Continuous Time
Markov Decision Processes:
Theory, Applications and Computational Algorithms

Peter Buchholz,
Informatik IV, TU Dortmund, Germany
Overview 1

- Continuous Time Markov Decision Processes (CTMDPs)
  - Definition
  - Formalization
  - Applications
- Infinite Horizons
  - Result Measures
  - Optimal Policies
  - Computational Methods
Overview 2

- Finite horizons
  - Result Measures
  - Optimal Policies
  - Computational Methods

- Advanced Topics
  - Model Checking CTMDPs
  - Infinite State Spaces
  - Transition Rate Bounds
  - ....
Continuous Time Markov Chains (CTMCs) with Rewards

Components:
- States
- Initial Probabilities
- Rate Rewards
- Transitions
- Transition Rates
- Impulse Rewards
Behavior of the CTMC with rewards:

- Process stays an exponentially distributed time in a state (if it is not absorbing)
- Performs a transition into a successor state (according to the transition rates)
- Accumulates (rate) rewards per time unit in a state and impulse reward per transition

Result measures:

- Average reward accumulated per unit of time in the long run
- Discounted reward
- Accumulated reward during a finite interval
- Accumulated reward before entering a state or subset of states
Extension: Choose between different actions in a state

Decision between red or blue in the states 2 and 3!

Continuous Time Markov Decision Process
Markov Decision Process (with finite state and action spaces)

- State space $S = \{1, \ldots, n\}$ ($S = \mathbb{Z}_+$ in the countable case)
- Set of decisions $D_i = \{1, \ldots, m_i\}$ for $i \in S$
- Vector of transition rates $q_i^u \in \mathbb{R}^{1,n}$
  - where $q_i^u(j) < \infty$ is the transition rate from $i$ to $j$ ($i \neq j$, $i, j \in S$) under decision $u \in D_i$
  - $\Rightarrow$ exponential sojourn time in state $i$ with rate $-q_i^u(i) = \sum_{i \neq j} q_i^u(j)$
    - if decision $u$ is taken, afterwards transition into $j$ with probability $q_i^u(j) / -q_i^u(i)$
- $r_u(i) \in \mathbb{R}_+$ the (non-negative, decision dependent) reward in state $i$, we assume $r_u(i) < \infty$
- $s_u(i,j)$ the (non-negative, decision dependent) reward of a transition from state $i$ into state $j$; we assume $s_u(i,j) < \infty$ and $s_u(i,j) = 0$ for $i=j$ or $q_i^u(j) = 0$
Goal: Analysis and control of the system in the interval \([0, T]\)

\((T = \infty\) is included\)

- \(d_t\) is the decision vector at time \(t\) where \(d_{t}(i) \in D_i\)
- Decision space \(D = \times_{i=1..n} D_i\) \((\text{size } \prod_{i=1..n} m_i)\)
- \(Q^d \in \mathbb{R}^{n,n}\) with \(Q^d(i,j) = q_{i}^{d(i)}(j)\) transition matrix of the CTMC under decision vector \(d\)
- \(r^d \in \mathbb{R}^{n,1}\) rate reward vector under decision vector \(d\)
- \(S^d \in \mathbb{R}^{n,n}\) impulse reward matrix under decision vector \(d\)
- \(p_0\) is the initial distribution of the CTMDP at time \(t=0\)
Control of CTMDP via policies:

A policy $\pi$ is a measurable function from $[0,T]$ into $D$ (set of all policies $\mathcal{M}$)

$\Rightarrow$ decisions can depend on the time and the state (but not on the history)

$\pi_t$ defines $d_t$, the decision vector taken at time $t$

$\pi_{t,T}$ is the policy $\pi$ restricted to the interval $[t,T]$

Policy $\pi$ is

- **piecewise constant,** iff $0 = t_0 < t_1 < \ldots < t_m = T$ exist such that $d_t = d_{t'}$
  for $t,t' \in (t_{i-1}, t_i]$

- **constant,** iff $d$ is independent of $t$
  (i.e., decisions depend only on the state)
Other forms of policies:

- **randomized policy**: $\pi_t$ defines a probability distribution over $\mathcal{D}$
- **history dependent**: $\pi_t$ depends on $(x_{t'}, a_{t'})$ for $0 \leq t' < t$
  
  restricted forms of history dependency

- **reward dependent**: $\pi_t$ depends on reward accumulated in $[0,t)$

Policies types can be combined:

E.g., piecewise constant history dependent, constant randomized …. 
CTMDP with a fixed policy $\pi$: Stochastic Process $(X_t, A_t)$

- $X_t$ state process
- $A_t$ action/decision process

Both processes together define

- $G_t$ gain/reward process (i.e., accumulated reward in $[0,t]$)

Behavior of $G_t$:

- Changes with rate $r^a(i)$ if $X_t = i$ and $A_t = a$
- Makes a jump of height $s^a(i,j)$ if $X_t$ jumps at time $t$ from $i$ to $j$ and $A_t = a$
Policy:
- accept if $t < 2$ and $\text{popul} < 3$
- reject otherwise

Rewards:
- 3-$\text{popul}$ per time unit
- 1 per service
Mathematical Basics:
Assume for the moment that policy $\pi$ is known

