{\tau_i}}{2\sigma^2_\mu} \geq \frac{\sum_i P_i^2 \sigma^2_{\tau_i}}{2\sigma^2_\mu}$ because $P_i \geq P_i^2$, for any i. For the forth term, we have:

$$\begin{aligned} &\frac{\sum_i P_i (\lambda_{\tau_i} - \lambda_\mu)^2}{2\sigma^2_\mu} - \frac{(\sum_i P_i \lambda_{\tau_i} - \lambda_\mu)^2}{2\sigma^2_\mu} \\ =& \frac{\sum_i P_i (\lambda^2_{\tau_i} + \lambda^2_\mu - 2\lambda_{\tau_i}\lambda_\mu)}{2\sigma^2_\mu} - \frac{\lambda^2_\mu + (\sum_i P_i \lambda_{\tau_i})^2 - 2\lambda_\mu \sum_i P_i \lambda_{\tau_i}}{2\sigma^2_\mu} \\ =& \frac{\lambda^2_\mu + \sum_i P_i \lambda^2_{\tau_i} - 2\lambda_\mu \sum_i P_i \lambda_{\tau_i}}{2\sigma^2_\mu} - \frac{\lambda^2_\mu + (\sum_i P_i \lambda_{\tau_i})^2 - 2\lambda_\mu \sum_i P_i \lambda_{\tau_i}}{2\sigma^2_\mu} \\ =& \frac{\sum_i P_i \lambda^2_{\tau_i} - (\sum_i P_i \lambda_{\tau_i})^2}{2\sigma^2_\mu} = \frac{\mathbb{V}[\lambda_\eta]}{2\sigma^2_\mu} \geq 0 \end{aligned} \tag{50}$$

So the loss of using the policy of a transformed state as the target is an upper bound of using the average policy as the target:

$$\hat{\mathbb{E}}_\eta\big[D_{KL}(\pi_{\theta,sg}(\cdot \,|\, f_\eta(s)) \,\|\, \pi_\theta(\cdot \,|\, f_\mu(s))\big] \geq D_{KL}(\pi_{avg}(\cdot \,|\, s) \,\|\, \pi_\theta(\cdot \,|\, f_\mu(s)) \tag{51}$$

D KL DIVERGENCE

Given a transformation $f_\nu(s)$ on state s, considering that the KL divergence is not symmetric, we discuss the differences between two kinds of KL regularization here:

$$D_{KL}\big(\pi_{\theta,sg}(s))||\pi_\theta(f_\nu(s))\big) \text{ and } D_{KL}\big(\pi_\theta(f_\nu(s))||\pi_{\theta,sg}(s)\big), \tag{52}$$

Detach First

$$\begin{aligned} D_{KL}\big(\pi_{\theta,sg}(s))||\pi_\theta(f_\nu(s))\big) &= \int_a \pi_{\theta,sg}(a|s) \log \frac{\pi_{\theta,sg}(a|s)}{\pi_\theta(a|f_\nu(s))} \\ &= \int_a \Big(\pi_{\theta,sg}(a|s) \log \pi_{\theta,sg}(a|s) - \pi_{\theta,sg}(a|s) \log \pi_\theta(a|f_\nu(s))\Big) \\ &= -H(\pi_{\theta,sg}(s)) - \int_a \pi_{\theta,sg}(a|s) \log \pi_\theta(a|f_\nu(s)) \end{aligned} \tag{53}$$

Detach Second

$$
\begin{aligned}
D_{KL}\big(\pi_\theta(f_\nu(s))||\pi_{\theta,sg}(s)\big) &= \int_a \pi_\theta(a|f_\nu(s)) \log \frac{\pi_\theta(a|f_\nu(s))}{\pi_{\theta,sg}(a|s)}, \\
&= \int_a \Big(\pi_\theta(a|f_\nu(s)) \log \pi_\theta(a|f_\nu(s)) - \pi_\theta(a|f_\nu(s)) \log \pi_{\theta,sg}(a|s)\Big) \\
&= -H(\pi_\theta(f_\nu(s))) - \int_a \pi_\theta(a|f_\nu(s)) \log \pi_{\theta,sg}(a|s),
\end{aligned} \tag{54}
$$

in which H represent the entropy for a distribution.

"Detach second" introduces an entropy term for the policy of the transformed state. This regularization not only makes the policy of the augmented state and the original state close, but also maximizes the entropy of the policy of the transformed state. However, in "detach first", the entropy term with the sign sg is not used to update the policy.

E VARIANCE UNDER DATA AUGMENTATION

E.1 MORE AUGMENTED SAMPLES REDUCE THE VARIANCE OF THE CRITIC LOSS

Considering one transition (s, a, r, s') and M transformations $\{f_{\tau_m} \mid m = 1, ..., M\}$, K transformation $\{(f_{\tau'_k} \mid k = 1, ..., K\}$ respectively on s and s', the Q-values and target Q-values for the transformed samples are

$$
\begin{aligned}
Q_m &= Q(f_{\tau_m}(s), a), \\
y_k &= r + \gamma Q_{\bar{\phi}}(f_{\tau'_k}(s'), a') - \alpha \log \pi(a'|f_{\tau'_k}(s'))|_{a' \sim \pi(\cdot|f_{\tau'_k}(s'))},
\end{aligned} \tag{55}
$$

where $m \in \{1, ..., M\}, k \in \{1, ..., K\}$.

RAD+ loss becomes

$$
\ell_{RAD+} = \frac{1}{M} \cdot \sum_{m=1}^{M} (Q_m - y_k)^2 \tag{56}
$$

DrQ loss becomes

$$
\begin{aligned}
\ell_{DrQ} &= \frac{1}{M} \cdot \sum_{m=1}^{M} (Q_m - \frac{1}{K} \sum_{k=1}^{K} y_k)^2 \\
&= \frac{1}{M} \cdot \sum_{m=1}^{M} \Big(Q_m^2 + (\frac{1}{K} \sum_{k=1}^{K} y_k)^2 - 2Q_m \cdot \frac{1}{K} \sum_{k=1}^{K} y_k\Big) \\
&= \frac{1}{M} \cdot \sum_{m=1}^{M} Q_m^2 + \frac{1}{M} \cdot \sum_{m=1}^{M} (\frac{1}{K} \sum_{k=1}^{K} y_k)^2 - \frac{1}{M} \cdot \sum_{m=1}^{M} 2Q_m \cdot \frac{1}{K} \sum_{k=1}^{K} y_k \\
&= \frac{1}{M} \cdot \sum_{m=1}^{M} Q_m^2 + \frac{1}{K^2} (\sum_{k=1}^{K} y_j)^2 - \frac{2}{M \cdot K} \cdot \sum_{m=1}^{M} \sum_{k=1}^{K} Q_m \cdot y_k
\end{aligned} \tag{57}
$$

If all the combinations of above Q and y values are used for estimation, the loss becomes:

$$\begin{aligned}\ell_{all} &= \frac{1}{M \cdot K} \cdot \sum_{m=1}^{M}\sum_{k=1}^{K}(Q_m - y_k)^2 \\ &= \frac{1}{M \cdot K} \cdot (\sum_{m=1}^{M} K \cdot Q_m^2 + \sum_{k=1}^{K} M \cdot y_k^2 - 2\sum_{m=1}^{M}\sum_{k=1}^{K} Q_m \cdot y_k) \\ &= \frac{1}{M} \cdot \sum_{m=1}^{M} Q_m^2 + \frac{1}{K}(\sum_{k=1}^{K} y_k)^2 - \frac{2}{M \cdot K} \cdot \sum_{m=1}^{M}\sum_{k=1}^{K} Q_m \cdot y_k\end{aligned} \tag{58}$$

The second terms in both Equation 57 and 58 can be ignored because the gradients of target values y_j with respect to critic parameters are stopped.

Obviously, ℓ_{all} and ℓ_{DrQ} have same gradients with respect to trainable parameters of the critic. The comparison between ℓ_{DrQ} and ℓ_{RAD+} is exactly the comparison between ℓ_{all} and ℓ_{RAD+}. For one transition in one gradient step, $M \cdot K$ pairs of Q_m and y_k are used to formulate ℓ_{all} while only M pairs of Q_m and y_k are used to formulate ℓ_{RAD+}. Therefore, we can find out that DrQ outperforms RAD by leveraging more augmented samples and the averaged target. These operations indeed reduce the variance of the estimated critic loss.

E.2 KL REDUCES THE VARIANCE OF ACTOR LOSS

SAC actor loss Given data augmentation $f_\nu|\nu \sim \mathrm{P}$ on state s, if $Q(f_\nu(s), a)$ is invariant with respect to ν for all $a \in \mathcal{A}$, the variance of the actor loss $\mathbb{V}_\nu[\ell_\theta^I(s,\nu)]$ is bounded by a term that depends on the KL divergence $D_{\eta,\nu} = D_{KL}(\pi(\cdot \mid f_\eta(s))\|\pi(\cdot \mid f_\nu(s)))$ for $\nu, \eta \sim \mathrm{P}$:

$$\mathbb{V}_\nu[\ell_\theta^I(s,\nu)] \leq \frac{1}{n}\mathbb{E}_\nu\Big[\Big(\mathbb{E}_\eta[D_{\eta,\nu} + c(f_\nu(s))\sqrt{2D_{\eta,\nu}}]\Big)^2\Big], \tag{59}$$

where $c(f_\nu(s)) > 0, n$ is the number of samples to estimate the empirical mean $\ell_\theta^I(s,\nu)$.

