Accelerating Reinforcement Learning with Skill Priors

Karl Pertsch, Youngwoon Lee, Joseph J. Lim

Accelerating Reinforcement Learning with Skill Priors

Karl Pertsch, Youngwoon Lee, Joseph J. Lim

Motivation

- The goal of our work is to leverage prior experience for accelerated learning of downstream tasks

Approach

- We present SPiRL, an approach for leveraging large, unstructed datasets to accelerate downstream learning of unseen tasks
  - We propose a deep latent variable model that jointly learns an embedding space of skills and a prior over these skills from offline data
  - We then extend maximum-entropy RL algorithms to incorporate both skill embedding and skill prior for efficient downstream learning

Results & New finding

- We show that learned skill priors accelerate learning of new tasks across three simulated navigation and robot manipulation tasks

Discussion & Comments

- Future work can combine learned skill priors with methods for extracting semantic skills of flexible length from unstructed data

Background

- (Problem) The majority of this data (about robot) is unstructed, without clear task or reward definitions, making it difficult to use for learning new tasks
- (Solution) One flexible way to utilize unstructed prior experience is by extracting skills, temporally extended actions that represent useful behaviors, which can be repurposed to solve downstream tasks

- (Problem) The large skill libraries extracted from rich datasets can, somewhat paradoxically, lead to worse learning efficiency on the downstream task
- (Solution) The key idea of this work is to learn prior over skills along with the skill library to guide the exploration in skill space and enable efficient downstream learning, even with the large skill spaces

Suggestion

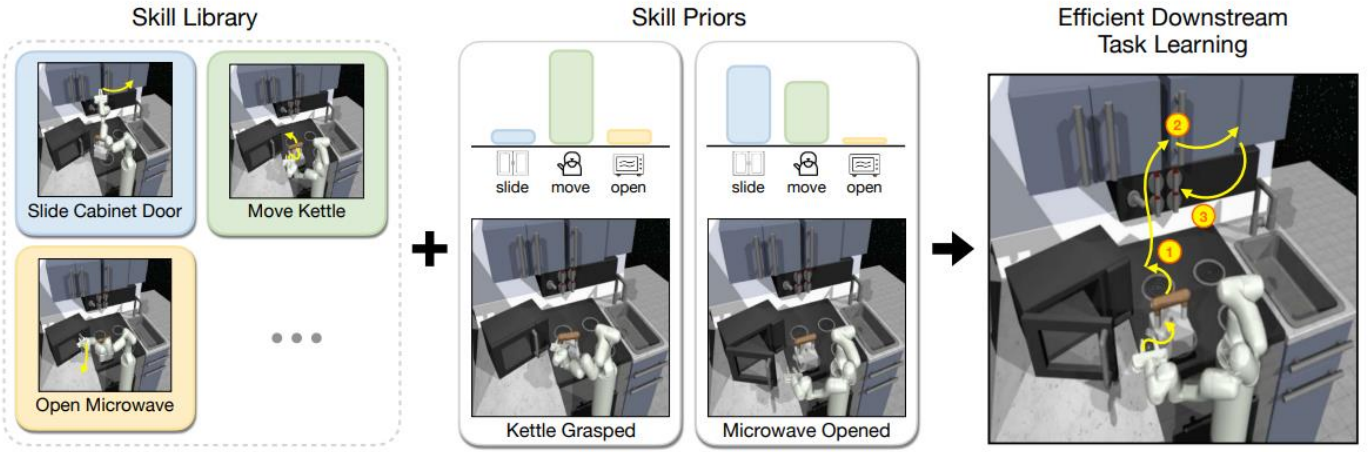

Figure 1: Intelligent agents can use a large library of acquired skills when learning new tasks. Instead of exploring skills uniformly, they can leverage *priors over skills* as guidance, based e.g., on the current environment state. Such priors capture which skills are promising to explore, like moving a kettle when it is already grasped, and which are less likely to lead to task success, like attempting to open an already opened microwave. In this work, we propose to jointly learn an embedding space of skills and a prior over skills from unstructured data to accelerate the learning of new tasks.

