Supplementary Information

Here, we provide additional information on different parts of the paper. In particular, in section 1 we introduce and discuss two chain models in polymer physics. In section 2, we provide the theoretical proofs of Theorems 3 and 4, Lemma 2, and Corollary 5 in the manuscript. In section 3, we present the action-sampling algorithm, and in section 4 we provide additional baseline results in the standard MuJoCo tasks. Finally, in section 5, we provide the network architecture of the learning methods, as well as the PolyRL hyper parameters used in the experimental section.

1 Polymer Models

In the field of *Polymer Physics*, the conformations and interactions of polymers that are subject to thermal fluctuations are modeled using principles from statistical physics. In its simplest form, a polymer is modeled as an *ideal chain*, where interactions between chain segments are ignored. The *no-interaction* assumption allows the chain segments to cross each other in space and thus these chains are often called *phantom chains* [1]. In this section, we give a brief introduction to two types of ideal chains.

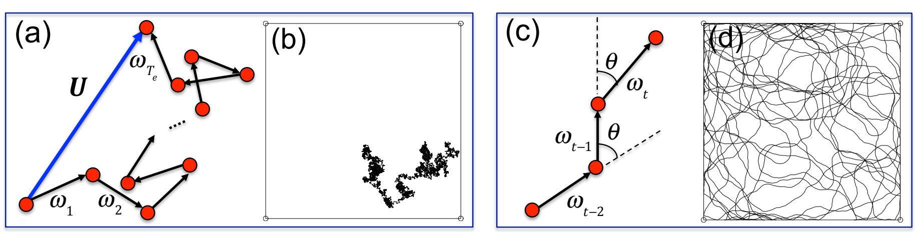

Figure 1: A chain (or trajectory) is shown as a sequence of T_e random bond vectors $\{\boldsymbol{\omega}_i\}_{i=1..T_e}$. In a freely-jointed chain (a), the orientation of the bond vectors are independent of one another. The end-to-end vector of the chain is depicted by $\boldsymbol{U}$. In a freely-rotating chain (c), the correlation angle θ is invariant between every two consecutive bond vectors, which induces a finite stiffness in the chain. (b, d) A qualitative comparison between an FJC (b) and an FRC with $\theta \approx 5.7°$ (d), in a 2D environment of size 400×400 for 20000 number of moves.

Two main ideal chain models are: 1) freely-jointed chains (FJCs) and 2) freely-rotating chains (FRCs) [1]. In these models, chains of size T_e are demonstrated as a sequence of T_e random vectors $\{\boldsymbol{\omega}_i\}_{i=1..T_e}$, which are as well called *bond vectors* (See Figure 1). FJC is the simplest proposed model and is composed of mutually independent random vectors of the same size (Figure 1(a)). In other words, an FJC chain is formed via uniform random sampling of vectors in space, and thus is a *random walk* (RW). In the FRC model, on the other hand, the notion of *correlation angle* is introduced, which is the angle θ between every two consecutive bond vectors. The FRC model, fixes the correlation angle θ (Figure 1(c)), thus the vectors in the chain are temporally correlated. The vector sampling strategy in the FRCs induces *persistent chains*, in the sense that the orientation of the consecutive vectors in the space are preserved for certain number of time steps (a.k.a. persistence number), after which the correlation is broken and the bond vectors *forget* their original orientation. This feature introduces a finite *stiffness* in the chain, which induces what we call *local* self avoidance, leading to faster expansion of the chain in the space (Compare Figures 1 (b) and (d) together). Below, we discuss two important properties of the FJCs and the FRCs, and subsequently formally introduce the *locally self-avoiding random walks* (LSA-RWs) in Definition 1.

FJCs (*Property*) **-** In the Freely-Jointed Chains (FJCs) or the flexible chains model, the orientations of the bond vectors in the space are mutually independent. To measure the expected end-to-end length of a chain $\tilde{U}$ with T_e bond vectors of constant length b_o given the end-to-end vector $\mathbf{U} = \sum_{i=1}^{T_e} \boldsymbol{\omega}_i$ (Figure 1 (a)) and considering the mutual independence between bond vectors of an FJC, we can write [1],

$$\mathbb{E}[\|\mathbf{U}\|^2] = \sum_{i,j=1}^{T_e} \mathbb{E}[\boldsymbol{\omega}_i.\boldsymbol{\omega}_j] = \sum_{i=1}^{T_e} \mathbb{E}[\boldsymbol{\omega}_i^2] + 2\sum_{i>j} \mathbb{E}[\boldsymbol{\omega}_i.\boldsymbol{\omega}_j] = T_e b_o^2, \tag{1}$$

where $\mathbb{E}[.]$ denotes the ensemble average over all possible conformations of the chain as a result of thermal fluctuations. Equation 1 shows that the expected end-to-end length $\tilde{U} = \mathbb{E}[\|\mathbf{U}\|^2]^{1/2} = b_o\sqrt{T_e}$, which reveals random-walk behaviour as expected.

FRCs (*Property*) **-** In the Freely-Rotating Chains (FRCs) model, we assume that the angle θ (correlation angle) between every two consecutive bond vectors is invariant (Figure 1 (c)). Therefore, bond vectors $\boldsymbol{\omega}_{i:1,\ldots,T_e}$ are not mutually independent. Unlike the FJC model, in the FRC model the bond vectors are correlated such that [1],

$$\mathbb{E}[\boldsymbol{\omega}_i.\boldsymbol{\omega}_j] = b_o^2\,(\cos\theta)^{|i-j|} = b_o^2 e^{-\frac{|i-j|}{Lp}}, \tag{2}$$

where $Lp = \frac{1}{|\log(\cos\theta)|}$ is the correlation length (persistence number). Equation 2 shows that the correlation between bond vectors in an FRC is a decaying exponential with correlation length Lp.

