Google DeepMind

What type of inference is planning?

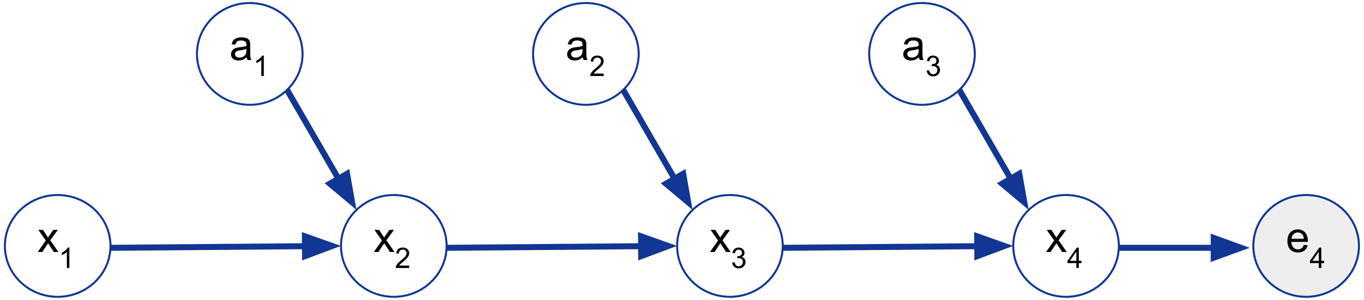
Miguel Lázaro-Gredilla, Li Yang Ku, Kevin P. Murphy, Dileep George

{lazarogredilla, liyangku, kpmurphy, dileepgeorge}@google.com

Problem setup

- We want to plan from known, factorized dynamics and rewards, expressed as a factor graph.
- Rewards are hard to reach with random shooting.
- Demonstrations are not available.
- Dynamics are stochastic.

Planning problem with one reward as a factor graph



$$f(\mathbf{x}, \mathbf{a}) = R(x_T) p(x_1) \prod_{t=1}^{T-1} p(x_{t+1} | x_t, a_t)$$

