KIT
Karlsruhe Institute of Technology

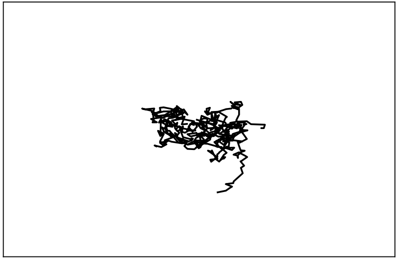
ALR
Autonomous Learning Robots Lab

A Temporally Coherent Policy for Reinforcement Learning

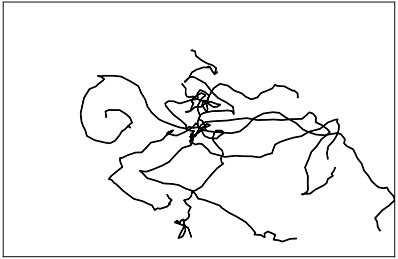
Ricardo Dominguez-Olmedo | October 29, 2021

Effective exploration and temporal coherence

- Effective exploration is required to learn optimal behaviours.
- Temporally coherent action trajectories result in more effective exploration.



Uncoherent exploration



Coherent exploration

Gaussian policy

- Stochastic, $\pi(a_t|s_t) = \mathcal{N}\left(a_t; \mu(s_t), \Sigma(s_t)\right)$
- Actions are not very correlated... $a_t = \mu(s_t) + L(s_t)\epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, I), \quad \epsilon_t \perp \epsilon_{t+k}\, \forall k \neq 0$
- ... particularly in the initial learning stages $\mathbb{C}\left[a_t, a_{t+k}\right] = \mathbb{C}\left[\mu_0 + L_0 e_t, \mu_0 + L_0 e_{t+k}\right] = 0 \quad \forall k \neq 0$

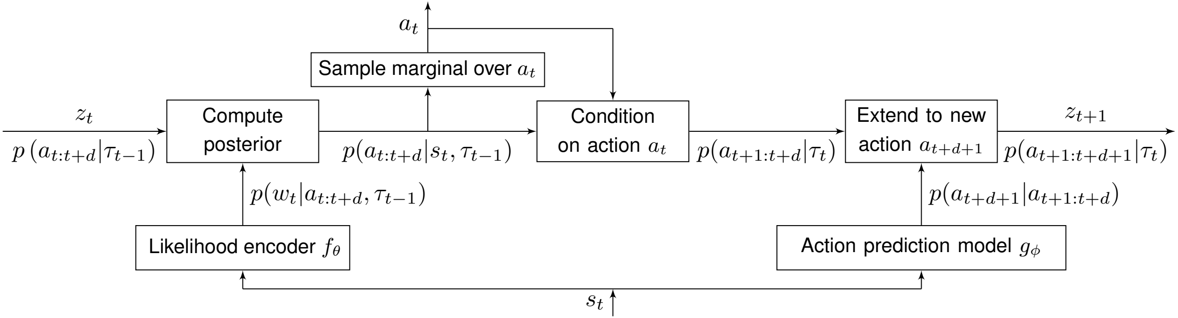
A Gaussian distribution over multiple actions

- Plan d steps ahead $\pi(a_{t:t+d}|s_t, \tau_{t-1}) = \mathcal{N}\left(a_{t:t+d}; \mu, \Sigma\right)$
- Assume perfectly planning... $\pi(a_{t:t+d}|\tau_{t-1}) = \pi(a_{t:t+d}|s_t, \tau_{t-1})$
- ... then $a_{t+k} \sim \mathcal{N}\left(\mu_k, \Sigma_{kk}\right)$ and $\mathbb{C}\left[a_t, a_{t+k}\right] = \Sigma^t_{1(k+1)}, \; 1 \leq k \leq d$
- Let $\Sigma_{mn} = \alpha^{|m-n|}\sqrt{\Sigma_{mm} \odot \Sigma_{nn}}$ with parameter $\alpha \in (0, 1)$...
- Then the correlation coefficient between two actions is $\rho_{a_t, a_{t+k}} = \alpha^{|k|} \quad 1 < k \leq d$

