A **Graphical Model** Perspective on **Multi-Task** and **Meta-RL**

CS 330

The Plan

Variational inference review

Control as inference

Control as variational inference

Meta-RL as variational inference

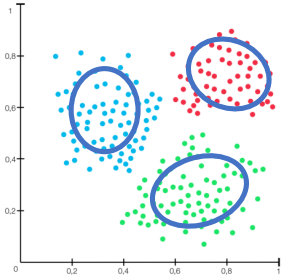
Disclaimer

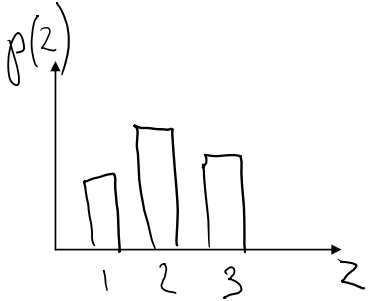
There will be quite a bit of math and derivations

This is an active research topic

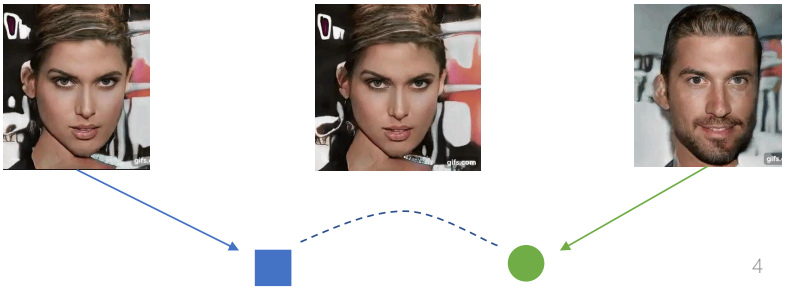
Why Variational Inference?

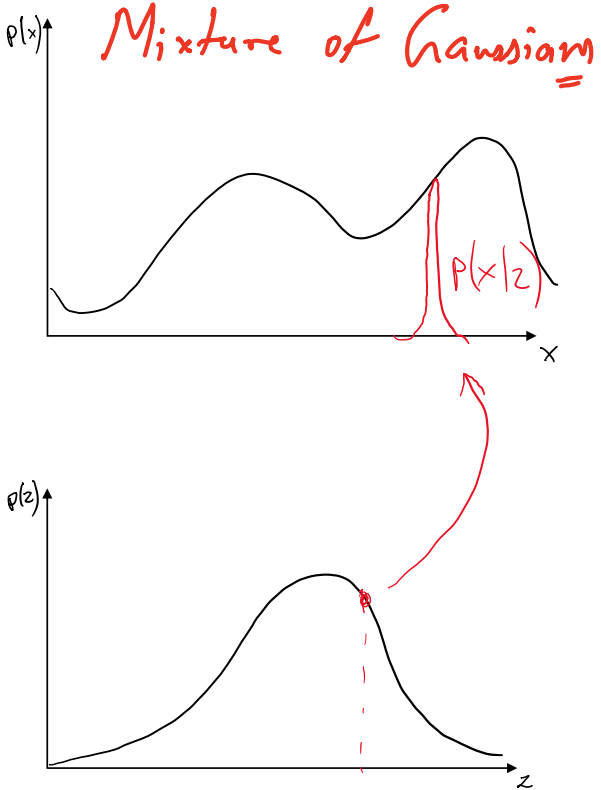
Main tool to cope with latent variable models





$$p(x) = \sum_z p(x|z)p(z)$$

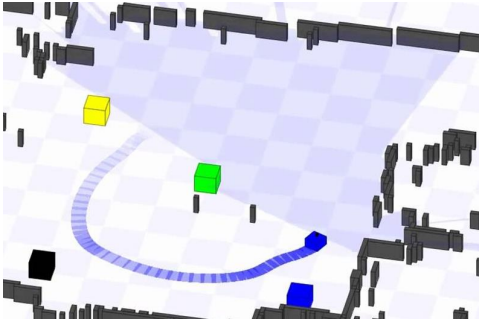


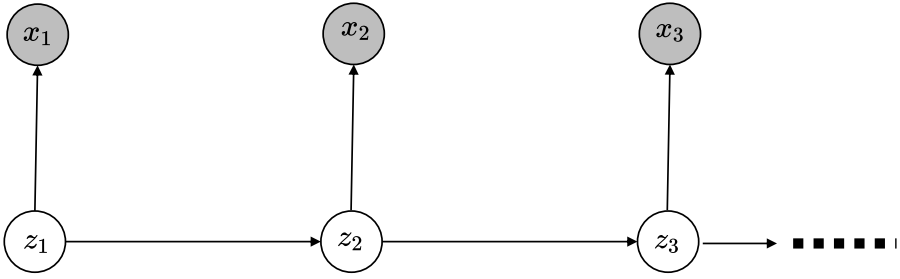


$$p(x) = \int p(x|z)p(z)dz$$

Latent Variable Models in Control

Kalman Filter

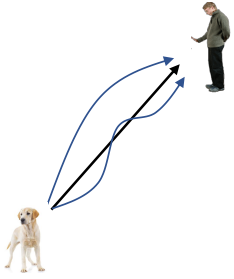




Kalman, A New Approach to Linear Filtering, 1960

Why Control as Inference?

