IASD M2 at Paris Dauphine

Deep Reinforcement Learning

15: Exploration (Part 2)

Eric Benhamou Thérèse Des Escotais

Acknowledgement

These materials are based on the seminal course of Sergey Levine CS285

What's the problem?

this is easy (mostly)

this is impossible

Unsupervised learning of diverse behaviors

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

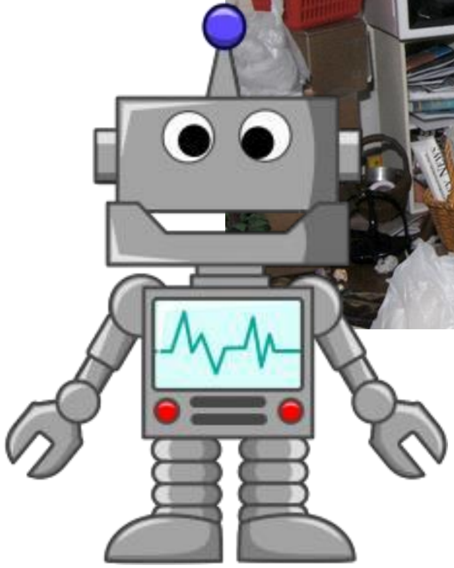
Why?

➢ *Learn skills without supervision, then use them to accomplish goals*

➢ *Learn sub-skills to use with hierarchical reinforcement learning*

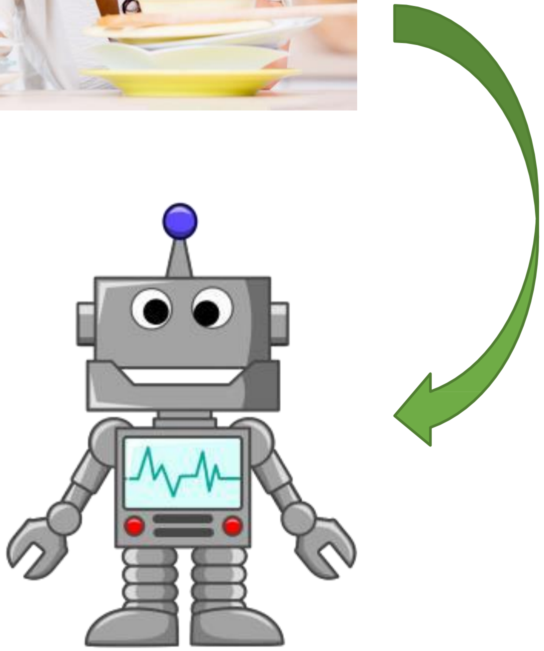
➢ *Explore the space of possible behaviors*

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...

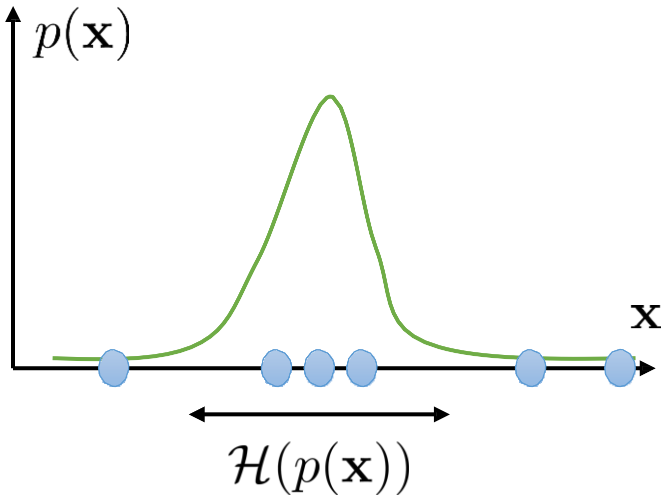
- ➢ Definitions & concepts from information theory
- ➢ Learning without a reward function by reaching goals
- ➢ A *state distribution-matching* formulation of reinforcement learning
- ➢ Is coverage of valid states a *good* exploration objective?
- ➢ Beyond state covering: covering the *space of skills*

Some useful identities

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

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

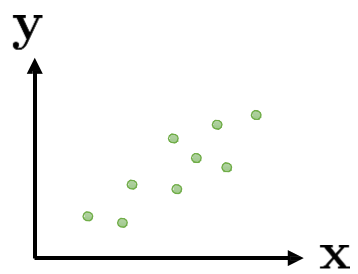
entropy – how "broad" $p(\mathbf{x})$ is



Some useful identities

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

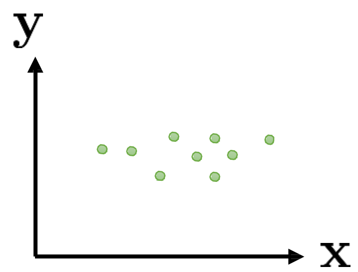
$$\mathcal{H}(p(\mathbf{x})) = -E_{\mathbf{x}\sim p(\mathbf{x})}[\log p(\mathbf{x})]$$