Evolution of the state process $X_t$:
Let $0 \leq t \leq u \leq T$

$$V_{t,t}^\pi = I \quad \text{and} \quad \frac{d}{du} V_{t,u}^\pi = V_{t,u}^\pi Q^{d_u}$$

$$V_{t,u}^\pi(i,j) := \text{Prob. of being in } j \text{ at } u, \text{ if the process is in } i \text{ at } t$$

Let $V_t^\pi = V_{0,t}^\pi$

We have $p_u = p_t V_{t,u}^\pi$ and $p_t = p_0 V_t^\pi$
Evolution of the gain process $G_t$:

Let $W^d_t$ be a $n \times n$ matrix with:

$$W^d_t(i,j) = \begin{cases} r^d_t(i)(i) & \text{if } i = j \\ s^d_t(i)(i,j)q^d_t(j) & \text{otherwise} \end{cases}$$

Then $-\frac{d}{dt}g^\pi_t = w^\pi + Q^d_t g^\pi_t$ (backwards in time!)

$g^\pi_{t,T}$ is the accumulated gain in the interval $[t,T]$ and $g^\pi_T$ the final gain at time $T$

such that $g^\pi_{t,T} = V^\pi_{t,T}g^\pi_T + \int_t^T V^\pi_{u,T}w^\pi du$
For \((T - t) \to \infty\) usually also \(g^\pi_{t,T}(i) \to \infty\) holds

Results for \([0, \infty]\) are not meaningful in this case!

Alternative ways of defining the gain vector

- **Time averaged gain**

\[
 g^\pi_{t,T} = \frac{1}{T - t} \left( V^\pi_{t,T} g^\pi_T + \int_t^T V^\pi_{u,T} w^\pi du \right) \text{ for } T > t
\]

- **Discounted gain**

\[
 g^\pi_{t,T} = e^{-\beta(T-t)} V^\pi_{t,T} g^\pi_T + \int_t^T e^{-\beta(u-t)} V^\pi_{u,T} w^\pi du
\]

for discount factor \(\beta > 0\)
From analysis to optimization problems:

Find a policy $\pi$ (from a specific class) that maximizes/minimizes the gain

Optimal gain

$$g_{0,T}^+ = \sup_{\pi \in \mathcal{M}} (g_{0,T}^\pi) \quad \text{(or } g_{0,T}^- = \inf_{\pi \in \mathcal{M}} (g_{0,T}^\pi)\text{)}$$

Optimal policy

$$\pi^+ = \arg \sup_{\pi \in \mathcal{M}} (g_{0,T}^\pi) \quad \text{(or } \pi^- = \arg \inf_{\pi \in \mathcal{M}} (g_{0,T}^\pi)\text{)}$$

$\pi^+/\pi^-$ need not be unique!

Policy $\pi$ is $\varepsilon$-optimal iff

$$\left\| g_{0,T}^{\pi^\pm} - g_{0,T}^{\pi} \right\| \leq \varepsilon$$
Examples (queuing networks)

Routing

Resource allocation/deallocation

Scheduling

© Peter Buchholz 2011
Examples (reliability, availability)

Components
- Working
- Maintenance
- Failure

System
- Find a maintenance policy to minimize system down time

© Peter Buchholz 2011

CTMDP Basics
Examples (security)

Attack tree
(specifies sequence of attack steps)

Adversary: Find attack steps to reach goal
Defender: Find mechanisms to defend the system
Examples (OR)

Inventory control

- orders
- delivery
- sold goods

Inventory

Airline yield management

- arrival rate $\lambda(t)$
- reject
- accept
- cancellation
- no show

CTMDP Basics
Examples (AI)

Agent and environment are modeled as CTMDPs
Optimal policy corresponds to an “optimal” behavior of the agent
CTMDPs on Infinite Horizons

We consider a CTMDP in the interval \([0, \infty)\).

- Result Measures
  - Reward to Absorption
  - Average Reward
  - Discounted Reward
- Optimal Policies
- Computational Methods
Reward to Absorption

Assumptions:

1. \( w^d(i) \geq 0 \) for \( i \in S_t \) and \( 0 \) for \( i \in S_a \) and all \( d \in D \)

2. \[ \lim_{t \to \infty} \sum_{j \in S_a} V_t^\pi(i, j) = 1 \] for all \( i \in S \) and all \( \pi \in \mathbb{M} \)

Goal:

Find a policy \( \pi \) such that \( g^\pi = \int_0^\infty V_t^\pi h^\pi dt \) is minimal/maximal in all components
Average Reward

Find a policy $\pi$ such that

$$g^\pi = \lim_{T \to \infty} \frac{1}{T} \left( \int_0^T V_t^\pi w^\pi dt \right)$$

is maximal/minimal in all components

No further assumptions necessary ($\lim_{t \to \infty} (V_t^\pi)$ exists in our case!)

