Control as a Probabilistic Inference

Kiyoung Om

Table of Contents

Motivation of understanding control as probabilistic inference framework

- **Maximum Entropy RL**

$$J_{\mathrm{MaxEnt}}(\pi; p, r) \triangleq \mathbb{E}_{\mathbf{a_t}\sim\pi(\mathbf{a_t}|\mathbf{s_t}),\mathbf{s_{t+1}}\sim p(\mathbf{s_{t+1}}|\mathbf{s_t},\mathbf{a_t})}\left[\sum_{t=1}^{T} r\left(\mathbf{s_t}, \mathbf{a_t}\right) + \alpha\mathcal{H}_{\pi}\left[\mathbf{a_t} \mid \mathbf{s_t}\right]\right],$$

  - Representative algorithms: Soft-Q Learning(**SQL**), Soft Actor Critic(**SAC**)

- **KL-Divergence Constraints for Policy Search**

$$\max_{\theta}\ \mathbb{E}_{s\sim\rho_{\theta_{\mathrm{old}}},\, a\sim\pi_{\theta_{\mathrm{old}}}}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\mathrm{old}}}(a|s)} A^{\pi_{\theta_{\mathrm{old}}}}(s,a)\right] \qquad \mathbb{E}_{s\sim\rho_{\theta_{\mathrm{old}}}}\left[D_{\mathrm{KL}}\left(\pi_{\theta_{\mathrm{old}}}(\cdot|s) \,\|\, \pi_\theta(\cdot|s)\right)\right] \le \delta$$

  - Representative algorithms: Trust Region Policy Optimization(**TRPO**),
    Maximum a posteriori policy optimization (**MPO**), Advantage weighted regression (**AWR**),

- **Return conditioned trajectory planning**

$$p\left(\tau \mid \mathcal{O}_{1:T}\right) \propto p(\tau)\exp\left(\sum_{t=1}^{T} r\left(s_t, a_t\right)\right)$$

  - Representative algorithms:
    Planning with diffusion for flexible behavior synthesis (**Diffuser**)

Theory of control as probabilistic inference

Standard RL Policy Search Problem

- Find θ that maximizes cumulative rewards, with parameterization: $p(a_t|s_t,\theta) = \pi_\theta(a_t|s_t)$

$$\theta^* = \arg\max_\theta \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim p(s_t,a_t|\theta)} \left[r(s_t,a_t)\right]$$

- In other words,

$$p(\tau|\theta) = p(s_1, a_t, \ldots, s_T, a_T \mid \theta) = p(s_1) \prod_{t=1}^{T} p(a_t \mid s_t, \theta) p(s_{t+1} \mid s_t, a_t).$$

$$\theta^* = \arg\max_\theta \mathbb{E}_{\tau\sim p(\tau|\theta)} \left[\sum_{t=1}^{T} r(s_t,a_t)\right]$$

To make RL policy search problem into probabilistic inference framework:

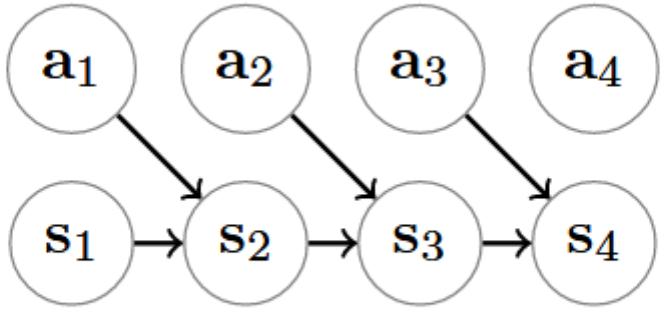
- Draw a **Probabilistic Graphical Model (PGM)**, where:

  Most probable trajectory: Trajectory from the **optimal policy**.

  Inference of posterior action conditional $p(a_t|s_t,\theta)$: Gives the **optimal policy**.

Theory of control as probabilistic inference

How to draw the PGM?



(a) graphical model with states and actions

- Simplest graphical model to express control problem.
- When we see the graphical model: **Decompose the Joint Distribution!**

$$p(s_1, a_1, s_2, a_2, \ldots, a_T, s_{T+1}) = p(s_1) \prod_{t=1}^{T} p(a_t) \cdot p(s_{t+1} \mid s_t, a_t)$$

- Is it enough to satisfy below questions?