- (1) Intuitively, the prior over skills is not uniform (current state can hint which skills are promising to explore)
  - we design a stochastic latent variable model that learns a continuous embedding space of skills and a prior distribution over these skills from unstructed agent experience

Suggestion

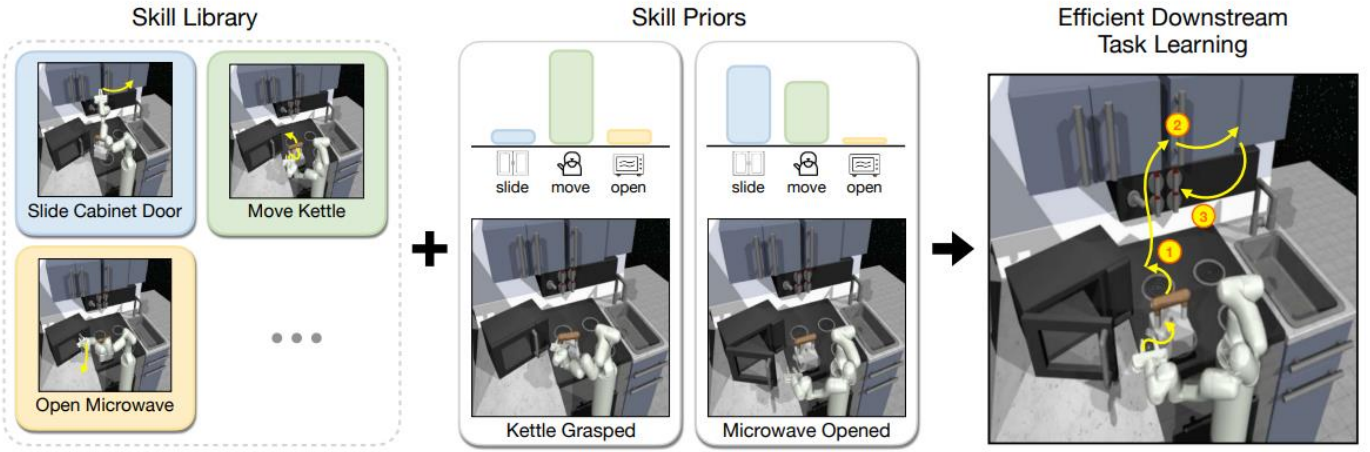

Figure 1: Intelligent agents can use a large library of acquired skills when learning new tasks. Instead of exploring skills uniformly, they can leverage *priors over skills* as guidance, based e.g., on the current environment state. Such priors capture which skills are promising to explore, like moving a kettle when it is already grasped, and which are less likely to lead to task success, like attempting to open an already opened microwave. In this work, we propose to jointly learn an embedding space of skills and a prior over skills from unstructured data to accelerate the learning of new tasks.

- (2) We then show how to naturally incorporate the learned skill prior into maximum-entropy RL algorithms for efficient learning of downstream tasks

1) Meta-learning approaches

- They aim to extract useful priors from previous experience to improve the learning efficiency for unseen tasks
- (limitation) they require a defined set of training tasks and online data collection during pre-training and therefore cannot leverage large offline datasets

2) Offline reinforcement learning

- (limitation) these approaches require the experience to be annotated with rewards for the downstream task, which are challenging to provide for large, real-world datasets, especially when the experience is collected across a wide range of tasks

3) Transferring skills between tasks

- When using powerful latent variable models, these approaches are able to represent a very large number of skills in a compact embedding space
- (limitation) the exploration of such a rich skill embedding space can be challenging, leading to inefficient downstream task learning

4) Learned behavior priors

- are commonly used to guide task learning in offline RL approaches in order to avoid value overestimation for actions outside of the training data distribution

We decompose the problem of prior-guided skill transfer into two sub-problems

- (1) the extraction of skill embedding and skill prior from offline data
- (2) the prior-guided learning of downstream tasks with a hierarchical policy