Discounted Reward

Find a policy $\pi$ such that

$$g^\pi = \lim_{T \to \infty} \left( \int_0^T e^{-\beta t} V_t^\pi w^\pi dt \right)$$

is maximal/minimal in all components

(discount factor $\beta \geq 0$)
Optimal Policies:

Stationary policies $\pi_1$ and $\pi_2$ exist such that

$$g^{\pi_1} = \sup_{\pi \in \mathcal{M}} (g^\pi) \quad \text{and} \quad g^{\pi_2} = \inf_{\pi \in \mathcal{M}} (g^\pi)$$

in the cases we consider here

The optimal policies might not be unique such that other criteria can be applied to rank policies with identical gain vectors (we do not go in this direction here)
Computational Methods

- Basic step is the transformation of the CTMDP into an equivalent DTMDP (plus a Poisson process) using uniformization

- Afterwards methods for computing the optimal policy/gain vector in DTMDPs can be applied (Poisson process is not needed)
The uniformization approach:

1) Poisson process with rate \( \alpha \geq \max_{d \in D} \max_{i \in S} (-Q^d(i,i)) \) for the timing

2) DTMC \( P^d = Q^d / \alpha + I \) for transitions

3) Transformed rewards
\[
W^d(i) = \frac{w^d(i)}{\alpha} \quad \text{or} \quad \frac{w^d(i)}{\alpha + \beta}
\]

discount factor \( \beta' = \frac{\alpha}{\alpha + \beta} \)
Uniformization transforms a CTMDP $= (Q^d, p_0, w^d, S, D)$ into a DTMDP $= (P^d, p_0, w'^d, S, D)$ with an identical
- optimal gain vector $g^*$ and
- optimal stationary policy $\pi^*$ (described by vector $d^*$)

The optimal gain vector is the solution of the following equation for the discounted and total reward

$$g^*(i) = \max_{u \in D_i} \left( w'^u(i) + \beta' \sum_{j=1}^{n} P^u(i, j) g^*(j) \right) \text{ for } i = 1, \ldots, n$$

Vector $d^*$ results from setting $d^*(i)$ to $argmax$ in the equation
For the optimization of average rewards, we restrict the model class to unichain models:

A DTMDP is unichain if for every $\mathbf{d} \in \mathcal{D}$ matrix $\mathbf{P}^d$ contains a single recurrent class of states (plus possibly some transient states)

For unichain DTMDPs the average reward is constant for all states and observes the following equations

$$\rho^* + h^*(i) = \max_{u \in \mathcal{D}_i} \left( w^{nu}(i) + \sum_{j=1}^{n} P^{u}(i, j) h^*(j) \right) \text{ for } i = 1, \ldots, n$$
Numerical methods to compute the optimal gain vector + policy

- linear programming
- value iteration
- policy iteration + combinations of value and policy iteration

Practical problems

- Curse of dimensionality, state space explosion
- Slow convergence, long solution times
**Infinite Horizons**

**Informatik IV**

### Linear Programming

#### Discounted Reward Case

\[ g \leq \max_{d \in D} \left( w'^d + \beta' P^d g \right) \leq g^* = w'^* + \beta' P^* g^* \]

\( g^* \) is the largest \( g \) that satisfies

\[ g \leq \max_{d \in D} \left( w'^d + \beta' P^d g \right) \]

**LP**

\[ \max \left( \sum_{i \in S} x_i \right) \]

subject to

\[ x_i \leq w'^u(i) + \beta' \sum_{j=1}^{n} P^u(j, i)x_j \]

for all \( i \in S, u \in D_j \)

---

#### Average Reward Case

\[ \max \left( \sum_{i \in S} \sum_{u \in D_i} x_i^u w'^u(i) \right) \]

subject to

\[ \sum_{u \in D_i} x_i^u - \sum_{j \in S} \sum_{u \in D_i} P^u(j, i)x_j^u = 0 \]

\[ \sum_{i \in S} \sum_{u \in D_i} x_i^u = 1 \]

for all \( i \in S, u \in D_i \)
- $d^*$ and $g^*$ ($\rho^*$ and $h^*$) can be derived from the results of the LP $x_i^u$

- An LP with
  - $n$ or $\sum_{i=1}^{n} |D_i|$ variables and
  - $\sum_{i=1}^{n} |D_i|$ or $n+1$ constraints

  has to be solved (this can be time and memory consuming!)

**Usually Linear Programming is not the best choice to compute optimal policies for MDP problems!**
Value iteration

Initialize $k=0$ and $g^{(k)} \geq 0$, then iterate until convergence

$$g^{(k+1)}(i) = \max_{u \in \mathcal{D}_i} \left( w'(i) + \beta' \sum_{j=1}^{n} P^u(i,j)g^{(k)}(j) \right)$$

or

$$h^{(k+1)}(i) = \max_{u \in \mathcal{D}_i} \left( w'(i) + \sum_{j=1}^{n} P^u(i,j)h^{(k)}(j) \right)$$

for all $i = 1, \ldots, n$

Repeated vector matrix products
Easy to implement

Convergence towards the optimal gain vector and policy
stop if policy does not change and \( \| g^{(k)} - g^{(k-1)} \| < \varepsilon \)

Effort per iteration: \( n \cdot nz \cdot \sum_{i=1}^{n} m_i \)

where \( nz \) is the average number of non-zeros in a row of \( P^d \)

Often slow convergence and a huge effort even for CTMDPs of a moderate size
Policy iteration

Initialize $k=0$ and initial policy $d_0$, then iterate until convergence

$$g^{(k)} = w' d_k + \beta' P d_k g^{(k)} \quad \text{or}$$

$$h^{(k)} + \rho^k e = w' d_k + \beta' P d_k h^{(k)} \quad \text{and} \quad h^{(k)}(n) = 0$$

and

$$d_{k+1} = \arg \max_{d \in \mathcal{M}} \left( w' d + \beta' P d g^{(k)} \right) \quad \text{or}$$

$$d_{k+1} = \arg \max_{d \in \mathcal{M}} \left( w' d + P d h^{(k)} \right)$$

until $d_{k+1} = d_k$
- Optimal policy and gain vector is computed after finitely many steps

