An Inference Perspective on Model-Based Reinforcement Learning

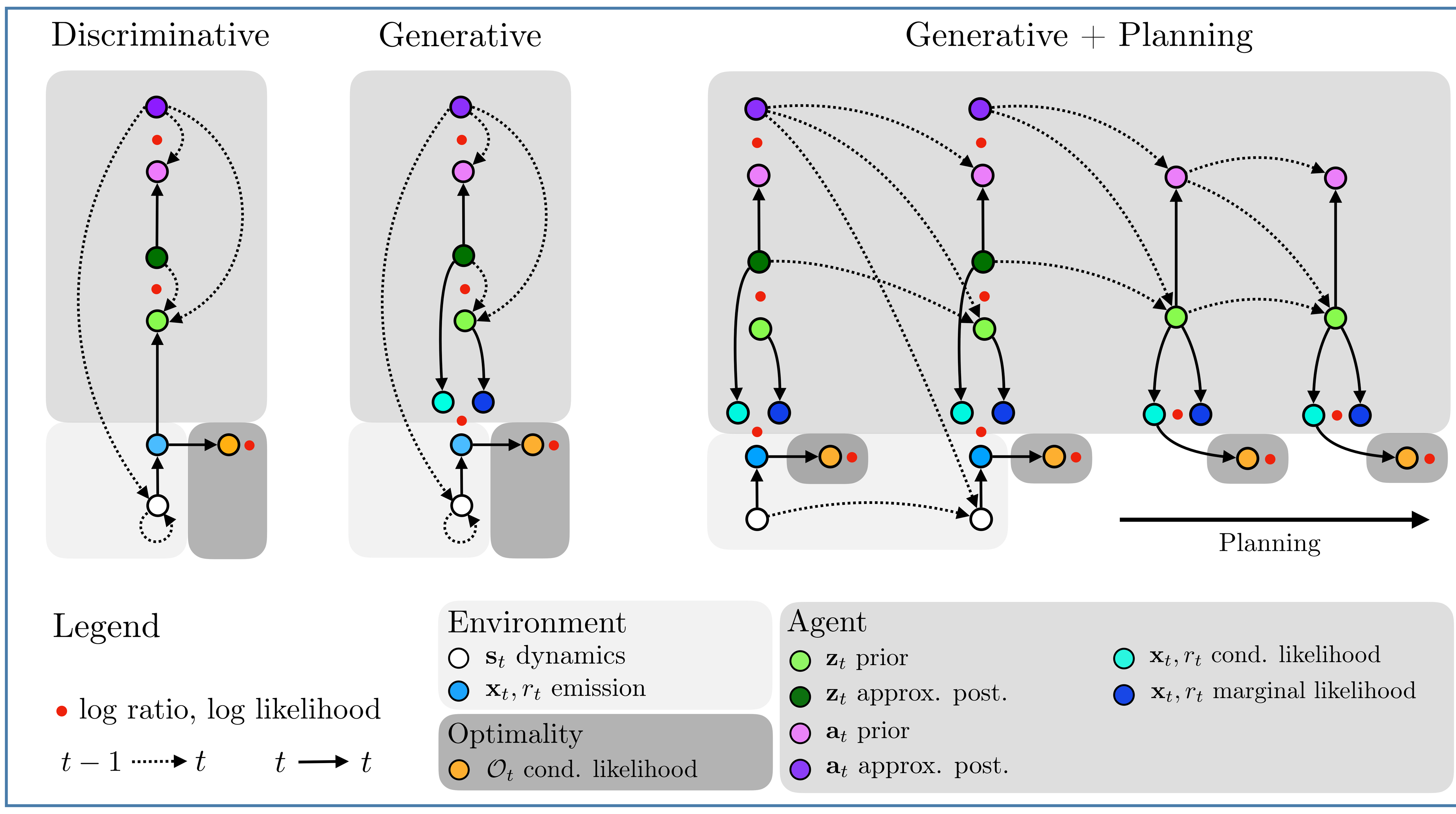
Joseph Marino and Yisong Yue
California Institute of Technology