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

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

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



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

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

Information theoretic quantities in RL

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

quantifies *coverage*

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

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

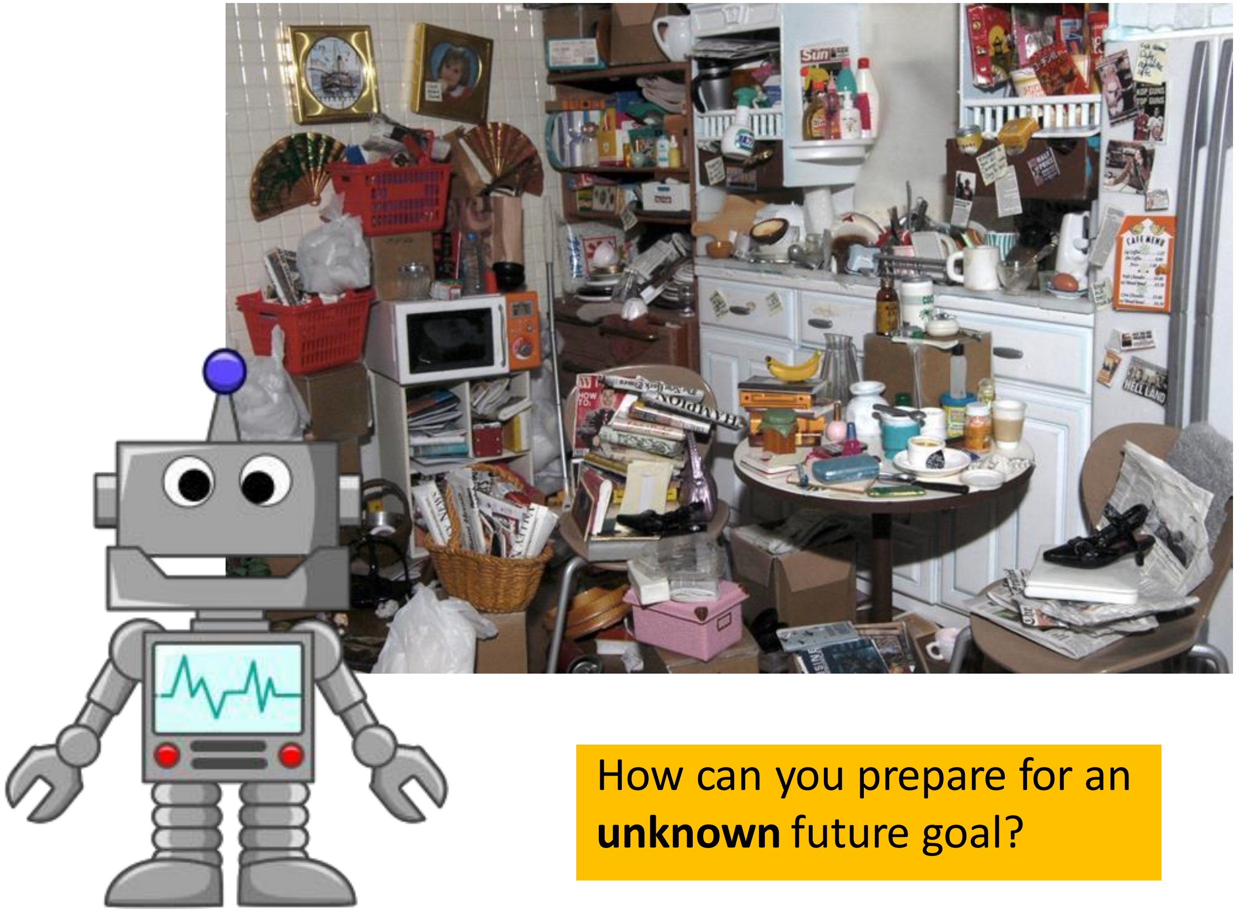
$$\mathcal{I}(\mathbf{s}_{t+1};\mathbf{a}_t) = \mathcal{H}(\mathbf{s}_{t+1}) - \mathcal{H}(\mathbf{s}_{t+1}|\mathbf{a}_t)$$

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

In this lecture…

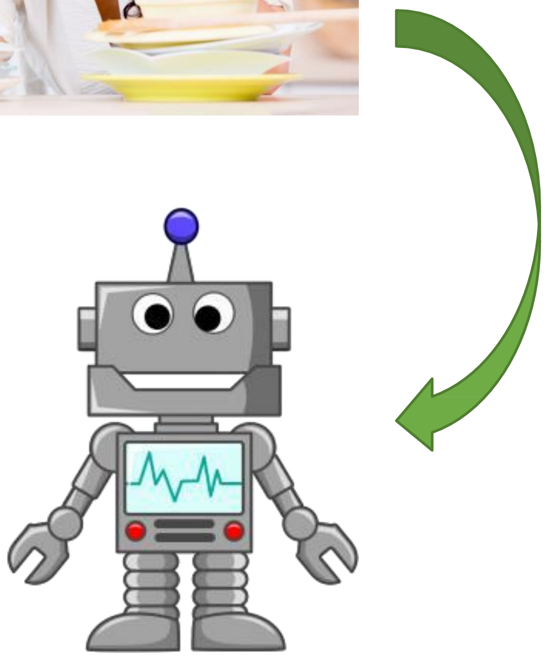
- ➢ Definitions & concepts from information theory
- ➢ Learning without a reward function by reaching goals
- ➢ A *state distribution-matching* formulation of reinforcement learning
- ➢ Is coverage of valid states a *good* exploration objective?
- ➢ Beyond state covering: covering the *space of skills*