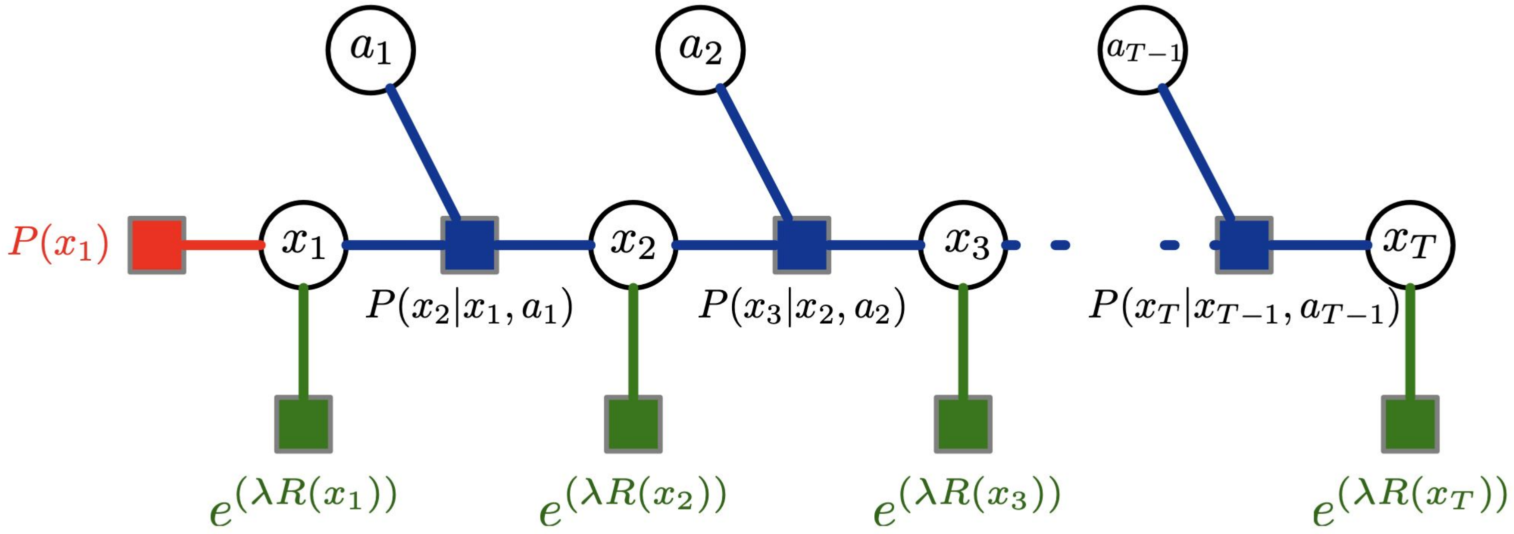
$$R(x_4) = p(e_4 = 1 | x_4)$$

Planning problem

$$\max_{\pi} \sum_{\mathbf{x},\mathbf{a}} f(\mathbf{x}, \mathbf{a}) \pi(\mathbf{a} | \mathbf{x})$$

Standard RL, can be solved exactly using T steps of value iteration.

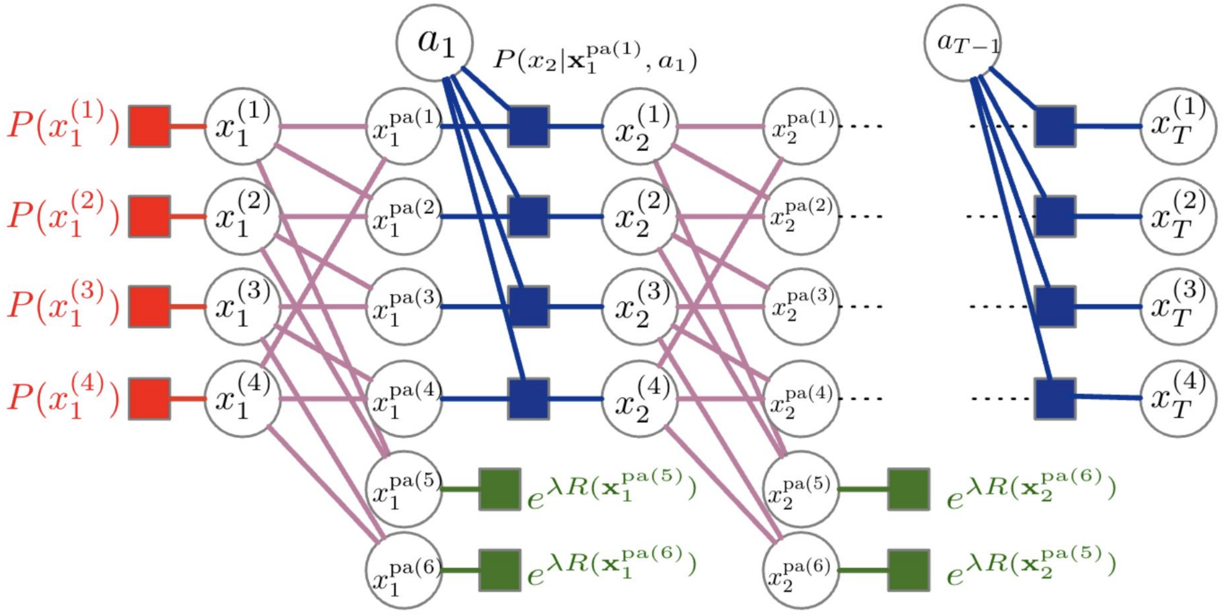
Introducing multiple rewards and the λ trick



$$F_\lambda^{\text{planning}} = \frac{1}{\lambda}\log \max_\pi \sum_{\mathbf{x},\mathbf{a}} f(\mathbf{x},\mathbf{a})\pi(\mathbf{a}|\mathbf{x}) = \max_\pi \frac{1}{\lambda}\log \mathbb{E}_\pi\left[\exp\left(\lambda\sum_{t=1}^{T-1} R(\mathbf{x}_t)\right)\right]$$

When λ→0, we get additive rewards $\lim_{\lambda\to 0} F_\lambda^{\text{planning}} = \max_\pi \mathbb{E}_\pi\left[\sum_{t=1}^{T-1} R(\mathbf{x}_t)\right]$

Planning: just approx. inference in a factor graph!



- Leverage research on approximate inference in factor graphs.
- Which type of inference, though? Marginal? MAP? Marginal MAP?