- Effort per iteration: \( O(n^3) + O(n \cdot n z \cdot \sum_{i=1}^{n} m_i) \)

- For larger state spaces huge effort
  (solution of a set of linear equations of order n)
Improvements by combining policy and value iteration

Initialize $k=0$ and initial policy $d_0$, then iterate until convergence

\[ g^{(k)} \leftarrow \text{iterate} \left( (I - \beta' P^d_k) \ g = w'd_k \right) \quad \text{or} \]

\[ h^{(k)} + \rho^k \leftarrow \text{iterate} \left( (I - P^d_k) \ h = w'd_k \quad \text{subject to} \ h^{(k)}(n) = 0 \right) \]

where iterate is some advanced iteration techniques like SOR, ML, GMRES….and

\[ d_{k+1} = \arg \max_{d \in \mathcal{M}} \left( w'd + \beta' P^d g^{(k)} \right) \quad \text{or} \]

\[ d_{k+1} = \arg \max_{d \in \mathcal{M}} \left( w'd + P^d h^{(k)} \right) \]

until $d_{k+1} = d_k$ and $\|g^{(k)} - g^{(k-1)}\| < \varepsilon$ or $\|h^{(k)} - h^{(k-1)}\| < \varepsilon$
Example: Exponential class switching delay and service times

Poisson or IPP arrivals

Finite buffer capacity over all classes

Nonlinear reward function:

\[(n_1 + n_2 + n_3)^{1.2} + n_1 + 1.5n_2 + 2n_3\]

Determination of the best non-preemptive scheduling strategy in steady state

Methods

- Value iteration
- Policy iteration
- Combined approach with BiCGStab or GMRES + ILU preconditioning
Exponential interarrival times (fast convergence of all solvers)

No need for advanced solvers
value iteration works fine!
IPP interarrival times (systems are much harder to solve)

Advanced solvers are much more efficient for larger configurations!
Some remarks:

- Analysis for infinite horizons in principle well understood
- Complex models with larger state spaces require sophisticated solution methods
- Work on numerical methods less developed than in CTMC/DTMC analysis
  - Which method to use?
  - Which preconditioner?
  - How many iterations?
  - …….
CTMDPs on Finite Horizons

We consider a CTMDP in the interval \([0,T]\) \((T < \infty)\)

- Result Measures
  - Accumulated Reward
  - Accumulated Discounted Reward

- Optimal Policies
- Computational Methods
Many problems are defined naturally on a finite horizon, e.g.,

- yield management (for a specific tour)
- scheduling (for a specific set of processes)
- maintenance (for a system with finite operational periods)
- …

use of the optimal stationary policy is suboptimal

(e.g., maintenance for a machine just before it is shut down)
Optimal policies for DTMDPs on a horizon of T steps

Compute iteratively

\[ g^{(k+1)}(i) = \max_{u \in D_i} \left( w^u(i) + \beta \sum_{j=1}^{n} P^u(i, j) g^{(k)}(j) \right) \]

for \( k=1,2,\ldots,T \) starting with \( g^{(0)}(i) = 0 \) for all \( i=1,\ldots,n \)

where \( \beta = 1.0 \) for the case without discounting

Effort \( O(T \cdot n \cdot nz \cdot (\sum_{i=1..n} m_i)) \)
How about applying the DTMC resulting from uniformization for the transient analysis of CTMDPs?

- Times between transitions in the uniformized DTMC are exponentially distributed with rate \( \alpha \) rather than constant.
- Substitution of the distribution by the mean as done in the stationary case does not work.
- It has been shown that the approach computes the optimal policy for uniformization parameter \( \alpha \to \infty \) (but this results in \( O(\alpha T) \) steps to cover the interval \([0, T]\)).
Poisson process with rate
\[ \alpha \geq \max_{d \in D} \max_{i \in S} (-Q^d(i,i)) \]
for the timing.

Probability for \( k \) events in \([t-\delta, t]\)
\[ \gamma(t, k) = e^{-\alpha t} t^k / k! \]
Probability for \( >k \) events
\[ 1 - \sum_{l=0..k} \gamma(t, k) \]

Uniformization revisited

Each event equals a transition in the DTMC.

Finite Horizons
Known results for CTMDPs on finite horizons
(by Miller 1968, Lembersky 1974, Lippman 1976 all fairly old!)

- A policy is optimal if it maximizes for almost all \( t \in [0,T] \)

\[
\max_{\pi \in \mathcal{M}} (Q^\pi g_t + w^\pi) \quad \text{where} \quad -\frac{d}{dt} g_t = Q^{d_t} g_t + w^{d_t} \quad \text{and} \quad g_T \geq 0
\]

- There exists a piecewise constant policy \( \pi^* \) which results in vector \( g^*_t \) and maximizes the equation

(a policy is piecewise constant, if \( m < \infty \) and

\( 0 = t_0 < t_1 < \ldots < t_m = T \) exist and \( d_i \) is the decision vector in \( [t_{i-1}, t_i) \))

i.e., the optimal policy depends on the time and the state but changes only finitely often in \( [0,T] \)
Selection of an optimal policy (using results of Miller 1968)

Assume that $g^*_t$ is known, then the following selection procedure select $d^*$ which is optimal in $(t-\delta^*, t)$ for some $\delta^* > 0$