A Gaussian distribution over multiple actions

- Plan d steps ahead $\quad \pi(a_{t:t+d}|s_t, \tau_{t-1}) = \mathcal{N}(a_{t:t+d}; \mu, \Sigma)$
- Assume perfectly planning... $\pi(a_{t:t+d}|\tau_{t-1}) = \pi(a_{t:t+d}|s_t, \tau_{t-1})$
- ... then $a_{t+k} \sim \mathcal{N}(\mu_k, \Sigma_{kk})$ and $\mathbb{C}[a_t, a_{t+k}] = \Sigma^t_{1(k+1)}, \quad 1 \leq k \leq d$
- Let $\Sigma_{mn} = \alpha^{|m-n|}\sqrt{\Sigma_{mm} \odot \Sigma_{nn}}$ with parameter $\alpha \in (0, 1)$...
- Then the correlation coefficient between two actions is $\rho_{a_t, a_{t+k}} = \alpha^{|k|} \quad 1 < k \leq d$

Our proposed policy



Training the policy

$$\max_{\theta,\phi} \ \mathrm{RL}_{\text{loss}} - \mathbb{E}_t \left[\lambda_1 P_t^{(1)} + \lambda_2 P_t^{(2)} \right] \tag{1}$$

- Any policy search method which admits recurrent policies could be used (e.g. PPO)
- Penalize large updates to the prior $\quad P_t^{(1)} = \mathsf{KL}\left(p(a_{t:t+d} | s_t, \tau_{t-1}) \,||\, p(a_{t:t+d} | \tau_{t-1})\right)$
- Regularize the posterior variance $\quad P_t^{(2)} = \mathsf{KL}\left(\mathcal{N}\left(\mu_t^+, \Sigma_t^+\right) || \, \mathcal{N}\left(\mu_t^+, \Sigma_t^*\right)\right)$

  such that actions are sufficiently correlated $\quad \Sigma_{mn}^* = \alpha^{|m-n|} \sqrt{\Sigma_{mm}^+ \odot \Sigma_{nn}^+} \quad m \neq n$

Experimental setting

- Simple continuous control environment.
- Linear dynamics

$$\begin{bmatrix} x_{t+1} \\ v_{t+1} \end{bmatrix} = \begin{bmatrix} I & \Delta t \\ 0 & I \end{bmatrix} \begin{bmatrix} x_t \\ v_t \end{bmatrix} + \begin{bmatrix} 0 \\ \Delta t \end{bmatrix} a_t \tag{2}$$

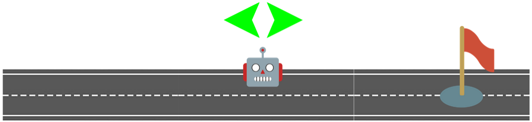
- Reward function

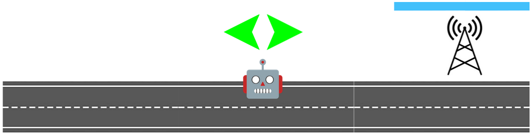
$$r(x_t, v_t, a_t) = \begin{cases} -0.01 a_t & \text{if } t \neq T \\ -v_t - \mathcal{R}_T(x_t) & \text{if } t = T \end{cases} \tag{3}$$

- Two settings: denser reward and semi-sparse

$$\mathcal{R}_T^{\text{DENSE}}(x_T) = x_T - \hat{x}_T \tag{4}$$

$$\mathcal{R}_T^{\text{SPARSE}}(x_T) = \max\{x_T - \hat{x}_T, D_{\max}\} \tag{5}$$