An Example Scenario



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

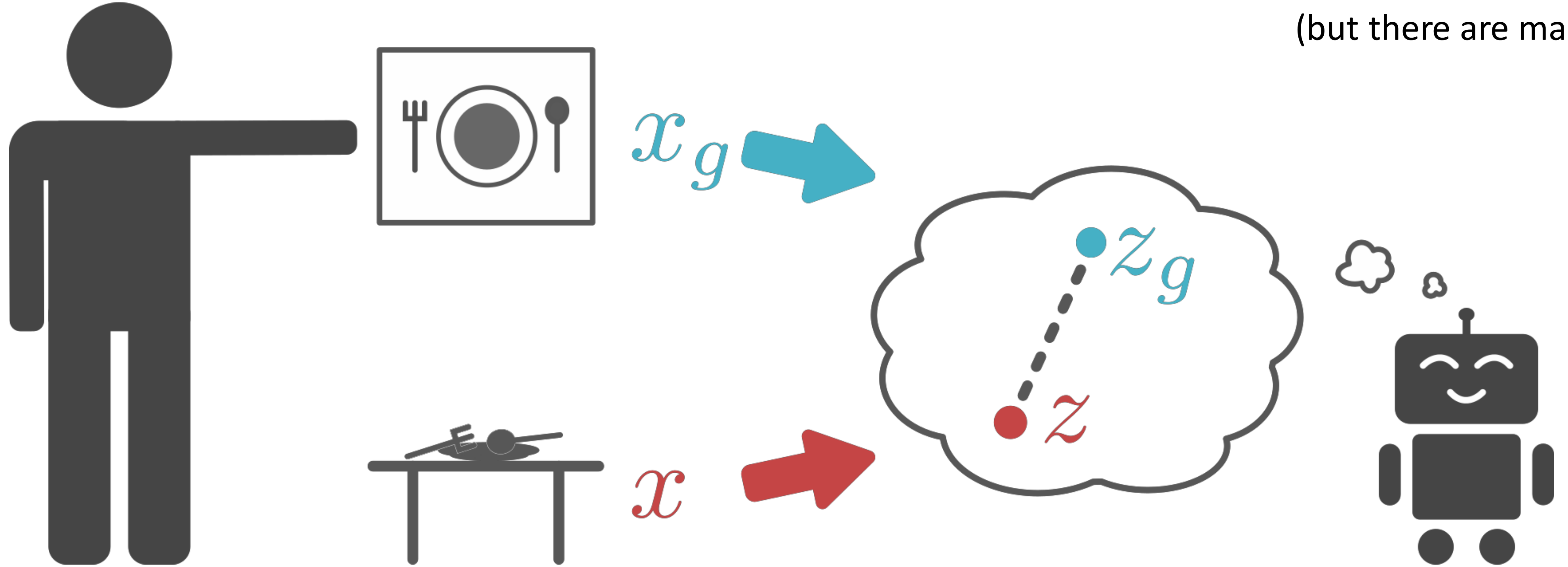
training time: unsupervised

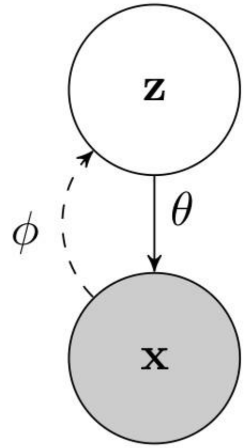


Learn without any rewards at all

VAE (Kingma & Welling '13)

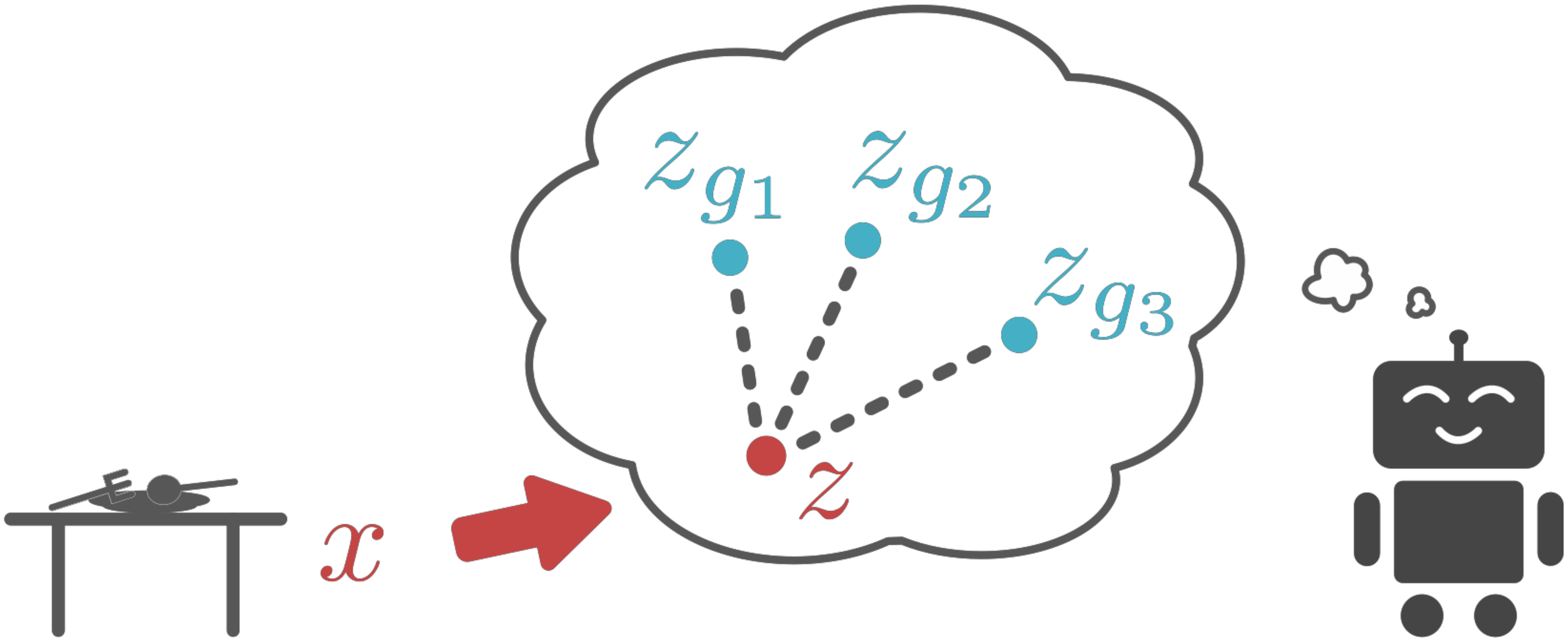
(but there are many other choices)





Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. '18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** '19

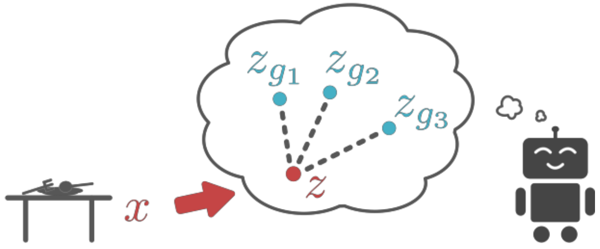
Learn without any rewards at all

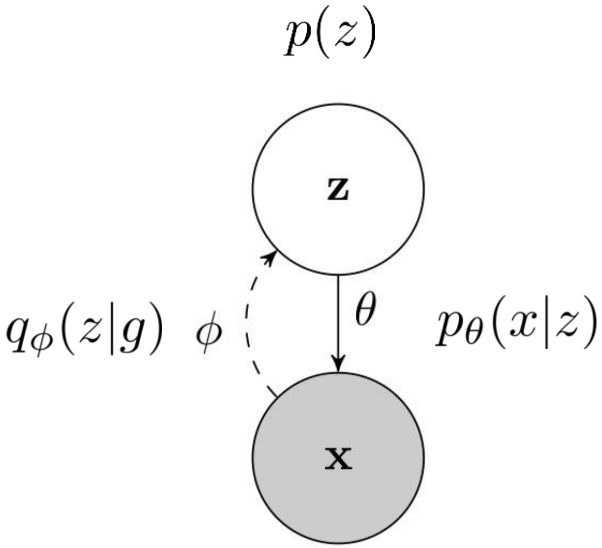


Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. ’18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** ‘19

Learn without any rewards at all

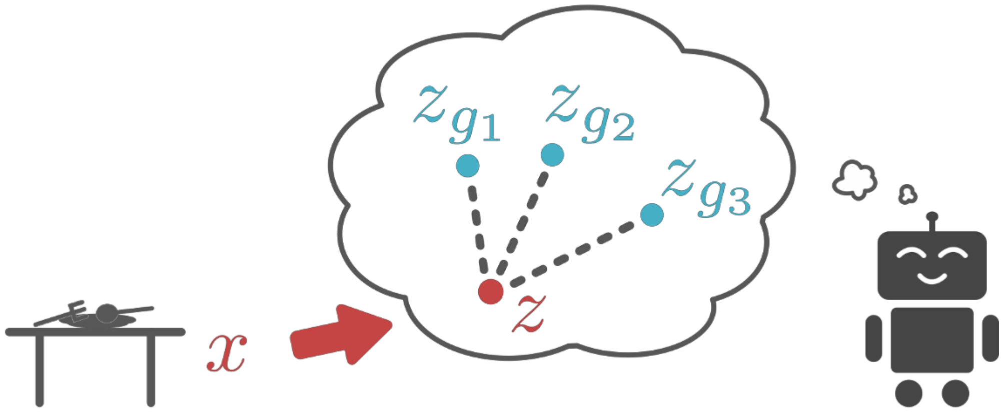
1. Propose goal: $z_g \sim p(z)$, $x_g \sim p_\theta(x_g|z_g)$
2. Attempt to reach goal using $\pi(a|x, x_g)$, reach $\bar{x}$
3. Use data to update π
4. Use data to update $p_\theta(x_g|z_g)$, $q_\phi(z_g|x_g)$

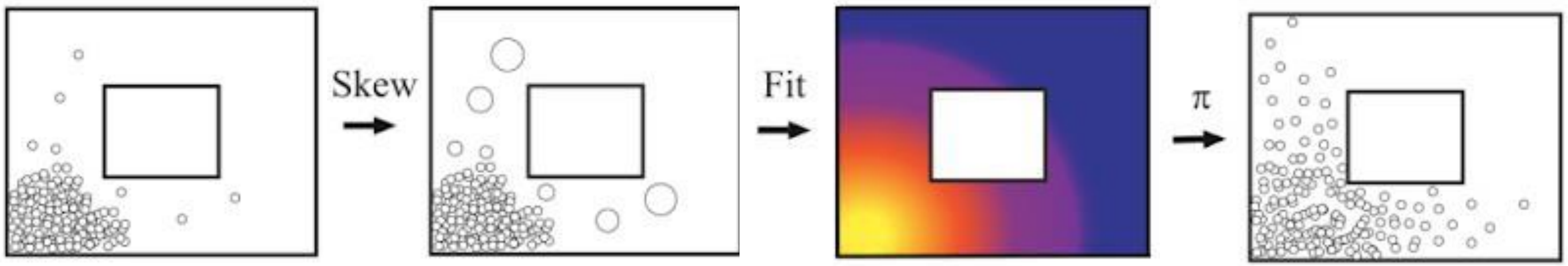




Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. '18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** '19

How do we get diverse goals?

