REINFORCEMENT LEARNING AND CONTROL AS PROBABILISTIC INFERENCE: TUTORIAL AND REVIEW

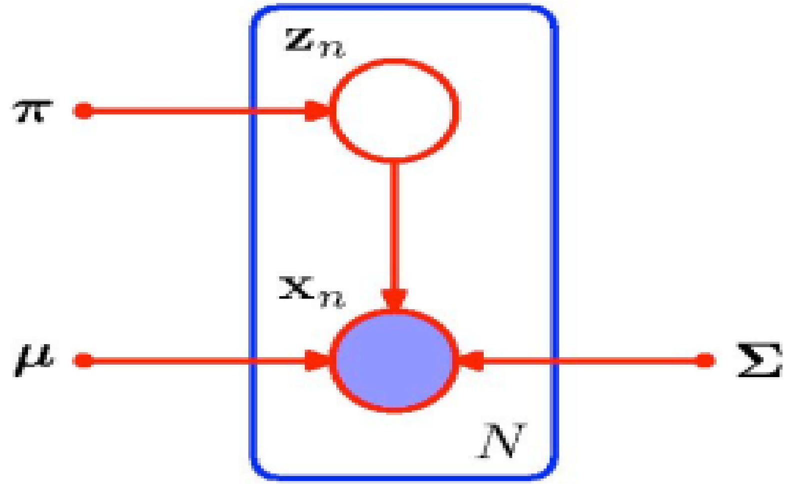
Alireza Nouri

Table of contents

An Alternative View of EM

- The goal of EM is to find maximum likelihood solutions for models with latent variables.
- We represent the observed dataset as an N by D matrix **X**.
- Latent variables will be represented and an N by K matrix **Z**.
- The set of all model parameters is denoted by θ.
- The log-likelihood takes form:

$$\ln p(\mathbf{X}|\theta) = \ln \left[\sum_{Z} p(\mathbf{X}, \mathbf{Z}|\theta) \right].$$



- Note: even if the joint distribution belongs to exponential family, the marginal typically does not!
- We will call:

$\{\mathbf{X}, \mathbf{Z}\}$ as complete dataset.

$\{\mathbf{X}\}$ as incomplete dataset.

Variational Bound

- Given a joint distribution p(**Z**,**X**|θ) over observed and latent variables governed by parameters θ, the goal is to maximize the likelihood function p(**X**|θ) with respect to θ:

$$p(\mathbf{X}|\theta) = \sum_{Z} p(\mathbf{X}, \mathbf{Z}|\theta).$$

- We will assume that **Z** is discrete, although derivations are identical if **Z** contains continuous, or a combination of discrete and continuous variables.
- For any distribution q(**Z**) over latent variables we can derive the following variational lower bound:

$$\ln p(\mathbf{X}|\theta) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) = \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}$$

Jensen's inequality

$$\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} = \mathcal{L}(q, \theta).$$

The variational approximation

$$\log p(x_i) \geq \overbrace{E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] + \mathcal{H}(q_i)}^{\mathcal{L}_i(p, q_i)}$$

what makes a good $q_i(z)$?

intuition: $q_i(z)$ should approximate $p(z|x_i)$

approximate in what sense?

compare in terms of KL-divergence: $D_{\mathrm{KL}}(q_i(z)\|p(z|x))$

why?

$$D_{\mathrm{KL}}(q_i(x_i)\|p(z|x_i)) = E_{z\sim q_i(z)}\left[\log \frac{q_i(z)}{p(z|x_i)}\right] = E_{z\sim q_i(z)}\left[\log \frac{q_i(z)p(x_i)}{p(x_i, z)}\right]$$

$$= -E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] + E_{z\sim q_i(z)}[\log q_i(z)] + E_{z\sim q_i(z)}[\log p(x_i)]$$

$$= -E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] - \mathcal{H}(q_i) + \log p(x_i)$$

$$= -\mathcal{L}_i(p, q_i) + \log p(x_i)$$

$$\log p(x_i) = D_{\mathrm{KL}}(q_i(z)\|p(z|x_i)) + \mathcal{L}_i(p, q_i)$$

$$\log p(x_i) \geq \mathcal{L}_i(p, q_i)$$

The variational approximation

$$\log p(x_i) \geq \overbrace{E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] + \mathcal{H}(q_i)}^{\mathcal{L}_i(p,q_i)}$$

$$\log p(x_i) = D_{\mathrm{KL}}(q_i(z)\|p(z|x_i)) + \mathcal{L}_i(p,q_i)$$

$$\log p(x_i) \geq \mathcal{L}_i(p,q_i)$$

$$D_{\mathrm{KL}}(q_i(z)\|p(z|x_i)) = E_{z\sim q_i(z)}\left[\log \frac{q_i(z)}{p(z|x_i)}\right] = E_{z\sim q_i(z)}\left[\log \frac{q_i(z)p(x_i)}{p(x_i,z)}\right]$$

$$= \underbrace{-E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] - \mathcal{H}(q_i)}_{-\mathcal{L}_i(p,q_i)} + \log p(x_i)$$

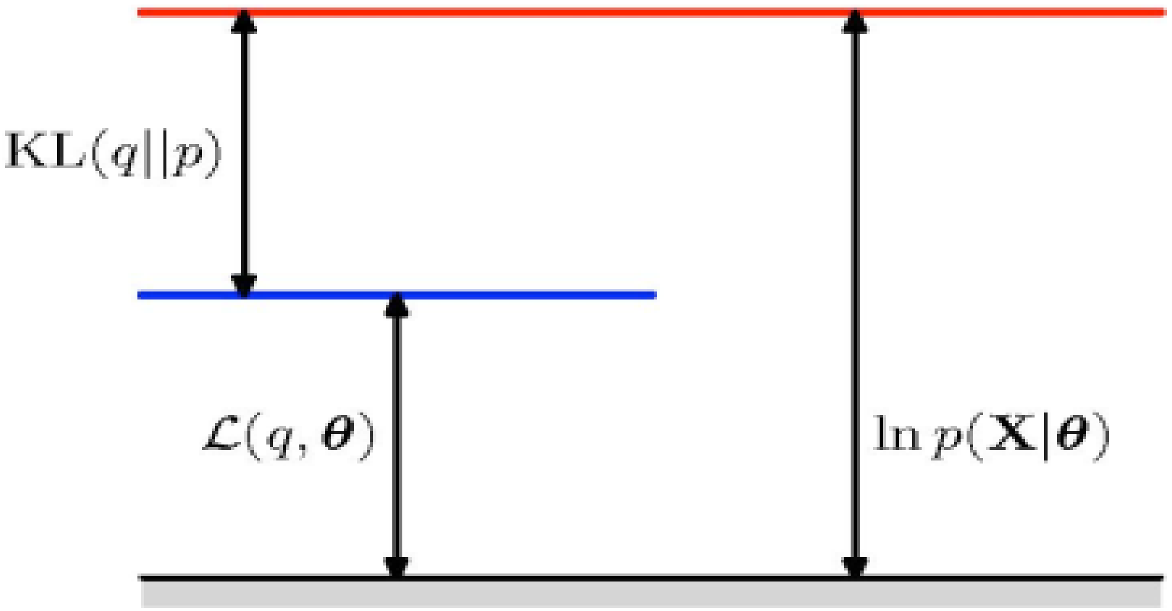
independent of q_i!

$\Rightarrow$ maximizing $\mathcal{L}_i(p,q_i)$ w.r.t. q_i minimizes KL-divergence!

Decomposition

- Illustration of the decomposition which holds for any distribution q(**Z**).

