Multi-Objective POMDPs with Lexicographic Reward Preferences

Kyle Hollins Wray and Shlomo Zilberstein
College of Information and Computer Sciences
University of Massachusetts, Amherst, MA 01003
{wray, shlomo}@cs.umass.edu
Motivation

- Multiple objectives are common in practice
Motivation

- Multiple objectives are common in practice
  - Semi-Autonomous Driving Domain
Motivation

• Multiple objectives are common in practice
  – Semi-Autonomous Driving Domain

Minimize time
Motivation

- Multiple objectives are common in practice
  - Semi-Autonomous Driving Domain

Minimize time                           Maximize autonomy

Source: http://video.pbs.org/video/2365404300/
Motivation

- Multiple objectives are common in practice
- Scalarization
  - Apply *scalarization function*
    - Maps multiple objectives to a single objective
Motivation

- Multiple objectives are common in practice

Scalarization

- Apply \textit{scalarization function}
  - Maps multiple objectives to a single objective

- Issues
  - Function and/or weight selection
  - Differing units among objective functions
Motivation

- Multiple objectives are common in practice
- Scalarization
- Preference orderings natural in many domains
  - **Optimize** objective functions in sequence
Motivation

- Multiple objectives are common in practice
- Scalarization
- Preference orderings natural in many domains
  - **Optimize** objective functions in sequence
  - **Restrict actions** available each time
Motivation

• Multiple objectives are common in practice
• Scalarization
• Preference orderings natural in many domains
  - *Optimize* objective functions *in sequence*
  - *Restrict actions* available each time
  - *Slack liberates* more *actions* for remaining objectives
Lexicographic POMDP

- Multi-Objective POMDP
  - $S$ is a set of $n$ states
  - $A$ is a set of $m$ actions
  - $\Omega$ is a set of $z$ observations
  - $T : S \times A \times S \to [0, 1]$ is a state transition function
  - $O : A \times S \times \Omega \to [0, 1]$ is an observation function
  - $R = [R_1, ..., R_k]^T$ is a vector of $k$ reward functions
    - $R_i : S \times A \to R$ is reward $i$
Lexicographic POMDP

- Multi-Objective POMDP
  - $S$ is a set of $n$ states
  - $A$ is a set of $m$ actions
  - $\Omega$ is a set of $z$ observations
  - $T : S \times A \times S \to [0, 1]$ is a state transition function
  - $O : A \times S \times \Omega \to [0, 1]$ is an observation function
  - $R = [R_1, ..., R_k]^T$ is a vector of $k$ reward functions
    - $R_i : S \times A \to \mathbb{R}$ is reward $i$
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
  - Prefer to maximize *expected value* for $R_1$, then $R_2$, etc.
Lexicographic POMDP

- Preference ordering over rewards $R$
  - Prefer to maximize **expected value** for $R_1$, then $R_2$, etc.
  - Issue
    - Belief state tie-breaking likely rare
Lexicographic POMDP

- Preference ordering over rewards $\mathbb{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
Lexicographic POMDP

- Preference ordering over rewards $\mathbb{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
  - **Local Slack** $\eta = <\eta_1, ..., \eta_k>$
    - One-step deviation from optimal action at a belief state
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
  - **Local Slack** $\eta = \langle \eta_1, \ldots, \eta_k \rangle$
    - One-step deviation from optimal action at a belief state

“Navigate to the destination as fast as possible, and take alternate roads we encounter on which we can drive autonomously, so as long as they would add less than 1 minute to our ride.”
Lexicographic POMDP

- Preference ordering over rewards $\mathbb{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
  - **Local Slack** $\eta = <\eta_1, ..., \eta_k>$
    - One-step deviation from optimal action at a belief state
  - **Global Slack** $\delta = <\delta_1, ..., \delta_k>$
    - Accumulated deviation from optimal over all time
Lexicographic POMDP

- Preference ordering over rewards $R$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
  - **Local Slack** $\eta = <\eta_1, \ldots, \eta_k>$
    - One-step deviation from optimal action at a belief state
  - **Global Slack** $\delta = <\delta_1, \ldots, \delta_k>$
    - Accumulated deviation from optimal over all time

“Navigate to the destination as fast as possible, and if you can drive autonomously 'more' throughout the trip, then do so as long as it adds less than 10 minutes to our ride.”
Lexicographic POMDP