Proof. For image-based control tasks, a data augmentation f parameterized by $\nu \sim \mathcal{P}$ is applied on the observations. The actor loss of SAC becomes

$$\ell_\theta^I(s,\nu) = \hat{\mathbb{E}}_\nu\Big[D_{KL}\Big(\pi_\theta(\cdot|f_\nu(s))||\exp(\frac{1}{\alpha}Q_\phi(f_\nu(s),\cdot) - \log Z(f_\nu(s)))\Big)\Big] \tag{60}$$

Let $g(f_\nu(s),\cdot) = \exp(\frac{1}{\alpha}Q_\phi(f_\nu(s),\cdot) - \log Z(f_\nu(s)))$.

The variance of empirical mean can be derived as the true variance divided by the number of samples n.

$$\begin{aligned}\mathbb{V}[\hat{\mathbb{E}}[x]] &= \mathbb{E}[(\hat{\mathbb{E}}(x) - \mathbb{E}[x])^2] \\ &= \mathbb{E}[(\frac{1}{n}(x_1 - \mathbb{E}[x] + x_2 - \mathbb{E}[x] + ... + x_n - \mathbb{E}[x]))^2] \\ &= \frac{1}{n^2}n \cdot \mathbb{V}[x] \\ &= \frac{1}{n}\mathbb{V}[x]\end{aligned} \tag{61}$$

The variance of $\ell_\theta^I(s,\nu)$ with respect to ν for a given number of samples n is

$$\begin{aligned}&\mathbb{V}_\nu[\ell_\theta^I(s,\nu)] \\ &= \frac{1}{n}\mathbb{V}_\nu[D_{KL}(\pi_\theta(\cdot|f_\nu(s))||g(f_\nu(s),\cdot))] \\ &= \frac{1}{n}\mathbb{E}_\nu\Big[\Big(D_{KL}(\pi_\theta(\cdot|f_\nu(s))||g(f_\nu(s),\cdot)) - \mathbb{E}_\eta[D_{KL}(\pi_\theta(\cdot|f_\eta(s))||g(f_\eta(s),\cdot))]\Big)^2\Big] \\ &= \frac{1}{n}\mathbb{E}_\nu\Big[\Big(D_{KL}(\pi_\theta(\cdot|f_\nu(s))||g(f_\nu(s),\cdot)) - \sum_\eta \mathcal{P}(\eta)D_{KL}(\pi_\theta(\cdot|f_\eta(s))||g(f_\eta(s),\cdot))\Big)^2\Big] \\ &= \frac{1}{n}\mathbb{E}_\nu\Big[\Big(\sum_\eta \mathcal{P}(\eta)(D_{KL}(\pi_\theta(\cdot|f_\eta(s))||g(f_\eta(s),\cdot)) - D_{KL}(\pi_\theta(\cdot|f_\nu(s))||g(f_\nu(s),\cdot)))\Big)^2\Big]\end{aligned} \tag{62}$$

For the term inside the above equation, we can further derive:

$$
\begin{aligned}
&D_{KL}(\pi_\theta(\cdot|f_\eta(s))||g(f_\eta(s),\cdot)) - D_{KL}(\pi_\theta(\cdot|f_\nu(s))||g(f_\nu(s),\cdot)) \\
&= \int_a \pi_\theta(\cdot|f_\eta(s)) \log \frac{\pi_\theta(\cdot|f_\eta(s))}{g(f_\eta(s),\cdot)} - \pi_\theta(\cdot|f_\nu(s)) \log \frac{\pi_\theta(\cdot|f_\nu(s))}{g(f_\nu(s),\cdot)} \\
&= \int_a \pi_\theta(\cdot|f_\eta(s)) \log \pi_\theta(\cdot|f_\eta(s)) - \pi_\theta(\cdot|f_\eta(s)) \log g(f_\eta(s),\cdot) \\
&- \pi_\theta(\cdot|f_\nu(s)) \log \pi_\theta(\cdot|f_\nu(s)) + \pi_\theta(\cdot|f_\nu(s)) \log g(f_\nu(s),\cdot) \\
&= \int_a \pi_\theta(\cdot|f_\eta(s)) \log \frac{\pi_\theta(\cdot|f_\eta(s))}{\pi_\theta(\cdot|f_\nu(s))} - \pi_\theta(\cdot|f_\eta(s)) \log g(f_\eta(s),\cdot) \\
&- (\pi_\theta(\cdot|f_\nu(s)) - \pi_\theta(\cdot|f_\eta(s))) \log \pi_\theta(\cdot|f_\nu(s)) + \pi_\theta(\cdot|f_\nu(s)) \log g(f_\nu(s),\cdot) \\
&= \int_a \pi_\theta(\cdot|f_\eta(s)) \log \frac{\pi_\theta(\cdot|f_\eta(s))}{\pi_\theta(\cdot|f_\nu(s))} - \pi_\theta(\cdot|f_\eta(s)) \log g(f_\nu(s),\cdot) + \pi_\theta(\cdot|f_\eta(s)) \log \frac{g(f_\nu(s),\cdot)}{g(f_\eta(s),\cdot)} \\
&- (\pi_\theta(\cdot|f_\nu(s)) - \pi_\theta(\cdot|f_\eta(s))) \log \pi_\theta(\cdot|f_\nu(s)) + \pi_\theta(\cdot|f_\nu(s)) \log g(f_\nu(s),\cdot) \\
&= D_{KL}(\pi_\theta(\cdot|f_\eta(s))||\pi_\theta(\cdot|f_\nu(s))) \\
&+ \int_a \Big(\pi_\theta(\cdot|f_\eta(s)) - \pi_\theta(\cdot|f_\nu(s))\Big) \cdot \Big(\log \pi_\theta(\cdot|f_\nu(s)) - \log g(f_\nu(s),\cdot)\Big) \\
&+ \int_a \pi_\theta(\cdot|f_\eta(s)) \log \frac{g(f_\nu(s),\cdot)}{g(f_\eta(s),\cdot)}
\end{aligned}
\tag{63}
$$

Then plug the above results into the equation of $\mathbb{V}_\nu[\ell_\theta^I(s,\nu)]$.

$$
\begin{aligned}
&\mathbb{V}_\nu[\ell_\theta^I(s,\nu)] \\
&= \frac{1}{n}\mathbb{E}_\nu\Big[\Big(\sum_\eta \mathcal{P}(\eta)(D_{KL}(\pi_\theta(\cdot|f_\eta(s))||g(f_\eta(s),\cdot)) - D_{KL}(\pi_\theta(\cdot|f_\nu(s))||g(f_\nu(s),\cdot)))\Big)^2\Big] \\
&= \frac{1}{n}\mathbb{E}_\nu\Big[\Big(\sum_\eta \mathcal{P}(\eta) D_{KL}(\pi_\theta(\cdot|f_\eta(s))||\pi_\theta(\cdot|f_\nu(s))) \\
&+ \sum_\eta P(\eta) \int_a (\pi_\theta(\cdot|f_\eta(s)) - \pi_\theta(\cdot|f_\nu(s)))(\log \pi_\theta(\cdot|f_\nu(s)) - \log g(f_\nu(s),\cdot)) \\
&+ \sum_\eta P(\eta) \int_a \pi_\theta(\cdot|f_\eta(s)) \log \frac{g(f_\nu(s),\cdot)}{g(f_\eta(s),\cdot)}\Big)^2\Big] \\
&= \frac{1}{n}\mathbb{E}_\nu\Big[\Big(\sum_\eta \mathcal{P}(\eta) D_{KL}(\pi_\theta(\cdot|f_\eta(s))||\pi_\theta(\cdot|f_\nu(s))) \\
&+ \sum_\eta P(\eta) \int_a (\pi_\theta(\cdot|f_\eta(s)) - \pi_\theta(\cdot|f_\nu(s))) \cdot (\log \pi_\theta(\cdot|f_\nu(s)) - \log g(f_\nu(s),\cdot)) \\
&+ \frac{1}{\alpha}\sum_\eta P(\eta) \int_a \pi_\theta(\cdot|f_\eta(s))(Q_\phi(f_\nu(s),a) - Q_\phi(f_\eta(s),a)) \\
&+ \sum_\eta P(\eta) \int_a \pi_\theta(\cdot|f_\eta(s)) \log \frac{Z(f_\nu(s))}{Z(f_\eta(s))}\Big)^2\Big]
\end{aligned}
\tag{64}
$$

For the second term on the right hand side of Equation 64, by applying Pinsker's inequality, we get

$$\begin{aligned}
&\sum_{\eta} P(\eta) \int_{a} (\pi_\theta(\cdot|f_\eta(s)) - \pi_\theta(\cdot|f_\nu(s))) \cdot (\log \pi_\theta(\cdot|f_\nu(s)) - \log g(f_\nu(s), \cdot)) \\
&\leq \sum_{\eta} \mathcal{P}(\eta) \int_{a} |\pi_\theta(a|f_\eta(s)) - \pi_\theta(a|f_\nu(s))| \cdot \max_a |\log \pi_\theta(a|f_\nu(s)) - \log g(f_\nu(s), a)| \\
&\leq \sum_{\eta} \mathcal{P}(\eta) \sqrt{2 D_{KL}\Big(\pi_\theta(\cdot|f_\eta(s)) || \pi_\theta(\cdot|f_\nu(s))\Big)} \cdot \max_a |\log \pi_\theta(a|f_\nu(s)) - \log g(f_\nu(s), a)|
\end{aligned} \tag{65}$$

For the third and fourth terms of Equation 64, given data augmentation $f_\nu|\nu \sim \mathrm{P}$ on state s, if $Q(f_\nu(s), a)$ is invariant with respect to ν for all $a \in \mathcal{A}$, both the third and the fourth terms of $\hat{\mathbb{V}}_\nu[\ell^I_\theta(s, \nu)]$ are zero.