Define the sets

$F_1(g^*_t) = \{d \in D \mid d \text{ maximizes } q^{(1)}(d)\}$

$F_2(g^*_t) = \{d \in F_1(g^*_t) \mid d \text{ maximizes } -q^{(2)}(d)\}$

$\ldots$

$F_{n+1}(g^*_t) = \{d \in F_n(g^*_t) \mid d \text{ maximizes } (-1)^n q^{(n+1)}(d)\}$

where $q^{(1)}(d) = Q^d g^*_t + w^d$

$q^{(j)}(d) = Q^d q^{(j-1)}$ where $q^{(j-1)} = q^{(j-1)}(d)$ for any $d \in F_{j-1}(g^*_t)$

Select the lexicographically smallest vector from $F_{n+1}(g^*_t)$
Constructive proof in Miller 1968 defines a base for an algorithm:

1. Set \( t' = T \) and initialize \( g_T \);
2. Select \( d_t \) as described;
3. Obtain \( g_t \) for \( t \leq t' \leq T \) by solving
   \[
   -\frac{d}{dt} g_t = Q^{d_t} g_t + w^{d_t}
   \]
   with terminal condition \( g_{t'} \);
4. Set \( t'' = \inf \{ t \mid d_t \text{ satisfies the selection procedure} \} \);
5. If \( t'' \leq 0 \) terminate, else goto 2. with \( t' = t'' \);

Not implementable!
From exact to approximate optimal policies:

A policy $\pi$ is $\epsilon$-optimal if $\|g^* - g^\pi\|_\infty \leq \epsilon$ for all $t \in [0, T]$

Discretization approach:

Define for $h$: $P^d_h = I + hQ^d$

for $h$ small enough this defines a stochastic matrix

Let $g_{t-h} = P^d_h g_t + hw^d + o(h)$

Representation of a DTMDP for policy optimization

- For $h \rightarrow 0$ the computed policy is $\epsilon$-optimal with $\epsilon \rightarrow 0$
- Value of $\epsilon$ is unknown, effort $O(h^{-1} \cdot n \cdot nz \cdot (\sum_{i=1}^n m_i))$
Idea of the uniformization based approach

- Idealized policies non realizable
- Best reward
- $d^*$ locally best policy

Graph showing reward over time with $t$, $t-\delta^*$, and $t-\delta$.
A uniformization based approach

Basic steps

1. Start with $t=T$ and $g^T = g^- _T = g^+ _T = 0$;
2. Compute an optimal decision vector $d$ based on $g^- _t$;
3. Compute a lower bound for $g^- _{t-\delta}$ using decision $d$ in $(t-\delta, t)$;
4. Compute an upper bound $g^+ _{t-\delta}$ for any policy in $(t-\delta, t)$;
5. Choose $\delta$ such that $||g^+ _{t-\delta} - g^- _{t-\delta}|| < \varepsilon (t - \delta) / T$;
6. Store $d$ and $t-\delta$;
7. If $t-\delta > 0$ then set $t = t-\delta$ and goto 2; else terminate;

All this has to be computable!
Computation of the lower bound $g_{t-\delta}$ using decision $d$ in $[t-\delta,t)$;

$d$ is the optimal decision at $t$ based on $g_t$.

Since $d$ is constant, this equals the transient analysis of a CTMC which can be computed using uniformization.

Consider up to $K$ Poisson steps (Prob. known!)
Formally, we get:

\[ v^{(k)}_\neg = P^d v^{(k-1)}_\neg \quad \text{with} \quad v^{(0)}_\neg = g_t^- \quad \text{and} \quad w^{(k)}_\neg = P^d w^{(k-1)}_\neg \quad \text{with} \quad w^{(0)}_\neg = w^d \]

Then

\[ g^-_{t-\delta} = \sum_{k=0}^K \gamma(\alpha\delta, k) v^{(k)}_\neg + \frac{1}{\alpha} \sum_{k=0}^K \left( 1 - \sum_{l=0}^{k} \gamma(\alpha\delta, l) \right) w^{(k)}_\neg + \eta(\alpha\delta, w^d, g^-_t) \]

where \( \eta(\alpha\delta, w^d, g^-_t) \) bounds the missing Poisson probabilities.

Effort

- for vector computation in \( O(K \cdot n \cdot nz) \)
- for evaluation of a new \( \delta \) in \( O(K \cdot n) \)
Computation of the upper bound $g^+_{t-\delta}$ based on two ideas:

1. Compute separate policies for accumulated reward and reward gained at the end

2. Assume that we can choose at every jump of the Poisson process a new policy

Compute for $k = 0, 1, \ldots, K$ steps in the interval

Reward gained + accumulated in $k+1, k+2, \ldots$ steps (upper bound)
Formally

\[ v^{(k)}_+ = \max_{d \in D} \left( P^d v^{(k-1)}_+ \right) \text{ with } v^{(0)} = g_t^+ \]

and

\[ w^{(k)}_+ = \max_{d \in D} \left( P^d w^{(k-1)}_+ \right) \text{ with } w^{(0)} = \max_{d \in D} (w^d) \] (identical for all intervals)

For \( g^+_t \geq g^*_t \) we obtain

\[ g^{+}_{t-\delta} = \sum_{k=0}^{K} \gamma(\alpha \delta, k) \left( v^{(k)}_+ \right) + \left( 1 - \sum_{l=0}^{k} \gamma(\alpha \delta, l) \right) w^{(k)}_+ + \nu(\alpha \delta, w^d, g^+_t) \geq g^*_{t-\delta} \]