A framework to describe stochastic optimal behavior



Knowledge from Probabilistic Graphical Models for control

- Better exploration
- Hierarchical RL
- Skill discovery

The Plan

Variational inference review

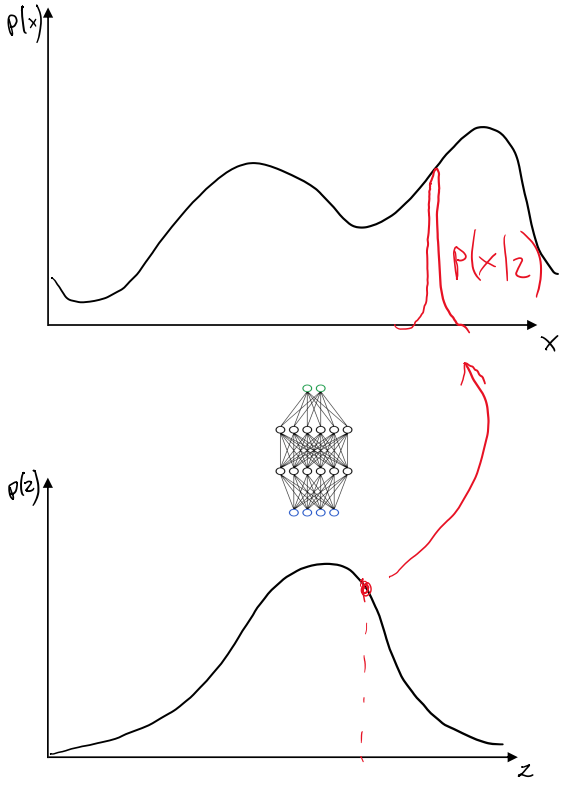
Control as inference

Control as variational inference

Meta-RL as variational inference

Variational Inference

$$p(x) = \int p(x|z)p(z)dz$$

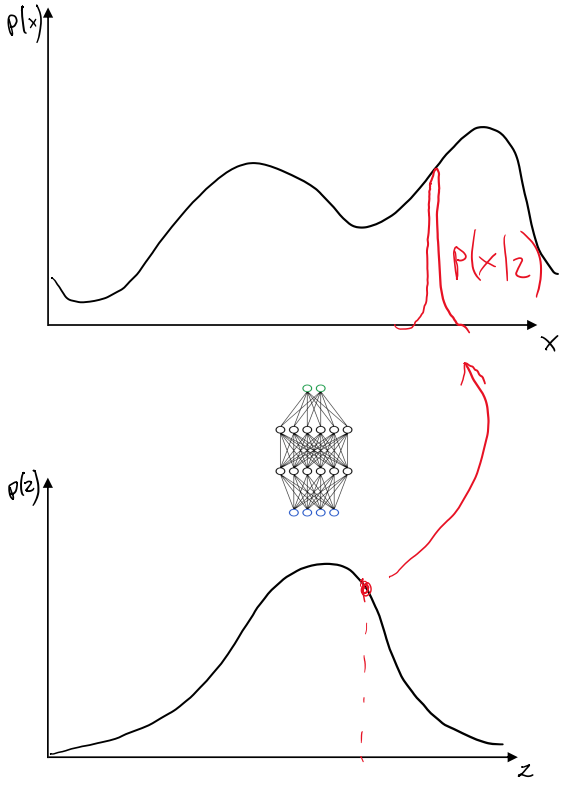


Variational Inference

$$p(x) = \int p(x|z)p(z)dz$$

$$\arg\max_{\theta} \prod_{x_i} p_\theta(x_i) = \arg\max_{\theta} \sum_{x_i} \log p_\theta(x_i)$$

$$\log p_\theta(x_i) = \log \int p(x_i|z)p(z)dz$$



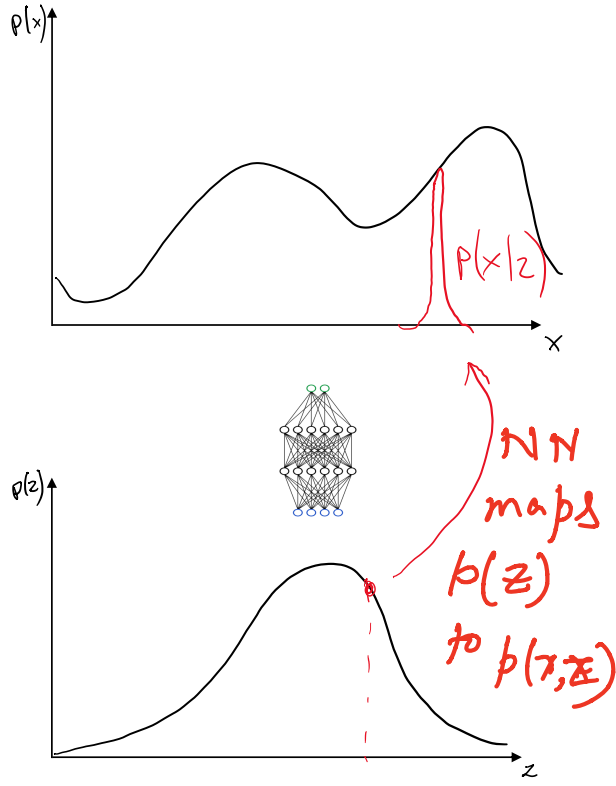
Variational Inference

$$p(x) = \int p(x|z)p(z)dz$$

$$\log p_\theta(x_i) \geq \mathbb{E}_{z\sim q_i(z)} \log p(x_i|z) - \mathbb{D}_{KL}(q_i(z)||p(z))$$