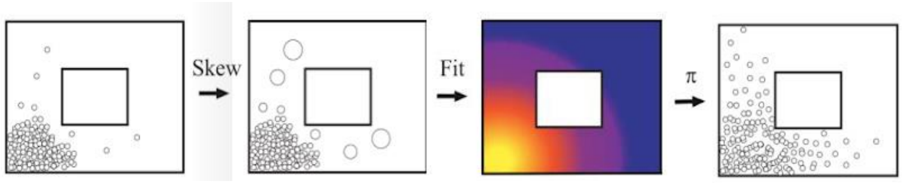




Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. '18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** '19

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 $p_\theta(x_g|z_g)$, $q_\phi(z_g|x_g)$

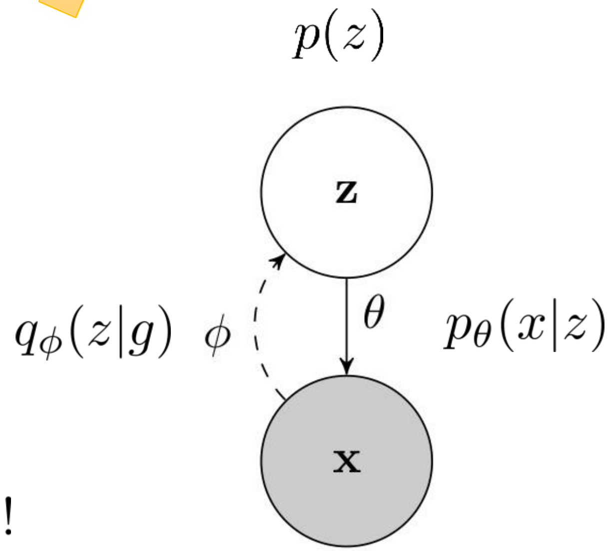


standard MLE: $\theta, \phi \leftarrow \arg\max_{\theta,\phi} E[\log p(\bar{x})]$

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!



Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. '18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** '19

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

what does RL do?

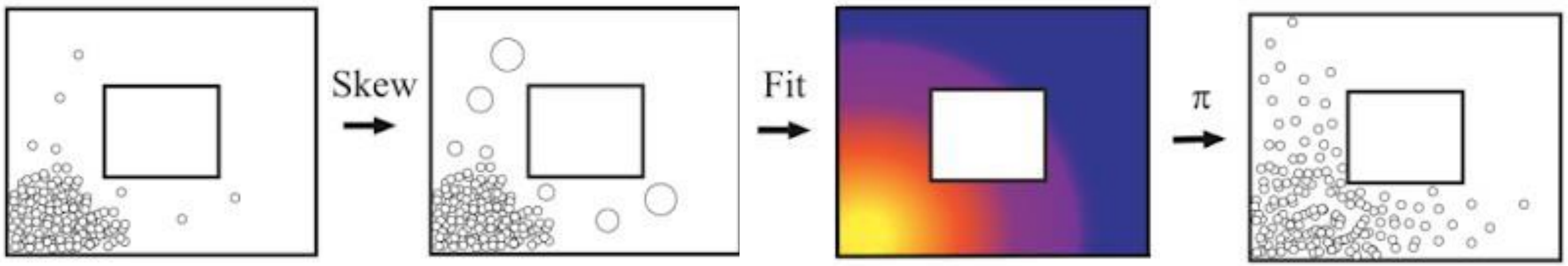
$\pi(a|S, G)$ trained to reach goal G

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

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

goal final state

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



Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. ’18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** ‘19

How do we get diverse goals?

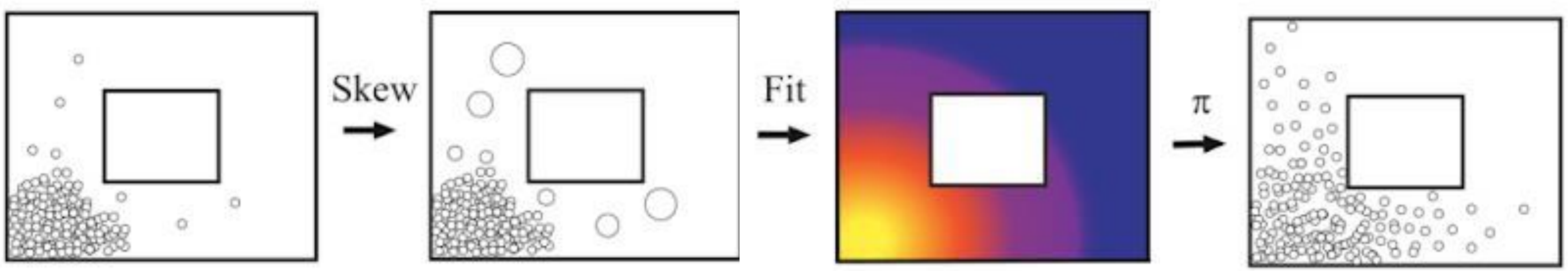
what is the objective?