Lemma 1. *[1] Given an FRC characterized by end-to-end vector $\boldsymbol{U}$, bond-size b_o and number of bond vectors T_e, we have $\mathbb{E}[\|\boldsymbol{U}\|^2] = b^2 T_e$, where $b^2 = b_o^2 \frac{1+cos\theta}{1-cos\theta}$ and b is called the effective bond length.*

Lemma 1 shows that FRCs obey random walk statistics with step-size (bond length) $b > b_o$. The ratio $b/b_o = \frac{1+\cos\theta}{1-\cos\theta}$ is a measure of the stiffness of the chain in an FRC.

FRCs have high expansion rates compared to those of FJCs, as presented in Proposition 2 below.

Proposition 2 (Expanding property of LSA-RW)**.** *[1] Let τ be a LSA-RW with the persistence number $Lp_\tau > 1$ and the end-to-end vector $\boldsymbol{U}(\tau)$, and let τ' be a random walk (RW) and the end-to-end vector $\boldsymbol{U}(\tau')$. Then for the same number of time steps and same average bond length for τ and τ', the following relation holds,*

$$\frac{\mathbb{E}[\|\boldsymbol{U}(\tau)\|]}{\mathbb{E}[\|\boldsymbol{U}(\tau')\|]} = \frac{1 + e^{-1/Lp_\tau}}{1 - e^{-1/Lp_\tau}} > 1, \tag{3}$$

where the persistence number $Lp_\tau = \frac{1}{|\log \cos \theta|}$, with θ being the average correlation angle between every two consecutive bond vectors.

Proof. This proposition is the direct result of combining Equations 2.7 and 2.14 in [1]. Equation 2.7 provides the expected T_e time-step length of the end-to-end vector with average step-size b_o associated with FJCs and Equation 2.14 provides a similar result for FRCs. Note that in the FRC model, since the bond vectors far separated in time on the chain are not correlated, they can cross each other. □

Radius of Gyration (*Formal Definition*) [2] The square radius of gyration $U_g^2(\tau)$ of a chain τ of size T_e is defined as the mean square distance between position vectors $\boldsymbol{t} \in \tau$ and the chain center of mass ($\bar{\boldsymbol{\tau}}$), and is written as,

$$U_g^2(\tau) := \frac{1}{T_e} \sum_{i=1}^{T_e} ||\boldsymbol{t}_i - \bar{\boldsymbol{\tau}}||^2, \tag{4}$$

where $\bar{\boldsymbol{\tau}} = \frac{1}{T_e} \sum_{i=1}^{T_e} \boldsymbol{t}_i$. When it comes to selecting a measure of coverage in the space where the chain resides, radius of gyration U_g is a more proper choice compared with the end-to-end distance $||\boldsymbol{U}||$, as it signifies the size of the chain with respect to its center of mass, and is proportional to the radius of the sphere (or the hyper sphere) that the chain occupies. Moreover, in the case of chains that are circular or branched, and thus cannot be assigned an end-to-end length, radius of gyration proves to be a suitable measure for the size of the corresponding chains [2]. For the case of fluctuating chains, the square radius of gyration is usually ensemble averaged over all possible chain conformations, and is written as [2],

$$\mathbb{E}[U_g^2(\tau)] := \frac{1}{T_e} \sum_{i=1}^{T_e} \mathbb{E}[||\boldsymbol{t}_i - \bar{\boldsymbol{\tau}}||^2]. \tag{5}$$

Remark 1. *The square radius of gyration U_g^2 is proportional to the square end-to-end distance $||\boldsymbol{U}||^2$ in ideal chains (*e.g.* FJCs and FRCs) with a constant factor [2]. Thus, Proposition 2 and Equation 3, which compare the the end-to-end distance of LSA-RW and RW with each other, similarly hold for the radius of gyration of the respective models, implying faster expansion of the volume occupied by LSA-RW compared with that of RW.*

2 The Proofs

In this section, the proofs for the theorems and Lemma 2 in the manuscript are provided.

2.1 The proof of Lemma 2 in the manuscript

Lemma 2 statement: Let $\tau_{\mathcal{S}} = (s_0, \ldots, s_{T_e-1})$ be the trajectory of visited states, s_{T_e} be a newly visited state and $\boldsymbol{\omega}_i = s_i - s_{i-1}$ be the bond vector that connects two consecutive visited states s_{i-1} and s_i. Then we have,

$$\|s_{T_e} - \bar{\tau}_{\mathcal{S}}\|^2 = \|\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right]\|^2. \tag{6}$$

Proof. Using the relation $\bar{\tau}_{\mathcal{S}} := \frac{1}{T_e}\sum_{s\in\tau_{\mathcal{S}}} s$ as well as the definition of bond vectors (Equation (3) in the manuscript), we can write $s_{T_e} - \bar{\tau}_{\mathcal{S}}$ on the left-hand side of Equation (6) in the manuscript as,

$$\begin{aligned}
s_{T_e} - \bar{\tau}_{\mathcal{S}} =& s_{T_e} - \frac{1}{T_e}\sum_{s\in\tau_{\mathcal{S}}} s \\
=& s_{T_e} - s_{T_e-1} + s_{T_e-1} - \frac{1}{T_e}\sum_{s\in\tau_{\mathcal{S}}} s \\
=& \boldsymbol{\omega}_{T_e} + \frac{1}{T_e}(s_{T_e-1} - s_0) + (s_{T_e-1} - s_1) + (s_{T_e-1} - s_2) + \ldots \\
& + (s_{T_e-1} - s_{T_e-2})] \\
=& \boldsymbol{\omega}_{T_e} + \frac{1}{T_e}[(s_{T_e-1} - s_{T_e-2} + s_{T_e-2} - s_{T_e-3} + \ldots \\
& + s_2 - s_1 + s_1 - s_0) + (s_{T_e-1} - s_{T_e-2} + s_{T_e-2} - s_{T_e-3} + \ldots \\
& + s_3 - s_2 + s_2 - s_1) + \cdots + (s_{T_e-1} - s_{T_e-2})] \\
=& \boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[(\boldsymbol{\omega}_{T_e-1} + \cdots + \boldsymbol{\omega}_1) + (\boldsymbol{\omega}_{T_e-1} + \cdots + \boldsymbol{\omega}_2) + \cdots + \boldsymbol{\omega}_{T_e-1}\right] \\
=& \boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right]
\end{aligned} \tag{7}$$

$$\Rightarrow \|s_{T_e} - \bar{\tau}_{\mathcal{S}}\|^2 = \|\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right]\|^2 \tag{8}$$