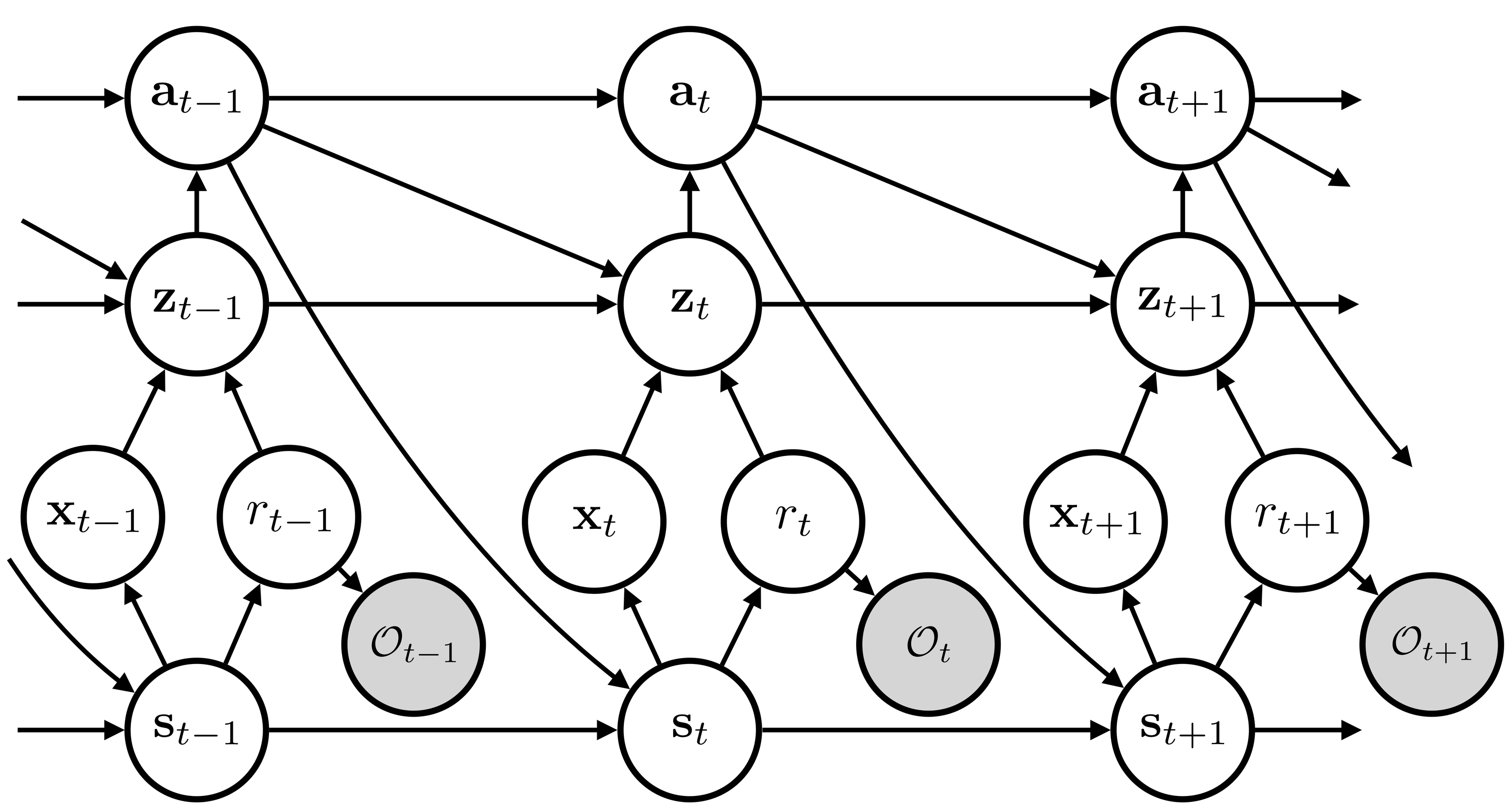
Caltech

abstract

We derive the model-based reinforcement learning objective from the perspective of probabilistic inference. Comparing with current approaches, this objective contains additional terms:

- **action prior**:
  - initialize planning,
  - roll-out policy,
  - consolidate model-based planning into a model-free policy
- **marginal log-likelihood of observations and reward**:
  - restrict the model for task-relevance,
  - bias planning toward confident states





set-up

We frame reinforcement learning as probabilistic inference and learning (Levine, 2018). This is achieved by "observing" maximum reward, then inferring actions that increase the likelihood of this outcome.