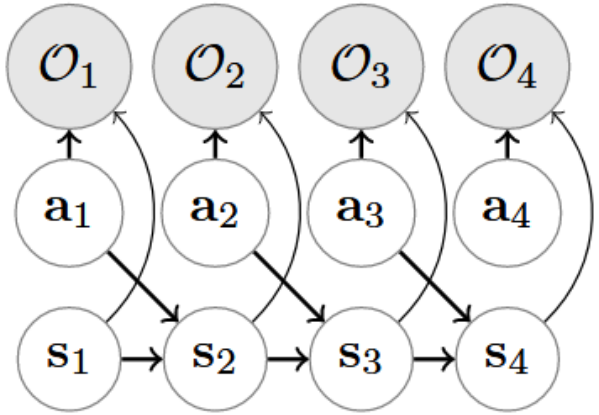
  Most probable trajectory: Trajectory from the **optimal policy**.

  Inference of posterior action conditional $p(a_t|s_t, \theta)$: Gives the **optimal policy.**

Cannot induce reward/cost information here.

Theory of control as probabilistic inference

How to draw the PGM?



(b) graphical model with optimality variables

- Introduced additional random variable $\mathcal{O}_t = \begin{cases} 1 & \text{if timestep } t \text{ is optimal} \\ 0 & \text{otherwise} \end{cases}$
- When we see the graphical model: **Decompose the Joint Distribution!**

$$p(\tau, \mathcal{O}_{1:T}) = p(s_1) \prod_{t=1}^{T} p(a_t) p(\mathcal{O}_t = 1 \mid s_t, a_t)\, p(s_{t+1} \mid s_t, a_t)$$

Decision Choice 1: As we were not interested in action prior (online), we can take it as simplest: $p(a_t) = p(a_t|s_t) = \frac{1}{|\mathcal{A}_t|}$

Decision Choice 2: We define conditional distribution: $p(\mathcal{O}_t = 1|s_t, a_t) = \exp(r(s_t, a_t))$

Theory of control as probabilistic inference

Most probable trajectory as trajectory from the optimal policy

- From the two Decision Choices, we can factorize posterior:

Decision 1 ⟹ $p(\tau \mid \mathcal{O}_{1:T}) \propto p(\tau, \mathcal{O}_{1:T}) = p(s_1) \prod_{t=1}^{T} p(\mathcal{O}_t = 1 \mid s_t, a_t)\, p(s_{t+1} \mid s_t, a_t)$

Decision 2 ⟹ $= p(s_1) \prod_{t=1}^{T} \exp(r(s_t, a_t))\, p(s_{t+1} \mid s_t, a_t)$

$= \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t)\right] \exp\left(\sum_{t=1}^{T} r(s_t, a_t)\right).$

RHS term should be $\propto$;
We can incorporate this defining reward $r(s_t, a_t)$

So we can set the target distribution as:

- Product distribution of **Dynamics** and **exponential of Reward Sum (stochastic dynamics)**

$$p(\tau|o_{1:T}) \propto \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t)\right] \exp\left(\sum_{t=1}^{T} r(s_t, a_t)\right).$$

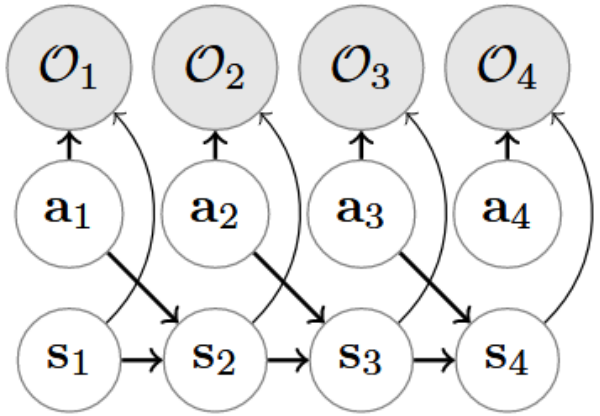
- Product distribution of **Feasible dynamics** and **exponential of Reward Sum (deterministic dynamics)**

$$p(\tau|o_{1:T}) \propto \mathbb{1}[p(\tau) \neq 0] \exp\left(\sum_{t=1}^{T} r(s_t, a_t)\right)$$

Theory of control as probabilistic inference

Inference of posterior action conditional $p(a_t|s_t, \theta)$: Gives the optimal policy.

- Now lets find out how to recover **optimal policy** $\mathbf{p}(\mathbf{a_t}|\mathbf{s_t}, \mathcal{O}_{\mathbf{t:T}} = \mathbf{1})$ from designed PGM.



