時間整合的マルコフ決定過程

恐神 貴行

IBM東京基礎研究所

森村哲郎, Mikael Onsjoe（IBM東京基礎研究所）との共同研究
Today’s route selection is limited

Shortest path algorithm finds an optimal path

For car navigation

For traffic simulation

Prescriptive analytics for drivers

Descriptive model of drivers

cf. sophisticated drivers select routes dynamically depending on latest conditions
Outline

- Limitations of traditional route selection
- Difficulties in selecting objective functions
- Time-consistent Markov decision processes
- Effectiveness of time-consistent Markov decision processes
Consider an example with three paths

<table>
<thead>
<tr>
<th>Path</th>
<th>Expected time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. W – C – mountain road – E</td>
<td>2h</td>
</tr>
<tr>
<td>2. W – C – popular highway – E</td>
<td>2h</td>
</tr>
<tr>
<td>3. W – E</td>
<td>1.7h</td>
</tr>
</tbody>
</table>

Path diagram:

- **Path 1**: W → C → mountain road → E; 1.5h before 8am; 2.5h after 8am
- **Path 2**: W → C → popular highway → E; 0.5h before 8am; 2.5h after 8am
- **Path 3**: W → E; 1.7h

7:30am
No path is better than a dynamic strategy

Expected time with the dynamic strategy is
$0.5h + 0.5 \times 0.5h + 0.5 \times 1.5h = 1.5h$
Expectation is obviously limited in representing drivers’ preference under risk

- Two routes have same expected travel time
- Some drivers prefer Route A
- Others prefer Route B
We study models for selecting dynamic strategies

• Interpretation of the dynamic strategy with a model of path selection is convoluted

To select the best dynamic strategy, the driver must be “unlucky” to select a suboptimal path

• Want to select optimal dynamic strategies with respect to a broad class of objective functions
  – personalized recommendation of dynamic strategies
  – realistic traffic simulation
Outline

- Limitations of traditional route selection
- Difficulties in selecting objective functions
- Time-consistent Markov decision processes
- Effectiveness of time-consistent Markov decision processes
Expected exponential utility is the standard objective of risk-sensitive Markov decision processes

\[
\min E[\exp(\gamma X)] \iff \min \text{ERM}_\gamma[X] \equiv \frac{1}{\gamma} \ln E[\exp(\gamma X)]
\]

(wworks only for \( \gamma > 0 \))

- \( \gamma > 0 \) \( \Rightarrow \) risk-averse
- \( \gamma < 0 \) \( \Rightarrow \) risk-seeking

Minimization of *expected exponential utility* is essentially equivalent to minimization of *entropic risk measure*
Which route would you take?

<table>
<thead>
<tr>
<th>Probability</th>
<th>Route P</th>
<th>Route Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9 (normal)</td>
<td>10 min.</td>
<td>Unif[0,20] min.</td>
</tr>
<tr>
<td>0.1 (busy)</td>
<td>Unif[20,80] min.</td>
<td>50 min.</td>
</tr>
</tbody>
</table>

![Diagram showing travel times and probabilities for Route P and Route Q. Route P has a higher probability and shorter travel time in normal conditions.](image)
Route P is never optimal with respect to any entropic risk measure

- Some drivers choose Route P
- Others choose Route Q
- They are all rational (e.g., $E[P] = E[Q]$)

\[ ERM_\gamma[T_P] \leq ERM_\gamma[T_Q], \forall \gamma \]
Expected utility is the standard objective function for decision making under risk

• Choose a dynamic strategy such that

\[ E[u(T)] \]

is minimized
- \( T \): travel time
- \( u \): utility function
cf. Expected utility theory (von Neumann & Morgenstern 1944)

• Entropic risk measure is a particular expected utility
  - \( u(x) = \exp(\gamma x) \)
For every path, does there exist a utility such that the path is optimal with respect to the expected utility?

<table>
<thead>
<tr>
<th></th>
<th>time</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_L$</td>
<td>20</td>
<td>1.0</td>
</tr>
<tr>
<td>$T_M$</td>
<td>10</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.2</td>
</tr>
<tr>
<td>$T_H$</td>
<td>10</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Expected utility is limited in representing driver’s preference

For any utility function, \( u \):
\[
E[u(T_M)] = 0.5 E[u(T_L)] + 0.5 E[u(T_H)]
\]

\[\Rightarrow\] We can have only

\[
E[u(T_L)] \leq E[u(T_M)] \leq E[u(T_H)]
\]
or

\[
E[u(T_H)] \leq E[u(T_M)] \leq E[u(T_L)]
\]

Never choose M with expected utility
Conditional tail expectation is a popular risk measure in finance