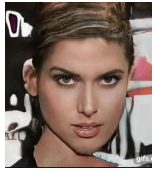
Amortized VI

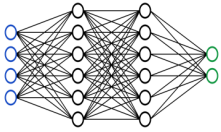
$$\log p_\theta(x_i) \geq \mathbb{E}_{z\sim q_\phi(z|x_i)} \log p_\theta(x_i|z) - \mathbb{D}_{KL}(q_\phi(z|x_i)||p(z))$$



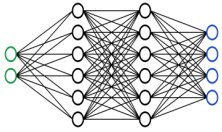
Variational Autoencoder (VAE)

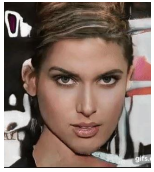
x_i





$q_\phi(z|x_i)$

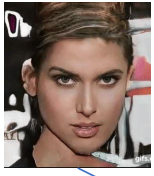




$p_\theta(x_i|z)$

$$\max_{\phi,\theta} \frac{1}{N} \sum_i \mathbb{E}_{z\sim q_\phi(z|x_i)} \log p_\theta(x_i|z) - \mathbb{D}_{KL}(q_\phi(z|x_i)||p(z))$$

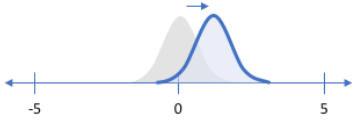
$\mathcal{N}(0, I)$



Variational Autoencoder (VAE) – what did we just do?

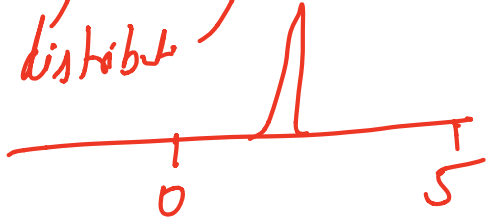
$$\max_{\phi,\theta} \frac{1}{N} \sum_i \mathbb{E}_{z\sim q_\phi(z|x_i)} \log p_\theta(x_i|z) - \mathbb{D}_{KL}(q_\phi(z|x_i)||p(z))$$

Penalizing reconstruction loss encourages the distribution to describe the input



Our distribution deviates from the prior to describe some characteristic of the data

Without regularization our network can "cheat" by learning narrow distribut



Penalizing KL divergence acts as a regularizer

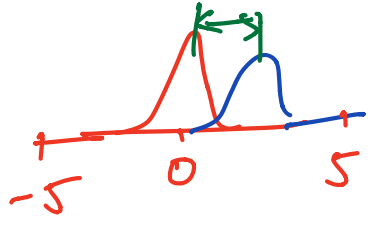


Image credit: Jeremy Jordan

The Plan

Variational inference review

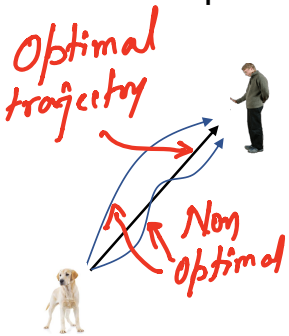
Control as inference

Control as variational inference

Meta-RL as variational inference

Control as Inference

A framework to describe stochastic optimal behavior



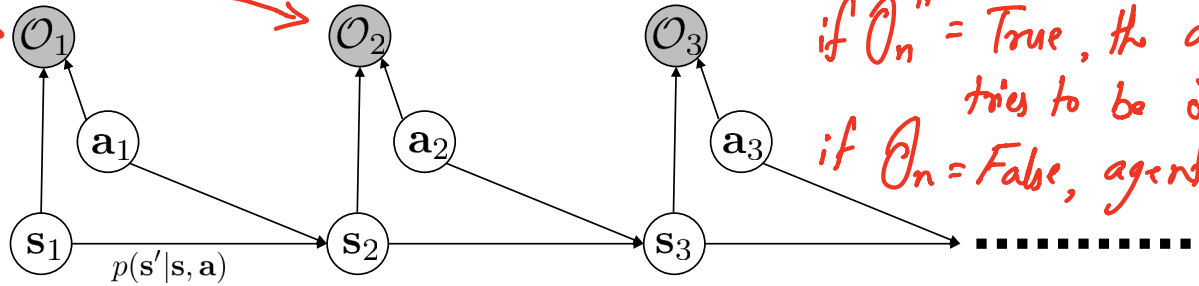
$\tau \rightarrow$ trajectory

$$p(\underbrace{\mathbf{s}_{1:T}, \mathbf{a}_{1:T}}_{\tau})$$

$$p(\tau | \mathcal{O}_{1:T})$$

$$\boxed{p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) = \exp(r(\mathbf{s}_t, \mathbf{a}_t))}$$

