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

Prescriptive analytics for drivers

For traffic simulation

Descriptive model of drivers

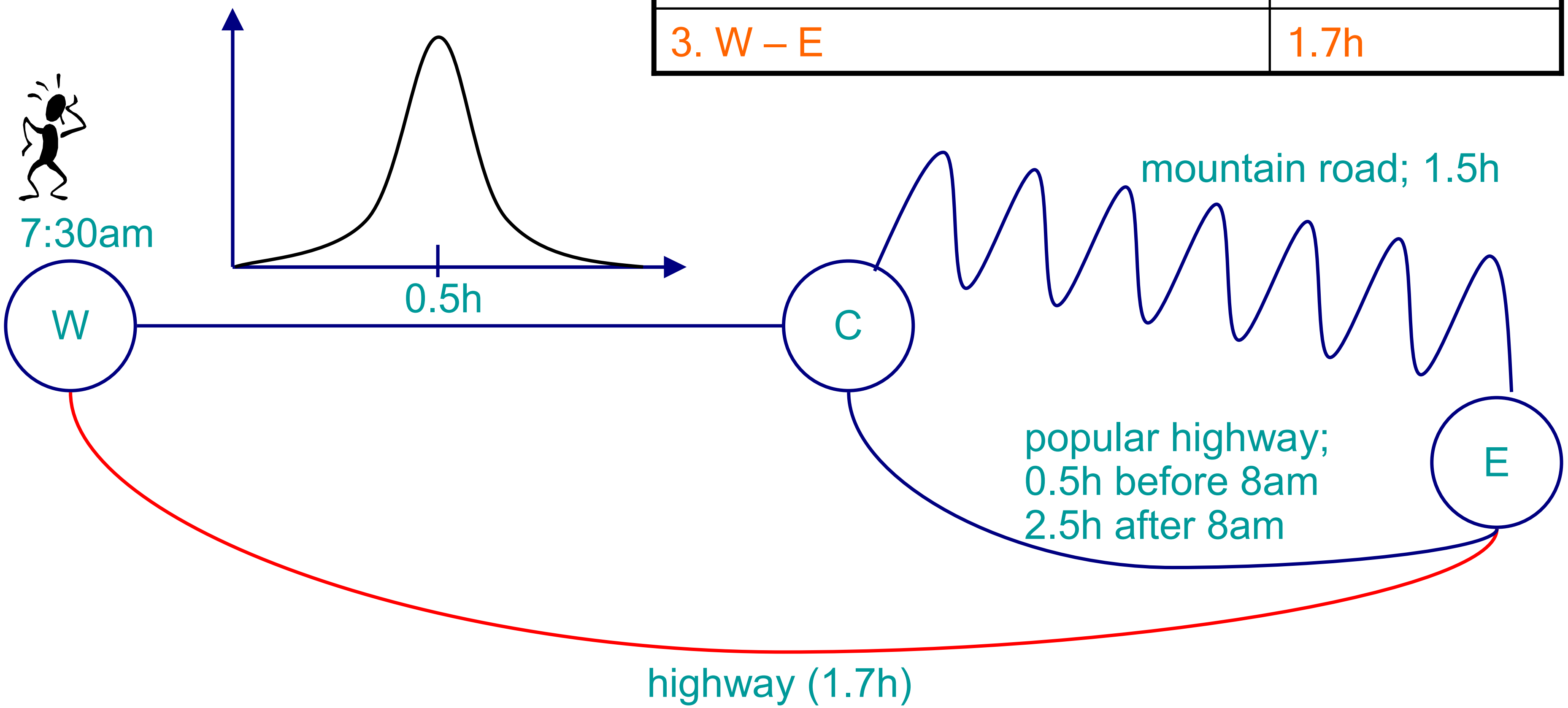
cf. sophisticated drivers select routes dynamically depending on latest conditions

Outline

- Limitations of traditional route selection
- Difficulties in selecting objective functions
- Time-consistent Markov decision processes
- Effectiveness of time-consistent Markov decision processes

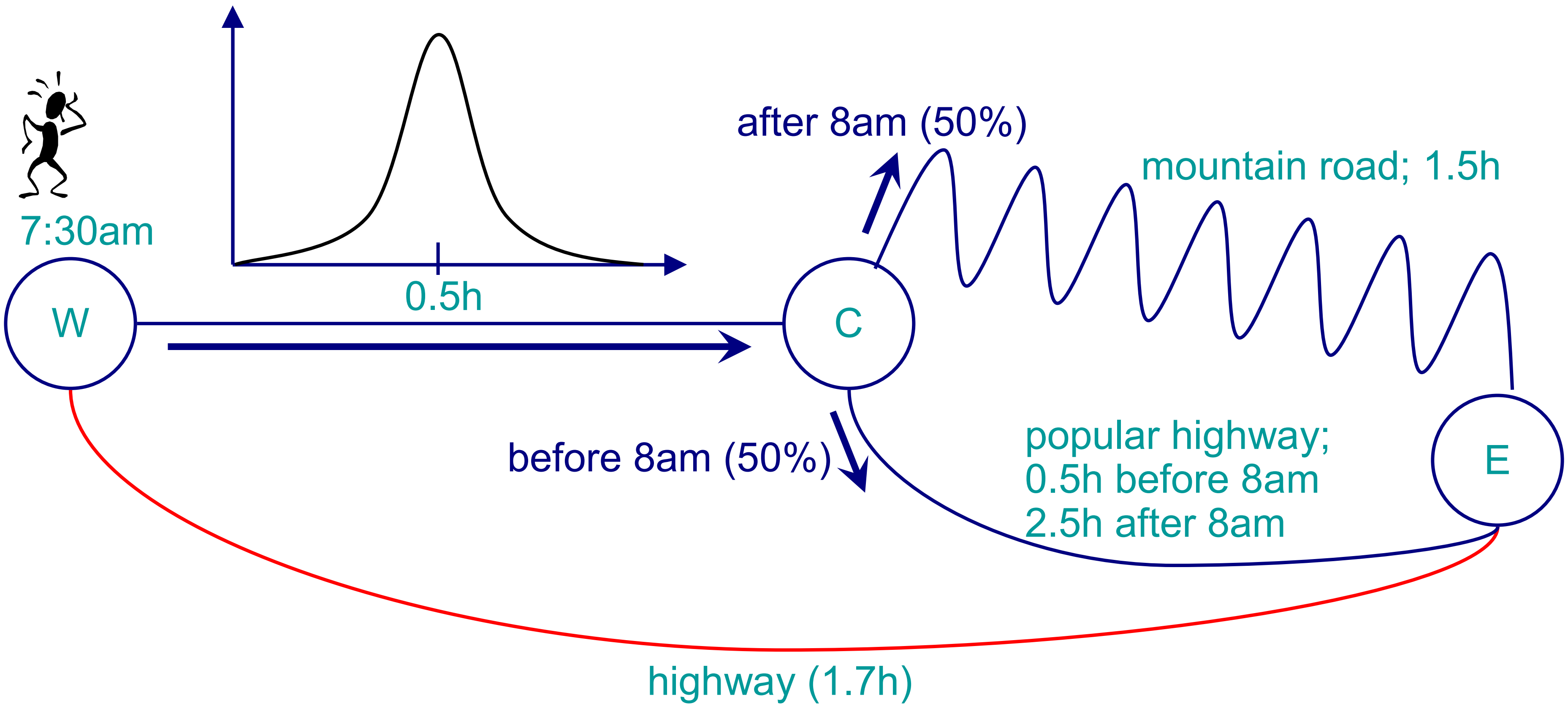
Consider an example with three paths

Path	Expected time
1. W – C – mountain road – E	2h
2. W – C – popular highway – E	2h
3. W – E	1.7h



