

UC Berkeley·CS285 | Deep Reinforcement Learning (2020)

CS285 (2020)· 课程资料包 @ShowMeAI



视频

中英双语字幕



课件

一键打包下载



笔记

官方笔记翻译



代码

作业项目解析



视频·B 站 [扫码或点击链接]

<https://www.bilibili.com/video/BV12341167kL>



课件 & 代码·博客 [扫码或点击链接]

<http://blog.showmeai.tech/cs285>

Berkeley

actor-critic

reinforcement learning

exploration

元学习

exploitation

梯度策略

探索与利用

DQN

迁移学习

逆强化学习

Q-learning / Q 学习

deep q network

meta learning

Awesome AI Courses Notes Cheatsheets 是 [ShowMeAI](#) 资料库的分支系列，覆盖最具知名度的 **TOP20+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets· 持续更新中			
知识图谱	图机器学习	深度强化学习	自动驾驶
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094



微信公众号

资料下载方式 2: 扫码点击**底部菜单栏**

称为 **AI 内容创作者**? 回复 [添砖加瓦]

Exploration (Part 2)

CS 285

Instructor: Sergey Levine
UC Berkeley

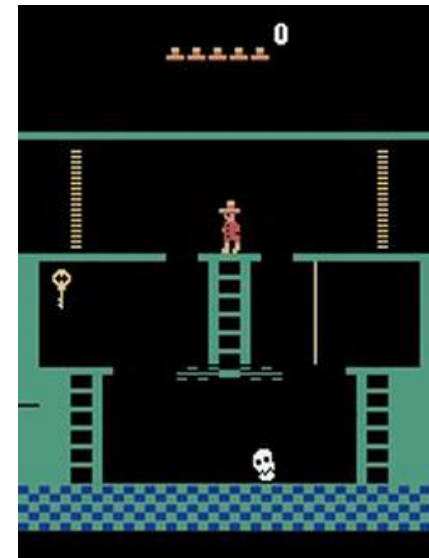


Recap: what's the problem?

this is easy (mostly)



this is impossible



Why?

Unsupervised learning of diverse behaviors

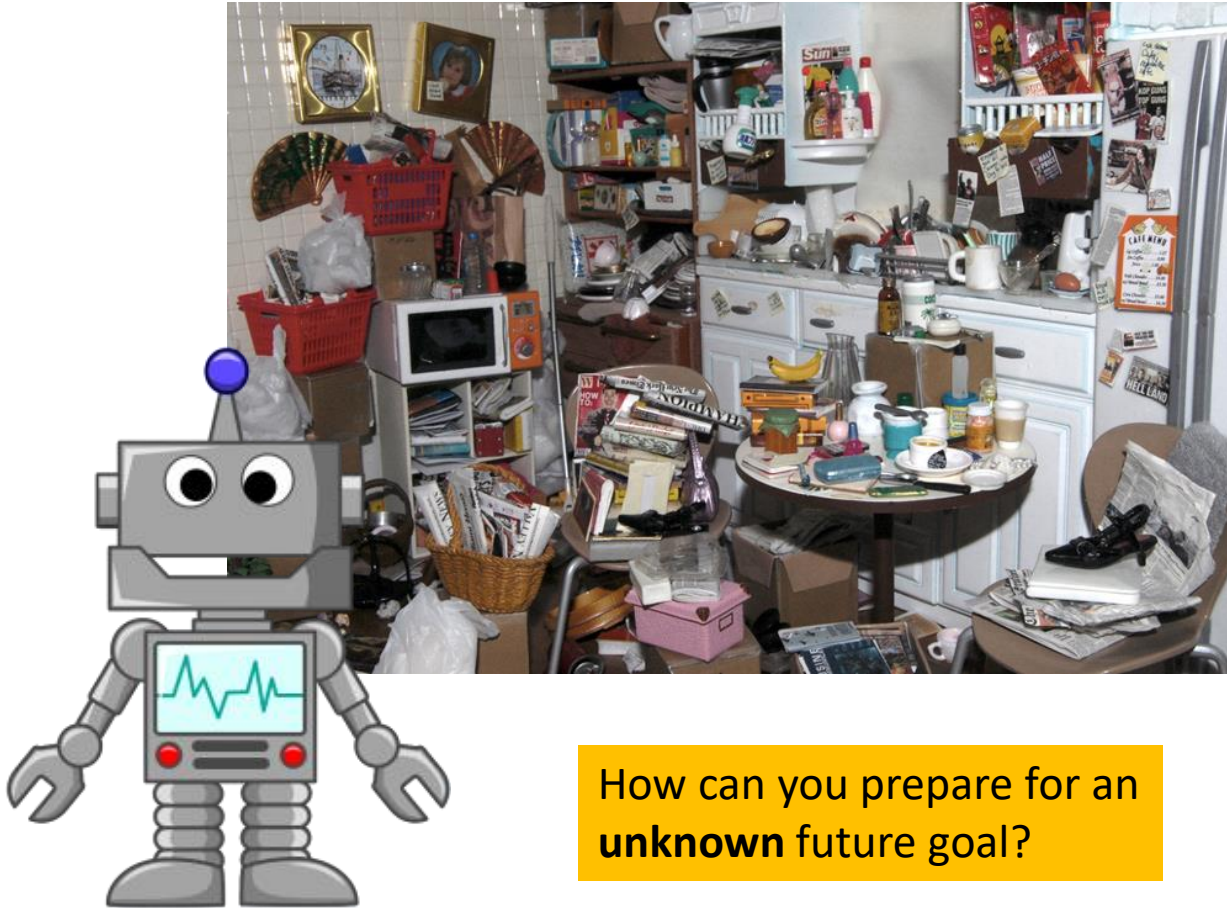
What if we want to recover diverse behavior **without any reward function at all?**



Why?

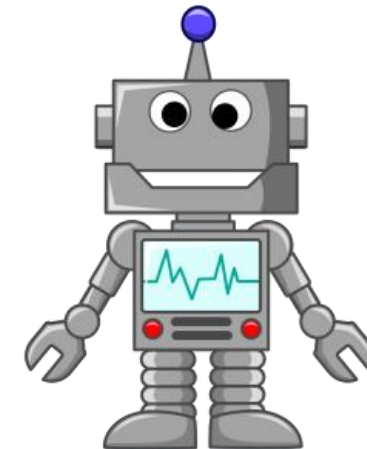
- *Learn skills without supervision, then use them to accomplish goals*
- *Learn sub-skills to use with hierarchical reinforcement learning*
- *Explore the space of possible behaviors*

An Example Scenario



How can you prepare for an **unknown** future goal?

training time: unsupervised



In this lecture...

- Definitions & concepts from information theory
- Learning without a reward function by reaching goals
- A *state distribution-matching* formulation of reinforcement learning
- Is coverage of valid states a *good* exploration objective?
- Beyond state covering: covering the *space of skills*

In this lecture...

