Utility Theory for Sequential Decision Making

Mehran Shakerinava [1,2] Siamak Ravanbakhsh [1,2]

[1] McGill University

[2] Mila Quebec AI Institute

ICML 2022

Motivation

- In Reinforcement Learning (RL) the objective of the agent is to maximize expected sum of rewards.

Motivation

- In Reinforcement Learning (RL) the objective of the agent is to maximize expected sum of rewards.
- The reward hypothesis: *"That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*

Motivation

- In Reinforcement Learning (RL) the objective of the agent is to maximize expected sum of rewards.
- The reward hypothesis: *"That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*
- Von Neumann-Morgenstern (VNM) utility theory offers a principled approach.

Motivation

- In Reinforcement Learning (RL) the objective of the agent is to maximize expected sum of rewards.
- The reward hypothesis: *"That all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*
- Von Neumann-Morgenstern (VNM) utility theory offers a principled approach.
- We extend this theory to sequential decision making.

Von Neumann-Morgenstern (VNM) Utility Theory

- $\mathcal{O}$: set of outcomes

Von Neumann-Morgenstern (VNM) Utility Theory

- $\mathcal{O}$: set of outcomes
- $\mathcal{L}$: set of all lotteries of outcomes

Von Neumann-Morgenstern (VNM) Utility Theory

- $\mathcal{O}$: set of outcomes
- $\mathcal{L}$: set of all lotteries of outcomes

Example

$$\mathcal{O} = \{\square, \circ, \triangle\}$$

$$M = p_1\square + p_2 \circ + p_3\triangle \qquad (p_1 + p_2 + p_3 = 1)$$

$$N = q_1\triangle + q_2 M \qquad (q_1 + q_2 = 1)$$

Von Neumann-Morgenstern (VNM) Utility Theory

- $\mathcal{O}$: set of outcomes
- $\mathcal{L}$: set of all lotteries of outcomes

Example

$$\mathcal{O} = \{\square, \circ, \triangle\}$$

$$M = p_1\square + p_2 \circ + p_3\triangle \qquad (p_1 + p_2 + p_3 = 1)$$

$$N = q_1\triangle + q_2 M \qquad (q_1 + q_2 = 1)$$

- $\succsim$: preference relation defined over $\mathcal{L}$

Von Neumann-Morgenstern (VNM) Utility Theory

- $\mathcal{O}$: set of outcomes
- $\mathcal{L}$: set of all lotteries of outcomes

Example

$$\mathcal{O} = \{\square, \circ, \triangle\}$$

$$M = p_1\square + p_2 \circ + p_3\triangle \qquad (p_1 + p_2 + p_3 = 1)$$

$$N = q_1\triangle + q_2 M \qquad (q_1 + q_2 = 1)$$

- $\succsim$: preference relation defined over $\mathcal{L}$

Definition (Utility function)

A function $u : \mathcal{L} \to \mathbb{R}$, such that for all $M, N \in \mathcal{L}$,

$$M \succsim N \iff u(M) \geq u(N).$$

Von Neumann-Morgenstern (VNM) Utility Theory (cont.)

VNM Rationality Axioms

Von Neumann-Morgenstern (VNM) Utility Theory (cont.)

VNM Rationality Axioms

- ***Completeness***: *For all* $M, N \in \mathcal{L}$, $M \succsim N$ *or* $N \succsim M$.

Von Neumann-Morgenstern (VNM) Utility Theory (cont.)

VNM Rationality Axioms

- ***Completeness***: *For all* $M, N \in \mathcal{L}$, $M \succsim N$ *or* $N \succsim M$.
- ***Transitivity***: *For all* $M, N, K \in \mathcal{L}$, *if* $M \succsim N$ *and* $N \succsim K$, *then* $M \succsim K$.

Von Neumann-Morgenstern (VNM) Utility Theory (cont.)

VNM Rationality Axioms

- ***Completeness***: *For all* $M, N \in \mathcal{L}$, $M \succsim N$ *or* $N \succsim M$.
- ***Transitivity***: *For all* $M, N, K \in \mathcal{L}$, *if* $M \succsim N$ *and* $N \succsim K$, *then* $M \succsim K$.
- ***Continuity***: *For all lotteries* $M \succsim N \succsim K$, *there exists* $p \in [0, 1]$ *such that* $pM + (1 - p)K \approx N$.