□

2.2 The proof of Theorem 3 in the manuscript

Theorem 3 statement (Upper-Bound Theorem) Let $\beta \in (0,1)$ and $\tau_{\mathcal{S}}$ be an LSA-RW in $\mathcal{S}$ induced by PolyRL with the persistence number $Lp_{\tau_{\mathcal{S}}} > 1$ within episode N, $\omega_{\tau_{\mathcal{S}}} = \{\boldsymbol{\omega}_i\}_{i=1}^{T_e-1}$ be the sequence of corresponding bond vectors, where $T_e > 0$ denotes the number of bond vectors within $\tau_{\mathcal{S}}$, and b_o be the average bond length. The upper confidence bound for $ULS_{Ug^2}(\tau_{\mathcal{S}})$ with probability of at least $1-\delta$ is,

$$UB = \Lambda(T_e, \tau_{\mathcal{S}}) + \frac{1}{\delta}\left[\Gamma(T_e, b_o, \tau_{\mathcal{S}}) + \frac{2b_o^2}{T_e^2}\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp\tau_{\mathcal{S}}}}\right], \tag{9}$$

where,

$$\Lambda(T_e, \tau_{\mathcal{S}}) = -\frac{1}{T_e - 1}U_g^2(\tau_{\mathcal{S}}) \tag{10}$$

$$\Gamma(T_e, b_o, \tau_{\mathcal{S}}) = \frac{b_o^2}{T_e} + \frac{\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2}{T_e^3} \tag{11}$$

Proof. If we replace the term ${U_g}^2(\tau'_{\mathcal{S}})$ in Equation (5) in the manuscript with its incremental representation as a function of ${U_g}^2(\tau_{\mathcal{S}})$, we get

$$\begin{aligned} ULS_{Ug^2}(\tau_{\mathcal{S}}) &= \sup_{s_{T_e}\in\Omega}\left(\frac{T_e-2}{T_e-1}{U_g}^2(\tau_{\mathcal{S}}) + \frac{1}{T_e}\|s_{T_e} - \bar{\tau}_{\mathcal{S}}\|^2 - {U_g}^2(\tau_{\mathcal{S}})\right) \\ &= -\frac{1}{T_e-1}{U_g}^2(\tau_{\mathcal{S}}) + \sup_{s_{T_e}\in\Omega}\frac{1}{T_e}\|s_{T_e} - \bar{\tau}_{\mathcal{S}}\|^2. \end{aligned} \tag{12}$$

Therefore, the problem reduces to the calculation of

$$\frac{1}{T_e}\sup_{s_{T_e}\in\Omega}\|s_{T_e} - \bar{\tau}_{\mathcal{S}}\|^2. \tag{13}$$

Using Lemma 2 in the manuscript, we can write Equation (13) in terms of bond vectors $\boldsymbol{\omega}_i = s_i - s_{i-1}$ as,

$$\frac{1}{T_e}\sup_{s_{T_e}\in\Omega}\|s_{T_e} - \bar{\tau}_{\mathcal{S}}\|^2 = \frac{1}{T_e}\sup_{s_{T_e}\in\Omega}\|\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right]\|^2. \tag{14}$$

From now on, with a slight abuse of notation, we will treat $\boldsymbol{\omega}_{T_e} = S_{T_e} - s_{T_e-1}$ as a random variable due to the fact that S_{T_e} is a random variable in our system. Note that $\boldsymbol{\omega}_i$ for $i = 1, 2, \ldots, T_e - 1$ is fixed, and thus is not considered a random variable. We use high-probability concentration bound techniques to calculate Equation (13). For any $\delta \in (0, 1)$, there exists $\alpha > 0$, such that

$$\Pr[\|\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right]\|^2 < \alpha | S_{T_e} \in \Omega] > 1 - \delta. \tag{15}$$

We can rearrange Equation 15 as,

$$\Pr[\|\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right]\|^2 \geq \alpha | S_{T_e} \in \Omega] \leq \delta. \tag{16}$$

Multiplying both sides by T_e^2 and expanding the squared term in Equation 16 gives,

$$\Pr[T_e^2\|\boldsymbol{\omega}_{T_e}\|^2 + 2T_e(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i) + \|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 \geq T_e^2\alpha | S_{T_e} \in \Omega] \leq \delta. \quad (17)$$

By Markov's inequality we have,

$$\begin{aligned}
&\Pr\left[T_e^2\|\boldsymbol{\omega}_{T_e}\|^2 + 2T_e(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i) + \|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 \geq T_e^2\alpha\right] \\
&\quad \leq \frac{\mathbb{E}\left[T_e^2\|\boldsymbol{\omega}_{T_e}\|^2 + 2T_e(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i) + \|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right]}{T_e^2\alpha} = \delta \\
&\implies \alpha = \frac{1}{\delta T_e^2}\left[T_e^2\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2\right] + 2T_e\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right] + \|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right] \\
&\underbrace{\implies}_{\text{by Def. 1}} \alpha = \frac{1}{\delta T_e^2}\left[T_e^2 b_o^2 + 2T_e\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right] + \|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right]
\end{aligned}$$

Note that all expectations $\mathbb{E}$ in the equations above are over the transition kernel $\mathcal{P}$ of the MDP. Using the results from Lemma 3 below, we conclude the proof. $\square$