(1) Learning Continuous Skill Embedding and Skill Prior

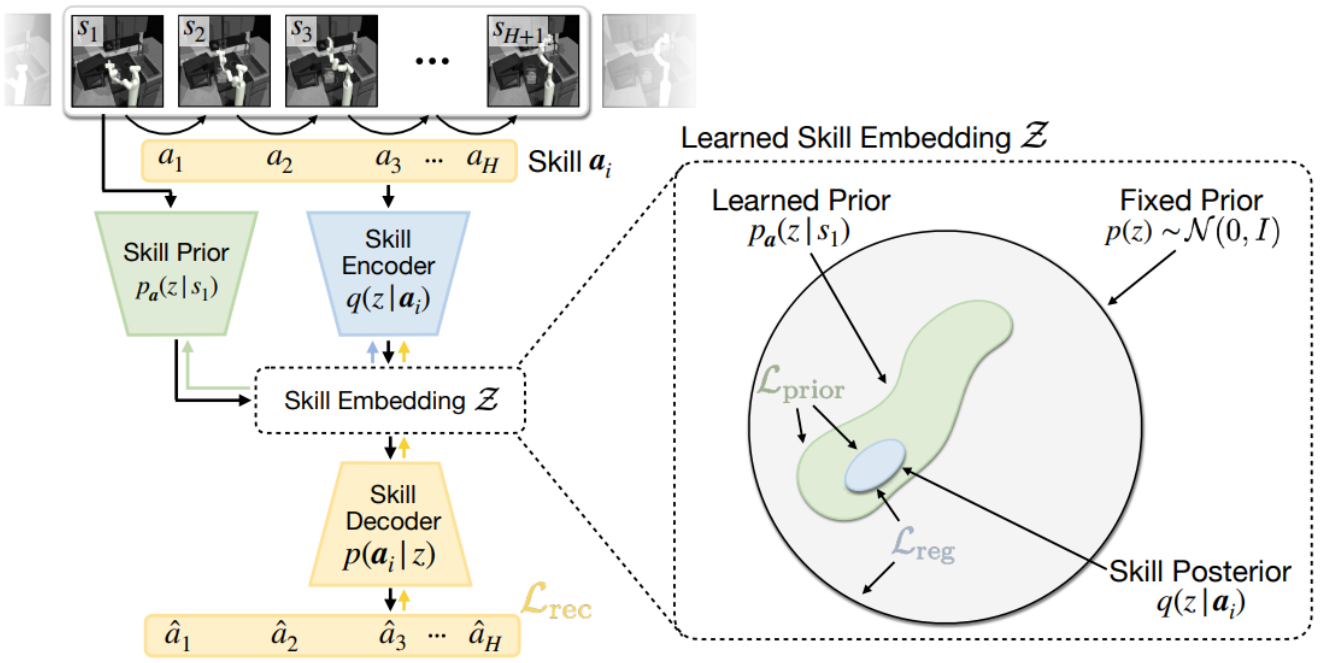

Figure 2: Deep latent variable model for joint learning of skill embedding and skill prior. Given a state-action trajectory from the dataset, the skill encoder maps the action sequence to a posterior distribution $q(z|\boldsymbol{a}_i)$ over latent skill embeddings. The action trajectory gets reconstructed by passing a sample from the posterior through the skill decoder. The skill prior maps the current environment state to a prior distribution $p_{\boldsymbol{a}}(z|s_1)$ over skill embeddings. Colorful arrows indicate the propagation of gradients from reconstruction, regularization and prior training objectives.