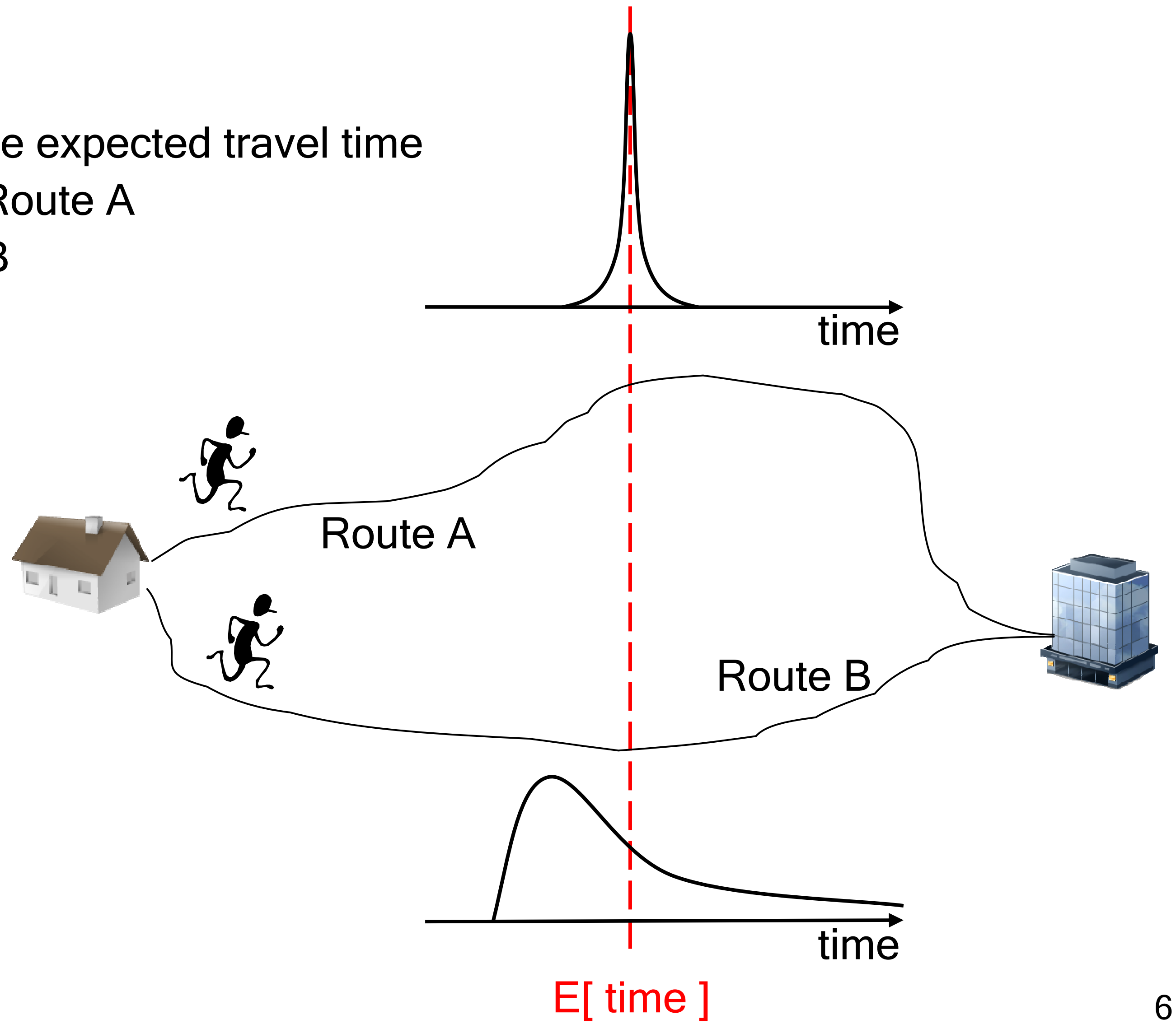
No path is better than a dynamic strategy

Expected time with the dynamic strategy is
0.5h + 0.5 x 0.5h + 0.5 x 1.5h = 1.5h



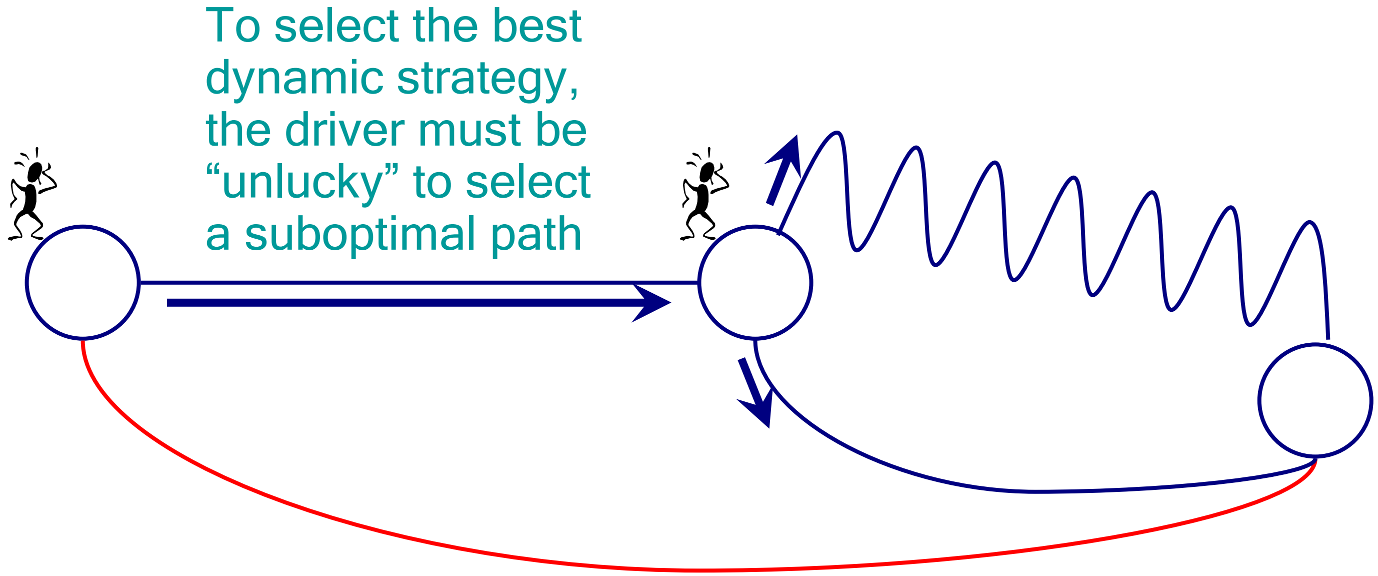
Expectation is obviously limited in representing drivers' preference under risk

- 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

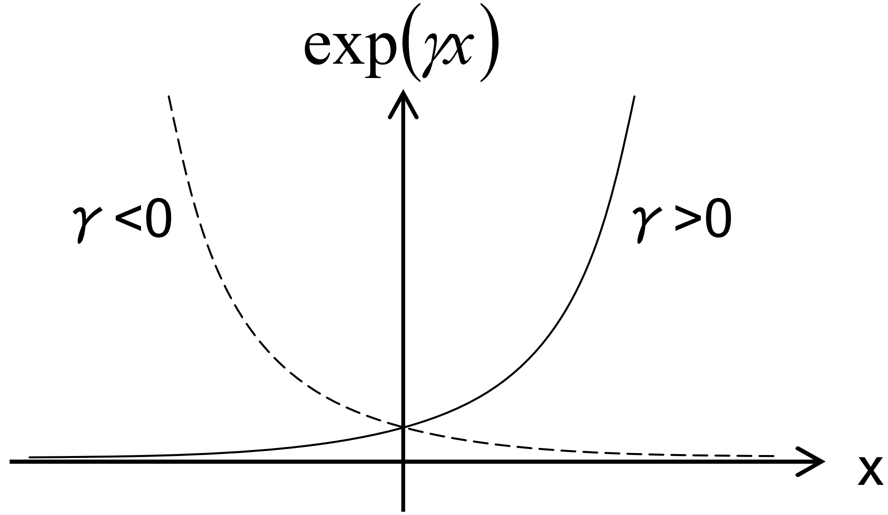


- Want to select optimal dynamic strategies with respect to a broad class of objective functions
  - personalized recommendation of dynamic strategies
  - realistic traffic simulation