Lemma 3. *Let $\tau_\mathcal{S}$ denote the sequence of states observed by PolyRL and S_{T_e} be the new state visited by PolyRL. Assuming that $\tau'_\mathcal{S} := (\tau_\mathcal{S}, S_{T_e})$ (Equation (2) in the manuscript) follows the LSA-RW formalism with the persistence number $Lp_{\tau_\mathcal{S}} > 1$, we have*

$$\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right] = b_0^2\sum_{i=1}^{T_e-1} ie^{\frac{-|T_e-i|}{Lp_{\tau_\mathcal{S}}}} \quad (18)$$

Proof.

$$\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right] = \mathbb{E}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] = \sum_{i=1}^{T_e-1} i\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right]. \quad (19)$$

Here, the goal is to calculate the expectation in Equation 19 under the assumption that $\tau'_\mathcal{S}$ is LSA-RW with persistence number $Lp_{\tau_\mathcal{S}} > 1$. Note that if $\tau'_\mathcal{S}$ is LSA-RW and $Lp_{\tau_\mathcal{S}} > 1$, the chain of states visited by PolyRL prior to visiting s_{T_e} is also LSA-RW with $Lp_{\tau_\mathcal{S}} > 1$. Now we focus on the expectation in Equation 19. We compute $\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right]$ using the LSA-RW formalism (Definition 1 in the manuscript) as following,

$$\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] = b_0^2 e^{\frac{-|T_e-i|}{Lp_{\tau_\mathcal{S}}}}$$

Therefore, we have,

$$\sum_{i=1}^{T_e-1} i\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] = \sum_{i=1}^{T_e-1} ib_0^2 e^{\frac{-(T_e-i)}{Lp_{\tau_\mathcal{S}}}} = b_0^2\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp_{\tau_\mathcal{S}}}}$$

$\square$

2.3 The proof of Theorem 4 in the manuscript

Theorem 4 statement (Lower-Bound Theorem) Let $\beta \in (0,1)$ and $\tau_{\mathcal{S}}$ be an LSA-RW in $\mathcal{S}$ induced by PolyRL with the persistence number $Lp_{\tau_{\mathcal{S}}} > 1$ within episode N, $\omega_{\tau_{\mathcal{S}}} = \{\boldsymbol{\omega}_i\}_{i=1}^{T_e-1}$ be the sequence of corresponding bond vectors, where $T_e > 0$ denotes the number of bond vectors within $\tau_{\mathcal{S}}$, and b_o be the average bond length. The lower confidence bound for $LLS_{Ug^2}(\tau_{\mathcal{S}})$ at least with probability $1-\delta$ is,

$$LB = \Lambda(T_e, \tau_{\mathcal{S}}) + (1 - \sqrt{2-2\delta}) \left[\Gamma(T_e, b_o, \tau_{\mathcal{S}}) + \frac{(T_e-1)(T_e-2)}{T_e^2} b_0^2 e^{\frac{-|T_e-1|}{Lp_{\tau_{\mathcal{S}}}}} \right], \quad (20)$$

where,

$$\Lambda(T_e, \tau_{\mathcal{S}}) = -\frac{1}{T_e - 1} U_g^2(\tau_{\mathcal{S}}) \quad (21)$$

$$\Gamma(T_e, b_o, \tau_{\mathcal{S}}) = \frac{b_o^2}{T_e} + \frac{\| \sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i \|^2}{T_e^3} \quad (22)$$

Proof. Using the definition of radius of gyration and letting $d = L_2$-norm in Equation (4) in the manuscript, we have

$$\begin{aligned} LLS_{Ug^2}(\tau_{\mathcal{S}}) &= \inf_{s_{T_e} \in \Omega} \frac{T_e - 2}{T_e - 1} U_g^2(\tau_{\mathcal{S}}) + \frac{1}{T_e} \| s_{T_e} - \bar{\tau}_{\mathcal{S}} \|^2 - U_g^2(\tau_{\mathcal{S}}) \\ &= -\frac{1}{T_e - 1} U_g^2(\tau_{\mathcal{S}}) + \inf_{s_{T_e} \in \Omega} \frac{1}{T_e} \| s_{T_e} - \bar{\tau}_{\mathcal{S}} \|^2 \end{aligned} \quad (23)$$

To calculate the high-probability lower bound, first we use the result from Lemma 2 in the manuscript. Thus, we have

$$\inf_{s_{T_e} \in \Omega} \frac{1}{T_e} \| s_{T_e} - \bar{\tau}_{\mathcal{S}} \|^2 = \frac{1}{T_e} \inf_{s_{T_e} \in \Omega} \| \boldsymbol{\omega}_{T_e} + \frac{1}{T_e} \left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i \right] \|^2. \quad (24)$$

We subsequently use the second moment method and Paley–Zygmund inequality to calculate the high-probability lower bound. Let $Y = \| \boldsymbol{\omega}_{T_e} + \frac{1}{T_e} \left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i \right] \|^2$, for the finite positive constants c_1 and c_2 we have,

$$\Pr[Y > c_2 \beta] \geq \frac{(1-\beta)^2}{c_1} \quad (25)$$

where,

$$\begin{aligned} \mathbb{E}\left[Y^2\right] &\leq c_1 \mathbb{E}\left[Y\right]^2 \\ \mathbb{E}\left[Y\right] &\geq c_2. \end{aligned} \quad (26)$$