Von Neumann-Morgenstern (VNM) Utility Theory (cont.)

VNM Rationality Axioms

- ***Completeness***: *For all* $M, N \in \mathcal{L}$, $M \succsim N$ *or* $N \succsim M$.
- ***Transitivity***: *For all* $M, N, K \in \mathcal{L}$, *if* $M \succsim N$ *and* $N \succsim K$, *then* $M \succsim K$.
- ***Continuity***: *For all lotteries* $M \succsim N \succsim K$, *there exists* $p \in [0, 1]$ *such that* $pM + (1-p)K \approx N$.
- ***Independence***: *For all* $M, N, K \in \mathcal{L}$ *and for all* $p \in [0, 1]$,

$$M \succsim N \iff (1-p)M + pK \succsim (1-p)N + pK.$$

Von Neumann-Morgenstern (VNM) Utility Theory (cont.)

VNM Rationality Axioms

- ***Completeness***: *For all* $M, N \in \mathcal{L}$, $M \succsim N$ *or* $N \succsim M$.
- ***Transitivity***: *For all* $M, N, K \in \mathcal{L}$, *if* $M \succsim N$ *and* $N \succsim K$, *then* $M \succsim K$.
- ***Continuity***: *For all lotteries* $M \succsim N \succsim K$, *there exists* $p \in [0, 1]$ *such that* $pM + (1 - p)K \approx N$.
- ***Independence***: *For all* $M, N, K \in \mathcal{L}$ *and for all* $p \in [0, 1]$,

$$M \succsim N \iff (1 - p)M + pK \succsim (1 - p)N + pK.$$

Theorem (VNM Utility Theorem)

$\succsim$ *satisfies the VNM axioms* $\iff$ *there exists a utility function* u *such that*

$$u\left(\sum_{x \in \mathcal{O}} p(x)x\right) = \sum_{x \in \mathcal{O}} p(x)u(x).$$

Extension to Sequential Decision Making

- The agent's actions (stochastically) determine a **trajectory** in a state-space $\mathcal{S}$.

Extension to Sequential Decision Making

- The agent's actions (stochastically) determine a **trajectory** in a state-space $\mathcal{S}$.
- $\mathcal{O} = \{\text{all trajectories}\}$

Extension to Sequential Decision Making

- The agent's actions (stochastically) determine a **trajectory** in a state-space $\mathcal{S}$.
- $\mathcal{O} = \{\text{all trajectories}\}$
- Preferences are defined over *lotteries of trajectories*.

Extension to Sequential Decision Making

- The agent's actions (stochastically) determine a **trajectory** in a state-space $\mathcal{S}$.
- $\mathcal{O} = \{\text{all trajectories}\}$
- Preferences are defined over *lotteries of trajectories*.
- So far, the structure of the decision process is not taken into account.

Extension to Sequential Decision Making

- The agent's actions (stochastically) determine a **trajectory** in a state-space $\mathcal{S}$.
- $\mathcal{O} = \{\text{all trajectories}\}$
- Preferences are defined over *lotteries of trajectories*.
- So far, the structure of the decision process is not taken into account.
- **Notation**
  - transitions: $t, t_1, t_2, \ldots$
  - trajectories: τ, τ_1, τ_2
  - lotteries: M, N, J, K

Memoryless Sequential Decision Making

Axiom (Memorylessness)

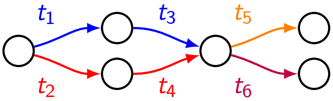
$\tau \cdot M \succsim \tau \cdot N \iff M \succsim N$, *where* $\cdot$ *denotes concatenation.*

Memoryless Sequential Decision Making

Axiom (Memorylessness)

$\tau \cdot M \succsim \tau \cdot N \iff M \succsim N$, *where* $\cdot$ *denotes concatenation.*

- Example:



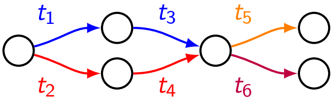
$$\langle t_1, t_3, t_5 \rangle \succsim \langle t_1, t_3, t_6 \rangle \iff \langle t_5 \rangle \succsim \langle t_6 \rangle$$

Memoryless Sequential Decision Making

Axiom (Memorylessness)

$\tau \cdot M \succsim \tau \cdot N \iff M \succsim N$, where $\cdot$ denotes concatenation.