(b) graphical model with optimality variables

- Definition 1. $\beta_t(s_t, a_t) = p(\mathcal{O}_{t:T}|s_t, a_t)$
- Definition 2. $\beta_t(s_t) = p(\mathcal{O}_{t:T}|s_t)$

Theory of control as probabilistic inference

Inference of posterior action conditional $p(a_t|s_t, \theta)$: Gives the optimal policy.

Definition 1. $\beta_t(s_t, a_t) = p(\mathcal{O}_{t:T}|s_t, a_t)$ Definition 2. $\beta_t(s_t) = p(\mathcal{O}_{t:T}|s_t)$

- **Formulation 1**; Marginalizing $\beta_t(s_t)$ with a_t

$$\beta_t(s_t) = p(\mathcal{O}_{t:T} \mid s_t) = \int_{\mathcal{A}} p(\mathcal{O}_{t:T}, a_t|s_t)da_t \quad (1)$$

$$= \int_{\mathcal{A}} \frac{p(\mathcal{O}_{t:T}, a_t|s_t)}{p(a_t|s_t)} p(a_t|s_t)da_t \quad (2)$$

$$= \int_{\mathcal{A}} p(\mathcal{O}_{t:T} \mid s_t, a_t)p(a_t \mid s_t)\, da_t \quad (3)$$

$$= \int_{\mathcal{A}} \beta_t(s_t, a_t)p(a_t \mid s_t)\, da_t. \quad (4)$$

Decision 1 ⟹ $\beta_t(s_t) = \int_{\mathcal{A}} \beta_t(s_t, a_t)da_t$

$p(a_t) = p(a_t|s_t) = \frac{1}{|\mathcal{A}_t|}$

Theory of control as probabilistic inference

Inference of posterior action conditional $p(a_t|s_t,\theta)$: Gives the optimal policy.

Definition 1. $\beta_t(s_t,a_t)=p(\mathcal{O}_{t:T}|s_t,a_t)$ Definition 2. $\beta_t(s_t)=p(\mathcal{O}_{t:T}|s_t)$

- **Formulation 2**; Marginalizing $\beta_t(s_t,a_t)$ with s_{t+1}

$$\beta_t(s_t,a_t)=p(\mathcal{O}_{t:T}\mid s_t,a_t) \tag{1}$$

$$=\int_{\mathcal{S}}\beta_{t+1}(s_{t+1})\,p(s_{t+1}\mid s_t,a_t)\,p(\mathcal{O}_t\mid s_t,a_t)\,ds_{t+1}. \tag{2}$$

- **Formulation 3**; Optimal target policy from $\beta_t(s_t,a_t),\ \beta_t(s_t)$

Decision 1

$$p(a_t\mid s_t,\mathcal{O}_{t:T})=\frac{p(s_t,a_t|\mathcal{O}_{t:T})}{p(s_t|\mathcal{O}_{t:T})}=\frac{p(\mathcal{O}_{t:T}|s_t,a_t)\,p(a_t|s_t)\,p(s_t)}{p(\mathcal{O}_{t:T}|s_t)\,p(s_t)}$$

$$\propto\frac{p(\mathcal{O}_{t:T}|s_t,a_t)}{p(\mathcal{O}_{t:T}|s_t)}=\frac{\beta_t(s_t,a_t)}{\beta_t(s_t)},$$

Theory of control as probabilistic inference

Inference of posterior action conditional $p(a_t|s_t, \theta)$: Gives the optimal policy.

- Definition 3 (Soft Q): $Q(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$
- Definition 4 (Soft V): $V(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$
- From formulation 1:
  $\beta_t(s_t) = \int_{\mathcal{A}} \beta_t(s_t, a_t) da_t$

If the scale of Q is large,
High Q-value dominates the other actions.
Soft Maximization

$$V(\mathbf{s}_t) = \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t \approx \max_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t).$$

- From formulation 2:

$$\beta_t(s_t, a_t) = \int_{\mathcal{S}} \beta_{t+1}(s_{t+1}) \, p(s_{t+1} \mid s_t, a_t) \, p(\mathcal{O}_t \mid s_t, a_t) \, ds_{t+1}.$$

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

- From formulation 3:

$$p(a_t \mid s_t, \mathcal{O}_{t:T}) = \frac{\beta_t(s_t, a_t)}{\beta_t(s_t)} = \exp(Q(s_t, a_t) - V(s_t))$$

Theory of control as probabilistic inference

Problem of the drawn PGM; Overly optimistic Q function

- **Stochastic dynamics** Leads to **overly optimistic Q function, risk seeking property.**

$$V(\mathbf{s}_t) = \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_t, \mathbf{a}_t))\, d\mathbf{a}_t \approx \max_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t).$$

Another soft maximization term

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

Where original bellman expectation equation:

$$V(s_t) = \max_{a_t} Q(s_t, a_t)$$

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

- EX): Two states for s_{t+1},
  - Win the lottery with $p(s_{t+1} = win | s_t, a_t = buy) = 0.000001$
  - Lose the lottery with $p(s_{t+1} = lose\ | s_t, a_t = buy) = 0.999999$

As $V(s_{t+1} = win)$ dominates, always choose action to buy lottery.