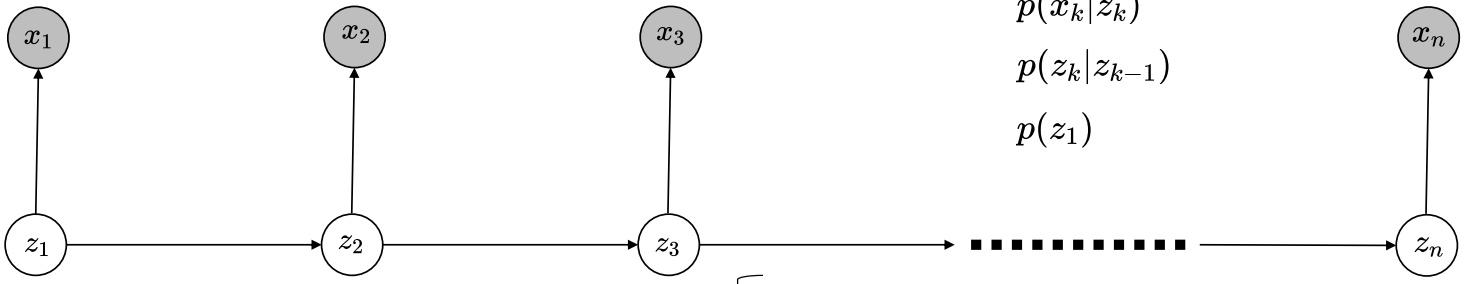
Binary random variable (optimality variable)



if "$\mathcal{O}_n$" = True, the agent tries to be optimal

if $\mathcal{O}_n$ = False, agent doesn't care

Hidden Markov Models



$p(x_k|z_k)$

$p(z_k|z_{k-1})$

$p(z_1)$

Forward-backward algorithm: compute $p(z_k|x_{1:n})$

Forward algorithm: compute $p(z_k|x_{1:k-1})$

Backward algorithm: compute $p(x_{k:n}|z_k)$

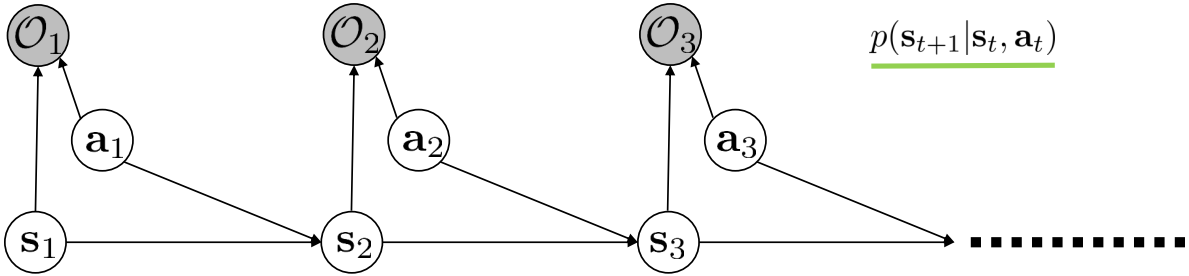
$$p(z_k|x_{1:n}) = \frac{p(z_k, x_{1:n})}{p(x_{1:k-1})p(x_{k:n})} = \frac{p(x_{k:n}|z_k, \cancel{x_{1:k-1}})p(z_k|x_{1:k-1})\cancel{p(x_{1:k-1})}}{\cancel{p(x_{1:k-1})}p(x_{k:n})} = \frac{p(x_{k:n}|z_k)p(z_k|x_{1:k-1})}{p(x_{k:n})}$$

$$\propto p(x_{k:n}|z_k)p(z_k|x_{1:k-1})$$

Backward messages

$p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) \propto \exp(r(\mathbf{s}_t, \mathbf{a}_t))$

$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$



probability that we can be optimal at steps t through T given that we take action $\mathbf{a}_t$ in state $\mathbf{s}_t$

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)$$
$$= \int p(\mathcal{O}_{t:T}, \mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) d\mathbf{s}_{t+1}$$
$$= \int p(\mathcal{O}_{t+1:T}|\mathbf{s}_{t+1}) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) d\mathbf{s}_{t+1}$$

$$p(\mathcal{O}_{t+1:T}|\mathbf{s}_{t+1}) = \int \underbrace{p(\mathcal{O}_{t+1:T}|\mathbf{s}_{t+1}, \mathbf{a}_{t+1})}_{\beta_t(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})} \cancel{p(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})} d\mathbf{a}_{t+1}$$

which actions are likely *a priori*
(assume uniform for now)

for $t = T - 1$ to 1:

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}[\beta_{t+1}(\mathbf{s}_{t+1})]$$

$$\beta_t(\mathbf{s}_t) = E_{\mathbf{a}_t \sim p(\mathbf{a}_t|\mathbf{s}_t)}[\beta_t(\mathbf{s}_t, \mathbf{a}_t)]$$

Slide adapted from Sergey Levine

A closer look at the backward pass

for $t = T-1$ to 1:

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}[\beta_{t+1}(\mathbf{s}_{t+1})]$$

$$\beta_t(\mathbf{s}_t) = E_{\mathbf{a}_t \sim p(\mathbf{a}_t|\mathbf{s}_t)}[\beta_t(\mathbf{s}_t, \mathbf{a}_t)]$$

for stochastic case we can get rid of the expectation

"optimistic" transition (something is a little off)

let $V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \overbrace{\log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]}$$

let $Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$

deterministic transition: $Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + V_{t+1}(\mathbf{s}_{t+1})$

we'll come back to the stochastic case later!