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



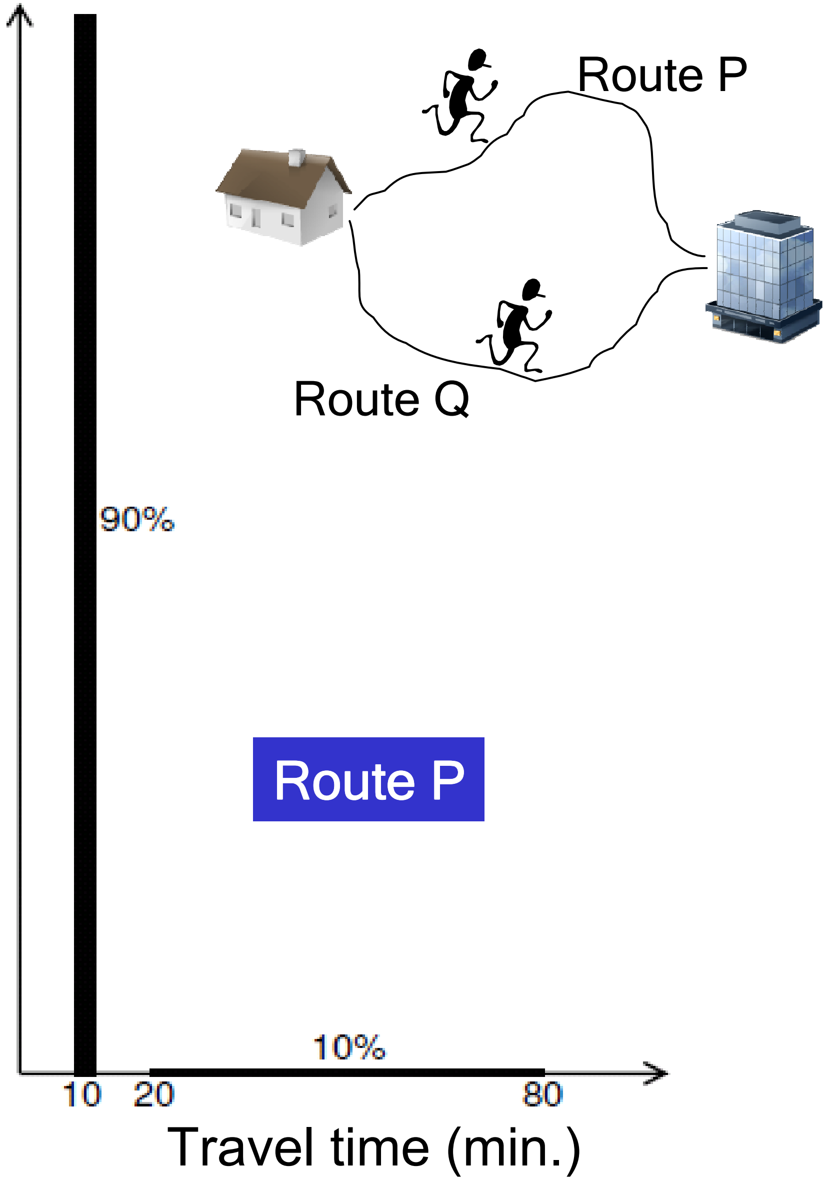
$$\min.\ \mathrm{E}[\exp(\gamma X)] \quad \Leftrightarrow \quad \min.\ \mathrm{ERM}_{\gamma}[X] \equiv \frac{1}{\gamma} \ln \mathrm{E}[\exp(\gamma X)]$$

(works only for γ >0)

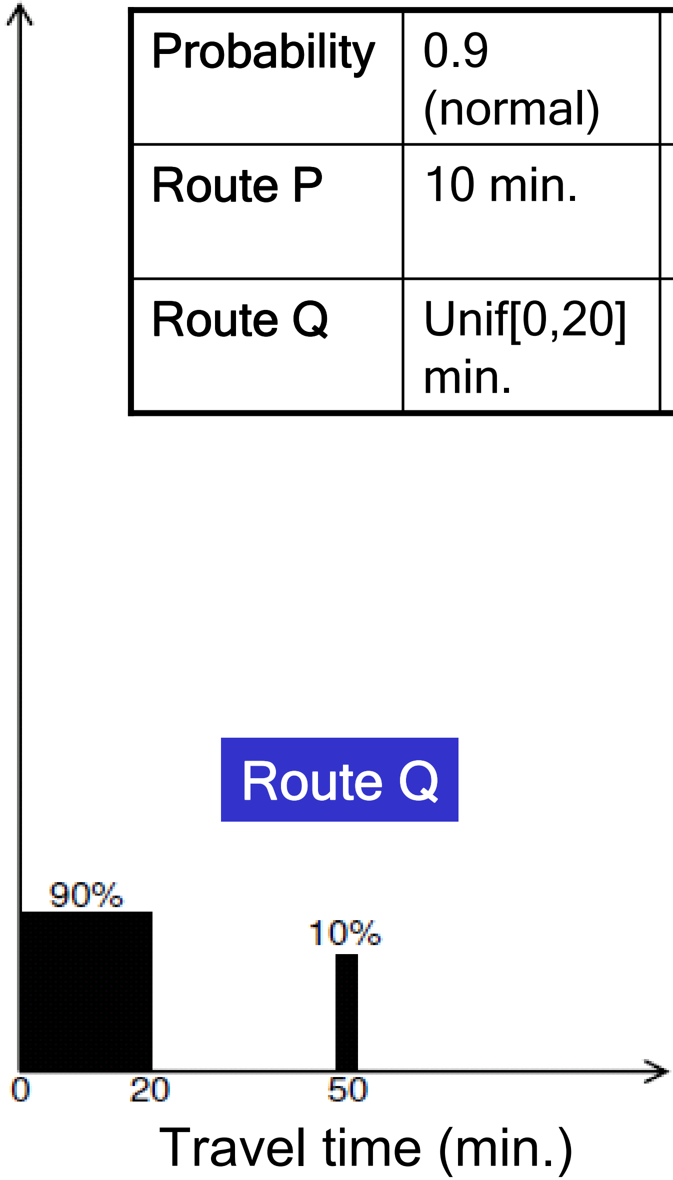
- γ >0 ⇒ risk-averse
- γ <0 ⇒ risk-seeking

Minimization of *expected exponential utility* is essentially equivalent to minimization of *entropic risk measure*

Which route would you take?

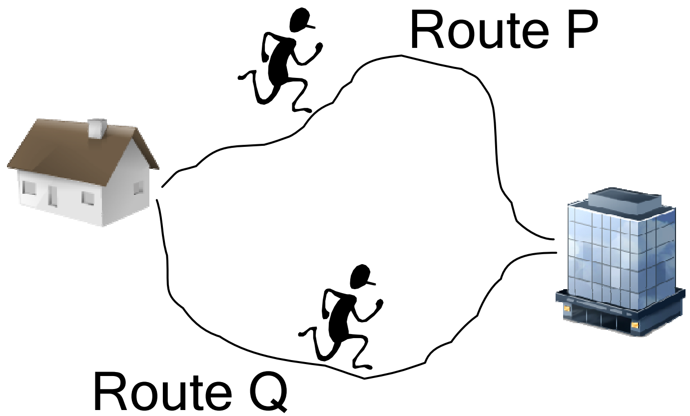


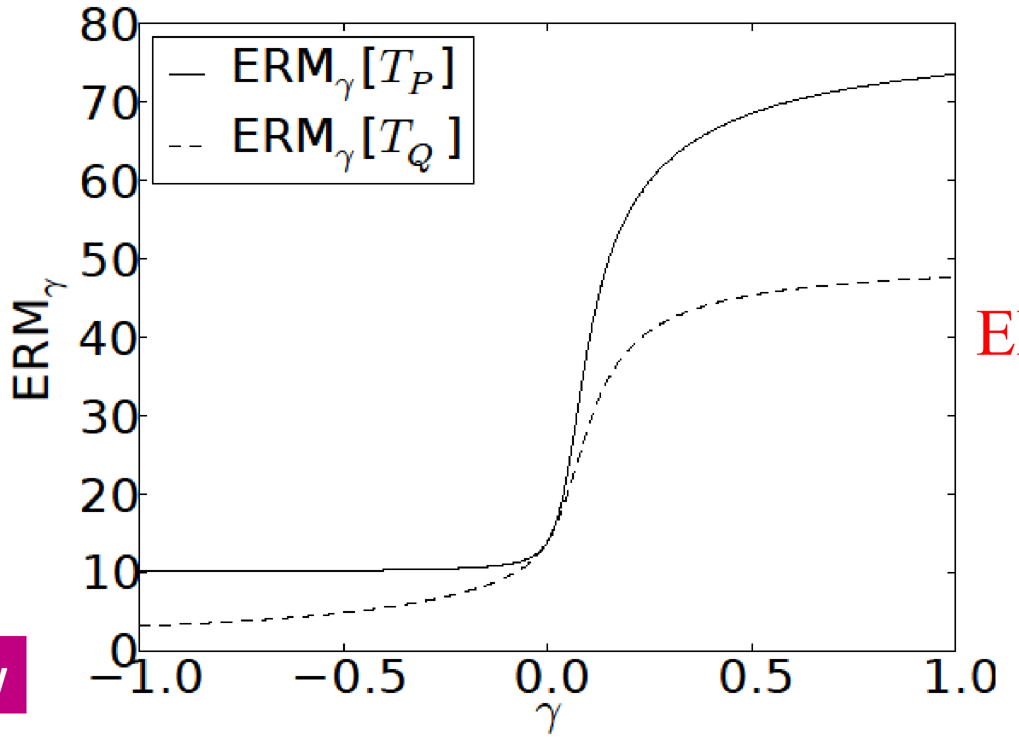
Probability	0.9 (normal)	0.1 (busy)
Route P	10 min.	Unif[20,80] min.
Route Q	Unif[0,20] min.	50 min.



