Optimal mixing of suboptimal decision rules for MDP control

Dinard van der Laan

Department of Econometrics and Operations Research
Vrije University Amsterdam

YEQT-IV 2010, November 26

Outline

Infinite horizon Markov decision problem (MDP)

$T = \{1, 2, \ldots\}$ is set of decision epochs

Infinite horizon Markov decision problem (MDP)

$T = \{1, 2, \ldots\}$ is set of decision epochs

Optimization criterion:
Maximize expected average reward

Infinite horizon Markov decision problem (MDP)

$T = \{1, 2, \ldots\}$ is set of decision epochs

Optimization criterion:
Maximize expected average reward

Markov policy π is determined by infinite sequence of decision rules,
$\pi = (d_1, d_2, \ldots)$.

Difficulties in maximizing the reward

State space becomes (too) large

Curse of dimensionality

Difficulties in maximizing the reward

State space becomes (too) large
Curse of dimensionality

Optimal decision rule may be hard to implement

Difficulties in maximizing the reward

State space becomes (too) large
Curse of dimensionality

Optimal decision rule may be hard to implement

Imperfect current state information

Difficulties in maximizing the reward

State space becomes (too) large
Curse of dimensionality

Optimal decision rule may be hard to implement

Imperfect current state information

Observing current state: costly,
time-consuming, impossible

Mixing decision rules

Idea: Restrict to some easy implementable (suboptimal) decision rules

Mixing decision rules

Idea: Restrict to some easy implementable (suboptimal) decision rules

Optimization Problem: Optimize control policy $\pi = (d_1, d_2, \ldots)$ under restriction $d_t \in \mathcal{D}$ for all $t \in T$ where $\mathcal{D}$ is a given finite set of Markov decision rules

Example $\mathcal{D}$ restricted problem

Queueing system: Route arriving jobs to heterogeneous servers/machines to minimize the average waiting time

Current state information: Number of jobs waiting in each queue

Example $\mathcal{D}$ restricted problem

Queueing system: Route arriving jobs to heterogeneous servers/machines to minimize the average waiting time
Current state information: Number of jobs waiting in each queue

$\mathcal{D} = \{d^1, d^2, d^3\}$ where

Example $\mathcal{D}$ restricted problem

Queueing system: Route arriving jobs to heterogeneous servers/machines to minimize the average waiting time
Current state information: Number of jobs waiting in each queue

$\mathcal{D} = \{d^1, d^2, d^3\}$ where

- d_1: Choose server with shortest waiting queue

Example $\mathcal{D}$ restricted problem

Queueing system: Route arriving jobs to heterogeneous servers/machines to minimize the average waiting time
Current state information: Number of jobs waiting in each queue

$\mathcal{D} = \{d^1, d^2, d^3\}$ where

- d_1: Choose server with shortest waiting queue
- d_2: Choose the server with highest service rate

Example $\mathcal{D}$ restricted problem

Queueing system: Route arriving jobs to heterogeneous servers/machines to minimize the average waiting time
Current state information: Number of jobs waiting in each queue

$\mathcal{D} = \{d^1, d^2, d^3\}$ where

- d_1: Choose server with shortest waiting queue
- d_2: Choose the server with highest service rate
- d_3: Choose a server at random

Improving performance by mixing suboptimal decision rules

Two approaches:

- Randomized stationary policies
- Deterministic non-stationary policies

Improving performance by mixing suboptimal decision rules

Two approaches:

- Randomized stationary policies
- Deterministic non-stationary policies

Let $\mathcal{D} = \{d^1, d^2, \ldots, d^k\}$ be the set of available decision rules,
P_i is transition matrix induced by d^i for $i = 1, 2, \ldots, k$

Bernoulli policies

At any decision epoch choose rule d^i with probability θ_i, $\sum_{i=1}^{k} \theta_i = 1$