- Choose a dynamic strategy such that

\[
\text{CTE}_\alpha[T] \equiv \frac{(1 - \beta)E[T \mid T > Q_\alpha] + (\beta - \alpha)Q_\alpha}{1 - \alpha}
\]

is minimized
  - T: travel time

\[
Q_\alpha \equiv \min\{t \mid \Pr(T \leq t) \geq \alpha\}
\]

\[
\beta \equiv \Pr(T \leq V_\alpha)
\]

- When T is continuous,

\[
\text{CTE}_\alpha[T] \equiv E[T \mid T > Q_\alpha]
\]
Choose either the road to B or that to B’ when we leave A to reach C

B-C is normal iff B’-C is normal
A-B’-C has smaller risk than A-B-C with respect to $\text{CTE}_{0.99}$

<table>
<thead>
<tr>
<th></th>
<th>$\text{CTE}_{0.99}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B-C</td>
<td>130 min.</td>
</tr>
<tr>
<td>A-B’-C</td>
<td>[210 \times 0.001 + 110 \times 0.009 \div 0.01 = 120 \text{ min.}]</td>
</tr>
</tbody>
</table>
If traffic conditions are normal at B’, B’-B-C appears to have smaller risk than B’-C with respect to $\text{CTE}_{0.99}$.

<table>
<thead>
<tr>
<th></th>
<th>$\text{CTE}_{0.99}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B’-C</td>
<td>110</td>
</tr>
<tr>
<td>A-B’-B-C</td>
<td>100</td>
</tr>
</tbody>
</table>

B-C is normal iff B’-C is normal.
If traffic conditions are busy at B’, B’-B-C appears to have smaller risk than B’-C with respect to $\text{CTE}_{0.99}$.

<table>
<thead>
<tr>
<th></th>
<th>$\text{CTE}_{0.99}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B’-C</td>
<td>210</td>
</tr>
<tr>
<td>A-B’-B-C</td>
<td>160</td>
</tr>
</tbody>
</table>

B-C is normal iff B’-C is normal.
Following “optimal” directions, we end up in taking a poor route surely

B-C is normal iff B’-C is normal
Outline

• Limitations of traditional route selection

• Difficulties in selecting objective functions

• Time-consistent Markov decision processes

• Effectiveness of time-consistent Markov decision processes
We define a time-consistent MDP as the MDP whose objective is to minimize an iterated risk measure.

X: travel time with Route A-B'-C

Evaluate riskiness of X by evaluating riskiness of \( \text{CTE}_{0.8}[X] \) to be evaluated tomorrow (known as iterated CTE).

\[
\text{CTE}_{0.8}[\text{CTE}_{0.8}[X]] = 88.75
\]

Want to evaluate riskiness of X

Will evaluate X with \( \text{CTE}_{0.8}[X] \)

Normal at B'

Busy at B'

Evaluate riskiness of X with \( \text{CTE}_{0.8}[X] \)

\[
\text{CTE}_{0.8}[X] = 115
\]

(0.1)

(0.99)

(0.01)

X=210

X=110

X=110

Evaluate riskiness of X with \( \text{CTE}_{0.8}[X] \)

\[
\text{CTE}_{0.8}[X] = 62.5
\]

(0.1)

(0.99)

(0.01)

X=60
Formally, an iterated risk measure is a dynamic risk measure having a recursive structure

- \((\Omega, \mathcal{F}, \mathbb{P})\): Filtered probability space
  - \(F_0 \subseteq F_1 \subseteq \ldots \subseteq F_N = \mathcal{F}\)
- \(Y\): \(\mathcal{F}\)-measurable random variable

We say that \(\rho\) is an iterated risk measure if
- \(\rho_N[Y] = Y\)
- \(\rho_n[Y] = r_n[\rho_{n+1}[Y]]\)
  - \(r_n\): conditional risk measure mapping \(F_{n+1}\)-measurable random variable to \(F_n\)-measurable random variable
Recursive definition implies dynamic programming finds the optimal policy for time-consistent MDP.

For each action \(a\), calculate \(r^a[V]\)

\[
v = \min_a r[V_a]
\]

\(a^* = \arg\min_a r[V_a]\)

\(r\): risk measure

\(V_a\): (random) value at next state given \(a\)

### Example

- \(s\):
  - \(p_a = 0.9\)
  - \(v = 3.6\), \(a^* = \text{“go straight”}\)

- \(s_1\):
  - \(p_a = 0.06\)
  - \(v = \) unassigned

- \(s_2\):
  - \(p_a = 0.4\)
  - \(v = 1.4\), \(a^* = \text{“turn right”}\)

- \(s_3\):
  - \(p_a = 0.04\)
  - \(v = 0.8\), \(a^* = \text{“go straight”}\)