Planning as inference in the literature (roughly)

Learning (planning as maximum likelihood)

- Probabilistic inference for solving discrete and continuous state Markov Decision Processes (2006)

MAP inference

- Planning by probabilistic inference (2003)
- Reinforcement learning and control as probabilistic inference: Tutorial and review (2018)

Marginal inference

- Factored MCTS for large scale stochastic planning (2015)
- Uniqueness and Complexity of Inverse MDP Models (2023)
- Reinforcement learning and control as probabilistic inference: Tutorial and review (2018)

Marginal MAP inference

- Online Symbolic Gradient-Based Optimization for Factored Action MDPs (2016)
- Stochastic planning with lifted symbolic trajectory optimization (2019)
- Approximate Inference for Stochastic Planning in Factored Spaces (2022)

(and long etc)

Which type of inference is adequate for planning?

Should maximize **expected reward**,
but standard inference finds…

$$\max_{\pi} \sum_{\mathbf{x},\mathbf{a}} f(\mathbf{x},\mathbf{a}) \pi(\mathbf{a}|\mathbf{x})$$

…the partition function (**marginal**),

$$\sum_{\mathbf{x},\mathbf{a}} f(\mathbf{x},\mathbf{a})$$

…the maximum (**MAP**),

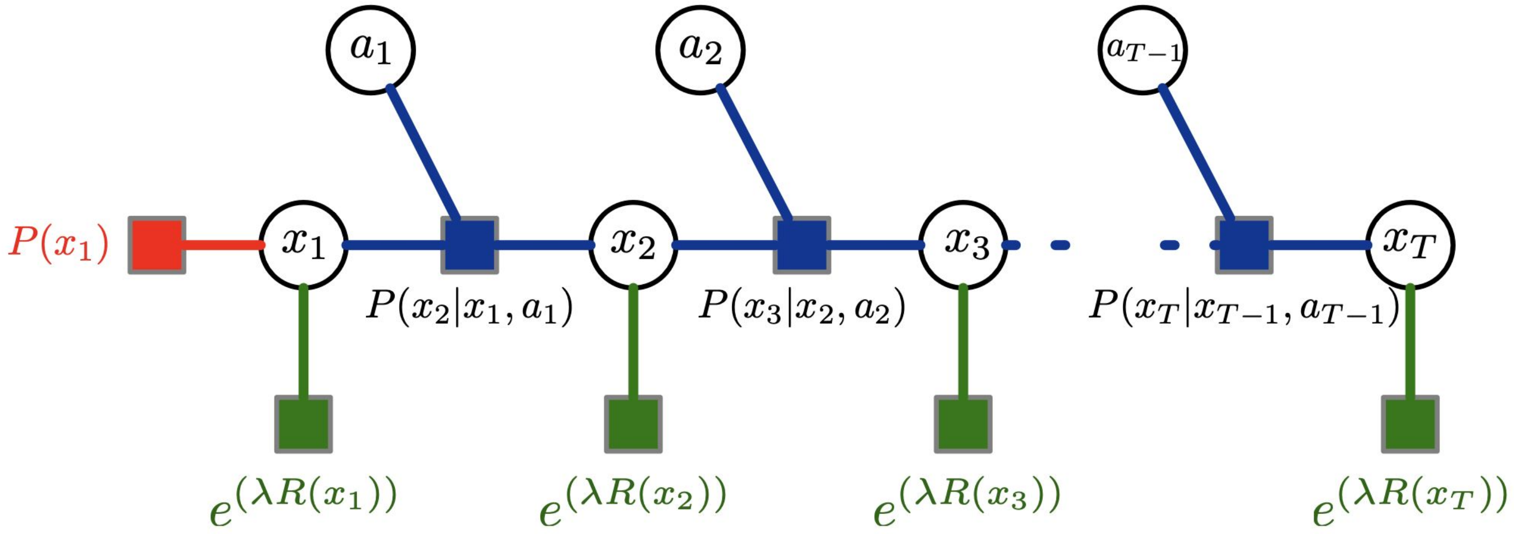
$$\max_{\mathbf{x},\mathbf{a}} f(\mathbf{x},\mathbf{a})$$

…the maximum partition function action sequence (**marginal MAP**).

$$\max_{\mathbf{a}} \sum_{\mathbf{x}} f(\mathbf{x},\mathbf{a})$$

Only first is exact in the stochastic case. But it isn't a standard inference type.