- Preference ordering over rewards $\mathbb{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
- **Action restriction** limits actions at belief states
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
- **Action restriction** limits actions at belief states

\[
A_{i+1}(b) = \{ a \in A_i(b) | \max_{a' \in A_i(b)} Q_i(b, a') - Q_i(b, a) \leq \eta_i \}
\]
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
- **Action restriction** limits actions at belief states

$$A_{i+1}(b) = \{ a \in A_i(b) \mid \max_{a' \in A_i(b)} Q_i(b, a') - Q_i(b, a) \leq \eta_i \}$$

Keep action
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
- **Action restriction** limits actions at belief states

$$A_{i+1}(b) = \{ a \in A_i(b) \mid \max_{a' \in A_i(b)} Q_i(b, a') - Q_i(b, a) \leq \eta_i \}$$

Keep action only if **best** value
Lexicographic POMDP

- Preference ordering over rewards $\mathbf{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
- **Action restriction** limits actions at belief states

$$A_{i+1}(b) = \{ a \in A_i(b) \mid \max_{a' \in A_i(b)} Q_i(b, a') - Q_i(b, a) \leq \eta_i \}$$

Keep action only if best value is reduced by
Lexicographic POMDP

- Preference ordering over rewards \( R \)
- **Slack** \((\eta, \delta)\) – allowable deviation from optimal
- **Action restriction** limits actions at belief states

\[
A_{i+1}(b) = \{ a \in A_i(b) \mid \max_{a' \in A_i(b)} Q_i(b, a') - Q_i(b, a) \leq \eta_i \}
\]

Keep action only if best value is reduced by less than local slack
Lexicographic POMDP

- Preference ordering over rewards $\mathbb{R}$
- **Slack** $(\eta, \delta)$ – allowable deviation from optimal
- **Action restriction** limits actions at belief states
Lexicographic POMDP Solutions

- Lexicographic Value Iteration (LVI)
  - Full Policy Tree Values with Action Restriction

- Lexicographic Point-Based Value Iteration (LPBVI)
  - Belief Point Values with Action Restriction
Lexicographic POMDP Solutions

• Lexicographic Value Iteration (LVI)
  – Full Policy Tree Values with Action Restriction

• Lexicographic Point-Based Value Iteration (LPBVI)
  – Belief Point Values with Action Restriction

• Two Cases
  – Local Slack Given
  – Global Slack Given
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
  - Accumulated worst-case error from local slack
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
  - Accumulated worst-case error from local slack

\[
\sum_{t=0}^{\infty} \gamma^t \eta_i = \frac{\eta_i}{1 - \gamma}
\]
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
  - Accumulated worst-case error from local slack
  - Proposition #1:
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
  - Accumulated worst-case error from local slack
  - Proposition #1: Given prior local action restrictions,

\[
\text{If } \eta_i = (1 - \gamma) \delta_i, \text{ then } \forall b \in B, V_i^{\eta}(b) - V_i^{\pi}(b) \leq \delta_i.
\]
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
  - Accumulated worst-case error from local slack
  - Proposition #1: Given prior local action restrictions,

\[
\text{If } \eta_i = (1 - \gamma)\delta_i, \text{ then } \forall b \in B, V_i^{\eta}(b) - V_i^{\pi}(b) \leq \delta_i.
\]

Local slack assignment
Lexicographic POMDP Solution: LVI

- Full Policy Tree Values with Action Restriction
  - Accumulated worst-case error from local slack
  - Proposition #1: \textit{Given prior local action restrictions,}

\[
\text{If } \eta_i = (1 - \gamma) \delta_i, \text{ then } \forall b \in B, \ V_i^\eta(b) - V_i^\pi(b) \leq \delta_i.
\]

Local slack assignment \hspace{3.7in} Final policy error is bounded by global slack
Lexicographic POMDP Solution: LPBVI

- Belief Point Values with Action Restriction
Lexicographic POMDP Solution: LPBVI

- Belief Point Values with Action Restriction
  - Additional error from non-representative beliefs
Lexicographic POMDP Solution: LPBVI

- Belief Point Values with Action Restriction
  - Additional error from non-representative beliefs
  - Proposition #2:
Lexicographic POMDP Solution: LPBVI