Bernoulli policies

At any decision epoch choose rule d^i with probability θ_i, $\sum_{i=1}^{k} \theta_i = 1$

Bernoulli policy: Randomized and Stationary

Bernoulli policies

At any decision epoch choose rule d^i with probability θ_i, $\sum_{i=1}^{k} \theta_i = 1$

Bernoulli policy: Randomized and Stationary

Let $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ be the vector of probabilities determining the Bernoulli policy. The corresponding Bernoulli policy π_θ induces a stationary Markov chain with transition matrix $P_\theta = \sum_{i=1}^{k} \theta_i P_i$

Performance computation

Transition matrix $P_\theta = \sum_{i=1}^{k} \theta_i P_i$ induces aperiodic unichain MC if all P_i do.

Performance computation

Transition matrix $P_\theta = \sum_{i=1}^{k} \theta_i P_i$ induces aperiodic unichain MC if all P_i do.

Let p_θ be the unique stationary distribution for P_θ

Performance computation

Transition matrix $P_\theta = \sum_{i=1}^{k} \theta_i P_i$ induces aperiodic unichain MC if all P_i do.

Let p_θ be the unique stationary distribution for P_θ

For $i = 1, 2, \ldots, k$ let $r(d^i)$ be the expected immediate reward (or cost) vector if decision rule d^i is applied.
Performance $g^\pi(\theta)$ of Bernoulli policy π_θ:
$g^\pi(\theta) = \sum_{i=1}^{k} \theta_i (p_\theta \cdot r(d^i))$

Performance optimization

$g(\theta)$ is the expected Césaro average profit (costs) for Bernoulli policy of rate vector θ.

Performance optimization

$g(\theta)$ is the expected Césaro average profit (costs) for Bernoulli policy of rate vector θ.

- $g(\theta)$ is relatively easy to compute/approximate

Performance optimization

$g(\theta)$ is the expected Césaro average profit (costs) for Bernoulli policy of rate vector θ.

- $g(\theta)$ is relatively easy to compute/approximate
- $g(\theta)$ is independent of initial state distribution if all P_i are unichain and aperiodic.

Performance optimization

$g(\theta)$ is the expected Césaro average profit (costs) for Bernoulli policy of rate vector θ.

- $g(\theta)$ is relatively easy to compute/approximate
- $g(\theta)$ is independent of initial state distribution if all P_i are unichain and aperiodic.
- $g(\theta)$ is a smooth function

Performance optimization

$g(\theta)$ is the expected Césaro average profit (costs) for Bernoulli policy of rate vector θ.

- $g(\theta)$ is relatively easy to compute/approximate
- $g(\theta)$ is independent of initial state distribution if all P_i are unichain and aperiodic.
- $g(\theta)$ is a smooth function
- Techniques for computing/approximating optimal rate vector θ^* are available

Improvement by non-stationary mixing policies

Could a non-stationary Markov policy do better than the optimal Bernoulli policy?

Improvement by non-stationary mixing policies

Could a non-stationary Markov policy do better than the optimal Bernoulli policy?

For example for $\mathcal{D} = \{d^1, d^2, d^3\}$ suppose the optimal Bernoulli rate vector is (close to) (0.50,0.25,0.25)

Improvement by non-stationary mixing policies

Could a non-stationary Markov policy do better than the optimal Bernoulli policy?

For example for $\mathcal{D} = \{d^1, d^2, d^3\}$ suppose the optimal Bernoulli rate vector is (close to) (0.50,0.25,0.25)

Policy $\pi = (d^1, d^2, d^1, d^3, d^1, d^2, d^1, d^3, \ldots)$ (periodic with period 4) could very well be an improvement

Improvement by non-stationary mixing policies

Could a non-stationary Markov policy do better than the optimal Bernoulli policy?