- Example:



$$\langle t_1, t_3, t_5 \rangle \succsim \langle t_1, t_3, t_6 \rangle \iff \langle t_5 \rangle \succsim \langle t_6 \rangle$$

Theorem

Utilities take the form $u(t \cdot \tau) = r(t) + m(t)u(\tau)$, where r is the reward function and m is the reward multiplier function.

An Axiom for Markov Decision Processes

Axiom (Additivity)

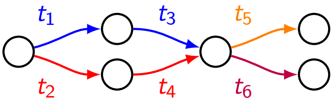
$$p(\tau_1 \cdot M) + (1-p)J \succsim p(\tau_1 \cdot N) + (1-p)K$$
$$\iff p(\tau_2 \cdot M) + (1-p)J \succsim p(\tau_2 \cdot N) + (1-p)K$$

An Axiom for Markov Decision Processes

Axiom (Additivity)

$$p(\tau_1 \cdot M) + (1-p)J \succsim p(\tau_1 \cdot N) + (1-p)K$$
$$\iff p(\tau_2 \cdot M) + (1-p)J \succsim p(\tau_2 \cdot N) + (1-p)K$$

- Example:



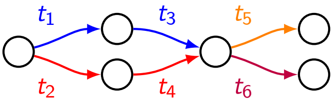
$$\langle t_1, t_3\rangle \succsim \langle t_2, t_4\rangle \text{ and } \langle t_5\rangle \succsim \langle t_6\rangle \implies \langle t_1, t_3, t_5\rangle \succsim \langle t_2, t_4, t_6\rangle$$

An Axiom for Markov Decision Processes

Axiom (Additivity)

$$p(\tau_1 \cdot M) + (1-p)J \succsim p(\tau_1 \cdot N) + (1-p)K$$
$$\iff p(\tau_2 \cdot M) + (1-p)J \succsim p(\tau_2 \cdot N) + (1-p)K$$

- Example:



$$\langle t_1, t_3 \rangle \succsim \langle t_2, t_4 \rangle \text{ and } \langle t_5 \rangle \succsim \langle t_6 \rangle \implies \langle t_1, t_3, t_5 \rangle \succsim \langle t_2, t_4, t_6 \rangle$$

Theorem

Utilities take the form $u(\tau) = \sum_{t \in \tau} r(t)$, where r is the reward function.

Discussion

"That all of what we mean by ***goals and purposes*** *can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*

- What exactly does "***goals and purposes***" mean in the reward hypothesis?

Discussion

*"That all of what we mean by **goals and purposes** can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*

- What exactly does "***goals and purposes***" mean in the reward hypothesis?
- If the goal is to achieve a desired policy $\pi^\star$, then we simply set

$$r(s, a, s') = \begin{cases} +1 & a = \pi^\star(s) \\ -1 & \text{otherwise.} \end{cases}$$

Discussion

*"That all of what we mean by **goals and purposes** can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*

- What exactly does "***goals and purposes***" mean in the reward hypothesis?
- If the goal is to achieve a desired policy $\pi^\star$, then we simply set

$$r(s, a, s') = \begin{cases} +1 & a = \pi^\star(s) \\ -1 & \text{otherwise.} \end{cases}$$

- goals and purposes = *rational preferences*

Discussion

"That all of what we mean by ***goals and purposes*** *can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*

- What exactly does "***goals and purposes***" mean in the reward hypothesis?
- If the goal is to achieve a desired policy $\pi^\star$, then we simply set

$$r(s, a, s') = \begin{cases} +1 & a = \pi^\star(s) \\ -1 & \text{otherwise.} \end{cases}$$

- goals and purposes = *rational preferences*
- Any two behaviours can be compared.

Discussion

*"That all of what we mean by **goals and purposes** can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."*

- What exactly does "***goals and purposes***" mean in the reward hypothesis?
- If the goal is to achieve a desired policy $\pi^\star$, then we simply set

$$r(s, a, s') = \begin{cases} +1 & a = \pi^\star(s) \\ -1 & \text{otherwise.} \end{cases}$$

- goals and purposes = *rational preferences*
- Any two behaviours can be compared.
- If a given task can be represented as a *rational and additive* preference relation, then it can be modeled as an MDP.