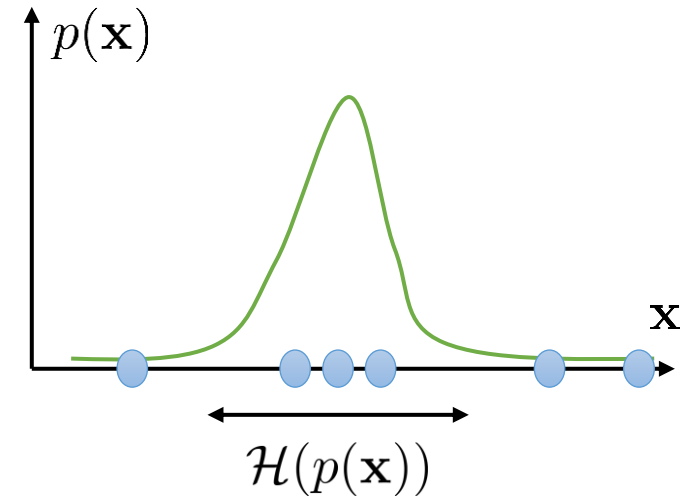
- Definitions & concepts from information theory
- Learning without a reward function by reaching goals
- A *state distribution-matching* formulation of reinforcement learning
- Is coverage of valid states a *good* exploration objective?
- Beyond state covering: covering the *space of skills*

Some useful identities

$p(\mathbf{x})$ distribution (e.g., over observations \mathbf{x})

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

entropy – how “broad” $p(\mathbf{x})$ is



Some useful identities

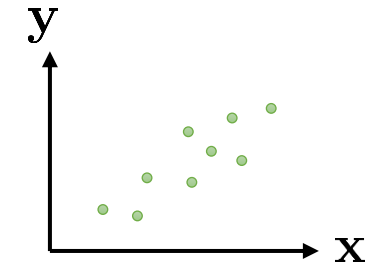
entropy – how “broad” $p(\mathbf{x})$ is

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

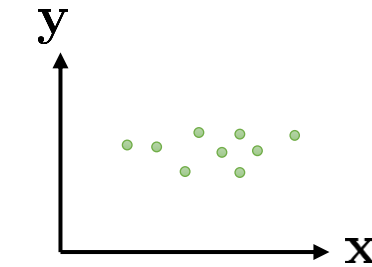
$$\mathcal{I}(\mathbf{x}; \mathbf{y}) = D_{\text{KL}}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y}))$$

$$= E_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} \left[\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right]$$

$$= \mathcal{H}(p(\mathbf{y})) - \mathcal{H}(p(\mathbf{y}|\mathbf{x}))$$



high MI: \mathbf{x} and \mathbf{y} are *dependent*



low MI: \mathbf{x} and \mathbf{y} are *independent*

Information theoretic quantities in RL

$\pi(\mathbf{s})$ state *marginal* distribution of policy π

$\mathcal{H}(\pi(\mathbf{s}))$ state *marginal* entropy of policy π  quantifies *coverage*

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

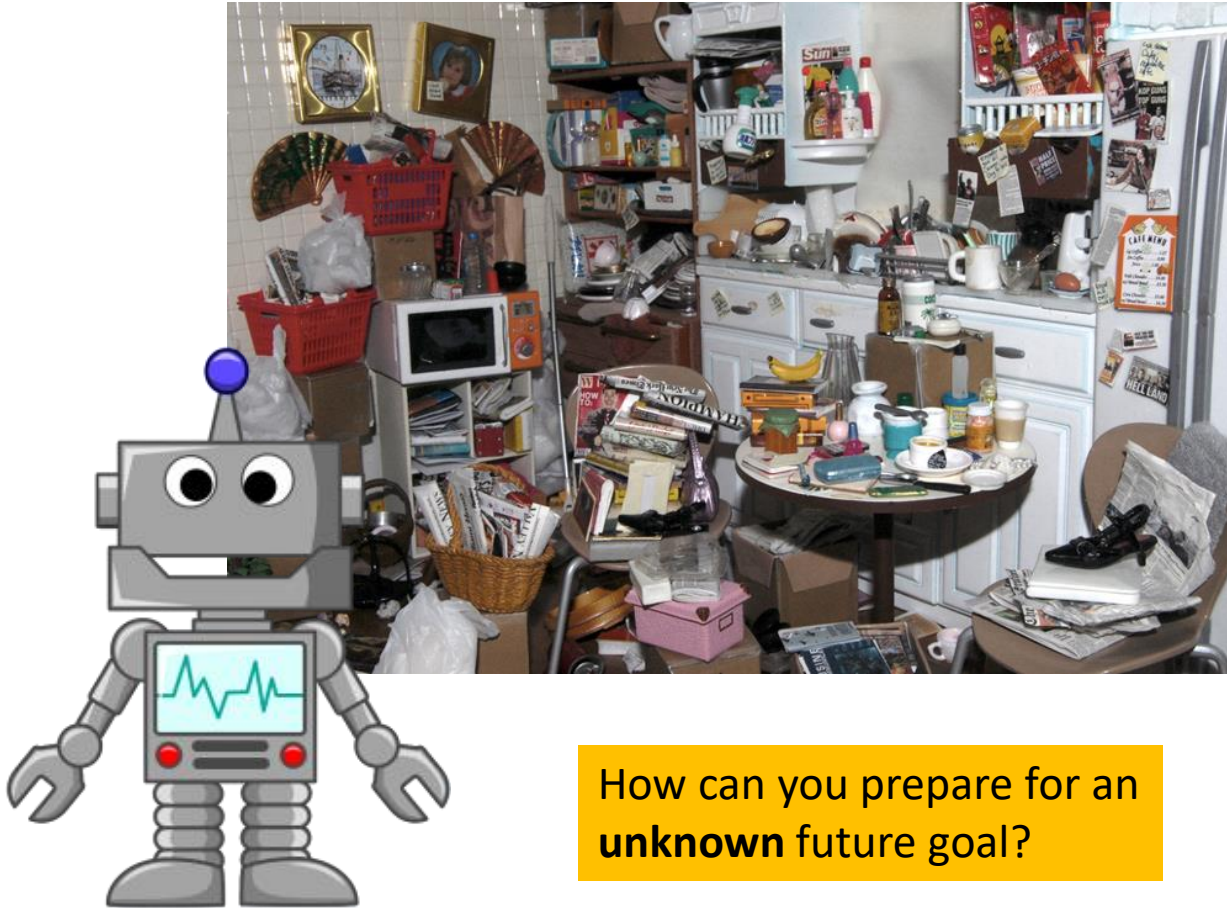
$$\mathcal{I}(\mathbf{s}_{t+1}; \mathbf{a}_t) = \mathcal{H}(\mathbf{s}_{t+1}) - \mathcal{H}(\mathbf{s}_{t+1} | \mathbf{a}_t)$$

can be viewed as quantifying “control authority” in an information-theoretic way

In this lecture...