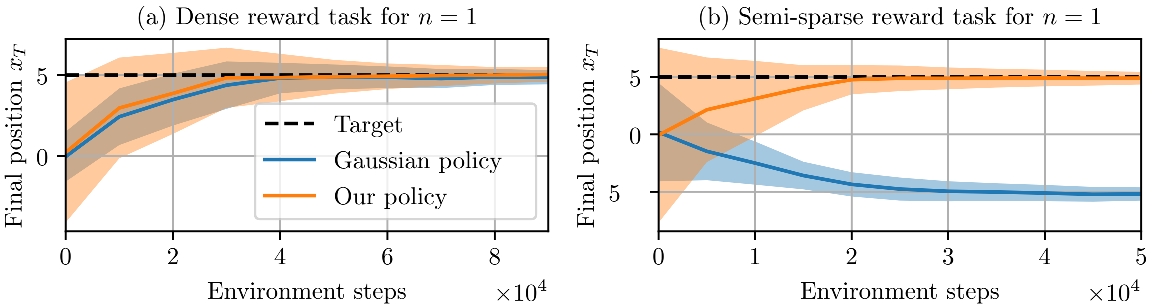


Results: temporal coherence

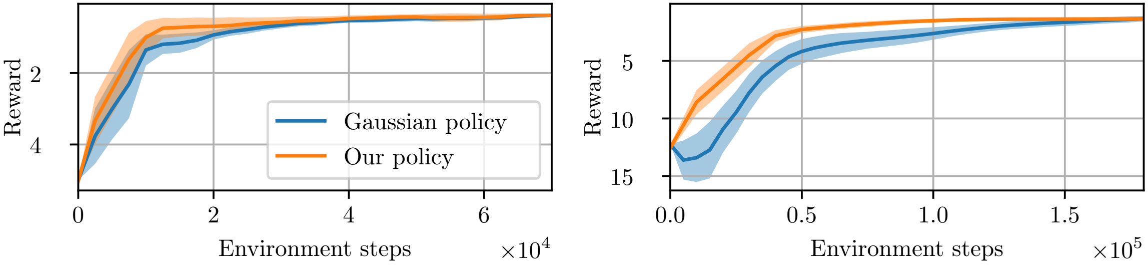
Average Pearson correlation coefficient between subsequent actions throughout policy training.

	Dense reward task		Semi-sparse reward task	
	$n = 1$	$n = 6$	$n = 1$	$n = 6$
Gaussian policy	0.12 ± 0.02	0.12 ± 0.02	0.08 ± 0.01	0.09 ± 0.01
Our policy	**0.20 ± 0.09**	**0.80 ± 0.12**	**0.25 ± 0.18**	**0.63 ± 0.07**

Results: exploration effectiveness



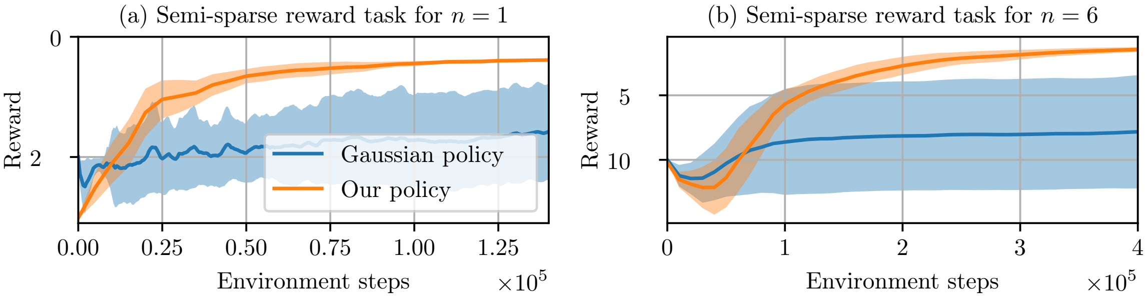
Results: dense reward setting



(a) Dense reward task for $n = 1$

(b) Dense reward task for $n = 6$

Results: semi-sparse reward setting



Limitations

- Right amount of correlation α is task-specific.
- Policy training is...
  - Very fragile to choice of hyperparameters.
  - Numerically unstable for large planing-horizon d or time-horizon d.
  - Computationally expensive.

Future work

- Consider more complex environments.
- Compare with previous approaches on coherent exploration.
- Policy with a latent representation to avoid matrix inverses.

Conclusion

- We propose a recurrent policy parametrizing a distribution over future actions.
- The policy is regularized such that contiguous actions are sufficiently correlated.
- For the environment considered:
  - More effective exploration.
  - More sample-efficient.
  - Better asymptotic performance.

Thank you

Questions?

Related work on temporally coherent exploration

- Cummulative sum of policy output.
- Motion primitives.
- Hierarchical RL.
- Episode-based exploration in parameter space.
- Step-based exploration in action space.
- Model-based planning.