(1) Learning Continuous Skill Embedding and Skill Prior

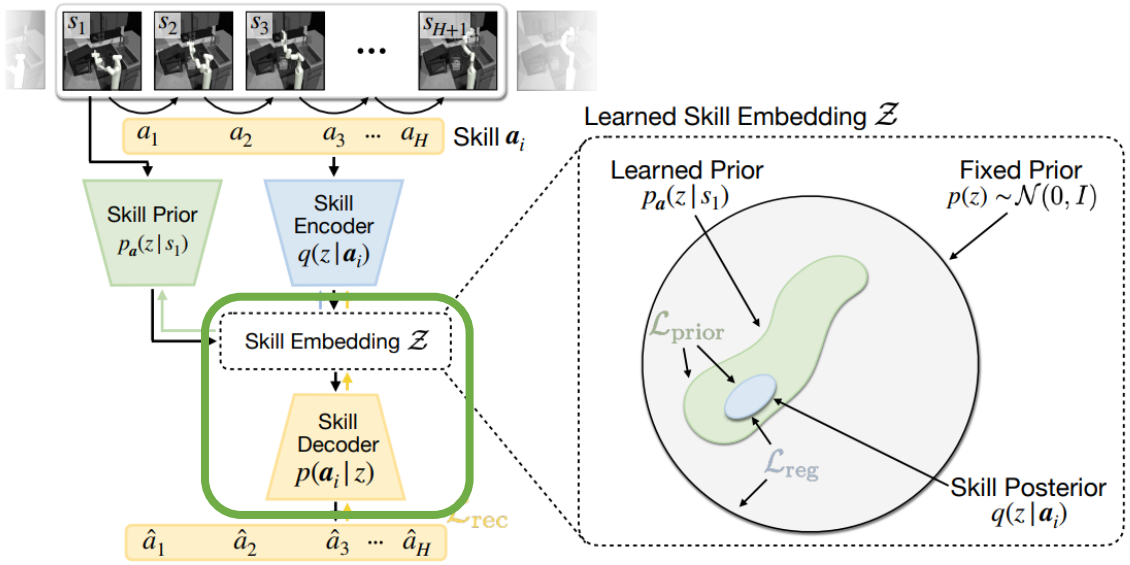

Figure 2: Deep latent variable model for joint learning of skill embedding and skill prior. Given a state-action trajectory from the dataset, the skill encoder maps the action sequence to a posterior distribution $q(z|\boldsymbol{a}_i)$ over latent skill embeddings. The action trajectory gets reconstructed by passing a sample from the posterior through the skill decoder. The skill prior maps the current environment state to a prior distribution $p_{\boldsymbol{a}}(z|s_1)$ over skill embeddings. Colorful arrows indicate the propagation of gradients from reconstruction, regularization and prior training objectives.

- (1) To learn a low-dimensional skill embedding space Z, we train a stochastic latent variable model $\boldsymbol{p(a_i|z)}$ of skills using the offline dataset

(1) Learning Continuous Skill Embedding and Skill Prior

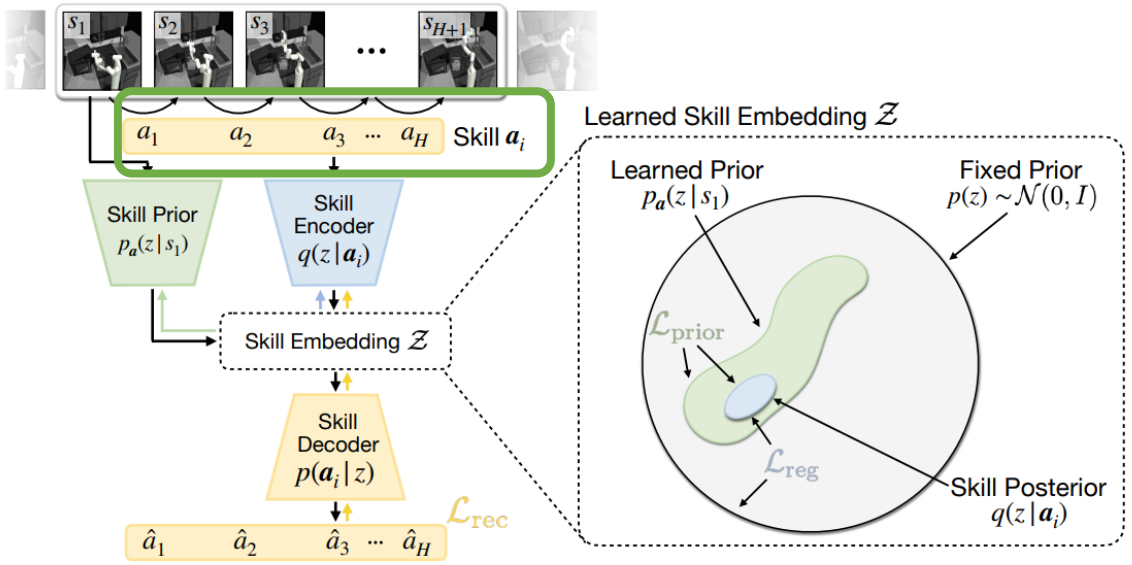

Figure 2: Deep latent variable model for joint learning of skill embedding and skill prior. Given a state-action trajectory from the dataset, the skill encoder maps the action sequence to a posterior distribution $q(z|\boldsymbol{a}_i)$ over latent skill embeddings. The action trajectory gets reconstructed by passing a sample from the posterior through the skill decoder. The skill prior maps the current environment state to a prior distribution $p_{\boldsymbol{a}}(z|s_1)$ over skill embeddings. Colorful arrows indicate the propagation of gradients from reconstruction, regularization and prior training objectives.

