Limitation of Markov Models and Event-Based Learning & Optimization

Plenary Presentation
at
2008 Chinese Control and Decision Conference
July 2, 2008    Yaitai, China

Xi-Ren Cao

Hong Kong University of Science and Technology
A Typical Formulation of a Control Problem
(Continuous Time Continuous State Model)

\[ \frac{dx}{dt} = Ax + Bu + w \]

- \( x \): State
- \( u \): Control variable
- \( w \): Random noise

Control \( u \) depends on state \( x \)
A policy \( u(x) \): \( x \rightarrow u \)

Performance measure
\[ \eta = \frac{1}{T} \int_{0}^{T} E\{f[x(t),u(t)]\} dt \]

LQG problem
\[ \eta = \frac{1}{T} \int_{0}^{T} E\{x^\top Ax + u^\top Bu\} dt \]
Discrete-time Discrete State Model (I)
- an example

A random walk of a robot

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Probabilities
\[ p + q = 1 \]

Reward function
\[ f(0) = 0 \]
\[ f(1) = f(4) = 100 \]
\[ f(2) = f(3) = -100 \]

Performance measure
\[ \eta = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} f(X_t) \]
Shannon Mouse (Theseus)
Discrete Model (II)
- the dynamics

A random walk of a robot

A Sample Path (system dynamics):
Discrete Model (III)
- the Markov model

System dynamics:
- $X = \{X_n, n=1,2,\ldots\}$, $X_n$ in $S = \{1,2,\ldots,M\}$
- Transition Prob. Matrix $P= [p(i,j)]_{i,j=1,\ldots,M}$

System performance:
- Reward function: $f=(f(1),\ldots,f(M))^T$
- Performance measure:
  \[
  \eta = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} f(X_t) = \pi^T f = \sum_{i \in S} \pi(i) f(i)
  \]

Steady-state probability:
- Steady-state probability:
  \[
  \pi=(\pi(1), \pi(2),\ldots,\pi(M)).
  \]
  \[
  \pi(I-P)=0, \quad \pi e=1
  \]
  $I$: identity matrix, $e=(1,\ldots,1)^T$
Control of Transition Probabilities

- move to left
- move to right

Turn on red with prob. $\alpha$

Turn on green with prob. $1-\alpha$
Discrete Model (IV)
- Markov decision processes (MDPs)
- the Control Model

\[ \alpha = d(x) \]

**System dynamics:**
Markov model

\[ \frac{dx}{dt} = Ax + Bu + w \]

- \( \alpha \): Action controls transition probabilities
- \( p^\alpha(i,j) \): governs the system dynamics
- \( \alpha = d(x) \): policy (state based)

Performance depend on policies, \( \pi^d \), \( \eta^d \), etc

\[ \eta^d = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} f(X_t^d) \]

**Goal of Optimization:**
Find a policy \( d \) that maximizes \( \eta^d \) in policy space
0. Review: Optimization Problems (state-based policies)

1. Event-Based Optimization
   - Limitation of the state-based formulation
   - Events and event-based policies
   - Event-Based Optimization
The policy space is too large

\[ M = 100 \text{ states}, \quad N = 2 \text{ actions}, \quad N^M = 2^{100} = 10^{30} \text{ policies} \]

(10GHZ \( \Rightarrow \) 3* 10^{12} \text{ years to count!})

- Special structures not utilized
- May not perform well
Limitation of State-Based Formulation (II)

Example: Random walk of a robot

Choose $\alpha$ to maximize the average performance
Limitation of State-Based Formulation (III)

Transition probabilities:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$p\alpha$</td>
<td>$p(1-\alpha)$</td>
<td>$q\alpha$</td>
<td>$q(1-\alpha)$</td>
</tr>
</tbody>
</table>

- **At state 0,**
  - if moves top, $\alpha$ needs to be as large as possible
  - if moves down, $\alpha$ needs to be as small as possible
- **Let $p = q = 1/2,$**
  - Average perf in next step = 0, no matter what $\alpha$ you choose (best you can do with a state-based model)
We can do better!

- Group two up transitions together as an event “a” and two down transitions as event “b”.
- When “a” happens, choose the largest $\alpha$.
  When “b” happens, choose the smallest $\alpha$.
- Average performance = 100, if $\alpha = 1$. 
Events and Event-Based Policies

An event is defined as a set of state transitions

Event-based optimization:
- May lead to a better performance than the state-based formulation
- MDP model may not fit:
  - Only controls a part of transitions
  - An event may consist of transitions from many states
- May reflect and utilize special structures

Questions:
- Why it may be better?
- How general is the formulation?
- How to solve event-based optimization problems?
Notations:

- A single transition \(<i,j>\),
  
  \(i,j \text{ in } S = \{1,2, \ldots, M\}\)

- An event: a set of transitions, \(2^M\) sets
  
  \(a = \{<0,1>, <0,2>\}\)
  
  \(b = \{<0,3>, <0,4>\}\)

Why it is better?