Value of iterated risk measure \(r[r[r[X]...]]\) given that optimal actions are taken from \(s_3\).
More precisely, risk measures must be monotonic

<table>
<thead>
<tr>
<th>Properties of a risk measure, $r$</th>
<th>Optimal policy for MDPs with respect to the corresponding iterated risk measure</th>
</tr>
</thead>
</table>
| Monotonic: $X \leq Y \Rightarrow r(X) \leq r(Y)$ | Can be found with dynamic programming  
Needed augmented states  
state := (state, accumulated cost) |
| Monotonic & Translation invariant:  
$r(X + c) \leq r(X) + c$ | No need for augmented states  
Cannot discount future cost |
| Monotonic & Translation invariant & Positive homogeneity:  
$r(ax) \leq ar(X)$ | No need for augmented states  
Can discount future cost |
Dynamic programming with monotonic and translation-invariant iterated risk measures

- Markov decision process
  - $S_n$: State at time $n$ (random variable, $F_n$-measurable)
  - $A_n$: Action at time $n$ (random variable, $F_n$-measurable)
  - $C_n$: Cost between time $n$ and time $n+1$, depending on $S_n$, $A_n$, $S_{n+1}$ (random variable, $F_{n+1}$-measurable)
  - $S_n$: State space at time $n$ (set)
  - $A(s)$: Action space from state $s$ (set)
  - $\Pi$: Set of candidate policies (set)

- Find $\pi$ that minimizes $\rho_n \left[ \sum_{\ell=0}^{N-1} C_{\ell} \mid S_n = s, \pi \right]$ or equivalently $\rho_n \left[ \sum_{\ell=n}^{N-1} C_{\ell} \mid S_n = s, \pi \right]$ for every $s \in S_n$, $n=0,...,N-1$

$$V_n^*(s) \equiv \min_{\pi \in \Pi} \rho_n \left[ \sum_{\ell=n}^{N-1} C_{\ell} \mid S_n = s, \pi \right]$$

New

$$V_N^*(s) = 0$$
$$V_n^*(s) = \min_{a \in A(s)} r_n \left[ C_n + V_{n+1}^*(S_{n+1}) \mid S_n = s, A_n = a \right] \quad \forall s \in S_n$$
Outline

• Limitations of traditional route selection

• Difficulties in selecting objective functions

• Time-consistent Markov decision processes

• Effectiveness of time-consistent Markov decision processes
Iterated risk measures overcome limitations of expected utility and other risk measures.
A driver takes only extreme routes if his decisions follow expected utility.

For any utility function, Route 0 (equivalent to doing nothing) or Route 10 (riskiest) is most preferable.

<table>
<thead>
<tr>
<th>Route</th>
<th>0</th>
<th>1</th>
<th>...</th>
<th>i</th>
<th>...</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net gain</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>+1</td>
<td>...</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>Probability</td>
<td>1</td>
<td>0.04</td>
<td>0.9</td>
<td>0.06</td>
<td>...</td>
<td>0.04i</td>
<td>0.9 - 0.1i</td>
</tr>
</tbody>
</table>

For any utility function, Route 0 (equivalent to doing nothing) or Route 10 (riskiest) is most preferable.
Iterated risk measures can represent the preference that cannot be represented with any expected utility.

\[ \rho_t[X] \equiv r_t[\rho_{t+1}[X]] \quad \text{s.t.} \quad r_t[Y] \equiv E[Y] - 20(\Pr(Y \leq 0) - \alpha) I\{\Pr(Y \leq 0) \geq \alpha\} \]

\[ \alpha = 0 \]

\[ \alpha = 1 \]

X = capital after 8 travels (no travel after ruined)

initial capital = 2

Ruin probability cannot be in (0, 0.16) for any expected utility.
Takeaways

- “Shortest path” is limited
  - Route selection follows a dynamic strategy
- Expected utility can only represent limited preferences for
  - Personalized recommendation of dynamic strategies
  - Realistic traffic simulation
- Traditional risk measures lead to inconsistent decisions
  - Inconsistent decision maker can surely lose infinite capital against rational decision maker
- Time-consistent MDP is defined with iterated risk measures
  - Can represent broad preferences with consistent decisions
  - Optimal policy found with dynamic programming
時間整合的マルコフ決定過程

恐神 貴行

IBM東京基礎研究所

森村哲郎, Mikael Onsjoe （IBM東京基礎研究所）との共同研究

References:
T. Osogami, "Iterated risk measures for risk-sensitive Markov decision processes with discounted cost", in Proceedings of UAI 2011.
T. Osogami and T. Morimura, Time-consistency of optimization problems, manuscript, 2011.
T. Osogami and M. Onsjoe, Overcoming a limitation of expected utility, manuscript 2011.