Therefore, if $Q(f_\nu(s), a)$ is invariant with respect to ν for all $a \in \mathcal{A}$, the variance of the augmented actor loss $\mathbb{V}_\nu[\ell^I_\theta(s, \nu)]$ is bounded by the KL divergence $D_{\eta,\nu} = D_{KL}(\pi(\cdot \mid f_\eta(s)) \mid \pi(\cdot \mid f_\nu(s)))$ for $\nu, \eta \sim \mathrm{P}$:

$$\mathbb{V}_\nu[\ell^I_\theta(s, \nu)] \leq \frac{1}{n} \mathbb{E}_\nu \Big[\Big(\mathbb{E}_\eta [D_{\eta,\nu} + c(f_\nu(s)) \sqrt{2 D_{\eta,\nu}}] \Big)^2 \Big] \tag{66}$$

where $c(f_\nu(s)) = \max_a |\log \pi_\theta(a|f_\nu(s)) - \log g(f_\nu(s), a)| > 0$, n is the number of samples to estimate the empirical mean. □

DDPG actor loss Based on Equation 31, the DDPG actor loss $\ell^I_\theta(s, \mu)$ becomes,

$$\ell^I_\theta(s, \mu) \approx -\frac{1}{2} \hat{\mathbb{E}}_\mu \Big[(\pi_\theta(f_\mu(s)) - \pi^*(f_\mu(s)))^T H_\mu (\pi_\theta(f_\mu(s)) - \pi^*(f_\mu(s))) \Big]. \tag{67}$$

The variance of the actor loss is reduced if we minimize the mean squared error between two deterministic actions $||\pi_\theta(f_\eta(s')) - \pi_\theta(f_\nu(s'))||^2$, where $\eta, \nu \sim \mathrm{P}$.

Proof. Let $M_\mu = \pi_\theta(f_\mu(s)) - \pi^*(f_\mu(s))$. Assuming that the Hessian matrix H_μ have a a lower bound and upper bound:

$$l_\mu \mathbf{I} \preceq H_\mu \preceq L_\mu \mathbf{I}, \tag{68}$$

we have

$$l_\mu \|M_\mu\|^2 \leq \ell^I_\theta(s, \mu) \leq L_\mu \|M_\mu\|^2. \tag{69}$$

$$
\begin{aligned}
&\mathbb{V}_\nu[\ell^I_\theta(s,\nu)]\\
&=\frac{1}{n}\mathbb{E}_\nu[(\ell^I_\theta(s,\nu)-\mathbb{E}_\eta[\ell^I_\theta(s,\eta)])^2]=\frac{1}{n}\mathbb{E}_\nu[(\sum_\eta \mathcal{P}(\eta)\ell^I_\theta(s,\nu)-\sum_\eta \mathcal{P}(\eta)\ell^I_\theta(s,\eta))^2]\\
&=\frac{1}{n}\mathbb{E}_\nu[(\sum_\eta \mathcal{P}(\eta)(\ell^I_\theta(s,\nu)-\ell^I_\theta(s,\eta)))^2]\le\frac{1}{n}\mathbb{E}_\nu[(\sum_\eta \mathcal{P}(\eta)(\ell^I_\theta(s,\nu)-\ell^I_\theta(s,\eta))^2)]\\
&=\frac{1}{n}\mathbb{E}_{\nu,\eta}[(\ell^I_\theta(s,\nu)-\ell^I_\theta(s,\eta))^2]\le\frac{1}{n}\cdot(\max_\nu \ell^I_\theta(s,\nu)-\min_\eta \ell^I_\theta(s,\eta))^2\\
&\le\frac{1}{n}\cdot(\max_\nu L_\nu\|M_\nu\|^2-\min_\eta l_\eta\|M_\eta\|^2)^2=\frac{1}{n}\cdot(L_{\nu_{\max}}\|M_{\nu_{\max}}\|^2-l_{\eta_{\min}}\|M_{\eta_{\min}}\|^2)^2
\end{aligned}
$$

Let $\nu_{\max}=\arg\max_\nu \ell^I_\theta(s,\nu), \eta_{\min}=\arg\min_\eta \ell^I_\theta(s,\eta)$.

$$
\begin{aligned}
&\mathbb{V}_\nu[\ell^I_\theta(s,\nu)]\\
&\le\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2+l_{\eta_{\min}}(\|M_{\nu_{\max}}\|^2-\|M_{\eta_{\min}}\|^2))^2\\
&=\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}(\|\pi_\theta(f_{\nu_{\max}}(s))-\pi^*(s)\|^2-\|\pi_\theta(f_{\eta_{\min}}(s))-\pi^*(s)\|^2)^2\\
&=\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}(\sum_i(\pi_\theta(f_{\nu_{\max}}(s))_i-\pi^*(s)_i)^2-\sum_i(\pi_\theta(f_{\eta_{\min}}(s))_i-\pi^*(s)_i)^2)^2\\
&=\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}(\sum_i\Big((\pi_\theta(f_{\nu_{\max}}(s))_i-\pi^*(s)_i)^2-(\pi_\theta(f_{\eta_{\min}}(s))_i-\pi^*(s)_i)^2\Big))^2\\
&\le\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}\sum_i\Big((\pi_\theta(f_{\nu_{\max}}(s))_i-\pi^*(s)_i)^2-(\pi_\theta(f_{\eta_{\min}}(s))_i-\pi^*(s)_i)^2\Big)^2\\
&=\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}\sum_i\Big(\pi_\theta(f_{\nu_{\max}}(s))_i+\pi_\theta(f_{\eta_{\min}}(s))_i-2\pi^*(s)_i\Big)^2\Big(\pi_\theta(f_{\nu_{\max}}(s))_i-\pi_\theta(f_{\eta_{\min}}(s))_i\Big)^2\\
&=\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}\sum_i\Big(\pi_\theta(f_{\nu_{\max}}(s))_i+\pi_\theta(f_{\eta_{\min}}(s))_i-2\pi^*(s)_i\Big)^2\Big(\pi_\theta(f_{\nu_{\max}}(s))_i-\pi_\theta(f_{\eta_{\min}}(s))_i\Big)^2\\
&=\frac{1}{n}\cdot((L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2\\
&+l_{\eta_{\min}}\sum_i\Big(\pi_\theta(f_{\nu_{\max}}(s))_i+\pi_\theta(f_{\eta_{\min}}(s))_i-2\pi^*(s)_i\Big)^2\Big(\pi_\theta(f_{\nu_{\max}}(s))_i-\pi_\theta(f_{\eta_{\min}}(s))_i\Big)^2
\end{aligned}
\tag{70}
$$

Since

$$\begin{aligned}
&(a-b)^2(a+b-2c)^2 \\
&\leq \frac{(a-b)^4+(a+b-2c)^4}{2} \\
&= \frac{(a-b)^4+((a-c+b-c)^2)^2}{2} \\
&\leq \frac{(a-b)^4+(2(a-c)^2+2(b-c)^2)^2}{2} \\
&= \frac{(a-b)^4+4((a-c)^2+(b-c)^2)^2}{2} \\
&\leq \frac{(a-b)^4+8(a-c)^4+8(b-c)^4}{2}
\end{aligned} \tag{71}$$

we have

$$\begin{aligned}
&\mathbb{V}_\nu[\ell_\theta^I(s,\nu)] \\
&\leq \frac{1}{n}\cdot(L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2 \\
&+ l_{\eta_{\min}}\sum_i\Big(\pi_\theta(f_{\nu_{\max}}(s))_i+\pi_\theta(f_{\eta_{\min}}(s))_i-2\pi^*(s)_i\Big)^2\Big(\pi_\theta(f_{\nu_{\max}}(s))_i-\pi_\theta(f_{\eta_{\min}}(s))_i\Big)^2 \\
&\leq \frac{1}{n}\cdot(L_{\nu_{\max}}-l_{\eta_{\min}})\|M_{\nu_{\max}}\|^2 \\
&+ \frac{1}{2}l_{\eta_{\min}}\|\pi_\theta(f_{\nu_{\max}}(s))-\pi_\theta(f_{\eta_{\min}}(s))\|^4 \\
&+ 4l_{\eta_{\min}}\|\pi_\theta(f_{\nu_{\max}}(s))-\pi^*(s))\|^4 \\
&+ 4l_{\eta_{\min}}\|\pi_\theta(f_{\eta_{\min}}(s))-\pi^*(s))\|^4
\end{aligned} \tag{72}$$

□

In Equation 72, the third and fourth terms are minimized by the actor loss. If we minimize the second term of Equation 72 by minimizing the mean squared error between two deterministic actions $||\pi_\theta(f_\eta(s'))-\pi_\theta(f_\nu(s'))||^2$ in the case of DDPG, the variance of the actor loss is reduced.

E.3 KL REDUCES THE VARIANCE OF THE TARGET Q-VALUE

DDPG target values For DDPG, when we compute target values, we add Ornstein-Uhlenbeck noise to deterministic actions for exploration. Then the policy can be regarded as a probability distribution π.