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])





$$\mathrm{ERM}_\gamma[T_Q] \le \mathrm{ERM}_\gamma[T_P], \forall \gamma$$

New

Expected utility is the standard objective function for decision making under risk

- Choose a dynamic strategy such that

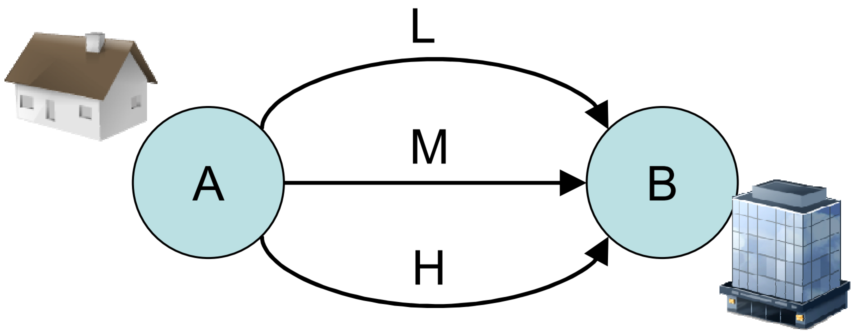
$$\mathrm{E}[u(T)]$$

  is minimized
  - T: travel time
  - u: utility function

cf. Expected utility theory
(von Neumann & Morgenstern 1944)

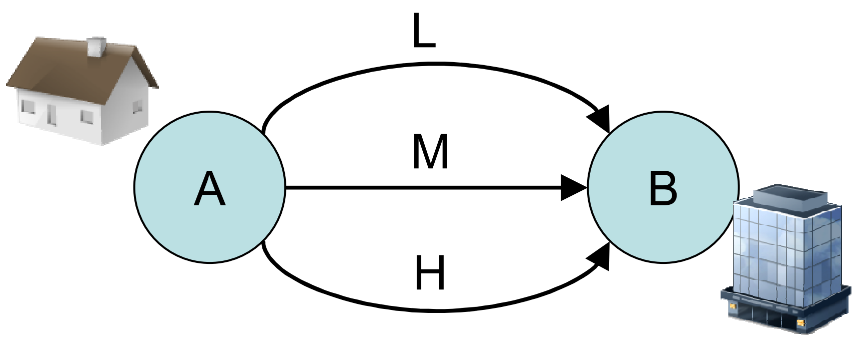
- Entropic risk measure is a particular expected utility
  - u(x) = exp(γ x)

For every path, does there exist a utility such that the path is optimal with respect to the expected utility?



	time	probability
T_L	20	1.0
T_M	10	0.3
	20	0.5
	30	0.2
T_H	10	0.6
	30	0.4

Expected utility is limited in representing driver's preference



	time	probability
T_L	20	1.0
T_M	10	0.3
	20	0.5
	30	0.2
T_H	10	0.6
	30	0.4

For any utility function, u:

$$\mathrm{E}[u(T_M)] = 0.5\,\mathrm{E}[u(T_L)] + 0.5\,\mathrm{E}[u(T_H)]$$

$\Rightarrow$ We can have only

$$\mathrm{E}[u(T_L)] \le \mathrm{E}[u(T_M)] \le \mathrm{E}[u(T_H)]$$

or

$$\mathrm{E}[u(T_H)] \le \mathrm{E}[u(T_M)] \le \mathrm{E}[u(T_L)]$$

Never choose M with expected utility

Conditional tail expectation is a popular risk measure in finance

- Choose a dynamic strategy such that

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

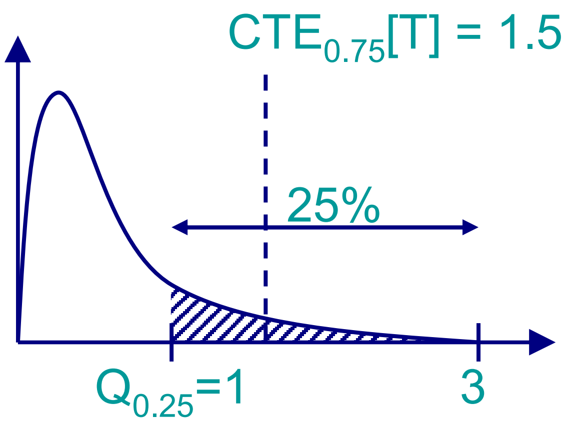
  is minimized
  - T: travel time

$$Q_\alpha \equiv \min\{t \mid \Pr(T \le t) \ge \alpha\}$$

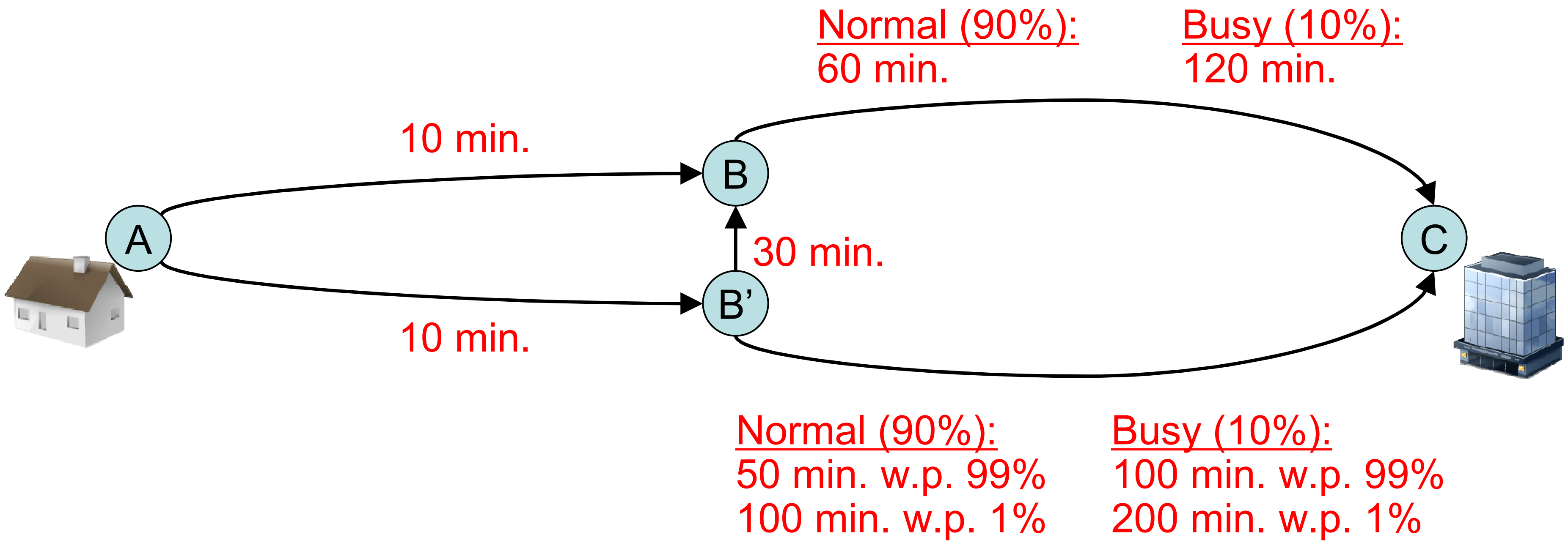
$$\beta \equiv \Pr(T \le V_\alpha)$$

- When T is continuous,

$$\mathrm{CTE}_\alpha[T] \equiv \mathrm{E}[T \mid T > Q_\alpha]$$

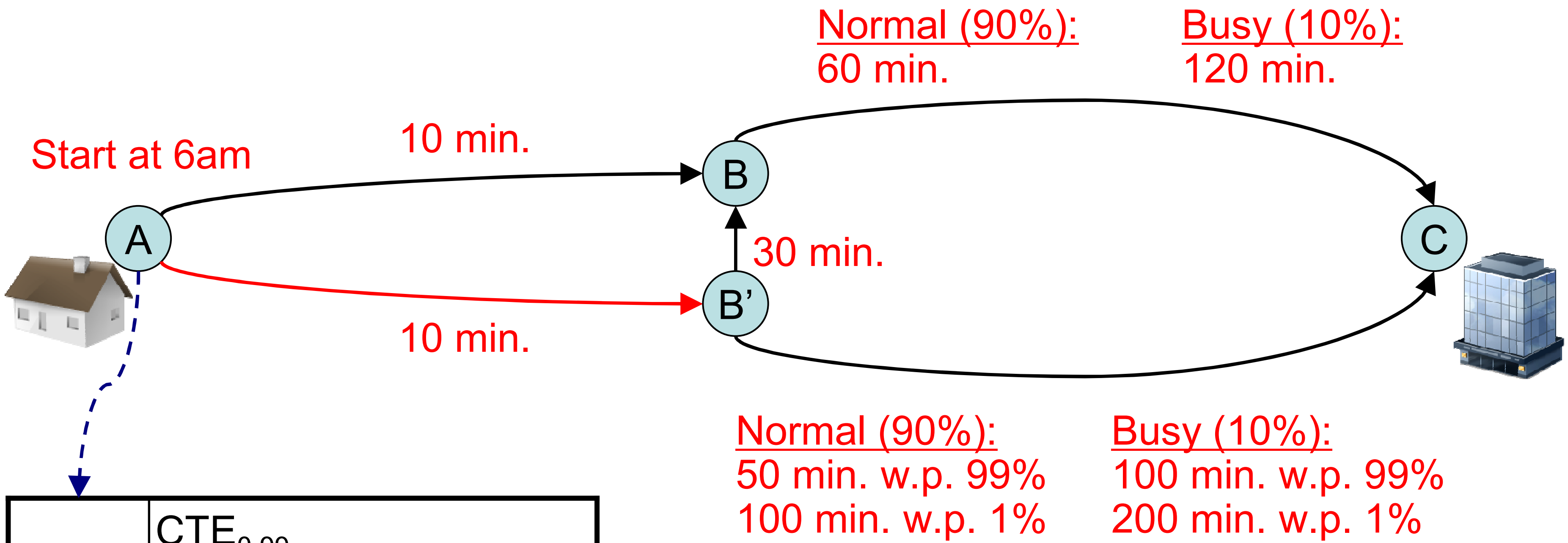


Choose either the road to B or that to B' when we leave A to reach C



B-C is normal iff B'-C is normal

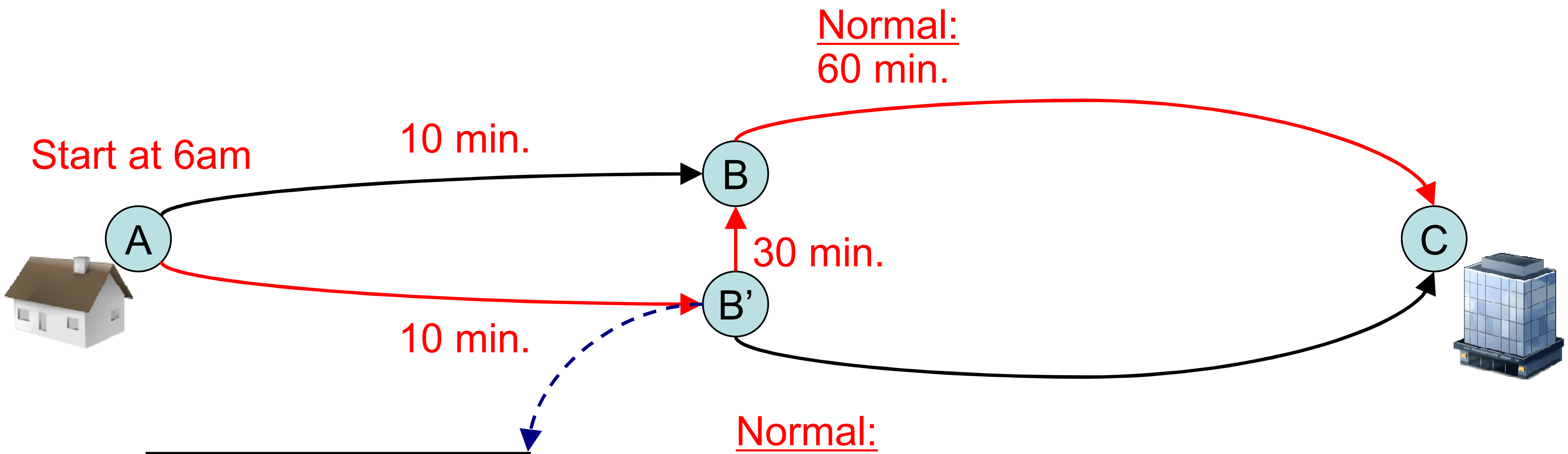
A-B'-C has smaller risk than A-B-C with respect to $CTE_{0.99}$



	$CTE_{0.99}$
A-B-C	130 min.
A-B'-C	$\frac{210 \times 0.001 + 110 \times 0.009}{0.01}$ = 120 min.

B-C is normal iff B'-C is normal

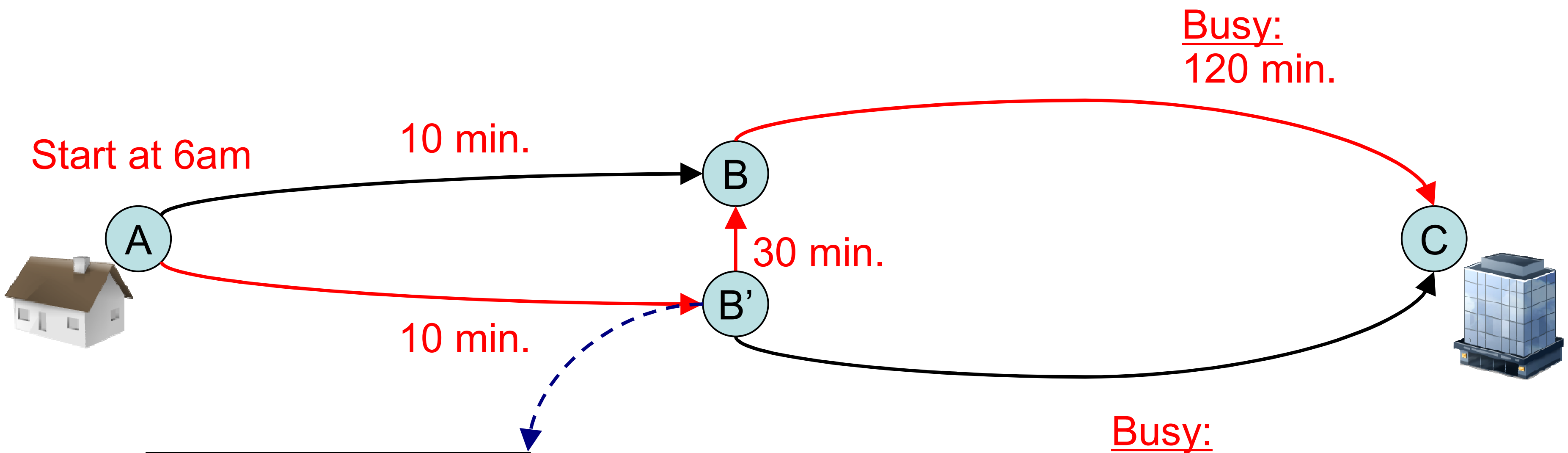
If traffic conditions are normal at B', B'-B-C appears to have smaller risk than B'-C with respect to $CTE_{0.99}$