ENVIRONMENT

$$p_e(\mathbf{x}_{1:T}, r_{1:T}, \mathbf{s}_{1:T}|\mathbf{a}_{1:T-1}) = \prod_{t=1}^{T} p_e(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_{t-1}) p_e(\mathbf{x}_t|\mathbf{s}_t) p_e(r_t|\mathbf{s}_t)$$

AGENT

$$p_a(\mathbf{a}_{1:T}, \mathbf{z}_{1:T}|\mathbf{x}_{1:T}, r_{1:T}) = \prod_{t=1}^{T} p_a(\mathbf{a}_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t}) p_a(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t})$$

OPTIMALITY

$$p(\mathcal{O}_{1:T}|r_{1:T}) = \prod_{t=1}^{T} p(\mathcal{O}_t|r_t)$$

where $p(\mathcal{O}_t|r_t) = \text{Bernoulli}(\exp(r_t))$, so $\log p(\mathcal{O}_t = 1|r_t) = \log(\exp(r_t)) = r_t$.

maximum likelihood training objective

$$\boxed{\theta^* = \arg\max_\theta \log p(\mathcal{O}_{1:T} = \mathbf{1}).}$$

where $p(\mathcal{O}_{1:T}) = \int p(\mathbf{x}_{1:T}, r_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}, \mathbf{z}_{1:T}, \mathcal{O}_{1:T}) d\mathbf{x}_{1:T} dr_{1:T} d\mathbf{s}_{1:T} d\mathbf{a}_{1:T} d\mathbf{z}_{1:T}$

variational inference & learning

Lower bound $\log p(\mathcal{O}_{1:T} = \mathbf{1})$ using variational inference:

APPROXIMATE POSTERIOR

$$q(\mathbf{z}_{1:T}, \mathbf{a}_{1:T}|\mathbf{x}_{1:T}, r_{1:T}, \mathcal{O}_{1:T}) = \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T}) \cdot q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})$$

EVIDENCE LOWER BOUND (*Discriminative Agent*)

$$\mathcal{L} = \mathbb{E}_{\substack{\mathbf{s},\mathbf{x},r\sim p_e \\ \mathbf{z},\mathbf{a}\sim q}}\left[\sum_{t=1}^{T} r_t - \log \frac{q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_a(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t})} - \log \frac{q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_a(\mathbf{a}_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})}\right]$$

To get a model, invert the agent's internal state prior using *Bayes' Rule*:

$$p_a(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}) = \frac{p_a(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{<t}, r_{<t}) p_a(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}{p_a(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}$$

EVIDENCE LOWER BOUND (*Generative Agent*)

$$\mathcal{L} = \mathbb{E}_{\substack{\mathbf{s},\mathbf{x},r\sim p_e \\ \mathbf{z},\mathbf{a}\sim q}}\left[\sum_{t=1}^{T} r_t + \overbrace{\log \frac{p_a(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{<t}, r_{<t})}{p_a(\mathbf{x}_t, r_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}}^{\textit{information gain}} \overbrace{- \log \frac{q(\mathbf{z}_t|\mathbf{z}_{<t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_a(\mathbf{z}_t|\mathbf{a}_{<t}, \mathbf{z}_{<t}, \mathbf{x}_{<t}, r_{<t})}}^{\textit{internal state consistency}} \underbrace{- \log \frac{q(\mathbf{a}_t|\mathbf{z}_{\leq t}, \mathbf{a}_{<t}, \mathbf{x}_{\leq t}, r_{\leq t}, \mathcal{O}_{t+1:T})}{p_a(\mathbf{a}_t|\mathbf{a}_{<t}, \mathbf{z}_{\leq t}, \mathbf{x}_{\leq t}, r_{\leq t})}}_{\textit{action consistency}}\right]$$

PLANNING

$$\hat{\mathcal{L}}_{t+1:T} = \mathbb{E}_{\mathbf{x},r,\mathbf{z},\mathbf{a}\sim p_a}\left[\sum_{\tau=t+1}^{T} r_\tau + \overbrace{\log \frac{p_a(\mathbf{x}_\tau, r_\tau|\mathbf{a}_{<\tau}, \mathbf{z}_{\leq\tau}, \mathbf{x}_{<\tau}, r_{<\tau})}{p_a(\mathbf{x}_\tau, r_\tau|\mathbf{a}_{<\tau}, \mathbf{z}_{<\tau}, \mathbf{x}_{<\tau}, r_{<\tau})}}^{\textit{mutual information}}\right]$$

discussion

Many recent works have combined latent variable models with RL (Buesing *et al.*, 2018; Igl *et al.*, 2018; Ha & Schmidhuber, 2018; Hafner *et al.*, 2019; Zhang *et al.*, 2019). In comparison, the objective here contains:

- **action prior**: this facilitates consolidation of planning into a model-free policy (Weber *et al.*, 2017; Nagabandi *et al.*, 2018; Kurutach *et al.*, 2018; Buesing *et al.*, 2018), acts as a roll-out policy (Silver *et al.*, 2016), and initializes planning.
- **marginal log-likelihood of observations and reward**: this restricts the internal state to task-relevant information during training. During planning, this appears in the mutual information between the internal state and inputs, biasing planning toward higher confidence states.