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + \mathrm{KL}(q||p),$$

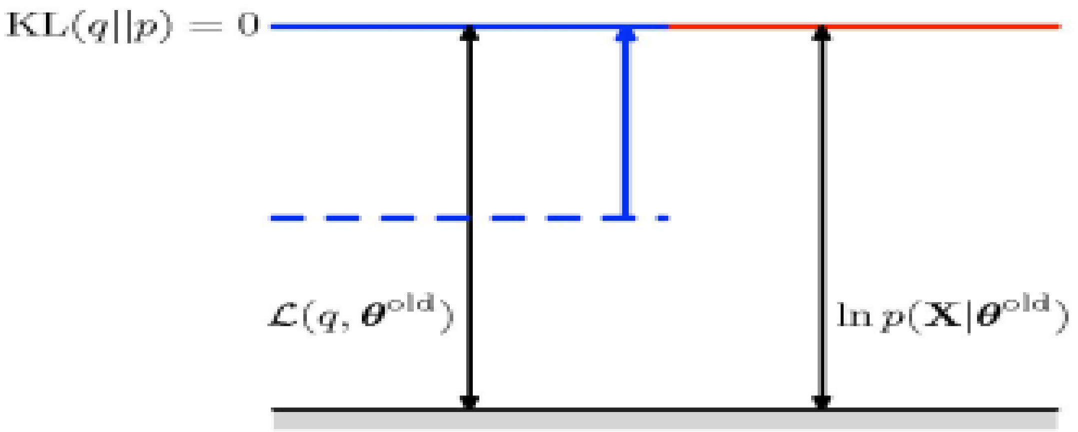


E-step

- Suppose that the current value of the parameter vector is θ^{old}.
- In the E-step, we maximize the lower with respect to q while holding parameters θ^{old} fixed.

$$\mathcal{L}(q, \theta^{old}) = \ln p(\mathbf{X}|\theta^{old}) - \mathbf{KL}(q||p).$$

does not
depend on q



- The lower-bound is maximized when KL term turns to zero.
- In other words, when q(**Z**) is equal to the true posterior:

$$q(\mathbf{Z}) = \mathbf{p}(\mathbf{Z}|\mathbf{X}, \theta^{\mathbf{old}}).$$

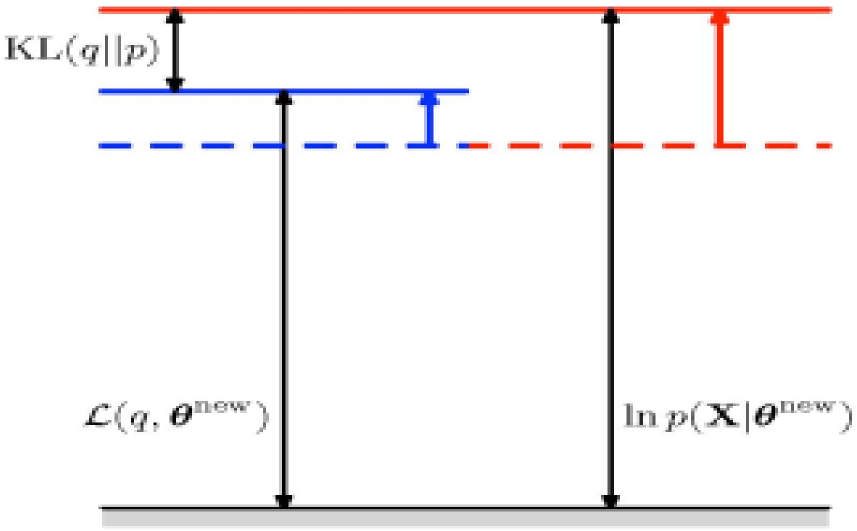
- The lower bound will become equal to the log-likelihood.

M-step

- In the M-step, the lower bound is maximized with respect to parameters θ while holding the distribution q fixed.

does not depend on θ.

$$\mathcal{L}(q,\theta) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln p(\mathbf{X},\mathbf{Z}|\theta) + \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\theta^{old}) \ln \frac{1}{p(\mathbf{Z}|\mathbf{X},\theta^{old})}.$$



$$\mathcal{L}(q,\theta) = Q(\theta,\theta^{old}) + \text{const.}$$

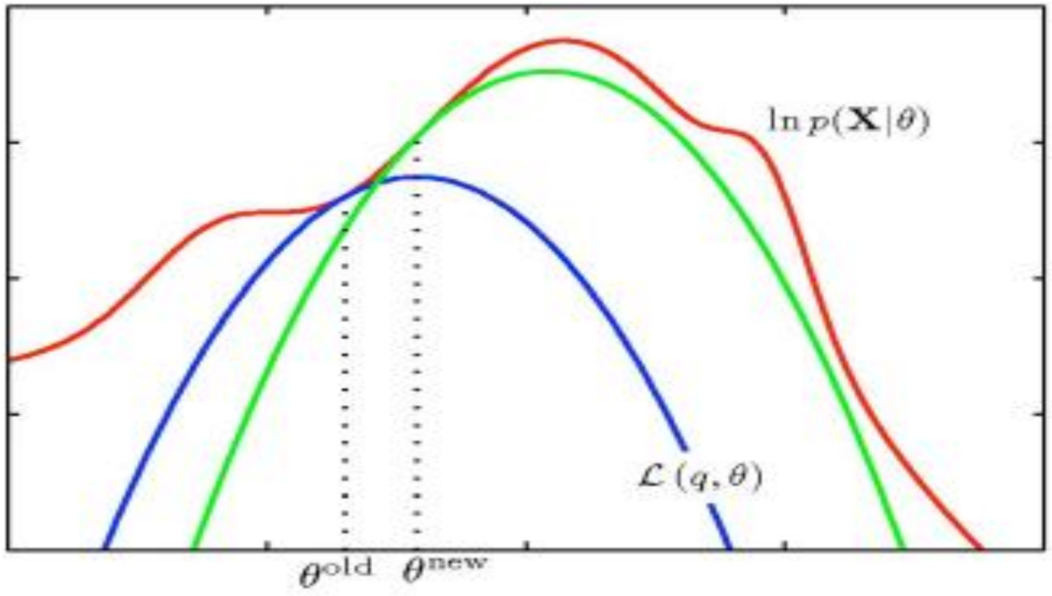
- Hence the M-step amounts to maximizing the expected complete log-likelihood.

$$\theta^{new} = \arg\max_{\theta} Q(\theta,\theta^{old}).$$

- Because KL divergence is non-negative, this causes the log-likelihood log p(**X** | θ) to increase by at least as much as the lower bound does.

