Hierarchical RL and Skill Discovery

CS 330
Reminders

Monday (Nov 8\textsuperscript{th}):

Homework 4 (optional) is due

Wednesday (Nov 10\textsuperscript{th}):

Project milestone is due

Two guest lectures:

- Colin Raffel on big language models (virtual)
- Jascha Sohl-Dickstein on learning optimizers
Recall: RL so far

We knew what we wanted
Short-horizon behaviors
Well defined tasks/rewards
Why Skill Discovery?

What if we want to discover interesting behaviors?

[The construction of movement by the spinal cord, *Tresch et al.*, 1999]

Why Skill Discovery? More practical version

Coming up with tasks is tricky...

Task ideas for a tabletop manipulation scenario

[Meta-World, Yu, Quillen, He, Julian, et al., 2019]
Why Hierarchical RL?

Performing tasks at various levels of abstractions

Bake a cheesecake
Buy ingredients
Go to the store
Walk to the door
Take a step

Contract muscle X

Exploration
The Plan

Information-theoretic concepts

Skill discovery

Using discovered skills

Hierarchical RL
The Plan

Information-theoretic concepts

Skill discovery

Using discovered skills

Hierarchical RL
Entropy

\[ p(x) \] distribution (e.g., over observations \( x \))

\[ \mathcal{H}(p(x)) = -E_{x \sim p(x)}[\log p(x)] \]

entropy – how “broad” \( p(x) \) is

Slide adapted from Sergey Levine
KL-divergence

Distance between two distributions

$$D_{KL}(q||p) = E_q \left[ \log \frac{q(x)}{p(x)} \right] = E_q \log q(x) - E_q \log p(x) = -E_q \log p(x) - H(q(x))$$
Mutual information

\[ I(x; y) = D_{KL}(p(x, y) || p(x)p(y)) \]
\[ = E_{(x,y) \sim p(x,y)} \left[ \log \frac{p(x, y)}{p(x)p(y)} \right] \]
\[ = \mathcal{H}(p(y)) - \mathcal{H}(p(y|x)) = \mathcal{H}(p(x)) - \mathcal{H}(p(x|y)) \]

High MI?

- it rains tomorrow, y – streets are wet tomorrow
- it rains tomorrow, y – we find life on Mars tomorrow

Slide adapted from Sergey Levine
Mutual information

\[ I(x; y) = D_{KL}(p(x, y) \| p(x)p(y)) \]
\[ = E_{(x,y) \sim p(x,y)} \left[ \log \frac{p(x, y)}{p(x)p(y)} \right] \]
\[ = H(p(y)) - H(p(y|x)) = H(p(x)) - H(p(x|y)) \]

example of mutual information: “empowerment” (Polani et al.)

\[ I(s_{t+1}; a_t) = H(s_{t+1}) - H(s_{t+1}|a_t) \]
The Plan

Information-theoretic concepts

Skill discovery

Using discovered skills

Hierarchical RL
Soft Q-learning

Objective:

\[
\sum_t E(s_t, a_t) \sim \rho \left[ r(s_t, a_t) + \mathcal{H}(q(a_t | s_t)) \right]
\]

\[
\pi(a_t | s_t) = \mathcal{N}(\mu(s_t), \Sigma)
\]

\[
Q(s_t, a_t)
\]

\[
\pi(a_t | s_t) \propto \exp Q(s_t, a_t)
\]

Q-learning

1. collect dataset \( \{(s_i, a_i, s'_i, r_i)\} \)
2. set \( y_i \leftarrow r(s_i, a_i) + \gamma \max_{a'_i} Q(\phi)(s'_i, a'_i) \)
3. set \( \phi \leftarrow \arg\min_\phi \frac{1}{K} \sum_i \| Q(\phi)(s_i, a_i) - y_i \|^2 \)

\[
\pi(a | s) = \arg\max_a Q(\phi)(s, a)
\]

Soft Q-learning

1. collect dataset \( \{(s_i, a_i, s'_i, r_i)\} \)
2. set \( y_i \leftarrow r(s_i, a_i) + \gamma \max_{a'_i} Q(\phi)(s'_i, a'_i) \)
3. set \( \phi \leftarrow \arg\min_\phi \frac{1}{K} \sum_i \| Q(\phi)(s_i, a_i) - y_i \|^2 \)

\[
\pi(a | s) = \arg\max_a \exp \left( \frac{1}{\zeta} \sum_t \phi(s_t, a_t) \right)
\]
Soft Q-learning

\[ \pi(a_t | s_t) = \mathcal{N}(\mu(s_t), \Sigma) \]