The goal is to find two constants c_1 and c_2 such that Equation (26) is satisfied and then we find $\beta \in (0,1)$ in Equation (25) using δ. We start by finding c_2 ,

$$\begin{aligned}
\mathbb{E}\left[Y\right] &= \mathbb{E}\left[(\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right] s).(\boldsymbol{\omega}_{T_e} + \frac{1}{T_e}\left[\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right])\right] \\
&= \mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2 + \frac{2}{T_e}(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i) + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right] \\
&= \mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2\right] + \frac{2}{T_e}\sum_{i=1}^{T_e-1} i\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 \\
&= b_o^2 + \frac{2}{T_e}b_0^2\sum_{i=1}^{T_e-1} ie^{\frac{-|T_e-i|}{Lp_{\tau_{\mathcal{S}}}}} + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2,
\end{aligned} \tag{27}$$

therefore,

$$\begin{aligned}
\mathbb{E}\left[Y\right] &= b_o^2 + \frac{2}{T_e}b_0^2\sum_{i=1}^{T_e-1} ie^{\frac{-|T_e-i|}{Lp_{\tau_{\mathcal{S}}}}} + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 \\
&\geq b_o^2 + \frac{2}{T_e}b_0^2 e^{\frac{-|T_e-1|}{Lp_{\tau_{\mathcal{S}}}}}\sum_{i=1}^{T_e-1} i + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 \\
&= \underbrace{b_o^2 + \frac{(T_e-1)(T_e-2)}{T_e}b_0^2 e^{\frac{-|T_e-1|}{Lp_{\tau_{\mathcal{S}}}}} + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2}_{=c_2}
\end{aligned} \tag{28}$$

To find c_1, we have

$$\mathbb{E}\left[Y^2\right] \leq c_1\mathbb{E}\left[Y\right]^2.$$

$$
\begin{aligned}
\mathbb{E}\left[Y^2\right] &= \mathbb{E}\left[\left(\|\boldsymbol{\omega}_{T_e}\|^2 + \frac{2}{T_e}(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i) + \frac{1}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right)^2\right] \\
&= \mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^4\right] + \frac{4}{T_e^2}\mathbb{E}\left[(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)^2\right] \\
&+ \frac{1}{T_e^4}\mathbb{E}\left[\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^4\right] + \frac{4}{T_e}\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)\right] \\
&+ \frac{2}{T_e^2}\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right] + \frac{2}{T_e^3}\mathbb{E}\left[(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right] \\
&= \mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^4\right] + \frac{4}{T_e^2}\mathbb{E}\left[(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)^2\right] \\
&+ \frac{1}{T_e^4}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^4 + \frac{4}{T_e}\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)\right] \\
&+ \frac{2b_o^2}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 + \frac{2}{T_e^3}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\mathbb{E}\left[(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)\right]
\end{aligned} \tag{29}
$$

We calculate the expectations appearing in Equation (29) to conclude the proof.

$$
\begin{aligned}
\frac{4}{T_e^2}\mathbb{E}\left[(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)^2\right] &\leq \frac{4}{T_e^2}\mathbb{E}\left[(\|\boldsymbol{\omega}_{T_e}\|\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|)^2\right] \\
&= \frac{4\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2}{T_e^2}\mathbb{E}\left[(\|\boldsymbol{\omega}_{T_e}\|)^2\right] \\
&= \frac{4\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 b_o^2}{T_e^2}
\end{aligned} \tag{30}
$$

$$
\begin{aligned}
\frac{4}{T_e}\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i)\right] &= \frac{4}{T_e}\mathbb{E}\left[\sum_{i=1}^{T_e-1} i\|\boldsymbol{\omega}_{T_e}\|^2\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] \\
&= \frac{4}{T_e}\sum_{i=1}^{T_e-1} i\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^2\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] \\
&= \frac{4\max_{s,s'}\|\boldsymbol{\omega}(s,s')\|^2}{T_e}\sum_{i=1}^{T_e-1} i\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] \\
&= \frac{4b_o^2\max_{s,s'}\|\boldsymbol{\omega}(s,s')\|^2}{T_e}\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{L p_\tau S}}
\end{aligned} \tag{31}
$$

where $\boldsymbol{\omega}(s, s')$ denotes the bond vector between two states s and s'.

To calculate $\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^4\right]$, we let $Z \sim \mathcal{N}(0,1)$ and using definition 1, w.l.o.g. we assume $\|\boldsymbol{\omega}_{T_e}\|^2 \sim \mathcal{N}(b_o^2, \sigma^2)$ with $\sigma < \infty$. Thus, we have

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\omega}_{T_e}\|^4\right] &= \mathbb{E}\left[\sigma^2 Z^2 + 2b_o^2\sigma Z + b_o^2\right] \\ &\underbrace{=}_{\text{Binomial Theorem and linearity of expectation}} \sigma^2 + b_o^4\end{aligned} \tag{32}$$

$$\mathbb{E}\left[\left(\boldsymbol{\omega}_{T_e}.\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\right)\right] = \mathbb{E}\left[\left(\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right)\right] = \sum_{i=1}^{T_e-1} i\mathbb{E}\left[\boldsymbol{\omega}_{T_e}.\boldsymbol{\omega}_i\right] = \sum_{i=1}^{T_e-1} ib_o^2 e^{\frac{-|T_e-i|}{Lp\tau_S}} \tag{33}$$

Substitution of the expectations in Equation (29) with Equations (32), (30), (31) and (33) gives,

$$\begin{aligned}\mathbb{E}\left[Y^2\right] \le \sigma^2 + b_o^4 + \frac{4b_o^2}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 + \frac{1}{T_e^4}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^4 + \frac{4b_o^2\max_{s,s'}\|\boldsymbol{\omega}(s,s')\|^2}{T_e}\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp\tau_S}} \\ + \frac{2b_o^2}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 + \frac{2b_o^2}{T_e^3}\sum_{i=1}^{T_e-1} ie^{\frac{-|T_e-i|}{Lp\tau_S}}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\end{aligned} \tag{34}$$

Equation (27) gives,

$$\begin{aligned}\mathbb{E}\left[Y\right]^2 = b_o^4 + \frac{4b_o^4}{T_e^2}\left(\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp\tau_S}}\right)^2 + \frac{1}{T_e^4}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^4 + \frac{2b_o^4}{T_e^2}\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp\tau_S}} \\ + \frac{2b_o^2}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 + \frac{2b_o^2}{T_e^3}\sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)|}{Lp\tau_S}}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\end{aligned} \tag{35}$$