\( \nu(\alpha \delta, w^d, g^+_t) \) bounds the truncated Poisson probabilities

Policy is better than any realizable policy!!
Effort

- for vector computation in $O(K \cdot n \cdot nz \cdot (\sum_{i=1}^{n} m_i))$
- for evaluation of a new $\delta$ in $O(K \cdot n)$
  can be applied in a line search to find an appropriate $\delta$

Error proportional to the length of the subinterval

Choose $\delta$ such that $\| g^+_{t-\delta} - g^-_{t-\delta} \| < \varepsilon (t - \delta) / T$;
If $\delta$ is known and $s$ is independent of the decision vector, the upper bound can be improved by computing a non-realizable bounding policy for accumulated and gained reward

$$x_+^{(k)} = \max_{d_1^k, \ldots, d_k^k \in \mathcal{D}} \left( \prod_{i=1}^{k} p_d^i (\eta_k g_t^+ + \zeta_k w) \right)$$

where $\eta_k = \gamma(\alpha\delta, k)$ and $\zeta_k = \alpha^{-1} \left( 1 - \sum_{l=0}^{k} \gamma(\alpha\delta, l) \right)$

We have then

$$\sum_{k=0}^{\infty} x_+^{(k)} \geq g_{t-\delta}$$

Effort $O(K^2 \cdot n \cdot nz \cdot (\Sigma_{i=1..n} m_i))$
- Overall effort $O(K\cdot n\cdot \text{nz}\cdot \left(\sum_{i=1}^{n} m_i\right))$
  
  if the optimal policy is selected from $F_i(g_t)$ for some small $i$
  
  (usually the case)

- Local error $O(\delta^2) \Rightarrow$ Global error $O(\delta) \Rightarrow$
  
  for any $\varepsilon > 0$ (theoretically) the appropriate policy can be computed
For $T \geq 70.5058$:

- $b, b$ in $[T - 4.11656, T]$
- $b, r$ in $(T - 70.5058, T - 4.11656]$
- $r, r$ in $[0, T - 70.5058)$

$g_0^T = (20.931, 20.095, 19.138, 20.107, 8.6077)$

Bounds with $\varepsilon = 1.0e-6$
4 processors, 2 buses, 3 memories, 1 repair unit
Prioritized repair $\Rightarrow$ Computation of the availability
CTMDP with 60 states

<table>
<thead>
<tr>
<th>$\varepsilon$</th>
<th>iter</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>iter</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0e-2</td>
<td>1080</td>
<td>0.986336</td>
<td>0.995790</td>
<td>872</td>
<td>0.987109</td>
<td>0.995386</td>
</tr>
<tr>
<td>1.0e-4</td>
<td>2026</td>
<td>0.995633</td>
<td>0.995722</td>
<td>1419</td>
<td>0.995188</td>
<td>0.995282</td>
</tr>
<tr>
<td>1.0e-6</td>
<td>5896</td>
<td>0.995721</td>
<td>0.995721</td>
<td>21089</td>
<td>0.995279</td>
<td>0.995280</td>
</tr>
<tr>
<td>1.0e-8</td>
<td>34273</td>
<td>0.995721</td>
<td>0.995721</td>
<td>1513361</td>
<td>0.995280</td>
<td>0.995280</td>
</tr>
</tbody>
</table>
CTMDP with 121 states
Goal maximization of the throughput in [0,100]

<table>
<thead>
<tr>
<th>ε</th>
<th>iter</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>Sw. Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0e+0</td>
<td>6098</td>
<td>97.4599</td>
<td>98.4556</td>
<td>68.2561</td>
</tr>
<tr>
<td>1.0e-1</td>
<td>45866</td>
<td>97.4878</td>
<td>97.5875</td>
<td>68.3037</td>
</tr>
<tr>
<td>1.0e-2</td>
<td>334637</td>
<td>97.4879</td>
<td>97.4979</td>
<td>68.3099</td>
</tr>
<tr>
<td>1.0e-3</td>
<td>2981067</td>
<td>97.4881</td>
<td>97.4891</td>
<td>68.3102</td>
</tr>
</tbody>
</table>
Effort

- Linear in $\alpha t$ and in $\varepsilon^{-1}$ (and in $n\cdot nz$)
- Influence of $K$ depends on the model
  - $K$ too small results in small time steps due to the truncated Poisson probabilities
  - $K$ too large results in many unnecessary computations that are truncated due to the difference in the policy bounds
- Adaptive approach to choose $K$ such that fraction of the error due to the truncation of the Poisson probabilities remains constant
Extensions

- Method can be applied to discounted rewards after introducing small modifications.
- Method can be applied to CTMDPs with time dependent but piecewise constant rates and rewards.
- Method can be extended to CTMDPs with time dependent rates and rewards.
- Method can be extended to countable state spaces.
Advanced Topics