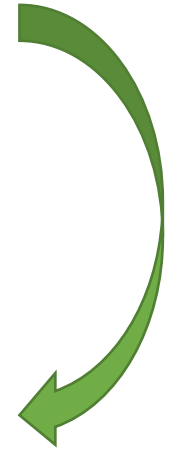
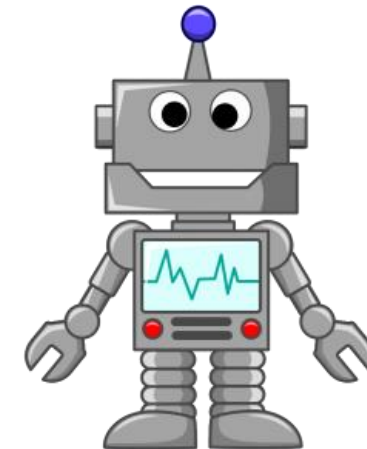
- Definitions & concepts from information theory
- Learning without a reward function by reaching goals
- A *state distribution-matching* formulation of reinforcement learning
- Is coverage of valid states a *good* exploration objective?
- Beyond state covering: covering the *space of skills*

An Example Scenario



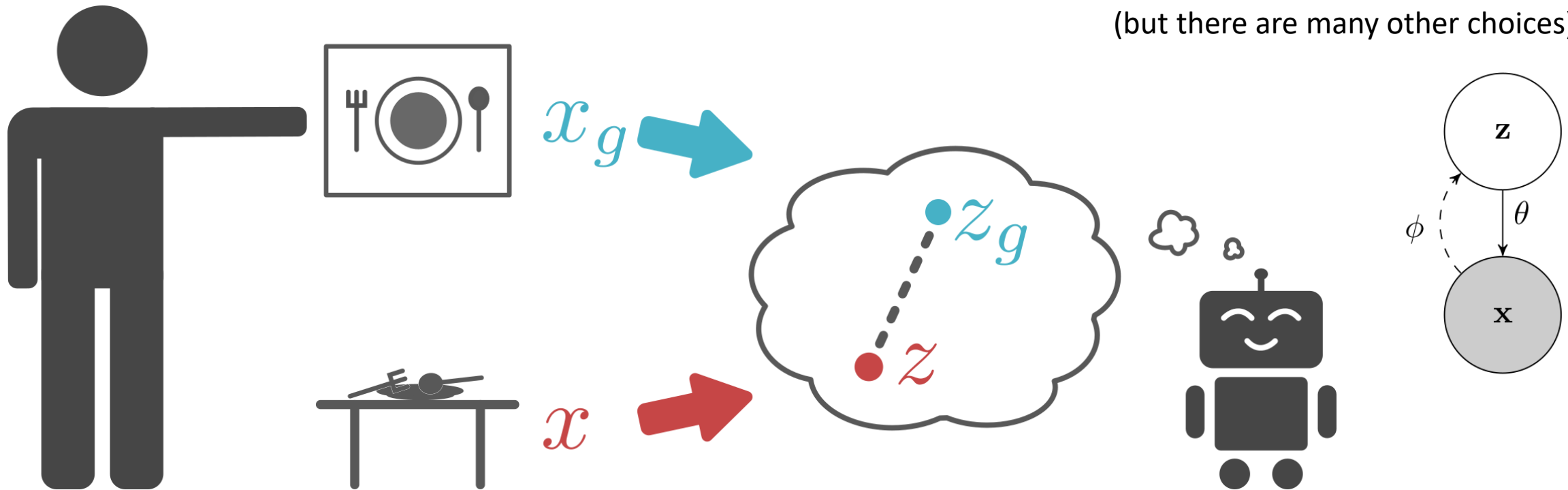
How can you prepare for an **unknown** future goal?

training time: unsupervised

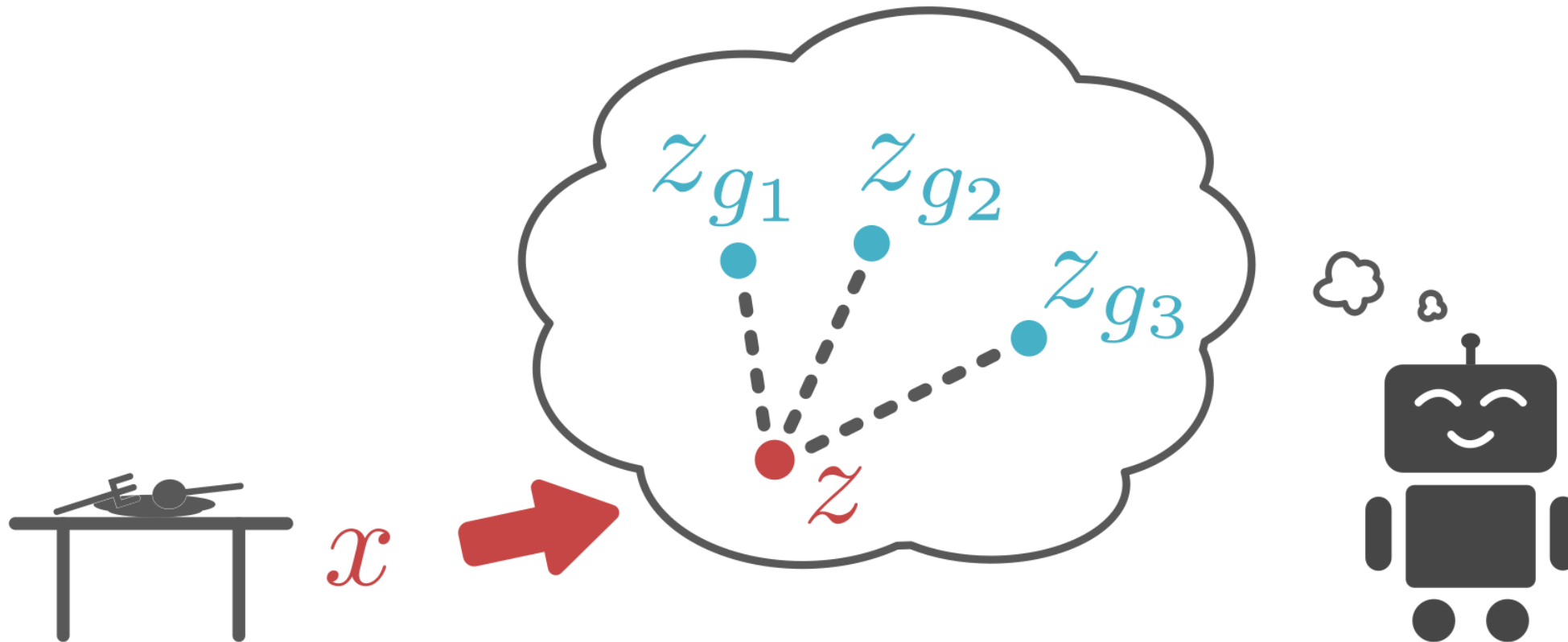


Learn without any rewards at all

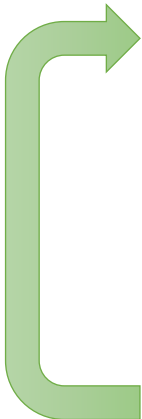
VAE (Kingma & Welling '13)
(but there are many other choices)