Theory of control as probabilistic inference

Problem of the drawn PGM; Overly optimistic Q function

- Deterministic dynamics much more desirable.

$$V(\mathbf{s}_t) = \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_t, \mathbf{a}_t))\, d\mathbf{a}_t \approx \max_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t).$$

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

Where original bellman expectation equation:

$$V(s_t) = \max_{a_t} Q(s_t, a_t)$$

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

- We can **safely** use deterministic update of bellman equation, extract the policy.

$$p(a_t \mid s_t, \mathcal{O}_{t:T}) = \frac{\beta_t(s_t, a_t)}{\beta_t(s_t)} = \exp(Q(s_t, a_t) - V(s_t))$$

Then, is this framework only available in deterministic settings?

Theory of control as probabilistic inference

So what this inference of optimal conditioned policy exactly optimizes?

- Lets get back to the optimality conditioned trajectory distribution (**Target**).

$$p(\tau|o_{1:T}) = \left[p\left(\mathbf{s}_1\right)\prod_{t=1}^{T} p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right)\right] \exp\left(\sum_{t=1}^{T} r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right) \quad (1)$$

- What we want **to model is policy**; So we formulate differently with above equation.

$$p(\tau \mid \mathcal{O}_{1:T}) = p(s_1 \mid \mathcal{O}_{1:T}) \cdot \prod_{t=1}^{T} p(a_t \mid s_t, \mathcal{O}_{1:T}) \cdot p(s_{t+1} \mid s_t, a_t, \mathcal{O}_{1:T})$$

- We can define approximate trajectory distribution closest to the target distribution.

$$\hat{p}(\tau) = p\left(\mathbf{s}_1 \mid \mathcal{O}_{1:T}\right)\prod_{t=1}^{T} p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T}\right)\pi\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \quad (2)$$

- Then make this optimization problem: (1), (2)

$$D_{\mathrm{KL}}(\hat{p}(\tau)\|p(\tau)) = -E_{\tau\sim\hat{p}(\tau)}[\log p(\tau) - \log \hat{p}(\tau)].$$

Theory of control as probabilistic inference

So what this inference of optimal conditioned policy exactly optimizes?

- In deterministic case,

$$p(\tau|o_{1:T}) = \left[p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right) \quad (1)$$

- Also most probable approximate distribution will be:

$$\hat{p}(\tau) = p(\mathbf{s}_1 \mid \mathcal{O}_{1:T}) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T})\, \pi(\mathbf{a}_t \mid \mathbf{s}_t)$$

$$= \hat{p}(\tau) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\, \pi(\mathbf{a}_t \mid \mathbf{s}_t) \quad (3)$$

- Then optimization problem solves: (1), (3)

$$D_{\mathrm{KL}}(\hat{p}(\tau)\|p(\tau)) = -E_{\tau\sim\hat{p}(\tau)}[\log p(\tau) - \log \hat{p}(\tau)].$$

Theory of control as probabilistic inference

So what this inference of optimal conditioned policy exactly optimizes?

$$
\begin{aligned}
-D_{\mathrm{KL}}(\hat{p}(\tau)\|p(\tau)) =& E_{\tau\sim\hat{p}(\tau)}\Bigg[\log p\left(\mathbf{s}_1\right)+\sum_{t=1}^{T}\left(\log p\left(\mathbf{s}_{t+1}\mid\mathbf{s}_t,\mathbf{a}_t\right)+r\left(\mathbf{s}_t,\mathbf{a}_t\right)\right) \\
& -\log p\left(\mathbf{s}_1\right)-\sum_{t=1}^{T}\left(\log p\left(\mathbf{s}_{t+1}\mid\mathbf{s}_t,\mathbf{a}_t\right)+\log\pi\left(\mathbf{a}_t\mid\mathbf{s}_t\right)\right)\Bigg] \\
=& E_{\tau\sim\hat{p}(\tau)}\left[\sum_{t=1}^{T} r\left(\mathbf{s}_t,\mathbf{a}_t\right)-\log\pi\left(\mathbf{a}_t\mid\mathbf{s}_t\right)\right] \\
=& \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim\hat{p}(\mathbf{s}_t,\mathbf{a}_t))}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right)-\log\pi\left(\mathbf{a}_t\mid\mathbf{s}_t\right)\right] \\
=& \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim\hat{p}(\mathbf{s}_t,\mathbf{a}_t)}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right)\right]+E_{\mathbf{s}_t\sim\hat{p}(\mathbf{s}_t)}\left[\mathcal{H}\left(\pi\left(\mathbf{a}_t\mid\mathbf{s}_t\right)\right)\right]
\end{aligned}
$$