- Model Checking CTMDPs
- Inifinite State Spaces
- CTMDPs with bounds on the transition rates
- Equivalence of CTMDPs
- Partially observable CTMDPs
- ……
CSL model checking of CTMDPs

joint work with Holger Hermanns, Ernst Moritz Hahn, Lijun Zhang

- Model checking of CTMCs is very popular
- Extension for CTMDPs define formulas that hold for a set of states and all schedulers / some scheduler
- Model checking means to compute for every state whether a formula holds/does not hold
- Validation of path formulas requires computation of minimal/maximal gain vectors for finite or infinite horizons
Syntax of CSL for CTMDPs:

$$\Phi ::= a \mid \neg \Phi \mid \Phi \land \Phi \mid P_J(\Phi \cup I \Phi) \mid S_I(\Phi) \mid \Pi_J^t(\Phi) \mid C_J^t(\Phi)$$

where I und J are closed intervals and t is some time point

$$a \mid \neg \Phi \mid \Phi \land \Phi$$ with the usual interpretation
$s \models \mathbb{P}_J(\Phi \cup^I \Psi)$ if the probability of all paths that start in state $s$ observe $\Phi$
unlabeled until $\Psi$ in $I$ falls into $J$ for all policies

$s \models \mathbb{S}_I(\Phi)$ if the process starts in state $s$ and has a time averaged stationary
reward over state observing $\Phi$ that lies in $I$ for all policies

$s \models \mathbb{I}^t_J(\Phi)$ if the process starts in state $s$ and has a instantaneous reward at
time $t$ in states observing $\Phi$ that lies in $J$ for all policies

$s \models \mathbb{C}^I_J(\Phi)$ if the process starts in state $s$ and has an accumulated reward
over states observing $\Phi$ in the interval $I$ that lies in $J$ for all policies
Validation of formulas, an example:

\[ s \models C_{[p_1, p_2]}^{[t_0, T]}(\Phi) \quad \text{iff} \quad a^- (s) \geq p_1 \land a^+ (s) \leq p_2 \]

where \( a^- = \inf_{\pi \in \mathcal{M}} (V_{0, t_0}^\pi g_{t_0, T}^\pi | \Phi) \) and \( a^+ = \sup_{\pi \in \mathcal{M}} (V_{0, t_0}^\pi g_{t_0, T}^\pi | \Phi) \)

Two step approach to compute \( V_{0, t_0}^\pi \) and \( g_{t_0, T}^\pi | \Phi \)

Computation of \( V_{0, t_0}^\pi \) with the standard approach, i.e.

\[ V_{t, t} = I \quad \text{and} \quad \frac{d}{du} V_{t, u} = V_{t, u} Q^d_u \]

Computation of \( g_{t_0, T}^\pi | \Phi \) using vectors \( s^\pi | \Phi \) and \( g_T | \Phi \) then

\[ g_{t, T}^\pi | \Phi = V_{t, T}^\pi g_{T}^\pi | \Phi + \int_t^T V_{u, T}^\pi w^\pi | \Phi du \]

where \( h | \Phi (s) = h(s) \text{ if } s \models \Phi \text{ and } 0 \text{ else} \)
A small example

$s_4$ is made absorbing to compute the probability of reaching $s_4$ in the interval $[3,7]$

\[
\sup_{\pi \in \mathcal{M}} \left( P_{s_1}^\pi \left( \text{true } U^{[3,7]} s_4 \right) \right) = P_{s_1}^{\max} \left( \diamond^{[3,7]} s_4 \right)
\]

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$\epsilon = 1.0e - 3$</th>
<th>$\epsilon = 6.0e - 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_1$</td>
<td>time bounded prob.</td>
<td>iter$_1$</td>
</tr>
<tr>
<td>9.0e - 4</td>
<td>0.97170</td>
<td>0.97186</td>
</tr>
<tr>
<td>5.0e - 4</td>
<td>0.97172</td>
<td>0.97186</td>
</tr>
<tr>
<td>1.0e - 4</td>
<td>0.97175</td>
<td>0.97185</td>
</tr>
<tr>
<td>1.0e - 5</td>
<td>0.97175</td>
<td>0.97185</td>
</tr>
</tbody>
</table>

Table 1: Bounds for reaching $s_4$ in $[3,7]$, i.e., $P_{s_1}^{\max}(\diamond^{[3,7]} s_4)$. 

© Peter Buchholz 2011

Advanced Topics
Infinite (countable) state spaces

- Infinite horizon
  - Discounted reward
  - Average reward
- Finite horizon

We assume that the control space per state remains finite, all rewards and transitions rates are bounded.
Infinite horizon discounted case

- We assume that the optimality equations (slide 28) for the original model have a solution

- We compute results on $\hat{S} = \{1, \ldots, n\} \subset S$

- Let $\hat{Q}^d$ be the $n \times n$ submatrix of $Q^d$ and $\hat{w}^d$ the $n$ dimensional subvector of $w^d$ restricted to states from $\hat{S}$

- Define $\hat{P}^d = \hat{Q}^d/\alpha + I$ and $\hat{w}'^d = w^d/(\alpha + \beta)$ (see slide 27)

This defines an DTMDP with $n$ states
Value iteration:

- Modified DTMDP can be analyzed with value iteration (see slide 33)

- If \( \lim_{k \to \infty} \| \hat{\mathbf{P}}_k^{d_k} \mathbf{g}^{(k-1)}|_{\hat{S}} - \left( \mathbf{P}_k^{d_k} \mathbf{g}^{(k-1)} \right) \|_{\infty} < \epsilon \) then
  \[
  \| \hat{\mathbf{g}}^* - \mathbf{g}^* \|_{\infty} < \frac{\epsilon}{1 - \beta'}
  \]