- (2) We randomly sample H-step trajectories from the training sequences and maximize the following evidence lower bound (ELBO)

Skill Decoder Skill Encoder Skill Prior

$$\log p(\boldsymbol{a}_i) \geq \mathbb{E}_q \Big[\underbrace{\log p(\boldsymbol{a}_i|z)}_{\text{reconstruction}} - \beta \Big(\underbrace{\log q(z|\boldsymbol{a}_i)}_{\text{regularization}} - \log p(z) \Big) \Big]. \quad (1)$$

(1) Learning Continuous Skill Embedding and Skill Prior

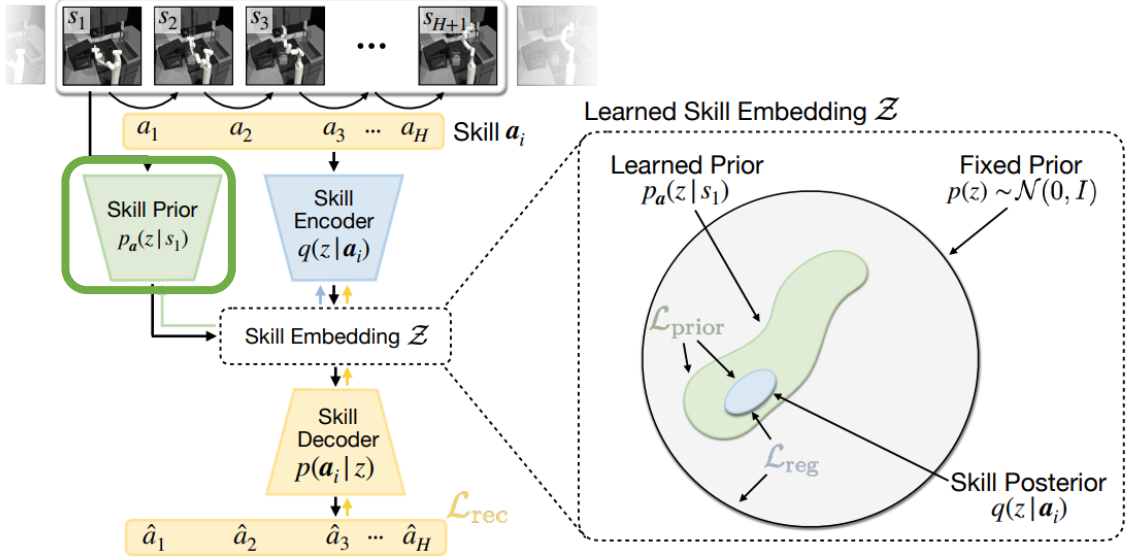

Figure 2: Deep latent variable model for joint learning of skill embedding and skill prior. Given a state-action trajectory from the dataset, the skill encoder maps the action sequence to a posterior distribution $q(z|\boldsymbol{a}_i)$ over latent skill embeddings. The action trajectory gets reconstructed by passing a sample from the posterior through the skill decoder. The skill prior maps the current environment state to a prior distribution $p_{\boldsymbol{a}}(z|s_1)$ over skill embeddings. Colorful arrows indicate the propagation of gradients from reconstruction, regularization and prior training objectives.