Exploration

\[ Q(s_t, a_t) \]

Fine-tunability

\[ \pi(a_t | s_t) \propto \exp Q(s_t, a_t) \]

Robustness

Haarnoja et al. RL with Deep Energy-Based Policies, 2017
Learning diverse skills

\[ \pi(a|s, z) \]

Why can’t we just use MaxEnt RL

1. **action** entropy is not the same as **state** entropy
   - agent can take very different actions, but land in similar states
2. MaxEnt policies are stochastic, but not always **controllable**
   - intuitively, we want **low** diversity for a fixed \( z \), high diversity across \( z \)’s

**Intuition:** different **skills** should visit different **state-space regions**

Eysenbach, Gupta, Ibarz, Levine. *Diversity is All You Need.*

Slide adapted from Sergey Levine
Diversity-promoting reward function

\[ \pi(a|s, z) = \arg\max_\pi \sum_z E_{s \sim \pi(s|z)} [r(s, z)] \]

reward states that are unlikely for other \( z' \neq z \)

\[ r(s, z) = \log p(z|s) \]

Eysenbach, Gupta, Ibarz, Levine. Diversity is All You Need. Slide adapted from Sergey Levine
Examples of learned tasks

Cheetah

Ant

Mountain car

Eysenbach, Gupta, Ibarz, Levine. *Diversity is All You Need.*
A connection to mutual information

\[ \pi(a|s, z) = \arg \max_{\pi} \sum_{z} E_{s \sim \pi(s|z)}[r(s, z)] \]

\[ r(s, z) = \log p(z|s) \]

\[ I(z, s) = H(z) - H(z|s) \]

maximized by using uniform prior \( p(z) \)

minimized by maximizing \( \log p(z|s) \)

Eysenbach, Gupta, Ibarz, Levine. Diversity is All You Need.
See also: Gregor et al. Variational Intrinsic Control. 2016

Slide adapted from Sergey Levine
The Plan

Information-theoretic concepts

Skill discovery

Using discovered skills

Hierarchical RL
How to use learned skills?

\[ \pi(a|s, z) \]

How can we use the learned skills to accomplish a task?

Learn a policy that operates on z's

Eysenbach, Gupta, Ibarz, Levine. *Diversity is All You Need.*
Results: hierarchical RL

Can we do better?

Eysenbach, Gupta, Ibarz, Levine. Diversity is All You Need.
What’s the problem?

Skills might not be particularly useful

It’s not very easy to use the learned skills

What makes a useful skill?
What’s the problem?

Consequences are **hard** to predict

Consequences are **easy** to predict
Slightly different mutual information

\[ I(z, s) = H(z) - H(z|s) \]

\[ \max \mathcal{I}(s', z | s) = \max \left( \mathcal{H}(s' | s) - \mathcal{H}(s' | s, z) \right) \]

\[ \mathcal{I}(x; y | z) \quad p(x, y | z) \quad p(x | z) \quad p(y | z) \]

Future hard to predict for different skills
Predictable future for a given skill

\[ I(s'; z | s) \geq \mathbb{E}_s \mathbb{E}_z \mathbb{E}_{p(s'|s, z)} \left[ \log \frac{q_\phi(s'|s, z)}{p(s'|s)} \right] \]

\[ \approx \mathbb{E}_s \mathbb{E}_z \mathbb{E}_{p(s'|s, z)} \left[ \log \frac{q_\phi(s'|s, z)}{\sum_{i=1}^L q_\phi(s'|s, z_i)} + \log L \right] \]

Sharma, Gu, Levine, Kumar, Hausman, DADS, 2019.
Skill-dynamics model

We are learning a skill-dynamics model \( q(s' \mid s, z) \)

compared to conventional global dynamics \( p(s' \mid s, a) \)

Skills are optimized specifically to make skill-dynamics easier to model

Sharma, Gu, Levine, Kumar, Hausman, DADS, 2019.
Algorithm 1: Dynamics-Aware Discovery of Skills (DADS)

