A Theoretical Perspective on Sequential Decision Making with Preference Feedback

Simone Drago, Marco Mussi, and Alberto Maria Metelli
{simone.drago, marco.mussi, albertomaria.metelli}@polimi.it

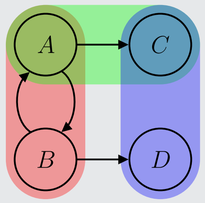
Politecnico Milano 1863

Setting

Preference

$$\text{PrefMDP} = (\underbrace{\mathcal{S}, \mathcal{A}, H, p, \mu}_{\text{MDP}\setminus\text{R}}, \preceq_{\mathcal{T}})$$

$\preceq_{\mathcal{T}} \subseteq \mathcal{T} \times \mathcal{T}$ is a **partial (pre)order** over the trajectory space $\mathcal{T}$



($\prec$) C is preferred over A
($\succeq$) A is equivalent to B
($\|$) C is incomparable to D

Assumption: Human expresses preferences based on an underlying (unknown) utility

Utility

$$\text{UtilMDP} = (\underbrace{\mathcal{S}, \mathcal{A}, H, p, \mu}_{\text{MDP}\setminus\text{R}}, \boldsymbol{u})$$

$\boldsymbol{u} : \mathcal{T} \to \mathbb{R}^m$ is a **multi-dimensional utility**

Expected utility of a policy $\pi \in \Pi$:
$\boldsymbol{J}(\pi; \boldsymbol{u}) \coloneqq \sum_{\tau \in \mathcal{T}} d_\pi(\tau) \boldsymbol{u}(\tau) = \langle d_\pi, \boldsymbol{u} \rangle$,
where d_π is the distribution over trajectories induced by π

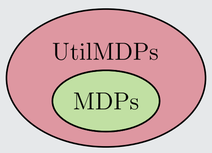
Assumption: Human expresses preferences based on an (unknown) Markovian reward

Reward

$$\text{MDP} = (\underbrace{\mathcal{S}, \mathcal{A}, H, p, \mu}_{\text{MDP}\setminus\text{R}}, \boldsymbol{r})$$

$\boldsymbol{r} \coloneqq (r_h)_{h \in [\![H]\!]}$ is a **stage-dependent multi-dimensional reward** function

Trajectory return:
$\boldsymbol{u_r}(\tau) \coloneqq \sum_{h=1}^{H} r_h(s_h, a_h)$
Expected policy return:
$\boldsymbol{J}(\pi; \boldsymbol{r}) \coloneqq \boldsymbol{J}(\pi; \boldsymbol{u_r})$



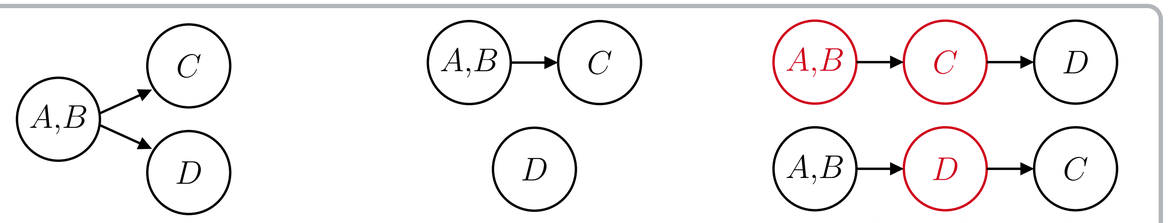
Compatible Utilities

Compatible Utility

$\boldsymbol{u}$ is compatible with $\preceq_{\mathcal{T}}$ if $\forall\, \tau, \tau' \in \mathcal{T}$:
$\tau \preceq_{\mathcal{T}} \tau' \Rightarrow \boldsymbol{u}(\tau) \preceq \boldsymbol{u}(\tau')$ (*element-wise*)

Realizer Dimension	Exists?	Computational Complexity
$< \dim(\preceq_{\mathcal{T}})$	✗	—
$= \dim(\preceq_{\mathcal{T}})$	✓	NP-hard
$> \dim(\preceq_{\mathcal{T}})$	✓	Poly($\lvert\mathcal{T}\rvert$)

where $\dim(\cdot)$ is the order dimension



(*i*) Represent the quotient set w.r.t. equivalence of $\preceq_{\mathcal{T}}$ as a DAG