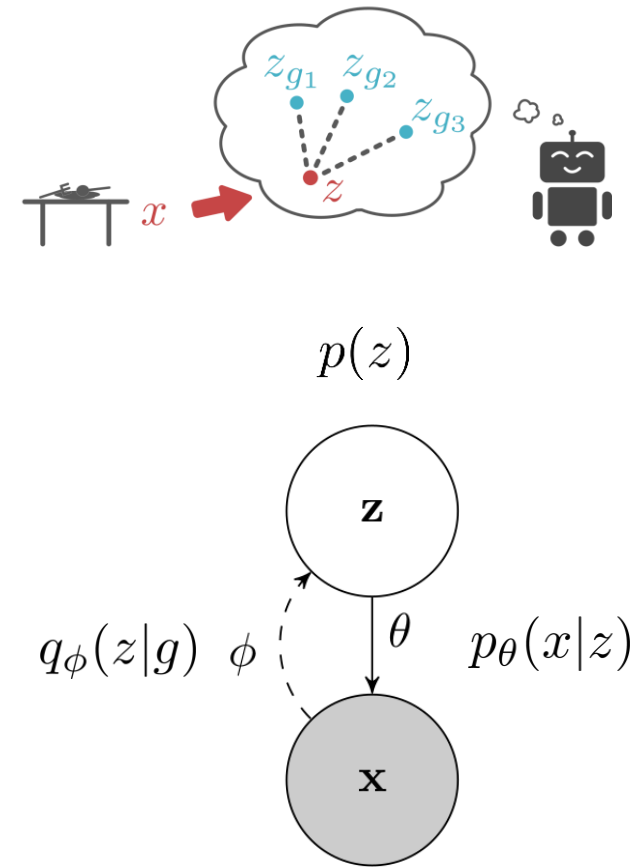


Learn without any rewards at all

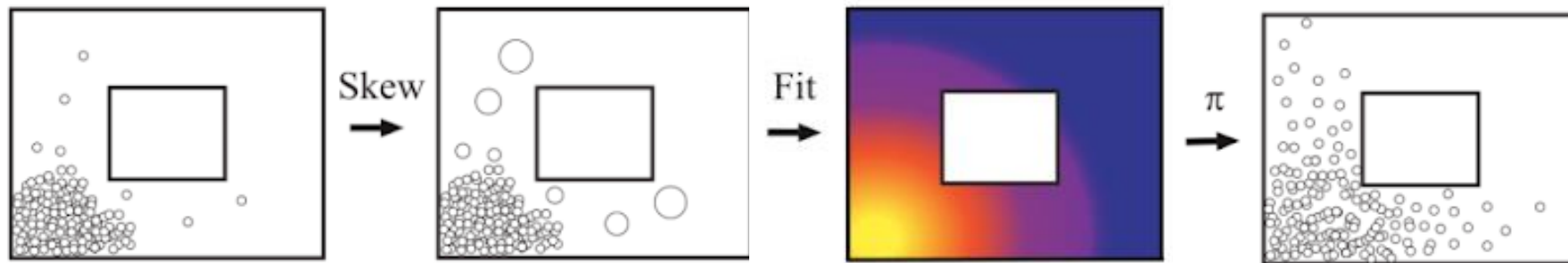
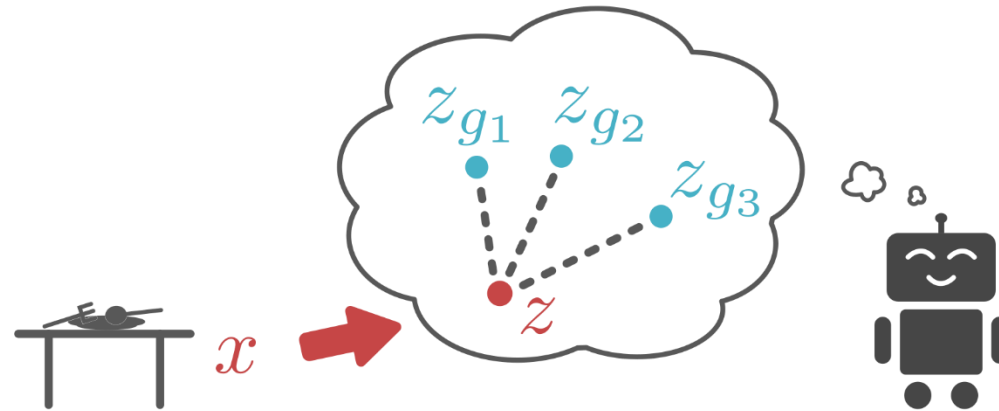


Learn without any rewards at all

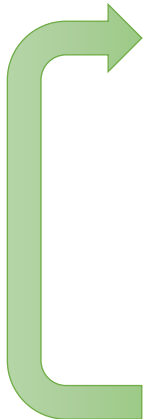
- 
1. Propose goal: $z_g \sim p(z)$, $x_g \sim p_\theta(x_g|z_g)$
 2. Attempt to reach goal using $\pi(a|x, x_g)$, reach \bar{x}
 3. Use data to update π
 4. Use data to update $p_\theta(x_g|z_g)$, $q_\phi(z_g|x_g)$



How do we get diverse goals?



How do we get diverse goals?

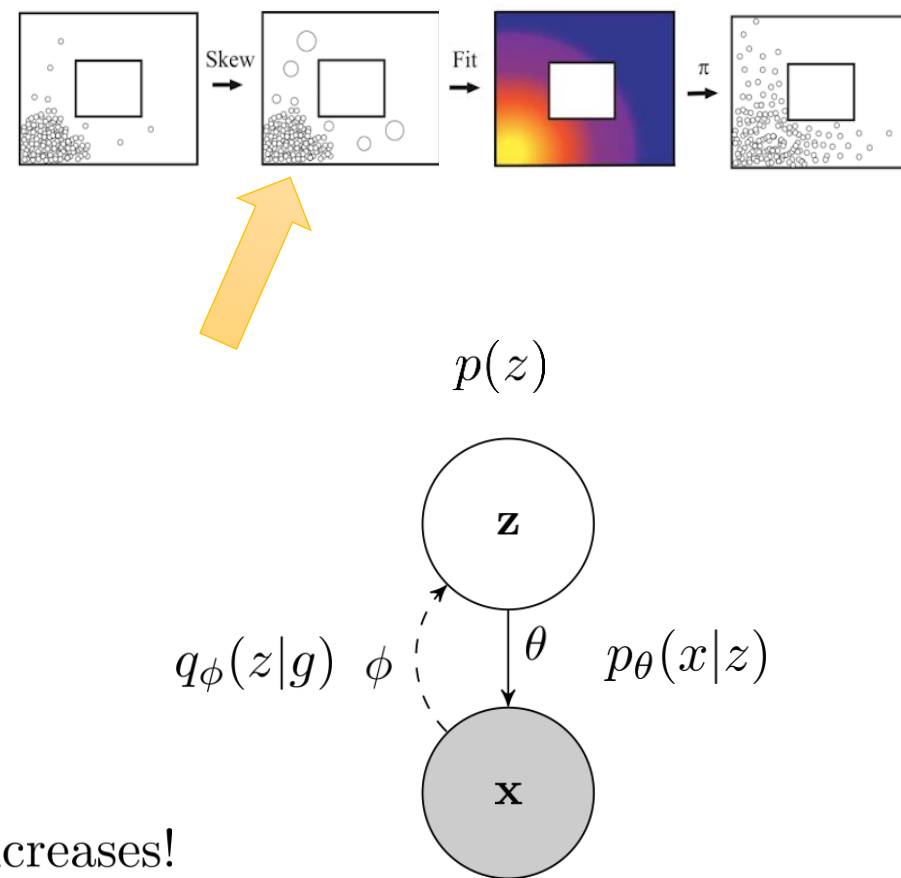
- 
1. Propose goal: $z_g \sim p(z)$, $x_g \sim p_\theta(x_g|z_g)$
 2. Attempt to reach goal using $\pi(a|x, x_g)$, reach \bar{x}
 3. Use data to update π
 4. Use data to update $p_\theta(x_g|z_g)$, $q_\phi(z_g|x_g)$