For example for $\mathcal{D} = \{d^1, d^2, d^3\}$ suppose the optimal Bernoulli rate vector is (close to) (0.50,0.25,0.25)

Policy $\pi = (d^1, d^2, d^1, d^3, d^1, d^2, d^1, d^3, \ldots)$ (periodic with period 4) could very well be an improvement

Intuitively decisions are better spaced-out

Difficulties for non-stationary mixing policies

- Applied decision rules are time-dependent inducing a non-stationary Markov chain

Difficulties for non-stationary mixing policies

- Applied decision rules are time-dependent inducing a non-stationary Markov chain
- Performance computation and optimization is harder than for Bernoulli policies

Difficulties for non-stationary mixing policies

- Applied decision rules are time-dependent inducing a non-stationary Markov chain
- Performance computation and optimization is harder than for Bernoulli policies

For non-stationary mixing policies the performance may depend on the initial state distribution even if all transition matrices P_i induce aperiodic unichain MC

Counterexample

Suppose $\mathcal{D} = \{d^1, d^2\}$

Counterexample

Suppose $\mathcal{D} = \{d^1, d^2\}$

$$P_1 = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, P_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}$$

Counterexample

Suppose $\mathcal{D} = \{d^1, d^2\}$

$$P_1 = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, P_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}$$

Both decision rules induce an aperiodic unichain MC

Counterexample

$$P_1 P_2 = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Counterexample

$$P_1 P_2 = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Policy $(d^1, d^2, d^1, d^2, \ldots)$ induces two closed classes of states

Counterexample

$$P_1 P_2 = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Policy $(d^1, d^2, d^1, d^2, \ldots)$ induces two closed classes of states

Also $P_2 P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.25 & 0.25 \\ 0 & 0 & 1 \end{pmatrix}$

gives two closed classes

Initial state (in)dependence

In counterexample performance of periodic mixing policy $\pi = (d^1, d^2, d^1, d^2, \ldots)$ depends on the initial state distribution

Initial state (in)dependence

In counterexample performance of periodic mixing policy $\pi = (d^1, d^2, d^1, d^2, \ldots)$ depends on the initial state distribution

To apply and optimize over non-stationary mixing policies we demand $(d^1, d^2, d^1, d^2, \ldots)$ and $(d^2, d^1, d^2, d^1, \ldots)$ to have the same performance and this performance to be independent of initial state

Criterion

General criterion: For any given infinite sequence of $\mathcal{D}$ decision rules $(d_1, d_2, \ldots)$, bounded reward vectors $\{r(d^i), d^i \in \mathcal{D}\}$ and positive integers n, m:

Criterion

General criterion: For any given infinite sequence of $\mathcal{D}$ decision rules $(d_1, d_2, \ldots)$, bounded reward vectors $\{r(d^i), d^i \in \mathcal{D}\}$ and positive integers n, m:
$\mathcal{D}$ mixing policies $(d_n, d_{n+1}, \ldots)$ and $(d_m, d_{m+1}, \ldots)$ should have equal performance (expected Césaro average reward)

Criterion

General criterion: For any given infinite sequence of $\mathcal{D}$ decision rules $(d_1, d_2, \ldots)$, bounded reward vectors $\{r(d^i), d^i \in \mathcal{D}\}$ and positive integers n, m:
$\mathcal{D}$ mixing policies $(d_n, d_{n+1}, \ldots)$ and $(d_m, d_{m+1}, \ldots)$ should have equal performance (expected Césaro average reward)

Look for condition(s) on the transition matrices P_i induced by $d^i \in \mathcal{D}$

Coefficient of ergodicity

Dobrushin's coefficient of ergodicity for $P = (p_{ij})$:

Coefficient of ergodicity

Dobrushin's coefficient of ergodicity for $P = (p_{ij})$:

$$\rho_0(P) = \frac{1}{2} max_{i,j} \sum_k |p_{ik} - p_{jk}|$$

Coefficient of ergodicity

Dobrushin's coefficient of ergodicity for $P = (p_{ij})$:

$$\rho_0(P) = \frac{1}{2} max_{i,j} \sum_k |p_{ik} - p_{jk}|$$

Property: MC is aperiodic unichain if and only if $\rho_0(P^N) < 1$ for some positive integer N