	$CTE_{0.99}$
A-B'-C	110
A-B'-B-C	100

B-C is normal iff B'-C is normal

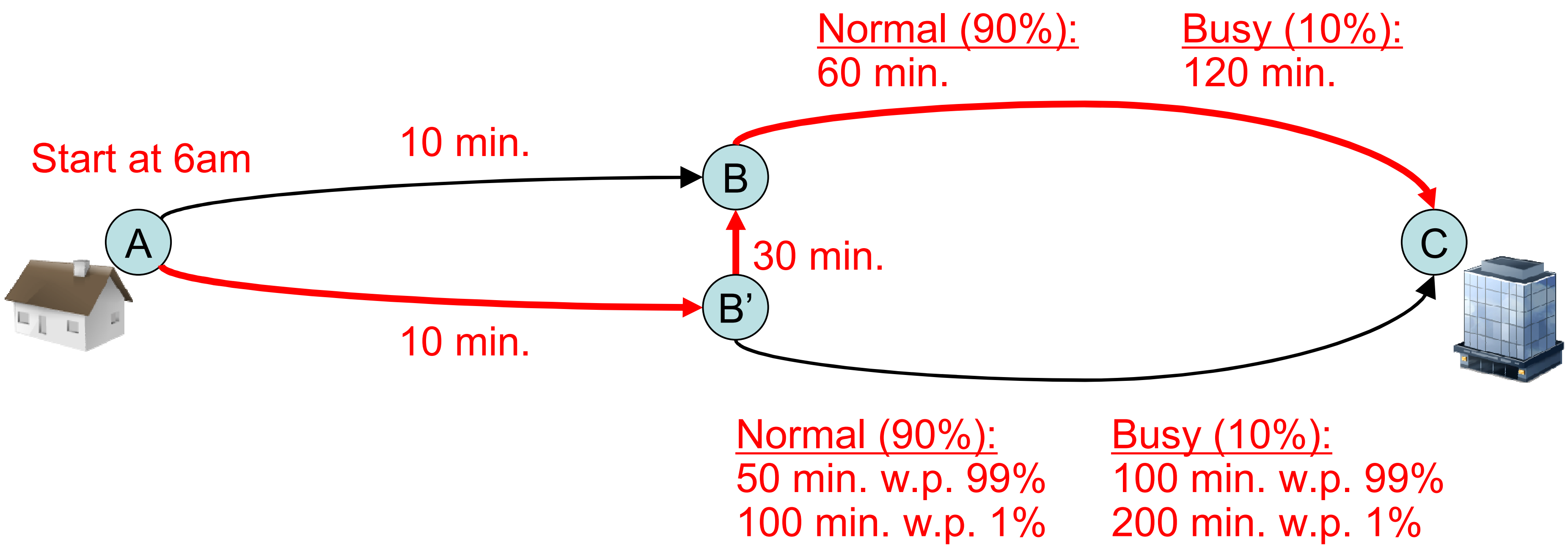
If traffic conditions are busy at B', B'-B-C appears to have smaller risk than B'-C with respect to $CTE_{0.99}$