standard MLE: $\theta, \phi \leftarrow \arg \max_{\theta, \phi} E[\log p(\bar{x})]$

weighted MLE: $\theta, \phi \leftarrow \arg \max_{\theta, \phi} E[w(\bar{x}) \log p(\bar{x})]$

$$w(\bar{x}) = p_\theta(\bar{x})^\alpha$$

key result: for any $\alpha \in [-1, 0)$, entropy $\mathcal{H}(p_\theta(x))$ increases!



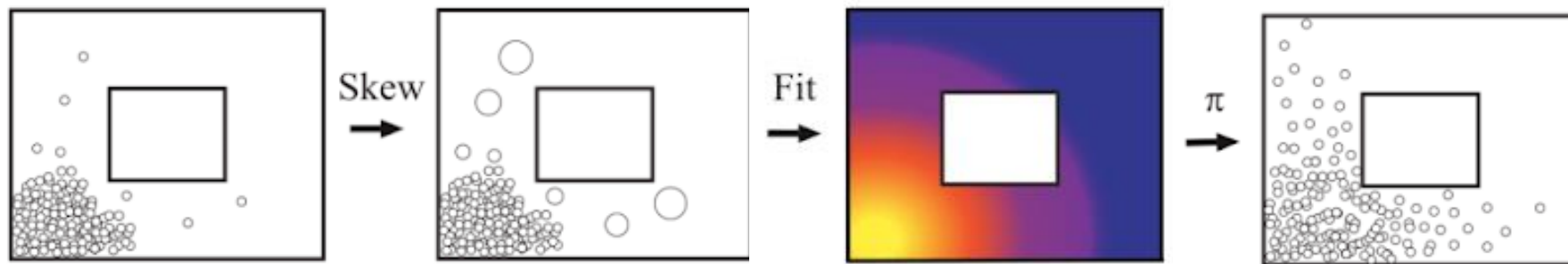
How do we get diverse goals?

what is the objective?

$$\max \mathcal{H}(p(G)) - \mathcal{H}(p(G|S))$$

goals get higher
entropy due to Skew-Fit

$$w(\bar{x}) = p_{\theta}(\bar{x})^{\alpha}$$
$$\alpha \in [-1, 0)$$



what does RL do?

$\pi(a|S, G)$ trained to reach goal G

as π gets better, final state S gets close to G

that means $p(G|S)$ becomes more deterministic!

goal final state

How do we get diverse goals?

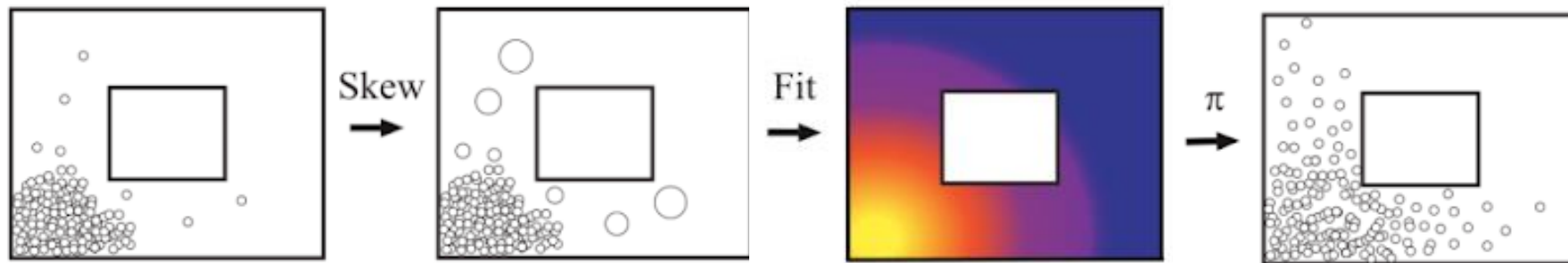
what is the objective?

$$\max \mathcal{H}(p(G)) - \mathcal{H}(p(G|S)) = \max \mathcal{I}(S; G)$$

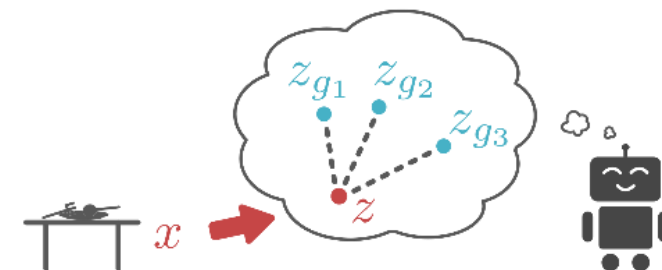
maximizing mutual information between S and G leads to

good exploration (state coverage) $-\mathcal{H}(p(G))$

effective goal reaching $-\mathcal{H}(p(G|S))$



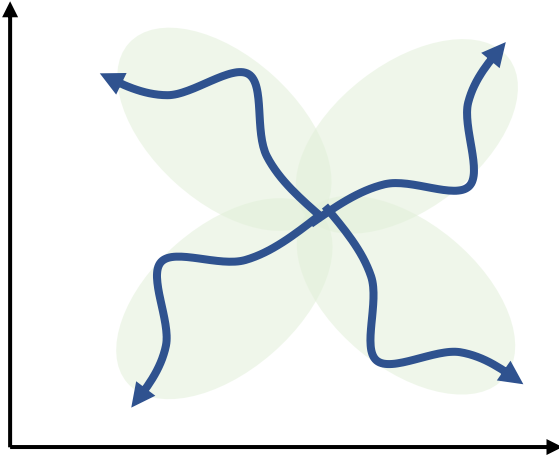
Reinforcement learning with *imagined* goals



In this lecture...