- Belief Point Values with Action Restriction
  - Additional error from non-representative beliefs
  - Proposition #2: Given prior local action restrictions,

\[
\eta_i = \max \left\{ 0, (1 - \gamma)\delta_i - \frac{R_i^{\text{max}} - R_i^{\text{min}}}{1 - \gamma}\delta_B \right\}
\]

then \( \forall b \in B, V_i^{\eta^B}(b) - V_i^{\pi^B}(b) \leq \delta_i \).
Lexicographic POMDP Solution: LPBVI

- Belief Point Values with Action Restriction
  - Additional error from non-representative beliefs
  - Proposition #2: Given prior local action restrictions,

\[
\eta_i = \max \left\{ 0, (1 - \gamma)\delta_i - \frac{R_i^{\text{max}} - R_i^{\text{min}}}{1 - \gamma}\delta_B \right\}
\]

*Local slack* assignment

Then \( \forall b \in B, V_i^{\eta B}(b) - V_i^{\pi B}(b) \leq \delta_i \).
Lexicographic POMDP Solution: LPBVI

- **Belief Point Values with Action Restriction**
  - **Additional error** from non-representative beliefs
  - Proposition #2: Given prior local action restrictions, if
    \[
    \eta_i = \max \left\{ 0, (1 - \gamma)\delta_i - \frac{R^\text{max}_i - R^\text{min}_i}{1 - \gamma} \delta_B \right\}
    \]
    then \( \forall b \in B, V_i^{\eta B}(b) - V_i^{\pi B}(b) \leq \delta_i \).
Lexicographic POMDP Solution: LPBVI

- Belief Point Values with Action Restriction
  - Additional error from non-representative beliefs
  - Proposition #2: Given prior local action restrictions,

\[
\eta_i = \max \left\{ 0, (1 - \gamma) \delta_i - \frac{R^\text{max}_i - R^\text{min}_i}{1 - \gamma} \delta_B \right\}
\]

\text{Local slack assignment}

then \( \forall b \in B, V_i^{\eta^B}(b) - V_i^{\pi^B}(b) \leq \delta_i \).
Lexicographic POMDP Solution: LPBVI

- **Belief Point Values with Action Restriction**
  - Additional error from non-representative beliefs
  - Proposition #2: *Given prior local action restrictions,*

\[
\eta_i = \max \left\{ 0, (1 - \gamma) \delta_i - \frac{R_i^{\max} - R_i^{\min}}{1 - \gamma} \delta_B \right\}
\]

**Local slack** assignment

Then \( \forall b \in B, \ V_i^{\eta^B} (b) - V_i^{\pi^B} (b) \leq \delta_i \).

**Final policy error is bounded** by **global slack**
Experimentation

- Semi-Autonomous Driving Domain
Experimentation

- Semi-Autonomous Driving Domain
  - Semi-Autonomous System (SAS)
    - Requires agent-human collaboration to achieve a goal
Experimentation

- Semi-Autonomous Driving Domain
  - Semi-Autonomous System (SAS)
  - Vehicle may be autonomous on autonomy-capable roads
Experimentation

- Semi-Autonomous Driving Domain
  - Semi-Autonomous System (SAS)
  - Vehicle may be autonomous on autonomy-capable roads
  - Driver may be attentive or tired
Experimentation

- **Semi-Autonomous Driving Domain**
  - Semi-Autonomous System (SAS)
  - Vehicle may be autonomous on **autonomy-capable** roads
  - Driver may be **attentive** or **tired**
    - True state **unknown**
    - **Monitored** by sensors
Experimentation

- Semi-Autonomous Driving Domain
  - Semi-Autonomous System (SAS)
  - Vehicle may be autonomous on **autonomy-capable** roads
  - Driver may be **attentive** or **tired**
  - Rewards **time** and **autonomy**
## Experimentation