Sufficient condition

Generalized ergodicity condition: Consider a $\mathcal{D}$ restricted MDP with $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$
Let $\mathcal{A} = \{P_1, P_2, \ldots, P_n\}$ be the set of n corresponding transition matrices

Sufficient condition

Generalized ergodicity condition: Consider a $\mathcal{D}$ restricted MDP with $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$
Let $\mathcal{A} = \{P_1, P_2, \ldots, P_n\}$ be the set of n corresponding transition matrices
There exists some $\gamma < 1$ and positive integer N such that for all n^N matrix products A of the form $A = \prod_{k=1}^{N} A_k$ with $A_k \in \mathcal{A}$ for $k = 1, 2, \ldots, N$ it holds that $\rho_0(A) \leq \gamma$

Sufficient condition

Generalized ergodicity condition: Consider a $\mathcal{D}$ restricted MDP with $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$
Let $\mathcal{A} = \{P_1, P_2, \ldots, P_n\}$ be the set of n corresponding transition matrices
There exists some $\gamma < 1$ and positive integer N such that for all n^N matrix products A of the form $A = \prod_{k=1}^{N} A_k$ with $A_k \in \mathcal{A}$ for $k = 1, 2, \ldots, N$ it holds that $\rho_0(A) \leq \gamma$

Claim: This generalized ergodicity condition is sufficient

The associated MDP

For optimizing over all $\mathcal{D}$ mixing policies the following associated continuous state space MDP could be considered

The associated MDP

For optimizing over all $\mathcal{D}$ mixing policies the following associated continuous state space MDP could be considered

- State space X is set of all probability distributions on original (finite) state space

The associated MDP

For optimizing over all $\mathcal{D}$ mixing policies the following associated continuous state space MDP could be considered

- State space X is set of all probability distributions on original (finite) state space
- Action space is $\mathcal{D}$ for all $x \in X$

The associated MDP

For optimizing over all $\mathcal{D}$ mixing policies the following associated continuous state space MDP could be considered

- State space X is set of all probability distributions on original (finite) state space
- Action space is $\mathcal{D}$ for all $x \in X$
- For all $x \in X$ and $d^i \in \mathcal{D}$ the immediate reward $r(x, d^i)$ is the inner product $x \cdot r(d^i)$ of x and reward vector $r(d^i)$

The associated MDP

For optimizing over all $\mathcal{D}$ mixing policies the following associated continuous state space MDP could be considered

- State space X is set of all probability distributions on original (finite) state space
- Action space is $\mathcal{D}$ for all $x \in X$
- For all $x \in X$ and $d^i \in \mathcal{D}$ the immediate reward $r(x, d^i)$ is the inner product $x \cdot r(d^i)$ of x and reward vector $r(d^i)$
- For all $d^i \in \mathcal{D}$ state transitions are given by state space mapping $x \to xP_i$

Associated sample paths

Let $(x_1, d_1, x_2, d_2, \ldots)$ be a sample path for the associated MDP. For $n = 1, 2, \ldots$ consider the corresponding $\mathcal{D}$ mixing policy $\pi_n := (d_n, d_{n+1}, \ldots)$

Associated sample paths

Let $(x_1, d_1, x_2, d_2, \ldots)$ be a sample path for the associated MDP. For $n = 1, 2, \ldots$ consider the corresponding $\mathcal{D}$ mixing policy
$\pi_n := (d_n, d_{n+1}, \ldots)$

Claim: If generalized ergodicity condition holds then for all mixing policies π_n, $n = 1, 2, \ldots$ the expected Césaro average reward is equal to the Césaro average reward induced by the given sample path
(which equals $\liminf_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} x_t \cdot r(d_t)$)