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

maximizing mutual information between S and G leads to

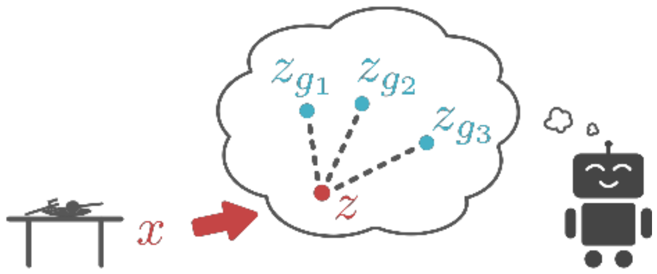
good exploration (state coverage) – $\mathcal{H}(p(G))$

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



Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. '18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** '19

Reinforcement learning with *imagined* goals



imagined goal

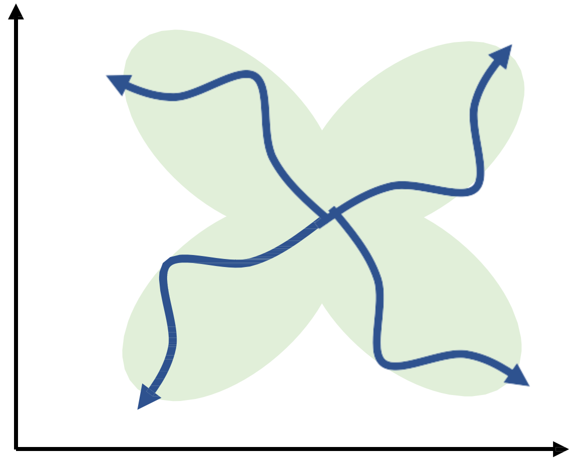
RL episode

Nair*, Pong*, Bahl, Dalal, Lin, L. **Visual Reinforcement Learning with Imagined Goals**. ’18
Dalal*, Pong*, Lin*, Nair, Bahl, Levine. **Skew-Fit: State-Covering Self-Supervised Reinforcement Learning.** ‘19

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: $\tilde{r}(\mathbf{s}) = r(\mathbf{s}) - \log p_\pi(\mathbf{s})$

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 minimze $D_{\mathrm{KL}}(p_\pi(\mathbf{s})\|p^\star(\mathbf{s}))$

idea: can we use intrinsic motivation?

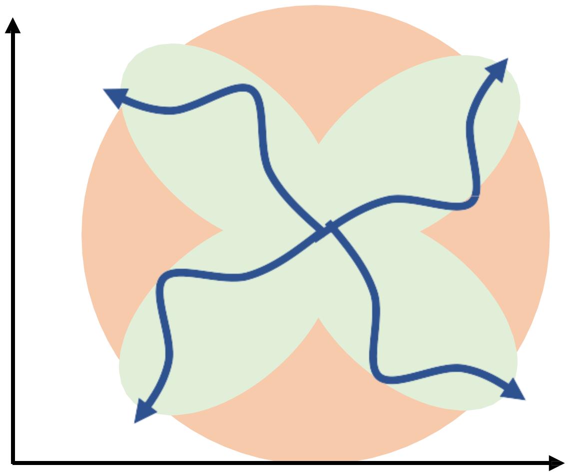
$$\tilde{r}(\mathbf{s}) = \log p^\star(\mathbf{s}) - \log p_\pi(\mathbf{s})$$

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 $\pi^\star(\mathbf{a}|\mathbf{s}) = \sum_k \pi^k(\mathbf{a}|\mathbf{s})$

this **does** perform marginal matching!

target state density



special case: $\log p^\star(\mathbf{s}) = C \Rightarrow$ *uniform* target

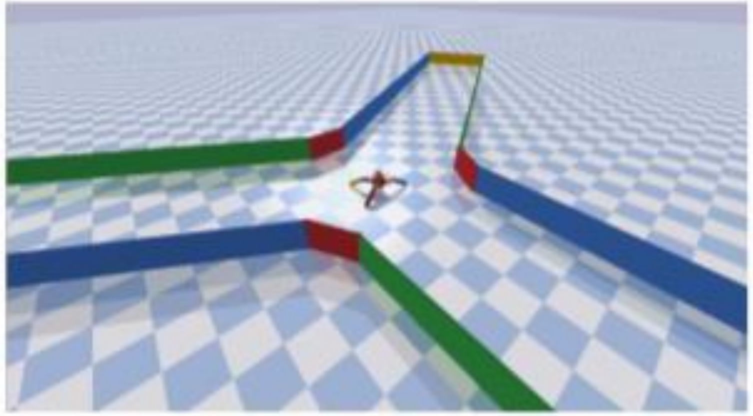
$$D_{\mathrm{KL}}(p_\pi(\mathbf{s})\|U(\mathbf{s})) = \mathcal{H}(p_\pi(\mathbf{s}))$$