- (3) To better guide downstream learning, we learn a prior over skills along with the skill embedding model
  - We therefore introduce another component in our model: the skill prior $\boldsymbol{p_a(z|\cdot)}$
  - In this work we focus on learning a state-conditioned skill prior $\boldsymbol{p_a(z|s_t)}$
- To train the skill prior we minimize the Kullback-Leibler divergence between the predicted prior and the inferred skill posterior: $\mathbb{E}_{(s,\boldsymbol{a}_i)\sim\mathcal{D}} D_{\mathrm{KL}}\left(q(z|\boldsymbol{a}_i), p_{\boldsymbol{a}}(z|s_t)\right)$

Jointly optimize the skill embedding model and the skill prior

(2) Skill Prior Regularized Reinforcement Learning

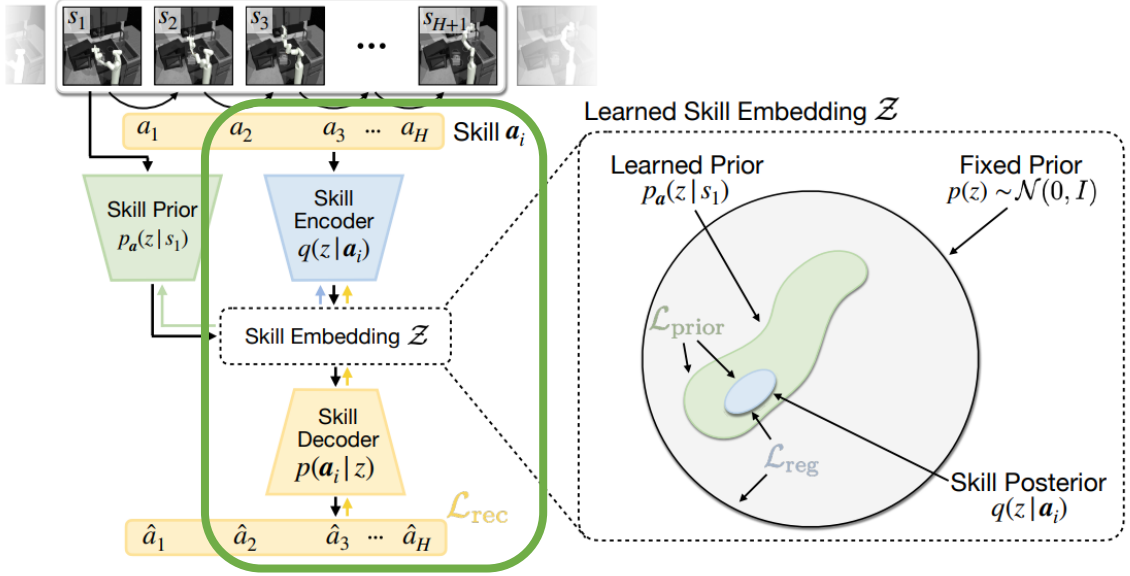

Figure 2: Deep latent variable model for joint learning of skill embedding and skill prior. Given a state-action trajectory from the dataset, the skill encoder maps the action sequence to a posterior distribution $q(z|\boldsymbol{a}_i)$ over latent skill embeddings. The action trajectory gets reconstructed by passing a sample from the posterior through the skill decoder. The skill prior maps the current environment state to a prior distribution $p_{\boldsymbol{a}}(z|s_1)$ over skill embeddings. Colorful arrows indicate the propagation of gradients from reconstruction, regularization and prior training objectives.

- To use the learned skill embedding for downstream task learning, we employ a hierarchical policy learning scheme by using the skill embedding space as action space of a high-level policy
  - (1) we learn a policy $\pi_\theta(\boldsymbol{z}|\boldsymbol{s_t})$ that outputs skill embeddings
  - (2) we decode skill embeddings into action sequences using the learned skill decoder $\{a_t^i, \ldots, a_{t+H-1}^i\} \sim p(\boldsymbol{a}_i|z)$
  - (3) We **execute these actions** for H steps before sampling the next skill from the high-level policy

(2) Skill Prior Regularized Reinforcement Learning

- When using a large offline dataset D with diverse behaviors, the number of embedded skills can grow rapidly, leading to a challenging exploration problem when training the high-level policy
  - → we propose to use the learned skill prior to guide the high-level policy

(2) Skill Prior Regularized Reinforcement Learning

How the skill prior can be naturally integrated into maximum-entropy RL algorithms?