	$CTE_{0.99}$
A-B'-C	210
A-B'-B-C	160

B-C is normal iff B'-C is normal

Following “optimal” directions, we end up in taking a poor route surely



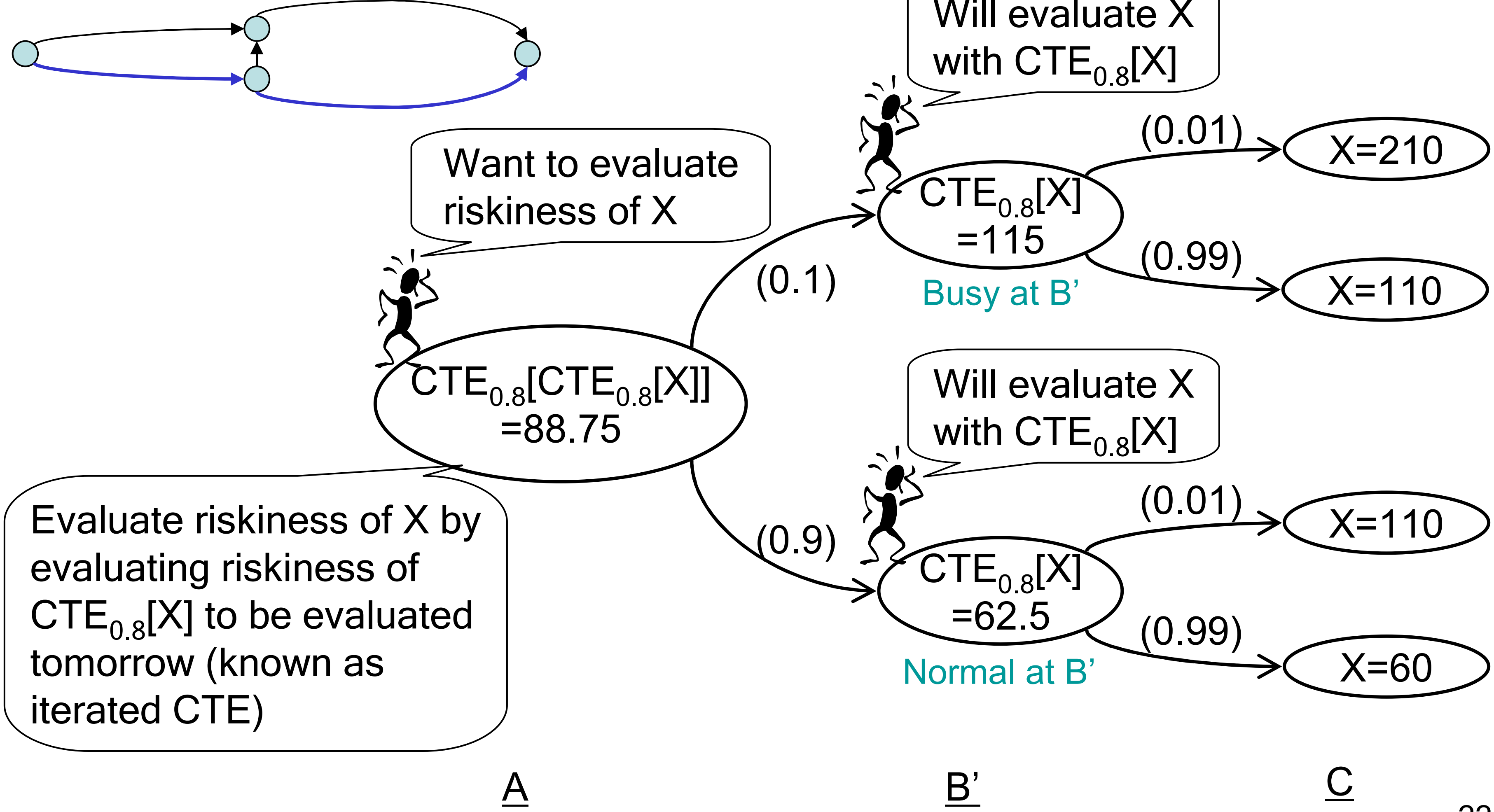
B-C is normal iff B’-C is normal

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



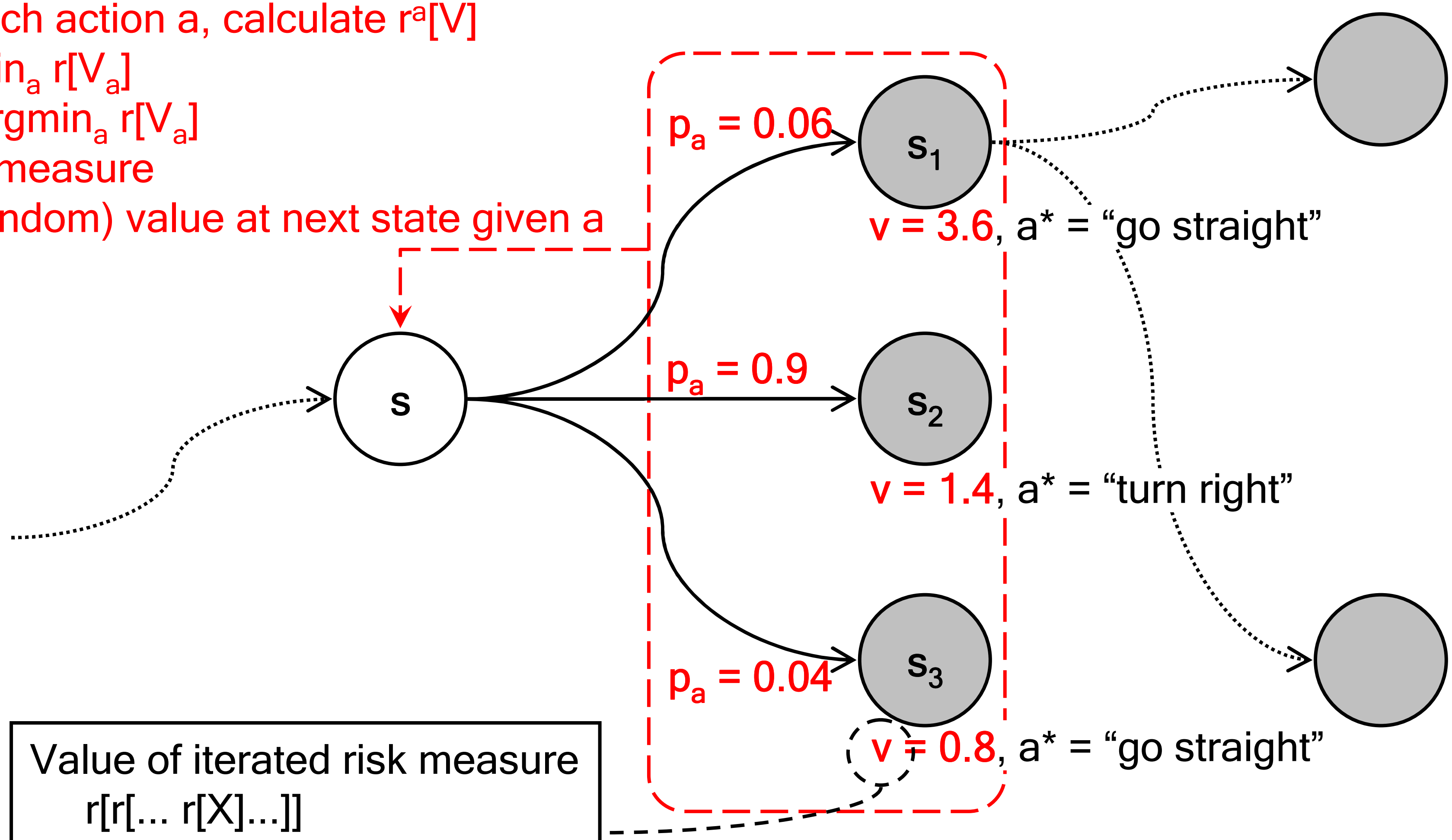
Formally, an iterated risk measure is a dynamic risk measure having a recursive structure

- (Ω, $\mathcal{F}$, P): Filtered probability space
  - $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_N = \mathcal{F}$
- Y: $\mathcal{F}$-measurable random variable

- We say that ρ 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 $\mathcal{F}_{n+1}$-measurable random variable to $\mathcal{F}_n$-measurable random variable

Recursive definition implies dynamic programming finds the optimal policy for time-consistent MDP

For each action a, calculate $r^a[V]$
$v = \min_a r[V_a]$
$a^* = \operatorname{argmin}_a r[V_a]$
r: risk measure
V_a: (random) value at next state given a



Value of iterated risk measure
r[r[... r[X]...]]
given that optimal actions are taken from s_3

New

More precisely, risk measures must be monotonic

Properties of a risk measure, r	Optimal policy for MDPs with respect to the corresponding iterated risk measure
Monotonic: $X \le Y \Rightarrow r(X) \le r(Y)$	Can be found with dynamic programming Need augmented states state := (state, accumulated cost)
Monotonic & Translation invariant: $r(X+c) \le r(X)+c$	No need for augmented states Cannot discount future cost
Monotonic & Translation invariant & Positive homogeneite: $r(aX) \le ar(X)$	No need for augmented states Can discount future cost

New

Dynamic programming with monotonic and translation-invariant iterated risk measures

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

- Find π 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 $s \in \mathbf{S}_n$, n=0,...,N-1

by translation-invariance
ρ (X+c)= ρ (X)+c

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

$$V_N^*(s) = 0 \qquad \forall s \in \mathbf{S}_N$$

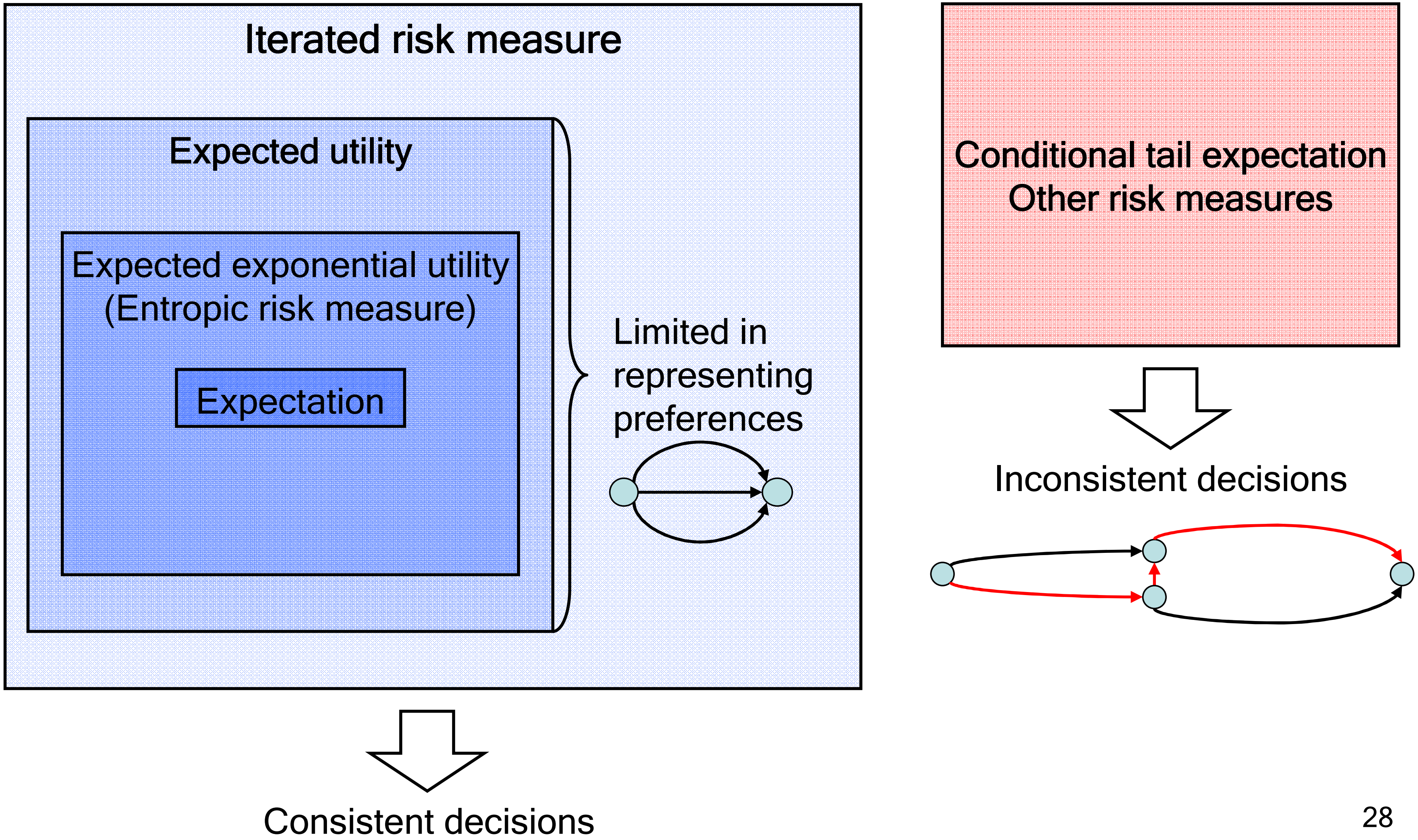
$$V_n^*(s) = \min_{a \in \mathbf{A}(s)} r_n\left[C_n + V_{n+1}^*(S_{n+1}) \middle| S_n = s, A_n = a\right] \qquad \forall s \in \mathbf{S}_n, n = 0,\ldots,N-1$$