In deterministic setting; Approximation of inference, equivalent to max entropy RL

Theory of control as probabilistic inference

So what this inference of optimal conditioned policy exactly optimizes?

- In stochastic case we cannot derive the same as deterministic.

$$p(\tau|o_{1:T}) = \left[p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right) \quad (1)$$

$$\hat{p}(\tau) = p(\mathbf{s}_1 \mid \mathcal{O}_{1:T}) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T})\, \pi(\mathbf{a}_t \mid \mathbf{s}_t) \quad (2)$$

- We **cannot erase optimality conditioned dynamics** terms.
- Directly optimize : $D_{\mathrm{KL}}(\hat{p}(\tau)\|p(\tau)) = -E_{\tau\sim\hat{p}(\tau)}[\log p(\tau) - \log \hat{p}(\tau)]$.
- Leads: $E_{\tau\sim\hat{p}(\tau)}\left[\log p(\mathbf{s}_1) + \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\right] + \mathcal{H}(\hat{p}(\tau))$

So, to modeling the approximation as (2), we cannot use model-free algorithms.

Theory of control as probabilistic inference

Direct Solution! Abuse PGM.

- Why not we just model our approxmate distribution with (3) rather than (2) ?

$$p(\tau|o_{1:T}) = \left[p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right) \quad (1)$$

$$\hat{p}(\tau) = p(\mathbf{s}_1 \mid \mathcal{O}_{1:T}) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T})\, \pi(\mathbf{a}_t \mid \mathbf{s}_t) \quad (2)$$

$$= \hat{p}(\tau) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\, \pi(\mathbf{a}_t \mid \mathbf{s}_t) \quad (3)$$

- Pros: We can **recover the maximum entropy RL objective**, like deterministic case.
  Agents cannot control dynamical systems like (2), not intuitive.
  (Is there anyone who can make environment to always win the lottery?)

- Cons: We abuse the PGM w.r.t. optimality -> No theoretical grounds (yet)

Theory of control as probabilistic inference

Connection to Structural Variational Inference

- **The theoretical grounds of abused solution!**
- Target distribution: $p(\tau|o_{1:T}) = \left[p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right)$ (1)
- Variational approximate distribution: $\hat{q}(\tau) = q(\mathbf{s}_1)\prod_{t=1}^{T} q(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\, q(\mathbf{a}_t \mid \mathbf{s}_t)\,.$ (2) *q, not p*
- With given observation, **ELBO with variational trajectory distribution:**

$$
\begin{aligned}
\log p(\mathcal{O}_{1:T}) &= \log \iint p(\mathcal{O}_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T})\, d\mathbf{s}_{1:T} d\mathbf{a}_{1:T} \\
&= \log \iint p(\mathcal{O}_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \frac{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} d\mathbf{s}_{1:T} d\mathbf{a}_{1:T} \\
&= \log E_{(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \sim q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\frac{p(\mathcal{O}_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T})}{q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\right] \\
&\geq E_{(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \sim q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\log p(\mathcal{O}_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}) - \log q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\right]
\end{aligned}
$$

Theory of control as probabilistic inference

Connection to Structural Variational Inference

- **The theoretical grounds of abused solution!**
- Target distribution: $p(\tau|o_{1:T}) = \left[p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right)$ (1)
- Variational approximate distribution: $\hat{q}(\tau) = q(\mathbf{s}_1)\prod_{t=1}^{T} q(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)\, q(\mathbf{a}_t \mid \mathbf{s}_t)$. (2) $q, not\ p$

$$\log p(\mathcal{O}_{1:T}) \geq E_{(\mathbf{s}_{1:T},\mathbf{a}_{1:T})\sim q(\mathbf{s}_{1:T},\mathbf{a}_{1:T})}\left[\log p(\mathcal{O}_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}) - \log q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})\right]$$