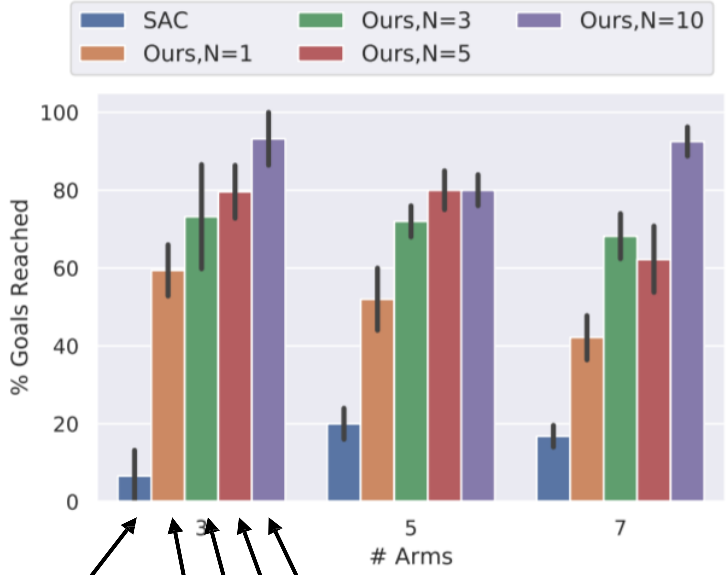
$p_\pi(\mathbf{s}) = p^\star(\mathbf{s})$ is Nash equilibrium of two player game between π^k and p_{π^k}

Lee*, Eysenbach*, Parisotto*, Xing, Levine, Salakhutdinov. **Efficient Exploration via State Marginal Matching**
See also: Hazan, Kakade, Singh, Van Soest. **Provably Efficient Maximum Entropy Exploration**

State marginal matching for exploration

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





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

more or less the same thing

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

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**

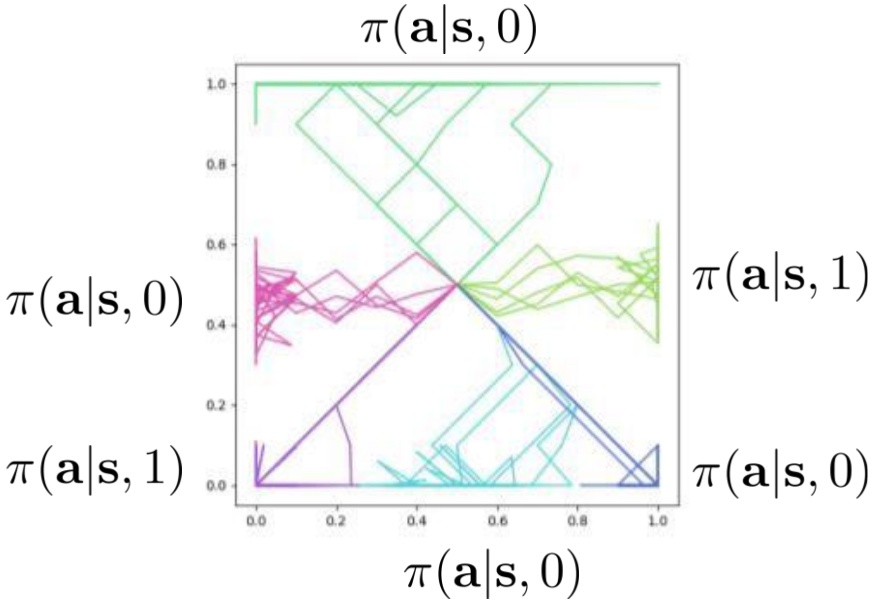
In this lecture…

- ➢ 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

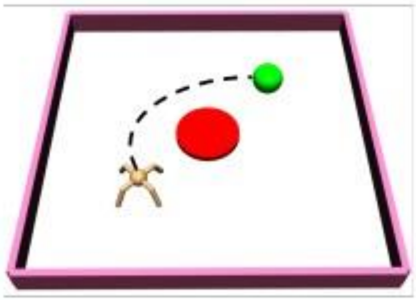
$$\pi(\mathbf{a}|\mathbf{s}, z)$$

task index



Reaching diverse **goals** is not the same as performing diverse **tasks**

not all behaviors can be captured by **goal-reaching**



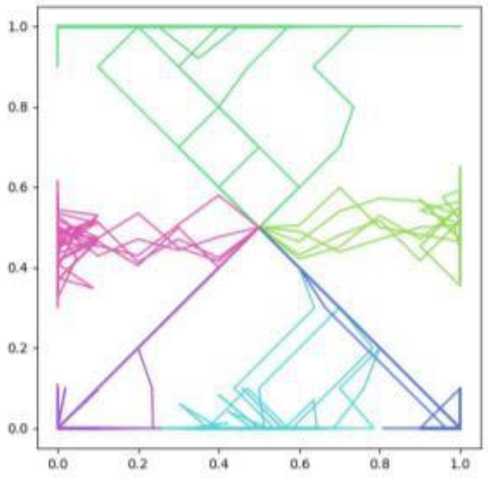
Intuition: different **skills** should visit different **state-space regions**