- Maximum entropy RL augments the training objective of the policy with a term that encourages maximization of the policy's entropy along with the return
- (1) Before
  - The added entropy term is equivalent to the negated KL divergence between the policy and a uniform action prior $U(a_t)$: $\mathcal{H}(\pi(a_t|s_t)) = -\mathbb{E}_\pi\big[\log \pi(a_t|s_t)\big] \propto -D_{\mathrm{KL}}(\pi(a_t|s_t), U(a_t))$

$$J(\theta) = \mathbb{E}_\pi\left[\sum_{t=1}^{T} \gamma^t r(s_t, a_t) + \alpha \mathcal{H}(\pi(a_t|s_t))\right] \quad (2)$$

Entropy Term

- (2) After
  - We aim to regularize the policy towards a non-uniform, learned skill prior to guide exploration in skill space
    - → We can therefore replace the entropy term with the negated KL divergence from the learned prior

$$J(\theta) = \mathbb{E}_\pi\left[\sum_{t=1}^{T} \tilde{r}(s_t, z_t) - \alpha D_{\mathrm{KL}}(\pi(z_t|s_t), p_{\boldsymbol{a}}(z_t|s_t))\right] \quad (3)$$

Negated KL Divergence

Summary

Algorithm 1 SPiRL: Skill-Prior RL

Inputs: H-step reward function $\tilde{r}(s_t, z_t)$, discount γ, target divergence δ, learning rates $\lambda_\pi, \lambda_Q, \lambda_\alpha$, target update rate τ.
Initialize replay buffer $\mathcal{D}$, high-level policy $\pi_\theta(z_t|s_t)$, critic $Q_\phi(s_t, z_t)$, target network $Q_{\bar{\phi}}(s_t, z_t)$
for each iteration **do**
  for every H environment steps **do**
    $z_t \sim \pi(z_t|s_t)$ ▷ sample skill from policy
    $s_{t'} \sim p(s_{t+H}|s_t, z_t)$ ▷ execute skill in environment
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{s_t, z_t, \tilde{r}(s_t, z_t), s_{t'}\}$ ▷ store transition in replay buffer
  for each gradient step **do**
    $\bar{Q} = \tilde{r}(s_t, z_t) + \gamma \left[Q_{\bar{\phi}}(s_{t'}, \pi_\theta(z_{t'}|s_{t'})) - \alpha D_{\text{KL}}\left(\pi_\theta(z_{t'}|s_{t'}), p_{\boldsymbol{a}}(z_{t'}|s_{t'})\right)\right]$ ▷ compute Q-target
    $\theta \leftarrow \theta - \lambda_\pi \nabla_\theta \left[Q_\phi(s_t, \pi_\theta(z_t|s_t)) - \alpha D_{\text{KL}}(\pi_\theta(z_t|s_t), p_{\boldsymbol{a}}(z_t|s_t))\right]$ ▷ update policy weights
    $\phi \leftarrow \phi - \lambda_Q \nabla_\phi \left[\frac{1}{2}\left(Q_\phi(s_t, z_t) - \bar{Q}\right)^2\right]$ ▷ update critic weights
    $\alpha \leftarrow \alpha - \lambda_\alpha \nabla_\alpha \left[\alpha \cdot (D_{\text{KL}}(\pi_\theta(z_t|s_t), p_{\boldsymbol{a}}(z_t|s_t)) - \delta)\right]$ ▷ update alpha
    $\bar{\phi} \leftarrow \tau\phi + (1-\tau)\bar{\phi}$ ▷ update target network weights
return trained policy $\pi_\theta(z_t|s_t)$

Experiments Goal

- (1) Can we leverage unstructured datasets to accelerate downstream task learning by transferring skills?
- (2) Can learned skill priors improve exploration during downstream task learning?
- (3) Are learned skill priors necessary to scale skill transfer to large datasets?

Conclusions

- We propose a deep latent variable model that jointly learns an embedding space of skills and a prior over these skills from offline data
- We then extend maximum-entropy RL algorithms to incorporate both skill embedding and skill prior for efficient downstream learning

Results

- We evaluate SPiRL on challenging simulated navigation and robotic manipulation tasks and show that both, skill embedding and skill prior are essential for effective transfer from rich datasets

Limitation and future work

- Future work can combine learned skill priors with methods for extracting semantic skills of flexible length from unstructured data
- Further, skill priors are important in safety-critical applications, like autonomous driving, where random exploration is dangerous
  - Skill priors learned e.g. from human demonstration, can guide exploration to skills that do not endanger the learner or other agents