$put\ (1), (2)$

$$\log p(\mathcal{O}_{1:T}) \geq E_{(\mathbf{s}_{1:T},\mathbf{a}_{1:T})\sim q(\mathbf{s}_{1:T},\mathbf{a}_{1:T})}\left[\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) - \log q(\mathbf{a}_t \mid \mathbf{s}_t)\right]$$

- Then we set dynamics distributions q with $p(s_1), p(s_{t+1}|s_t, a_t)$ → Recovers Max-Entropy RL

$$= \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim\hat{p}(\mathbf{s}_t,\mathbf{a}_t)}\left[r(\mathbf{s}_t, \mathbf{a}_t) + \mathcal{H}(\pi(\mathbf{a}_t \mid \mathbf{s}_t))\right]$$

- Abused distribution is one of the ELBO, derivation of our designed PGM (**Got Theoretical Grounds**.)

Theory of control as probabilistic inference

How can we extract policy from the abused settings (ELBO)

- We already discussed before **soft-Q, V and optimal policy** for **stochastic/deterministic** settings.
- What about **abused settings**, any difference with $\exp\big(Q(s_t, a_t) - V(s_t)\big)$?

From the max-entropy objective function:

$$\sum_{t=1}^{T} E_{(\mathbf{s}_t, \mathbf{a}_t) \sim \hat{p}(\mathbf{s}_t, \mathbf{a}_t)} \left[r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \mathcal{H}\left(\pi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right) \right]$$

If $t = T$:

$$\begin{aligned}
&\mathbb{E}_{(s_T, a_T) \sim \hat{p}(s_T, a_T)} \left[r(s_T, a_T) - \log \pi(a_T \mid s_T) \right] \\
&= \mathbb{E}_{s_T \sim \hat{p}(s_T)} \left[\mathbb{E}_{a_T \sim \pi(\cdot \mid s_T)} \left[r(s_T, a_T) - \log \pi(a_T \mid s_T) \right] \right] \\
&= \mathbb{E}_{s_T \sim \hat{p}(s_T)} \left[\mathbb{E}_{a_T \sim \pi(\cdot \mid s_T)} \left[-\log \pi(a_T \mid s_T) + r(s_T, a_T) \right] \right] \\
&= \mathbb{E}_{s_T \sim \hat{p}(s_T)} \left[-D_{\mathrm{KL}} \left(\pi(a_T \mid s_T) \,\|\, \frac{1}{Z(s_T)} \exp(r(s_T, a_T)) \right) + \log Z(s_T) \right] \\
&= \mathbb{E}_{s_T \sim \hat{p}(s_T)} \left[-D_{\mathrm{KL}} \left(\pi(a_T \mid s_T) \,\|\, \frac{1}{\exp(V(s_T))} \exp(r(s_T, a_T)) \right) + V(s_T) \right]
\end{aligned}$$

Define:

$V(s_T) = \log \int_{\mathcal{A}} \exp(r(s_T, a_t)) da_T$

$\pi^*(a_T | s_t) = \exp(r(s_T, a_T))$

Theory of control as probabilistic inference

How can we extract policy from the abused settings (ELBO)

If $t = t$ $(intermediate\ steps)$:

- Policy should also consider the **future expected value**. (Summation of $t = 1 \sim T$ recovers original objective.)

$$\mathbb{E}_{(s_t,a_t)\sim\hat{p}(s_t,a_t)}\Big[r(s_t,a_t) - \log\pi(a_t \mid s_t) + \underline{\mathbb{E}_{s_{t+1}\sim p(\cdot|s_t,a_t)}\big[V(s_{t+1})\big]}\Big]$$

$$\overset{(1)}{=} \mathbb{E}_{(s_t,a_t)\sim\hat{p}}\Big[\underbrace{r(s_t,a_t) + \mathbb{E}_{s_{t+1}}[V(s_{t+1})]}_{:=\ Q(s_t,a_t)} - \log\pi(a_t \mid s_t)\Big]$$

Define, same as Bellman Expectation Eq.
We can mitigate risk-seeking Q Problem!

$$= \mathbb{E}_{s_t\sim\hat{p}(s_t)}\Big[\mathbb{E}_{a_t\sim\pi(\cdot|s_t)}\big[Q(s_t,a_t) - \log\pi(a_t \mid s_t)\big]\Big]$$

$$\overset{(2)}{=} \mathbb{E}_{s_t\sim\hat{p}(s_t)}\Big[-D_{\mathrm{KL}}\Big(\pi(\cdot \mid s_t)\Big\|\tfrac{1}{Z(s_t)}\exp\big(Q(s_t,\cdot)\big)\Big) + \log Z(s_t)\Big]$$