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

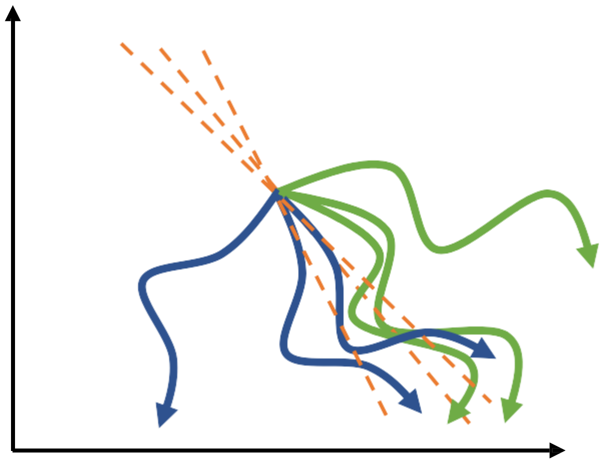
Diversity-promoting reward function

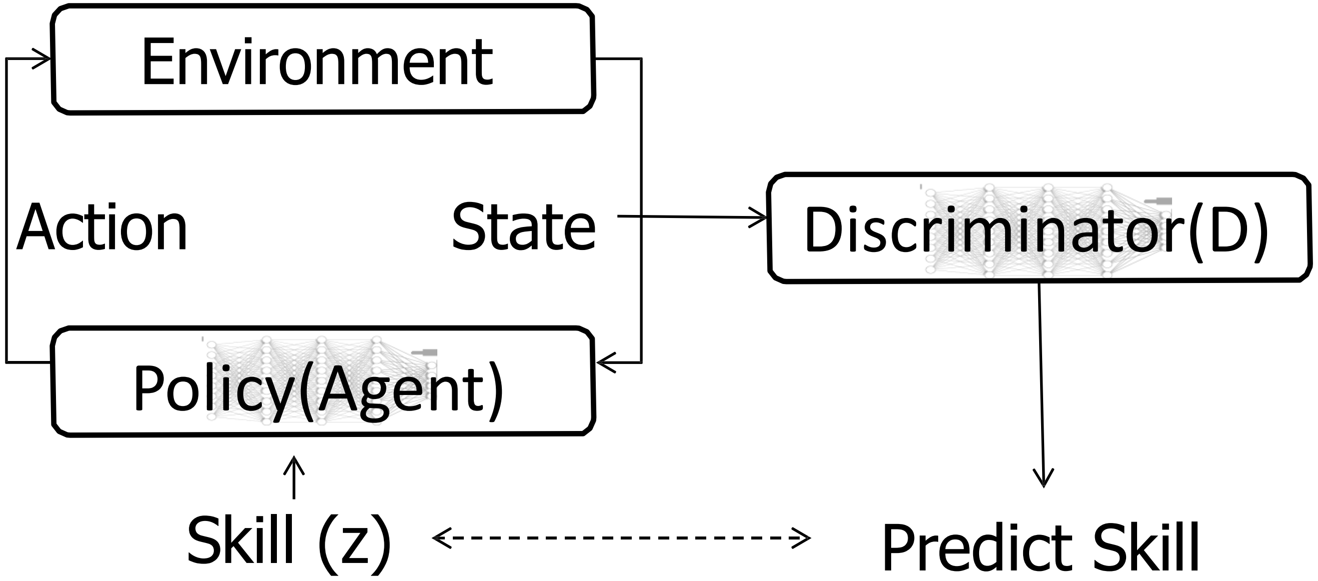
$$\pi(\mathbf{a}|\mathbf{s}, z) = \arg\max_{\pi} \sum_{z} E_{\mathbf{s}\sim\pi(\mathbf{s}|z)}[r(\mathbf{s}, z)]$$

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



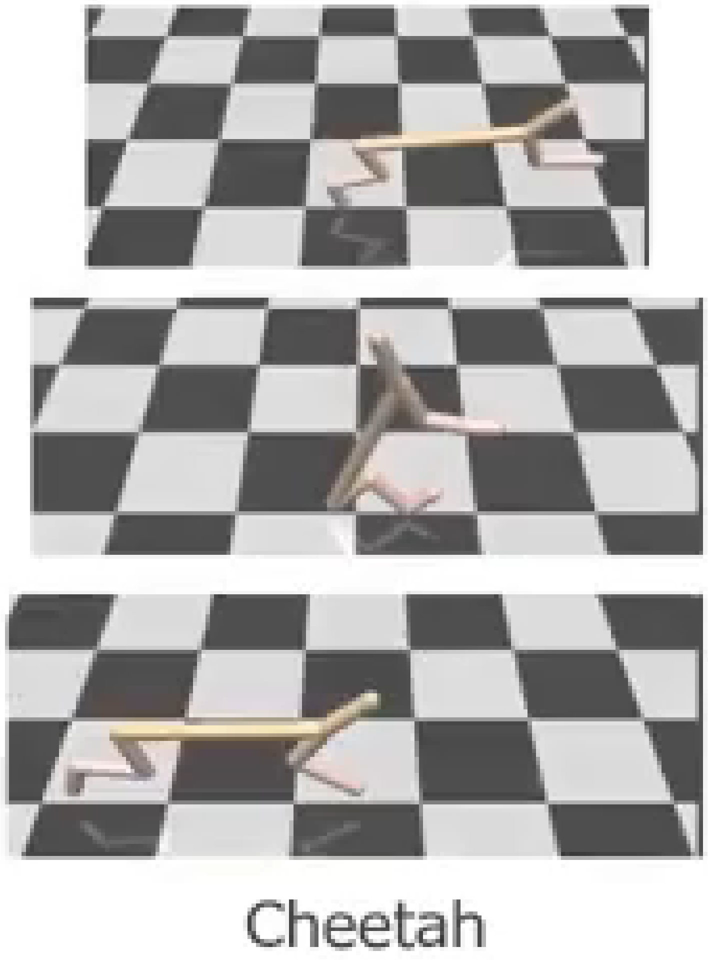
$$r(\mathbf{s}, z) = \log p(z|\mathbf{s})$$





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

Examples of learned tasks



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

A connection to mutual information

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

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

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

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