- Definitions & concepts from information theory
- Learning without a reward function by reaching goals
- ***A state distribution-matching*** formulation of reinforcement learning
- Is coverage of valid states a *good* exploration objective?
- Beyond state covering: covering the *space of skills*

Aside: exploration with intrinsic motivation



common method for exploration:

incentivize policy $\pi(\mathbf{a}|\mathbf{s})$ to explore diverse states


...before seeing any reward

reward visiting **novel** states

if a state is visited *often*, it is not *novel*

\Rightarrow add an exploration bonus to reward: $\tilde{r}(\mathbf{s}) = r(\mathbf{s}) - \log p_\pi(\mathbf{s})$

\uparrow
state density under $\pi(\mathbf{a}|\mathbf{s})$

- 
1. update $\pi(\mathbf{a}|\mathbf{s})$ to maximize $E_\pi[\tilde{r}(\mathbf{s})]$
 2. update $p_\pi(\mathbf{s})$ to fit state marginal

Can we use this for state marginal matching?

the state marginal matching problem: learn $\pi(\mathbf{a}|\mathbf{s})$ so as to minimize $D_{\text{KL}}(p_{\pi}(\mathbf{s})||p^*(\mathbf{s}))$

idea: can we use intrinsic motivation?

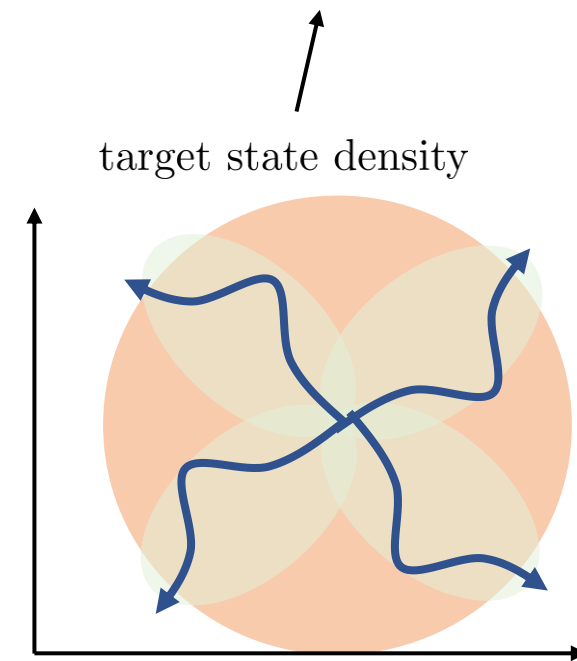
$$\tilde{r}(\mathbf{s}) = \log p^*(\mathbf{s}) - \log p_{\pi}(\mathbf{s})$$

this does **not** perform marginal matching!

- 1. learn $\pi^k(\mathbf{a}|\mathbf{s})$ to maximize $E_{\pi}[\tilde{r}^k(\mathbf{s})]$
- ~~2. update $p_{\pi^k}(\mathbf{s})$ to fit state marginal~~
- 2. update $p_{\pi^k}(\mathbf{s})$ to fit *all states seen so far*
- 3. return $\pi^*(\mathbf{a}|\mathbf{s}) = \sum_k \pi^k(\mathbf{a}|\mathbf{s})$

this **does** perform marginal matching!

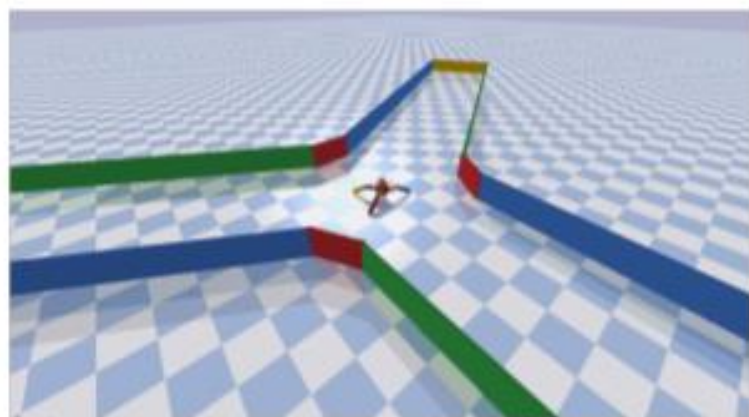
$p_{\pi}(\mathbf{s}) = p^*(\mathbf{s})$ is Nash equilibrium of two player game between π^k and p_{π^k}



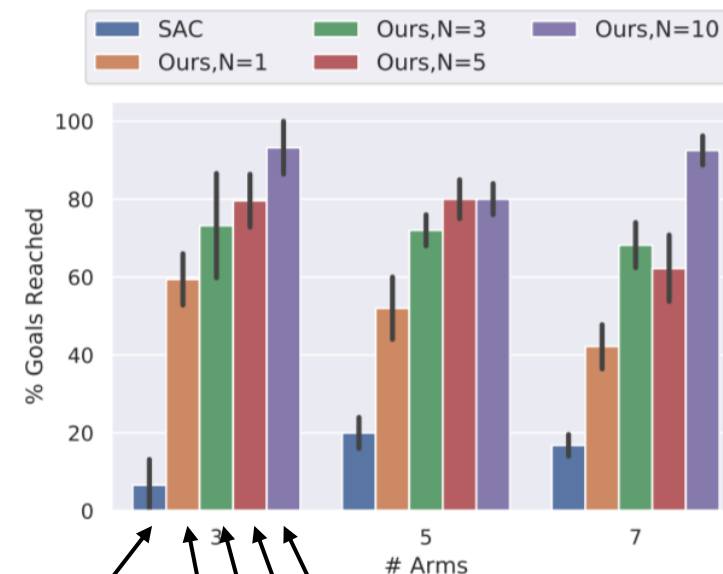
special case: $\log p^*(\mathbf{s}) = C \Rightarrow$ *uniform* target
 $D_{\text{KL}}(p_{\pi}(\mathbf{s})||U(\mathbf{s})) = \mathcal{H}(p_{\pi}(\mathbf{s}))$

State marginal matching for exploration

the state marginal matching problem: learn $\pi(\mathbf{a}|\mathbf{s})$ so as to minimize $D_{\text{KL}}(p_{\pi}(\mathbf{s})||p^*(\mathbf{s}))$



much better coverage!



MaxEnt on actions

variants of SMM

Lee*, Eysenbach*, Parisotto*, Xing, Levine, Salakhutdinov. **Efficient Exploration via State Marginal Matching**

See also: Hazan, Kakade, Singh, Van Soest. **Provably Efficient Maximum Entropy Exploration**

In this lecture...

- Definitions & concepts from information theory
- Learning without a reward function by reaching goals
- *A state distribution-matching* formulation of reinforcement learning
- Is coverage of valid states a *good* exploration objective?
- Beyond state covering: covering the *space of skills*

Is state entropy *really* a good objective?

Skew-Fit: $\max \mathcal{H}(p(G)) - \mathcal{H}(p(G|S)) = \max \mathcal{I}(S; G)$

SMM (special case where $p^*(\mathbf{s}) = C$): $\max \mathcal{H}(p_\pi(S))$

more or less the same thing

When is this a good idea?