$$V_t(\mathbf{s}_t) = \log \int \underbrace{\exp(Q_t(\mathbf{s}_t, \mathbf{a}_t))}d\mathbf{a}_t$$

dominated by the highest $Q_t(s_t, a_t)$ because of exponent

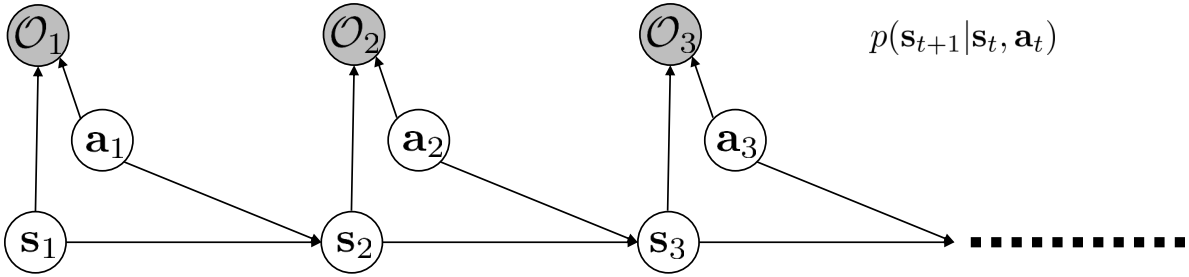
$V_t(\mathbf{s}_t) \to \max_{\mathbf{a}_t} Q_t(\mathbf{s}_t, \mathbf{a}_t)$ as $Q_t(\mathbf{s}_t, \mathbf{a}_t)$ gets bigger!

Slide adapted from Sergey Levine

Policy computation

$$p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) \propto \exp(r(\mathbf{s}_t, \mathbf{a}_t))$$

$$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$



2. compute policy $p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T})$

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)$$

$$\beta_t(\mathbf{s}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t)$$

$$\begin{aligned}
p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T}) &= \pi(\mathbf{a}_t|\mathbf{s}_t) \\
&= p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{t:T}) \\
&= \frac{p(\mathbf{a}_t, \mathbf{s}_t|\mathcal{O}_{t:T})}{p(\mathbf{s}_t|\mathcal{O}_{t:T})} \\
&= \frac{p(\mathcal{O}_{t:T}|\mathbf{a}_t, \mathbf{s}_t)p(\mathbf{a}_t, \mathbf{s}_t)/\cancel{p(\mathcal{O}_{t:T})}}{p(\mathcal{O}_{t:T}|\mathbf{s}_t)p(\mathbf{s}_t)/\cancel{p(\mathcal{O}_{t:T})}} \\
&= \frac{p(\mathcal{O}_{t:T}|\mathbf{a}_t, \mathbf{s}_t)}{p(\mathcal{O}_{t:T}|\mathbf{s}_t)}\frac{p(\mathbf{a}_t, \mathbf{s}_t)}{p(\mathbf{s}_t)} = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}\cancel{p(\mathbf{a}_t|\mathbf{s}_t)}
\end{aligned}$$

$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}$$

Slide adapted from Sergey Levine

Policy computation with value functions

for $t = T - 1$ to 1:

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]$$

$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t))\mathbf{a}_t$$

$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}$$

$$V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$$

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$$

$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t) - V_t(\mathbf{s}_t)) = \exp(\underbrace{A_t(\mathbf{s}_t, \mathbf{a}_t)}_{\text{advantage function}})$$

Slide adapted from Sergey Levine

The Plan

Variational inference review

Control as inference

Control as variational inference

Meta-RL as variational inference

Control via variational inference

let $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) q(\mathbf{a}_t|\mathbf{s}_t)$

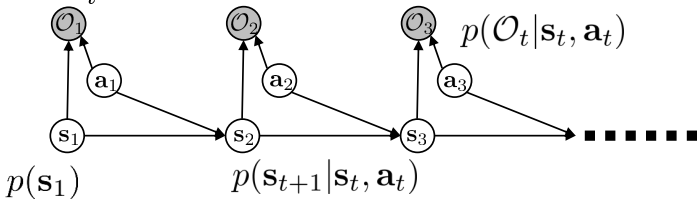
same dynamics and initial state as p

only new thing

let $\mathbf{x} = \mathcal{O}_{1:T}$ and $\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$

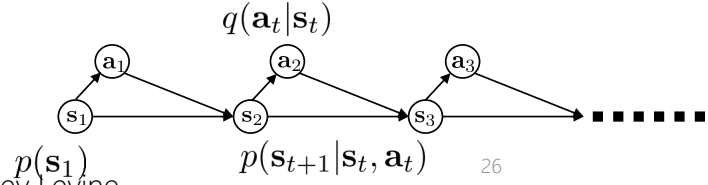
$$p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}, \mathcal{O}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t)$$

$p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}, \mathcal{O}_{1:T})$



$p(\mathbf{z}, \mathbf{x})$

$q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$



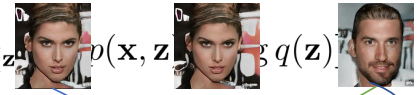
$q(\mathbf{z})$

Slide adapted from Sergey Levine

The variational lower bound

$$\log p(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}|\mathbf{z}) - \mathbb{D}_{KL}(q(\mathbf{z})||p(\mathbf{z}))$$

$$\log p(\mathbf{x}) \geq E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z})$$



let $\mathbf{x} = \mathcal{O}_{1:T}$ and $\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$

$$p(\mathbf{s}_{1:T},\mathbf{a}_{1:T},\mathcal{O}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t) p(\mathcal{O}_t|\mathbf{s}_t,\mathbf{a}_t)$$

let $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) q(\mathbf{a}_t|\mathbf{s}_t)$