New

Outline

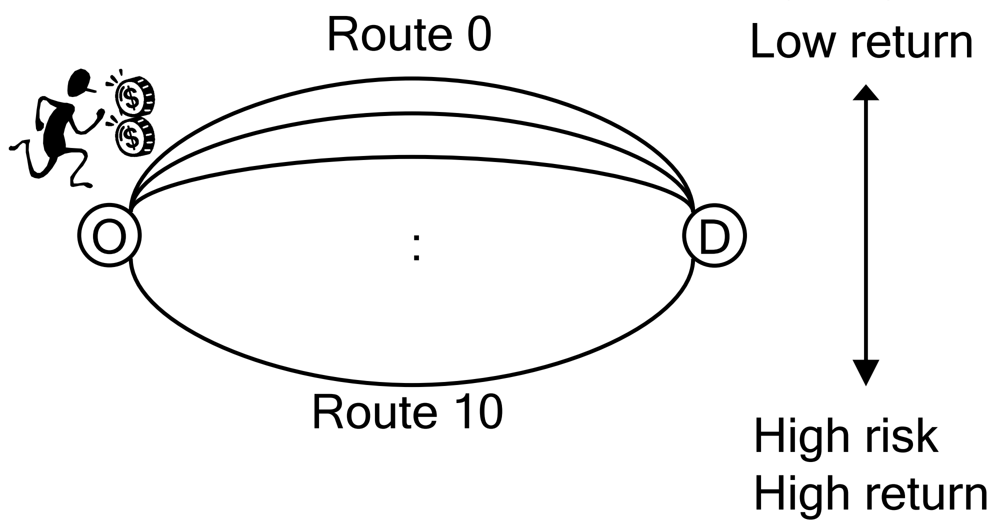
- Limitations of traditional route selection
- Difficulties in selecting objective functions
- Time-consistent Markov decision processes
- Effectiveness of time-consistent Markov decision processes

Iterated risk measures overcome limitations of expected utility and other risk measures



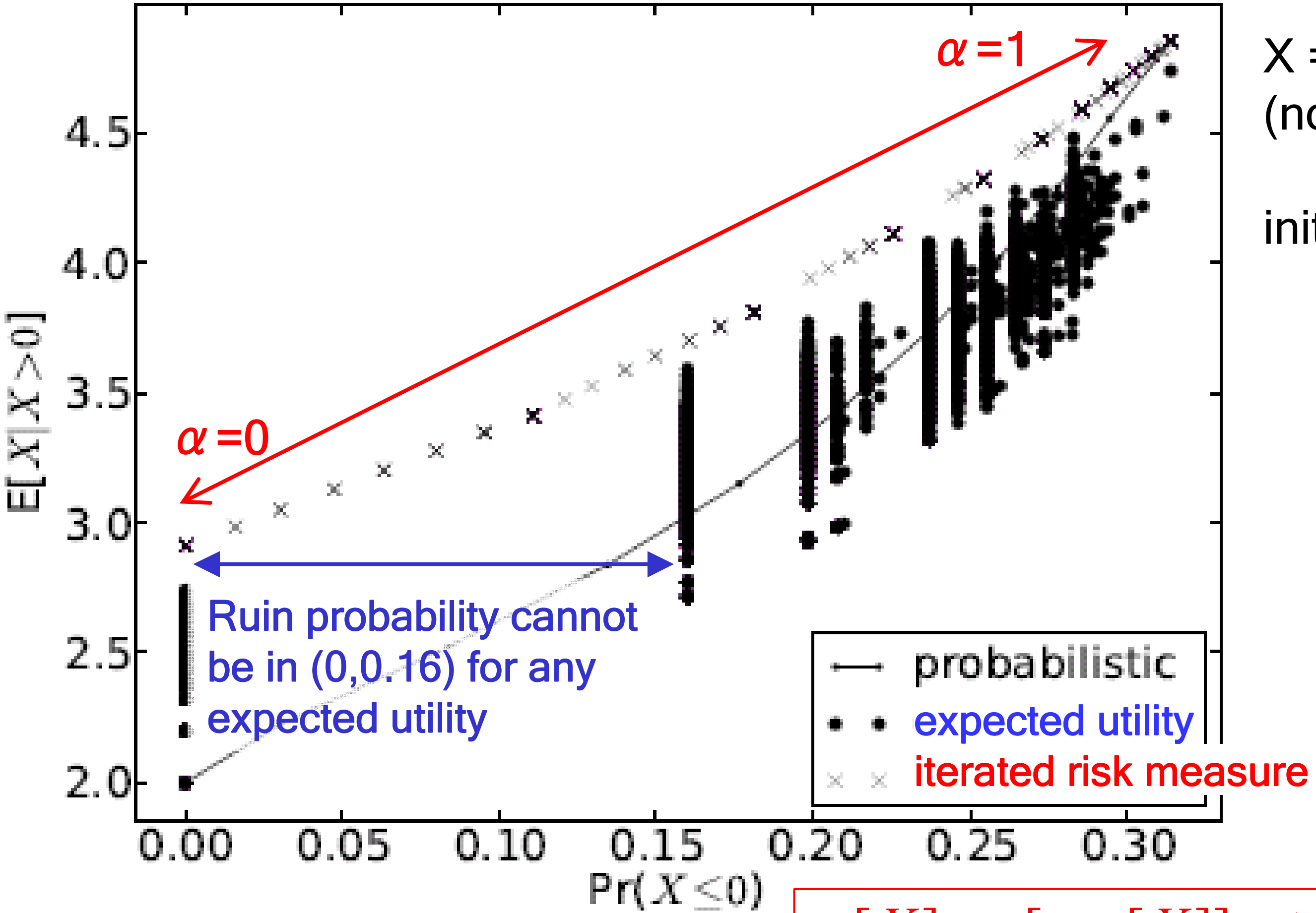
A driver takes only extreme routes if his decisions follow expected utility

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



Route	0	1			...	i			...	9			10	
Net gain	0	-1	0	+1	...	-1	0	+1	...	-1	0	+1	-1	+1
Probability	1	0.04	0.9	0.06	...	0.04i	1 - 0.1i	0.06i	...	0.36	0.1	0.56	0.4	0.6

Iterated risk measures can represent the preference that cannot be represented with any expected utility



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

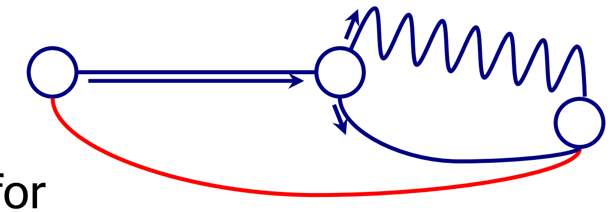
initial capital = 2

$$\rho_t[X] \equiv r_t[\rho_{t+1}[X]] \text{ s.t.}$$

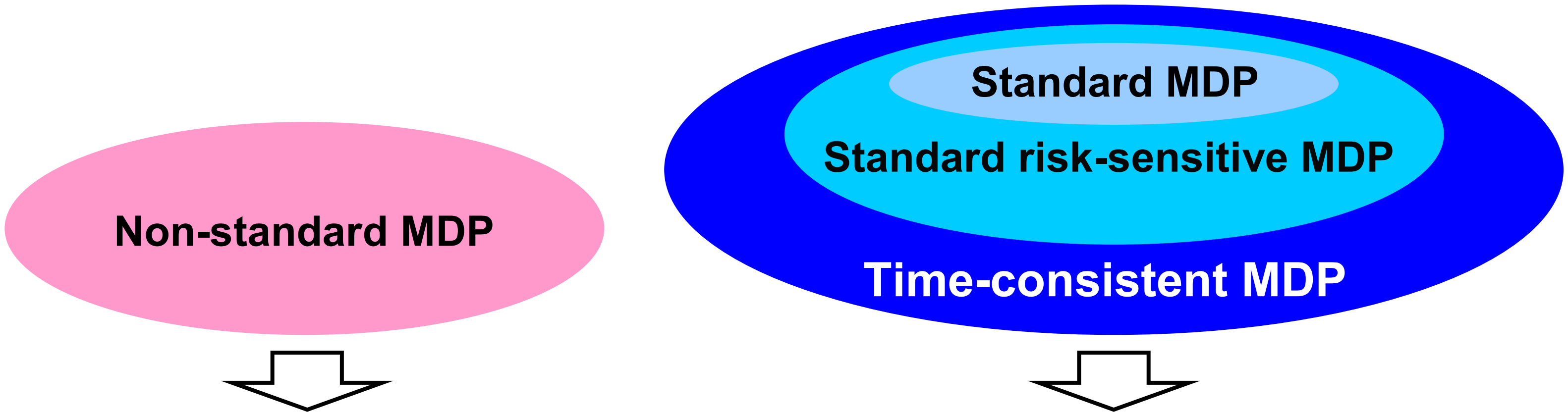
$$r_t[Y] \equiv \mathrm{E}[Y] - 20(\Pr(Y \le 0) - \alpha)\,\mathrm{I}\{\Pr(Y \le 0) \ge \alpha\}$$

New

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



Inconsistent decisions

Consistent decisions

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.