Bound Optimization

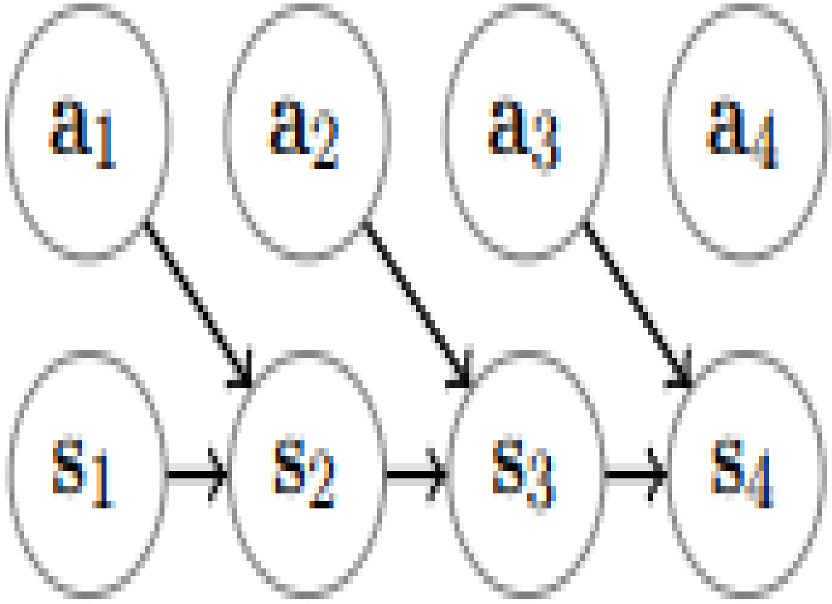
• The EM algorithm belongs to the general class of bound optimization methods:



• At each step, we compute:

- E-step: a lower bound on the log-likelihood function for the current parameter values. The bound is concave with unique global optimum.
- M-step: maximize the lower-bound to obtain the new parameter values.

HARD OPTIMIZATION IN RL



(a) graphical model with states and actions

$$\theta^\star = \arg\max_\theta \sum_{t=1}^{T} E_{(\mathbf{s}_t, \mathbf{a}_t) \sim p(\mathbf{s}_t, \mathbf{a}_t | \theta)} [r(\mathbf{s}_t, \mathbf{a}_t)].$$

$$p(\tau) = p(\mathrm{s}_1, \mathrm{a}_t, \ldots, \mathrm{s}_T, \mathrm{a}_T | \theta) = p(\mathrm{s}_1) \prod_{t=1}^{T} p(\mathrm{a}_t | \mathrm{s}_t, \theta) p(\mathrm{s}_{t+1} | \mathrm{s}_t, \mathrm{a}_t).$$

Policy

Transition

Some Notations

$$p(a_t|s_t) = \pi(a_t|s_t)$$

Policy

$$v_\pi(s) = \sum_{t=1}^{T} E(\gamma^t r(t+1)|S_t = s), \textit{ for all } s \epsilon \mathcal{S}$$

value function

$$q_\pi(s,a) = \sum_{t=1}^{T} E_\pi(\gamma^t r(t+1)|\, S_t = s, A_t = a)$$

action-value function

value iteration algorithm:

1. set $Q(\mathbf{s},\mathbf{a}) \leftarrow r(\mathbf{s},\mathbf{a}) + \gamma E[V(\mathbf{s}')]$
2. set $V(\mathbf{s}) \leftarrow \max_{\mathbf{a}} Q(\mathbf{s},\mathbf{a})$

Control as Approximate Inference in PGM

1. Does reinforcement learning and optimal control provide a reasonable model of human behavior?

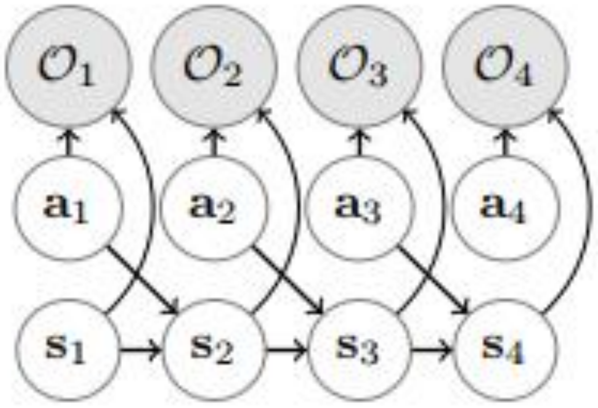
2. Is there a better explanation?

3. Can we derive optimal control, reinforcement learning, and planning as *probabilistic inference*?

4. How does this change our RL algorithms?

5. (next lecture) We'll see this is crucial for *inverse* reinforcement learning

- **Goals:**
  - **Understand the connection between inference and control**
  - **Understand how specific RL algorithms can be instantiated in this framework**
  - **Understand why this might be a good idea**



(b) graphical model with optimality variables

$$\mathbf{a}_1, \ldots, \mathbf{a}_T = \arg \max_{\mathbf{a}_1, \ldots, \mathbf{a}_T} \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)$$

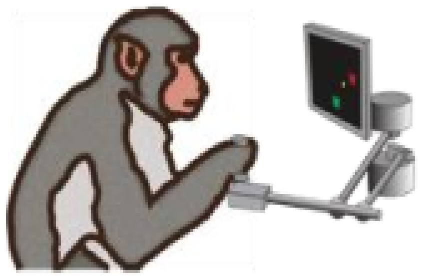
$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t)$$

optimize this to explain the data

$$\pi = \arg \max_{\pi} E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t), \mathbf{a}_t \sim \pi(\mathbf{a}_t|\mathbf{s}_t)} [r(\mathbf{s}_t, \mathbf{a}_t)]$$

$$\mathbf{a}_t \sim \pi(\mathbf{a}_t|\mathbf{s}_t)$$

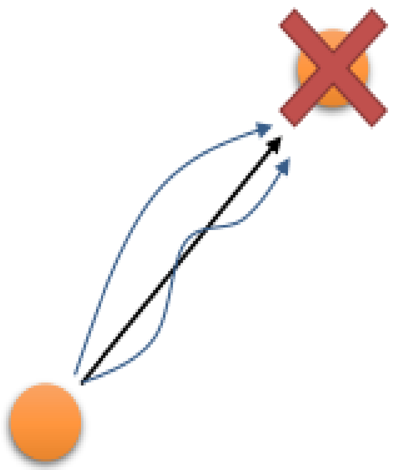
What if the data is **not** optimal?



some mistakes matter more than others!

behavior is **stochastic**

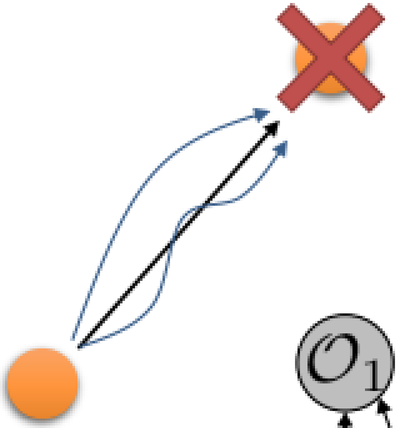
but good behavior is still the most likely



A probabilistic graphical model of decision making

$$\cancel{\mathbf{a}_1, \ldots, \mathbf{a}_T = \arg\max_{\mathbf{a}_1, \ldots, \mathbf{a}_T} \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)}$$

$$\cancel{\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t)}$$

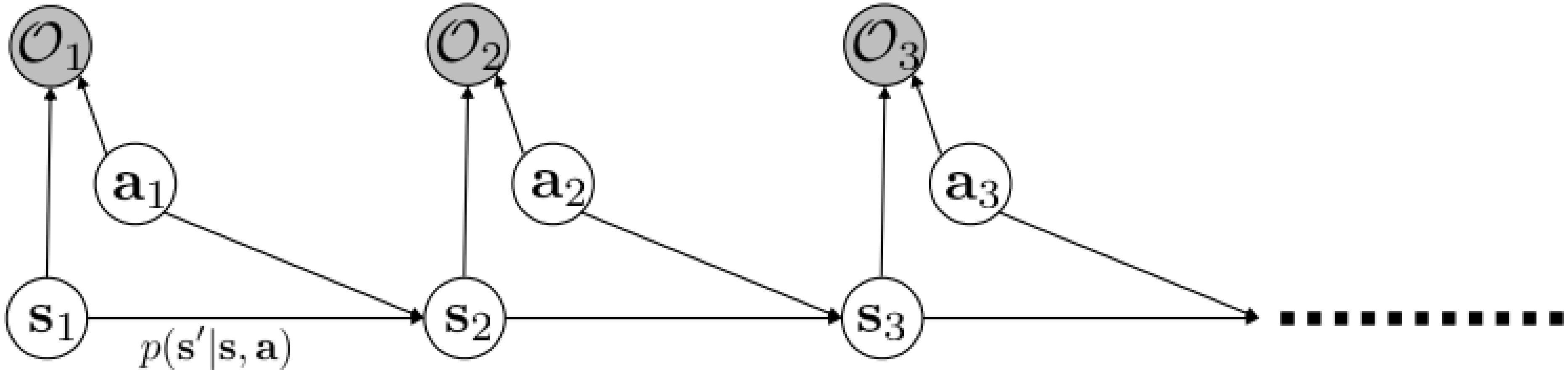


$p(\underbrace{\mathbf{s}_{1:T}, \mathbf{a}_{1:T}}_{\tau}) = ??$ no assumption of optimal behavior!