(*ii*) Solve a *minimum path cover* problem to obtain a set of $\text{width}(\preceq_{\mathcal{T}})$ chains

(*iii*) Extend each chain to obtain a linear extension of $\preceq_{\mathcal{T}}$, accounting for incomparabilities

Procedure to derive a realizer of size $\text{width}(\preceq_{\mathcal{T}}) \geqslant \dim(\preceq_{\mathcal{T}})$ in $\mathcal{O}(|\mathcal{T}|^3)$

Policy Dominance

Policy Dominance

Policy π $\preceq_{\mathcal{T}}$-strictly dominates policy π' ($\pi' \prec_\Pi \pi$) if it yields a strictly better expected utility

$\boldsymbol{J}(\pi; \boldsymbol{u}) - \boldsymbol{J}(\pi'; \boldsymbol{u}) > \boldsymbol{0}$ (*element-wise*)

for every compatible utility function $\boldsymbol{u}$

$\preceq_{\mathcal{T}}$-Pareto Optimality

Set of $\preceq_{\mathcal{T}}$-Pareto optimal policies:

$\Pi^*(\preceq_{\mathcal{T}}) \coloneqq \{\pi \in \Pi : \nexists \pi' \in \Pi \text{ s.t. } \pi \prec_\Pi \pi'\}$

How to Evaluate?

$\pi' \preceq_\Pi \pi$ can be verified by evaluating if:

$$\forall\, n \in [\![|\mathcal{T}|]\!] : \sum_{i=1}^{n} (d_\pi(i) - d_{\pi'}(i)) \geqslant 0$$

holds for every linear extension of $\preceq_{\mathcal{T}}$, where index i represents the i-th trajectory sorted w.r.t. the linear extension

Open Question

Is there an efficient evaluation method?

Trivial solution requires the evaluation of $\mathcal{O}(|\mathcal{T}|!)$ linear extensions of $\preceq_{\mathcal{T}}$

Approximation via Markovian Rewards

Why the Need for Approximation?

Preferences	+	Most general type of feedback
	−	Intractable without introducing a structure
Utilities	+	Assign numerical signals to each trajectory
	−	Complexity of learning is exponential in H
Rewards	+	Enable efficient learning
	−	Less representational power

Human-friendliness ↕ Tractability

(Convex) Quadratic Program

Goal: Find $\boldsymbol{u}$ and $\boldsymbol{r}$ that best represent $\preceq_{\mathcal{T}}$
Input: A realizer $\{\leqslant_{\mathcal{T},i}\}_{i \in [\![m]\!]}$ of $\preceq_{\mathcal{T}}$ of size m
Idea: Jointly choose $\boldsymbol{u}$ compatible with the realizer and $\boldsymbol{r}$ as Markovian approximation of $\boldsymbol{u}$
Define: $\boldsymbol{B} \in \{0,1\}^{|\mathcal{T}| \times |\mathcal{S}||\mathcal{A}|H}$ (binary encoding of $\mathcal{T}$) and $\boldsymbol{A} \coloneqq \boldsymbol{I}_{|\mathcal{T}|} - \boldsymbol{B}(\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top$ (OLS)

$$\eta^* \coloneqq \min \|\boldsymbol{A}\boldsymbol{u}\|_F^2$$
$$\text{s.t. } u_j(i+1) \leqslant u_j(i) - \varepsilon \;\; \forall i \in [\![|\mathcal{T}| - 1]\!],\; j \in [\![m]\!]$$

If $\eta^* = 0 \to \boldsymbol{u_r} = \boldsymbol{u} \to$ Preferences derive from a Markovian reward function

If $\eta^* > 0 \to$ Preferences cannot be expressed via Markovian rewards
$\to$ Using $\boldsymbol{r}$, we learn how to solve a simpler surrogate problem
$\to$ Such an approximation introduces a suboptimality in terms of the performance of the optimal policy bounded by $2\sqrt{m\eta^*}$