$$\log p(\mathcal{O}_{1:T}) \geq E_{(\mathbf{s}_{1:T},\mathbf{a}_{1:T})\sim q}\Big[\cancel{\log p(\mathbf{s}_1)} + \sum_{t=1}^{T} \cancel{\log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)} + \sum_{t=1}^{T} \log p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t)$$

$$- \cancel{\log p(\mathbf{s}_1)} - \sum_{t=1}^{T} \cancel{\log p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)} - \sum_{t=1}^{T} \log q(\mathbf{a}_t|\mathbf{s}_t)\Big]$$

$$= E_{(\mathbf{s}_{1:T},\mathbf{a}_{1:T})\sim q}\left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) - \log q(\mathbf{a}_t|\mathbf{s}_t)\right]$$

$$= \sum_t E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q}\left[r(\mathbf{s}_t, \mathbf{a}_t) + \mathcal{H}(q(\mathbf{a}_t|\mathbf{s}_t))\right] \longleftarrow \text{maximize reward and maximize action entropy!}$$

Slide adapted from Sergey Levine

Summary

Objective:

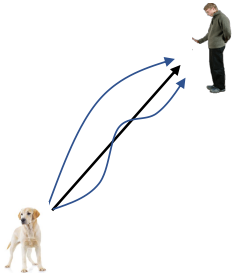
$$\sum_t E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q}\left[r(\mathbf{s}_t,\mathbf{a}_t)+\mathcal{H}(q(\mathbf{a}_t|\mathbf{s}_t))\right]$$

Value-, Q-functions, and the policy

$$V_t(\mathbf{s}_t)=\log\int\exp(Q_t(\mathbf{s}_t,\mathbf{a}_t))\mathbf{a}_t$$

$$\cancel{Q_t(\mathbf{s}_t,\mathbf{a}_t)=r(\mathbf{s}_t,\mathbf{a}_t)+\log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]} \longrightarrow Q_t(\mathbf{s}_t,\mathbf{a}_t)=r(\mathbf{s}_t,\mathbf{a}_t)+E[(V_{t+1}(\mathbf{s}_{t+1})]$$

$$\pi(\mathbf{a}_t|\mathbf{s}_t)=\exp(Q_t(\mathbf{s}_t,\mathbf{a}_t)-V_t(\mathbf{s}_t))=\exp(A_t(\mathbf{s}_t,\mathbf{a}_t))$$



Q-learning

1. collect dataset $\{(\mathbf{s}_i,\mathbf{a}_i,\mathbf{s}_i',r_i)\}$
2. set $\mathbf{y}_i\leftarrow r(\mathbf{s}_i,\mathbf{a}_i)+\gamma\max_{\mathbf{a}_i'}Q_\phi(\mathbf{s}_i',\mathbf{a}_i')$
3. set $\phi\leftarrow\arg\min_\phi\frac{1}{2}\sum_i\|Q_\phi(\mathbf{s}_i,\mathbf{a}_i)-\mathbf{y}_i\|^2$

$K\times$

$$\pi(\mathbf{a}|\mathbf{s})=\arg\max_{\mathbf{a}}Q_\phi(\mathbf{s},\mathbf{a})$$

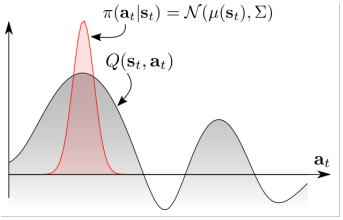
Soft Q-learning

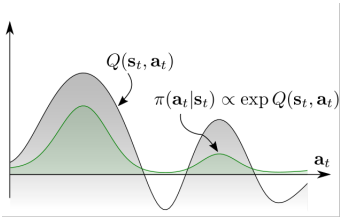
1. collect dataset $\{(\mathbf{s}_i,\mathbf{a}_i,\mathbf{s}_i',r_i)\}$
2. set $\mathbf{y}_i\leftarrow r(\mathbf{s}_i,\mathbf{a}_i)+\gamma\,\overset{\text{softmax}}{\cancel{\max}}_{\mathbf{a}_i'}Q_\phi(\mathbf{s}_i',\mathbf{a}_i')$
3. set $\phi\leftarrow\arg\min_\phi\frac{1}{2}\sum_i\|Q_\phi(\mathbf{s}_i,\mathbf{a}_i)-\mathbf{y}_i\|^2$

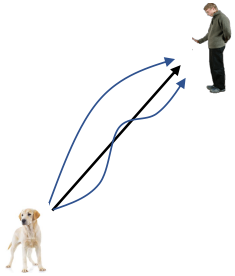
$K\times$

$$\pi(\mathbf{a}|\mathbf{s})=\cancel{\arg\max_{\mathbf{a}}Q_\phi(\mathbf{s},\mathbf{a})}\quad \exp(A_t(s_t,a_t))$$