For image-based control tasks, a data augmentation f parameterized by $\mu \sim \mathcal{P}$ is applied on the observations. Then the target value y for a given transition (s,a,r,s') is

$$y(f_\mu(s'),a') = r+\gamma Q_{\bar{\phi}}(f_\mu(s'),a'), \text{where } a'\sim\pi(\cdot|f_\mu(s')). \tag{73}$$

The expectation of $y(f_\mu(s'),a')$ with respect to $a'\sim\pi(\cdot|f_\mu(s'))$ is

$$\begin{aligned}
\mathbb{E}_{a'}[y(f_\mu(s'),a')] &= r+\gamma\mathbb{E}_{a'}[Q_{\bar{\phi}}(f_\mu(s'),a')] \\
&= r+\gamma\sum_{a'}\pi(a'|f_\mu(s'))Q_{\bar{\phi}}(f_\mu(s'),a')
\end{aligned} \tag{74}$$

The expectation of $y(f_\mu(s'),a')$ with respect to $\mu\sim\mathcal{P}$ and $a'\sim\pi(\cdot|f_\mu(s'))$ is

$$\begin{aligned}
&\mathbb{E}_{\mu,a'}[y(f_\mu(s'),a')] \\
&= r+\gamma\mathbb{E}_{\mu,a'}[Q_{\bar{\phi}}(f_\mu(s'),a')] \\
&= r+\gamma\sum_\mu\mathcal{P}(\mu)\sum_{a'}\pi(a'|f_\mu(s'))Q_{\bar{\phi}}(f_\mu(s'),a')
\end{aligned} \tag{75}$$

We create two tables to better illustrate the meanings of $\mathbb{E}_{a'}[y(f_\mu(s'), a')]$ and $\mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')]$.

a' \ μ	...	$f_{\tau_m}(s')$ with $\mathcal{P}(\mu = \tau_m)$	...
a'_1	...	$y(f_{\tau_m}(s'), a'_1)$ with $\mathcal{P}(\mu = \tau_m) \cdot \pi_\theta(a'_1 \vert f_{\tau_m}(s'))$	...
a'_2	...	$y(f_{\tau_m}(s'), a'_2)$ with $\mathcal{P}(\mu = \tau_m) \cdot \pi_\theta(a'_2 \vert f_{\tau_m}(s'))$	...
...	...	...	...
a'_n	...	$y(f_{\tau_m}(s'), a'_n)$ with $\mathcal{P}(\mu = \tau_m) \cdot \pi_\theta(a'_n \vert f_{\tau_m}(s'))$	...
...	...	...	...
$\mathbb{E}[y]$ wrt. a'	...	$\mathbb{E}_{a'}[y(f_{\tau_m}(s'), a')]$	...

$f_{\tau_1}(s')$ with $\mathcal{P}(\mu = \tau_1)$	...	$\mathbb{E}[y]$ wrt. a and μ
$\mathbb{E}_{a'}[y(f_{\tau_1}(s'), a')]$	...	$\mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')]$

The variance of $y(f_\mu(s'), a')$ with respect to μ and a' is

$$
\begin{aligned}
&\mathbb{V}_{\mu,a'}[y(f_\mu(s'), a')] \\
&= \sum_\mu \mathcal{P}(\mu) \sum_{a'} \pi(a' \vert f_\mu(s')) \Big[(y(f_\mu(s'), a') - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')])^2 \Big] \\
&= \sum_\mu \mathcal{P}(\mu) \sum_{a'} \pi(a' \vert f_\mu(s')) \\
&\Big[(y(f_\mu(s'), a') - \mathbb{E}_{a'}[y(f_\mu(s'), a')] + \mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')])^2 \Big] \\
&= \sum_\mu \mathcal{P}(\mu) \sum_{a'} \pi(a' \vert f_\mu(s')) \Big[(y(f_\mu(s'), a') - \mathbb{E}_{a'}[y(f_\mu(s'), a')])^2 \\
&+ 2(y(f_\mu(s'), a') - \mathbb{E}_{a'}[y(f_\mu(s'), a')]) \cdot (\hat{\mathbb{E}}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')]) \\
&+ (\mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')])^2 \Big]
\end{aligned} \tag{76}
$$

The first term of Equation 76 is the expectation of squared advantage.

The second term of Equation 76 is 0 because

$$
\begin{aligned}
&\sum_\mu \mathcal{P}(\mu) \sum_{a'} \pi(a' \vert f_\mu(s')) \Big[2(y(f_\mu(s'), a') - \mathbb{E}_{a'}[y(f_\mu(s'), a')]) \\
&\cdot (\mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')]) \Big] \\
&= 2 \sum_\mu \Big[\mathcal{P}(\mu) \cdot (\mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')]) \\
&\cdot \Big(\sum_{a'} \pi(a' \vert f_\mu(s'))(y(f_\mu(s'), a') - \mathbb{E}_{a'}[y(f_\mu(s'), a')]) \Big) \Big] \\
&= 2 \sum_\mu \Big[\mathcal{P}(\mu) \cdot (\mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\mu,a'}[y(f_\mu(s'), a')]) \\
&\cdot \Big(\mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{a'}[y(f_\mu(s'), a')]) \Big) \Big] \\
&= 0
\end{aligned} \tag{77}
$$

The third term of Equation 76 is the variance of $\mathbb{E}_{a' \sim \pi(\cdot \vert f_\mu(s'))}[y(f_\mu(s'), a')]$ with respect to μ. Both the variance $\mathbb{V}_\mu[\mathbb{E}_{a' \sim \pi(\cdot \vert f_\mu(s'))}[y(f_\mu(s'), a')]]$ and the variance of the empirical mean

$\mathbb{V}_\mu[\hat{\mathbb{E}}_\mu[\mathbb{E}_{a'\sim\pi_\theta(\cdot|f_\mu(s'))}[y(f_\mu(s'),a')]]]$ are bounded by the KL divergence $D_{\eta,\mu} = D_{KL}(\pi(\cdot \mid f_\eta(s')) \mid \pi(\cdot \mid f_\mu(s')))$ for $\mu,\eta\sim\mathcal{P}$ if $Q_{\bar{\phi}}(f_\mu(s'),a')$ is invariant with respect to μ for all $a'\in\mathcal{A}$.

Proof.

$$\mathbb{V}_\mu[\mathbb{E}_{a'\sim\pi(\cdot|f_\mu(s'))}[y(f_\mu(s'),a')]] = \mathbb{E}_\mu\Big[(\mathbb{E}_{a'}[y(f_\mu(s'),a')] - \mathbb{E}_{\eta,a'}[y(f_\eta(s'),a')])^2\Big] \tag{78}$$

$$\begin{aligned}
&\mathbb{E}_{a'}[y(f_\mu(s'),a')] - \mathbb{E}_{\eta,a'}[y(f_\eta(s'),a')] \\
&= \gamma\Big((\sum_{a'}\pi(a'|f_\mu(s'))Q_{\bar{\phi}}(f_\mu(s'),a')) - (\sum_\eta \mathcal{P}(\eta)\int_{a'}\pi(a'|f_\eta(s'))Q_{\bar{\phi}}(f_\eta(s'),a'))\Big) \\
&= \gamma\Big((\sum_\eta \mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\mu(s'))Q_{\bar{\phi}}(f_\mu(s'),a')) - (\sum_\eta \mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\eta(s'))Q_{\bar{\phi}}(f_\eta(s'),a'))\Big) \\
&= \gamma\Big(\sum_\eta \mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\mu(s'))Q_{\bar{\phi}}(f_\mu(s'),a') - \pi(a'|f_\eta(s'))Q_{\bar{\phi}}(f_\eta(s'),a')\Big) \\
&= \gamma\Big(\sum_\eta \mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\mu(s'))Q_{\bar{\phi}}(f_\mu(s'),a') - \pi(a'|f_\eta(s'))Q_{\bar{\phi}}(f_\mu(s'),a') \\
&+ \pi(a'|f_\eta(s'))Q_{\bar{\phi}}(f_\mu(s'),a') - \pi(a'|f_\eta(s'))Q_{\bar{\phi}}(f_\eta(s'),a')\Big) \\
&= \gamma\Big(\sum_\eta \mathcal{P}(\eta)\sum_{a'}(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s')))Q_{\bar{\phi}}(f_\mu(s'),a') \\
&+ \pi(a'|f_\eta(s'))(Q_{\bar{\phi}}(f_\mu(s'),a') - Q_{\bar{\phi}}(f_\eta(s'),a'))\Big)
\end{aligned} \tag{79}$$

The second term of Equation 79 $\gamma\sum_\eta \mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\eta(s'))(Q_{\bar{\phi}}(f_\mu(s'),a') - Q_{\bar{\phi}}(f_\eta(s'),a'))$ is related to the difference of $Q_{\bar{\phi}}(f_\eta(s'),a')$ and $Q_{\bar{\phi}}(f_\mu(s'),a')$, which is governed by the critic loss. When $Q_{\bar{\phi}}(f_\mu(s'),a')$ is invariant with respect to μ for all $a'\in\mathcal{A}$, this term is zero.