$p(\tau|\mathcal{O}_{1:T})$ $\quad$ $p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) = \exp(r(\mathbf{s}_t, \mathbf{a}_t))$

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

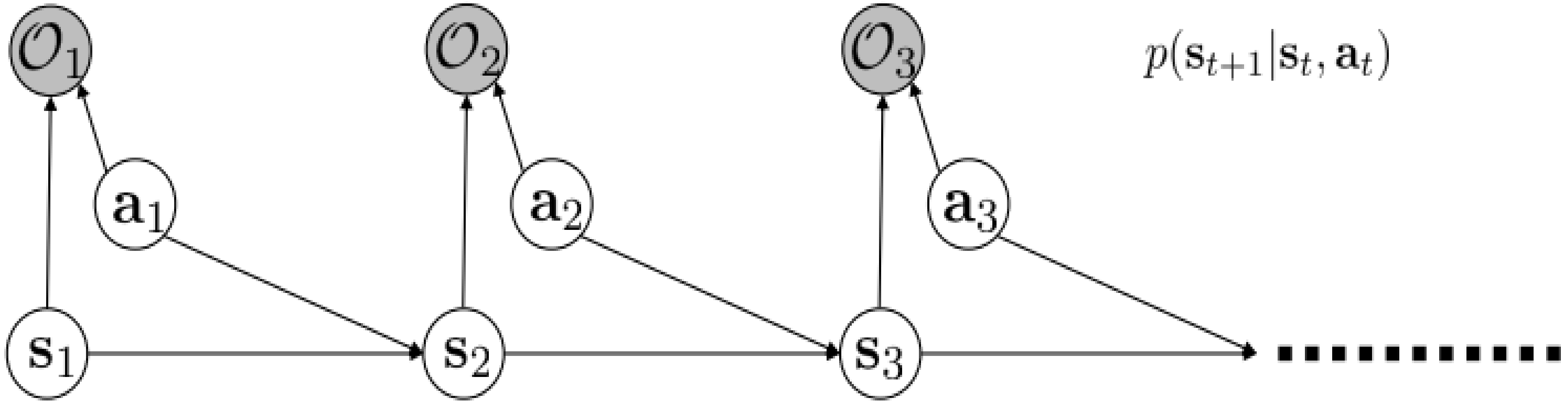
$$\propto p(\tau) \prod_t \exp(r(\mathbf{s}_t, \mathbf{a}_t)) = p(\tau) \exp\left(\sum_t r(\mathbf{s}_t, \mathbf{a}_t)\right)$$



Inference = planning

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

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



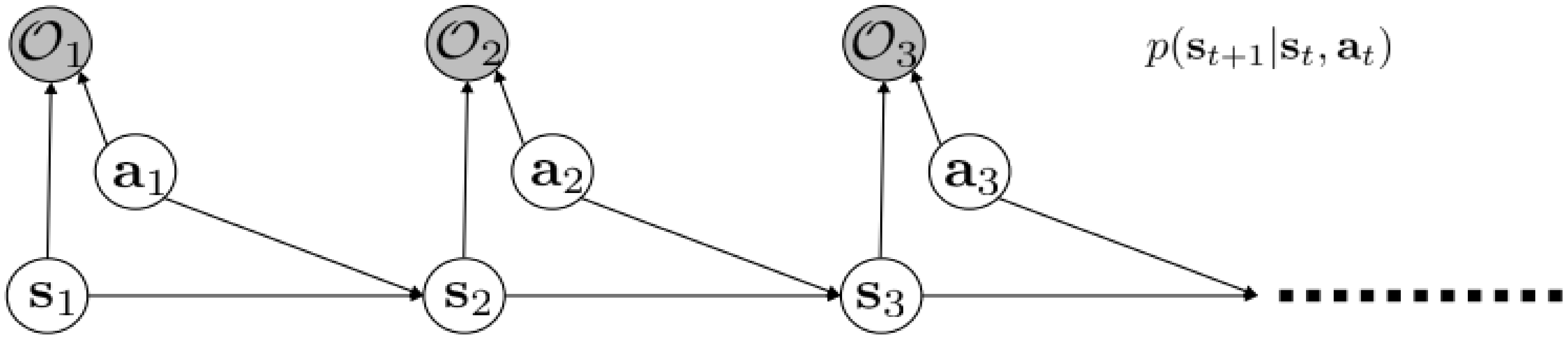
how to do inference?

1. compute backward messages $\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)$
2. compute policy $p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T})$
3. compute forward messages $\alpha_t(\mathbf{s}_t) = p(\mathbf{s}_t|\mathcal{O}_{1:t-1})$

Inference = planning

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

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



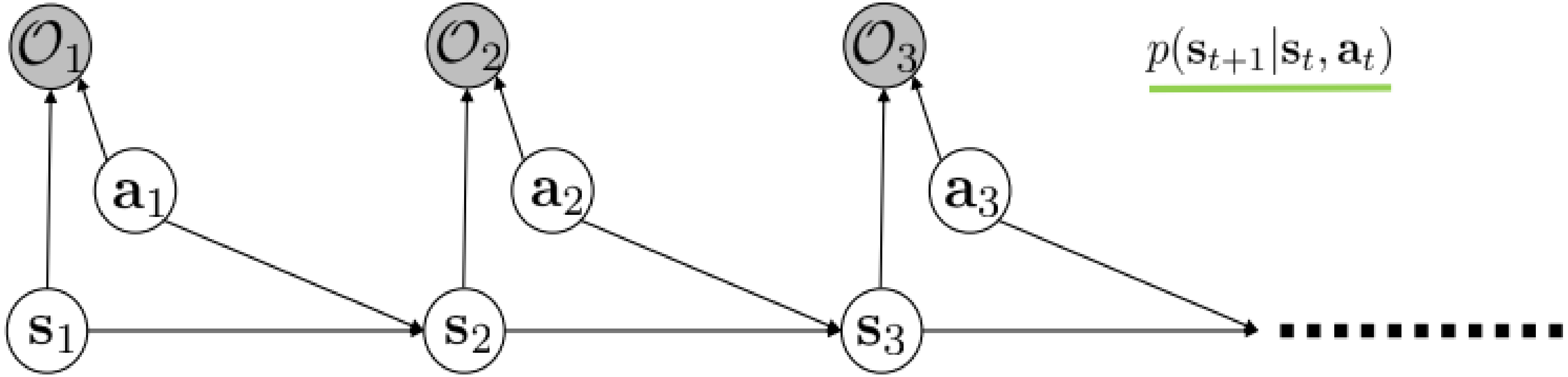
how to do inference?