Soft Q-learning



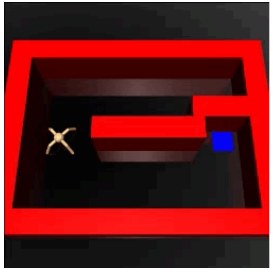


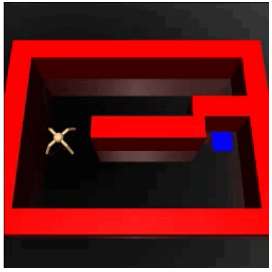


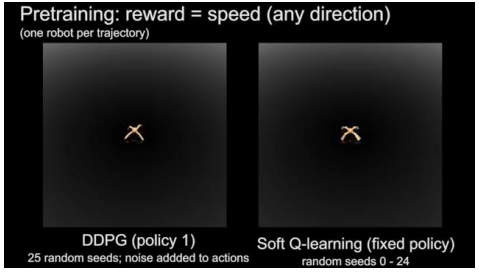
Exploration

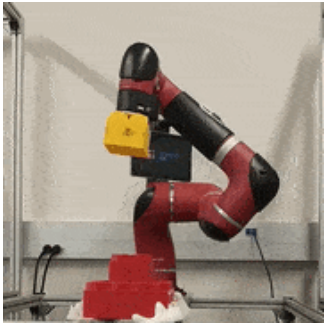
Fine-tunability

Robustness









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

The Plan

Variational inference review

Control as inference

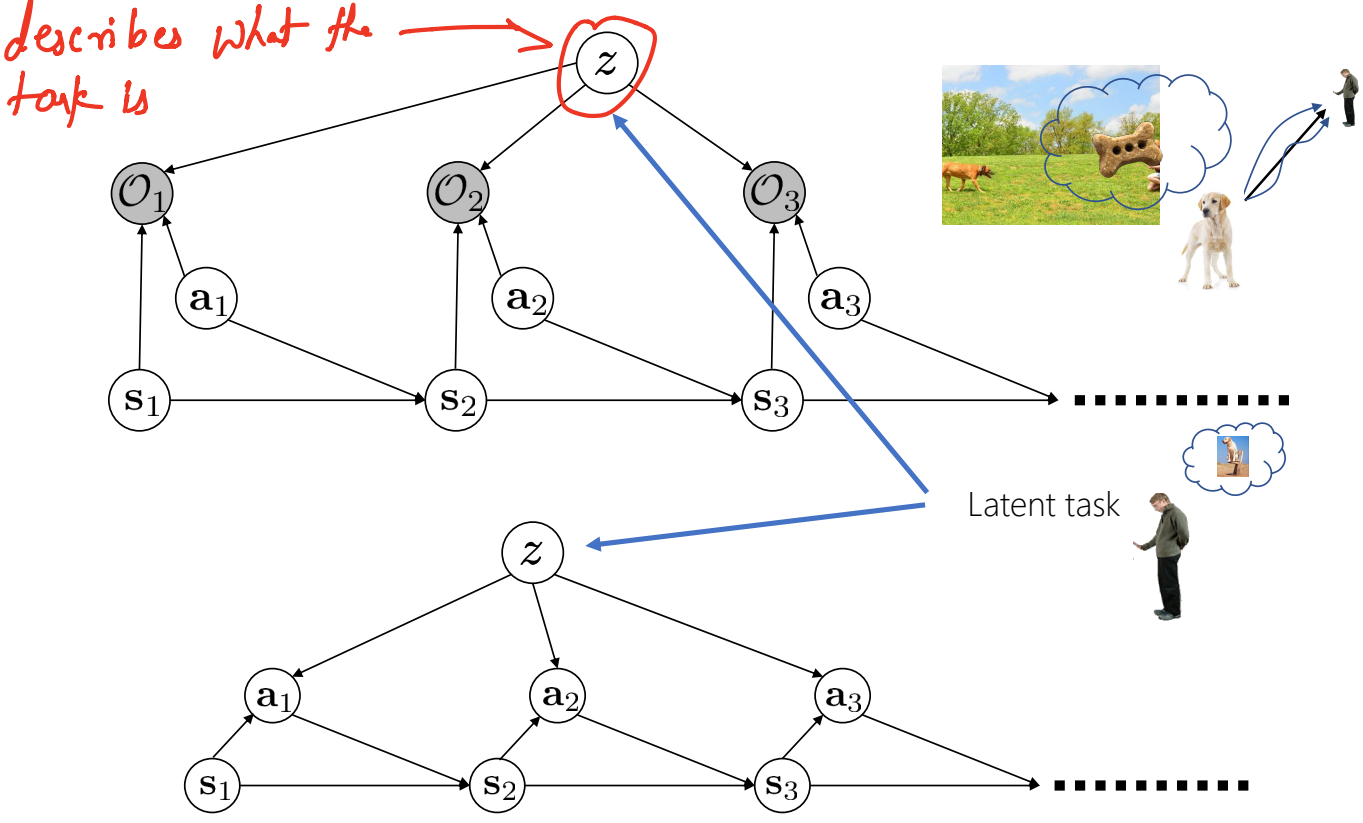
Control as variational inference

Meta-RL as variational inference

Meta-RL via Variational Inference

What's a good strategy for the dog?

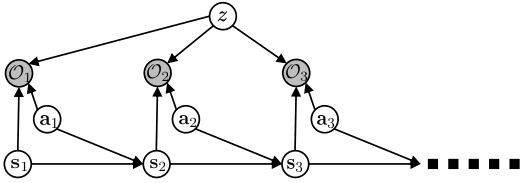
$$\max_{\phi,\theta} \frac{1}{N} \sum_i \mathbb{E}_{z \sim q_\phi(z|x_i)} \log p_\theta(x_i|z) - \mathbb{D}_{KL}(q_\phi(z|x_i)||p(z))$$