An event contains information about the future!

(compared with the state-based policies)
How general is the formulation?

- Event: a customer arrival finding population $n$
- Action: accept or reject
  - Only applies when an event occurs
- MDP does not apply: Same action is applied for different state with the same population

$n$: population
$N$: network capacity

$\lambda$: arrival rate
$\alpha(n)$, $1-\alpha(n)$: probabilities of accepting or rejecting
$q_{0i}$, $q_{ij}$: transition rates
Riemann Sampling vs. Lebesgue Sampling

Sample the system whenever the signal reaches a certain prespecified level, and control is added then.
$dX(t) = b(t, X(t))dt + \sigma(t, X(t))dw(t) + \int \gamma(t, X(t-), z)N(dt, dz).$

$w(t):$ Brownian motion;  $N(dt, dz):$ Poisson random measure
$X(t):$ Ito-Levy process
2. Sensitivity-Based Approach to Optimization
   - A unified framework for optimization
   - Extensions to event-based optimization

3. Summary
An overview of the paths to the top of a hill

\[ \frac{d\eta}{d\delta} = \pi(A_P)g \]

\[ \eta' - \eta = \pi(A_P)g \]
A Sensitivity-Based View of Optimization

- Continuous Parameters (perturbation analysis)
- Discrete Policy Space (policy iteration)

\[
\frac{d \eta}{d \delta} = \pi Q_g
\]

\[
\eta' - \eta = \pi' Q_g
\]

\(\eta\): performance
\(\pi\): steady-state prob
\(g\): perf. potentials
\(Q=P' - P\)
Poisson Equation

\[ g(i) = \text{potential contribution of state } i \ (\text{potential, or bias}) \]
\[ = \text{contribution of the current state } f(i) - \eta \]
\[ + \text{expected long term contribution after a transition} \]

\[ g(i) = f(i) - \eta + \sum_{j=1}^{M} p(i, j) g(j) \]

In matrix (Poisson equation):

\[ (I - P)g + \eta e = f \]

Potential is relative: if \( g(i) \) is solution, \( i=1, \ldots, M \), so is \( g(i) + c \), \( c \): constant

Physical interpretation:

\[ g(i) = E\left\{ \sum_{l=0}^{\infty} \left[ f(X_l) - \eta \right] \mid X_0 = i \right\} \]

\[ g(4) \approx \text{average of } \sum_{i=0}^{\infty} f(X_i) \]

starting from \( X_0 = 4 \)
Two Sensitivity Formulas

For two Markov chains \( P, \eta, \pi \) and \( P', \eta', \pi' \), let \( Q = P' - P \)

**Performance difference:**

\[
\eta' - \eta = \pi' Q g = \pi' (P' - P) g
\]

**One line simple derivation:**

\[\times \pi': (I - P) g + \eta e = f\]

**Performance derivative:**

\[
P \text{ is a function of } \theta: P(\theta)
\]

\[
\frac{d\eta(\theta)}{d\theta} = \pi \frac{dP(\theta)}{d\theta} g = \frac{d}{d\theta}[\pi P(\theta) g]
\]

Derivative = average change in expected potential at next step

Perturbation analysis: choose the direction with the largest average change in expected potential at next step
Policy Iteration

\[ \eta' - \eta = \pi'Qg = \pi'(P' - P)g \]

1. \( \eta' > \eta \) if \( P'g > Pg \) (Fact: \( \pi' > 0 \))

2. **Policy iteration:**
   At any state find a policy \( P' \) with \( P'g > Pg \)

   **Policy iteration:** Choose the action with largest changes in expected potential at next step

3. **Reinforcement learning**
   (Stochastic approximation algorithms)
Multi-Chain MDPs
Perf./ Bias/ Blackwell Optimization

With perf. difference formulas, we can derive a simple, intuitive approach without discounting

$D$: Policy space
$D_0$: Perf. optimal policies
$D_1$: (1st) Bias optimal policies
$D_2$: 2nd Bias optimal policies
$\ldots$
$D_M$: Blackwell optimal policies

Bias measures transient behavior
Two policies: $P, P'$, $Q = P' - P$
Steady-state prob: $\pi, \pi'$
Long-run ave. perf: $\eta, \eta'$
Poisson eq: $(I - P + e \pi)g = f$

RL: reinforcement learning, Neuro-DP ..

(online estimate)

Potentials $g$

$\frac{d \eta}{d \delta} = \pi Qg$

$\eta' - \eta = \pi' Qg$

Gradient-based PI

SAC: stochastic adaptive cont.

PA: perturbation analysis

MDP: Markov decision proc.