1. compute backward messages $\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)$
2. compute policy $p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T})$
3. compute forward messages $\alpha_t(\mathbf{s}_t) = p(\mathbf{s}_t|\mathcal{O}_{1:t-1})$

Backward messages

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

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



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

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

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

for $t = T-1$ to 1:

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

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

A closer look at the backward pass

value iteration algorithm:

1. set $Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E[V(\mathbf{s}')]$
2. set $V(\mathbf{s}) \leftarrow \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

for $t = T-1$ to 1:

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

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

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

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

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \overbrace{\log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]}^{\text{“optimistic” transition (not a good idea!)}}$$

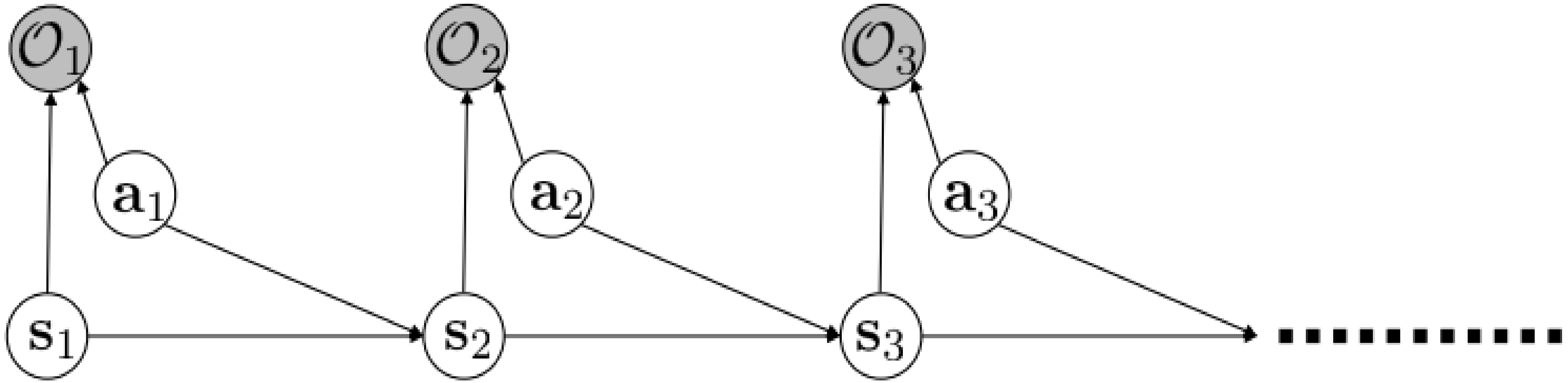
deterministic transition: $Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + V_{t+1}(\mathbf{s}_{t+1})$

we'll come back to the stochastic case later!

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

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

Backward pass summary



$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t)$ probability that we can be optimal at steps t through T given that we take action $\mathbf{a}_t$ in state $\mathbf{s}_t$

for $t = T - 1$ to 1:

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

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

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

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

log of β_t is "Q-function-like"

The action prior

remember this?

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

("soft max")

what if the action prior is not uniform?

$$V(\mathbf{s}_t) = \log \int \exp(Q(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t))\mathbf{a}_t$$

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

$$\text{let } \tilde{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t) + \log E[\exp(V(\mathbf{s}_{t+1}))]$$

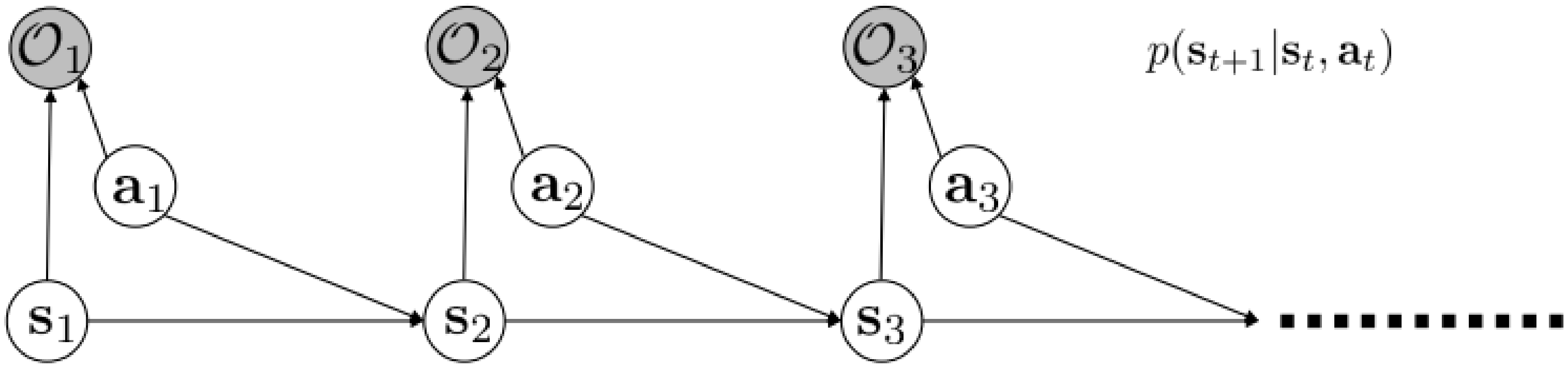
$$V(\mathbf{s}_t) = \log \int \exp(\tilde{Q}(\mathbf{s}_t, \mathbf{a}_t))\mathbf{a}_t \quad \Leftrightarrow \quad V(\mathbf{s}_t) = \log \int \exp(Q(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t))\mathbf{a}_t$$

can **always** fold the action prior into the reward! uniform action prior
can be assumed without loss of generality

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

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

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

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

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

Policy computation with value functions

for $t = T - 1$ to 1:

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

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

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

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

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

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

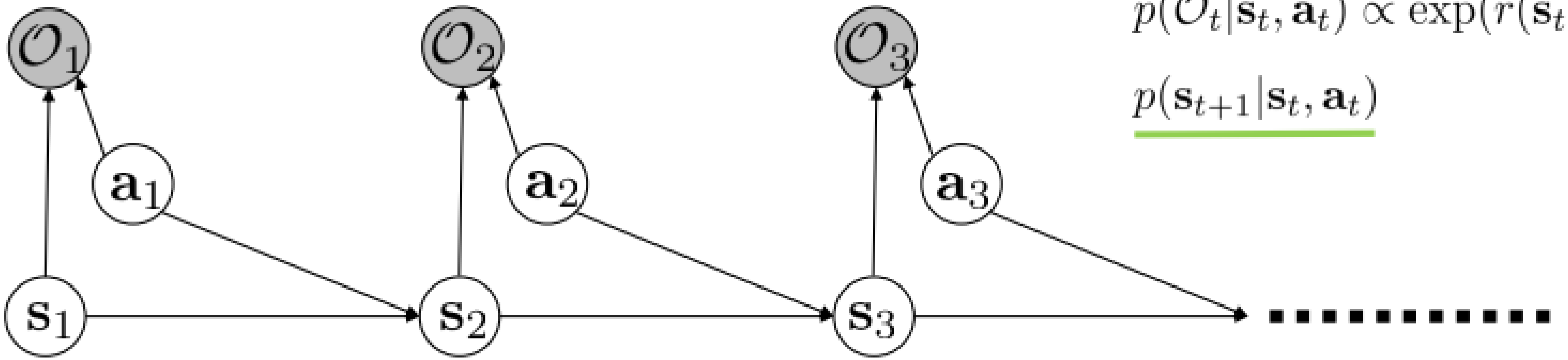
Policy computation summary

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

$$\text{with temperature: } \pi(\mathbf{a}_t|\mathbf{s}_t) = \exp(\tfrac{1}{\alpha}Q_t(\mathbf{s}_t, \mathbf{a}_t) - \tfrac{1}{\alpha}V_t(\mathbf{s}_t)) = \exp(\tfrac{1}{\alpha}A_t(\mathbf{s}_t, \mathbf{a}_t))$$