Initialize $\pi, q_\phi$; 
while not converged do 
    Sample a skill $z \sim p(z)$ every episode; 
    Collect new $M$ on-policy samples; 
    Update $q_\phi$ using $K_1$ steps of gradient descent on $M$ transitions; 
    Compute $r_z(s, a, s')$ for $M$ transitions; 
    Update $\pi$ using any RL algorithm; 
end
Using learned skills

Use skill-dynamics for model-based planning
Plan for skills not actions
Tasks can be learned zero-shot
Summary

- Two skill discovery algorithms that use mutual information
- Predictability can be used as a proxy for “usefulness”
- Method that optimizes for both, predictability and diversity
- Model-based planning in the skill space
- Opens new avenues such as unsupervised meta-RL
  - Gupta et al. *Unsupervised Meta-Learning for RL*, 2018
The Plan

Information-theoretic concepts

Skill discovery

Using discovered skills

Hierarchical RL
Why Hierarchical RL?

Performing tasks at various levels of abstractions

Bake a cheesecake
Buy ingredients
Go to the store
Walk to the door
Take a step

Contract muscle X
Hierarchical RL – design choices

Design choices:
- goal-conditioned vs not
- pre-trained vs e2e
- self-terminating vs fixed rate
- on-policy vs off-policy
Learning Locomotor Controllers

Design choices:
- goal-conditioned vs not
- pre-trained vs e2e
- self-terminating vs fixed rate
- on-policy vs off-policy

Option Critic

A Markovian option $\omega \in \Omega$ is a triple $(I_\omega, \pi_\omega, \beta_\omega)$ in which $I_\omega \subseteq S$ is an initiation set, $\pi_\omega$ is an intra-option policy, and $\beta_\omega : S \rightarrow [0, 1]$ is a termination function. We also assume that $\forall s \in S, \forall \omega \in \Omega : s \in I_\omega$ (i.e., all options are available everywhere).

- Option is a self-terminating mini-policy
- Everything trained together with policy gradient

Design choices:
- goal-conditioned vs not
- pre-trained vs e2e
- self-terminating vs fixed rate
- on-policy vs off-policy

Relay Policy Learning

Design choices:
- goal-conditioned vs not
- pre-trained vs e2e
- self-terminating vs fixed rate
- on-policy vs off-policy

Relay Policy Learning

Design choices:
- goal-conditioned vs not
- pre-trained vs end-to-end
- self-terminating vs fixed rate
- on-policy vs off-policy

Goal-conditioned policies with relabeling
- Demonstrations to pre-train everything
- On-policy

Design choices:
- goal-conditioned vs not
- pre-trained vs e2e
- self-terminating vs fixed rate
- on-policy vs off-policy
HRL Summary

- Multiple design choices and frameworks
- Helps with exploration and temporally extended tasks
- Can be difficult to get it to work
- Seems like a natural direction for harder RL problems

Design choices:
- goal-conditioned vs not
- pre-trained vs e2e
- self-terminating vs fixed rate
- on-policy vs off-policy

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Experiments</th>
<th>Important?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H1) Temporal training</td>
<td>Figures 2, 3</td>
<td>Yes, but only for the use of multi-step rewards (n-step returns).</td>
</tr>
<tr>
<td>(H2) Temporal exploration</td>
<td>Figures 2, 4</td>
<td>Yes, and this is important even for non-hierarchical exploration.</td>
</tr>
<tr>
<td>(H3) Semantic training</td>
<td>Figure 3</td>
<td>No.</td>
</tr>
<tr>
<td>(H4) Semantic exploration</td>
<td>Figure 4</td>
<td>Yes, and this is important even for non-hierarchical exploration.</td>
</tr>
</tbody>
</table>

Figure 5: A summary of our conclusions on the benefits of hierarchy.

The Plan

Information-theoretic concepts

Skill discovery

Using discovered skills

Hierarchical RL
Reminders

Monday (Nov 8th):
Homework 4 (optional) is due

Wednesday (Nov 10th):
Project milestone is due

Next week:
Two guest lectures:
- Colin Raffel on big language models (virtual)
- Jascha Sohl-Dickstein on learning optimizers
The week after next

Can the agent **learn continuously** over their life-time?

Lifelong learning – Nov 15th