Reinforcement Learning with History-Dependent Dynamic Contexts

Guy Tennenholtz, Nadav Merlis,
Lior Shani, Martin Mladenov, Craig Boutilier

Google Research

Motivation

- Many real-world settings are inherently history-dependent
- Challenging credit assignment for long-term histories
- We introduce a Logistic DCMDPs:
  - Inspired by Rescorla-Wagner model
  - Account for long-term history dependence
  - Allow for efficient credit assignment and exploration
- We provide theoretical regret guarantees and a practical algorithm

Dynamic Contextual MDP (DCMDP)

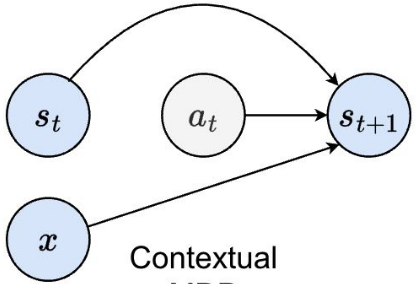
- Defined by the tuple $(\mathcal{X}, \mathcal{S}, \mathcal{A}, r, P, H)$
- DCMDP dynamics are **history-dependent**
  - Agent interacting with an environment.
  - Generating a sequence of states, actions, and contexts.
- Performance is measured in terms of value and regret

$$V_h^\pi(s, \tau) = \mathbb{E}_\pi\left[\sum_{t=h}^{H} r(s_t, a_t, x_t) \,\middle|\, s_h = s, \tau_h = \tau\right]$$

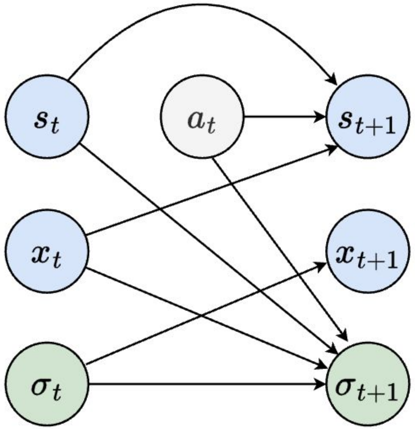
$$\mathrm{Reg}(K) = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k)$$

Special Cases of DCMDPs

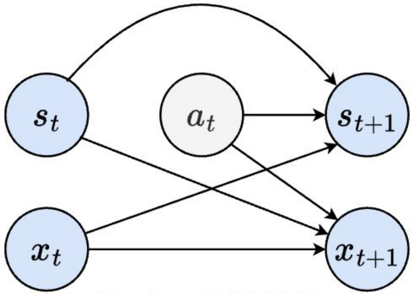
- Contextual MDPs: context remains fixed across transitions.
- Markov DCMDPs: context transitions are Markov.
  - Can be reduced to MDP
- Logistic DCMDPs (next)



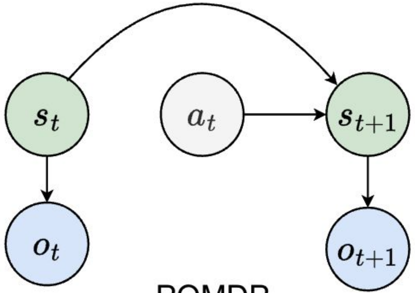
Contextual MDP



Logistic DCMDP



Markov DCMDP



POMDP

Logistic DCMDPs

General class of DCMDPs where history dependence is structured via an aggregation of state-action-context-dependent features.

$$P_{\boldsymbol{f}^*}(x_h^{(i)}|\tau_h) = z_i\left(\sum_{t=0}^{h-1}\alpha^{h-t-1}\boldsymbol{f}_t^*(s_t, a_t, x_t))\right)$$

Context transition function

State, action, context history at time step h

Softmax function

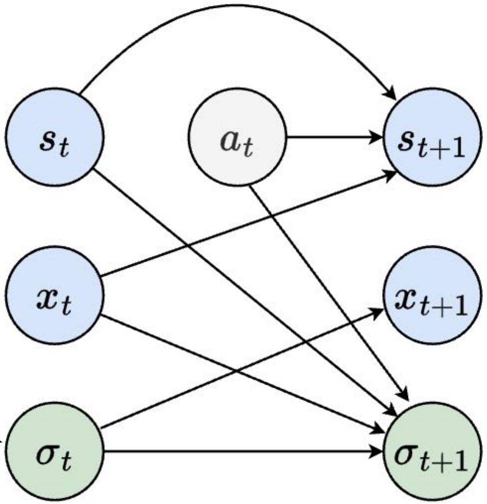
Discount over history (can equal 1 for no discount)

Unknown vector valued functions

Logistic DCMDPs

Sufficient Statistic

$$\boxed{\sum_{t=0}^{h-1} \alpha^{h-t-1} \boldsymbol{f}_t^*(s_t, a_t, x_t))}$$



Strong Theoretical Results

- A general RL method for logistic DCMDPs with unknown features.
  - Utilizes estimates of rewards, transitions and projected estimates of features.
  - Incorporates optimism to account for uncertainty.
- We address computational complexity:
  - We develop a local confidence bound for every state-action-context triple.
  - We construct an optimistic planner using a novel threshold mechanism.
- We prove statistically efficient regret guarantees.

DCZero

- Inspired by MuZero, DCZero incorporates representation, transition, and prediction networks for learning and acting.
- Unique to DCZero, an additional ensemble of networks estimates the unknown features using cross-entropy.
- Optimistic value is trained using our thresholding technique.
- We demonstrate the efficiency of DCZero on a difficult movie recommendation task with long history dependence.