“Eysenbach’s Theorem” (not really what it’s called)

(follows trivially from classic maximum entropy modeling)

at test time, an *adversary* will choose the *worst* goal G

which goal distribution should you use for *training*?

answer: choose $p(G) = \arg \max_p \mathcal{H}(p(G))$

See also: Hazan, Kakade, Singh, Van Soest. **Provably Efficient Maximum Entropy Exploration**

Gupta, Eysenbach, Finn, Levine. **Unsupervised Meta-Learning for Reinforcement Learning**

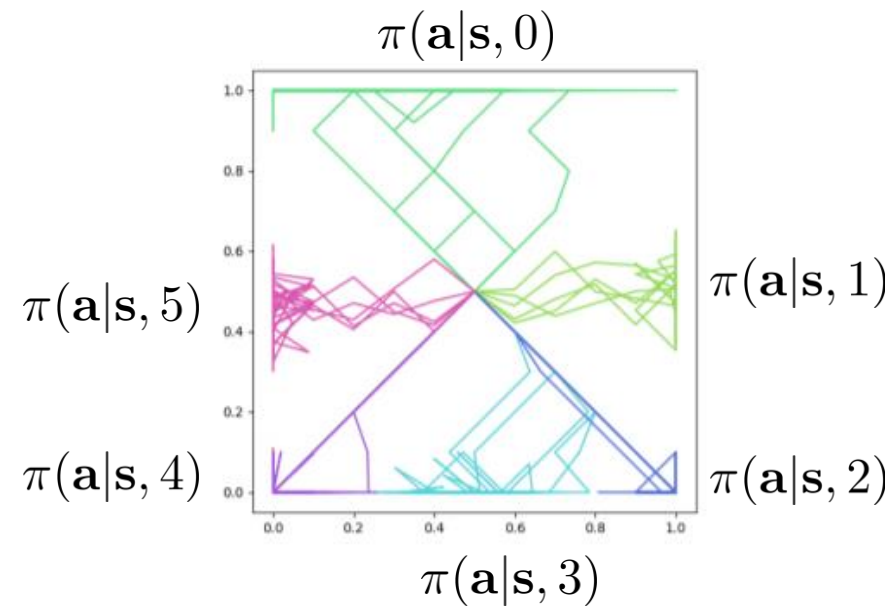
In this lecture...

- Definitions & concepts from information theory
- A *distribution-matching* formulation of reinforcement learning
- Learning without a reward function by reaching goals
- A *state distribution-matching* formulation of reinforcement learning
- Is coverage of valid states a *good* exploration objective?
- Beyond state covering: covering the *space of skills*

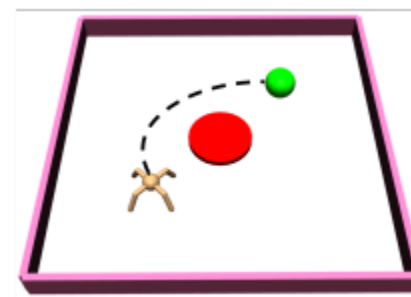
Learning diverse skills

$$\pi(\mathbf{a}|\mathbf{s}, z)$$

↑
task index



Reaching diverse **goals** is not the same as performing diverse **tasks**
not all behaviors can be captured by **goal-reaching**



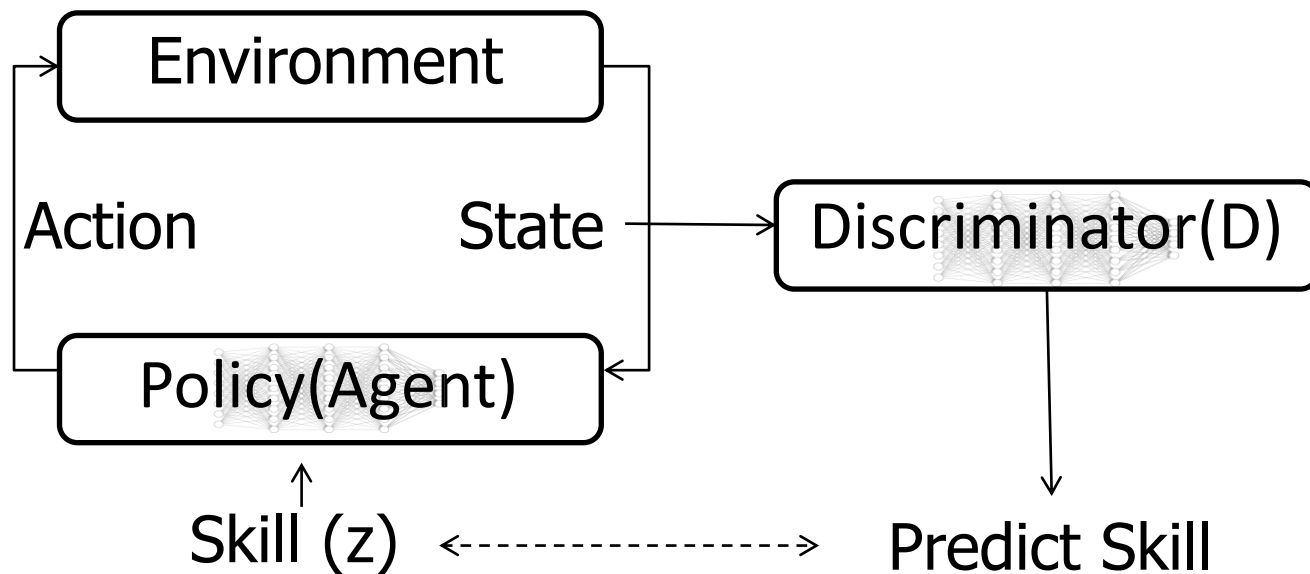
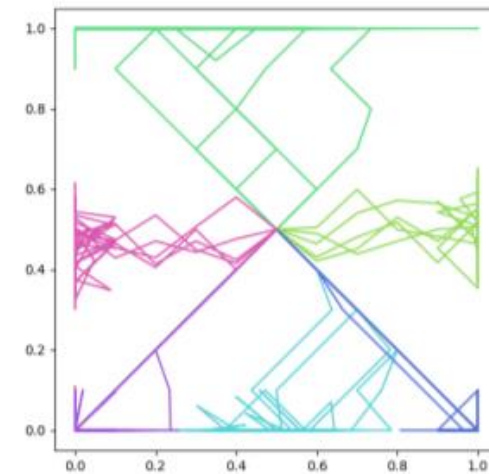
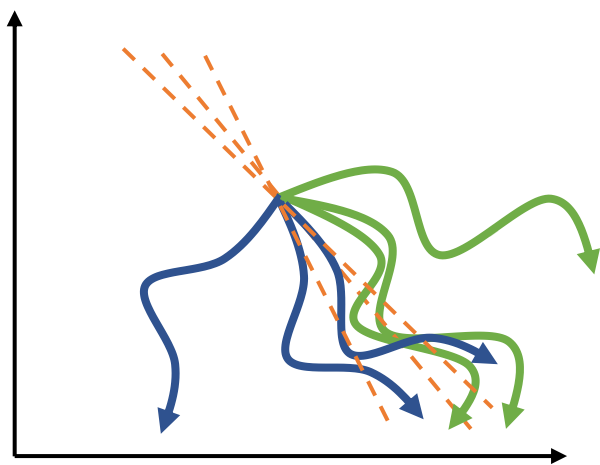
Intuition: different **skills** should visit different **state-space regions**