Variational inference



$$\max_{q(\boldsymbol{x},\boldsymbol{a})} \langle \log f(\boldsymbol{x}, \boldsymbol{a}) \rangle_{q(\boldsymbol{x},\boldsymbol{a})} + H_q^{\text{type}}(\boldsymbol{x}, \boldsymbol{a})$$

Claim: All relevant inference types correspond to different *weightings* of entropy terms.

The different types of inference

Type of inference	↓ Closed form for quant. of interest $F_\lambda = \max_{\boldsymbol{q}} F_\lambda(\boldsymbol{q})$	↓ Entropy term $H^{\text{type}}(\boldsymbol{q})$ for variational bound $F_\lambda(\boldsymbol{q}) = \frac{1}{\lambda}(-E_\lambda(\boldsymbol{q}) + H^{\text{type}}(\boldsymbol{q}))$	Tr
Marginal[4]	$\frac{1}{\lambda} \log \sum_{\boldsymbol{x},\boldsymbol{a}} P(\boldsymbol{x}\vert\boldsymbol{a}) e^{\lambda R(\boldsymbol{x},\boldsymbol{a})}$	$H_q(x_1) + \sum_{t=1}^{T-1} H_q(x_{t+1}, a_t \vert x_t)$	✓
Planning	$\frac{1}{\lambda} \max_{\boldsymbol{\pi}} \log \langle e^{\lambda R(\boldsymbol{x},\boldsymbol{a})} \rangle_{P(\boldsymbol{x}\vert\boldsymbol{a})\pi(\boldsymbol{a}\vert\boldsymbol{x})}$	$H_q(x_1) + \sum_{t=1}^{T-1} H_q(x_{t+1} \vert a_t, x_t)$	✓
M. MAP	$\frac{1}{\lambda} \max_{\boldsymbol{a}} \log \sum_{\boldsymbol{x}} P(\boldsymbol{x}\vert\boldsymbol{a}) e^{\lambda R(\boldsymbol{x},\boldsymbol{a})}$	$H_q(x_1) + \sum_{t=1}^{T-1} H_q(x_{t+1}, a_t \vert x_t) - H_q(a_t)$	✗
MAP	$\frac{1}{\lambda} \max_{\boldsymbol{x},\boldsymbol{a}} \log P(\boldsymbol{x}\vert\boldsymbol{a}) e^{\lambda R(\boldsymbol{x},\boldsymbol{a})}$	0	✓
Marginal[U]	$\frac{1}{\lambda} \log \sum_{\boldsymbol{x},\boldsymbol{a}} P(\boldsymbol{x}\vert\boldsymbol{a}) \frac{1}{N_a^{T-1}} e^{\lambda R(\boldsymbol{x},\boldsymbol{a})}$	$H_q(x_1) + \sum_{t=1}^{T-1} (H_q(x_{t+1}, a_t \vert x_t) - \log N_a)$	✓

The energy term is the same for all of these objective functions:

$$E_\lambda(\boldsymbol{q}) = -\langle \log P(x_1) \rangle_{q(x_1)} - \sum_{t=1}^{T-1} \langle \log P(x_{t+1} | x_t, a_t) + \lambda R_t(x_t, a_t, x_{t+1}) \rangle_{q(x_{t+1}, x_t, a_t)}$$

Ranking different types of inference for planning

- For a given posterior, the bounds can be ordered monotonically

$$\left.\begin{matrix} F_\lambda^{\text{MAP}}(\boldsymbol{q}) \\ F_\lambda^{\text{marginal}^{\text{U}}}(\boldsymbol{q}) \end{matrix}\right\} \leq F_\lambda^{\text{MMAP}}(\boldsymbol{q}) \leq F_\lambda^{\textbf{planning}}(\boldsymbol{q}) \leq F_\lambda^{\text{marginal}}(\boldsymbol{q})$$