For the first term of Equation 79,

$$\begin{aligned}
&\sum_\eta \mathcal{P}(\eta)\sum_{a'}(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s')))Q_{\bar{\phi}}(f_\mu(s'),a') \\
&\leq \sum_\eta \mathcal{P}(\eta)\sum_{a'}|\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))|Q_{\bar{\phi}}(f_\mu(s'),a') \\
&\leq \max_{a'} Q_{\bar{\phi}}(f_\mu(s'),a')\cdot\sum_\eta \mathcal{P}(\eta)\sum_{a'}|\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))| \\
&\leq \max_{a'} Q_{\bar{\phi}}(f_\mu(s'),a')\cdot\sum_\eta \mathcal{P}(\eta)\sqrt{2D_{KL}(\pi(\cdot|f_\eta(s'))||\pi(\cdot|f_\mu(s')))}
\end{aligned} \tag{80}$$

where in the first inequality absolute values $|\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))|$ are applied, in the second inequality $Q_{\bar{\phi}}(f_\mu(s'),a')$ is replaced with $\max_{a'} Q_{\bar{\phi}}(f_\mu(s'),a')$ and Pinsker's inequality is applied in the third inequality.

Similarly, a lower bound can be derived.

$$\begin{aligned}
&\sum_\eta \mathcal{P}(\eta)\sum_{a'}(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s')))Q_{\bar{\phi}}(f_\mu(s'),a') \\
&\geq -\max_{a'} Q_{\bar{\phi}}(f_\mu(s'),a')\cdot\sum_\eta \mathcal{P}(\eta)\sqrt{2D_{KL}(\pi(\cdot|f_\eta(s'))||\pi(\cdot|f_\mu(s')))}
\end{aligned} \tag{81}$$

Therfore,

$$\mathbb{V}_\mu[\mathbb{E}_{a'\sim\pi(\cdot|f_\mu(s'))}[y(f_\mu(s'),a')]] \leq \mathbb{E}_\mu\Big[\gamma^2\Big(\max_{a'} Q_{\bar{\phi}}(f_\mu(s'),a')\mathbb{E}_\eta\Big[\sqrt{2D_{\eta,\mu}}\Big]\Big)^2\Big] \tag{82}$$

Let $\hat{Y}(s', \mu) = \hat{\mathbb{E}}_\mu[\mathbb{E}_{a' \sim \pi_\theta(\cdot|f_\mu(s'))}[y(f_\mu(s'), a')]]$.

From Equation 61,

$$\mathbb{V}_\mu[\hat{Y}(s', \mu)] = \frac{1}{n}\mathbb{V}_\mu[\mathbb{E}_{a' \sim \pi_\theta(\cdot|f_\mu(s'))}[y(f_\mu(s'), a')]], \quad \text{where } n \text{ is the number of samples to estimate the empirical mean } \hat{Y}(s', \mu). \tag{83}$$

Therefore, if $Q_{\bar{\phi}}(f_\mu(s), a')$ is invariant with respect to μ for all $a' \in \mathcal{A}$, the variance of $\hat{Y}(s', \mu)$ with respect to μ is bounded by the KL divergence $D_{\eta,\mu} = D_{KL}(\pi(\cdot \mid f_\eta(s')) \mid \pi(\cdot \mid f_\mu(s')))$ for $\mu, \eta \sim \mathcal{P}$.

$$\mathbb{V}_\mu[\hat{Y}(s', \mu)] \leq \frac{1}{n}\mathbb{E}_\mu\Big[\gamma^2\Big(\max_{a'} Q_{\bar{\phi}}(f_\mu(s'), a')\mathbb{E}_\eta\Big[\sqrt{2D_{\eta,\mu}}\Big]\Big)^2\Big] \tag{84}$$

For DDPG, minimizing the KL divergence between policy distributions of two augmented states $D_{KL}(\pi(\cdot \mid f_\eta(s')) \mid \pi(\cdot \mid f_\mu(s')))$ is equivalent to minimizing the mean squared error between two deterministic actions $||\bar{\pi}(f_\eta(s')) - \bar{\pi}(f_\mu(s'))||^2$. □

SAC target value with the entropy term If the entropy term is added to the target value, the variance of the empirical mean $\mathbb{V}_\mu[\hat{\mathbb{E}}_\mu[\mathbb{E}_{a' \sim \pi_\theta(\cdot|f_\mu(s'))}[y(f_\mu(s'), a')]]]$ is still bounded by the KL divergence $D_{\eta,\mu} = D_{KL}(\pi(\cdot \mid f_\eta(s')) \mid \pi(\cdot \mid f_\mu(s')))$ for $\mu, \eta \sim \mathcal{P}$ if $Q_{\bar{\phi}}(f_\mu(s'), a')$ is invariant with respect to μ for all $a' \in \mathcal{A}$.

$$\mathbb{V}_\mu[\hat{\mathbb{E}}_\mu[\mathbb{E}_{a' \sim \pi_\theta(\cdot|f_\mu(s'))}[y(f_\mu(s'), a')]]] \leq \frac{1}{n}\mathbb{E}_\mu\Big[\Big(\mathbb{E}_\eta\Big[\max_{a'}(y(f_\mu(s'), a') - r)\sqrt{2D_{\eta,\mu}} + \alpha \cdot D_{\eta,\mu}\Big]\Big)^2\Big] \tag{85}$$

where n is the number of samples to estimate the empirical mean, r is the reward of this transition and α is the entropy coefficient.

Proof. After we add the entropy term, the target value becomes

$$y(f_\mu(s'), a') = r + \gamma Q_{\bar{\phi}}(f_\mu(s'), a') - \alpha \log \pi(a'|f_\mu(s')), \tag{86}$$

where $a' \sim \pi(\cdot|f_\mu(s'))$ and α is the entropy coefficient.

Let

$$y_1(f_\mu(s'), a') = y(f_\mu(s'), a') - r = \gamma Q_{\bar{\phi}}(f_\mu(s'), a') - \alpha \log \pi(a'|f_\mu(s')) \tag{87}$$

Since r is a constant value, we can drop r when calculating the variance.

$$\begin{aligned}
&\mathbb{V}_\mu[\mathbb{E}_{a' \sim \pi(\cdot|f_\mu(s'))}[y(f_\mu(s'), a')]] \\
&= \mathbb{V}_\mu[\mathbb{E}_{a' \sim \pi(\cdot|f_\mu(s'))}[y_1(f_\mu(s'), a')]] \\
&= \mathbb{E}_\mu\Big[(\mathbb{E}_{a'}[\gamma Q_{\bar{\phi}}(f_\mu(s'), a') - \alpha \log \pi(a'|f_\mu(s'))] - \mathbb{E}_{\eta,a'}[\gamma Q_{\bar{\phi}}(f_\eta(s'), a') - \alpha \log \pi(a'|f_\eta(s'))])^2\Big]
\end{aligned} \tag{88}$$

$$
\begin{aligned}
&\mathbb{E}_{a'}[y_1(f_\mu(s'),a')] - \mathbb{E}_{\eta,a'}[y_1(f_\eta(s'),a')] \\
&= \sum_{a'} \pi(a'|f_\mu(s'))y_1(f_\mu(s'),a') - \sum_{\eta}\mathcal{P}(\eta)\int_{a'}\pi(a'|f_\eta(s'))y_1(f_\eta(s'),a') \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\mu(s'))y_1(f_\mu(s'),a') - \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\eta(s'))y_1(f_\eta(s'),a') \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\mu(s'))y_1(f_\mu(s'),a') - \pi(a'|f_\eta(s'))y_1(f_\eta(s'),a') \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\mu(s'))y_1(f_\mu(s'),a') - \pi(a'|f_\eta(s'))y_1(f_\mu(s'),a') \\
&+ \pi(a'|f_\eta(s'))y_1(f_\mu(s'),a') - \pi(a'|f_\eta(s'))y_1(f_\eta(s'),a') \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}y_1(f_\mu(s'),a')(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))) \\
&+ \pi(a'|f_\eta(s'))(y_1(f_\mu(s'),a') - y_1(f_\eta(s'),a')) \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}y_1(f_\mu(s'),a')(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))) \\
&+ \pi(a'|f_\eta(s'))(\gamma Q_{\bar{\phi}}(f_\mu(s'),a') - \gamma Q_{\bar{\phi}}(f_\eta(s'),a')) \\
&+ \pi(a'|f_\eta(s'))(\alpha\log\pi(a'|f_\eta(s')) - \alpha\log\pi(a'|f_\mu(s'))) \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}y_1(f_\mu(s'),a')(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))) \\
&+ \pi(a'|f_\eta(s'))(\gamma Q_{\bar{\phi}}(f_\mu(s'),a') - \gamma Q_{\bar{\phi}}(f_\eta(s'),a')) \\
&+ \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\eta(s'))(\alpha\log\pi(a'|f_\eta(s')) - \alpha\log\pi(a'|f_\mu(s'))) \\
&= \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}y_1(f_\mu(s'),a')(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s'))) \\
&+ \gamma\sum_{\eta}\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\eta(s'))(Q_{\bar{\phi}}(f_\mu(s'),a') - Q_{\bar{\phi}}(f_\eta(s'),a')) \\
&+ \sum_{\eta}\mathcal{P}(\eta)\alpha\cdot D_{KL}(\pi(a'|f_\eta(s'))|\pi(a'|f_\mu(s')))
\end{aligned}
\tag{89}
$$

Similar to Equation 80 and Equation 81, we apply Pinsker's inequality and obtain the lower and the upper bounds for the first term of Equation 89.

$$
\begin{aligned}
&-\max_{a'} y_1(f_\mu(s'),a')\cdot\sum_{\eta}\mathcal{P}(\eta)\sqrt{2D_{KL}(\pi(\cdot|f_\eta(s'))||\pi(\cdot|f_\mu(s')))} \\
&\leq \sum_{\eta}\mathcal{P}(\eta)\sum_{a'}(\pi(a'|f_\mu(s')) - \pi(a'|f_\eta(s')))y(f_\mu(s'),a') \\
&\leq \max_{a'} y_1(f_\mu(s'),a')\cdot\sum_{\eta}\mathcal{P}(\eta)\sqrt{2D_{KL}(\pi(\cdot|f_\eta(s'))||\pi(\cdot|f_\mu(s')))}
\end{aligned}
\tag{90}
$$

The second term of Equation 89 $\gamma\sum_\eta\mathcal{P}(\eta)\sum_{a'}\pi(a'|f_\eta(s'))(Q_{\bar{\phi}}(f_\mu(s'),a') - Q_{\bar{\phi}}(f_\eta(s'),a'))$ is related to the difference of $Q_{\bar{\phi}}(f_\eta(s'),a')$ and $Q_{\bar{\phi}}(f_\mu(s'),a')$, which is governed by the critic loss. When $Q_{\bar{\phi}}(f_\mu(s'),a')$ is invariant with respect to μ for all $a'\in\mathcal{A}$, this term is zero. □

Therefore,

$$\begin{aligned}
&\sum_{\eta} \mathcal{P}(\eta) \cdot \Big(-\max_{a'} y_1(f_\mu(s'), a') \cdot \sqrt{2D_{KL}(\pi(\cdot|f_\eta(s'))||\pi(\cdot|f_\mu(s')))} + \alpha \cdot D_{KL}(\pi(a'|f_\eta(s'))|\pi(a'|f_\mu(s'))) \Big) \\
&\leq \mathbb{E}_{a'}[y(f_\mu(s'), a')] - \mathbb{E}_{\eta, a'}[y(f_\eta(s'), a')] \\
&\leq \sum_{\eta} \mathcal{P}(\eta) \cdot \Big(\max_{a'} y_1(f_\mu(s'), a') \cdot \sqrt{2D_{KL}(\pi(\cdot|f_\eta(s'))||\pi(\cdot|f_\mu(s')))} + \alpha \cdot D_{KL}(\pi(a'|f_\eta(s'))|\pi(a'|f_\mu(s'))) \Big)
\end{aligned} \tag{91}$$

Plug the above inequalities into Equation 88, we obtain

$$\mathbb{V}_\mu[\mathbb{E}_{a' \sim \pi(\cdot|f_\mu(s'))}[y(f_\mu(s'), a')]] \leq \mathbb{E}_\mu\Big[\Big(\mathbb{E}_\eta\Big[\max_{a'} y_1(f_\mu(s'), a')\sqrt{2D_{\eta,\mu}} + \alpha \cdot D_{\eta,\mu}\Big]\Big)^2\Big] \tag{92}$$

If $Q_{\bar{\phi}}(f_\mu(s), a')$ is invariant with respect to μ for all $a' \in \mathcal{A}$, the variance of $\hat{Y}(s', \mu)$ with respect to μ is bounded by the KL divergence $D_{\eta,\mu} = D_{KL}(\pi(\cdot \mid f_\eta(s')) \mid \pi(\cdot \mid f_\mu(s')))$ for $\mu, \eta \sim \mathcal{P}$.

$$\mathbb{V}_\mu[\hat{Y}(s', \mu)] \leq \frac{1}{n}\mathbb{E}_\mu\Big[\Big(\mathbb{E}_\eta\Big[\max_{a'}(y(f_\mu(s'), a') - r)\sqrt{2D_{\eta,\mu}} + \alpha \cdot D_{\eta,\mu}\Big]\Big)^2\Big] \tag{93}$$

F CALCULATING TARGET WITH COMPLEX DATA AUGMENTATION

In this section, we experimentally analyze using complex image transformations in calculating the target and show that cosine similarity of the augmented features at the early training stage can be used as a criteria for judging if an image transformation is complex or not. Sufficient updates is the key condition for good performance when using complex image transformations in calculating the target.

In contrast to the analysis in SVEA (Hansen et al., 2021), we observe that even using complex image transformation such as random conv in the target does not induce a large variance in the target. Instead, a much larger bias is observed for the trained agent, as shown in the Table 5. This can be solved by increasing the number of updates, as shown in Figure 4.

Furthermore, we test with other image transformations which are regarded as complex image transformations in SVEA (Hansen et al., 2021). In order to show whether it's easy to enforce the invariance of a image transformation, we record the cosine similarities of encoder outputs for two augmented images transformed by this image transformation, as shown in Table 4. For image transformations such as random overlay or gaussian blur, the invariance is easy to enforce and the cosine similarities are large. When using this kind of image transformation in calculating the target values, it won't hurt the performance. Otherwise, for image transformations such as random convolution or random rotation, the invariance is relatively harder to enforce during training and the cosine similarities are small. Then directly applying this kind of image transformations in calculating the target values will decrease the learning efficiency. To resolve this issue, we need more updates for each training step. The evaluation results for SVEA with random overlay, random convolution, random rotation and gaussian blur are shown in Figure 5.

G HYPERPARAMETERS

Hyperparameters used in experiments on DMControl (drq), DMControl (drqv2) and DMGB can be found in Table 6, Table 7 and 8. For experiments in DMControl(drqv2) and DMGB, when applicable, we adopt hyperparameters from the official implementation of drqv2 by Yarats et al. (2021) and SVEA by Hansen et al. (2021) respectively.

H ADDITIONAL RESULTS

H.1 ABLATION STUDY

The performance profile is shown in Figure 6 and the training curves in different environments are shown in Figure 7.

Statistics	svea(DA=blur)	svea(DA=overlay)	svea(DA=conv)	svea(DA=rotation)
actor sim (shift)	0.919±0.007	0.911±0.005	0.910±0.011	0.936±0.006
actor sim (DA)	0.998±0.001	0.906±0.006	0.854±0.019	0.536±0.069
critic sim (shift)	0.938±0.010	0.942±0.003	0.922±0.010	0.962±0.005
critic sim (DA)	0.998±0.001	0.939±0.003	0.883±0.014	0.660±0.057

Table 4: Recorded cosine similarity for latent features at 100k steps in walker walk environment for SVEA trained with different complex image transformations. Here, each column corresponds to SVEA with different image transformations and each row corresponds to a cosine similarity recorded at 100k steps. For example, the first number is calculated by the cosine similarity between the latent features $E_\pi(f_{shift}(s))$, in which E_π is the encoder of the actor in SVEA trained with gaussian blur and f_{shift} is the random shift. Considering that random shift is always applied in SVEA, the cosine similarity with respect to it is recorded as the baseline for comparisons. For gaussian blur and random overlay (second and third column), the cosine similarities of latent features are higher or similar to the cosine similarities between the latent features from two randomly shifted images which means they are not complex image transformations. In contrast, random conv and random rotation (last two columns) leads to smaller cosine similarities of latent features which indicates that they are relatively complex image transformations in this environment.

Statistics	Step 100k	Step 200k	Step 300k	Step 400k	Step 500k
Mean (w/ conv)	64.05	112.44	140.78	166.42	185.92
Mean (w/o conv)	84.96	148.46	197.10	221.67	239.35
Variance (w/ conv)	1.228	1.109	1.148	1.323	1.306
Variance (w/o conv)	0.882	0.812	0.885	0.757	0.795
Bias (w/ conv)	52.88	67.44	54.20	63.89	55.36
Bias (w/o conv)	77.40	60.48	50.40	43.22	28.75

Table 5: Mean, variance and bias of the target Q-values for the agent trained with/without using random conv in calculating the target. Here, the mean and variance are calculated by the mean and variance of a set of sampled target Q values and the bias is calculated by the mean-squared error between the targets used in the training and the true targets estimated by the sum of discounted rewards from sampled trajectories. At the end of the training, the increase in the variance when using random conv is not significant compared to the mean of the target. However, the bias is much larger at the end.

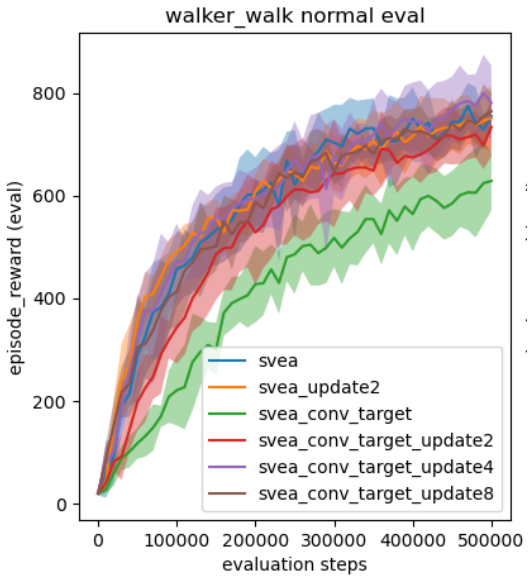

Figure 4: Performance of increasing the number of updates with/without using random conv in calculating the targets.

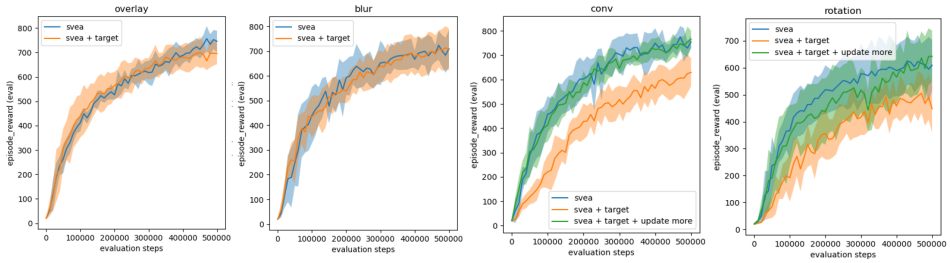

Figure 5: Performance of increasing the number of updates in walker walk environment when using complex image transformation in calculating the targets. For random convolution and random rotation, ”update more” stands for doing 4 updates for each training step.

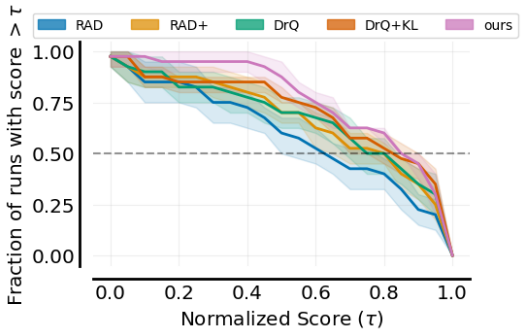

Figure 6: Performance profile of different methods.

Table 6: Hyperparameters used in experiments on DMControl (drq)

Hyperparameter	Value on DMControl
frame rendering	$84 \times 84 \times 3$
stacked frames	3
action repeat	2
replay buffer capacity	100,000
seed steps	1000
environment steps	250,000 in reacher easy 250,000 in finger spin 250,000 in ball 500,000 in others
batch size N	256
discount γ	0.99
optimizer (ϕ, θ)	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
optimizer (α of SAC)	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
learning rate (ϕ,θ)	1e-3
learning rate (α of SAC)	1e-3
target network update frequency	2
target network soft-update rate	0.01
actor update frequency κ	2
actor log stddev bounds	[-10,2]
init temperature α	0.1
tangent prop weight α_{tp}	0.1
actor KL weight α_{KL}	0.1

Table 7: Hyperparameters used in experiments on DMControl (drqv2)

Hyperparameter	Value on DMC
frame rendering	$84 \times 84 \times 3$
stacked frames	3
action repeat	2
replay buffer capacity	10^6
seed frames	4000
exploration steps	2000
n-step returns	3
batch size N	256
discount γ	0.99
optimizer (ϕ, θ)	Adam
learning rate (ϕ,θ)	1e-4
agent update frequency	2
target network soft-update rate	0.01
exploration stddev clip	0.3
exploration stddev schedule	linear(1.0, 0.1, 500000)
tangent prop weight α_{tp}	0.1
actor KL weight α_{KL}	0.1

H.2 CASE STUDY

The measures for invariance in the latent space are shown in Figure 8.

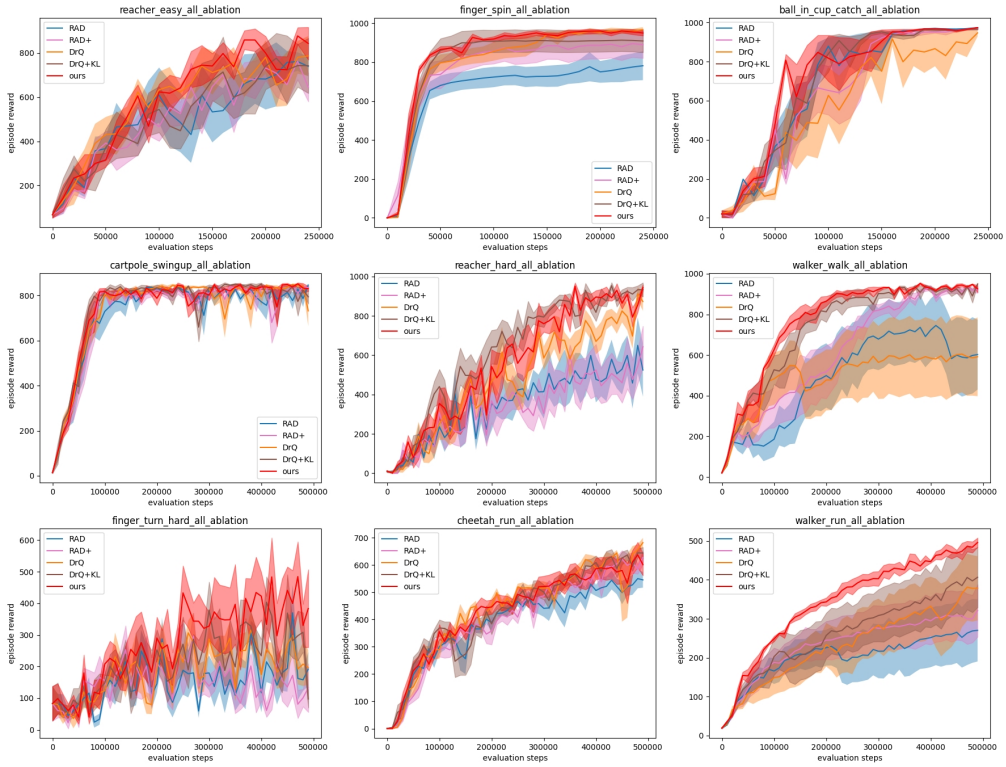

Figure 7: Full results of validating our propositions.

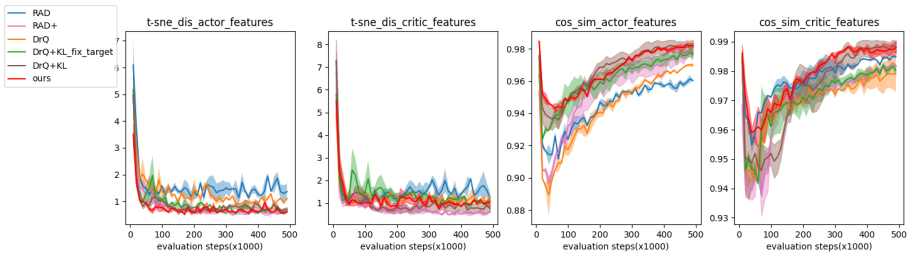

Figure 8: The figure shows the learned invariance in the feature space of the actor and critic. Two measures of the invariance are provided in this figure: the distances between projected points of the augmented features by t-SNE and the cosine similarities between augmented features.

Table 8: Hyperparameters used in experiments on DMControl Generalization Benchmark (DMGB)

Hyperparameter	Value on DMGB
frame rendering	$84 \times 84 \times 3$
stacked frames	3
action repeat	2(finger) 8(cartpole) 4(otherwise)
replay buffer capacity	500,000 / action repeat
seed steps	1000
environment steps	500,000
batch size N	128
discount γ	0.99
optimizer (ϕ, θ)	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
optimizer (α of SAC)	Adam ($\beta_1 = 0.5, \beta_2 = 0.999$)
learning rate (ϕ,θ)	1e-3
learning rate (α of SAC)	1e-4
target network update frequency	2
target network soft-update rate	0.01(critic) 0.05(encoder)
actor update frequency κ	2
actor log stddev bounds	[-10,2]
init temperature α	0.1
tangent prop weight α_{tp}	0.5
actor KL weight α_{KL}	0.1

H.3 More evaluations

Here, we include more evaluations of our proposition. The results of comparing our proposition with DrQ are shown in Figure 9. The results of comparing our proposition with DrQv2 are shown in Figure 10.

H.4 Results of generalization ability in DMControl Generalization Benchmark (DMGB)

The comparison of generalization performance in DMGB between SVEA and our method using random overlay as data augmentation is shown in Figure 11.

H.5 Results of recorded statistics

The curves for the recorded statistics, including standard deviation of the empirical critic loss, standard deviation of the target Q-values, and empirical mean of KL divergence between policies for two augmented samples along the training are shown in Figure 12 and Figure 13.

I Limitations

We try to provide some recommendations on how to apply theoretically-sound data augmentation method in DRL. However, the analysis can still be further refined to be more comprehensive such as including the theoretical analysis of using different distributions for the image transformation and providing a thorough analysis on tangent prop regularization. Moreover, our method naturally requires the knowledge of some effective image transformations for a given task. Without such knowledge, the invariant transformations for a problem would need to be learned, which is currently an active research direction. Finally, image transformation may rely on some implicit assumptions,

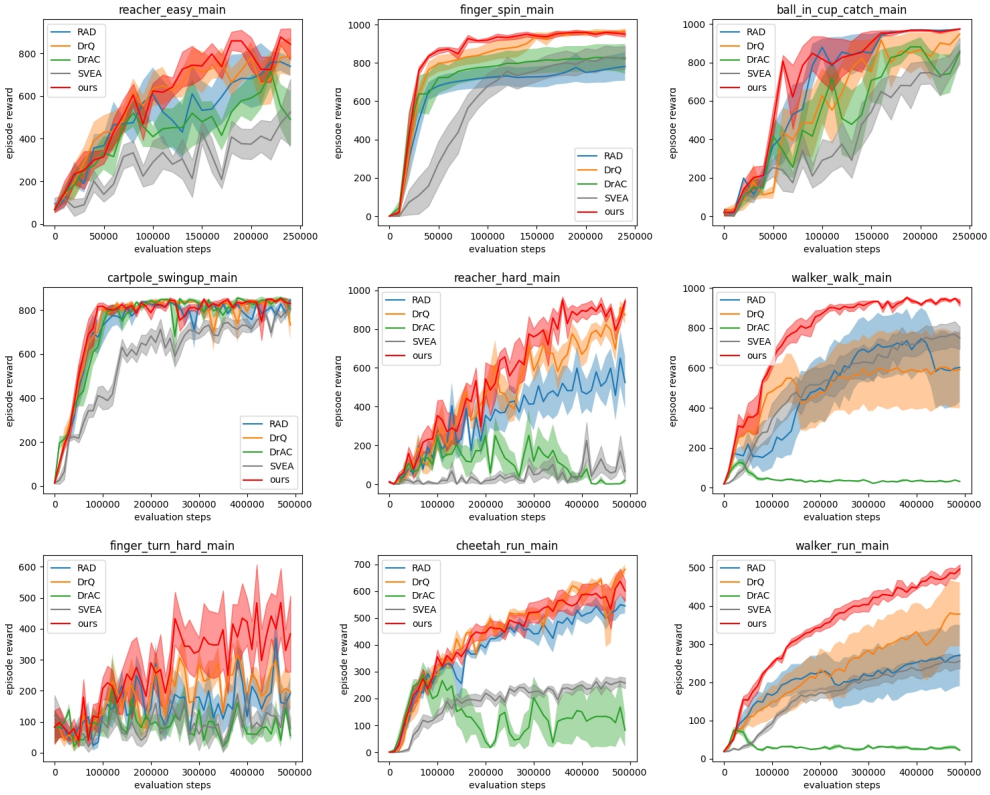

Figure 9: Comparison between different methods in DMControl with normal background.

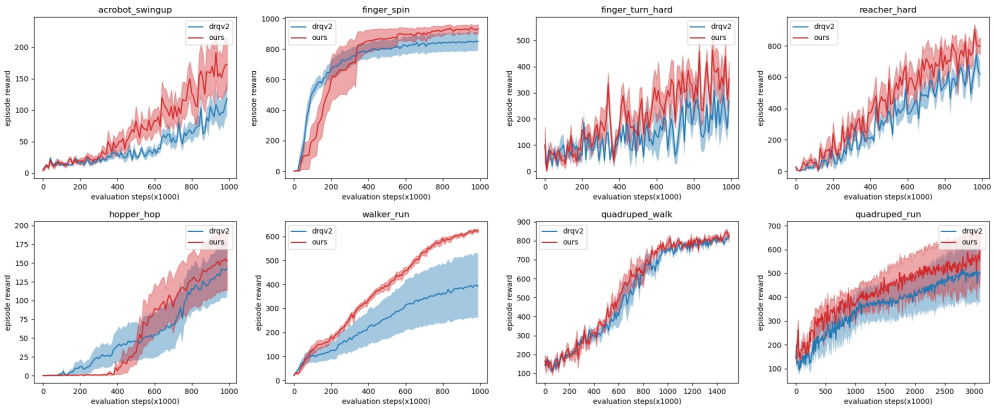

Figure 10: Results of running experiments with DDPG as base algorithm.

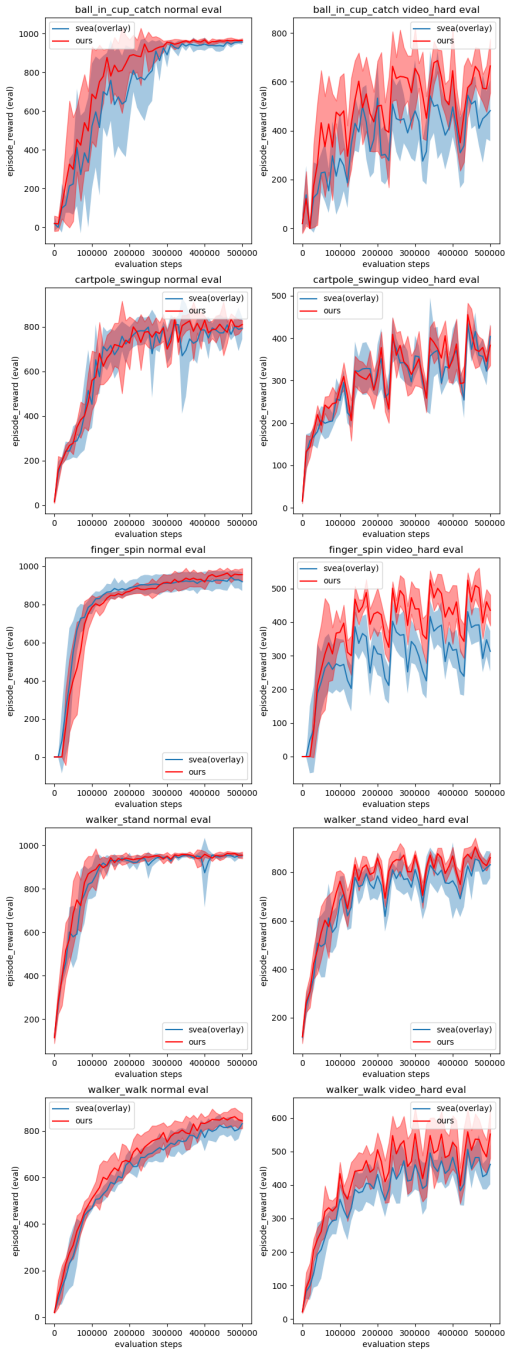

Figure 11: Comparison between SVEA and our method in DMControl with normal and video-hard backgrounds. Both methods use random overlay as image transformation. We can see the improvement in generalization ability especially in environments such as ball in cup catch, finger spin and walker walk. Since the evaluation curves are not stable even at the end of training, the recorded score in Table 2 is the average over the last 15 evaluation scores.

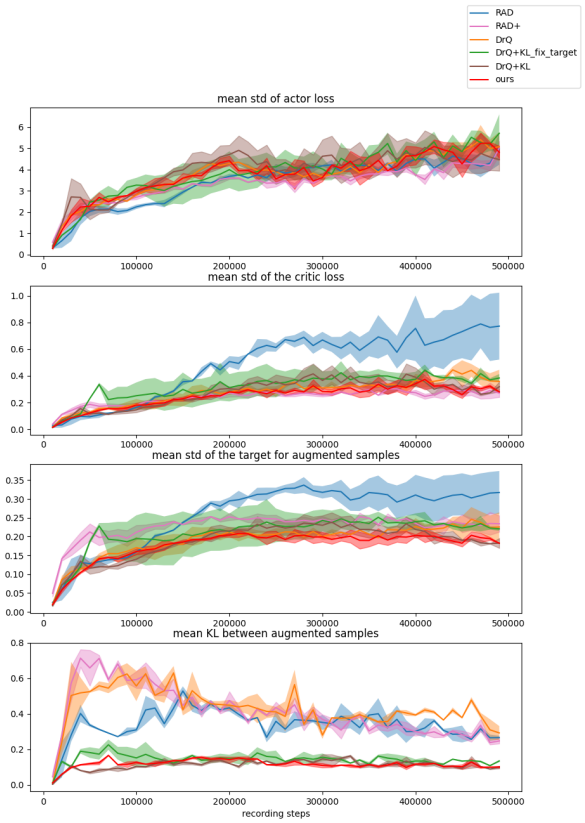

Figure 12: Some important statistics recorded along the training. The variance of critic loss and target values decreased after using more augmented samples in the training of the critic. Adding the KL divergence term to the loss can quickly enforce the invariance of the actor even at the beginning of the training.

which may lead to lower/bad performance if they are not satisfied in the real application domain. For instance, random shift/crop, which has been shown to be very effective in DMControl tasks, may yield worse performance if the agent is not well-centered in the image, according to the empirical results from Tomar et al. (2022). A better understanding of why a data augmentation transformation works in DRL is needed.

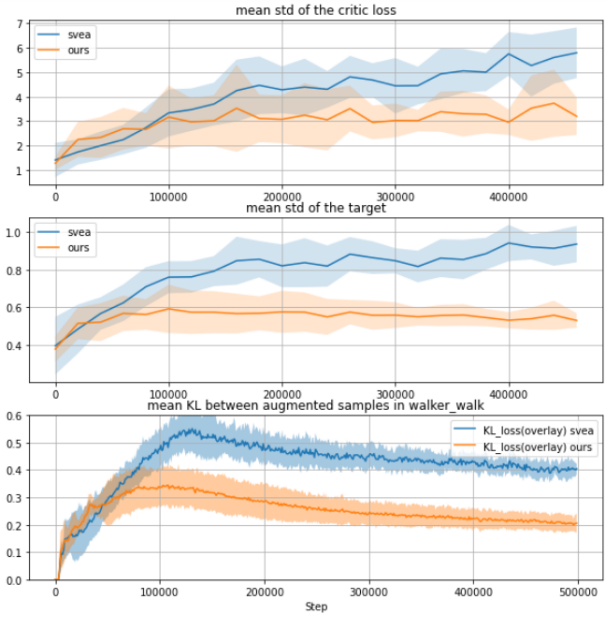

Figure 13: Some important statistics recorded along the training of SVEA and our method. With the help of KL loss and tangent prop loss, the variance of critic loss and target values are lower. Applying KL loss can quickly enforce the invariance of the actor.