- Natural interpretation: better actions are more probable
- Random tie-breaking
- Analogous to Boltzmann exploration
- Approaches greedy policy as temperature decreases

Forward messages



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

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

$\alpha_1(\mathbf{s}_1) = p(\mathbf{s}_1)$ (usually known)

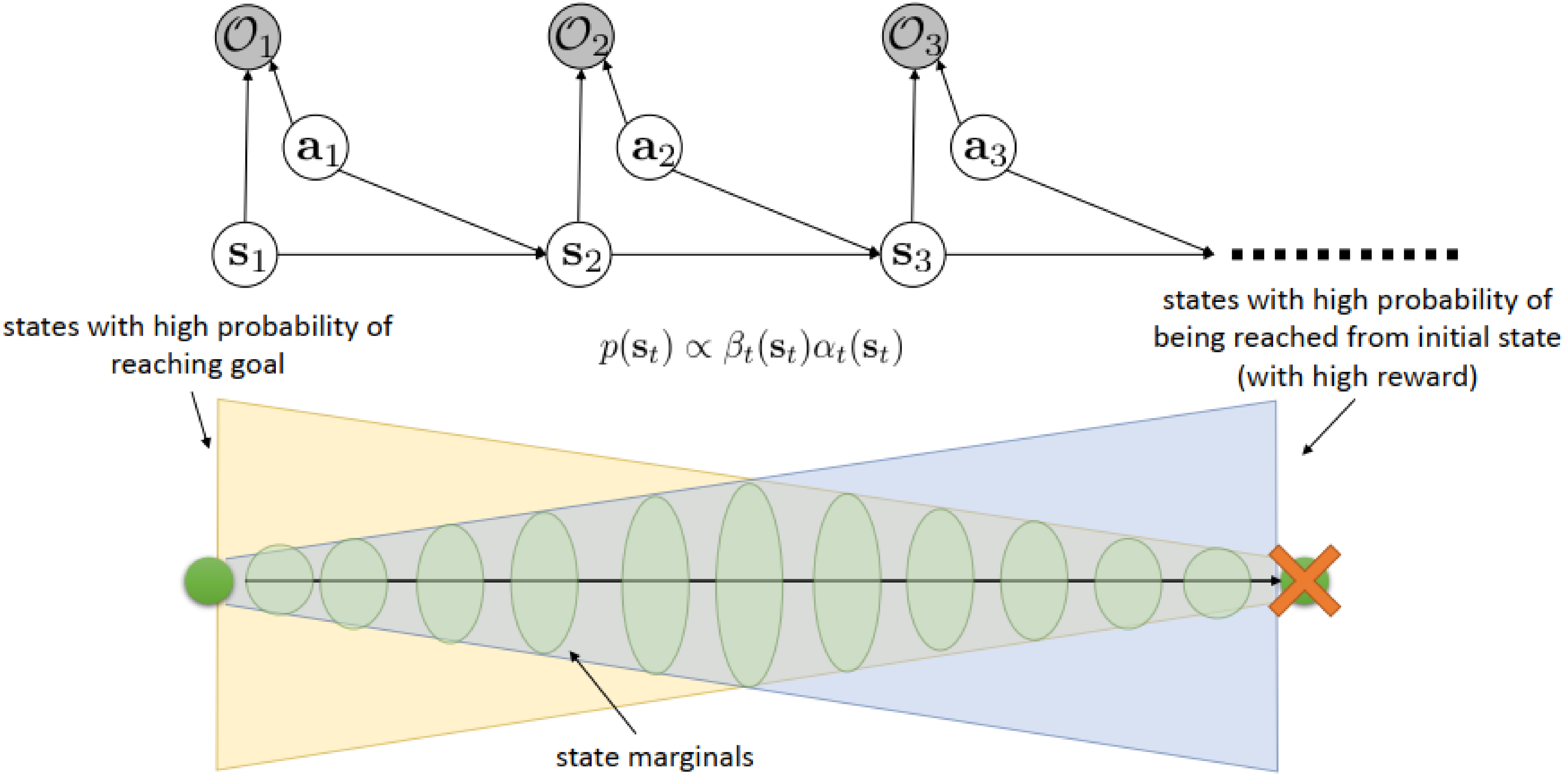
$$\begin{aligned}
\alpha_t(\mathbf{s}_t) &= p(\mathbf{s}_t|\mathcal{O}_{1:t-1}) \\
&= \int p(\mathbf{s}_t,\mathbf{s}_{t-1},\mathbf{a}_{t-1}|\mathcal{O}_{1:t-1})d\mathbf{s}_{t-1}d\mathbf{a}_{t-1} = \int p(\mathbf{s}_t|\mathbf{s}_{t-1},\mathbf{a}_{t-1},\cancel{\mathcal{O}_{1:t-1}})p(\mathbf{a}_{t-1}|\mathbf{s}_{t-1},\mathcal{O}_{1:t-1})p(\mathbf{s}_{t-1}|\mathcal{O}_{1:t-1})d\mathbf{s}_{t-1}d\mathbf{a}_{t-1} \\
&= \int \underline{p(\mathbf{s}_t|\mathbf{s}_{t-1},\mathbf{a}_{t-1})}p(\mathbf{a}_{t-1}|\mathbf{s}_{t-1},\mathcal{O}_{t-1})p(\mathbf{s}_{t-1}|\mathcal{O}_{1:t-1})d\mathbf{s}_{t-1}d\mathbf{a}_{t-1}
\end{aligned}$$

$$p(\mathbf{a}_{t-1}|\mathbf{s}_{t-1},\mathcal{O}_{t-1})p(s_{t-1}|\mathcal{O}_{1:t-1}) = \frac{p(\mathcal{O}_{t-1}|\mathbf{s}_{t-1},\mathbf{a}_{t-1})p(\mathbf{a}_{t-1}|\mathbf{s}_{t-1})}{p(\cancel{\mathcal{O}_{t-1}|\mathbf{s}_{t-1}})} \frac{p(\cancel{\mathcal{O}_{t-1}|\mathbf{s}_{t-1}})\overbrace{p(\mathbf{s}_{t-1}|\mathcal{O}_{1:t-2})}^{\alpha_{t-1}(\mathbf{s}_{t-1})}}{p(\mathcal{O}_{t-1}|\mathcal{O}_{1:t-2})}$$

what if we want $p(\mathbf{s}_t|\mathcal{O}_{1:T})$?