Optimal policies

Result: If generalized ergodicity condition holds for $\mathcal{D}$ then there exists some optimal stationary deterministic Markov policy for the associated MDP

Optimal policies

Result: If generalized ergodicity condition holds for $\mathcal{D}$ then there exists some optimal stationary deterministic Markov policy for the associated MDP

Such an optimal policy corresponds to mapping f from X to $\mathcal{D}$

Optimal policies

Result: If generalized ergodicity condition holds for $\mathcal{D}$ then there exists some optimal stationary deterministic Markov policy for the associated MDP

Such an optimal policy corresponds to mapping f from X to $\mathcal{D}$

Corollary: If $(x_1, d_1 = f(x_1), x_2, d_2 = f(x_2), \ldots)$ is a corresponding optimal sample path then for all $n = 1, 2, \ldots$ policy $\pi_n = (d_n, d_{n+1}, \ldots)$ is optimal among all $\mathcal{D}$ mixing policies

Solving an $\mathcal{D}$ restricted problem

Solving the associated continuous state space MDP is usually not tractable

Solving an $\mathcal{D}$ restricted problem

Solving the associated continuous state space MDP is usually not tractable

Optimizing over a subclass of $\mathcal{D}$ mixing policies may be tractable

Solving an $\mathcal{D}$ restricted problem

Solving the associated continuous state space MDP is usually not tractable

Optimizing over a subclass of $\mathcal{D}$ mixing policies may be tractable

Structural properties of optimal stationary policies and corresponding sample paths for the associated MDP could translate to specific structural properties of optimal $\mathcal{D}$ mixing policies

Solving an $\mathcal{D}$ restricted problem

Solving the associated continuous state space MDP is usually not tractable

Optimizing over a subclass of $\mathcal{D}$ mixing policies may be tractable

Structural properties of optimal stationary policies and corresponding sample paths for the associated MDP could translate to specific structural properties of optimal $\mathcal{D}$ mixing policies

For example: non-randomized, periodicity, threshold structures

Implications of threshold structures

Formulation of key result:

Let $I = [0,1]$ and $x_1, x^* \in I$ be given. Let $f_1, f_2 : I \longrightarrow I$ be given functions and $f : I \longrightarrow I$ be defined by $f(x) = \begin{cases} f_1(x) & \text{if } x \leq x^* \\ f_2(x) & \text{if } x > x^* \end{cases}$

Implications of threshold structures

Formulation of key result:

Let $I = [0, 1]$ and $x_1, x^* \in I$ be given. Let $f_1, f_2 : I \to I$ be given functions and $f : I \to I$ be defined by $f(x) = \begin{cases} f_1(x) & \text{if } x \leq x^* \\ f_2(x) & \text{if } x > x^* \end{cases}$

Consecutively for $n = 1, 2, \ldots$ determine u_n and x_{n+1} iteratively by

$u_n := \begin{cases} 0 & \text{if } x_n \leq x^* \\ 1 & \text{if } x_n > x^* \end{cases}$ and $x_{n+1} := f(x_n)$

Key result

Result: Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones generated as above with $f_1, f_2 : I \longrightarrow I$ both monotonically increasing and moreover, $f_1(f_2(x)) \geq f_2(f_1(x))$ for all $x \in I$

Key result

Result: Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones generated as above with $f_1, f_2 : I \to I$ both monotonically increasing and moreover,
$f_1(f_2(x)) \geq f_2(f_1(x))$ for all $x \in I$
Then U eventually coincides with a 0-1 billiard sequence

Conclusion

Result establishes under certain conditions for $\mathcal{D} = \{d^1, d^2\}$ the existence of an optimal mixing policy being representable as billiard sequence

Conclusion

Result establishes under certain conditions for $\mathcal{D} = \{d^1, d^2\}$ the existence of an optimal mixing policy being representable as billiard sequence

Can this result be generalized or other structural properties be obtained?