- … and also at the maximum

$$\left.\begin{matrix} F_\lambda^{\text{MAP}} \\ F_\lambda^{\text{marginal}^{\text{U}}} \end{matrix}\right\} \leq F_\lambda^{\text{MMAP}} \leq F_\lambda^{\textbf{planning}} \leq F_\lambda^{\text{marginal}}$$

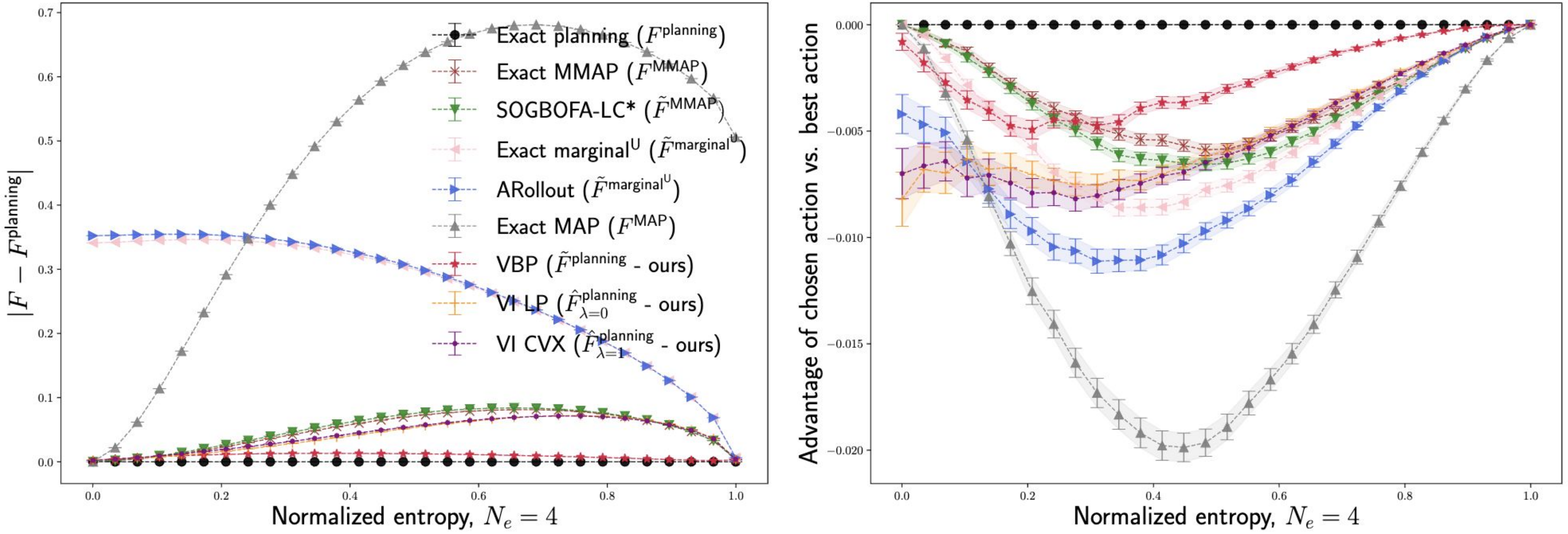
- Things simplify under deterministic dynamics

$$F_\lambda^{\text{marginal}^{\text{U}}} \leq F_\lambda^{\text{MAP}} = F_\lambda^{\text{MMAP}} = F_\lambda^{\textbf{planning}} \leq F_\lambda^{\text{marginal}}$$

Loopy BP recipe

- Replace the variational distribution with pseudomarginals (i.e., relax domain from marginal polytope to local polytope).
- Replace exact entropy with Bethe entropy.
  - For planning: Replace the *planning* entropy with the Bethe version of the *planning* entropy.
    - Non-concave problem → Value BP (loopy BP for planning).
    - It has a simple concave approximation.

The stochasticity of the dynamics is key



Other inference types lack *reactivity* in stochastic env.

- Example

- MMAP: Takes actions to move “left” in the above slider.
- VBP: Keeps the environment at the far “right” in the above slider.
    - Also, reacts to the environment and achieves maximum reward.

Summary

- All inference types correspond to VI with a modified entropy term.
- Planning is a distinct type of inference.
- This allows to compare the different types of inference for planning.
- LBP can be modified for planning, includes “backward reasoning”.
- Using these ideas, many inference algorithms can be adapted for planning, and vice versa.