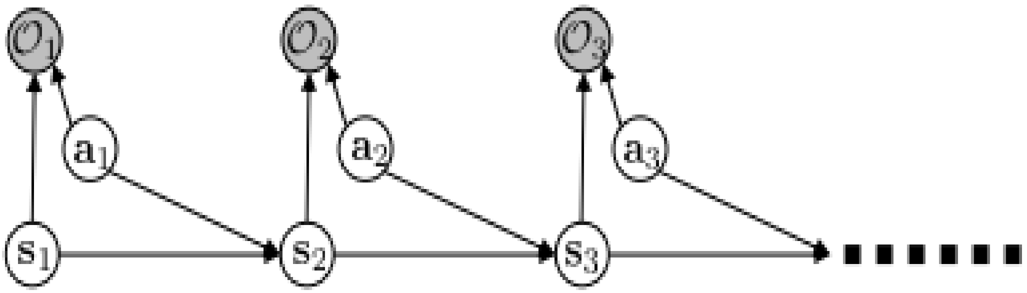
$$p(\mathbf{s}_t|\mathcal{O}_{1:T}) = \frac{p(\mathbf{s}_t,\mathcal{O}_{1:T})}{p(\mathcal{O}_{1:T})} = \frac{\overset{\beta_t(\mathbf{s}_t)}{p(\mathcal{O}_{t:T}|\mathbf{s}_t)}p(\mathbf{s}_t,\mathcal{O}_{1:t-1})}{p(\mathcal{O}_{1:T})} \propto \beta_t(\mathbf{s}_t)\underbrace{p(\mathbf{s}_t|\mathcal{O}_{1:t-1})}_{\alpha_t(\mathbf{s}_t)}\cancel{p(\mathcal{O}_{1:t-1})} \propto \beta_t(\mathbf{s}_t)\alpha_t(\mathbf{s}_t)$$

Forward/backward message intersection



Summary

1. Probabilistic graphical model for optimal control



2. Control = inference (similar to HMM, EKF, etc.)

3. Very similar to dynamic programming, value iteration, etc. (but "soft")

Control as Variational Inference

The optimism problem

for $t = T-1$ to 1:

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

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

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

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

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \overbrace{\log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]}^{\text{“optimistic” transition (not a good idea!)}}$$

why did this happen?

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

marginalizing and conditioning, we get: $p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T})$ (the policy)

“given that you obtained high reward, what was your action probability?”

marginalizing and conditioning, we get: $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T}) \neq p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

“given that you obtained high reward, what was your transition probability?”

Addressing the optimism problem

marginalizing and conditioning, we get: $p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T})$ (the policy) ⟵ we want this

"given that you obtained high reward, what was your action probability?"

marginalizing and conditioning, we get: $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T}) \neq p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$ ⟵ but not this!

"given that you obtained high reward, what was your transition probability?"

"given that you obtained high reward, what was your action probability,
given that your transition probability did not change?"

can we find another distribution $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$ that is close to $p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}|\mathcal{O}_{1:T})$ but has dynamics $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

where have we seen this before?

let $\mathbf{x} = \mathcal{O}_{1:T}$ and $\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$ find $q(\mathbf{z})$ to approximate $p(\mathbf{z}|\mathbf{x})$

let's try variational inference!

Control via variational inference

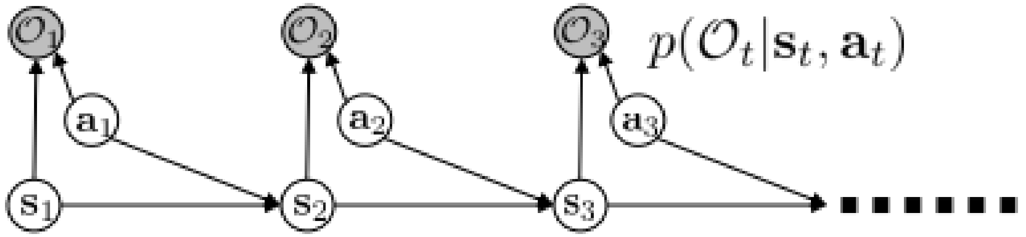
let $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) q(\mathbf{a}_t|\mathbf{s}_t)$

same dynamics and initial state as p

only new thing

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

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



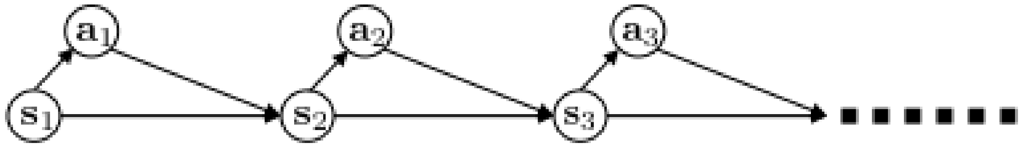
$p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t)$

$p(\mathbf{s}_1)$ $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

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

$q(\mathbf{a}_t|\mathbf{s}_t)$

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



$p(\mathbf{s}_1)$ $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

$q(\mathbf{z})$

The variational lower bound

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

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

the entropy $\mathcal{H}(q)$

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

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

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

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

maximize reward and maximize action entropy!

Optimizing the variational lower bound

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

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

base case: solve for $q(\mathbf{a}_T|\mathbf{s}_T)$:

$$q(\mathbf{a}_T|\mathbf{s}_T) = \arg\max E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} \left[E_{\mathbf{a}_T \sim q(\mathbf{a}_T|\mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T)] + \mathcal{H}(q(\mathbf{a}_T|\mathbf{s}_T)) \right]$$

$$= \arg\max E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} \left[E_{\mathbf{a}_T \sim q(\mathbf{a}_T|\mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T) - \log q(\mathbf{a}_T|\mathbf{s}_T)] \right]$$

optimized when $q(\mathbf{a}_T|\mathbf{s}_T) \propto \exp(r(\mathbf{s}_T, \mathbf{a}_T))$

$$q(\mathbf{a}_T|\mathbf{s}_T) = \frac{\exp(r(\mathbf{s}_T, \mathbf{a}_T))}{\int \exp(r(\mathbf{s}_T, \mathbf{a})) d\mathbf{a}} = \exp(Q(\mathbf{s}_T, \mathbf{a}_T) - V(\mathbf{s}_T))$$

$$V(\mathbf{s}_T) = \log \int \exp(Q(\mathbf{s}_T, \mathbf{a}_T)) d\mathbf{a}_T$$

$$E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} \left[E_{\mathbf{a}_T \sim q(\mathbf{a}_T|\mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T) - \log q(\mathbf{a}_T|\mathbf{s}_T)] \right] = E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} \left[E_{\mathbf{a}_T \sim q(\mathbf{a}_T|\mathbf{s}_T)} [V(\mathbf{s}_T)] \right]$$

Optimizing the variational lower bound

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

$$q(\mathbf{a}_T|\mathbf{s}_T) = \frac{\exp(r(\mathbf{s}_T,\mathbf{a}_T))}{\int \exp(r(\mathbf{s}_T,\mathbf{a}))d\mathbf{a}} = \exp(Q(\mathbf{s}_T,\mathbf{a}_T) - V(\mathbf{s}_T))$$

$$E_{\mathbf{s}_T\sim q(\mathbf{s}_T)}\left[E_{\mathbf{a}_T\sim q(\mathbf{a}_T|\mathbf{s}_T)}[r(\mathbf{s}_T,\mathbf{a}_T) - \log q(\mathbf{a}_T|\mathbf{s}_T)]\right] = E_{\mathbf{s}_T\sim q(\mathbf{s}_T)}\left[E_{\mathbf{a}_T\sim q(\mathbf{a}_T|\mathbf{s}_T)}[V(\mathbf{s}_T)]\right]$$

$$\begin{aligned}
q(\mathbf{a}_t|\mathbf{s}_t) &= \arg\max E_{\mathbf{s}_t\sim q(\mathbf{s}_t)}\left[E_{\mathbf{a}_t\sim q(\mathbf{a}_t|\mathbf{s}_t)}[r(\mathbf{s}_t,\mathbf{a}_t) + E_{\mathbf{s}_{t+1}\sim p(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)}[V(\mathbf{s}_{t+1})]] + \mathcal{H}(q(\mathbf{a}_t|\mathbf{s}_t))\right] \\
&= \arg\max E_{\mathbf{s}_t\sim q(\mathbf{s}_t)}\left[E_{\mathbf{a}_t\sim q(\mathbf{a}_t|\mathbf{s}_t)}[Q(\mathbf{s}_t,\mathbf{a}_t)] + \mathcal{H}(q(\mathbf{a}_t|\mathbf{s}_t))\right] \\
&= \arg\max E_{\mathbf{s}_t\sim q(\mathbf{s}_t)}\left[E_{\mathbf{a}_t\sim q(\mathbf{a}_t|\mathbf{s}_t)}[Q(\mathbf{s}_t,\mathbf{a}_t) - \log q(\mathbf{a}_t|\mathbf{s}_t)]\right]
\end{aligned}$$

regular Bellman backup

not optimistic

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

optimized when $q(\mathbf{a}_t|\mathbf{s}_t) \propto \exp(Q(\mathbf{s}_t,\mathbf{a}_t))$