- \( \hat{\mathbf{g}}^*, \mathbf{g}^* \) are the vectors to which value iteration applied to the reduced and original MDP converges

- \( d_k(i) \) equals the decision in state \( i \) during the \( k \)-th iteration of value iteration applied to the finite system if \( i \in \hat{S} \) and is arbitrary else

- \( g^{(k)}(i) \) equals the value in in the \( k \)-th iteration of value iteration applied to the finite system if \( i \in \hat{S} \) and is an upper bound for the value function otherwise.
Policy iteration:

- If for some policy $\pi$, vector $\tilde{g}^{(k)}|_{\hat{S}}$ is the gain vector restricted to states from $\hat{S}$ and $\hat{g}^{(k)}$ is the approximate solution computed for the finite system, such that $\|\tilde{g}^{(k)}|_{\hat{S}} - \hat{g}^{(k)}\|_\infty < \delta$ and

$$\|\hat{P}^{d_k+1}|_{\hat{S}} \hat{g}^{(k)}|_{\hat{S}} - \left( P^{d_k+1} \tilde{g}^{(k)}(k) \right)|_{\hat{S}} \|_\infty < \epsilon$$

where $d_{k+1}$ results from $\hat{g}_k$ and $\tilde{g}^{(k)}$ is an arbitrary extension of $\hat{g}^{(k)}$ to $S$,

then

$$\lim_{k \to \infty} \left( \|\hat{g}^{(k)} - g^*|_{\hat{S}}\|_\infty \right) < \frac{\epsilon + 2\beta' \delta}{(1 - \beta')^2}$$
Infinite horizon average case

- Additional conditions are required to assure existence of an optimal policy (many different conditions exist in the literature)

- Let for some policy $\pi$ be:
  - $C_\pi$ the expected gain of a cycle that starts in state 1 and ends when entering state 1 again
  - $N_\pi$ the expected number of visited states between two visits of state 1

  If for all policies $C_\pi$ is finite and the $N_\pi$ are uniformly bounded, then the Bellman equations have an optimal solution

Solution often via simulation
Finite horizon but infinite state spaces
(not much known from an algorithmic perspective!)

Assumptions:

- $S_0 = \{ i \mid p_0(i) > 0 \}$, let $|S_0| < \infty$
- Transition rates are bounded by $\alpha < \infty$

Define

- $S_k = \{ j \mid (\sum_{d \in D} P^d)^k(i,j) > 0 \text{ for some } i \in S_0 \}$
- $S_T(\varepsilon) = \bigcup_{k=0,\ldots,K_\varepsilon} S_k$ where $K_\varepsilon = \min_K \sum_{k=1\ldots K} \gamma(T,k) \geq 1-\varepsilon$
Algorithm for approximating/bounding the accumulated reward in \([0,T]\)

1. Define some finite subset \(\hat{S}\) of the countable state space \(S\) (e.g. using \(S_T(\varepsilon)\) for some appropriate \(\varepsilon\))

2. Define a new CTMDP with state space \(\hat{S} \cup \{0\}\)

   Matrices
   \[
   \hat{Q}^d = \begin{cases} 
   Q^d(i, j) & \text{if } i, j \neq 0 \\
   0 & \text{if } i = 0 \\
   \sum_{h \notin \hat{S}} Q^d(i, h) & \text{if } i \neq 0, j = 0
   \end{cases}
   \]

   Vectors
   \[
   \hat{w}^d(\pm) = \begin{cases} 
   w^d(i) & \text{if } i \neq 0 \\
   \max / \min_{j \in S, u \in D_j} (w^u(j)) & \text{if } i = 0
   \end{cases}
   \]

   \[
   \hat{g}^T(\pm) = \begin{cases} 
   g^T(\pm)(i) & \text{if } i \neq 0 \\
   \max / \min_{j \in S,} (\hat{g}^T(\pm)(j)) & \text{if } i = 0
   \end{cases}
   \]

3. Solve the resulting CTMDP to obtain bounds for the original one
Bounds on the transition rates

- Transition rates $Q^d(i,j)$ and reward vectors $s^d$ are not exactly known but we know $L^d(i,j) \leq Q^d(i,j) \leq U^d(i,j)$ and $l^d \leq w^d \leq u^d$
  (sometimes known as Bounded Parameter MDPs see e.g. Givan et al. 2000)
  realistic model if parameters result from measurements

- Goal: Find a policy that maximizes the minimal/maximal gain over an infinite or finite interval
Infinite horizon case:

We assume that the CTMDP is unichain for all $L^d$

- Uniformization can be used to transform the CTMDP in an equivalent DTMDP (as we did before)

- For a fixed decision vector $d$ the minimal/maximal average reward is obtained by a matrix $P \in [L, U]$ where in every row all except one elements are equal to the corresponding element in matrix $U$ or $L$
  
  $\Rightarrow$ only finitely many possibilities exist
  
  $\Rightarrow$ determination of the bounds is again an MDP problem
Infinite horizon case (continued):

- Overall solution is the combination of two nested MDP problems (Markov two person game)
- Some solution algorithms exist (stochastic min/max control) but advanced numerical techniques are rarely used (room for improvements remains!)

Finite horizon case

- Inherently complex to the best of my knowledge almost no results (even for the simpler DTMDP case!)
Thank you!
Bibliography
(incomplete and biased selection but there is too much to be exhaustive!)

Textbooks:

Papers:


Own papers:

