Uncertain MDP

The differences from the MDP.

- The transition matrices or/and the reward function are not fully known.
- The uncertainties can be presented in a different ways: as an ambiguity sets or distributions constructed from historical data.

Uncertain MDP

Various ways for decision making with imprecise probabilities:

- Admissibility
- Maximal expected utility
- Maximality
- Interval dominance
- E-admissibility
- Γ-maximax
- Γ-minimax

Uncertain MDP

Policy that was learned on imprecise or incorrectly estimated MDP can be highly sub-optimal.
There is line of works that study the loss of learning the policy on the wrong MDP.

Uncertain MDP

Most of the researchers concentrated on **risk averse** approaches.

The general case of the problem when the uncertainty is defined via ambiguity sets or probabilities are extremely time consuming and requires complicated non-linear optimizations.

Risk Averse Uncertain MDP

The names in the literature:

MDPIP

Markov Decision Process with Imprecise Transition Probabilities

OR the older one

MDP with Uncertain Transition Probabilities

Robust MDP

- Objective (in terms of cost): Minimax- Minimizes the expected discounted total cost under the most adversarial parameter scenarios.
- Parameters belongs to ambiguity set constructed from historical data.
- For certain types of ambiguity set may be solved effectively using LP, second order cone problem , etc.
- Doesn't take into account any statistical information after constructing ambiguity sets, so the sets need to be constructed very carefully.

MDPIP

Distributionally Robust MDP

- Similar to Robust MDP except the uncertainty modeling
- Reward and Transition Matrix are treated as random variables following unknown distribution Q. Q belongs to ambiguity set, constructed from historical data.

Robust MDP with another risk averse objectives

- Instead of minimax cost , optimizing minimax regret (special definitions of the regret)

Distributionally Robust MDP Cons of MDPIP:

- Overly conservative solutions
- Difficulty constructing good ambiguity sets
- Very time consuming algorithms for many cases
- Difficulty in incorporating probability - worst case scenario may be very unlikely

BUT

- Very reach literature about MDPIP
- Good algorithms for many cases

BMDP

One particular formulation of MDPIP called *Bounded-Parameter Markov decision processes* (BMDP) may be useful for us.

BMDP

General overview:

- Ambiguity set: The transition probabilities and/or the rewards are given as closed segment by upper and lower bound.
- The BMDP is viewed as family of exact MDPs .
- Another interpretation for a BMDP is that the states of the BMDP actually represent sets (aggregates) of more primitive states that we choose to group together.

BMDP

Solutions:

- Interval policy evaluation and interval value iteration are developed (similar to common interval policy evaluation and value iteration).
- These algorithms are generalization of the standard MDP algorithms successive approximation and value iteration, respectively.
- It's worth nothing that the resulting intervals doesn't provide us with precise way of selecting the action.

BMDP

Definition: $M_{\updownarrow} = (Q, A, F_{\updownarrow}, R_{\updownarrow})$
Defines the set of exact MDPs with transitions and rewards laying in the interval OR an instance of BMDP.

To ensure that the BMDP is well defined - we require that for any action and state the sum of lower bounds will be $\leq$ than 1. And for the upper bound $\geq$ than 1.

BMDP

Algorithm concepts sketch:

- Despite the fact that the number of the MDPs is in general uncountable , only **finite subset X_M of $M_\updownarrow$ is of particular interest**. I.e. for any policy π and $M \in M_\updownarrow$ the value of π is bracketed by two policies in X_M

BMDP

To define the optimal value function we should first rank the intervals in some way.

Two proposed ways, and pessimistic to define a partial lexicographical order:

- **Optimistic**: according to upper bound.
- **Pessimistic**: according to lower bound.

BMDP

Any initial interval value function produces a sequence of interval value functions that converges in a **polynomial number of steps** to the true value.

The algorithm to compute interval value is very similar to the standard MDP computation of VI, except that :

- We must now be able to select an MDP M from the family $M_{\updownarrow}$ that minimizes (maximizes) the value .
- Selecting transition values of the MDP M we need to assure that they are well defined (sums up to 1)

BMDP

Algorithm idea:

To compute the lower bounds, the idea is to sort the possible destination states q into increasing order according to their lower value, and then choose the transition probabilities within the intervals specified by $F_{\updownarrow}$ so as to **send as much probability mass to the states early in the ordering as possible**.

Bayesian approach

Much more rare approach is the one that try to balance between risk-neutral and risk-averse.

Bayesian approach

Percentile Optimization for Markov Decision Processes with Parameter Uncertainty (2006)

Bayesian point of view that considers the parameters as random variables and lead to a performance measure called the percentile criterion.

Bayesian approach

Problematic approach:

$$\boxed{\max_\pi E_{P'}(E_x(\sum_t \alpha^t R(x_t)|X_0, \pi, P'))}$$

- Because of the non-linear effect of P' on the expected return, evaluating the objective of this problem for a given policy is difficult.
- Most likely (or expected) parameters in the nominal problem leads to a strong bias in the performance of the chosen policy.

Bayesian approach

The optimization problem:

$$\begin{array}{l} \max_{\pi, y \in \mathbb{R}} \quad y \\ \text{s.t.} P_{P'} \langle E_x(\sum_t \alpha^t R(x_t) | X_0, \pi, P') \geq y \rangle \geq 1 - \epsilon \end{array}$$

For a given policy π, the above chance constrained problem gives us a $1 - \epsilon$ guarantee that π will perform better than y* (the optimal value of the problem), under the distribution of P'.

* If $\epsilon = 0$ the problem becomes robust MDP problem. * Similar to VaR (Value at risk) estimation

Bayesian approach

The problem is hard to solve in general , but using the **Dirichlet prior** we can generate near optimal policy given sufficient amount of Data.

Bayesian approach

In the end we are required to solve non-convex optimization problem for a relatively complicated expression. The authors did it using gradient descent.

Bayesian approach

Another corresponding approach.
In short the objective is :

$$\boxed{\min_\pi \rho_{P\sim\mu}(E^{P,\pi}(\sum_t \alpha^t R(s,a)))}$$

- When ρ is the risk functional applied to the expression with respect to uncertainty in P which is quantified by μ.
  ρ may be Var, Conditional VaR, mean-variance etc.

The approximation of the objective function is done by sampling and simulation (not analytically).