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

$$q(\mathbf{a}_t|\mathbf{s}_t) = \exp(Q(\mathbf{s}_t,\mathbf{a}_t) - V(\mathbf{s}_t))$$

Backward pass summary - variational

for $t = T-1$ to 1:

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

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

value iteration algorithm:

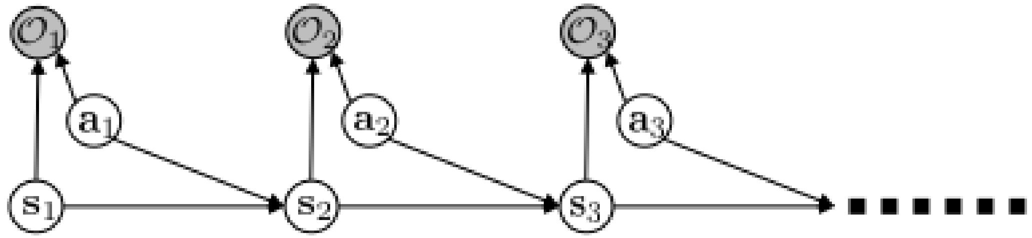
1. set $Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E[V(\mathbf{s}')]$
2. set $V(\mathbf{s}) \leftarrow \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

soft value iteration algorithm:

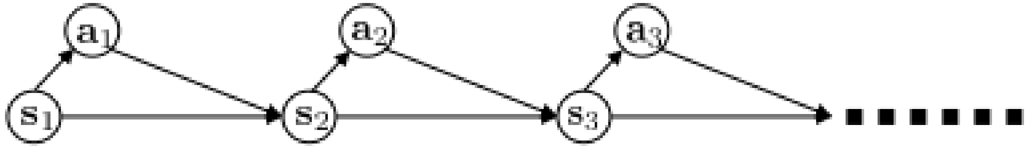
1. set $Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E[V(\mathbf{s}')]$
2. set $V(\mathbf{s}) \leftarrow \text{soft}\max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

Summary

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



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



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

variants:

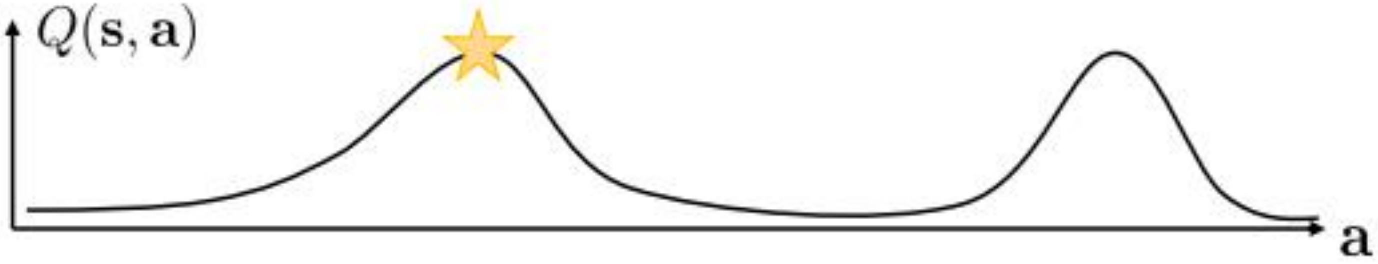
discounted SOC: $Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma E[V_{t+1}(\mathbf{s}_{t+1})]$

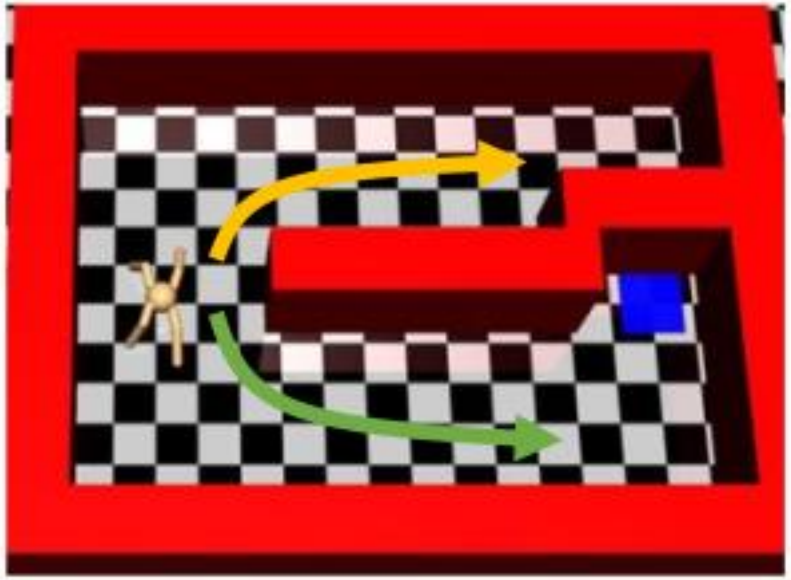
explicit temperature: $V_t(\mathbf{s}_t) = \alpha \log \int \exp\left(\frac{1}{\alpha} Q_t(\mathbf{s}_t, \mathbf{a}_t)\right) d\mathbf{a}_t$

For more details, see: Levine. (2018). Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review.

Stochastic energy-based policies

Q-function: $Q(\mathbf{s}, \mathbf{a}) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$





$$\pi(\mathbf{a}|\mathbf{s}) \propto \exp(Q(\mathbf{s}, \mathbf{a}))$$

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

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

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

Haarnoja*, Tang*, Abbeel, L., Reinforcement Learning with Deep Energy-Based Policies. ICML 2017

- Todorov. (2006). Linearly solvable Markov decision problems: one framework for reasoning about soft optimality.
- Todorov. (2008). General duality between optimal control and estimation: primer on the equivalence between inference and control.
- Kappen. (2009). Optimal control as a graphical model inference problem: frames control as an inference problem in a graphical model.
- Ziebart. (2010). Modeling interaction via the principle of maximal causal entropy: connection between soft optimality and maximum entropy modeling.
- Rawlik, Toussaint, Vijaykumar. (2013). On stochastic optimal control and reinforcement learning by approximate inference: temporal difference style algorithm with soft optimality.
- Haarnoja*, Tang*, Abbeel, L. (2017). Reinforcement learning with deep energy based models: soft Q-learning algorithm, deep RL with continuous actions and soft optimality
- Nachum, Norouzi, Xu, Schuurmans. (2017). Bridging the gap between value and policy based reinforcement learning.
- Schulman, Abbeel, Chen. (2017). Equivalence between policy gradients and soft Q-learning.
- Haarnoja, Zhou, Abbeel, L. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.
- Levine. (2018). Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review