| City         | $|S|$ | $|A|$ | $|\Omega|$ | $|B|$ | $V_1^n(b^0)$ | $V_2^n(b^0)$ | CPU ($h = 10$) |
|--------------|-----|-----|---------|-----|-------------|-------------|----------------|
| Austin       | 92  | 8   | 2       | 230 | 57.4        | 35.9        | 14.796         |
| San Franc.   | 172 | 8   | 2       | 430 | 97.8        | 53.8        | 51.641         |
| Denver       | 176 | 8   | 2       | 440 | 123.7       | 77.3        | 60.217         |
| Baltimore    | 220 | 8   | 2       | 550 | 56.2        | 43.9        | 104.031        |
| Pittsburgh   | 268 | 10  | 2       | 670 | 148.0       | 142.2       | 169.041        |
| L.A.         | 380 | 8   | 2       | 950 | 167.9       | 114.4       | 298.794        |
| Chicago      | 404 | 10  | 2       | 1010| 67.4        | 31.6        | 399.395        |
| Seattle      | 432 | 10  | 2       | 1080| 111.2       | 66.9        | 497.061        |
| N.Y.C.       | 1064| 12  | 2       | 2660| 108.1       | 73.7        | n/a            |
| Boston       | 2228| 12  | 2       | 5570| 109.3       | 79.2        | n/a            |
Experimentation

- GPU-Optimized PBVI
  - Parallelize the inner argmax over B, A, and Ω
  - Parallelize the outer argmax belief update equation
# Experimentation

| City          | $|S|$ | $|A|$ | $|\Omega|$ | $|B|$ | $V_1^n(b^0)$ | $V_2^n(b^0)$ | CPU ($h = 10$) | GPU ($h = 500$) |
|--------------|----|----|---------|----|-------------|-------------|---------------|----------------|
| Austin       | 92 | 8  | 2       | 230| 57.4        | 35.9        | 14.796        | 3.798          |
| San Franc.   | 172| 8  | 2       | 430| 97.8        | 53.8        | 51.641        | 8.056          |
| Denver       | 176| 8  | 2       | 440| 123.7       | 77.3        | 60.217        | 8.299          |
| Baltimore    | 220| 8  | 2       | 550| 56.2        | 43.9        | 104.031       | 11.782         |
| Pittsburgh   | 268| 10 | 2       | 670| 148.0       | 142.2       | 169.041       | 19.455         |
| L.A.         | 380| 8  | 2       | 950| 167.9       | 114.4       | 298.794       | 25.535         |
| Chicago      | 404| 10 | 2       | 1010|67.4        | 31.6        | 399.395       | 36.843         |
| Seattle      | 432| 10 | 2       | 1080|111.2       | 66.9        | 497.061       | 48.204         |
| N.Y.C.       | 1064|12 | 2       | 2660|108.1       | 73.7        | n/a           | 351.288        |
| Boston       | 2228|12 | 2       | 5570|109.3       | 79.2        | n/a           | 2424.961       |
Conclusion

- Proposed Lexicographic POMDP
- Introduced Two Algorithms: LVI and LPBVI
- Applied to Semi-Autonomous Driving Domain
- Developed GPU-based Optimization for (L)PBVI
Multi-Objective POMDPs with Lexicographic Reward Preferences

Kyle Hollins Wray and Shlomo Zilberstein
College of Information and Computer Sciences
University of Massachusetts, Amherst, MA 01003
{wray, shlomo}@cs.umass.edu
Appendix

Multi-Objective POMDPs with Lexicographic Reward Preferences
Appendix

![Attentive Policy](image1)

![Tired Policy](image2)
Appendix

\[ A_{i+1}(b) = \{ a \in A_i(b) \mid \max_{a' \in A_i(b)} Q_i(b, a') - Q_i(b, a) \leq \eta_i \} \]
Appendix

\[
V_{saw, \alpha}^t = \gamma \sum_{s' \in S} O(a, s', \omega) T(s, a, s') \alpha(s')
\]

\[
\Gamma_{a, \omega}^t = \{[V_{s_1a, \omega, \alpha}^t, \ldots, V_{s_na, \omega, \alpha}^t]^T, \forall \alpha \in \Gamma_{\cdot \omega}^{t-1}\}
\]

\[
\Gamma_{b}^t = \{[R(s_1, a), \ldots, R(s_n, a)]^T + \sum_{\omega \in \Omega} \arg\max_{\alpha \in \Gamma_{a, \omega}^t} \alpha \cdot b, \forall a \in A\}
\]

\[
\Gamma^t = \{\arg\max_{\alpha \in \Gamma_{b}^t} \alpha \cdot b, \forall b \in B\}
\]