Now defined on Q

$$V(s_t) = \log\int_{\mathcal{A}}\exp(Q(s_t,a_t))dat$$

$$= \boxed{\mathbb{E}_{s_t\sim\hat{p}(s_t)}\Big[-D_{\mathrm{KL}}\Big(\pi(a_t \mid s_t)\Big\|\frac{1}{\exp\big(V(s_t)\big)}\exp\big(Q(s_t,a_t)\big)\Big) + V(s_t)\Big]}$$

$$\pi^*(\mathbf{a}_t \mid \mathbf{s}_t) = \exp\left(Q(\mathbf{s}_t,\mathbf{a}_t) - V(\mathbf{s}_t)\right),$$

In Summary:

1. We want to transform the control problem -> Probabilistic Inference.

2. Draw PGM which can infer $p(a_t|s_t, O_{1:T})$, $p(\tau|O_{1:T})$ posteriors.

3. Decompose and **got optimal target distribution**: $p(\tau|O_{1:T})$

4. To find out optimal policy, defined **soft $\boldsymbol{Q}, \boldsymbol{V}$** functions and found **analytical form of $\boldsymbol{\pi^*(a_t|s_t)}$**
   Stochastic dynamics has risk-seeking problem with Q function.

5. With defining optimal approximate distribution (closest distribution),
   Transformed inference problem to optimization problem.
   Again, problem with stochastic dynamics setting has problem. (Only model-based works)

6. We abused the PGM, and modeled approximate distribution differently.
   This reveals that one of the **ELBO** of PGM := Abused distribution.

7. Abused distribution recovers Max Entropy RL, no problem of risk-seeking Q function.
   Most of the Max-Entropy RL algorithms built on this setting.

Applications of Control as Probabilistic Inference

Maximum Entropy Policy Gradients

- Policy Gradients with the Max-Entropy Objective

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_{t=1}^{T} \nabla_\theta E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[r\left(\mathbf{s}_t,\mathbf{a}_t\right) + \mathcal{H}\left(q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right)\right] \\ &= \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[\nabla_\theta \log q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \left(\sum_{t'=t}^{T} r\left(\mathbf{s}_{t'},\mathbf{a}_{t'}\right) - \log q_\theta\left(\mathbf{a}_{t'} \mid \mathbf{s}_{t'}\right) - 1 \right)\right] \\ &= \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[\nabla_\theta \log q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \left(\sum_{t'=t}^{T} r\left(\mathbf{s}_{t'},\mathbf{a}_{t'}\right) - \log q_\theta\left(\mathbf{a}_{t'} \mid \mathbf{s}_{t'}\right) - b\left(\mathbf{s}_{t'}\right) \right)\right],\end{aligned}$$

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T} E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)} \left[\nabla_\theta \log q_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \hat{A}\left(\mathbf{s}_t,\mathbf{a}_t\right)\right]$$

- We can just put $-\log q_\theta(a_t|s_t)$ **penalty** to the original reward, and use the policy gradient algorithms.

Applications of Control as Probabilistic Inference

Soft Q Learning

- With below two equations, definition of soft V, and its optimal policy $q(a_t|s_t)$.

$$V(\mathbf{s}_t) = \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_t, \mathbf{a}_t))\, d\mathbf{a}_t$$

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

- We can parameterize Q_ϕ only to express both equations.

- Training Q_ϕ:

$$\mathcal{E}(\phi) = E_{(\mathbf{s}_t,\mathbf{a}_t)\sim q(\mathbf{s}_t,\mathbf{a}_t)}\left[\left(r(\mathbf{s}_t,\mathbf{a}_t) + E_{q(\mathbf{s}_{t+1}|\mathbf{s}_t,\mathbf{a}_t)}\left[V_\psi(\mathbf{s}_{t+1})\right] - Q_\phi(\mathbf{s}_t,\mathbf{a}_t)\right)^2\right]$$

$$\phi \leftarrow \phi - \alpha E\left[\frac{dQ_\phi}{d\phi}(\mathbf{s}_t,\mathbf{a}_t)\left(Q_\phi(\mathbf{s}_t,\mathbf{a}_t) - \left(r(\mathbf{s}_t,\mathbf{a}_t) + \log\int_{\mathcal{A}} \exp(Q(\mathbf{s}_{t+1},\mathbf{a}_{t+1}))\, d\mathbf{a}_{t+1}\right)\right)\right].$$

$$\log \mathbb{E}_{q(a_t)}\frac{\exp(Q(s_{t+1},a_{t+1}))}{q(a_t)}$$

Importance Sampling with current policy

- Training π_θ:

Stein Variational Gradient Descent (**SVGD**) to match: $\pi_\theta(a_t|s_t) \propto \exp(Q_{soft}^{\phi}(s_t,\cdot))$

Applications of Control as Probabilistic Inference

Soft Actor Critic

- Explicit modeling of the policy network $\pi(a_t|s_t;\theta)$

$$\mathcal{T}^{\pi} Q\left(\mathbf{s}_t, \mathbf{a}_t\right) \triangleq r\left(\mathbf{s}_t, \mathbf{a}_t\right)+\gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[V\left(\mathbf{s}_{t+1}\right)\right]$$

$$V\left(\mathbf{s}_t\right)=\mathbb{E}_{\mathbf{a}_t \sim \pi}\left[Q\left(\mathbf{s}_t, \mathbf{a}_t\right)-\log \pi\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right]$$

⬅ Check lemma 1 from Haarnoja (2018)

- Train both soft V, Q with MSE loss:

$$\mathcal{E}(\phi)=E_{\left(\mathbf{s}_t, \mathbf{a}_t\right) \sim q\left(\mathbf{s}_t, \mathbf{a}_t\right)}\left[\left(r\left(\mathbf{s}_t, \mathbf{a}_t\right)+E_{q\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right)}\left[V_\psi\left(\mathbf{s}_{t+1}\right)\right]-Q_\phi\left(\mathbf{s}_t, \mathbf{a}_t\right)\right)^2\right]$$

$$\mathcal{E}(\psi)=E_{\mathbf{s}_t \sim q\left(\mathbf{s}_t\right)}\left[\left(E_{\mathbf{a}_t \sim q\left(\mathbf{a}_t \mid \mathbf{s}_t\right)}\left[Q_\phi\left(\mathbf{s}_t, \mathbf{a}_t\right)-\log q\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right]-V_\psi\left(\mathbf{s}_t, \mathbf{a}_t\right)\right)^2\right].$$

- Training π_θ:

$$J_\pi(\theta)=\mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta\left(\cdot \mid \mathbf{s}_t\right) \| \frac{\exp\left(Q_\phi\left(\mathbf{s}_t, \cdot\right)\right)}{Z_\theta\left(\mathbf{s}_t\right)}\right)\right].$$

$$J_\pi(\theta)=\mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}}\left[\log \pi_\theta\left(f_\theta\left(\epsilon_t ; \mathbf{s}_t\right) \mid \mathbf{s}_t\right)-Q_\phi\left(\mathbf{s}_t, f_\theta\left(\epsilon_t ; \mathbf{s}_t\right)\right)\right],$$

$$\begin{aligned}\hat{\nabla}_\theta J_\pi(\theta)= & \nabla_\theta \log \pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) \\ & +\left(\nabla_{\mathbf{a}_t} \log \pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)-\nabla_{\mathbf{a}_t} Q\left(\mathbf{s}_t, \mathbf{a}_t\right)\right) \nabla_\theta f_\theta\left(\epsilon_t ; \mathbf{s}_t\right)\end{aligned}$$

Applications of Control as Probabilistic Inference

Replacing Prior $p(a_t|s_t)$

- Decision Choice 1: As we were not interested in action prior (online), we can take it as simplest:

$$p(a_t) = p(a_t|s_t) = \frac{1}{|\mathcal{A}_t|}$$

- What if we assume the prior as $p(a_t|s_t) = \pi_{old}(a_t|s_t)$ or $\pi_{behavior}(a_t|s_t)$?

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} r\left(s_t, a_t\right) + \alpha\mathcal{H}\left(\pi\left(\cdot \mid s_t\right)\right)\right]$$

- Max Entropy objective leads to KL regularized objective. TRPO/PPO/AWR....

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} r\left(s_t, a_t\right)\right] - \alpha\mathbb{E}_{s\sim\rho_\pi}\left[D_{KL}\left(\pi(\cdot \mid s)\|\pi_{\text{old}}\left(\cdot \mid s\right)\right)\right]$$

- Trajectory planning. **Diffuser**

  Modeling Entire trajectory distribution with offline dataset.

$$p(\tau|\mathcal{O}_{1:T}) = p(\tau)\exp\left(\sum_{t=1}^{T} r(s_t, a_t)\right)$$

Future Plan of Research

1. Constrained black-box optimization with posterior inference. (Taeyoung Yun, Kyuil Sim) ~NIPS(2025)

2. Probabilistic Inference for sequential decision making. (Diffusion meets control/Control meets diffusion)

EOD