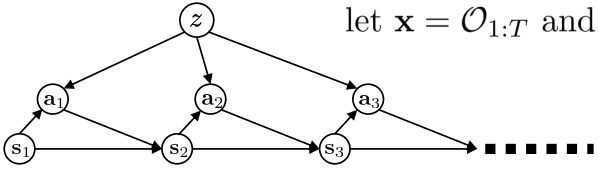

describes what the task is
z
$\mathcal{O}_1$
$\mathcal{O}_2$
$\mathcal{O}_3$
$\mathbf{a}_1$
$\mathbf{a}_2$
$\mathbf{a}_3$
$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_3$
Latent task
z
$\mathbf{a}_1$
$\mathbf{a}_2$
$\mathbf{a}_3$
$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_3$

Variational Inference Again!

let $\mathbf{x} = \mathcal{O}_{1:T}$ and $\cancel{\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}$

let $\mathbf{x} = \mathcal{O}_{1:T}$ and $\mathbf{z} =$ latent task





$$\log p_\theta(x_i) \geq \mathbb{E}_{z\sim q(z|x_i)} \log p(x_i|z) - \mathbb{D}_{KL}(q(z|x_i)||p(z))$$

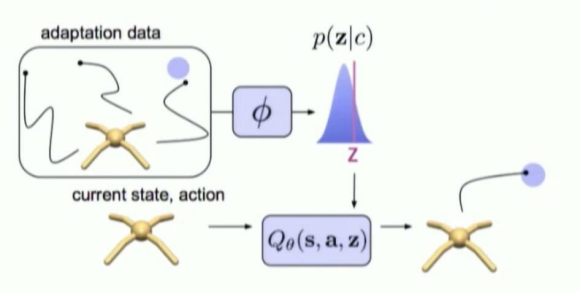
$$\log p(\mathcal{O}_{1:T}) \geq \mathbb{E}_{q(z|x_i)} \log p(\mathcal{O}_{1:T}|z) - \mathbb{D}_{KL}(q(z|x_i)||p(z))$$

$(O_{1:T}|z)$ (s_t, a_t, z) $r(s_t, a_t, z)$ $q(O_t|s_t, z)$

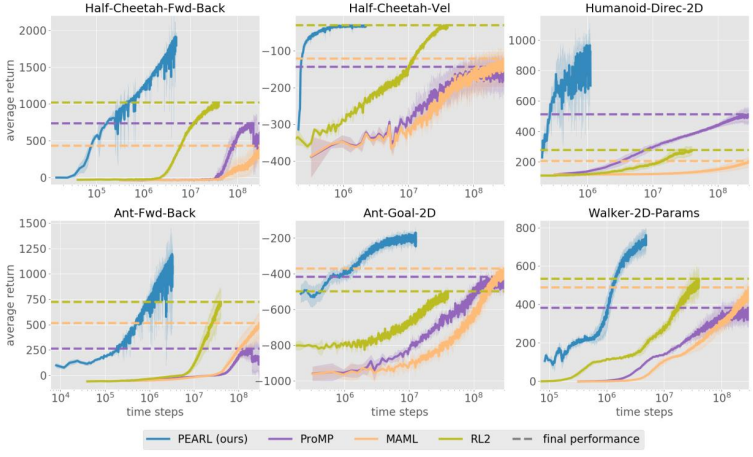
$$\log p(\cancel{\mathcal{O}_{1:T}}) \geq \sum_t E_{(\mathbf{s}_t, \cancel{\mathbf{a}_t})\sim q} [r(\cancel{\mathbf{s}_t}, \mathbf{a}_t) + \cancel{\mathcal{H}(q(\mathbf{a}_t|\mathbf{s}_t))}]$$

What's a good $\boldsymbol{x_i}$? → How about all the transitions seen so far?

PEARL

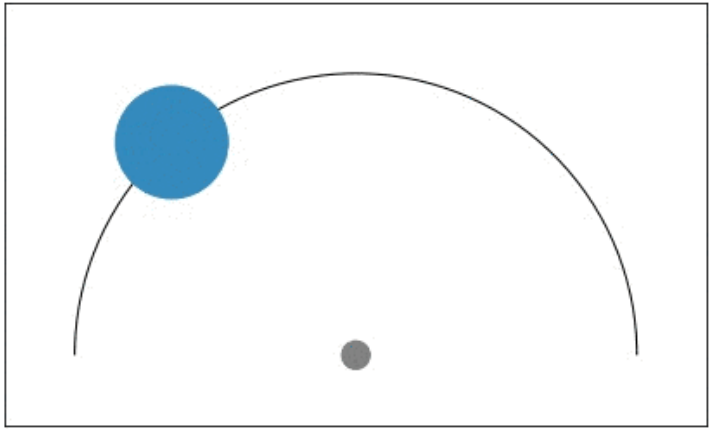


$$\mathbb{E}_{\mathcal{T}}[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{c}^{\mathcal{T}})}[R(\mathcal{T}, \mathbf{z}) + \beta D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{c}^{\mathcal{T}})||p(\mathbf{z}))]]$$



Rakelly et al. PEARL, 2019

PEARL



Rakelly et al. PEARL, 2019

The Plan

Variational inference review

Control as inference

Control as variational inference

Meta-RL as variational inference

Next Week

What if we want the agent to **come up with the tasks**?

What about **hierarchies** of tasks?

Hierarchical RL and Skill Discovery - **Nov 2**

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

Lifelong learning – **Nov 4**

Additional Resources

RL and Control as Probabilistic Inference, Levine, 2018

Learning to Learn with Probabilistic Task Embeddings, BAIR blog post,

Berkeley CS285: Deep Reinforcement Learning