Diversity-promoting reward function

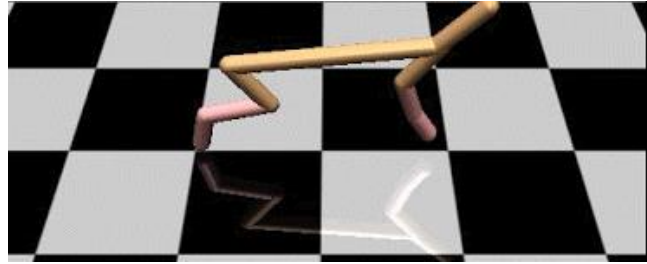
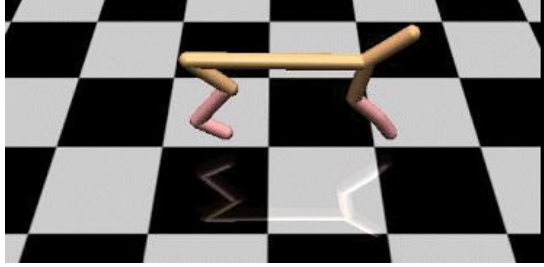
$$\pi(\mathbf{a}|\mathbf{s}, z) = \arg \max_{\pi} \sum_z E_{\mathbf{s} \sim \pi(\mathbf{s}|z)} [r(\mathbf{s}, z)]$$

reward states that are unlikely for other $z' \neq z$

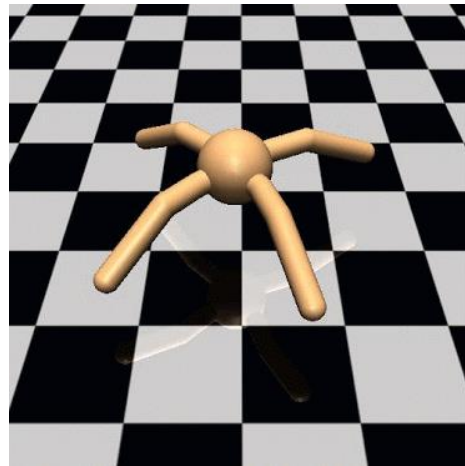
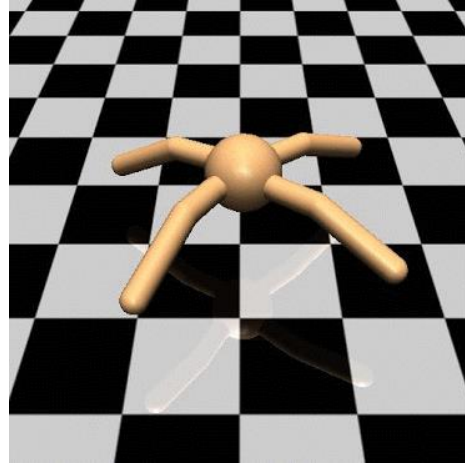
$$r(\mathbf{s}, z) = \log p(z|\mathbf{s})$$



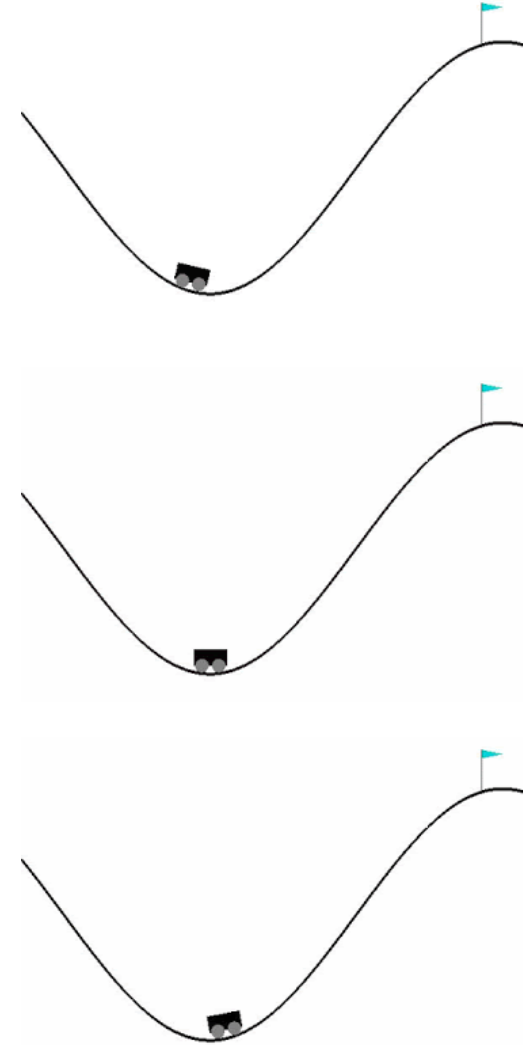
Examples of learned tasks



Cheetah



Ant



Mountain car

A connection to mutual information

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

$$r(\mathbf{s}, z) = \log p(z|\mathbf{s})$$

$$I(z, \mathbf{s}) = H(z) - H(z|\mathbf{s})$$

maximized by using uniform prior $p(z)$



minimized by maximizing $\log p(z|\mathbf{s})$



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

See also: Gregor et al. **Variational Intrinsic Control.** 2016

UC Berkeley·CS285 | Deep Reinforcement Learning (2020)

CS285 (2020)· 课程资料包 @ShowMeAI



视频

中英双语字幕



课件

一键打包下载



笔记

官方笔记翻译



代码

作业项目解析



视频·B 站 [扫码或点击链接]

<https://www.bilibili.com/video/BV12341167kL>



课件 & 代码·博客 [扫码或点击链接]

<http://blog.showmeai.tech/cs285>

Berkeley

actor-critic

reinforcement learning

exploration

元学习

exploitation

梯度策略

探索与利用

DQN

迁移学习

逆强化学习

Q-learning / Q 学习

deep q network

meta learning

Awesome AI Courses Notes Cheatsheets 是 [ShowMeAI](#) 资料库的分支系列，覆盖最具知名度的 **TOP20+** 门 AI 课程，旨在为读者和学习者提供一整套高品质中文学习笔记和速查表。

点击课程名称，跳转至课程**资料包**页面，**一键下载**课程全部资料！

机器学习	深度学习	自然语言处理	计算机视觉
Stanford · CS229	Stanford · CS230	Stanford · CS224n	Stanford · CS231n
# Awesome AI Courses Notes Cheatsheets· 持续更新中			
知识图谱	图机器学习	深度强化学习	自动驾驶
Stanford · CS520	Stanford · CS224W	UCBerkeley · CS285	MIT · 6.S094



微信公众号

资料下载方式 2: 扫码点击**底部菜单栏**

称为 **AI 内容创作者**? 回复 [添砖加瓦]