Online gradient based optimi

Online policy iteration

A Map of the L&O World

RL: reinforcement learning
PA: perturbation analysis
MDP: Markov decision proc.
SAC: stochastic adaptive cont.
Overview of State-Based Optimization

Introduction to Event-Based Optimization

Sensitivity-Based Approach to State-Based Optimization

Solution to Event-Based Optimization

Extension of the sensitivity-based approach to event-based optimization
Two sensitivity formulas
  • Performance derivatives
  • Performance differences

PA & PI
  • PA: Choose the direction with largest average change in expected potential at next step
  • PI: Choose the action with largest changes in expected potential at next step

Potentials are aggregated according to event structure
Solution to Random Walker Problem

Two policies:
\[
\begin{align*}
\alpha_a &= d(a), & \alpha_b &= d(b) \\
\alpha'_a &= d'(a), & \alpha'_b &= d'(b)
\end{align*}
\]

1. Performance diff:
\[
\eta' - \eta = \pi'(a)[(\alpha_a' - \alpha_a)g(a)] + \pi'(b)[(\alpha_b' - \alpha_b)g(b)]
\]
\[
g(a) = g(1) - g(2) \quad g(b) = g(3) - g(4)
\]
\[\pi'(a), \pi'(b): \text{perturbed steady-state prob. of events } a \text{ and } b\]

Choose the action with the largest changes
In expected potential at next step
\[g(a), \ g(b) \text{ aggregated}\]

2. Performance deriv:
\[
\frac{d\eta_{\theta}}{d\theta} = \pi_{\theta}(a)\frac{d\alpha_{a}(\theta)}{d\theta}[g_{\theta}(1) - g_{\theta}(2)] + \pi_{\theta}(b)\frac{d\alpha_{b}(\theta)}{d\theta}[g_{\theta}(3) - g_{\theta}(4)]
\]

Continuous with \(\theta\): \(\alpha_a(\theta), \ \alpha_b(\theta)\)
Solution to Admission Control Problem

Two policies: $\alpha(n)$ and $\alpha'(n)$

1. Performance diff:

$$\eta' - \eta = \sum_{n=0}^{N-1} \{p'(n)[\alpha'(n) - \alpha(n)]d(n)\}$$

$p(n)$: prob. of arrival finding $n$ cust.

Potential aggregation:

$$d(n) = \frac{1}{p(n)} \left\{ \sum_{i=1}^{M} q_{0i} \left[ \sum_{n_i=n} p(n)g(n_{i+1}) - \sum_{n_i=n} p(n)g(n) \right] - \sum_{n_i=n} p(n)g(n) \right\}$$

Choose the action with the largest changes
In expected potential at next step
$d(n)$: aggregated potential

2. Performance deriv:

$$\frac{d\eta}{d\delta} = \sum_{n=0}^{N-1} \{p(n)[\alpha'(n) - \alpha(n)]d(n)\}$$
Constructing New Sensitivity Eqs!

Sensitivity-Based Approaches to Event-Based Optimization

RL: reinforcement learning
PA: perturbation analysis
MDP: Markov decision proc.
SAC: stochastic adaptive cont.

Potentials $g$

$$\frac{d\eta}{d\delta} = \pi(e)Q(* | e)g_{agg}$$

$\eta' - \eta = \pi'(e)Q(* | e)g_{agg}$

PA
(Policy gradient)

MDP
(Policy iteration)

SAC

Gradient-based PI

Online gradient based optimi

Online policy iteration

RL: reinforcement learning
PA: perturbation analysis
MDP: Markov decision proc.
SAC: stochastic adaptive cont.
Summary
Advantages of the Event-Based Approach

1. **May have better performance**

2. **# of aggregated potentials** $d(n): N$
   - may be linear in system

3. **Actions at different states are correlated**
   - standard MDPs do not apply

4. **Special features captured by events**
   - action depends on future information

5. **Opens up a new direction**
   - to many engineering problems
   - POMDPs: observation $y$ as event
   - hierarchical control: mode change as event
   - network of networks: transitions among subnets as events
   - Lebesgue Sampling
Sensitivity-Based View of Optimization

1. A map of the learning and optimization world:
   Different approaches can be obtained from two sensitivity equations

2. Extension to event-based optimization
   Policy iteration, perturbation analysis
   reinforcement learning, time aggregation
   stochastic approximation, Lebesgue sampling
   ……

3. Simpler and complete derivation for MDPs
   Multi-chains, different perf. criteria
   Average performance with no discounting
   N-bias optimality – Blackwell optimality
Pictures to Remember (I)
Online gradient based optimisation

RL

Online policy iteration

PA

MDP

AC

Constructing New Sensitivity Eqs!

Potentials $g$

$\frac{d\eta}{d\delta} = \pi(e)Q^*(|e)g_{agg}$

$\eta' - \eta = \pi'(e)Q^*(|e)g_{agg}$

Gradient-based PI

Stochastic Approximation

Online gradient based optimisation

Online policy iteration

Pictures to Remember (II)
Limitation of State-Based Formulation (I)

0 Yautai
1 Alaska
2 Hawaii
Thank You!
Xi-Ren Cao:

Stochastic Learning and Optimization
- A Sensitivity Based Approach

9 Chapters, 566 pages
119 Figures, 27 Tables,
212 homework problems

Springer
October 2007