Now to find c_1, we use Equation (26),

$$\begin{aligned}\frac{4b_o^2}{T_e^2}\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 - \frac{4b_o^4}{T_e^2}\left(\sum_{i=1}^{T_e-1} ie^{\frac{-|T_e-i|}{Lp\tau_S}}\right)^2 &= \frac{4b_o^2}{T_e^2}\left(\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2 - \left(\sum_{i=1}^{T_e-1} ie^{\frac{-|T_e-i|}{Lp\tau_S}}\right)^2\right) \\ &\le \underbrace{\frac{4b_o^2}{T_e^2}\left(\|\sum_{i=1}^{T_e-1} i\boldsymbol{\omega}_i\|^2\right)}_{B}.\end{aligned}$$

$$\frac{4b_o^2 \max_{s,s'} \|\boldsymbol{\omega}(s,s')\|^2}{T_e} \sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp_{\tau_{\mathcal{S}}}}} - \frac{2b_o^4}{T_e^2} \sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp_{\tau_{\mathcal{S}}}}}$$
$$= \left(\frac{4b_o^2 \max_{s,s'} \|\boldsymbol{\omega}(s,s')\|^2}{T_e} - \frac{2b_o^4}{T_e^2} \right) \sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp_{\tau_{\mathcal{S}}}}}$$
$$\leq \underbrace{\left(\frac{4b_o^2 \max_{s,s'} \|\boldsymbol{\omega}(s,s')\|^2}{T_e} \right) \sum_{i=1}^{T_e-1} ie^{\frac{-(T_e-i)}{Lp_{\tau_{\mathcal{S}}}}}}_{A}$$

Thus, we have

$$\frac{\mathbb{E}\left[Y^2\right]}{\mathbb{E}\left[Y\right]^2} \leq 1 + \frac{\sigma^2 + A + B}{\mathbb{E}\left[Y\right]^2} \underbrace{\leq}_{\text{by comparing A and B with (35)}} 2 = c_1 \tag{36}$$

□

2.4 The proof of Corollary 5 in the manuscript

Corollary 5 statement: Given that assumption 1 is satisfied, any exploratory trajectory induced by PolyRL algorithm (ref. Algorithm 1 in the manuscript) with high probability is an LSA-RWs.

Proof. Given Assumption 1 in the manuscript, due to the Lipschitzness of the transition probability kernel w.r.t. the action variable, the change in the distributions of the resulting states are finite and bounded by the L_2 distance of the actions. Thus, given a locally self-avoiding chain $\tau_{\mathcal{A}} \in \mathcal{A}^{T_e}$ with persistence number $Lp_{\tau_{\mathcal{A}}}$, and $\forall i \in [T_e] : b_o^2 = \mathbb{E}[\|a_i\|^2]$, by the Lipschitzness of the transition probability kernel of the underlying MDP, there exists a finite empirical average bond vector among the states visited by PolyRL (*i.e.* the first condition in Definition 1 in the manuscript is satisfied).

On the other hand, the PolyRL action sampling method (Algorithm 2) by construction preserves the expected correlation angle $\theta_{\tau_{\mathcal{A}}}$ between the consecutive selected actions with finite L_2 norm, leading to a locally self-avoiding random walk in $\mathcal{A}$. Given the following measure of spread adopted by PolyRL and defined as,

$$U_g^2(\tau_{\mathcal{S}}) := \frac{1}{T_e - 1} \sum_{s \in \tau_{\mathcal{S}}} \|s - \bar{\tau}_{\mathcal{S}}\|^2, \tag{37}$$

and the results of Theorems 3 and 4 in the manuscript (LB and UB high probability confidence bounds on the sensitivity of $U_g^2(.)$), and considering that at each time step the persistence number of the chain of visited states $Lp_{\tau_{\mathcal{S}}}$ is calculated and the exploratory action is selected such that the stiffness of $\tau_{\mathcal{S}}$ is preserved, with probability $1 - \delta$ the correlation between the bonds in $\tau_{\mathcal{S}}$ is maintained (*i.e.* the second condition in Definition 1 in the manuscript is satisfied). Hence, with probability $1 - \delta$ the chain $\tau_{\mathcal{S}}$ induced by PolyRL is locally self avoiding. □

Corollary 3. *Under assumption 1 in the manuscript, with high probability the T_e time-step exploratory chain τ induced by PolyRL with persistence number Lp_τ provides higher space coverage compared with the T_e time-step exploratory chain τ' generated by a random-walk model.*

Proof. Results from Corollary 5 in the manuscript together with remark 1 conclude the proof. □

3 Action Sampling Method

In this section, we provide the action sampling algorithm (Algrithm 2), which contains the step-by-step instruction for sampling the next action. The action sampling process is also graphically presented in Figure 2 (a).

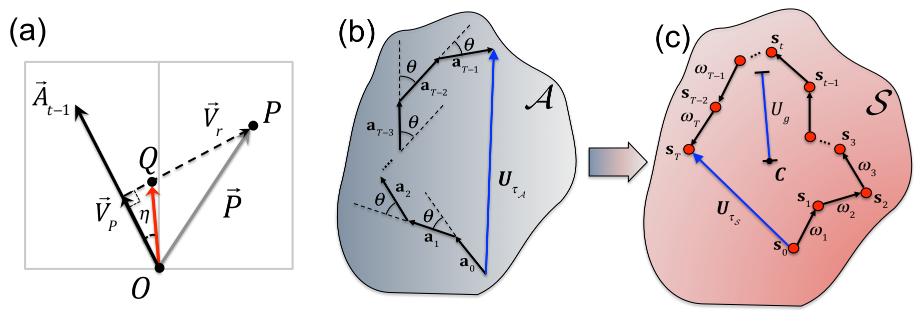

Figure 2: Schematics of the steps involved in the PolyRL exploration technique. (a) The action sampling method. In order to choose the next action $\vec{\mathbf{A}}_t$, a randomly chosen point P in $\mathcal{A}$ is projected onto the current action vector $\vec{\mathbf{A}}_{t-1}$, which gives $\vec{\mathbf{V}}_P$. The point Q is subsequently found on the vector $\vec{\mathbf{V}}_r = \vec{\mathbf{P}} - \vec{\mathbf{V}}_P$ using trigonometric relations and the angle η drawn from a normal distribution with mean θ. The resulting vector $\vec{OQ}$ (shown in red) gives the next action. Detailed instructions are given in Algorithm 2. (b) A schematic of action trajectory $\tau_\mathcal{A}$ with the mean correlation angle θ between every two consecutive bond vectors and the end-to-end vector $\mathbf{U}_{\tau_\mathcal{A}}$. (c) A schematic of state trajectory $\tau_\mathcal{S}$ with bond vectors $\boldsymbol{\omega}_i = \mathbf{s}_i - \mathbf{s}_{i-1}$. The radius of gyration and the end-to-end vector are depicted as U_g and $\mathbf{U}_{\tau_\mathcal{S}}$, respectively. Point $\mathbf{C}$ is the center of mass of the visited states.

Algorithm 2 Action Sampling

Require: Angle η and Previous action $\mathbf{A}_{t-1}$

Draw a random point P in the action space $(P_i \sim \mathcal{U}[-m, m]; i = 1, \ldots d)$ ▷ $\mathbf{P}$ is the vector from the origin to the point P

$D = \mathbf{A}_{t-1}.\mathbf{P}$

$\mathbf{V}p = \dfrac{D}{||\mathbf{A}_{t-1}||_2^2}\mathbf{A}_{t-1}$ ▷ The projection of $\mathbf{P}$ on $\mathbf{A}_{t-1}$

$\mathbf{V}r = \mathbf{P} - \mathbf{V}p$

$l = ||\mathbf{V}p||_2 \tan \eta$

$k = l/||\mathbf{V}r||_2$

$\mathbf{Q} = k\mathbf{V}r + \mathbf{V}p$

if $D > 0$ **then**

    $\mathbf{A}_t = \mathbf{Q}$

else

    $\mathbf{A}_t = -\mathbf{Q}$

end if

Clip $\mathbf{A}_t$ if out of action range

return $\mathbf{A}_t$

4 Additional Baseline Results

In this section, we provide the benchmarking results for DDPG-UC, DDPG-OU, DDPG-PARAM, DDPG-FiGAR (Figure 3), as well as SAC and OAC (Figure 4) algorithms on three standard MuJoco tasks. Moreover, the source code is provided here.

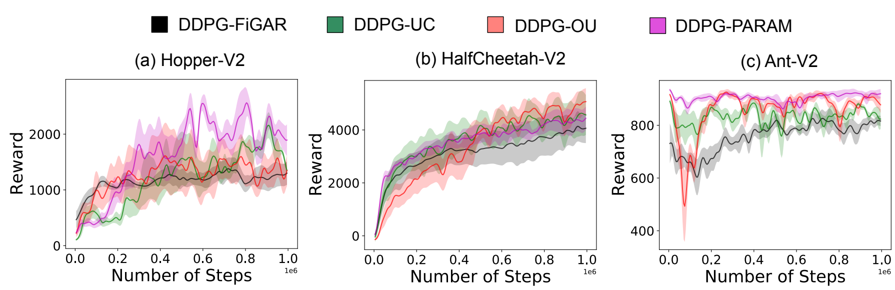

Figure 3: Performance of DDPG-UC, DDPG-OU, DDPG-PARAM, and DDPG-FiGAR algorithms across 3 MuJoCo domains. The plots are averaged over 5 random seeds. The test evaluation happens every 5k over 1 million time steps.

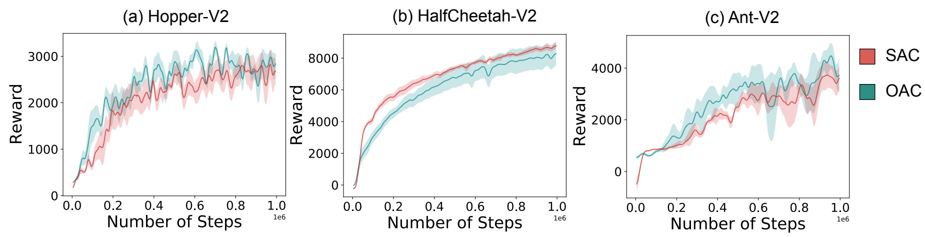

Figure 4: Performance of SAC and OAC algorithms across 3 MuJoCo domains. The plots are averaged over 5 random seeds. The test evaluation happens every 5k steps over 1 million time steps.

In order to Benchmark DDPG-FiGAR results, we let the action repetition set, defined as $W := \{1, 2, \ldots |W|\}$ ([3]), be equal to $\{1\}$. The results are expected to converge to those of DDPG-OU noise as depicted in Figure 5.

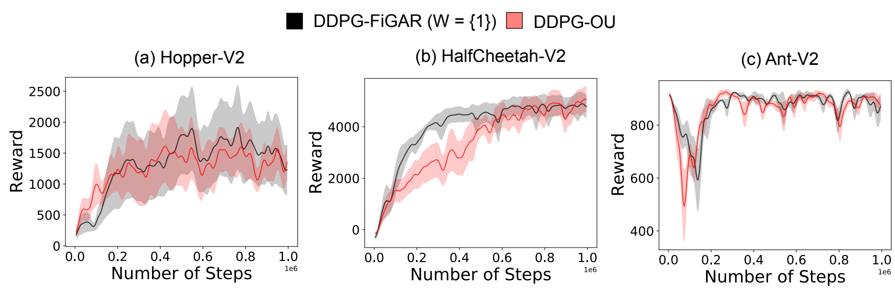

Figure 5: Benchmarking DDPG-FiGAR against DDPG-OU using action repetition set $W = \{1\}$ across 3 MuJoCo domains. The plots are averaged over 5 random seeds. The test evaluation happens every 5k steps over 1 million time steps.

5 Hyperparameters and Network Architecture

In this section, we provide the architecture of the neural networks (Table 1), as well as the PolyRL hyper parameters (Table 2) used in the experiments. Regarding the computing infrastructure, the experiments were run on a slurm-managed cluster with NVIDIA P100 Pascal (12G HBM2 memory) GPUs. The avergae run-time for DDPG-based and SAC-based models were around 8 and 12 hours, respectively.

Table 1: DDPG and SAC Network Architecture

Parameter	**Value**	
Optimizer	Adam	
Critic Learning Rate	1e−3 (DDPG)	3e−4 (SAC)
Actor Learning Rate	1e−4 (DDPG)	3e−4 (SAC)
Discount Factor	0.99	
Replay Buffer Size	1e+6	
Number of Hidden Layers (All Networks)	2	
Number of Units per Layer	400 (1^{st})- 300 (2^{nd}) (DDPG)	both 256 (SAC)
Number of Samples per Mini Batch	100	
Nonlinearity	ReLU	
Target Network Update Coefficient	5e−3	
Target Update Interval	1	

The exploration factor - The one important parameter in the PolyRL exploration method, which controls the exploration-exploitation trade-off is the exploration factor $\beta \in [0, 1]$. The factor β plays the balancing role in two ways: controlling (1) the range of confidence interval (Equations (7) and (11) in the manuscript; $\delta = 1 - e^{-\beta N}$); and (2) the probability of switching from the target policy π_μ to the behaviour policy π_{PolyRL}. Figure 6 illustrates the effect of varying β on the performance of a DDPG-PolyRL agent in the HalfCheetah-v2 environment. The heat maps (Figures 6 (a), (b) and (c)) show the average asymptotic reward obtained for different pairs of correlation angle θ and variance σ^2. The heat maps depict that for this specific task, the performance of DDPG-PolyRL improves as β changes from 0.0004 to 0.01. The performance plot for the same task (Figure 6 (d)) shows the effect of β on the amount of the obtained reward. The relation of β with the percentage of the moves taken using the target policy is illustrated in Figure 6 (e)). As expected, larger values of β lead to more exploitation and fewer exploratory steps.

Table 2: PolyRL Hyper parameters. Note that the parameters θ and σ are angles and their respective values in the table are in radian.

	Mean Correlation Angle $\boldsymbol{\theta}$	**Variance** $\boldsymbol{\sigma^2}$	**Exploration Factor** $\boldsymbol{\beta}$
DDPG-PolyRL			
SparseHopper-V2 ($\lambda = 0.1$)	0.035	0.00007	0.001
SparseHalfCheetah-V2 ($\lambda = 5$)	0.17	0.017	0.02
SparseAnt-V2 ($\lambda = 0.15$)	0.087	0.035	0.01
SAC-PolyRL			
SparseHopper-V2 ($\lambda = 3$)	0.35	0.017	0.01
SparseHalfCheetah-V2 ($\lambda = 15$)	0.35	0.00007	0.05
SparseAnt-V2 ($\lambda = 3$)	0.035	0.00007	0.01

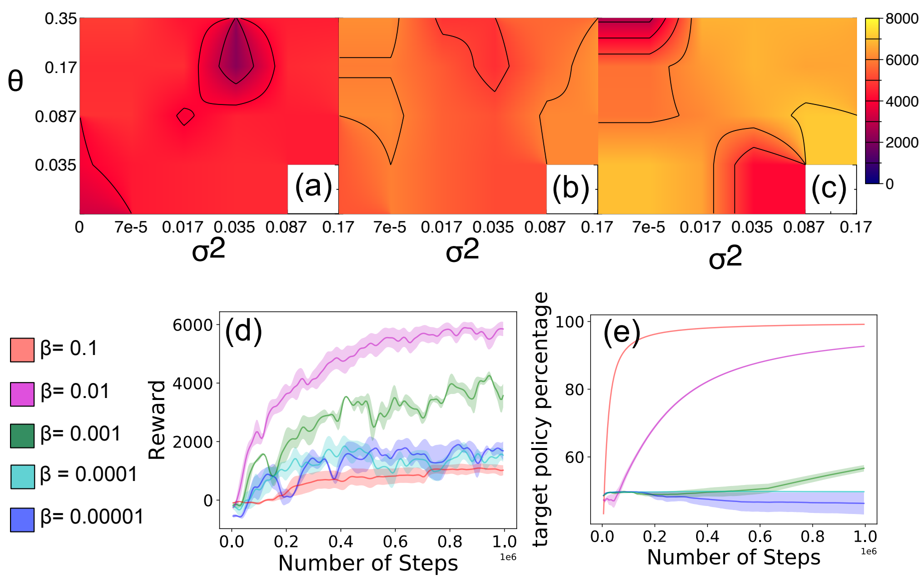

Figure 6: Performance of DDPG-PolyRL in HalfCheetah-v2 for different values of exploration factor β. (a-c) Heat maps depict the mean of the obtained asymptotic rewards after 3 million time steps over a range of correlation angle θ and the variance σ^2. The results are shown for $\beta = 0.0004$ (a), $\beta = 0.001$ (b), and $\beta = 0.01$ (c). (d) Performance of DDPG-PolyRL in HalfCheetah-v2 for the fixed values of $\theta = 0.035$ and $\sigma^2 = 0.00007$, and different values of β. (e) The percentage of the movements the DDPG-PolyRL agent behaves greedily. All values are averaged over four random seeds and the error bars show the standard error on the mean.

References

[1] M. Doi and S. F. Edwards. *The theory of polymer dynamics*, volume 73. oxford university press, 1988.

[2] M. Rubenstein and R. Colby. Polymer physics: Oxford university press, 2003.

[3] S. Sharma, A. S. Lakshminarayanan, and B. Ravindran. Learning to repeat: Fine grained action repetition for deep reinforcement learning. *arXiv preprint arXiv:1702.06054*, 2017.